THE

COMPLETE GUIDE TO STATISTICS

FOR THE SAAS EXECUTIVE

ProfitWell™

01
PART

THE FUTURE IS STATISTICAL

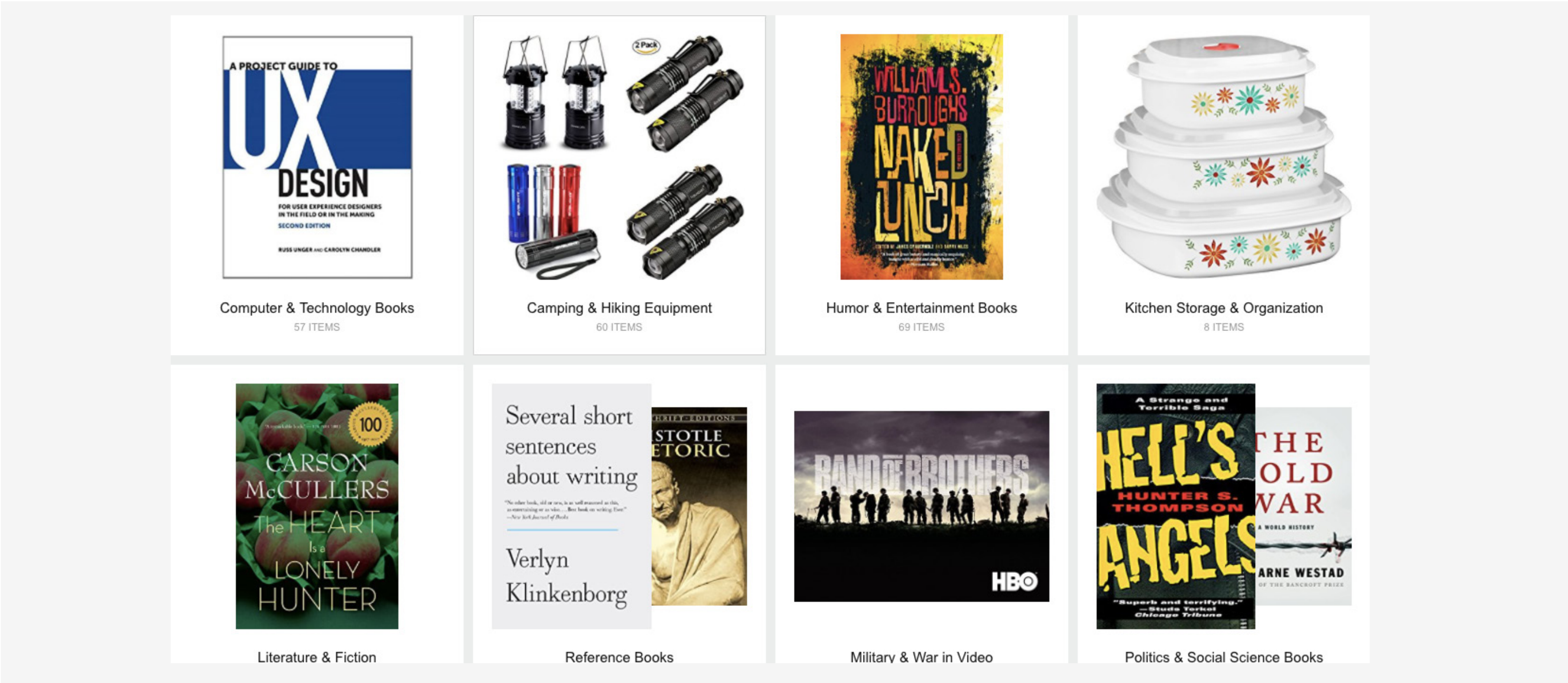
In my time at Google, I only once spoke to Eric Schmidt. But that one conversation changed my life. Talking about the future of mathematics, Schmidt told me,

“The world will be inherited by statisticians”

As a math geek, I was fortified, but his statement was already being proved all around me. Everything in tech is based in math—we are all part of a data machine. Statistics allows us to draw insight from the wealth of data this field is generating.

Stats has been applied to medicine, politics, and finance. Now it’s tech’s turn. Statistics is becoming more and more integral to the companies we run and the products we make. The technologies of the future that are starting to come online now—artificial intelligence, cryptocurrencies, predictive analytics—are all, deep down, based on statistics. They are at their core the regressions, probabilities, and p-values that you learned about in high school.

This is why we are writing this guide. To help SaaS executives get a better understanding of statistics so that they can a) understand their business now, and b) prepare their business for the future. In this first chapter of the complete guide to statistics for the SaaS executive, we want to show you exactly how the world will be inherited by statisticians.



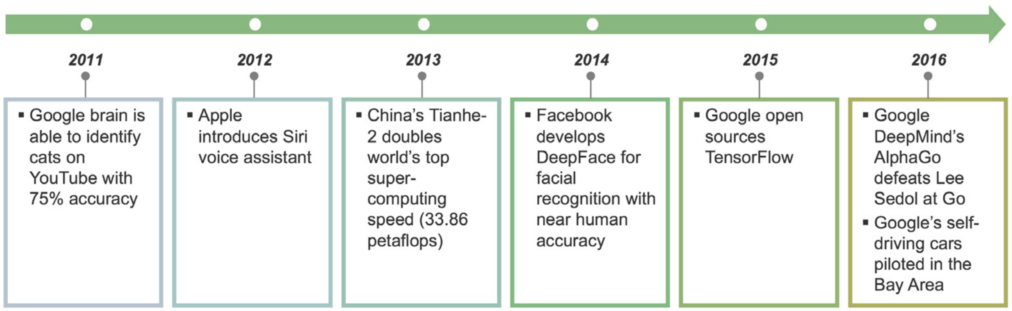
ARTIFICIAL INTELLIGENCE: HOW RECOMMENDERS USE MATH TO KNOW YOUR MIND

Amazon and Netflix seem to have windows to your soul, picking out just the right movie for a rainy day or a book about business you haven't yet heard of.

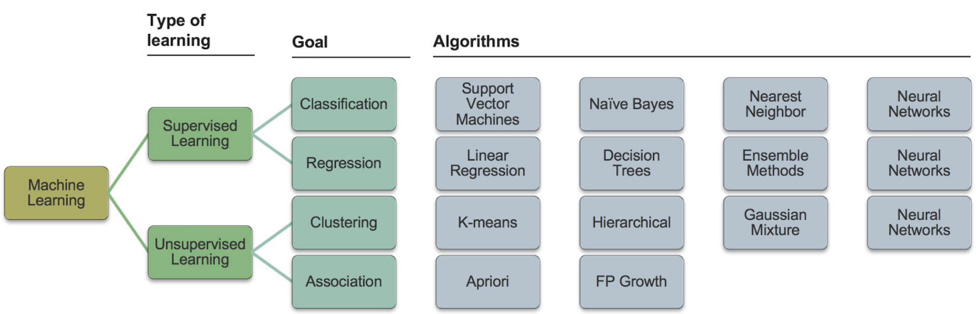
> *How'd they know that?*

They can do this because they do have a window into your soul—your actions. Every action you take on Google, Facebook, YouTube, Amazon, etc. is fed into their artificial intelligence algorithms to learn more about you and offer more of what (they think) you want.

AI isn't only the basis of recommender systems. The natural language processing of Siri or Alexa, the self-driving cars of Google or Tesla, even game-winning abilities of IBM's DeepBlue or DeepMind's AlphaGo—AI is now embedded in a massive range of technologies we already take for granted.



(Source: CBInsights)



(Source: CBInsights)

At the heart of this artificial intelligence is machine learning—the ability of programs to learn from previous data to predict future events. And machine learning, when you break it down, is just very fast, very advanced statistical modeling.

Support vector machines, linear regression, k-means, naive Bayes—all these are statistical techniques that you would learn in a college math class. Even the much-vaunted neural networks effectively work on probability. Amazon's machine-learning recommender is built on a similar statistical concepts, albeit with multiple equations and far more variables. In particular it is built on one of the simplest—regression analysis.

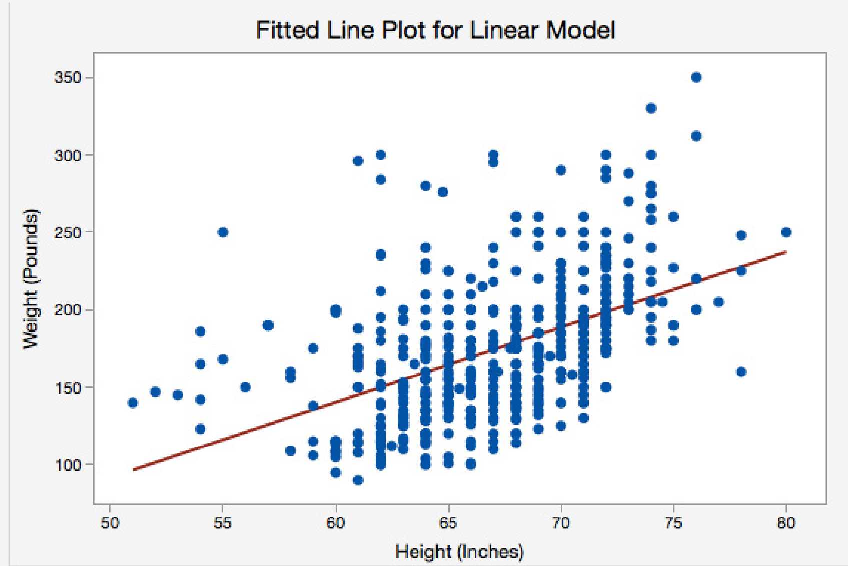
In statistics, regression analysis determines the relationships between points in a data set. The relationship is described by an equation which can be used to predict future outcomes. The equation is graphed to display a curve emerging from a smattering of data. A basic example is the linear correlation between height and weight.

Regression can be used for one independent variable and one or more dependent variables, growing more and more complex as more variables are added. In the height and weight example, weight (y) is dependent on height (x).

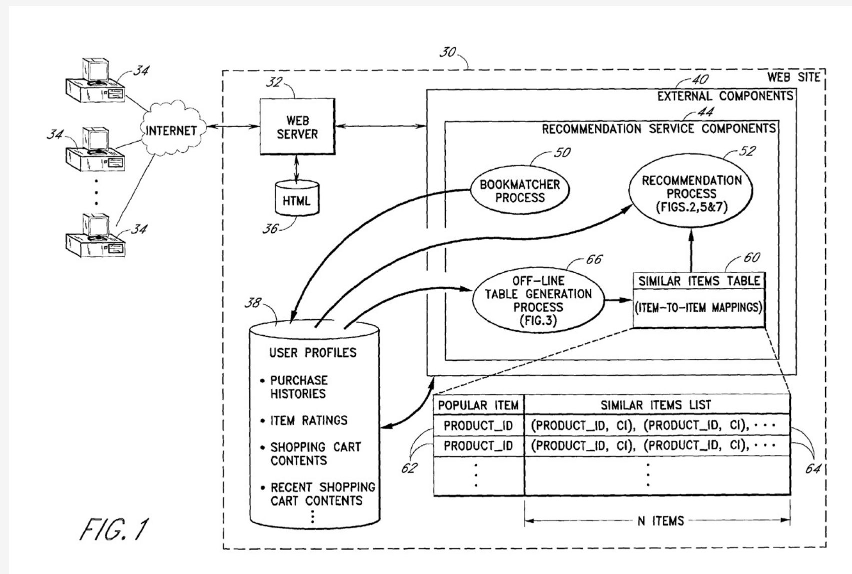
Here's an image from the patent for how Amazon's recommendation technology works:

The vital component here is the "recommendation process." This takes the inputs—user profiles, purchase histories, cart contents—and performs a regression analysis. As more and more variables are added, Amazon can draw patterns from the data and predict what items you might like from the trend curves.

In statistics, there's always a margin for error, and machine learning is all about reducing that margin. So if statistics is the way Amazon knows your purchase of a camping stove indicates you're likely to buy organic toothpaste, machine learning is the way those predictions get more accurate over time.



(source)



(source)

CRYPTOCURRENCIES: BIT-HEADS OR BIT-TAILS?

Cryptocurrencies and the underlying technology of the blockchain have the potential to upend the way sensitive information is stored and transmitted. While blockchain as a system is new, it's built on the backs of preexisting concepts in mathematics:

- **Elliptic Curve Digital Signature Algorithm (ECDSA)** is the main component of the blockchain. This allows people to sign their transactions and allows other people to verify those transactions. It is a complicated concept, but at the fundamental level is related to algebra and geometry.
- **Hash functions** are functions that take a variable length input and allow you to output something of fixed length. They can be designed to be one-way, meaning they are incredibly difficult to reverse. The RIPEMD-160 cryptographic hash function is used in Bitcoin.

Both of these relate to an even more fundamental component of statistics that is crucial to the success of any cryptocurrency: **probability.** Perhaps you've already invested in a fluctuating cryptocurrency or two, experimenting in this brave new world. Without the BoA controlling your money and The Fed backing them up, how do you know your money is safe? Because of statistics.

If you've ever bet on the likelihood of Bitcoin's success, then you've used statistical probability. What you probably didn't realize is that Bitcoin's existence hinges on probabilities *(as well as improbabilities)*, too.

Probability is easy to grasp, and factors into business all the time, from sales forecasts to risk events. You're subconsciously using probability all the time, whenever you say "the odds are..." Probability is the likelihood that a future event will occur. All probabilities are between 0 and 1 and can be expressed as fractions, decimals, and percentages.

The equation for determining the probability of an event A is:

> *P (A) = possible ways A can occur / total number of outcomes*

The simplest example is flipping a *physical* coin. If A is landing on "heads," then P (A) = 1/2 or 50%.

Probability in Bitcoin is not much more complicated than flipping a coin or rolling dice, though the numbers are much larger. Bitcoin wallets contain private keys which are only visible to the owner and are applied to every transaction the owner makes from that wallet. Each transaction also generates a public key, which is visible to bitcoin miners who record and publish transactions. Private keys must be unique. Otherwise, two owners with the same key would withdraw from the same pool of funds. *(source)*

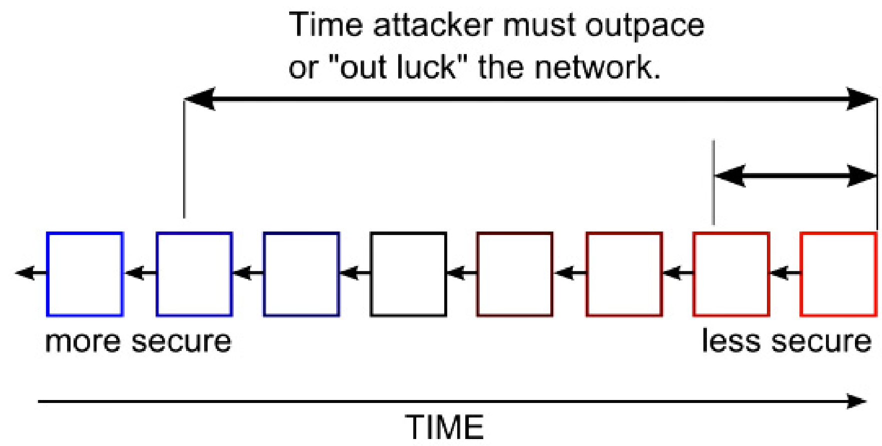
The way bitcoin creators designed it, there are approximately 2^160 private keys possible, making a collision statistically improbable, though not impossible. The odds are, even in a scenario with 1 billion users with 10 wallets each, less than 0.00000000000000000000000000000000000684%.

What if someone wanted to steal your key? Could they just guess it and take all your Bitcoin? Say they had this password cracking machine, capable of 350,000,000,000 guesses per second. It will crack your Windows password in six hours. How long for your Bitcoin private key?

1.32x10^29 years. Considering the life expectancy of the earth is 7.8x10^9 years you can probably say your Bitcoin keys statistically safe.

Probability also determines the verification of transactions through the order by which they are recorded.

Confirmation is the process by which blocks containing records of transactions are added to the blockchain. Once a miner has solved a block, it must be confirmed by the network. Confirmation takes about 10 minutes. This ten-minute figure is based on a probability calculation that a block that has been mined will be found and added. About 2/3 of transactions are confirmed within that 10-minute timeframe, and about 95% are confirmed within 30 minutes.



(source)

Security increases as more blocks are added.

Bitcoin's security is perhaps too reliant on these improbabilities. On the one hand, credit card numbers are a lot easier to guess, making fraud a lot easier with traditional transactions. On the other hand, because banks are centralized, reliable verification of payments is transparent and direct in comparison to Bitcoin's system.

Bitcoin's future remains unclear, but it has uncovered a wealth (maybe) of opportunities to look at current transactional systems and apply our latest knowledge of data to come up with innovative solutions.

PREDICTIVE ANALYTICS: DATA-DRIVEN EVERYTHING

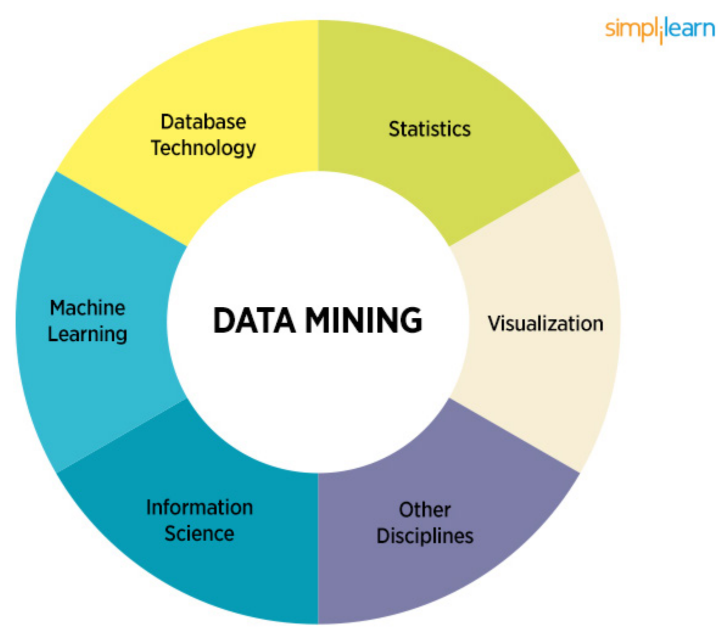
Companies' access to deep troves of data has never been greater. Personal data is used by Starbucks to offer a deal on a customer's preferred espresso drink and by fine-dining restaurants to engage customers in discussing sports. In B2B SaaS, sophisticated analytics platforms inform employees about customer attributes and behaviors, sales performance and growth, and countless other metrics.

Here we can use an example that is closer to home than Amazon's AI or Bitcoin's crypto. MadKudu, a B2B SaaS company specializes in predictive lead scoring. They use the signals coming in from your visitors and correlate that with your most successful customers. They use both behavioral data—the visitor downloaded an ebook or signed up for the freemium product—and demographic data—the visitor is a CMO or comes from a company with 100+ employees.

Using tracking and data enrichment services such as Clearbit this lead scoring can be completely frictionless to an incoming visitor. If they score highly (i.e., their behavioral and demographic traits correlate with success as a customer), then a member of the sales team will be notified and the little Drift box in the corner of the screen will magic into life. If they don't score high, then the sales process becomes no-touch. In this way, MadKudu helps companies minimize their CAC and improve their unit economics. All through statistics - in this case, correlation and data mining.

According to data mining pioneer Alex Zekulin, data mining is "the process of extracting previously unknown, valid, and actionable information from large databases and using it to make crucial business decisions."

In statistics, conclusions are drawn using predetermined models. In data mining, algorithms search for patterns in large, complex databases to form its models. The two very similar practices are



(source)

Data mining is an aggregation of several statistics-based practices. often used in tandem. Data mining is most powerful when multiple databases are combined. SaaS companies are particularly well-positioned to band together for superior data analysis.

MadKudu and Clearbit helped Geckoboard automate its inefficient lead scoring system into an entirely automated and highly accurate prediction machine.

When Geckoboard first employed Madkudu's algorithm, they were able to identify leads, but not the most valuable ones. Geckoboard's marketing lead instructed MadKudu's algorithm to analyze leads by LTV and so the score model adjusted for that, providing them with access to the specific customers they wanted to target.

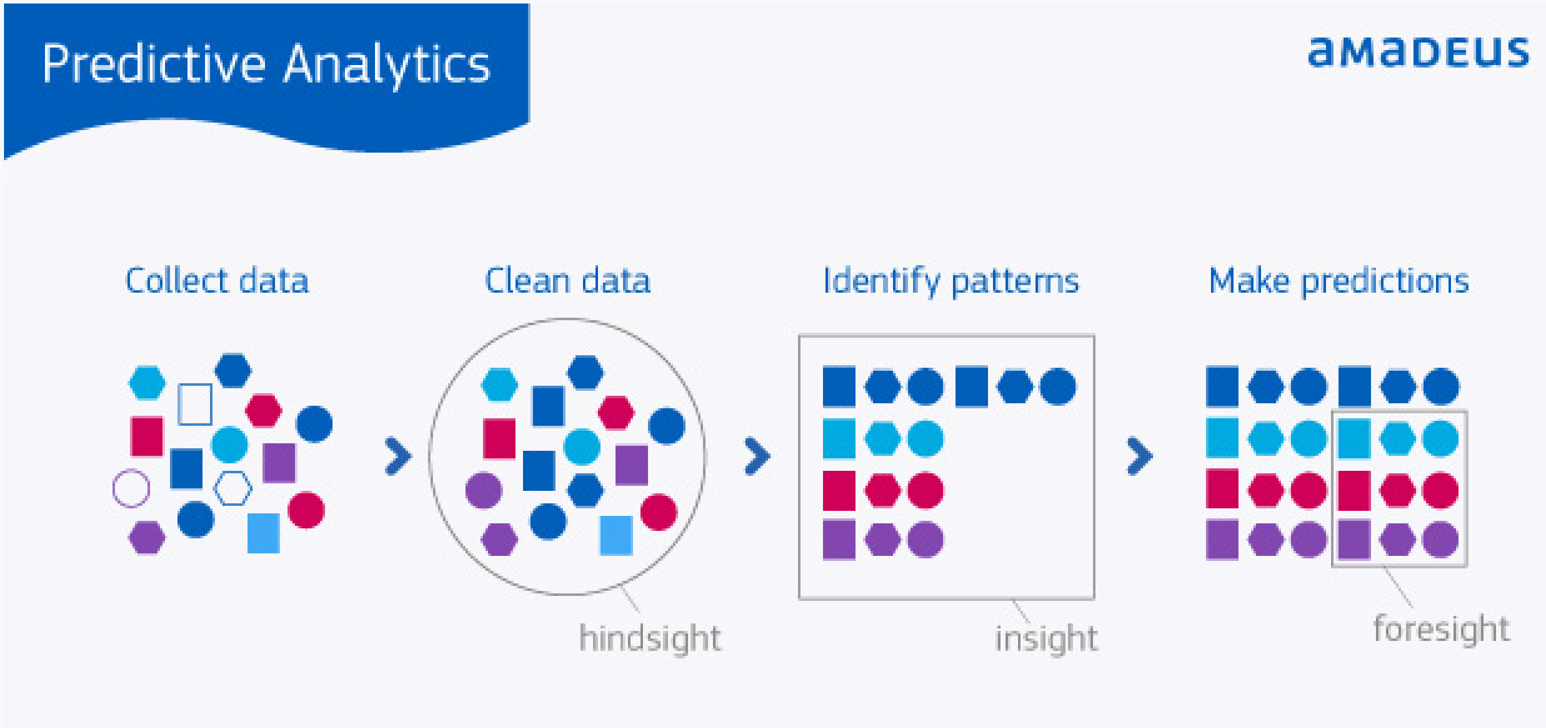
(source). Geckoboard's lead scoring system grew more complex and more accurate with advanced data analysis.

Clearbit gave the sales team information on the customers so they could segment the customers by demographics, industry, and job titles. By the end of their tinkerings, Geckoboard could predict 80% of their conversions from just 12% of their signups automatically.

ADVANCED POINT BASED LEAD SCORING

Including Behavioral Data

CRITERIA		POINTS
CREATED FIRST DASHBOARD		+25
CONNECTED A DATA SOURCE		+50
UPLOADED COMPANY LOGO		+15
NUMBER OF EMPLOYEES	500 ~1000	+40
NUMBER OF EMPLOYEES	50 ~ 500	+25
JOB TITLE	FOUNDER, OWNER, OR DIRECTOR	+10
INDUSTRY	SOFTWARE	+10



Source

In statistics, predictive analysis uses existing models to interpret limited amounts of data to predict future outcomes. For instance, a regression model can be used to predict the probability of a sale or churn. The analysis of several groupings of data can come together to form a larger picture of a day, a person, or a company, and accurately predict what might happen next.

DRESSED UP, BUT STILL STATISTICS

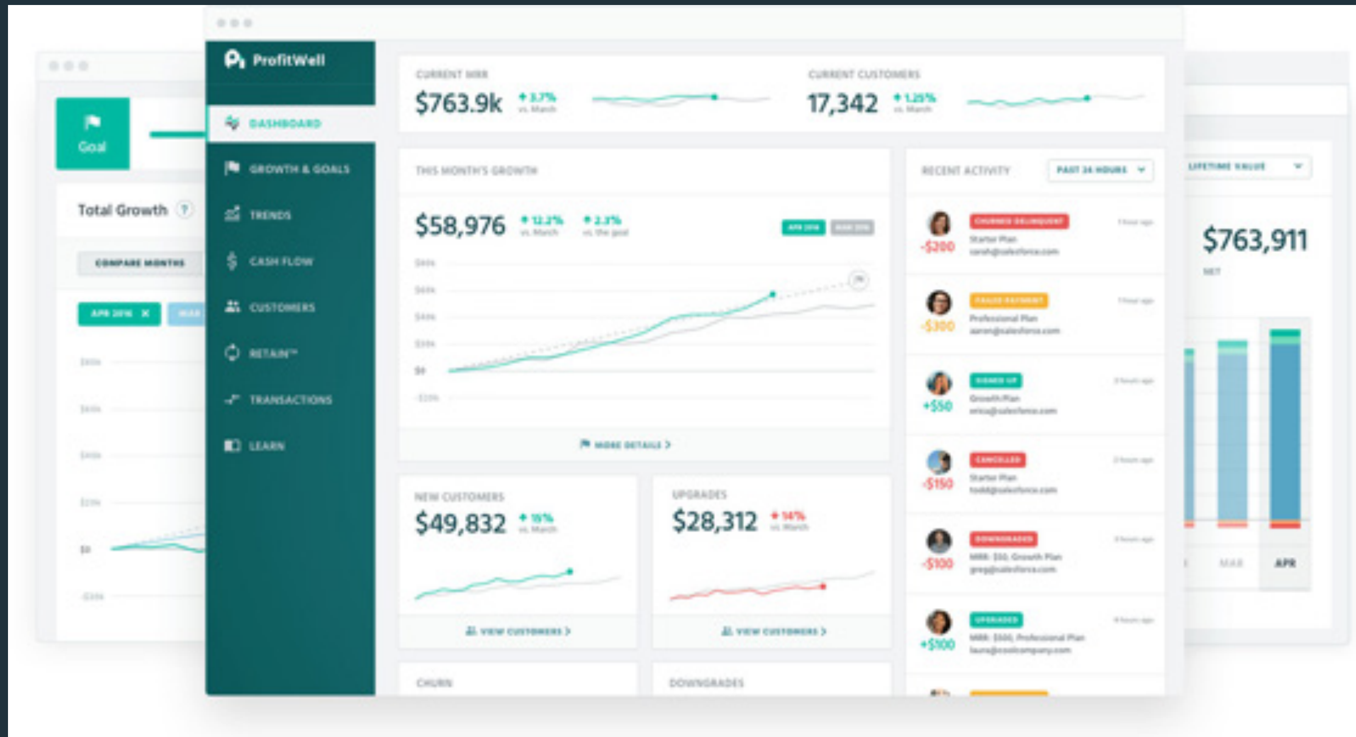
While at a glance, machine learning, bitcoin, and data mining all seem futuristic and fantastically complex, these techniques are just iterations of basic statistics for business. Statisticians will inherit the earth because of their curiosity about data. You're probably already curious about your own company's data, your competitor's data, and the greater trends in the marketplace. With a grasp of the narratives our data is telling us, we can learn more about our customers, our performance, and our future. We can also make the most of cutting-edge analysis and artificial intelligence. First, we have to look a bit closer at the math.

02
PART

THE

STATISTICAL BASICS OF YOUR BUSINESS

02 / 07

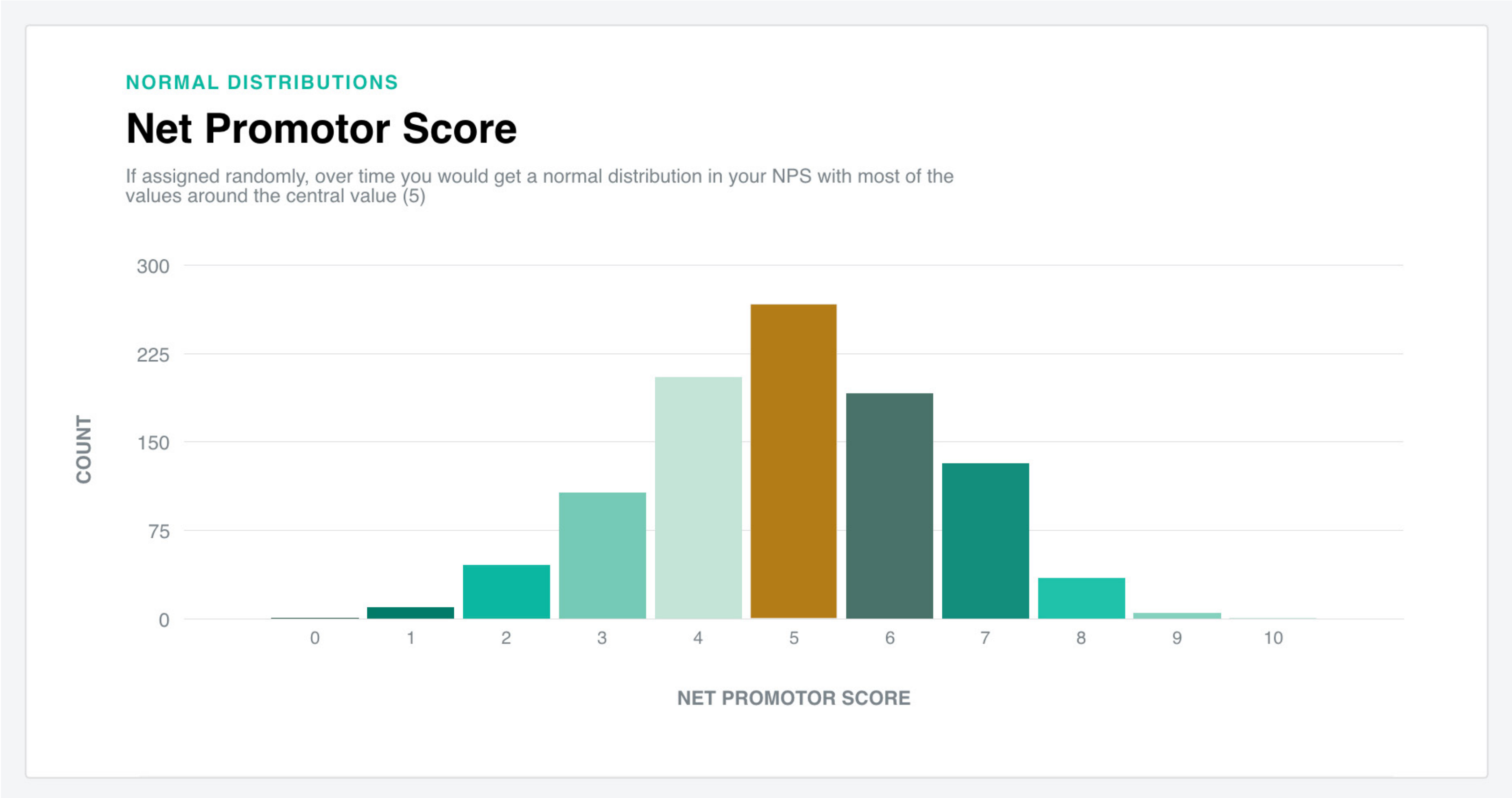


Just having these numbers in front of you is a great start. You can [define your compass metric.](#) **You can track your goals. You can, at a glance, gain a comprehensive understanding of your finances.**

But a lot of people leave these numbers there. They don't realize that with the right analysis, they can unlock massive amounts of information about you, your customers, and the future of your business.

That analysis is statistics. This is why people leave these numbers there. They either don't know or are flat-out frightened of the s-word. They left stats in the classroom and have shied away from its equations ever since. But stats aren't frightening, and aren't hard. With just a little bit of knowledge, you can turn these numbers into insight that can drive your company to further growth.

In this chapter of The ***Complete Guide to Statistics for the SaaS Executive,*** we want to introduce you to the basic components of statistics and start you on the road to using these metrics and other numbers as a basis for a greater understanding of the value of your product to customers. Here we are starting gently with the foundation of statistical understanding—the two most important stats, what's normal and how to describe it.



NORMAL IS HOW NATURE IS DISTRIBUTED, BUT NOT NECESSARILY BUSINESS

One day, because you are a masochist, you add an NPS survey pop-up to your product and get 1,000 responses. How many people score you 10? How many 0? How many right down the middle at 5? To find out, you plot all the responses. You could end up with a graph like this:

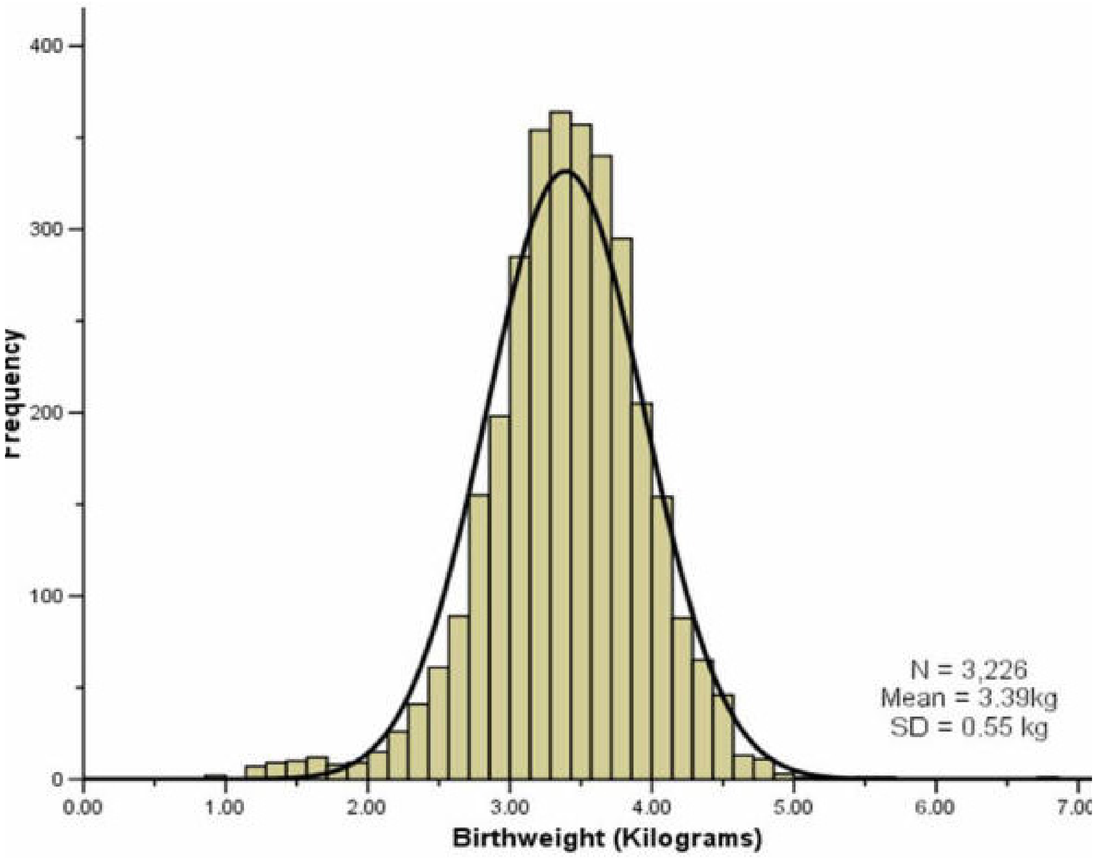
In this **distribution** (spread of data), most people are giving you a middling score. Then the values lower as you get towards the extremes. Most SaaS customer success teams getting these results would look on in horror. But this is one of the most common distributions of data—the normal distribution. This bell-shaped concentration of data is found in a ton of data distributions.

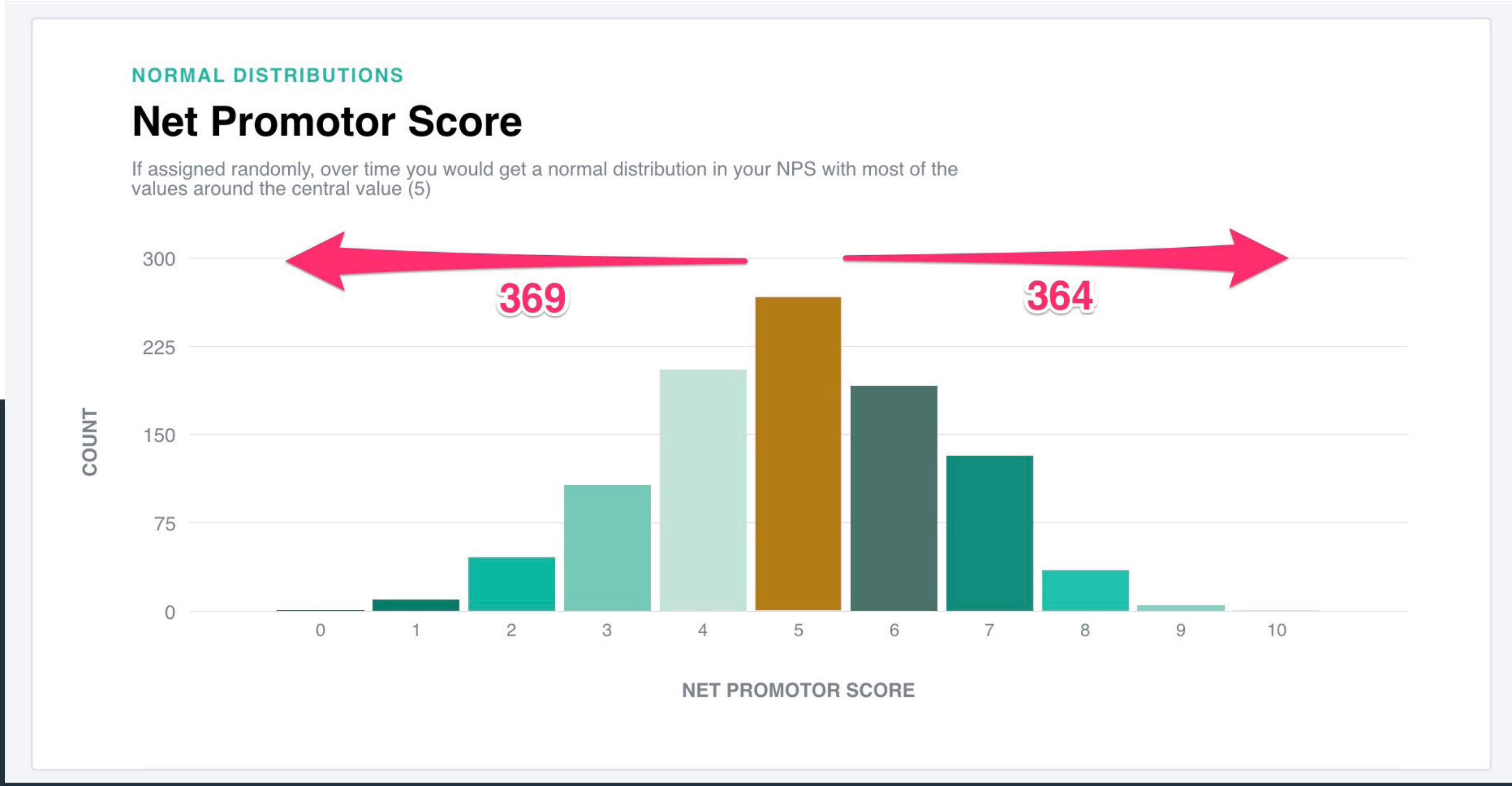
Heights, weights, salaries, test scores—all are normally distributed For instance, here's distribution of birth weight in 3,226 newborn babies:

A normal distribution has two main characteristics:

- It's symmetrical around the center. As much of the data in on the left side as on the right side.
- The mean, median, and mode are all the same.

Let's unpack what that means, starting with symmetry. We can see that the amount of data on either side of the graph is about the same, and that the center of the data lies in the center of the graph:





That center of the graph is also where the **mean**, **median** and **mode** lie.

- The **mean** is the average of all the values.
- The **median** is the middle value when you sort all the numbers.
- The **mode** is value that appears most frequently.

These three values are known as the central values. Mean is the most important from a statistical point of view.

Along with the standard deviation, which we'll come to below, this value is perhaps the most important single value in statistics. It is what every other value is measured by.

In the graph above, the mode is the clearest—5. It is the value that appeared in the NPS survey most frequently. The median is also fairly clear. As there are 1,000 responses in this survey, the median corresponds to the 500th response. Here are the actual numbers for that graph along with the cumulative sum:

	A	B	C
1	NPS Value	Count	Sum
2	0	1	1
3	1	10	11
4	2	46	57
5	3	107	164
6	4	205	369
7	5	267	636
8	6	191	827
9	7	132	959
10	8	35	994
11	9	5	999
12	10	1	1000

Mean is a bit more of challenge. This is where we have to start getting math involved. The equation for the mean is:

$$\bar{X} = \frac{\sum X}{N}$$

The mean (x with a little bar above it) is the sum (the big E, or capital sigma) of all the values (denoted by the X) divided by the number of values (denoted by N). In this case, that comes out to:

$$4.983 = \frac{4983}{1000} = \frac{(0\times1)+(1\times10)+(2\times46)\ldots}{1000}$$

Rounding up to the nearest whole number gives us: 5 again. So we can see that this data satisfies the conditions of a normal distribution.

The normal distribution is integral to statistics. When we try to understand how values are related, we often presume the data is normally distributed. A normal distribution is also described by the two most important values in statistics: **the mean and the standard deviation.** Let's take a closer look at these two numbers.

CENTERING YOUR BUSINESS AROUND STATISTICS

There are basically two branches of statistics:

- With **descriptive statistics** you can organize and summarize your data to get a better understanding of your business, for instance knowing your average revenue or churn.
- With **inferential statistics** you use the data you have to draw a wider conclusion, for instance about what your customers want or the success of your business model.

We'll get to inferences later. Descriptive statistics let you take all your data and transform it into single values that describe your data on the whole.

Think about what data you might commonly have:

- How many customers you have on day X
- Your MRR for month Y
- Your growth in week Z

These are all **univariate time series**. Univariate because only a single number is changing (customers, revenue, growth) and time series because this happens over time (for subscription companies, the unit of time is usually month). Technically, your data is multivariate because most of these numbers change together—the number of customers determines your revenue which determines your growth. But in descriptive statistics, these can be treated as independent to start.

You are already using descriptive statistics. At the very basic this is calculating central values. The most common central value to compute is the mean. So if you've ever calculated your average churn rate, you've used descriptive statistics. Let's say your user churn over the last 12 months is:

	A
1	5.9
2	6.2
3	5.2
4	5.8
5	5.3
6	5.6
7	4.6
8	4.3
9	4.1
10	4.6
11	3.8
12	3.2

The mean is 4.88—all the entries summed, then divided by the number of entries (12, in this case):

$$\bar{X} = \frac{\sum X}{N}$$

Though basic, the mean is such an important statistic that others build on, we wanted to make sure we didn't breeze through it.

The other two central values—**median** and **mode**—are less widely used. But both are important statistics that can help you understand your business.

The **median** is the midpoint value if you ordered from small to large. For the user churn above, the median value is 4.9 *(In an even set like this you take the average of the two middle numbers, 4.6 and 5.2 here).* The **mode** is the most commonly occurring value in the dataset. For this user churn dataset, it is 4.6 *(occurs twice).*

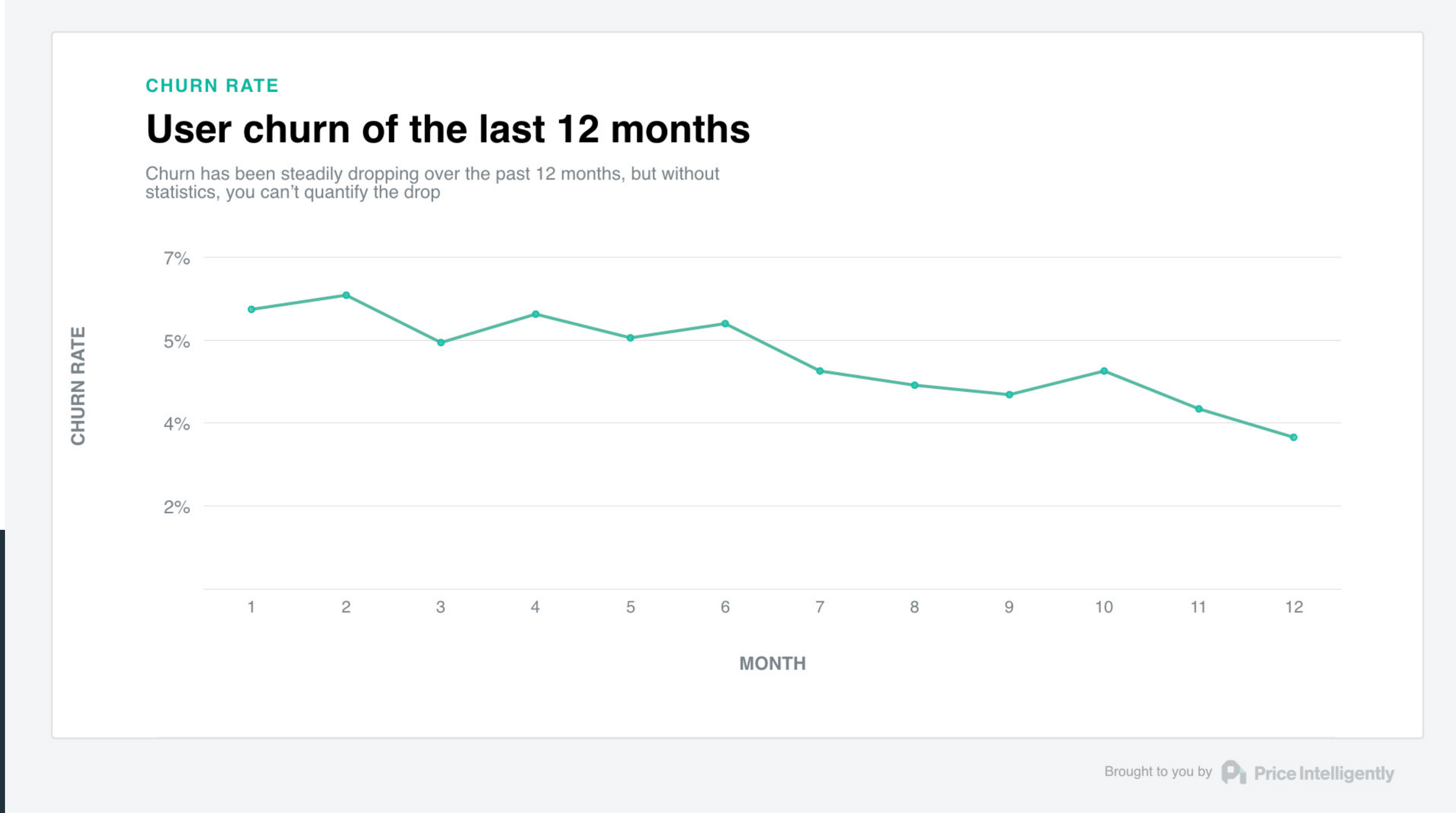
Each of these central values is slightly different, but similar enough that you might wonder why you need all three. This is almost identical to the mean, so why bother with the median?

THE SECRETS TO YOUR BUSINESS ARE THE OUTLIERS

Outliers. Outliers are the most important part of statistics. What you are looking for isn't these central values, but the values that aren't average—the outliers. These are the values that tell you something interesting. But they can also skew your results from the outset. Values that are much smaller or larger than the rest of the dataset can throw off the mean,so it is wise to also compute the median to understand if this is a problem, or when you know you have some massive outlier in the dataset.

Your average churn or customer usage is useful, but not what you are really interested in. What you want to ask your data is who is churning the most? Which customers are using the product the least?

You could do this just by eyeballing—looking at your data for large or small values. This is a great way to start but lacks the rigor needed to do anything with this information.

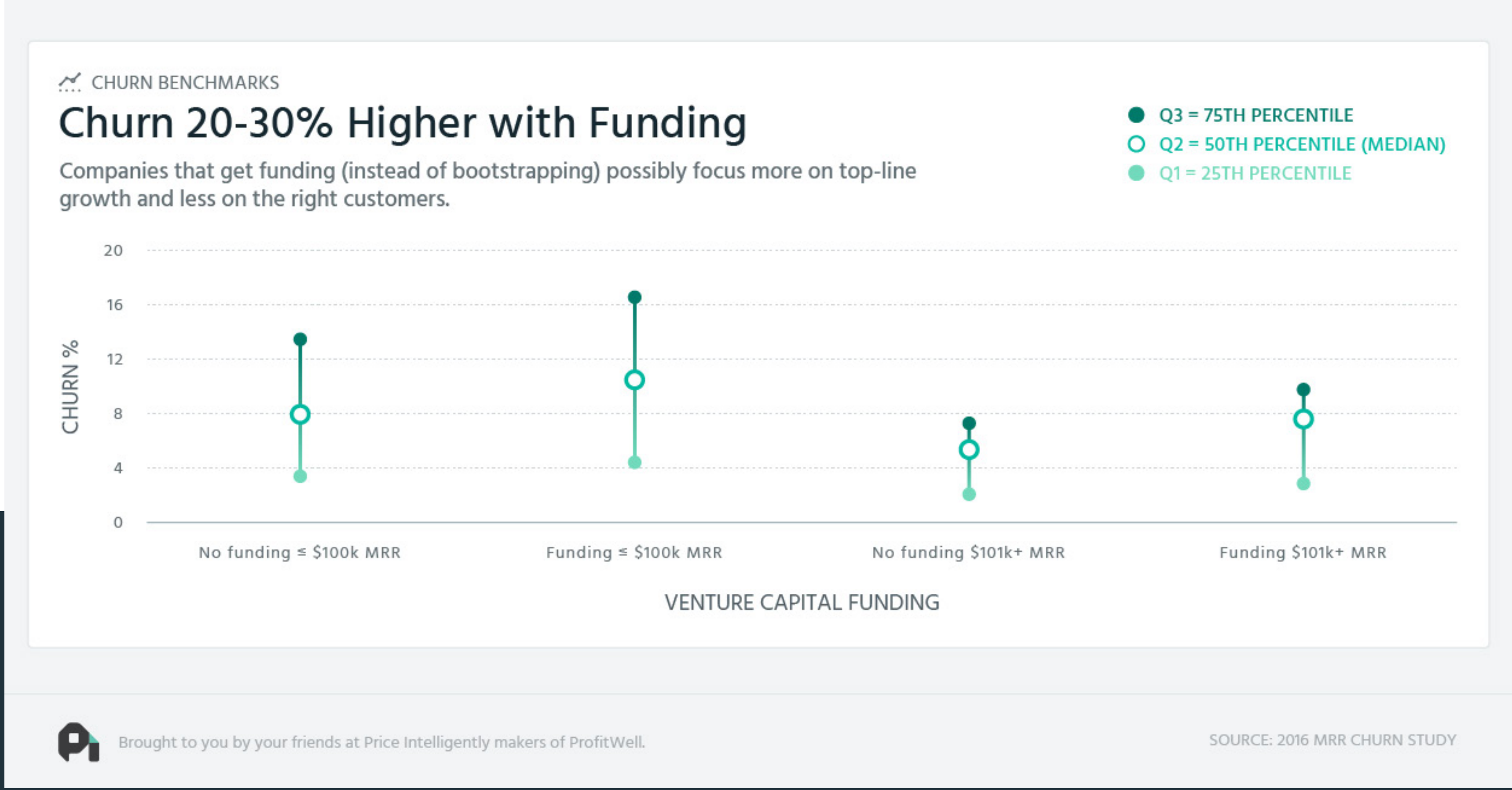


You can see the difference in the numbers, but is that difference substantial? Ultimately you want to find out why these numbers are different and you can't to do that in any systematic way unless you define what is an outlier mathematically.

The most statistically robust way to define your outliers is through standard deviation. **Standard deviation** might seem like a stat that you do not need to work with outside the classroom. But such a simple statistic can come in useful across many different use cases.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

It is a measure of the variance of your data. It tells you how spread out all your numbers are. Loosely this equation asks you to calculate the difference of each value from the mean, square it, add them all together, then divide by the total number of values *(or multiply by one over the total—same thing).*



If you do this for the user churn values above you get a standard deviation of 0.89. So one standard deviation less than the mean is a value of 3.99. One standard deviation above the mean is 5.77. Looking back at the data we can see there are some outliers:

- Months 1, 2 and 4 are all more than one standard deviation above the mean churn rate for the year
- Months 11 and 12 are lower than one standard deviation below the mean churn rate for the year

Immediately from this calculation you have some insight—you are moving in the right direction. The most recent two months are more than one sigma (one standard deviation) away from the mean. The most recent month is almost two sigma away.

You can also find ranges and outliers with the median, though in a less robust way. In our 2016 MRR churn study, we wanted to understand the relationship between churn and funding. For this we used median as it is less sensitive to outliers:

When using median, the limits are defined as some percentage of that central value, usually the 25th and 75th percentile. Any number 25% or less of the median is an outlier, as is any value 75% or more in this graph.

This is a quick and easy way to benchmark different companies. You can easily see whether your churn rate sits outside of these ranges without any difficult calculation.

HAVING CONFIDENCE IN YOUR DATA

But how do we know that any of this is correct? For instance, in the NPS survey, we got responses for 1,000 customers. But what if you actually have 5,000 customers. Then you only have a small sample of your customer base and you don't know if your sample NPS responders are actually representative of all your customers.

To determine whether they are, you need to know how confident you are in your data. For that you need confidence intervals.

Confidence intervals tell you how confident you can be that the true mean of your data, the mean for all your customers, lies within your sample data. To calculate this confidence interval you need:

- the number of samples. N = 1,000
- the mean. mu = 4.98
- the standard deviation. sigma = 2.37
- the Z-value. To be 95% confident, Z = 1.96

You just have to look these up. *(A surprising amount of statistics is just looking things up in tables—see more of this below).* The equation is then:

$$4.98 \pm 1.96 \frac{2.37}{\sqrt{1{,}000}} = 4.98 \pm 0.15$$

This means that you can be 95% sure that the mean for all your customers is between 4.83 and 5.13.

THE SIGNIFICANCE OF YOUR OUTLIERS

These are simple examples. The real power of statistics is when we ramp up to large datasets.

Say you have 10,000 users and want to understand to see if a specific cohort has significantly lower usage. The mean usage for the population is 100 events per month, with a standard deviation of 12. Our test cohort is compromised of 55 customers from a single company who has a mean usage of 96 events. Is this significantly different?

On the face of it, no. It is only a few events short each month. But the important factor is that this isn't a random assortment of customers. They all come from a single company. We want to understand whether this whole group is an outlier from the rest of the customers. For this, we need to use the mean and standard deviation from above to calculate the **standard error** and the z-score.
The z-score is:

$$z = \frac{M - \mu}{SE}$$

Where M is the mean of the test cohort (96 in this case), and u is the mean of the population (100). SE is the standard error, calculated using the standard deviation of the population and the square root of the number in the cohort:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

Putting this all together we get a z-score of -2.47:

$$z = \frac{96 - 100}{1.62} = -2.47$$

third significant figure

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

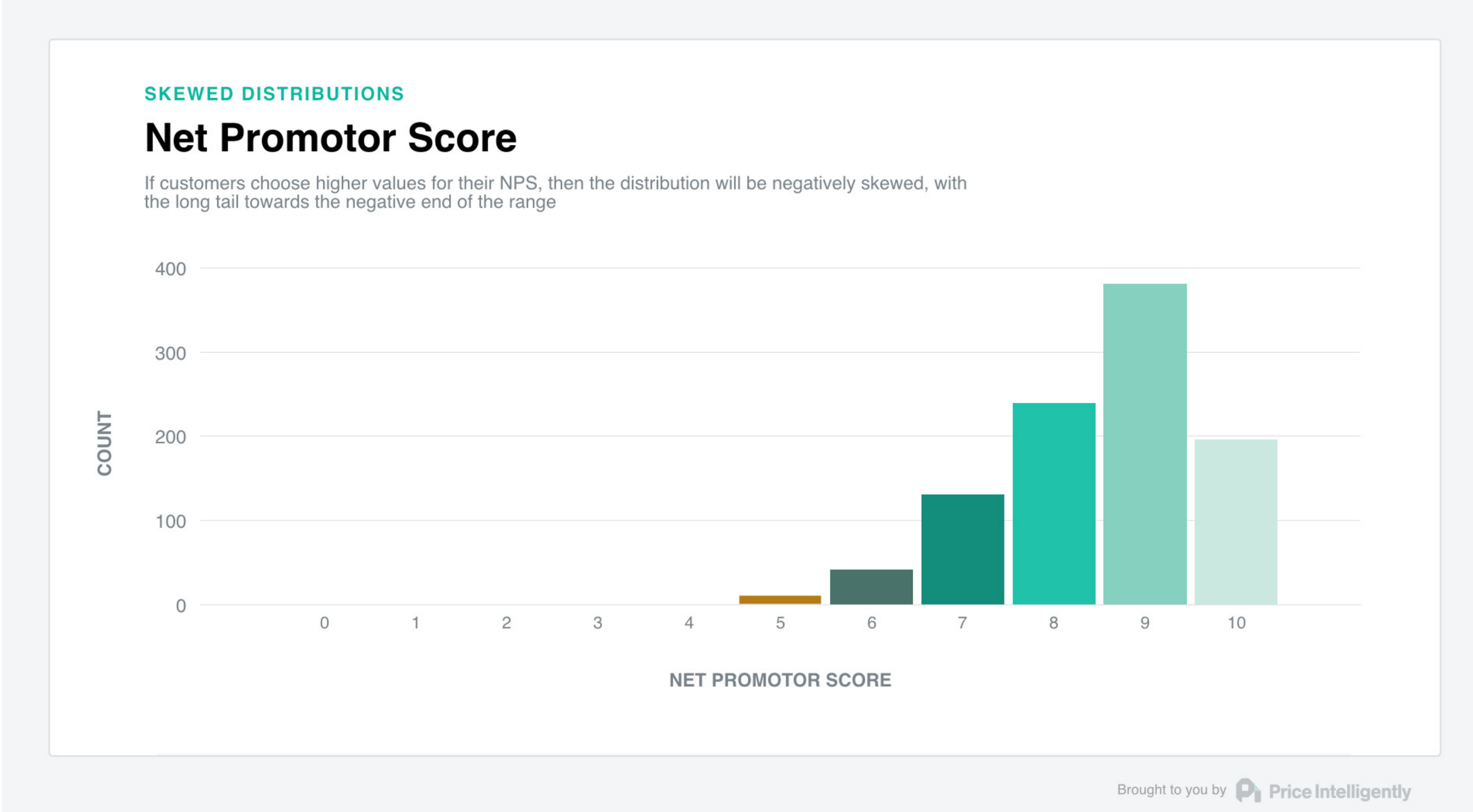
Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	$\boxed{.00676}$	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426

first 2 significant figures

We then use a z-score lookup table to find the one-sided p-value, which is 0.00676: We are presuming customer usage is a normal distribution—it has a bell-shape with most usage around the mean central value. So we need to double that p-value for the two-sided: 0.01352. So the probability that these customers are different from others is:

1 – 0.01352 = 098648 = 98.6%

These customers are statistically different. Though the mean difference is minor, the fact that a group has such a difference compared to a random sample means something is wrong. You know have the mathematical support for deciding what to do with these customers, be it more training or deciding that these are non-ideal customers that you have wrongly acquired.



NON-NORMAL DISTRIBUTIONS

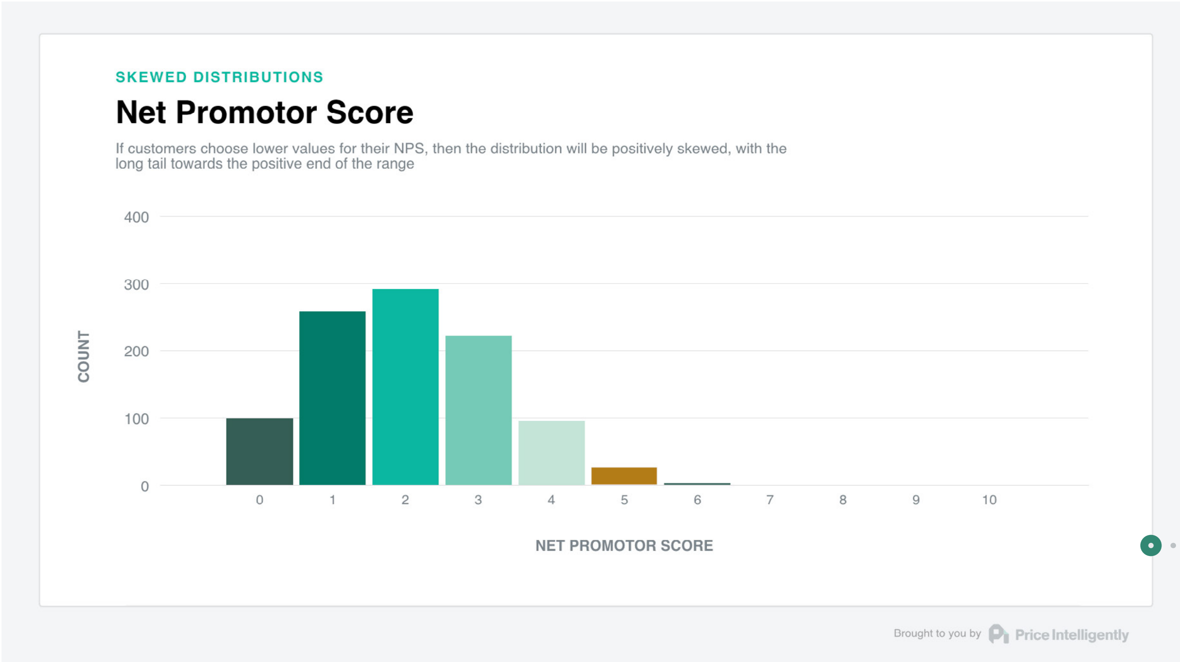
In business, sometimes your data won't be quite this nice and normal. For instance, does the above graph of NPS data ring true?

NPS isn't necessarily normally distributed. Your data could be spread in different ways. In fact, for NPS, you don't really want a bell-shaped curve for your distribution as it means most people are detractors (giving you a score of 6 or below). You want your NPS to be negatively skewed:

Is this a normal distribution? Just by eyeball you can tell it violates one of the rules for normality: symmetry. The central values for this graph are all similar though:

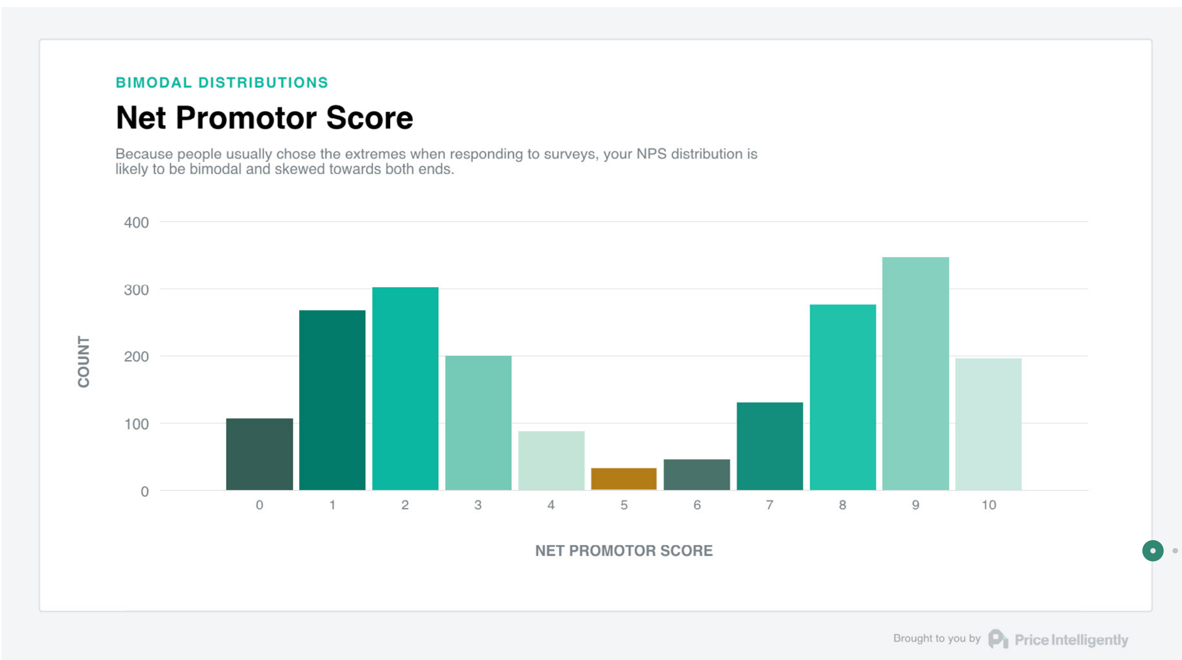
- **Mean: 8.525 (round up to 9)**
- **Median: 9**
- **Mode: 9**

Distributions can be skewed the other way as well, to be positively skewed:



Again, this lacks the symmetry needed for a normal distribution, but the mean, median, and mode are all ~2.

In reality, most products suffer from extreme responding—customers either love you or they hate you. So you end up with this type of distribution:

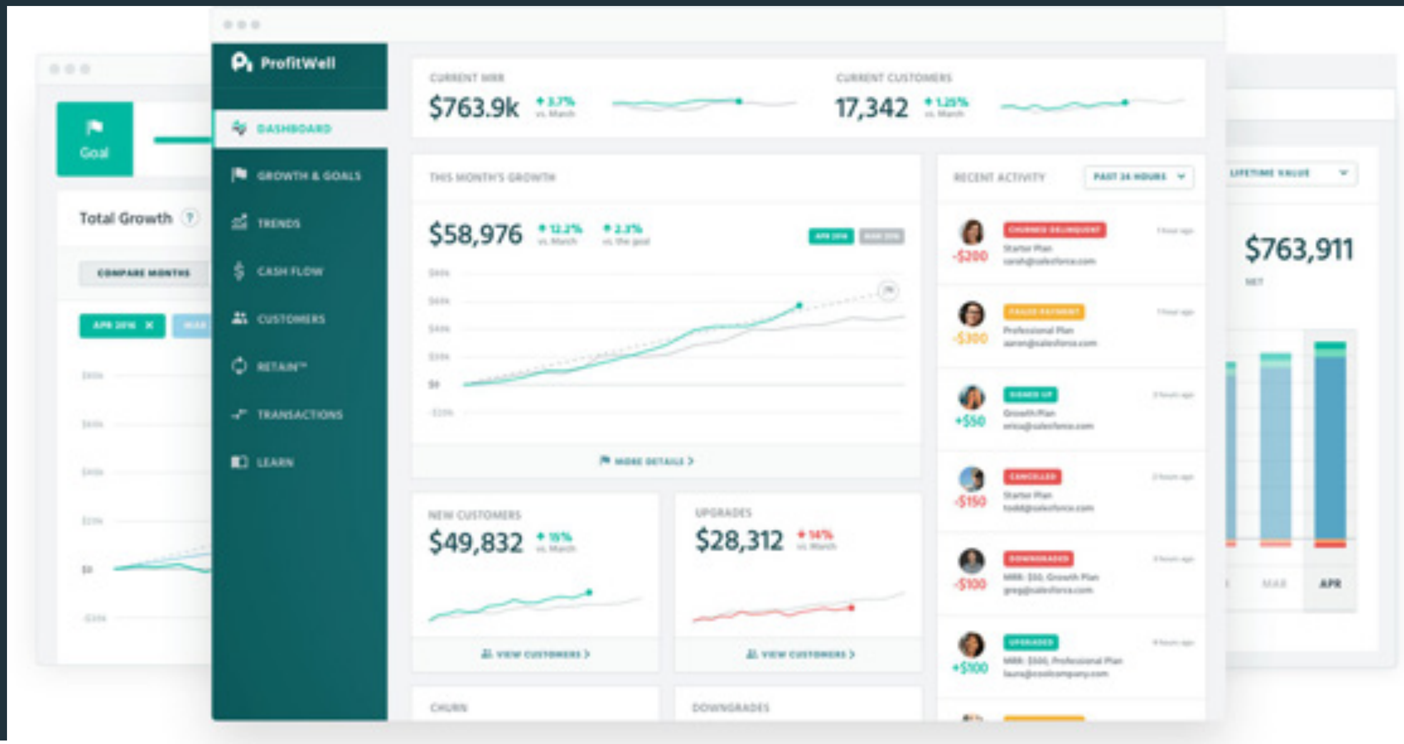


Hopefully not this bad, but something like this. In business this is good—you can learn from people on both ends about how to be better.

A BASIC UNDERSTANDING OF YOUR BUSINESS

In this series we will get to more sophisticated statistical methods. But just because you can use hardcore math doesn't mean you should. Often calculating the basic numbers is where you want to start. Looking at central values and deviations from them is the first step to understanding what your data means and whether there is something interesting in the outliers.

Even with such simple stats you can start to offer more value to your customers. If you see some are outliers, struggling with your product, you can prove this with statistics. This then gives you the authority to go and see what is wrong and start to help make their experience with you better and shifting the entire central value of your product up.



Every day you open up ProfitWell and are confronted with a wealth of data. MRR, churn, retention, cash flow, growth. All these numbers at your fingertips.

Just having these numbers in front of you is a great start. You can [define your compass metric](). You can, at a glance, gain a comprehensive understanding of your business and finances.

But they are all begging for analysis. A lot of people freeze at this point. They either don't think they can do anything or don't know what to do with these numbers to better understand their business.

The answer lies in statistics. In this chapter of The ***Complete Guide to Statistics for the SaaS Executive*** we want to introduce you to the basic components of statistics and start you on the road to using these metrics and other numbers as a basis for a greater understanding of the value of your product to customers. Here we are starting gently with the foundation of statistical understanding—descriptive statistics.

CENTERING YOUR BUSINESS AROUND STATISTICS

There are basically two branches of statistics:

- With descriptive statistics you can organize and summarize your data to get a better understanding of your business, for instance knowing your average revenue or churn.
- With inferential statistics you use the data you have to draw a wider conclusion, for instance about what your customers want or the success of your business model.

We'll get to inferences later. Descriptive statistics let you take all your data and transform it into single values that describe your data on the whole.

Think about what data you might commonly have:

- How many customers you have on day X
- Your MRR for month Y
- Your growth in week Z

These are all univariate time series. Univariate because only a single number is changing (customers, revenue, growth) and time series because this happens over time (for subscription companies, the unit of time is usually month). Technically, your data is multivariate because most of these numbers change together—the number of customers determines your revenue which determines your growth. But in descriptive statistics, these can be treated as independent to start.

	A
1	*5.9*
2	*6.2*
3	*5.2*
4	*5.8*
5	*5.3*
6	*5.6*
7	*4.6*
8	*4.3*
9	*4.1*
10	*4.6*
11	*3.8*
12	*3.2*

The mean is 4.88—all the entries summed, then divided by the number of entries (12, in this case):

$$\overline{X} = \frac{\sum X}{N}$$

Though basic, the mean is such an important statistic that other build on, we wanted to make sure we didn't breeze through it.

The other two central values—median and mode—are less widely used. But both are important statistics that can help you understand your business.

The median is the midpoint value if you ordered from small to large. For the user churn above, the median value is 4.9 (In an even set like this you take the average of the two middle numbers, 4.6 and 5.2 here). The mode is the most commonly occurring value in the dataset. For this user churn dataset, it is 4.6 (occurs twice).

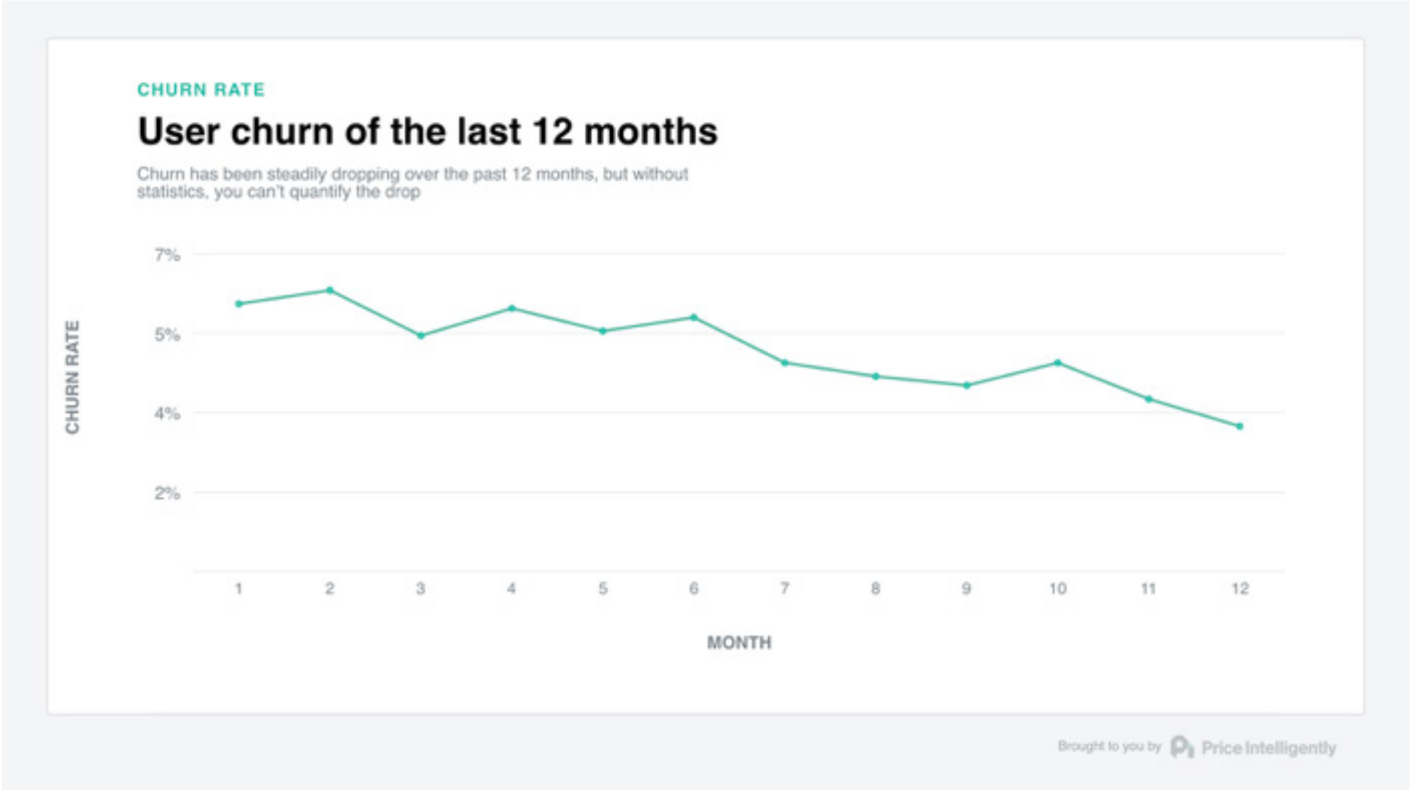
Each of these central values is slightly different, but similar enough that you might wonder why you need all three. This is almost identical to the mean, so why bother with the median?

THE SECRETS TO YOUR BUSINESS ARE THE OUTLIERS

Outliers. Outliers are the most important part of statistics. What you are looking for isn't these central values, but the values that aren't average—the outliers. These are the values that tell you something interesting. But they can also skew your results from the outset. Values that are much smaller or larger than the rest of the dataset can throw off the mean, so it is wise to also compute the median to understand if this is a problem, or when you know you have some massive outlier in the dataset.

Your average churn or customer usage is useful, but not what you are really interested in. What you want to ask your data is who is churning the most? Which customers are using the product the least?

You could do this just by eyeballing—looking at your data for large or small values. This is a great way to start but lacks the rigor needed to do anything with this information.

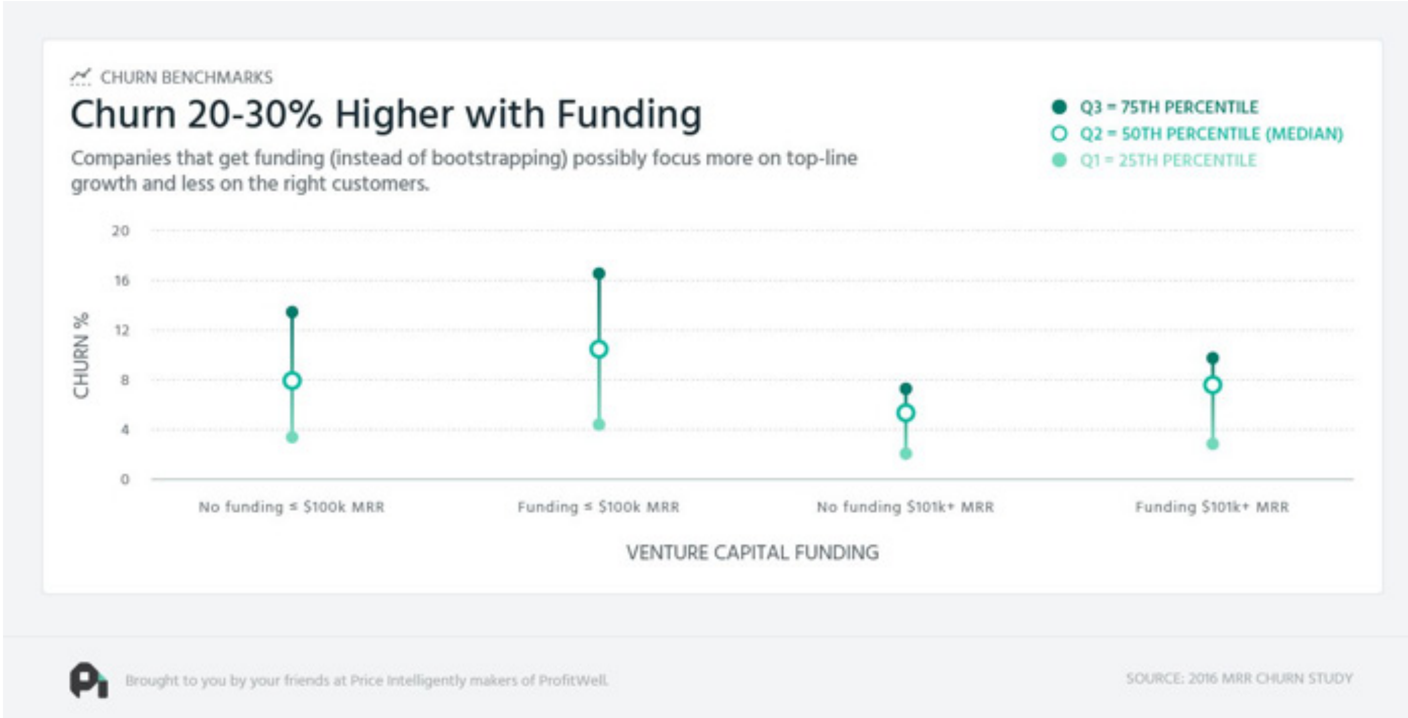


You can see the difference in the numbers, but is that difference substantial? Ultimately you want to find out why these numbers are different and you can't to do that in any systematic way unless you define what is an outlier mathematically.

The most statistically robust way to define your outliers is through standard deviation. Standard deviation might seem like a stat that you do not need to work with outside the classroom. But such a simple statistic can come in useful across many different use cases.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

It is a measure of the variance of your data. It tells you how spread out all your numbers are. Loosely this equation asks you to calculate the difference of each value from the mean, square it, add them all together, then divide by the total number of values (or multiply by one over the total—same thing).



If you do this for the user churn values above you get a standard deviation of 0.89. So one standard deviation less than the mean is a value of 3.99. One standard deviation above the mean is 5.77. Looking back at the data we can see there are some outliers:

- Months 1, 2 and 4 are all more than one standard deviation above the mean churn rate for the year
- Months 11 and 12 are lower than one standard deviation below the mean churn rate for the year

Immediately from this calculation you have some insight—you are moving in the right direction. The most recent two months are more than one sigma (one standard deviation) away from the mean. The most recent month is almost two sigma away.

You can also find ranges and outliers with the median, though in a less robust way. In our 2016 MRR churn study, we wanted to understand the relationship between churn and funding. For this we used median as it is less sensitive to outliers:

When using median, the limits are defined as some percentage of that central value, usually the 25th and 75th percentile. Any number 25% or less of the median is an outlier, as is any value 75% or more in this graph.

This is a quick and easy way to benchmark different companies. You can easily see whether your churn rate sits outside of these ranges without any difficult calculation.

THE SIGNIFICANCE OF YOUR OUTLIERS

These are simple examples. The real power of statistics is when we ramp up to large datasets.

Say you have 10,000 users and want to understand to see if a specific cohort has significantly lower usage. The mean usage for the population is 100 events per month, with a standard deviation of 12. Our test cohort is compromised of 55 customers from a single company who has a mean usage of 96 events. Is this significantly different?

On the face of it, no. It is only a few events short each month. But the important factor is that this isn't a random assortment of customers. They all come from a single company. We want to understand whether this whole group is an outlier from the rest of the customers. For this, we need to use the mean and standard deviation from above to calculate the **standard error** and the z-score. The z-score is:

$$z = \frac{M - \mu}{SE}$$

where M is the mean of the test cohort (96 in this case), and u is the mean of the population (100). SE is the standard error, calculated using the standard deviation of the population and the square root of the number in the cohort: Putting this all together we get a z-score of -2.47:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

$$z = \frac{96 - 100}{1.62} = -2.47$$

third significant figure

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426

first 2 significant figures

A BASIC UNDERSTANDING OF YOUR BUSINESS

In this series we will get to more sophisticated statistical methods. But just because you can use hardcore math doesn't mean you should. Often calculating the basic numbers is where you want to start. Looking at central values and deviations from them is the first step to understanding what your data means and whether there is something interesting in the outliers.

Even with such simple stats you can start to offer more value to your customers. If you see some are outliers, struggling with your product, you can prove this with statistics. This then gives you the authority to go and see what is wrong and start to help make their experience with you better and shifting the entire central value of your product up.

We are presuming customer usage is a normal distribution—it has a bell-shape with most usage around the mean central value. So we need to double that p-value for the two-sided: 0.01352. So the probability that these customers are different from others is:

$$1 - 0.01352 = 0.98648 = 98.6\%$$

03
PART

UNDERSTANDING TRENDS IN YOUR DATA

03 / 07

Up and to the right. For most, that is the only trend that matters in business. You want to see users, customers, and revenue grow over time.

But this isn't the only trend that governs your business. There is a delicate balance between all your metrics that needs to be understood. Appreciating the relationships between all your data allows you to not only know where your business is now, but gives you the foundation to predict the future, and change the future, if you need.

In this chapter of ***The Complete Guide to Statistics for the SaaS*** Executive we are going to show you the statistical basis for this understanding. By visualizing, correlating, and fitting trend lines to your data you can both see and explain the connection between your metrics and use these numbers as the bedrock for growth.

THE RELATIONSHIP BETWEEN YOUR X AND Y

In the previous chapter, we looked at univariate data—data with just one variable. But you don't have just a single variable in your company. All your numbers are interlinked. This is why you need to move beyond univariate analysis into multivariate analysis.

The simplest multivariate analysis is bivariate, where you look at the relationship between two variables. You already have an intuitive sense of how most of these relationships work:

- What is the relationship between ARPU and churn? If you have higher ARPU, you likely have longer contracts and sales and success teams to help customers. This lowers churn.
- What is the relationship between MRR and CAC? If you are pulling in more revenue per month, you can afford to spend more to acquire new customers.
- What is the relationship between MRR and time? Over time, you want to grow, generating more recurring revenue each month.

This last relationship will serve as our example as it is what people care about the most—up and to the right. Here is a quick model:

	A	B
1	Time (Months)	Revenue ($)
2	1	1,000
3	2	5,000
4	3	5,000
5	4	4,000
6	5	7,000
7	6	8,000
8	7	9,000
9	8	8,000
10	9	9,000
11	10	9,000
12	11	10,000
13	12	11,000

When looking at relationships between variables you want to start by plotting the data to visualize any underlying trend. In this case we'll use time as our independent variable (the x-axis) and revenue as our dependent variable (the y-axis) and create a scatterplot of the data:

Just by visualizing you can see the correlation—up and to the right. As time increases, so does revenue. But, like with the previous chapter, seeing isn't always believing.
You need mathematical proof of the trend if you want to experiment with your data. You want to know: how strong is this correlation?

To do this you can compute Pearson's r. Pearson's r, ρ (rho), or the Pearson correlation coefficient allows you to quantify the direction and strength of linear correlation with a single number:

$$\rho = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}}$$

For the above data this resolves to:

$$0.93 = \frac{12\times665{,}000 - 78\times86{,}000}{\sqrt{[12\times650-78^2][12\times708{,}000{,}000-86{,}000^2]}}$$

01

However, there is an easier way to calculate this number, using the z-scores from the previous chapter:

$$\rho = \frac{\sum Z_x Z_y}{n - 1}$$

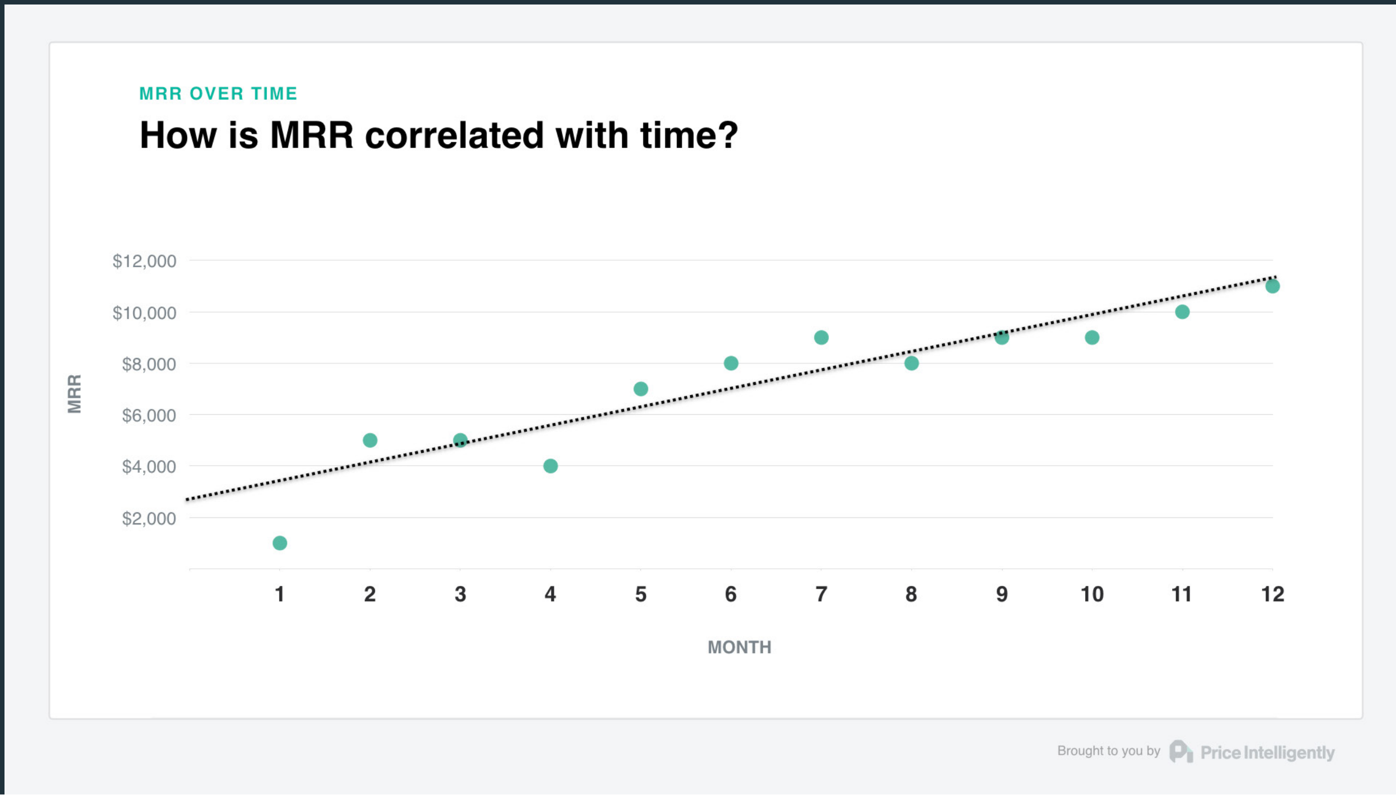
02

Here, for every data point we multiply the z-scores for the x (time) and y (revenue), then sum all of these together and divide by the number of entries minus one. Again, for this data this resolves to:

$$0.93 = \frac{10.18}{11}$$

03

The Pearson's r can be anywhere between -1, a perfect negative correlation and +1, a perfect positive correlation. An r = 0.93 shows that there is a strong positive correlation between time and revenue in this example. As time increases so does revenue.
An r = -0.93 would look not so good:



This shows a strong negative correlation where as time increases, revenue decreases. Growth = a positive Pearson's coefficient.

At this point, the correlation shows you that there is a strong relationship between time and revenue. It can even tell you the direction. But it doesn't really tell you how they are related. For that, we need to move beyond correlations and into linear algebra.

FINDING WHERE YOUR DATA IS TAKING YOU

The correlation coefficient tells you that the data are related. But you also want to know how they are related.

You can find the answer through regression analysis. The simplest version of regression is linear regression:

$$\hat{y} = a + bx$$

This is the equation Excel/Google Sheets uses when you ask them to fit a "Trendline" to your graph. Using this `equation we can fit one to the scatterplot above:

Any point on this line equals the point that the line crosses the y axis (a) plus the slope multiplied by the x value. This is the line of best fit, and is the line that you can draw through your data that minimizes the residuals—the difference between the true values of y (the points in the scatterplot) and the predicted values of y (the y with the little hat, known as y-hat).

You could imagine looking for every single line that could go through that data, working out the residuals and slowly working towards the line that minimizes that number. That's tedious. Luckily there is a much quicker way to work out a and b in the above equation. And because you have already calculated Pearson's r value for this data and you know from chapter 2 how to calculate the standard deviation for each variable, you already have the tools to solve this equation.

1. First, calculate b, using r and the standard deviations of each variable, x and y:

2. Solved for this data, we get a b value of 745.8:

3. This gives us the slope of the line. We should actually write this as $745.8. What this means is that for every step you take along the x-axis, each month, you increase your revenue by $745.8 linearly. We have the slope, but we don't yet know where the line starts, so we need to solve for a:

01 $$b = r\left(\frac{\sigma_y}{\sigma_x}\right)$$

02 $$745.8 = 0.93\left(\frac{2886.8}{3.6}\right)$$

03 $$a = \bar{y} - b\bar{x}$$

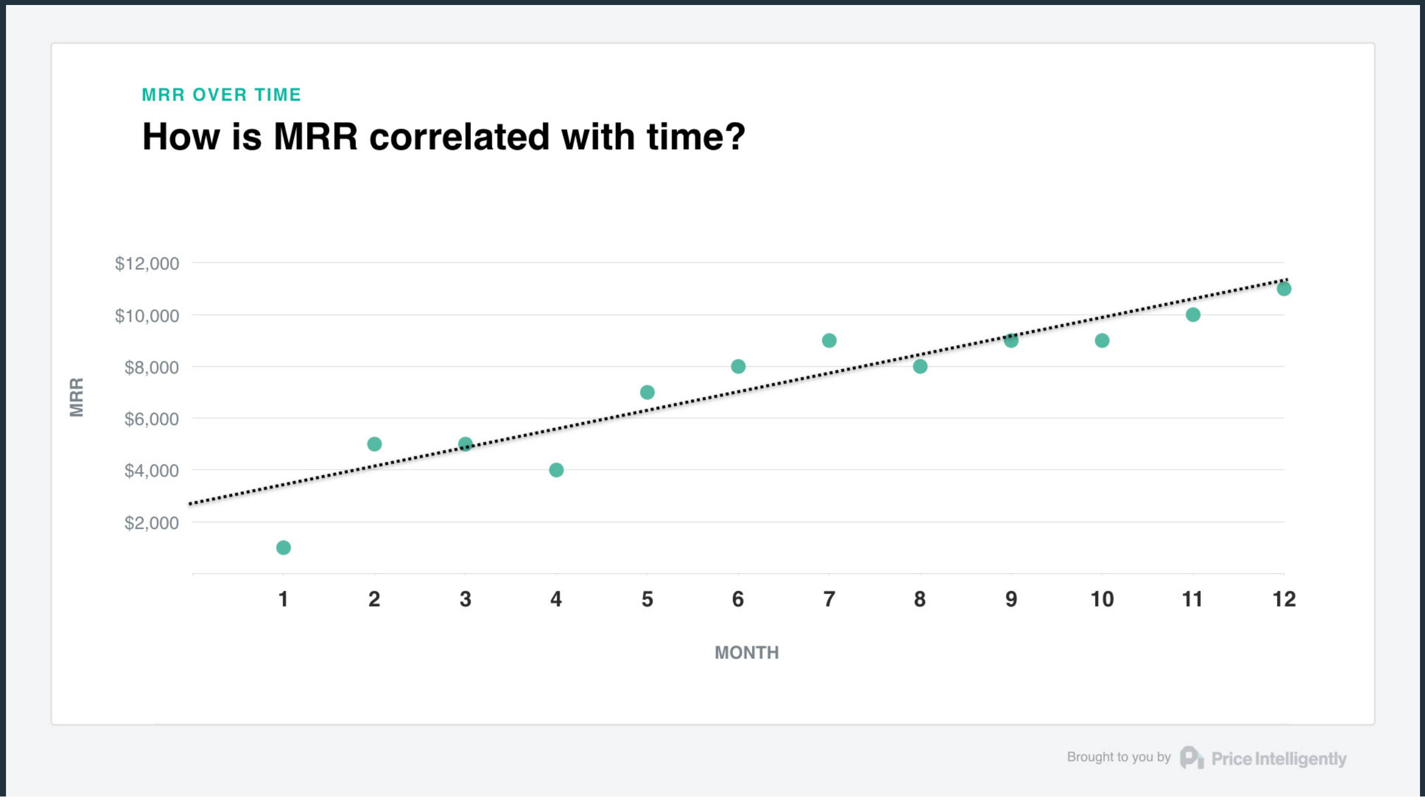
Here, we use the mean values of y and x along with the b value we just calculated to find the y-intercept:

$$2319 = 7166.7 - (745.8 \times 6.5)$$

We want to start our line at $2,319, in the graph above. Putting all this together we can resolve the linear equation above:

$$\hat{y} = 2319 + 745.8x$$

Now we can look back at the trendline we fit to our data with better intuition:



You can see the line starting at $2,319 and then going up ~$746 each month. A couple things should immediately spring out:

- According to this, you were making $2,319 in month 0. Congrats!
- None of the actual data points sit on the line. You made $9,000 in month 7, but the math says you made ~$7,500.

All models are an approximation. This line doesn't exactly describe the data. There is a nonlinear polynomial equation that will fit this data perfectly, but the equation would be insanely complicated and would be of no use inferentially. But this line fits well.

There is another bonus to adding a line of best fit to your scatterplot—it allows you to visualize outliers easier. In the previous chapter we talked about the importance of outliers to your business. Anything away from the mean is normally something very good or something very bad.

From the above chart we can now see clearly the "good" months and the "bad" months. Months one and four fall below the best fit line, so you can go back at look at other data from those months to see why MRR struggled. In this case, it could be something as simple as the company finding its feet in the first month. We can also see two other trends:

- Months five through seven were "good" months. What was so successful about these months?
- Months 10 through twelve have seen MRR dip a little again. Is this the beginning of the end for your customer lifecycle?

Simply be plotting this graph and quickly computing the linear regression line you have visual and mathematically confirmation that you might be in trouble in recent months. You can now deep-dive into these months, find causes, and push MRR back on track.

Measuring how far outlier lie is done by calculating the R-squared value:

$$R^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \overline{y})^2}$$

The numerator here is the sum of squares of residuals. Remember, residuals are the distances between the actual Y values and the predicted Y values (the y with the hat) on the best fit line. The denominator is the total sum of squares, which is the sum of the squares between the actual Y values and their mean (the y with the bar). The R-squared is 1 minus this ratio. For this data we get:

$$0.86 = 1 - \frac{13{,}093{,}240.09}{91{,}666{,}666.67}$$

GOING WHERE THE DATA TAKES YOU

Now that we have the basics of linear regression and correlation, we can start to expand this in two ways:

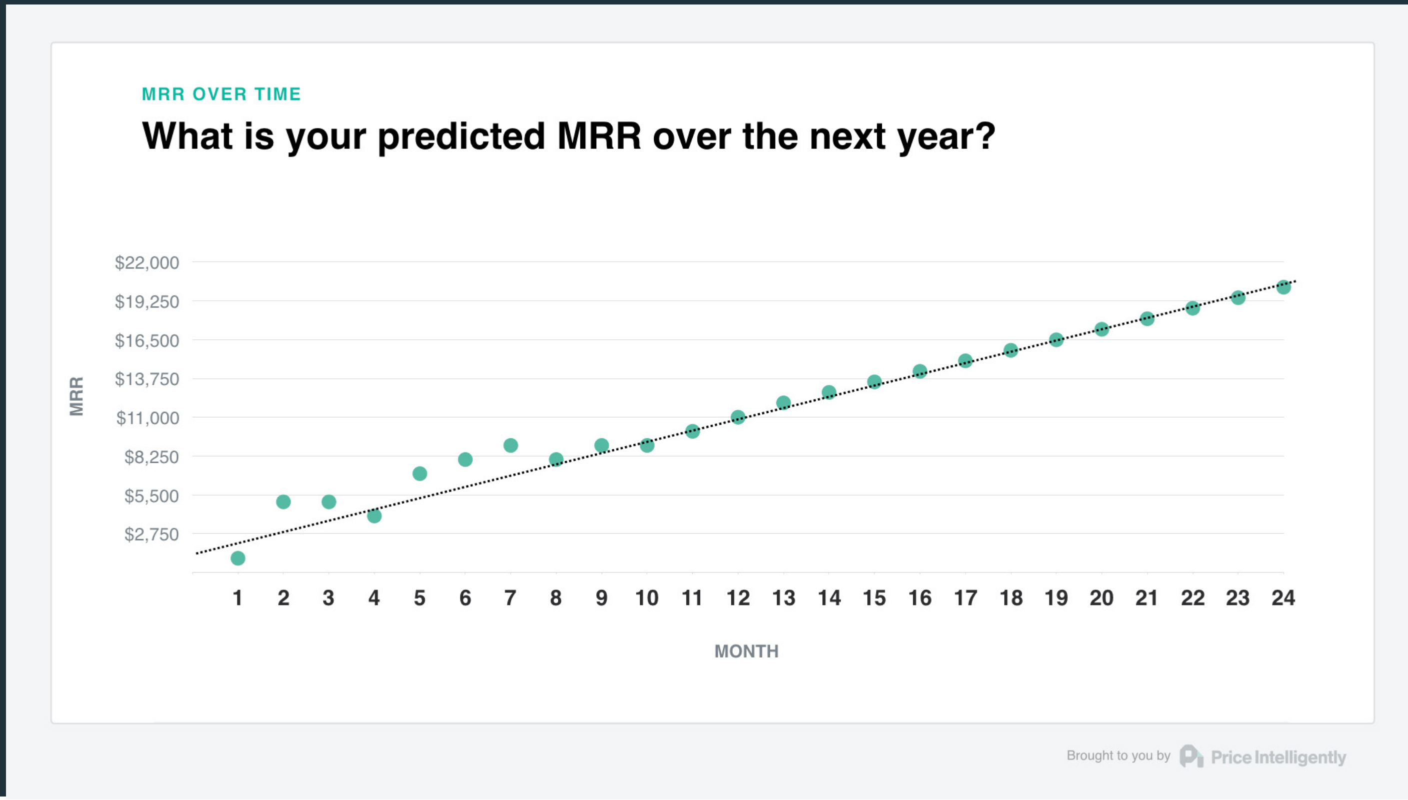
1. Expand this calculation from fitting to prediction
2. Increase dimensions and go from two-variable data to multivariable

The prediction point is one of the reasons we want to calculate R^2 for the regression line—it gives us a measure of the accuracy of our model. With an R^2 of 0.86, we can then feel confident about extrapolated prediction values from that line.

For instance, you might want to predict what your revenue will be after 18 months or 24 months. We can easily input these numbers into the linear regression equation and gain a prediction:

$$20{,}218.2 = 2{,}319 + (745.8 \times 24)$$

After 2 years predicted MRR will be $20,218.2. You can also extend the graph, extrapolating out your predicted MRR:



OK, so your growth is unlikely to be this linear, but this gives you an immediate guide as to where your current growth rate is taking you. You now know that MRR will be ~$20k after two years. Is there where your revenue needs to be for your business goals?

Having this understanding of where your current growth is taking you allows you to assess the next year and either plan with this number in mind or make the changes needed to increase.

The second expansion is along the dimensional axis. Here we've only looked at one variable related to another variable. But think about all the variables that impact your MRR—customers, churn, CAC, LTV. What you really need is a way to incorporate each of these into your regression equations together. You can do this through a general linear model.

Multiple linear regression is more complicated than linear regression, with the equation:

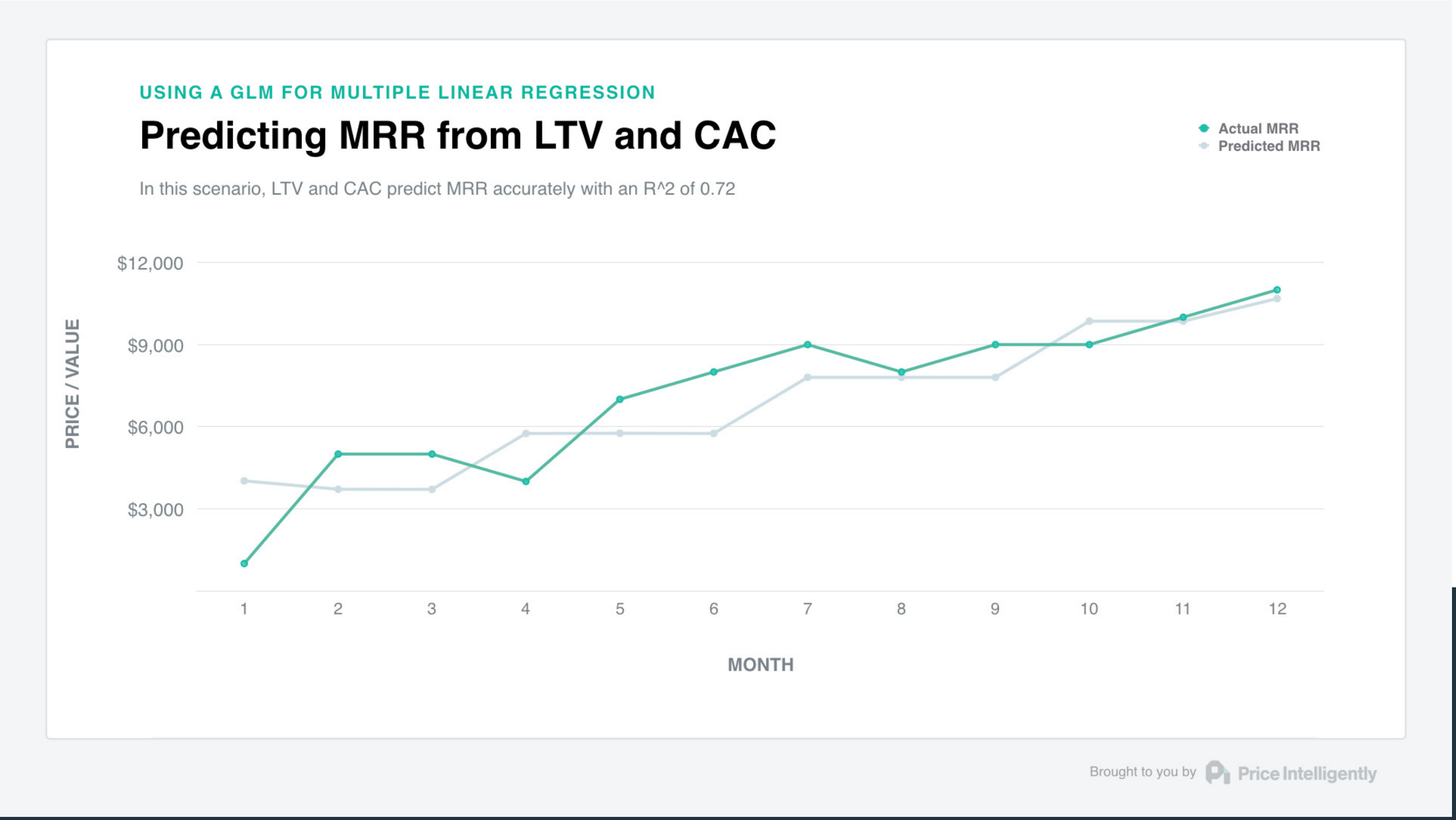
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

The variables of this equation look different to the simpler linear regression, but fundamentally are the same. You are multiplying the x variables by your estimated coefficients. You are just doing it for multiple variables instead of one. Using matrix algebra, you can determine the coefficients easily:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

You need matrix algebra because you are no longer dealing with vectors (one-dimensional arrays) but matrices, n-dimensional arrays. For instance, you might be trying to understand the relationship between LTV, CAC and MRR over the 12 months. So LTV and CAC would be in a single 2-dimensional array:

	A	B
1	CAC($)	LTV($)
2	100	300
3	100	300
4	100	300
5	200	400
6	200	400
7	200	400
8	300	500
9	300	500
10	300	500
11	400	600
12	400	600
13	400	700



Plugging these numbers and the MRR values from above into this matrix equation gives is an intercept value of $318 and two slope coefficients for each of CAC and LTV (in this case, 12.2 and 8.2 respectively). Putting this together allows us to try and model the MRR: This time, because of the extra dimensions, the line isn't straight. But it again explains a significant amount of the variance, with an R^2 of 0.72.

You can just imagine how complicated this could get. But even this simple regression technique can tell you a wealth of information about your data. Most importantly, it gives you a numerical foundation for your intuition. You might have known that ARPU and churn are related, but with r and R^2 you can put a number on that relationship.

THE ONE THING YOU ALREADY KNOW ABOUT CORRELATION

Calculating the correlation between your variables is a great way to get started understanding the trends and relationships in your data. But it isn't without danger. There are three ways that this type of analysis can go awry.

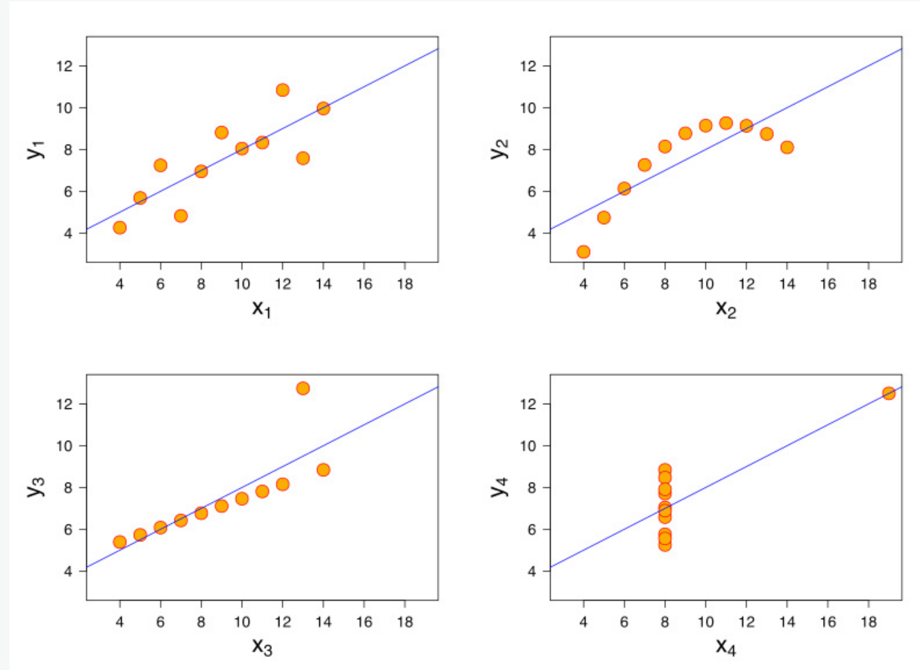
1. THINKING IT'S CAUSAL

We started with the time vs MRR relationship because it is a great lesson in the one thing about correlation that everyone has heard: correlation does not equal causation. Time isn't causing MRR to rise, even though they are highly correlated. It is what you are doing during that time that increases revenue. It is the underlying causes.

This is one of the reasons that correlations have to be treated with caution. They give you an understanding between variables but don't prove a causal relationship. In future chapters we will focus on how to prove causal relationships in your data with experimentation and analysis.

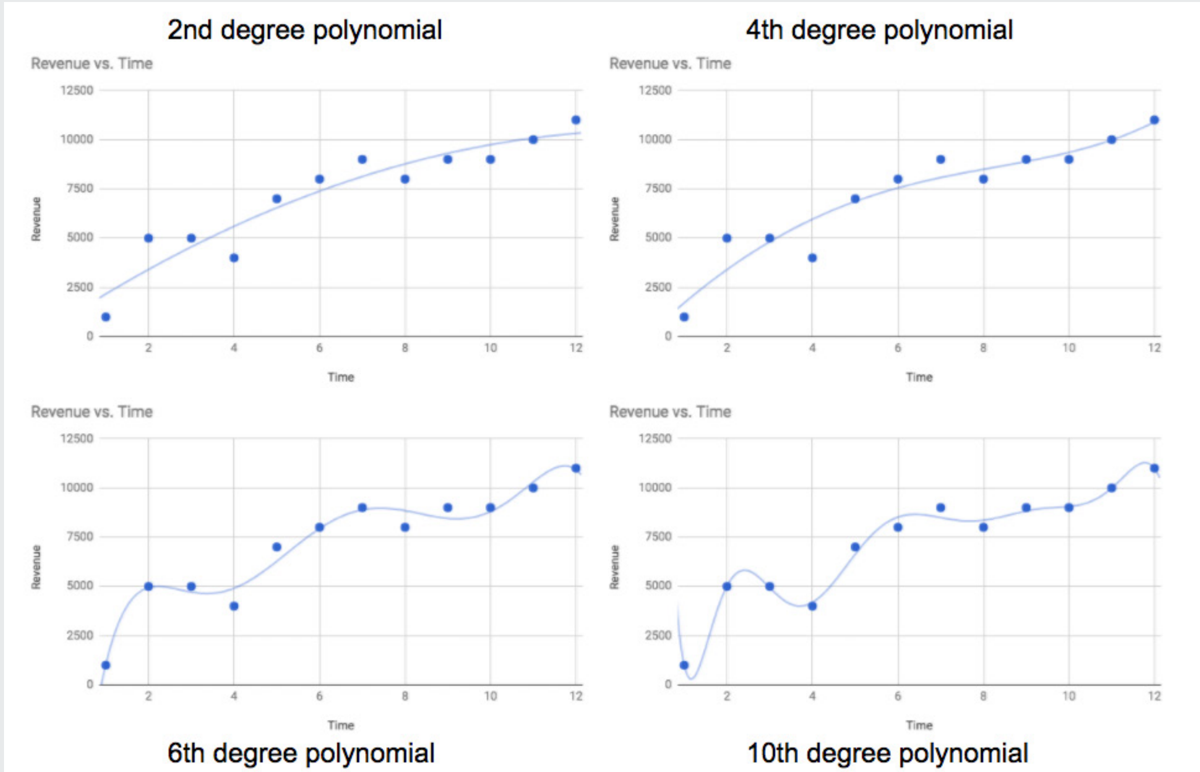
2. NOT GRAPHING THE UNDERLYING DATA

"Numerical calculations are exact, but graphs are rough" was a phrase that floated around statistical parties back in the seventies. Statistician Francis Anscombe didn't like it, so he published his Anscombe's Quartet to show the perils of only trusting numbers:



(Source: Wikipedia)

Each of these datasets has the same mean variance, correlation and liner regression line. Mathematically they are identical. But you immediately know differently once you look at them. Though you need the "numerical calculations" for rigor, you should still start with the "graphs" to visually understand your data and spot if any outliers are interesting/troubling.



3. THE CONCERNS OF OVERFITTING

Your growth is rarely linear so you can go beyond the linear equation to fit a line to your data. If you have an exponential growth rate, you would need a polynomial, curvilinear line to best describe the data.

But this comes with issues. If you try to describe the data too closely, you are in danger of overfitting. This is where the line fits your actual data closely, but is useless for predicting beyond.

Again, this is where graphing is crucial—you want to see whether your line "fits" before you use it to make any future predictions.

SIMPLE, POWERFUL TECHNIQUES

These simple techniques give you the mathematical foundation for understanding your data. You can know with a few quick equations just how closely related revenue is in your business to churn or CAC or any other metric. By starting at the foundation and graphing this data you can see the relationship, then you can use these equations (which normally come packaged in any spreadsheet or analytics package you use) to better understand this relationship, and where your data is taking you.

04

PART

4:3

2:4

6:36

1:10

1:2

2:4

+

UNDERSTANDING PROBABILITIES

04 / 07

Somewhere in your email inbox is an email telling you that you are owed $5,000,000. All you need to do is send over a few pieces of personal information so the Swiss bank can verify it's you and you can collect. In fact, you probably have multiple of these—millions of dollars just waiting to be collected. Yet you always miss these emails.

That is because they are in your spam folder. Your email provider has scanned that email and decided that it's probably not true.

Probably. Probabilities dictate a massive amount of our lives. In SaaS, that is even more true. What is the probability a visitor to your site will convert? What is the probability a customer will churn, retain, or expand? What is the probability that your company will survive?

In this part of ***The Complete Guide to Statistics for the SaaS Executive*** we want to dive into probabilities. We're going to show you not just how to calculate probabilities and how they fit together, but how they are foundational to the way we think and the way computers think, and how a single English minister made modern AI possible.

THE LIKELIHOOD OF CONVERSION

Probability. Likelihood. Chance. They all mean the same thing: How often is this event going to occur?

In math, this is expressed as a number between 0 and 1, though often people will use the equivalent percentage instead. In mathematical notation, the probability that event A will happen is expressed as P(A). So:

- When P(A) = 0, event A never happens
- When P(A) = 1, event A always happens
- When P(A) = 0.5, event A happens half the time

And all the shades in between. To compute P(A) (or P(B) and so on), all you need to know is the number of successful outcomes for event A and the number of all possible outcomes:

$$P(A) = \frac{\#\ of\ possibilities\ for\ event\ A}{\#\ of\ possibilities\ for\ all\ events}$$

The probability (or likelihood or chance) that event A will happen is the number of ways A can happen divided by the number of ways all events can happen. Let's run through this with a classic example first—a die.

When you throw a dice, there are six equally possible events: 1, 2, 3, 4, 5, 6. ***What is the probability you'll throw a one?*** Well, there is only one possibility for this, but there are six possible events, so:

$$P(one) = \frac{1}{6} = 0.16\ldots$$

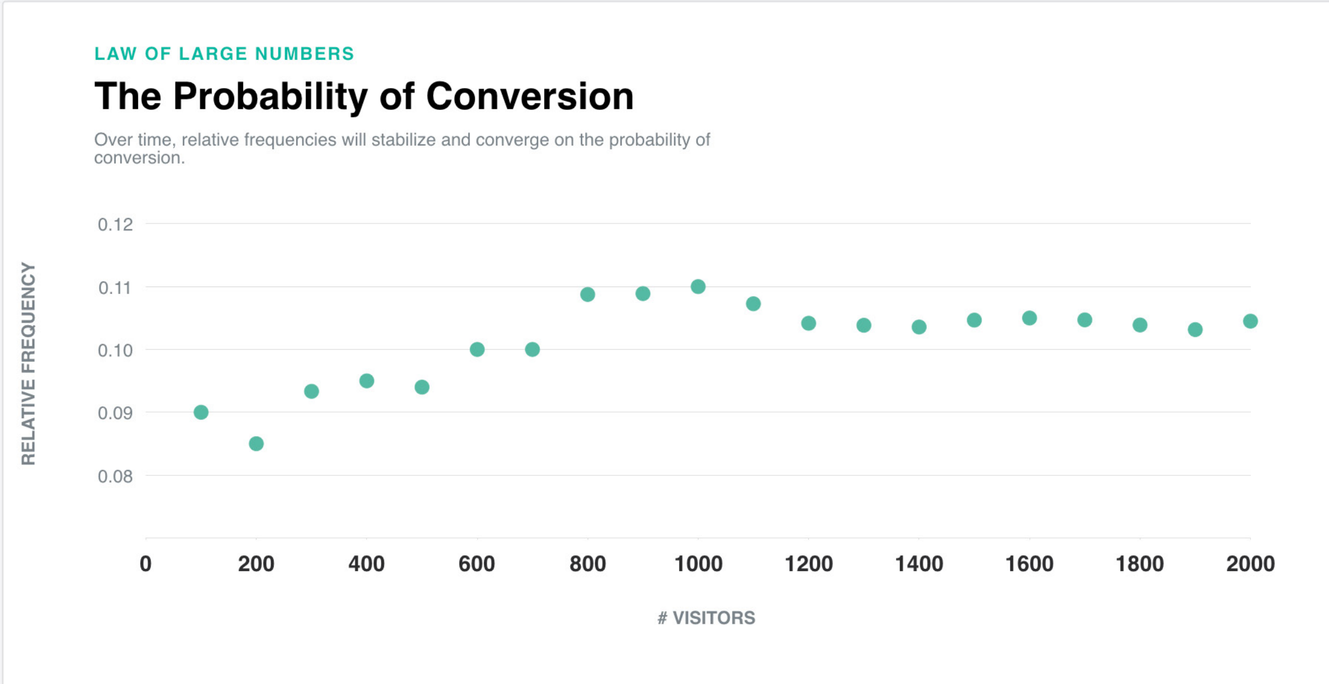
There is a one in six chance of rolling a one. In decimal, all probabilities are expressed as a number between 0 and 1. The closer to 1 the more probable the event. Here, the probability is 0.16 recurring, so not very likely.

What if we wanted to calculate the probability of rolling an odd number. These events are mutually exclusive—with a single die you can't roll a one and a five. When events are mutually exclusive you can simply add up the numerators from the probabilities of each exclusive event to get the **marginal probability**. In this case that would be:

$$P(odd) = P(1) + P(3) + P(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} = 0.5$$

Because probabilities are always between 0 and 1, it means that the sum of different probabilities related to the same outcome can't be more than 1. If you have two possible outcomes, **A and B, then:**

$$P(A) + P(B) = 1$$



As more and more customers came, the relative frequency, in the long run, would trend towards the probability of conversion:

THE LAW OF LARGE NUMBERS

Let's switch over to a more realistic business example. A visitor comes to your marketing site. What is the probability they will convert?

$$Conversion\ Rate = \frac{\#\ visitors\ who\ converted}{\#\ total\ visitors}$$

To decide that, you usually use the conversion rate of your site: But conversion rates themselves can be considered probabilities through the law of large numbers.

Say after every 100 visitors you look at how many converted so far. Each time you calculated this number you would be calculating the relative frequency of the conversion event. The relative frequency is how many times that event occurred divided by the total numbers of trials for that event:

- On the first day 100 customers came and 9 converted. The relative frequency would be 9/100, or 0.09.
- On the second day 100 visitors came and 8 converted. The relative frequency would be 8/100, or 0.08.

In fact, we can reform the formula for conversion rate into a probability equation.

01

$$P(Conversion) = \frac{\#\ of\ \ conversion\ \ events}{\#\ \ total\ \ events}$$

In this case, for the 2000 visitors, 209 have converted. So:

02

$$P(Conversion) = \frac{209}{2000} = 0.1045$$

It is easier to think of this if we go back to the dice example. The probability of throwing a six is always one in six. But the probability of throwing two sixes in sequence is much lower:

03

$$P(two\ \ sixes\ \ in\ \ a\ \ row) = \frac{1}{6} = \frac{1}{6} = \frac{1}{36} = 0.027...$$

We have a conversion rate, or conversion probability for each visitor, of 0.1045, or 10.45%. Every visitor that comes along has a 10.45% probability of converting and becoming a customer.

The conversion event is independent. That means that one event doesn't influence another. Say you have two visitors come to your site sequentially, visitor A and visitor B. Whether visitor B converts or not has no dependence on whether visitor A converts, and vice versa:

- ***The probability of conversion for visitor A is 10.45%***
- ***The probability of conversion for visitor B is 10.45%***

However, the probability of visitor A converting and then visitor B converting is just 1.09%. On the face of it, this seems counterintuitive—how can the probability of A and B converting be 10.45%, but the probability of A then B converting be 1.09%. It is because these are subtly different questions.

The intuition of this is easier to grasp if you consider rolling a six 10 times in a row. The probability of that is minute (0.000002%). It is obvious that this is unlikely to happen. The same goes for conversions. The probability of 2 visitors converting in a row is:

$$P(\text{two conversions in a row}) = 0.1045 \times 0.1045 = 0.0109$$

Probabilities start out easy, but as you consider more and more permutations, with each outcome either independent or dependent on the other outcomes around it, they get progressively more complicated.

THE PROBABILITY OF CHURN

The conversion of two random visitors is independent. The conversion of one has no bearing on the conversion of the other (for the most part). But not all events are independent.

You have five marbles in a bag. Two of them are blue and three of them are red. Therefore the probabilities of pulling out a marble of a certain color are:

- **P(blue) = 2/5 = 0.4**
- **P(red) = 3/5 = 0.6**

You put your hand in and pull out a blue. It's blue. Now what is the probability of pulling out a blue ball? Because of the first event, event A, the probabilities for the second event, event B, have changed. This is called conditional probability. The probability of the second event is conditional on what happens with the first event. In this case, the probabilities are now:

- **P(blue) = 1/4 = 0.25**
- **P(red) = 3/4 = 0.75**

The chance of pulling a red marble out has increased after the first event. In probability notation this is denoted as P(B|A), with the pipe | meaning "given." What is the probability of B given A? So P(B|A) = 0.25. This is the probability of pulling a blue ball out of the bag again (event B) given that we pulled a blue ball out of the bag the first time (event A).

01

To determine the probability of getting two blue marbles in a row, we go back to the multiplication from earlier, but here we include the conditional probability as well:

$$P(A\&B) = P(A) \times P(B|A) = 0.4 \times 0.25 = 0.1$$

There is a 10% chance that this sequence will occur.

02

A great factor of math is that you can swap around the components of an equation to calculate different aspects. Above we are calculating the probability of events A and B occurring. But we can move this formula about to calculate the conditional probability. If we were to divide the above equation by P(A), we would get:

$$\frac{P(A\&B)}{P(A)} = \frac{P(A) \times P(B|A)}{P(A)}$$

03

The P(A)s on the right side cancel each other out, so this becomes:

$$\frac{P(A\&B)}{P(A)} = P(B|A)$$

CONDITIONAL CHURN

You have a new onboarding video that you are using to help people understand your product and you want to understand how this influences retention. From behavioral cohort analysis, you know that the percentage of customers that watch the video and are retained is 68%. But this doesn't give you a true understanding of the effect of the video. You want to know the probability of retention given the video.

04

Again, from analytics, you know that 92% of customers watched the video. That is all the data you need.

$$P(B|A) = \frac{P(A\&B)}{P(A)} = \frac{0.68}{0.92} = 0.74$$

There is a 74% chance that a customer will be retained given they watched your onboarding video. Three in four customers will not churn after they watch your onboarding video. Through basic analytics and basic probability, you can call your onboarding video experiment a success.

THE PROBABILITY OF SPAM

> *When the Facts Change, I Change My Mind. What Do You Do, Sir?*

Apocryphally, this is what the economist John Maynard Keynes said to a government critic that accused him of equivocating. If true, he was conjuring up the spirit of an even older English thinker—the Reverend Thomas Bayes.

05

Bayes, an 18th-century statistician, developed this formula:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

At first glance, it looks very similar to the conditional probability equations above. And it is. But it includes a subtle difference that makes this one of the most important equations in mathematical history. Though this equation went quiet for two centuries, it is making a strong comeback and is now foundational to the modern machine learning world, all because of that subtle difference—it is updatable.

The common example for Bayes' theorem is cancer screening. It seems logical that everyone should be screened for cancer as often as possible. That way you can catch it early and treat it. But this thinking doesn't take into account two important factors—false positives and false negatives.

- A false positive is when cancer would be detected through screening but wasn't actually present.
- A false negative would be when cancer was present but the screen said it wasn't.

In cancer screening, both of these probabilities are fairly low:

- False positive rate = 9.6%
- False negative rate = 20%

That is why they are often discounted, especially when we look at the complement. 80% of tests will successfully detect cancer, and 90.4% of tests will successfully discount cancer. It seems obvious that, with these numbers, the false positives and negatives are a risk worth taking.

So what is the likelihood that someone has cancer if they test positive in screening? When tested on this question, most doctors will say 80% given that's the reliability of the screening.

06

The real answer is 7.8%. Bayesian probability can show us why. Again, the formula for Bayes' theorem is:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

06

Here we want to know P(A|B), where event A is the chance of having cancer given event B, testing positive. P(A) is the chance of having cancer in the population, which is 1%. P(B|A) is the chance of testing positive in the screening given you do have cancer. This is the true positive rate of 80%. So we can start to fill out this equation:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.01 \times 0.8}{P(B)}$$

07

However, we still don't know what P(B) is—the probability of testing positive regardless of having cancer. We can calculate P(B) through multiplying the true positive P(B|A) and true negative P(B|not A) probabilities with the actual probabilities of cancer P(A) or not P(not A), then adding these together:

- **P(B|A)*P(A) = 0.8 * 0.01 = 0.008**
- **P(B|not A)*P(not A) = 0.096 * 0.99 = 0.095**

Adding them together and fitting the answer into the equation gets us:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.01 \times 0.8}{0.008 + 0.095} = \frac{0.008}{0.103} = 0.078$$

If you go for cancer screening and test positive, though the test gets it right 80% of the time, you actually only have a 7.8% chance of having cancer. Though the test is pretty accurate, the initial chances of having cancer are small. In Bayesian probability, this is taken into consideration, and the output is highly weighted against the test by the large population that will be false positives.

UPDATING YOUR REASONING

What has this got to do with SaaS?

Everything, both from a mathematical standpoint and a theoretical one. The beauty of Bayesian probability isn't just in the initial calculation, it is in the updating.

In Bayes' theorem, you have priors and posteriors. Your priors are the initial values you use to set the math in motion. These are the true/false positives and negatives above, along with the actual population probabilities. The posterior is the output of the equation (7.8% in this case). But posteriors can become priors. In fact, this is what Bayesian thinking is really all about.

If we substitute the posterior for our initial prior of 1% chance of cancer in the population, we get this:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.078 \times 0.8}{0.062 + 0.095} = \frac{0.062}{0.157} = 0.395$$

Now we are saying that there is a 40% chance of this being a true result. If we then use that posterior as a prior, we get to 77%. Each time, the evidence that the test is correct is getting stronger. Bayes' isn't proof cancer screening is bad, it is proof that you can't rely on a single test. A test is just a test, and is fallible. You need to test over and over and update your priors constantly to get a true representation of the world.

This is how spam filtering works and is the basis of machine learning. P(A|B) is the probability that a message is spam (event A) given that a certain word occurs in the message (event B). Initially, the algorithm won't know which words mean a message is spam or not. The priors are low. But as you click "mark as spam" or "not spam" in your inbox, it gradually updates the probabilities and posteriors until some words or phrases such as "swiss bank" or "$5 million" will have a high probability of being spam. Learning from millions and millions of inboxes, it can perform this task with a high level of accuracy.

The mathematical version of Bayes' theorem can be useful in a business context. But it is the intuition of the concept that is really more interesting and valuable.

In their book, Superforecasting, about people who can accurately predict future events, Philip Tetlock and Dan Gardner said:

"The superforecasters are a numerate bunch: many know about Bayes' theorem and could deploy it if they felt it was worth the trouble. But they rarely crunch the numbers so explicitly. What matters far more to the superforecasters than Bayes' theorem is Bayes' core insight of gradually getting closer to the truth by constantly updating in proportion to the weight of the evidence."

This is the true value of Bayesian thinking. As evidence comes in—more people churning, less revenue, more support emails—you update the equation in your head, weighting the probabilities differently each time until you are sure.

THE PROBABILITIES OF BUSINESS

Nothing in business is certain—everything is a probability. Whether a visitor converts or a customer churns, every event or outcome in your company can be understood through the concept of probability. But probabilities can go deeper than just helping you quantify uncertainty. Understanding probabilities gives you a true insight into how the word really works and allows you to take in information and constantly update your thinking.

As a business leader, this skill is invaluable. If you can integrate information better and faster than the competition, then you will succeed while they are still stuck on their initial priors.

05
PART

+

THE POWER BEHIND A/B TESTING

05 / 07

A

B

Those two words ***"learn more"*** were worth $60M to the Obama campaign in 2008. They only found them because of A/B testing.

The Obama campaign put the success of the presidential election in the hands of statistics. Through A/B testing they were able to find just the right combination of images and language that resonated with their audience.

A/B testing can help you do this. Instead of making changes on a whim, you make changes based on data. In this chapter of our ***Complete Guide to Statistics for the SaaS Executive***, we are going to put a lot of what we've already learned together and show how stats can take hunches about your site and subject them to scientific rigor.

This technique pulls together basic statistics we've already learned—p-values, standard errors, and z-tests—into a single framework that allows you to put statistics to use throughout your site, emails, or product.

Let's "learn more."

A/B TESTING TURNS CORRELATION INTO CAUSATION

One night, you wake with a start from your sleep. A single thought fills your head: the CTA on your landing page shouldn't be green, it should be orange!

First thing next day, you change the hex codes and ship the new design. Within a week your conversion rate goes up from 2% to 3%. You pat yourself on the back for being the new Jonny Ives.

But are you? How do you know that the change isn't just due to chance? Or another change you made on the site? Or any other number of variables that could affect conversion? Within a week, your conversion rate settles back to 2%.

When you make changes on a whim, you are blind. This is the same whether you're altering your website, or switching up the copy in your emails, or introducing a new feature in your product. Without some kind of framework, you can't know what is causing what. You run into a problem that we identified in chapter three: *correlation doesn't mean causation.*

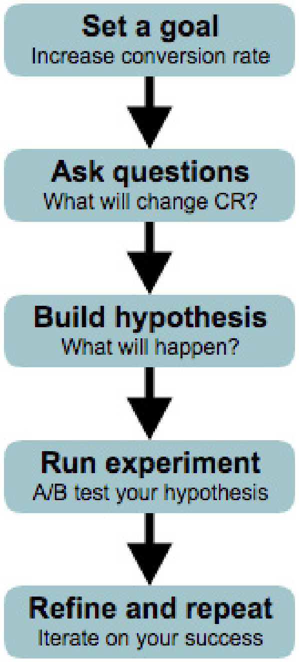
In this scenario there is a ***correlation*** between the color change and the uptick in conversion. But with no rigorous testing framework in place, you can't be sure of causation. Without a framework you can't iterate on your changes, constantly improving your site and your conversion rate. Without a framework, you can't get better and make more money.

Enter A/B testing. A/B testing lets you determine whether a change you made has had a statistically significant effect. We'll get into the details of this further down, but let's first look at what you need before you run your test—a hypothesis.

THE HYPOTHESIS IS THE BACKBONE OF YOUR EXPERIMENT.

Back when we helped you set up a bulletproof retention process we introduced you to the scientific method:

- **Set a goal:** What do you want the outcome of the experiment to be? If you are experimenting with your marketing site, the ultimate goal is going to be increased conversions. You might want to increase conversion overall by 20%, and set smaller conversion increase goals for different parts of the funnel.
- **Ask the right questions:** What is likely to change the conversion rate? At this point, you want to think about "moving big levers." Making large changes that will be obvious to visitors and have increased chances of showing a difference, but there is no limit to what you can test. If you think it matters, you can change your CTA button color from #000000 to #000001. You are asking a question of your website, your emails, or your product.
- If I change the copy of my website, will more people convert?
- If I change the subject line of my emails, will more people open them?
- If I change the flow of my onboarding, will more people come back?
- **Construct your hypothesis:** For each experiment, you need a hypothesis. What do you think will happen when you move one of your big levers?
- **Run your experiment:** Only at this point are you ready to run your A/B test on the questions that you want answering and to confirm or reject your hypotheses.
- **Refine, reject, repeat:** As you have a data-backed understanding of your conversion rates, you can now refine your hypothesis and experiments and run again to achieve higher conversion rates.



In terms of statistical rigor, the most important part of this process is right there in the middle—construct your hypothesis. A hypothesis is a formal declaration of what you expect from the testing. It is what the testing is measured against. Normally people think about hypotheses from the positive outcome standpoint:

> *Changing the CTA from green to orange will increase the conversion rate.*

Whereas this isn't wrong, it shows the confusion around statistical testing. With a statistics test, we aren't looking to prove a positive outcome, we are looking to reject the negative outcome, known as the null hypothesis. The null hypothesis here would be:

> *Changing the CTA from green to orange will have no effect on the conversion rate.*

This flip matters. When we say something is statistically significant, we are saying that the result is very unlikely if the **null hypothesis** is true. All statistics tests are measured by how far they differ from the status quo, not by how much they match a new theory. You decide whether something is significant through two values:

- **The significance level**, α. This can be considered the bar that your test has to get over for you to consider the test a success, usually set at 5% in the scientific community. Some online A/B testing tools have a default of 10%. Due to the misuse of statistics, some are pushing for a 1% threshold as standard. (Sometimes you'll see the complement of these listed instead: 1-α. So it will say a 95% significance level.)
- **The p-value.** This is the result of your test. When $p < \alpha$, you have a statistically significant result.

Finally, there is another variable related to the null hypothesis that you need to understand for running statistical tests: **statistical power.** This is the likelihood that your test will detect an effect when there is an effect. Effectively, how sensitive it is. It is affected by the "move big levers" idea—a bigger effect means you are more likely to detect it—and the sample size—more people means you are more likely to detect it.

RUNNING YOUR A/B TEST

Now you have the background on experimental design, we can get into the statistical weeds of A/B testing. The three important components of A/B testing are:

- Making sure the sample size gives you the statistical power required to draw conclusions
- Choosing the right statistical test to analyze your data
- Understanding the possible pitfalls of A/B testing and how to counteract these

Let's go through these one by one to understand their significance.

SAMPLE SIZE DETERMINES WHETHER YOU CAN LEARN FROM YOUR TEST

The most important people in any test are the ones not taking it—the controls.

You could have an A/B test where your A group see an orange button and your B group see a blue button. But if you have no previous data in either of these two colors you won't have anything to compare them against. You need a baseline where nothing changes so that you can determine if your improvements have a statistically significant effect.

(You can run an A/B/C test and test green/orange/blue at the same time, but here we'll stick with A/B testing.)

So in an A/B test you have two conditions: the control variation (A) and the experimental variation (B). As a visitor comes to your site, they are randomly assigned to one or the other group. Three important factors here:

- The word **"randomly" is crucial.** You want to control as many other variables that might affect conversion rate as possible. So you can't assign dependent on referral source, location, or other factors.
- The **two groups don't need to be the same size.** You can show your new variation to just 10% of total visitors if you want.
- However, the sample sizes for the **A and B** variations have to be large enough for the test to be significant.

Sample size is one of the most important factors in running an A/B test. Without a large enough dataset, you can't be confident in your findings. The easiest way to understand this is with an intuitive example. Consider two A/B tests. The first is with 100 visitors:

	A	B	C	D
1	Variation	Visitors	Converted visitors	Conversion Rate
2	A	50	3	6%
3	B	50	5	10%

The second is with 1,000 visitors:

	A	B	C	D
1	Variation	Visitors	Converted visitors	Conversion Rate
2	A	500	30	6%
3	B	500	50	10%

The conversion rate is identical in both. But you know that the second test is more valid. Statistics is a way to quantify this intuition.

Online optimization engines such as Optimizely or VWO will help you compute your sample size to reach significance. But let's take you through the calculation to show how it's done and what it means. You need to know a couple of numbers before you start:

- **The baseline conversion rate.** This is one of the reasons you need a known baseline. This is your current conversion rate without any changes.
- **The minimum detectable effect.** What is the goal?

Here is the calculation for estimating the sample size:

$$n = 16\frac{\sigma^2}{\delta^2}$$

The denominator is the minimum detectable effect, delta, squared. The numerator is sigma, the sample variance, squared. Of course, you don't know this beforehand, but can estimate it from your current conversion rate, p:

$$\sigma^2 = p \times (1-p)$$

Let's say you have a current conversion rate of 13% and want to see a 5% change. This becomes:

$$n = 16 \times \frac{0.1131}{0.0025} = 723.84$$

This is how many people you need in your experimental variation to achieve statistical power. The smaller your minimum detectable effect, the larger your sample size needs to be to have statistical power. If you have a 13% baseline but are only looking for a 1% change, this formula becomes:

$$n = 16 \times \frac{0.1131}{0.0001} = 18{,}096$$

It is not important that you have the same amount of people in each condition, but it is important you have a large enough sample size for statistical power in the experimental condition(s).

The "16" is a variable calculated from the statistical power and significance levels you are using. It is determined from another part of statistics, t-statistics. You can read more here.

A Z-TEST WILL GIVE YOU THE SIGNIFICANCE OF YOUR RESULT

Once you've run your test all the way through your sample, you need to analyze the data to determine if you did have a statistically significant difference. For this, there are a wide array of statistical tests you can use Student's t-test, Chi-squared, G-test, or an ANOVA—all these will produce a p-value. You can even go Bayesian.

But in our example we are going to use a test we've already come across, the z-test. This is an easy number to calculate and shows well how conversion rates and sample sizes come together to make a test significant. Let's go back to our A/B test from above with 1,000 visitors:

	A	B	C	D
1	Variation	Visitors	Converted visitors	Conversion Rate
2	A	500	30	6%
3	B	500	50	10%

Here, the formula for the z-test is going to be:

$$z = \frac{CR_B - CR_A}{\sqrt{\frac{CR_B(1-CR_B)}{N_B} + \frac{CR_A(1-CR_A)}{N_A}}}$$

Where the CRs are the conversion rates of A and B and N is the number of samples in A and B. Together they combine So the result for this test comes out as:

$$z = \frac{0.1 - 0.06}{\sqrt{\frac{0.1\times(1-0.1)}{500} + \frac{0.06(1-0.06)}{500}}} = 2.34$$

The z score for this test is 2.34. We can then use a lookup table to find the p-value. In this case, the p-value is 0.9904, meaning we can have 99% confidence that this result isn't due to chance. This is the p-value for a one-tailed test. When running A/B tests, two phrases you will come across are one-tailed and two-tailed.

- In a one-tailed test, you are only interested in a single direction for the result. Your hypothesis takes the form of the conversion rate of the experimental condition is significantly higher than the control.
- In a two-tailed test, you are interested in a difference either way. Your hypothesis takes the form of the conversion rate of the experimental condition is significantly higher or lower than the control.

A/B tests are commonly one-tailed tests. You want to know if your change is better not worse. You want to identify a winner. But a two-tailed test is valid as well, as long as you understand the math difference.

We can reject our null hypothesis and say that the B variation with the new orange button is the winner. You can confidently ship this color change to all your visitors to see a conversion rate increase.

As we can see from the formula, the z-test takes into account the sample size. So we can see what would happen if we had a lower sample size:

$$z = \frac{0.1 - 0.06}{\sqrt{\frac{0.1\times(1-0.1)}{50} + \frac{0.06(1-0.06)}{50}}} = 0.91$$

The only thing that has changed is the sample size for each variation. The z score has reduced to just 0.91. In the table for a one-tailed test this corresponds to a p-value of just 0.8186. This is below the usual thresholds for significance.

FOUR COMMON ERRORS IN A/B TESTING

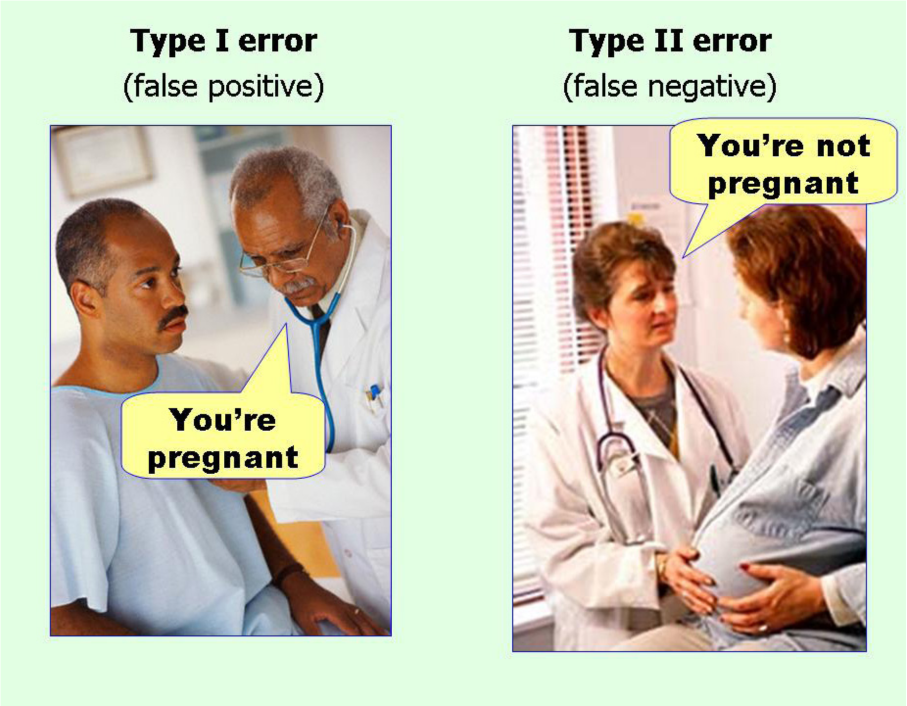
Even when using online A/B testing tools, errors can still arise. If you have the intuition about statistics and testing, you can be vigilant about these and draw the right conclusions from your tests.

- **Type I and Type II errors:** These two errors are closely related and are intrinsically linked to the way statistical testing works. When you run a test and get a p-value of < 5%, you can call it statistically significant. But that doesn't mean it's true. It just means that it is unlikely that the result would happen if the null hypothesis was true. Even in the above scenario, where we rejected the null hypothesis at the 1% level, that would still happen 1 in 100 times we ran the test, just by chance. This is a type I error. A type II error is the other way around:

- A type I error is when your research tells you that the alternative hypothesis is true, whereas in reality the null hypothesis is true. You think you have an interesting result, but you don't. This is a false positive.

- A type II error is when your research tells you the null hypothesis is true, but in reality the alternative hypothesis is true. You think you don't have an interesting result, but you do. This is a false negative. With a type I error, you say something is there when it's not. With a type II error you say something isn't there when it is.

Here is a visual representation:



- **Repeated significance error:** When you are running a significance test it can be tempting to follow the results along, looking at how well the conversion rates are tallying up. Do not do this. If you set the sample size ahead of time, you need to let the experiment run its course. Every time you look at the results before the test has concluded, you are increasing your chances of making a type I error.
- **Not understanding the results:** This is more general, but just as important. An A/B test is relative. You are testing B (the color orange) and A (the color green). They don't tell you that green is the best color for your CTA. They just tell you that green is better than orange for your CTA. Mathematically, we can say they find the local maxima rather than the global maxima. You have to try multiple tests to find all the maxima to increase your conversion rates.

SAAS FOR A/B TESTING

The purpose of this is to give you an understanding of how A/B testing works, the statistics behind it, and some of the concepts you have to look out for when designing these experiments.

But the glory of SaaS is that there is a tool for everything, and so it goes for A/B testing. Here are a few:

- **Optimizely:** Lets you easily run A/B tests throughout your site on entire pages or single design elements. It has a built-in statistical engine so that you can automatically perform all the tests above.
- **VWO: VWO** is an easy to use A/B testing engine, again that will perform the stats for you. You can build multiple variations to different changes at once. As with most SaaS optimization engines, VWO will split your audience for you so that each is randomly assigned to a variation.
- **Google Content Experiments:** If you want to run these types of tests yourself, you can use Google Content Experiments to perform A/B testing. Google performs A/B tests themselves with their own product.
- **Rankscience:** This steps away from traditional conversion-based A/B testing and into A/B testing SEO. Rankscience will A/B test elements on your site to improve your search ranking, automatically shipping the winner to all users and increasing ranking further.
- **Customer.io**: Alongside your site, you can also A/B test your emails. Customer.io allows you to A/B test different behavioral emails that you send, finding the subject lines, content, and copy that relate to the highest open and click rates.

A/B TEST (ALMOST) EVERYTHING

Actually, there is a limit to what you can test. We would advise not to test pricing using the A/B method. Two reasons why:

1. As we'll see, sample size is everything in A/B testing. As pricing is at the very end of your conversion funnel it inherently has a low sample size. A test will run for many months just to collect the right amount of data.

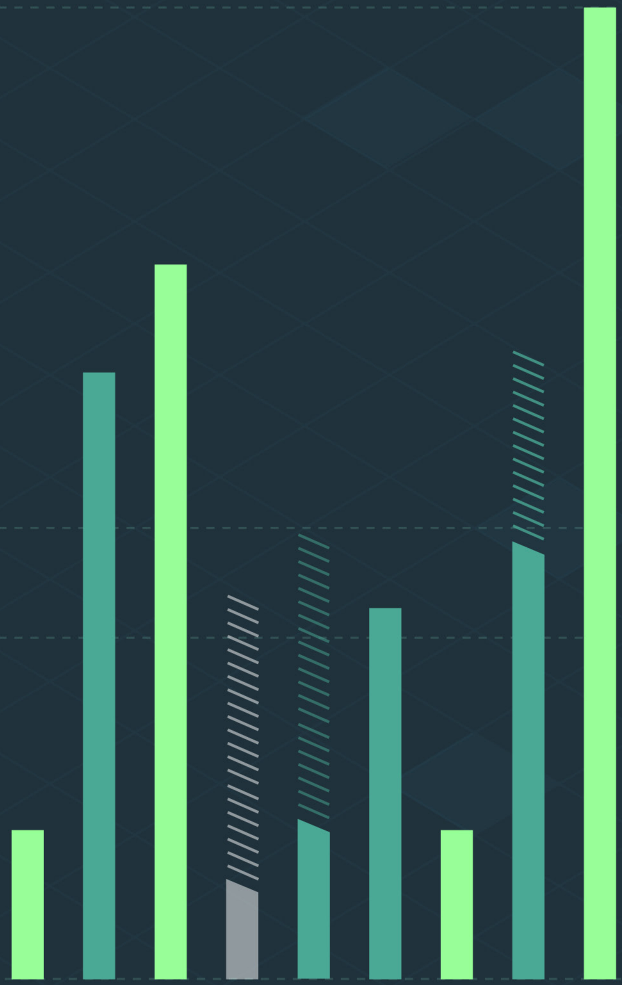
2. In those months, some customers will be randomly seeing, and paying, different prices. Not only will this make your unit economics more challenging to calculate, but it will also destroy customer satisfaction if half your customers find out they are paying more for your product just by random.

Test pricing on your entire cohort. This will tell you more about what works and what doesn't than an A/B test in this scenario.

A/B testing also doesn't take the place of market research and good design vision. Don't be a slave to the stats, instead use them as a guide for where you can improve. By doing this, and iterating on your results each time, you can optimize your site for your visitors and turn those visitors into customers. And stats will help you do it.

06
PART

MODELING YOUR GROWTH



THE ONLY ESSENTIAL THING IS GROWTH. EVERYTHING ELSE WE ASSOCIATE WITH STARTUPS FOLLOWS FROM GROWTH.

So says Paul Graham. Whether you buy into the nuances of Graham's argument or not, the fundamental is true. Every day, you are trying to grow your business. Whether that is the exponential growth of a unicorn, or the more workable growth of a bootstrapped company, understanding how growth works is integral to understanding how your business works.

In this chapter of ***The Complete Guide to Statistics for the SaaS Executive***, we want to take you through how you can model your growth using functions and a little bit of calculus. Though this deviates a little from statistics, these mathematical ideas are so important to understanding your business and determining your statistical foundations that every SaaS executive should understand what they are really showing when they present their up-and-to-the-right graph at the next board meeting.

We could write this out as part of the function like this:

$$f(4) = 4^2 = 4 \times 4 = 16$$

It’s common for startups to say they are growing exponentially. That is not entirely true. Exponential growth is a specific mathematical term, with a specific exponential function:

$$f(x) = ab^x$$

Here a and b are constants that are greater than 0 and 1, respectively. When we are talking about growth, a is your starting point. If we are calculating customer growth, it will be the number of customers you start with. b is your growth factor. If you are doubling your growth, this would be 2.In this function, x is the exponent. Another word for exponent is power or index. When we are talking growth, x increments with the time series. It is 1 on day/week/month one, 2 on day/week/month two, and so on. Let’s run through an example to see how this works. Our initial number of customers will be 1, our growth factor will be 2, and will increment by week. In week 0, when we start, we have just 1 customer:

$$f(0) = 1 \times 2^0 = 1$$

Input	**Relationship**	**Output**
1	1x1	1
2	2x2	4
3	3x3	9
4	4x4	16
5	5x5	25
10	10x10	100
28,627	28,627 x 28,627	819,505,129

Because of the zero power rule, 2 raised to the power of 0 is 1. 1 multiplied by 1 is 1. This is a bit of a funky start. The fun only really begins when we start increasing the exponent. After one day we get:

$$f(1) = 1 \times 2^1 = 2$$

Now we have two customers. At day two we have:

$$f(2) = 1 \times 2^2 = 4$$

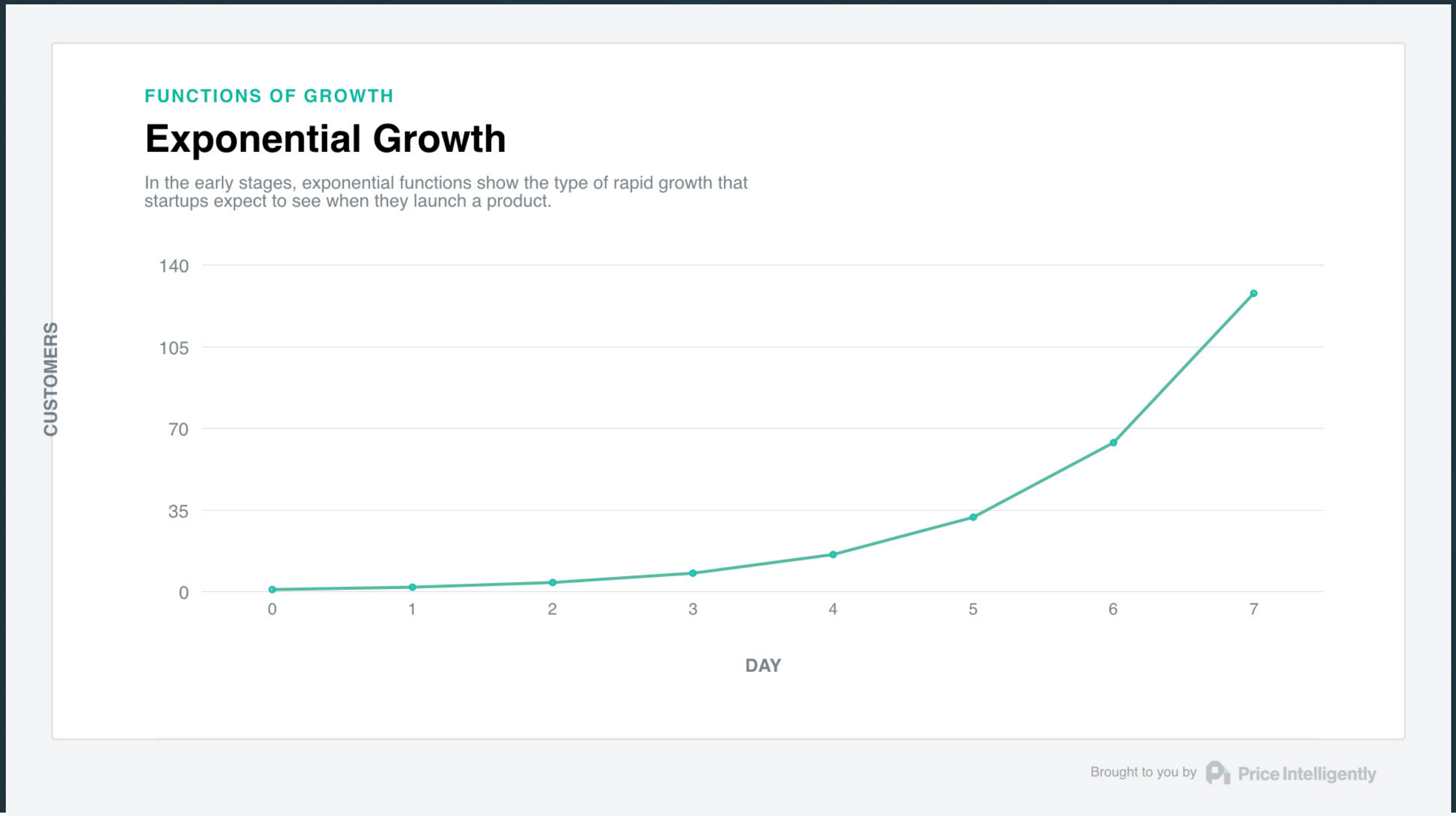
Because the exponent is increasing each day, we get exponential growth. On day three we are raising our growth factor to the power of three, so we get:

$$f(3) = 1 \times 2^3 = 8$$

You can see where this is going. At the end of the first week, the company already has 128 customers:

$$f(7) = 1 \times 2^7 = 128$$

This is just the fantastic kind of growth startups want—true exponential growth. It is also relatively realistic at this point. A company could gain ~100 customers in a week with the right marketing. This is the growth graphed:



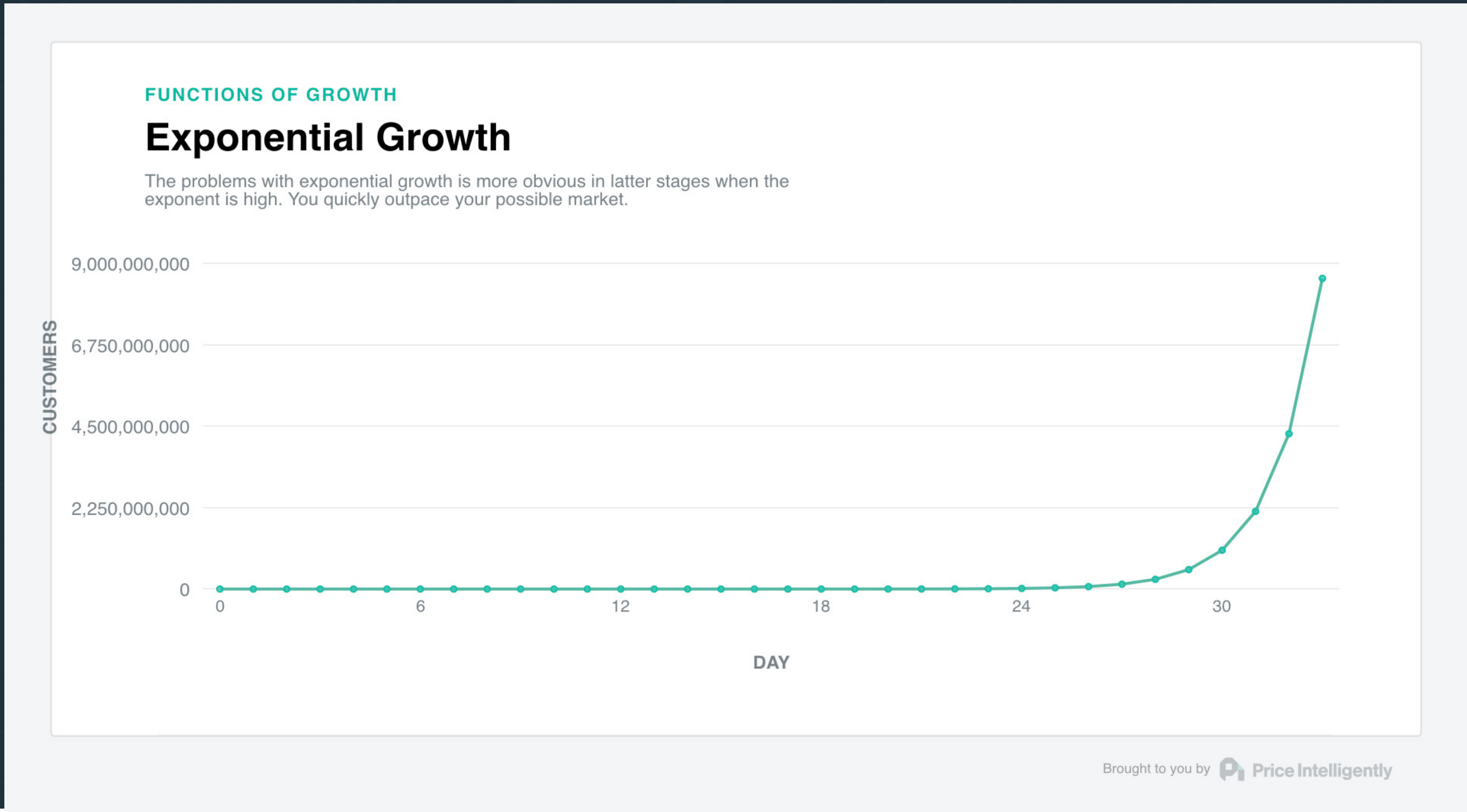
But in a limited population, exponential can't continue. On day 7, you have 128 customers. On day 30, just a month in with exponential growth, you have over one billion:

$$f(30) = 1 \times 2^{30} = 1{,}073{,}741{,}824$$

One day later, you surpass 2 billion customers and become bigger than Facebook, all within a month. By day 33, you hit this:

$$f(33) = 1 \times 2^{33} = 8{,}589{,}934{,}592$$

Congratulations! You now have more customers than people on earth, just 33 days after launch. This is that growth graphed:



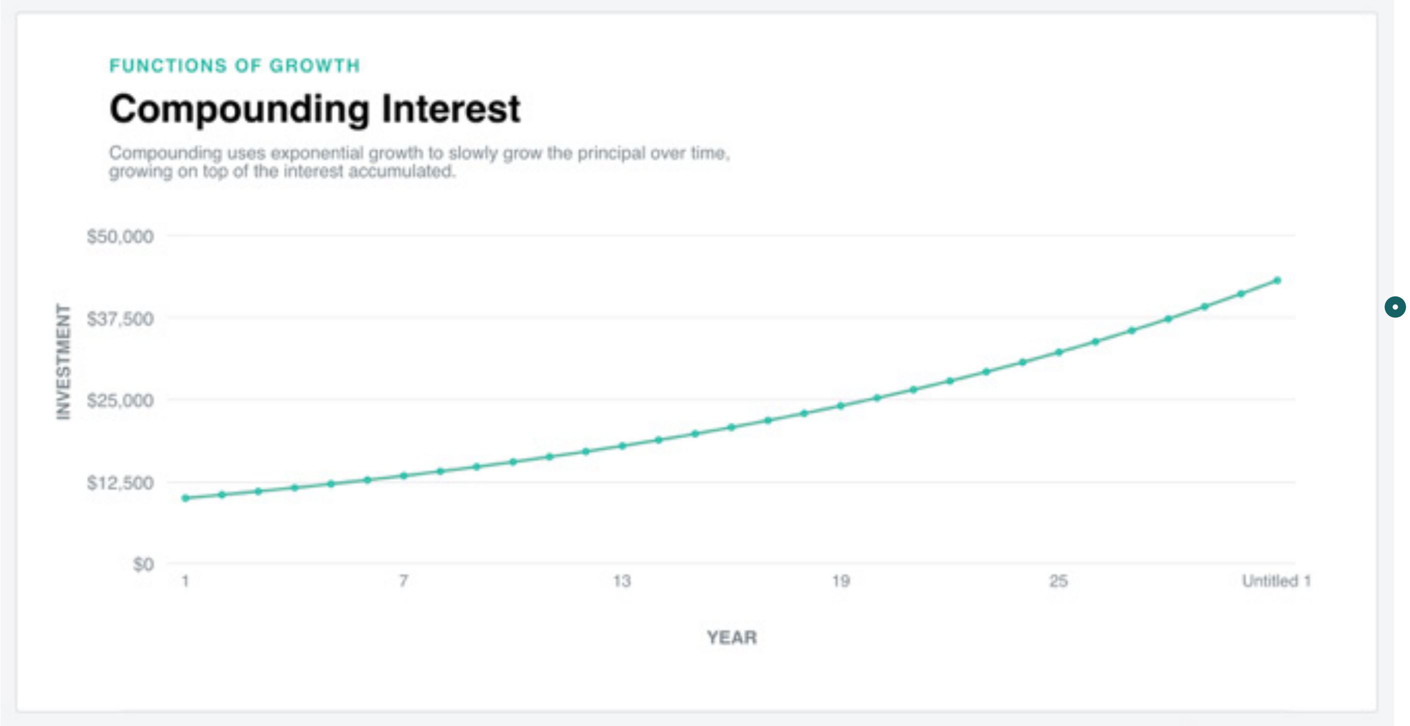
This is why it's a bad idea to go into a board meeting with investors and say you have exponential growth. They'll ask what planet you are on that can sustain this type of growth.

WHERE EXPONENTIAL GROWTH CAN WORK IN YOUR BUSINESS

With the exponential function, the growth factor is an important consideration. Pick the wrong one, and it completely skews your idea of growth.

Let's run through another example of exponential growth—compounding interest. This is the function for compounding interest:

$$f(x) = a\left(1+\frac{b}{n}\right)^{nx}$$



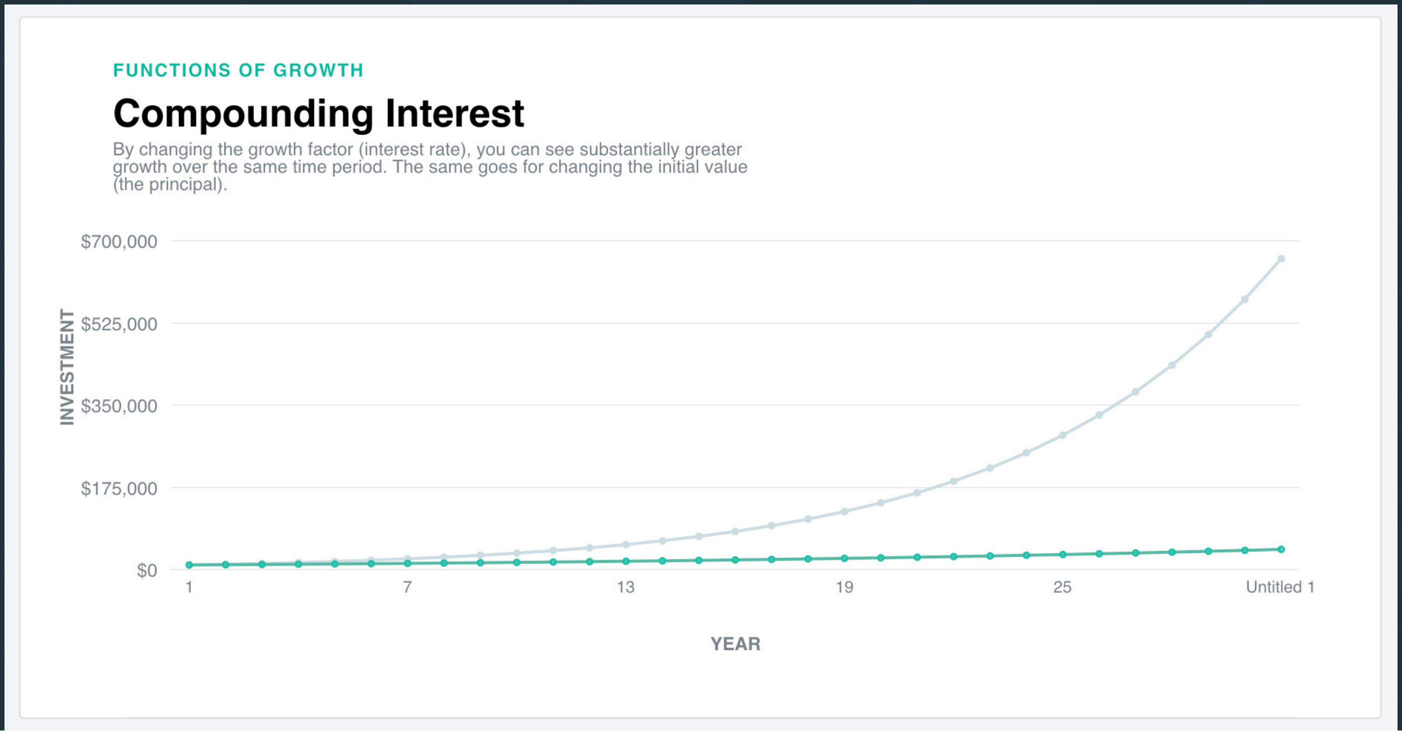
In this equation, a is the initial value that you begin with—the principal in banking parlance. b is the growth factor, or interest rate. n is an additional constant and is the number of times the interest will compound each year. If interest is compounded yearly, this is 1; if monthly, this is 12. The total number of years you are calculating for is x.

Say you put $10,000 in an account with an interest rate of 5% and leave it for 30 years—how much will you have when you cash out in 2047?

$$f(x) = a\left(1+\frac{b}{n}\right)^{nx} = \$10{,}000 \times \left(1+\frac{0.05}{1}\right)^{1 \times 30} = \$43{,}219.42$$

After those thirty years in the bank, you'll have $43k. You have 4X your investment and made a straight profit of $33,2109.42.

This is good, but not great for a 30-year investment. As with your growth rate, the growth of your investment is dependent on the initial value (the principal), growth factor (the interest rate), and how long you calculate for (the term of the investment). With the numbers above, this graphs out to:



This doesn't exactly look like exponential growth. It is, it just has a very low growth factor. The interest rate of this account is low. What about if we were to increase the growth factor by putting the investment in a high-interest account, say 15%:

$$f(x) = a\left(1+\frac{b}{n}\right)^{nx} = \$10{,}000 \times \left(1+\frac{0.15}{1}\right)^{1 \times 30} = \$662{,}117.72$$

That's much better. With a slightly higher growth rate we have now 66X our investment, made over $650,000, and have a more satisfying graph:

We can also change the principal. What if we increased our $10,000 investment to $100,000?

$$f(x) = a\left(1+\frac{b}{n}\right)^{nx} = \$100{,}000 \times \left(1+\frac{0.15}{1}\right)^{1 \times 30} = \$6{,}621{,}177.20$$

We still 66X our investment, but now make a return of over $6.5 million on the initial investment.

MODELING REAL GROWTH WITH OTHER FUNCTIONS

If we're running out of people to sell to in a month with an exponential function, what is the right function to use to model SaaS growth more accurately? By using functions to model out your growth, you can better understand what your growth really looks like and what is realistic for your company.

Here are five types of function that allow you to model growth more realistically to understand your business.

1. Exponential growth can work within strict parameters

OK, so we've just shown these are utterly unrealistic. But that is with two caveats:

1. **they are unrealistic in the long term, with high exponents**
2. **they are unrealistic with high growth factors.**

The first is obvious from the above walkthrough. Exponential growth worked well to start. If you launched well on Product Hunt, with a nice write up on Techcrunch, and an early customer showing you love on Reddit or Hacker News, exponential growth over the *short-term* is possible. You could go from 0 to 128 in a week.

Additionally, we were using a daily time scale for the example. If we said you got 128 customers over seven months, that would be slow growth. But still, the real problem is when we hit high exponents, no matter the timescale. You still can't get 8.5 billion customers after 30 years.

The second caveat is more subtle. In the example, our growth factor was 2. This meant we were effectively doubling customer numbers each day. But as long as $b > 1$, we could set that growth factor to anything for more realistic growth. If we set growth to just 1.1 (just growth to 1 is no growth, less than 1 is a decline in numbers) then after 30 days we only have 18 customers:

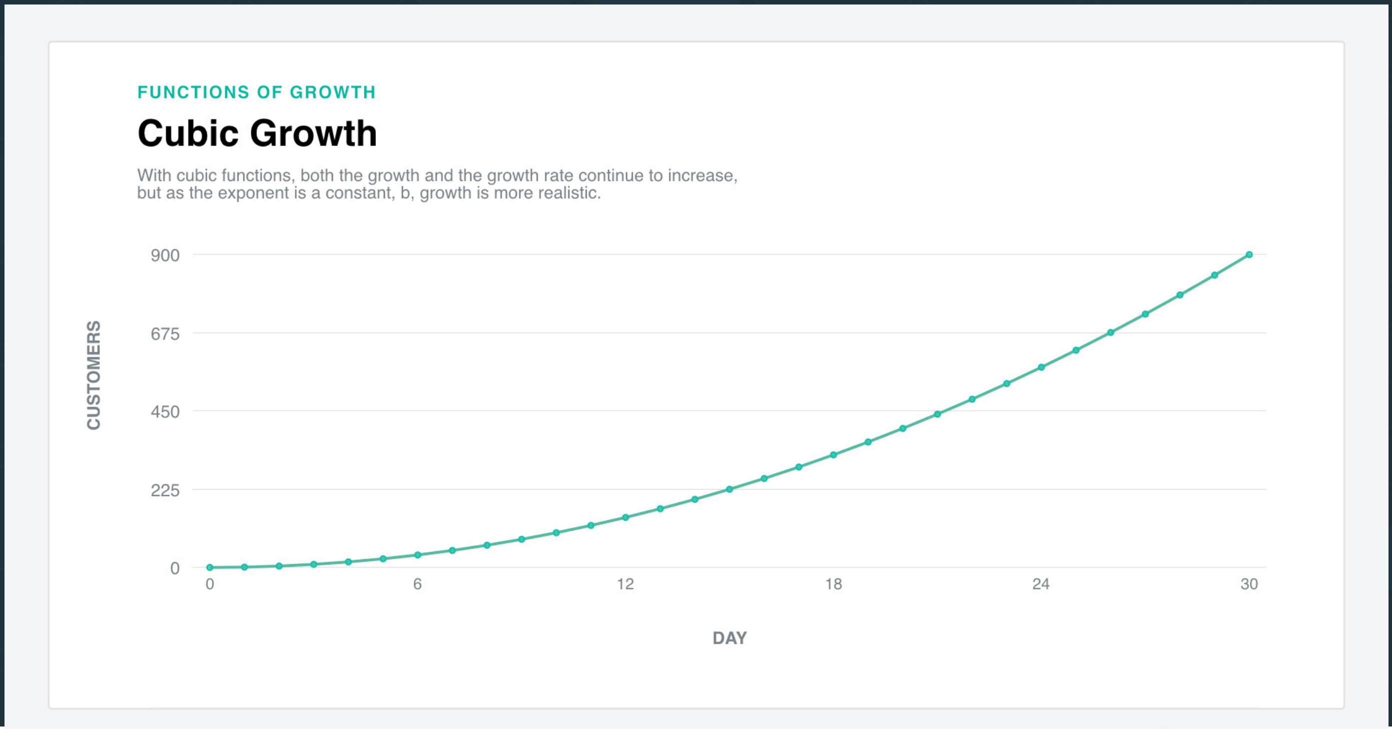
$$f(30) = 1 \times 1.1^{30} = 17.45$$

But still, **the real problem is when we hit high exponents.** The low growth factor is just delaying the inevitable. In this case until approximately two-thirds of a year into the new company when we outpace the planet's population:

$$f(240) = 1 \times 1.1^{240} = 8{,}594{,}971{,}441.07$$

Takeaway: You can use an exponential function to model real growth, but only for short time periods. When the exponents are low, this model can be realistic. When they become high, it no longer conforms to reality.

Cubic growth looks like exponential growth when graphed:



2. CUBIC GROWTH CAN BE MISTAKEN FOR EXPONENTIAL

Cubic functions are related to exponential functions, and the graphs look similar, but the terms in the function are rearranged. In an exponential function, the terms are:

$$f(x) = ab^x$$

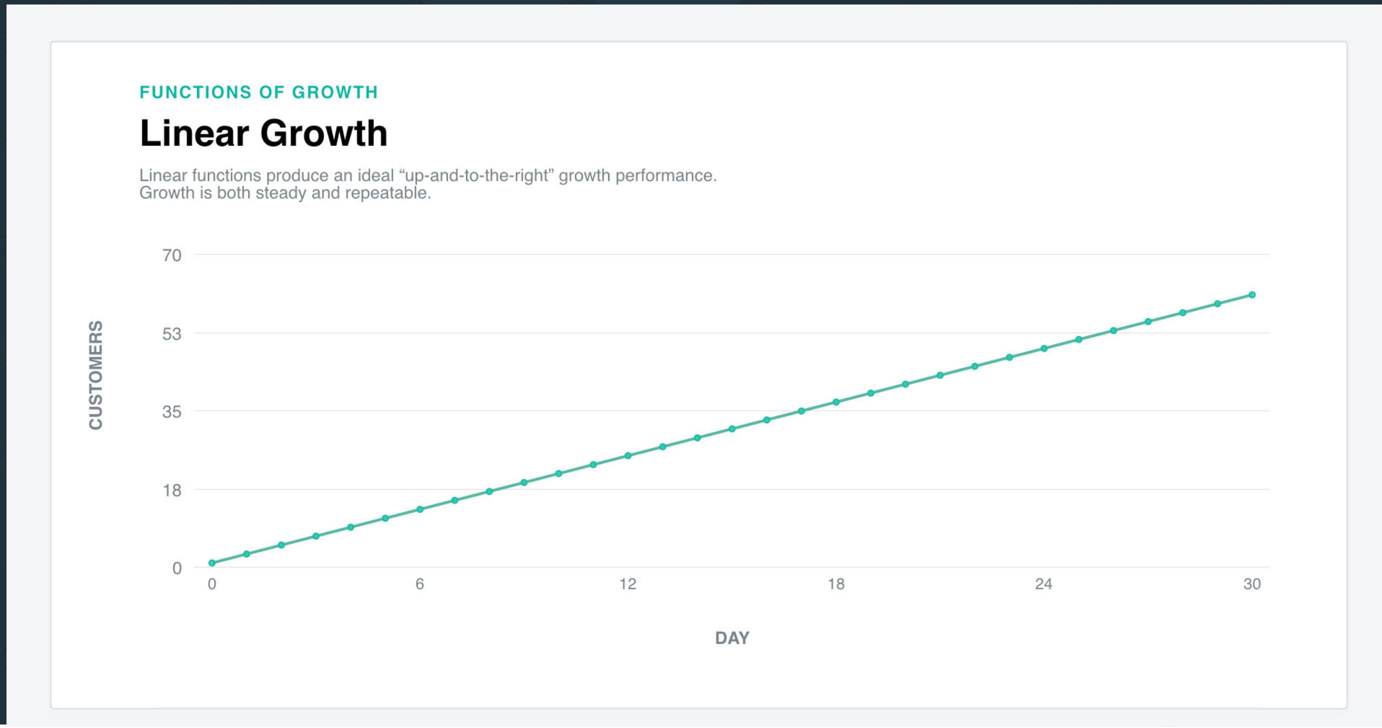
In a cubic function, the terms are: $f(x) = ax^b$

The function is no longer exponential. It still grows, but the rate is controlled by the growth factor b instead of the exponent x. This makes a huge difference in how the function scales. At day 30, we have 900 customers with a cubic function instead of 8 billion:

$$f(30) = 1 \times 30^2 = 900$$

This is one of the fundamental mistakes people make. They see a concave graph like this and presume exponential growth. In reality, it's ax^b rather than ab^x.

Takeaway: Cubic growth should be seen as good growth. Though not quite exponential, every day/week/month you are growing faster. The problems come from sustaining that growth without burning out or acquiring more customers than your company can support.



3. LINEAR GROWTH IS STRAIGHT AND STEADY

Compared to exponential or even cubic growth, linear growth can seem a letdown. But linear growth is the epitome of the "up and to the right" that you want to see in your customers, usage, or revenue. It is governed by the linear equation that we saw initially in chapter 3. This function is:

$$f(x) = a + bx$$

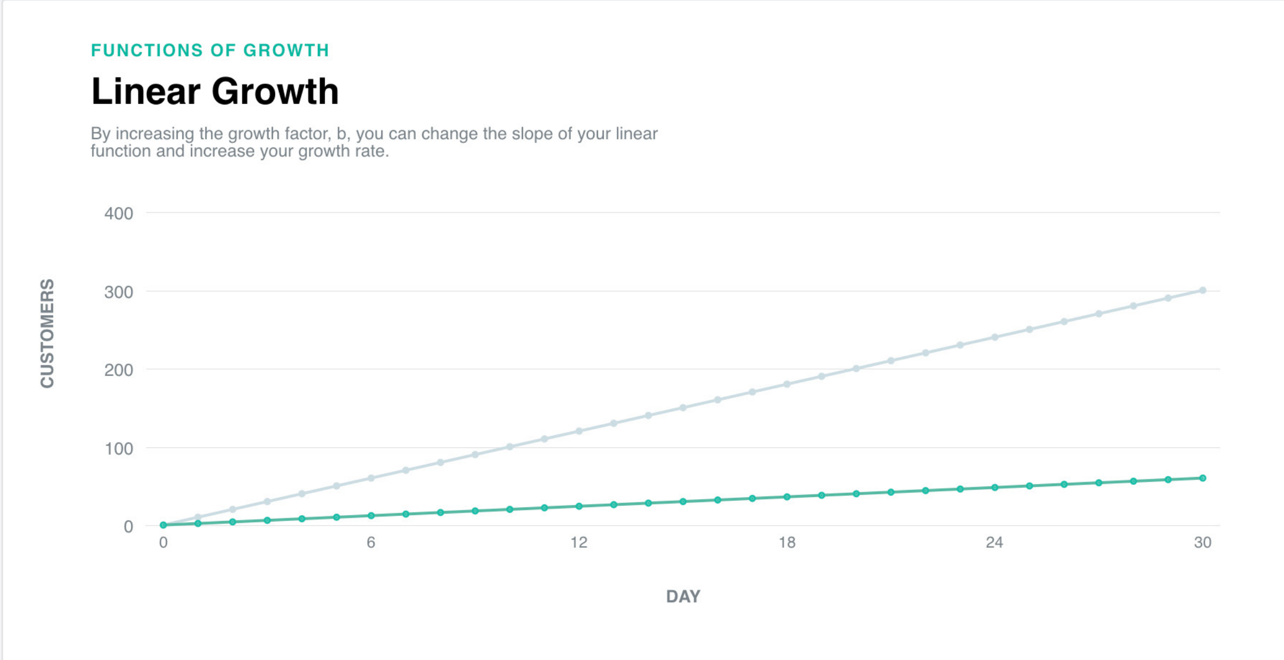
Where a is the starting number of customers and b is your growth factor. Each additional day/week/month (x) in your time series, you multiply the growth rate by x. Let's keep the initial value and growth factor as the same as above. On day 10, this becomes:

$$f(10) = 1 + (2 \times 10) = 21$$

Growth follows a straight line, so that on day 30 we get:

$$f(30) = 1 + (2 \times 30) = 61$$

If we increase the growth factor to 10, we can see the difference in the slope:



You can get strong linear growth and, unlike exponential growth, linear growth is both realistic and under your control.

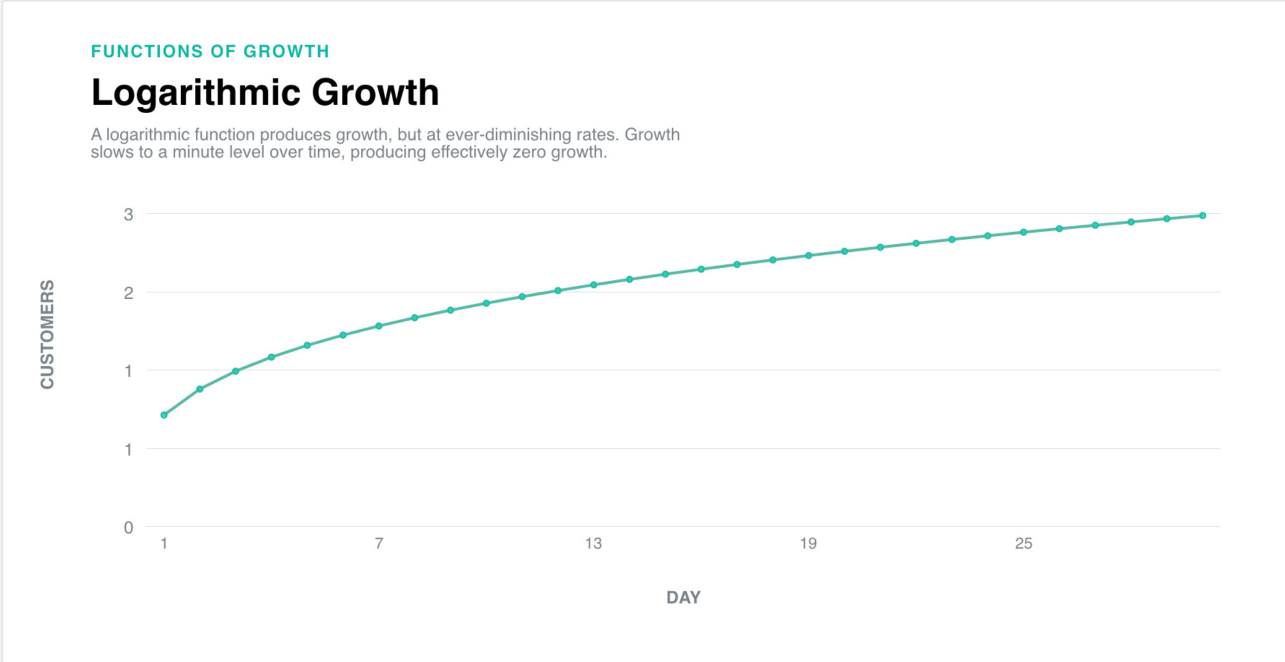
This does seem like a letdown. At the 30-day stage, exponential growth led to over one billion customers. Even cubic growth led to 900 customers. With the same parameters, linear growth gives you just 60 customers.

But because there is no exponential growth in the linear function, either through x or b, it is a truer representation of real growth. It also shines the spotlight on your growth factor. Unaided by exponentials, a growth factor of two doesn't look as good. Therefore you can look at this as an area of improvement to increase customer numbers.

The growth factor, b, controls the slope in a linear function. In the above, for every day that passes, 2 extra customers are added. When you increase your growth factor, you increase the slope and add more customers per day. You increase your growth factor by performing better as a business—marketing better, converting better, adding more value for your customers.

Takeaway: Linear growth is good. It shows you have built a repeatable, working growth machine. Look for ways to nudge the slope up, but apart from that—just keep doing what you are doing.

You do not want logarithmic growth. Graphing this out gives a clear idea of what is going on with this type of growth:



Logarithms are the inverse of exponents. It is the type of growth you don't want. To create a logarithmic version of your growth, you take the logarithm of x in the function:

$$f(x) = ab^{\log x}$$

Because we are taking the logarithm of x, as x increases the power increases logarithmically. For instance, when x = 1 with our usual parameters, the function returns 1:

$$f(1) = 1 \times 2^{\log(1)} = 1 \times 2^{0} = 1$$

But we have to wait until x = 10 for the power to increase to 1:

$$f(10) = 1 \times 2^{\log(10)} = 1 \times 2^{1} = 2$$

If x is days, then it has taken 10 days to go from one customer to two. By day 30, you've only incrementally increased, and haven't achieved the heady heights of three customers:

$$f(30) = 1 \times 2^{\log(30)} = 1 \times 2^{1.48} = 2.78$$

It takes 100 days to get to a power of 2, giving you four customers in total:

$$f(100) = 1 \times 2^{\log(100)} = 1 \times 2^{2} = 4$$

Because it is the inverse of exponential growth, instead of shooting off vertically, the graph shoots off horizontally. However, it is still technically growth—the line never plateaus. This isn't a horizontal asymptote, where the curve never reaches the line above. It's just that every day, your growth will get less and less.

Takeaway: If you find your growth following a logarithmic function, your growth is stalling. If you are a mature company, this might be OK, but if there is still market available, logarithmic growth is a red flag.

5. LOGISTIC GROWTH ALLOWS YOU TO PLAN FOR YOUR MARKET

So far, all the functions described have been fairly simple and followed a standard form with three variables, a, b, and x, just arranged differently.

Logistic functions are a bit different. They include two more constants, c and e.

The constant c is the carrying capacity of the population. This is the maximum that growth could ever be. For instance, if you were a social network such as Facebook, this really is the 7.4 billion people on Earth. Because you are a SaaS company, this is the size of your market. Say you only sell to other SaaS companies, the carrying capacity, c, is ~10,000.

The constant e is Euler's number (2.718281828...), a mathematical constant that adds a constant exponent to the function. The logistic function looks like this:

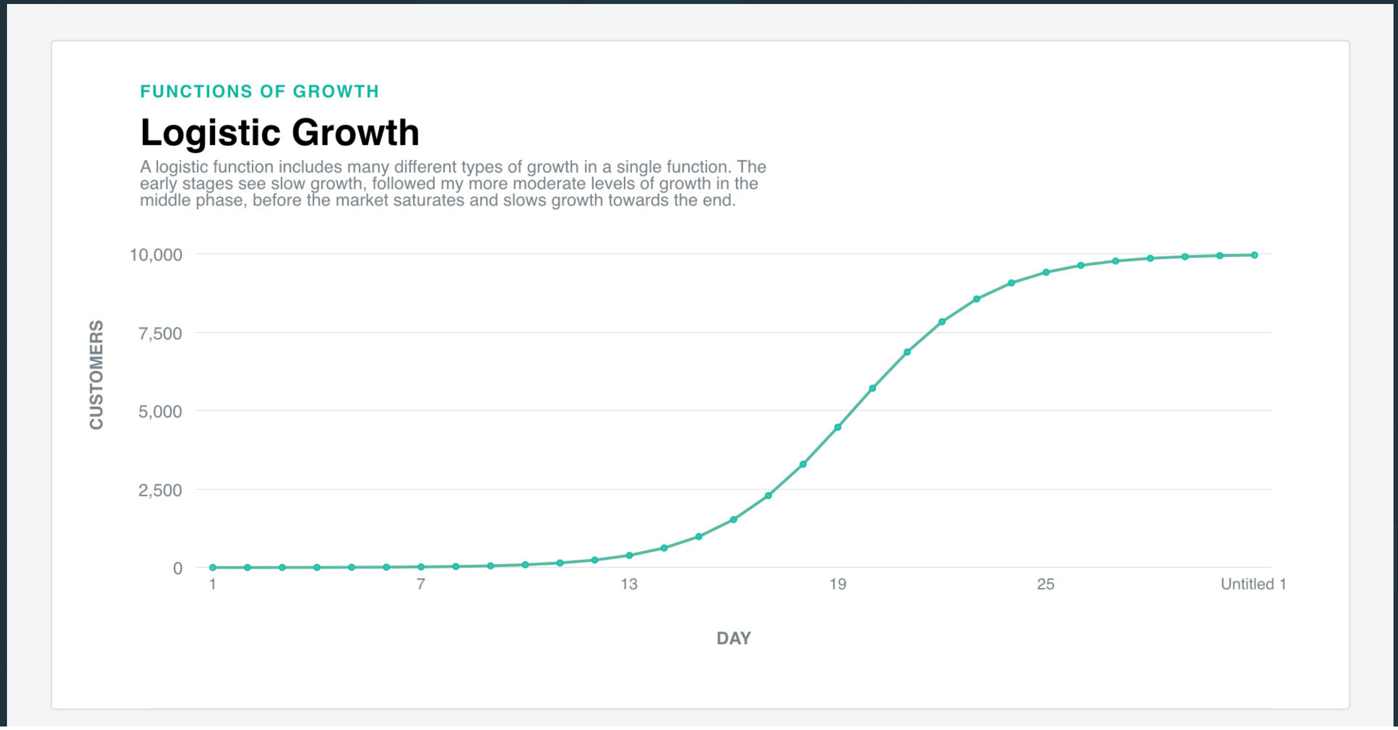
$$f(x) = \frac{c}{1+ae^{-bx}}$$

The parameters a and b are also subtly different here. The initial value that we have denoted with a above is calculated by c / 1+a here. The constant b is related to the growth factor, but not quite the same. In this example, we've set it as 0.5 purely for illustration. Though immediately more complicated than the other functions above, this complexity allows us to generate a more complex model of growth. When x = 0, the function returns 1:

$$f(0) = \frac{10{,}000}{1+9{,}999 \times 2.718^{-0.5 \times 0}} = \frac{10{,}000}{10{,}000 \times 1} = 1$$

Again, this falls prey to the zero power rule, setting our initial customer value as 1. On day one, we increase that number to 1.649:

$$f(1) = \frac{10{,}000}{1+9{,}999 \times 2.718^{-0.5 \times 1}} = \frac{10{,}000}{10{,}000 \times 0.607} = 1.649$$



As we move through the month, the ae^-bx component of the function gets less, effectively allowing more and more of the carrying capacity through. At day 15, this is already small:

$$f(15) = \frac{10{,}000}{1+9{,}999 \times 2.718^{-0.5 \times 15}} = \frac{10{,}000}{10{,}000 \times 0.00055} = 1531.33$$

By the time we get to the end of the month, this part of the function is practically non-existent, and the function when x = 30 returns almost the entire population:

$$f(30) = \frac{10{,}000}{1+9{,}999 \times 2.718^{-0.5 \times 30}} = \frac{10{,}000}{10{,}000 \times 0.0000003} = 9969.51$$

From these discrete points along the graph, it isn't clear exactly what has happened. That is clearer when we graph out each day along this month. The complexity of the function turns into a more interesting graph:

Curvy! The graph shows that with a logistic function there are three phases to growth:

1. A slow start phase
2. An exponential growth phase
3. A slow end phase

Logistic functions are often used to model natural ecological growth. As new species emerge, they suffer from slow growth initially as they are not perfected for the environment. As they evolve and specialize, they seem like explosive growth. This growth plateaus as they consume all the resources available.

This type of growth more closely correlates with what might be considered normal growth for a company. As you start out, it is difficult to get traction. But once you build your brand, improve your product, and start the growth flywheel, you then go through a period of accelerated growth. That ends once you saturate the market. It then becomes more expensive to acquire new customers.

The bound put on logistic growth by the carrying capacity makes it different and potentially more useful as a model than the other functions above. If you know your market size, you can set that as an upper bound. You then determine which phase you are currently in and how long it will take to hit your full potential.

Takeaway: Logistic growth is what really occurs in successful companies. By understanding your market, you can model your growth around this S-curve and determine the right time to grow and the right time to consolidate.

DETERMINING YOUR GROWTH FACTOR AND INITIAL VALUE

The functions above all need inputs and constants to work: a, b, and x. In time series such as growth curves, x is always going to be time. But what about a and b. How do we get realistic values for those constants?

With a and b above we've basically just picked two numbers out of thin air. Sure, they kind of make sense. For instance, it makes sense to have the initial customer value as 1. But you won't always be starting from the beginning when modeling growth. You might already be three years into your business and want to see what the next three years will bring.

For modeling any type of growth, you want to get these numbers as realistic as possible. There are two possible places to start instead of "thin air."

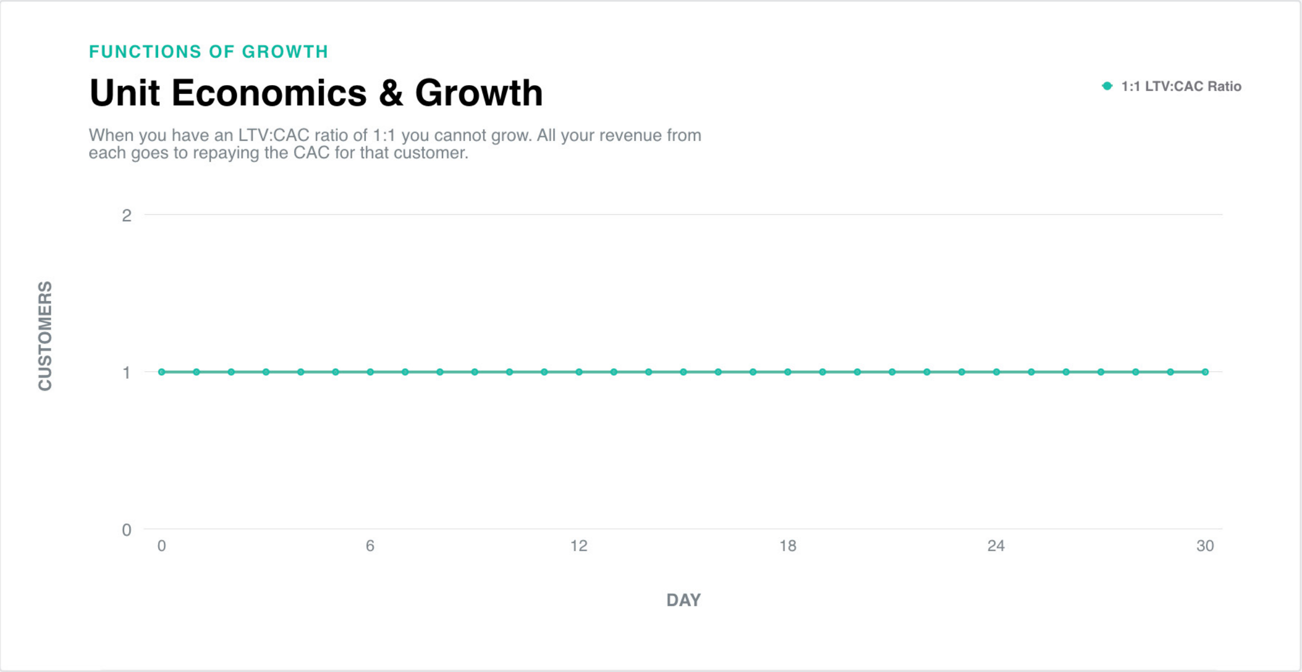
DETERMINING YOUR GROWTH FACTOR FROM YOUR UNIT ECONOMICS

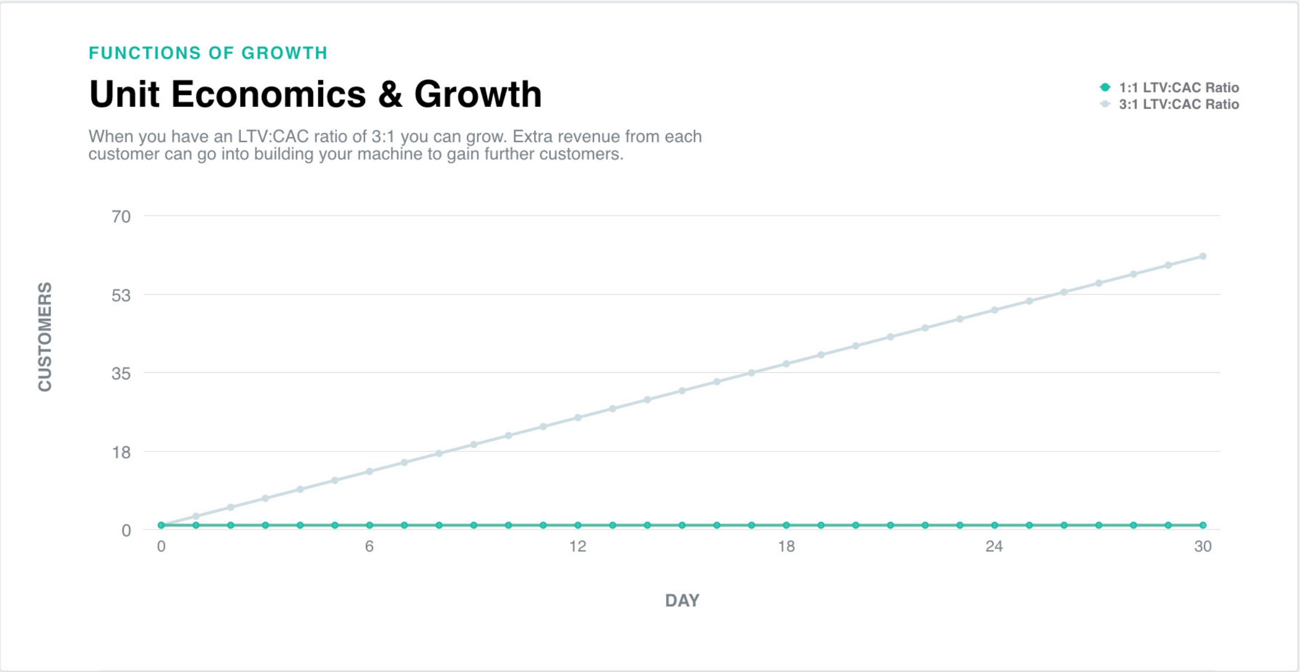
One of our fundamentals at Price Intelligently is that your LTV:CAC ratio needs to be at least 3:1 for you to grow. You can use this as a growth factor for each function to see what the curves look like with either different target LTV:CAC ratios or your current unit economics.

Let's use the linear function to model out our growth with different unit economics. We'll keep a = 1, but vary b as we shift our LTV:CAC ratios. This is our growth graph with a 1:1 ratio:

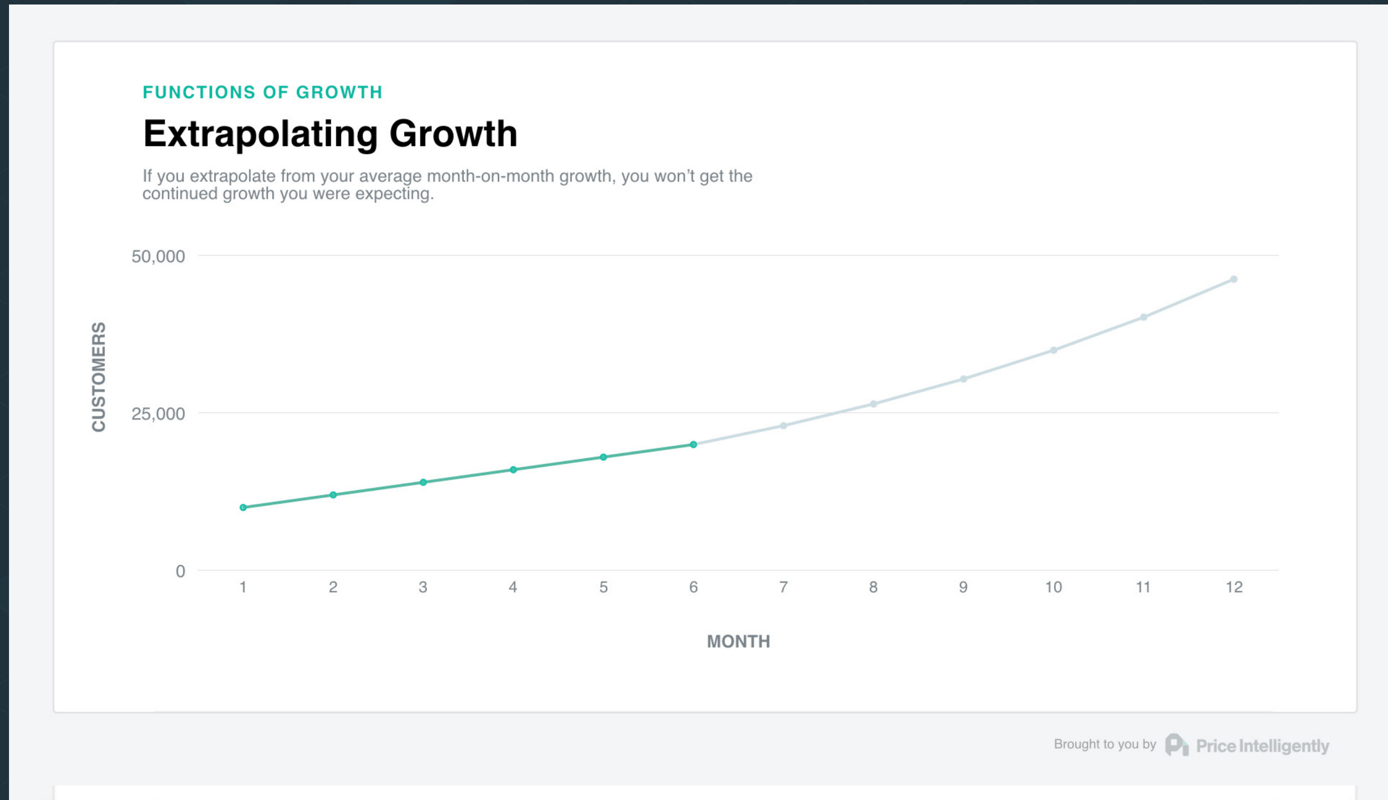
If we have a 1:1 ratio, b = 0 and there is no revenue available for growth. All you can do is stay level.

Now b = 2, so growth is the same as in the scenario above. But with good monetization and retention strategies, we've seen companies achieve LTV:CAC ratios of 11:1 or more:





Here b = 10, and we see even better growth. And this is with the most “boring” type of growth—linear growth. A growth factor such as this in cubic growth can push your acquisition sky high.

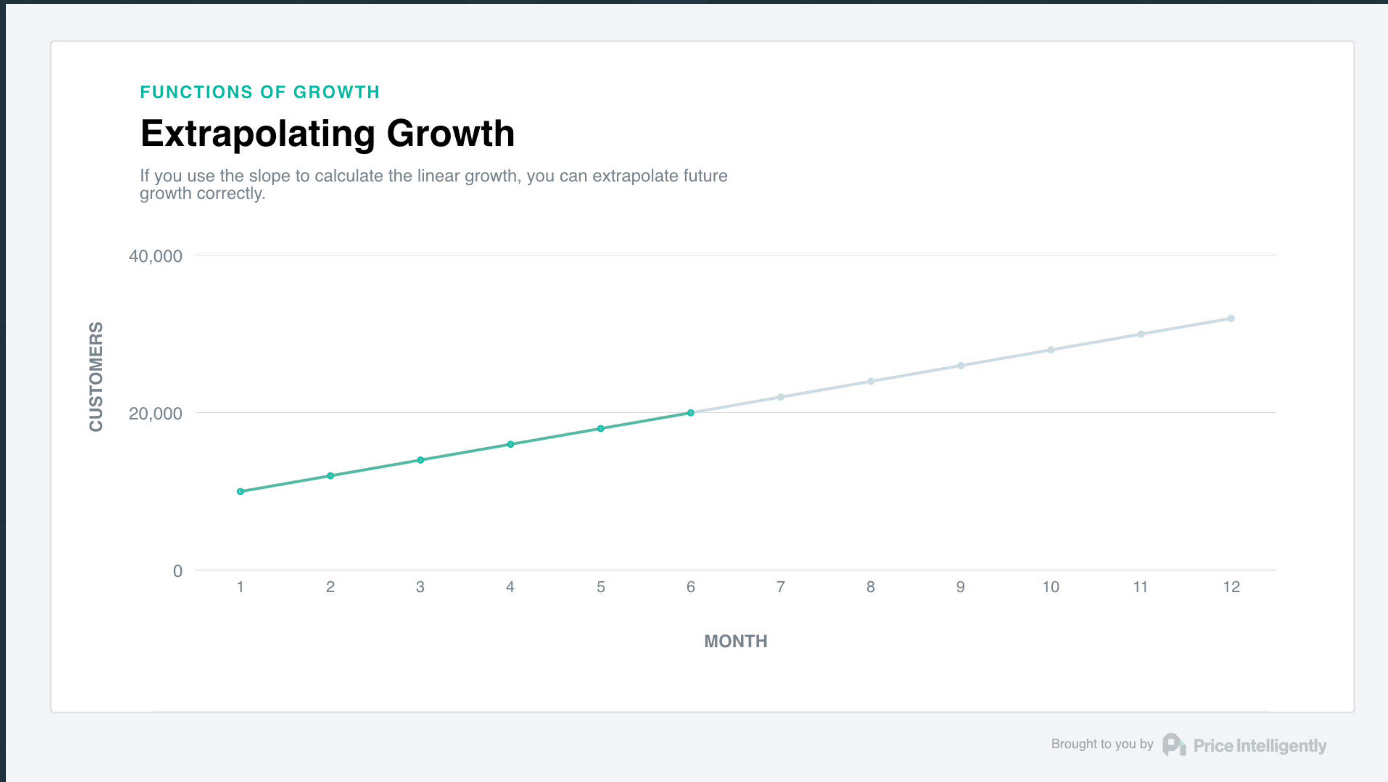


DERIVING YOUR GROWTH FACTOR FROM YOUR PAST GROWTH

We can use our past growth as the basis for these numbers. This second option lets us base our models in reality. It is at this point they can become more than just theoretical and allow us to plan out what future growth might look like.

The simplest way to do this would be to look at your average growth over a previous time period and then extrapolate that out. But this can come with errors. For instance, an intuitive way to do this would be to look over your previous six months of growth and then average your growth rate and extrapolate that out. Say this is your numbers for the past six months:

Input	1	2	3	4	5	6
Output	10,000	12,000	14,000	16,000	18,000	20,000



It seems obvious you are growing steadily so if you average out that growth, you get a 15%MoM growth rate. But when you plot that growth, your linear growth goes askew:

The first six months are real, the second six months are predicted. You have increased growth in the second half of the year. This is because your MoM growth has been slowing as you've got more customers. The MoM growth rate of month 1 was 20%, whereas the MoM growth rate of month 6 was just 11%. This averaged to 15%, but that isn't where your growth is going.

The better way would be to look at the slope of the line and continue along that slope. The slope is:

$$Slope = \frac{change\ \ in\ \ Y}{change\ \ in\ \ X} = \frac{\Delta y}{\Delta x} = \frac{20{,}000 - 10{,}000}{5} = \frac{10{,}000}{5} = 2{,}000$$

If we add 2,000 customers each month we get a better prediction of the future:

This is the basis of derivation in calculus. To do this properly, instead of looking for the slope over 6 months, you are looking for the slope over an infinitesimal amount of time, called dx. This is the basic formula for deriving any of the functions above:

It looks complicated (and can be depending on what you are trying to derive), but the concept is simple if not intuitive. Compute the change in x plus a measurable amount of dx, then remove dx to find the true derivative. You can find a better explained explanation than we can ever do on that site.

For the most part, the simpler version will suffice, as long as you visually check it makes sense. But using calculus, you can make your growth rate computations mathematically rigorous so you know they are the best prediction you can make.

UNDERSTAND YOUR GROWTH MATHEMATICALLY

You can think of your company as an abstract function. Input goes in—ideas, marketing, customers—and an output comes out—growth and revenues. The relationship between the two is the hard part.

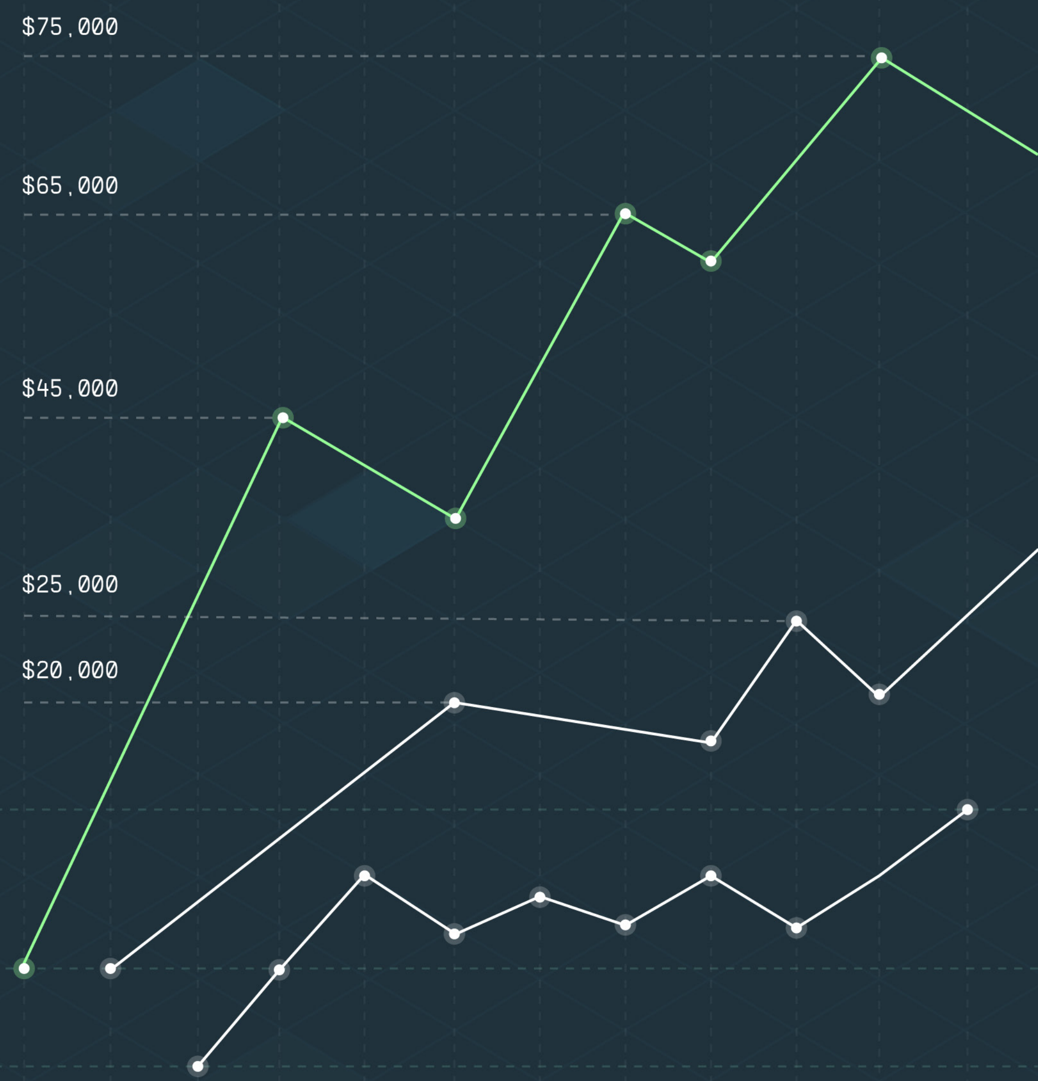
But functions should be about making a relationship simpler. When you figure out the relationship between an input and an output, you can package it in a function and make the calculation easily repeatable. Modeling out your growth like this allows you to put together realistic scenarios for your board and your team alike. Then you can get down to the real business of hitting these growth targets.

07
PART

+

MACHINE LEARNING

07 / 07



IN THE VERY FIRST CHAPTER WE SAID THE FUTURE IS STATISTICS. ONE OF THE EXAMPLES WE USED WAS AI: ARTIFICIAL INTELLIGENCE.

There is no better glimpse into how fast the future is approaching than recent advancements in artificial intelligence. Tesla's self-driving car, DeepMind's AlphaGo, Siri or Alexa. AI is already all around us.

There are a lot of mathematical techniques under the broad tent of AI, but the most common is machine learning. And at its heart, machine learning is applied statistics. In fact, Stanford statistician and ML expert Robert Tibshirani, calls machine learning "glorified statistics."

In the previous chapters of this guide, we've seen all the statistical tools at your disposal to help you understand your business better—central values, regressions, probabilities, functions, testing, and so on. Individually, these are all incredibly useful. But combining them into a single applied technique would allow you to significantly increase your understanding of your business.

In this chapter of ***The Complete Guide to Statistics for the SaaS Executive***, we are doing exactly that through machine learning. Though machine learning and its big sister, artificial intelligence, are seen as sophisticated computational techniques, at their heart they are still math, and a lot of the basics of ML is statistical modeling by another name.

Here, we're going to show you how machine learning works, how you might use it in your business, and how statistics allow computers to learn.

USING MACHINE LEARNING IN BUSINESS

Tom Tunguz identifies four ways machine learning can be used in SaaS. Each of these are learning problems:

Optimize - this morning, fastest way to travel from Sand Hill Road to South Park in San Francisco is highway 101. The job requisition for an account executive on our website uses too many clichés. To close more business, speak slower, talk about pricing later in the call, and use this case study.

Identify objects - the photograph you just took with your smartphone contains a cat. Find all red plaid woolen shorts in an ecommerce store. The CT scan shows high likelihood of Parkinson's Disease.

Detect anomalies - your credit card shows a $10,000 charge for a piano from a store in Nairobi. Your server cluster is operating at historically high CPU usage. Customers are responding to this morning's lead generation email at 25% greater rates than last week's campaign.

Segment data - customers who come to our product through the mobile app store show 15% higher engagement.

Let's say that you are a SaaS company, and want to understand how much different types of customers will pay for your product. There are two ways to do this:

- **You can ask them.** Through market research and willingness to pay surveys, you can find the price points that work for them.
- **You can use your data.** You already have data on your current customers and how much they pay, so you can deduce from that what other, similar customers would pay.

Deduce. What does that mean? For a human, it's pretty easy. If you sat down with all the data in a spreadsheet, you could probably look through and in five minutes have a rough idea of what different customers pay. You, personally, could deduce it.

But that's not exactly efficient. For one, it requires you to sit in front of a spreadsheet. As a SaaS exec, you have better things to do. Secondly, it only works for a small amount of data. Maybe you can deduce patterns in 20 current customers. But what about 200? Or 2,000? This data problem isn't just in the numbers of rows, but also the number of columns. Maybe you see the patterns in size of company and pricing, but what if you try to use all your variables—size, industry, usage, location, growth, and so on.

Finally, even if you were some superhuman, spreadsheet supremo, this vast knowledge and ability would all be locked in you. As your company scaled, you would still be the only one able to see the patterns in the data.

You can see how this takes in at least three of Tunguz's possible applications. We are trying to optimize our pricing, segment our customers, and potentially detect outliers (who might need custom pricing).

SaaS problems are ripe for machine learning because of the wealth of data we have. Every single customer and interaction is a new data point to aid our understanding of what customers want and what works, and what doesn't in our business.

Machine learning loves data. This is one of the beauties of the-approach—you can use all your data. The algorithm will naturally weight the most important variables stronger. This doesn't mean you should add everything into the pot and let your algorithm do its thing. The more noise you add, the more difficult to derive the signal. But you can add masses of data to see if it makes a difference. The algorithm can scale.

As with most things in SaaS, scaling is the problem.

USE MACHINES TO LEARN AT SCALE

As with most things scaling, machines are the answer.

The human brain is an excellent learning device. They far, far outpace computers in learning abilities. But the human brain does have a limit—it just isn't built to scale for large amounts of data.

Computers can scale to process vast amounts of data. But until recently, they have been pretty dumb. Machine learning allows them to get smarter. The ultimate aim for AI research is to develop machines that have a "feel" for the intuitive answer just as humans do. But they will be able to get to this answer using vast amounts of data covering millions of variables.

Let's go back to our initial problem—how can we predict how much a company will pay? You have certain information about your current customers that will help you work this out: size, usage, industry, and the price they currently pay:

Column 1	Column 2	Column 3	Column 4	Column 5
Company	Company Size	Usage	Industry	Price
A	10	2000	Tech	250
B	20	800	Media	300
C	20	850	Tech	150
D	30	550	Tech	100
E	40	2000	Sales	150

What you want to know is this:

Column 1	*Column 2*	*Column 3*	*Column 4*	*Column 5*
Company	*Company Size*	*Usage*	*Industry*	*Price*
F	*30*	*2000*	*Media*	*???*

For any new company, if you have the first few columns of data, can you predict the price?

Intuitively, your human learning tells you that the price is an outcome of all of those initial variables combined in a certain way. Say company size is the most important, then industry, then usage. Your brain is effectively following this procedure:

- Take each variable
- Multiply it by a weight that signals its importance
- Add all the variables*weights together to get the price

You are doing this in an abstract way, not with actual numbers. But that is what your brain is doing. As you get more data in, say more media companies sign up at certain price points, you update your "weighting" for that data. You have this constant computation happening in your brain. You can then predict the price of company F.

Machine learning does all this, but with actual numbers. It literally assigns weighting to each variable to try and get the right price. The "try" is the learning part. You let the machine learning algorithm try on some training data, where you already know the price (companies A through E in this case) again and again and again, until it gets it right, or as close to right as is possible.

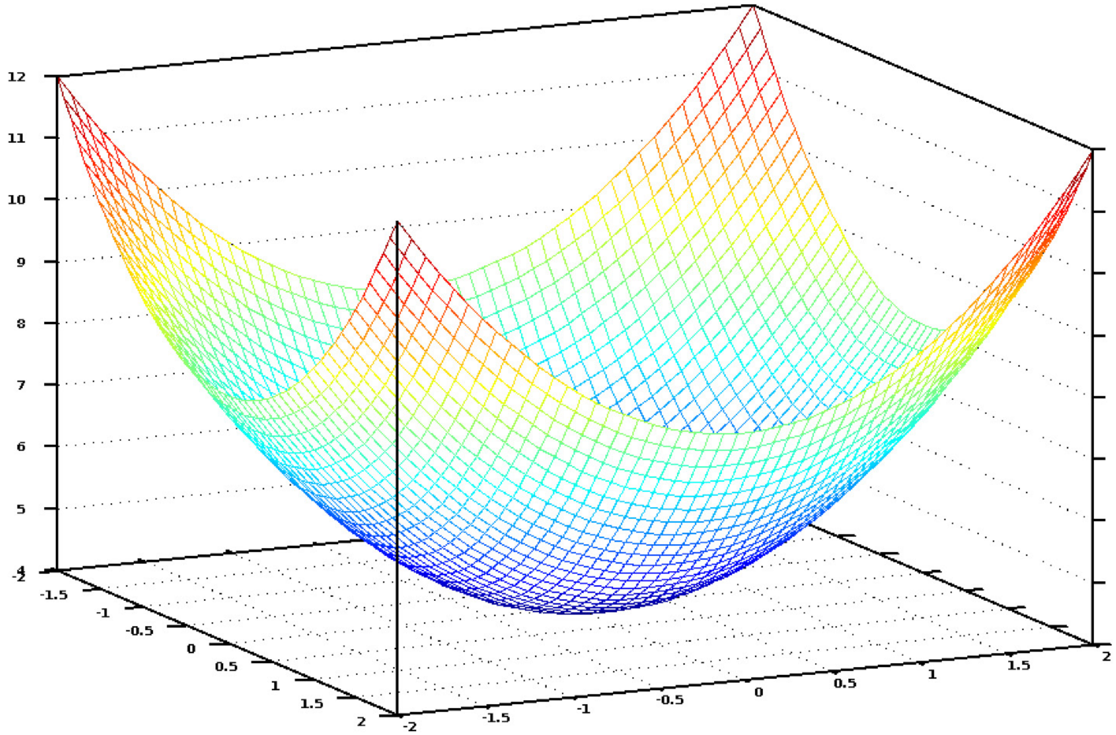
THE BASICS OF ML

A basic machine learning algorithm looks like this:

1. Start with random weights for each of the variables, say size_weight = 1, usage_weight = 1, and industry_weight = 1 and then compute the price with these weights.
2. Compute the cost function for these weights. This is the amount of error in the price with these weights from the right price.
3. Try it with all possible weights. You want to minimize the error, or cost, of your calculation. Whichever has the lowest cost, wins!

For a computer, trying millions of possible calculations is a trivial exercise. It can iterate through this in a few seconds. But we said all possible weights, which isn't millions of possible calculations. It is infinite possible calculations. Which will take more than a few seconds.

But math comes back with a trick for this as well—gradient descent. Gradient descent is the learning part of this algorithm. You can imagine for every combination of weights for size, usage, and industry, there is a respective cost. You could graph out some of these related to that cost and you would get a graph looking like this:



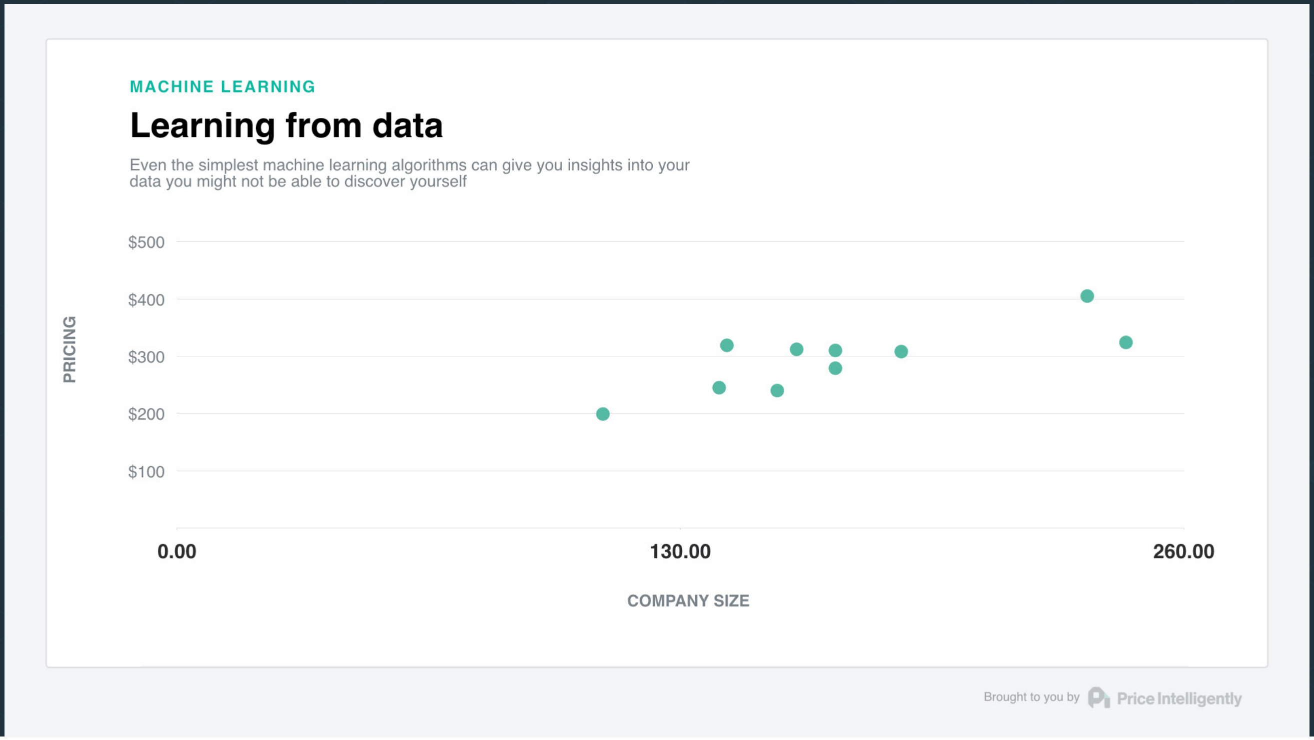
In this case the horizontal axes might be the size and usage weights, and the vertical axis would be the cost. You want the combination of weights that puts you in the deepest part of that bowl.

Like we say, you could try all possible weights. But that wouldn't be efficient. Instead, gradient descent allows you to start from wherever in that bowl and it will automatically find its way down-hill. To do this, the algorithm above becomes this:

(Source: machine learning is fun)

1. Initialize with some random weights
2. Compute the cost function for these weights
3. Calculate the partial derivatives for the cost function for each of these weights
4. Subtract those values from our weights, then recompute the cost function

You then continue steps 2 through 4 until you can't get your cost function any lower. The partial derivative tells the slope of the function, and it is from that the algorithm can work out which way is down.



LEARNING FROM YOUR DATA

Let's run through a simple example of this using just one variable—company size.

In this scenario, customers are paying on a usage tariff. We want to see whether we can discover a relationship between the company size and the price they ultimately pay. We have ten companies in our training set.

This is a good example of us trying to teach machines our natural intuition. Any person looking at that can draw a trendline through those points and interpolate/extrapolate the relationship between size and price. But to a machine, this is just single points of data—it doesn't understand the relationship, so it has to learn.

But the machine could follow the algorithm above to learn about 10,000,000 data points instead of 10, which a human couldn't do, and could do it with dozens of variables, which a human can't even visualize.

The next step for the machine is to initialize the weights, in this case a and b, the intersect and the slope (You might recognize this from chapter 3 on trends. Ultimately, this basic version of machine learning is just linear regression, which we met in chapter three. It is a special version, called multivariate linear regression. Instead of using just a single input variable, we use multiple to decide the outcome. Here we have condensed it down to a single variable again for easy visualization. But this algorithm would work no matter how many variables you used).

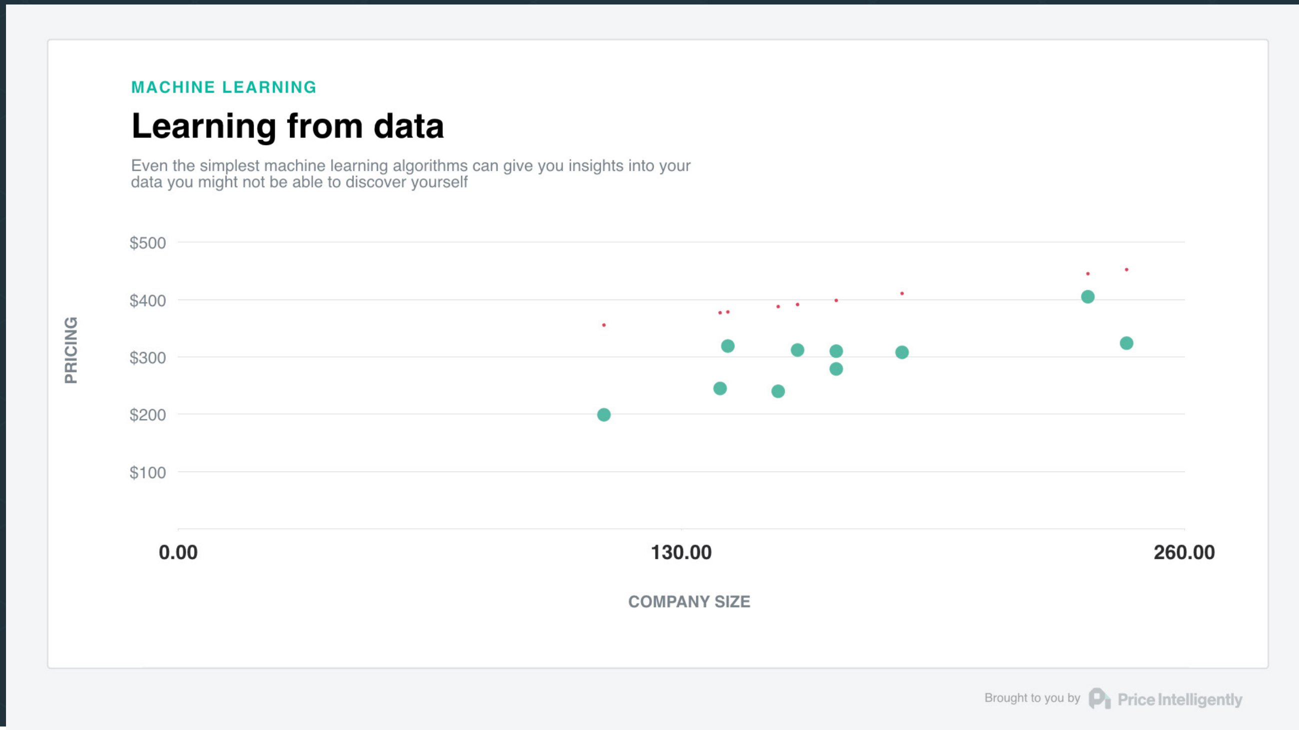
The great thing about this is that you don't need to have any prior knowledge to set these weights. They can be anything. The machine is going to learn they are wrong. Let's set a = 0.76 and b = 0.47 (At this point we want to normalize our data to a scale between 0 and 1 to make the calculations easier, which is why a and b are so low here. We do that using minmax normalization). We then use the linear equation to predict our first prices from the size data:

$$\hat{y} = a + bx$$

Where y-hat is our predicted value. If we did that for the data above, we'd end up with this data:

Column 1	*Column 2*
Company Size	*Pricing*
110	*$355.56*
140	*$377.08*
142	*$378.51*
155	*$387.83*
160	*$391.42*
170	*$398.59*
170	*$398.59*
187	*$410.78*
235	*$445.21*
245	*$452.38*

When we plot this data (the small red dots), we can see that they are too high, and the slope isn't quite aligned:



01

Next we want to compute the cost function, the error, between this data and the real data. The equation for the cost function is this:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

02

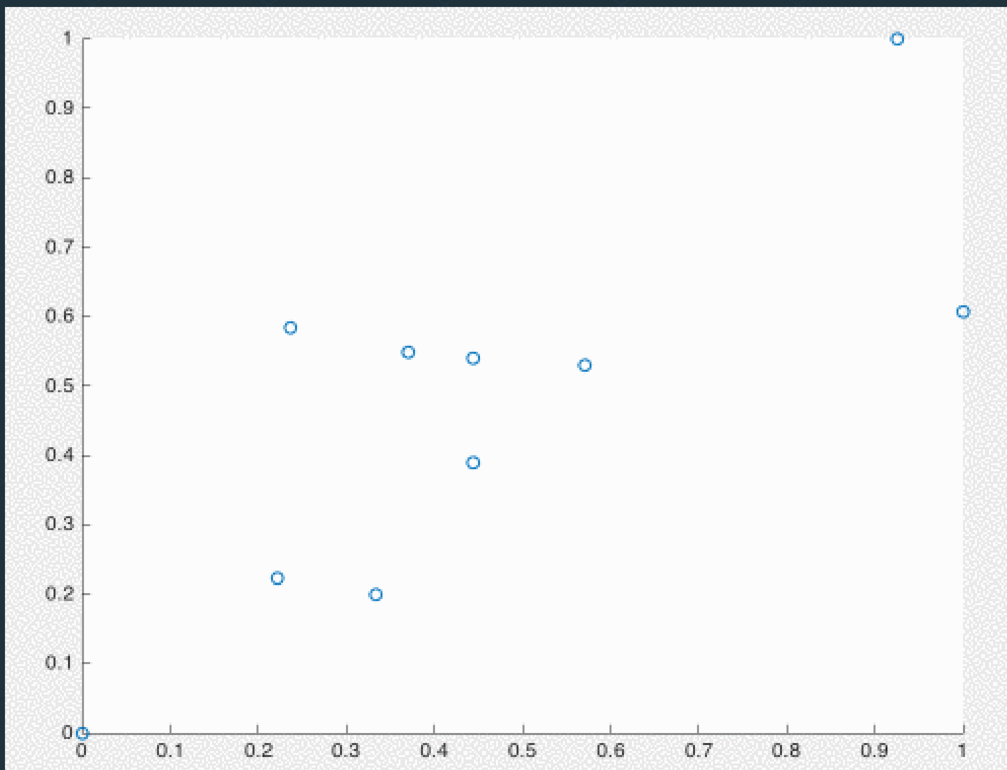
This looks confusing, but it isn't. It goes through each value, looks for the difference between what is actual and what is predicted, squares that difference, then sums up all of those differences for the ten values—the sum of the squared errors. This is what we want to minimize. To do this, we need to calculate the partial derivatives for a and b for this function. Again, this is easier than it sounds. The equations for this are:

$$\delta a = -(y - \hat{y}) \qquad \delta b = -(y - \hat{y})x$$

03

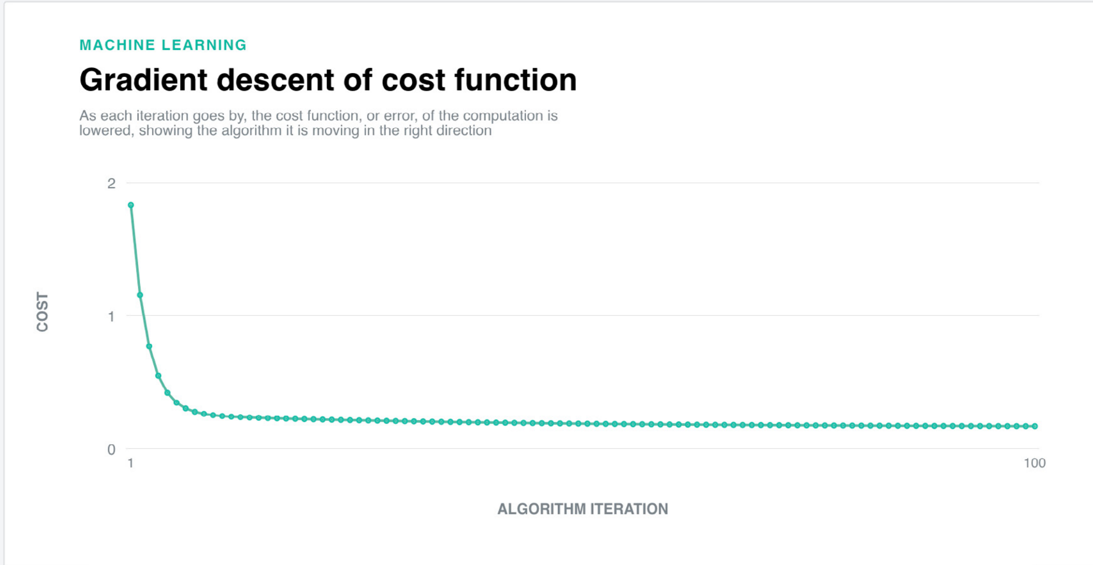
Once you have calculated these, you then use them to compute a and b again. So in this case a will become:

$$\delta a = -(y - \hat{y}) \qquad \delta b = -(y - \hat{y})x$$



The 0.01 is called the learning rate and sets how quickly the algorithm will move in each step (how quickly it can learn). Set it too high and it will miss the exact values that fit the equation best. Set it too low and the algorithm will take longer to converge (find the right answer).

So we've moved a, the intersect down slightly from 0.75 to 0.71, which is in the right direction. You then do the same for b. Then, you run the algorithm again, but with the new numbers. You do this over and over again, until the data aligns. Here is a gif of this computation being run 100 times on this data:



Brought to you by Price Intelligently

First the line drops to intersect with the data, then it shifts angles to align with the data. The machine is learning the line of best fit for this data.

The concept of gradient decent can be seen well if we plot the cost function for this algorithm:

That's it. You have trained a machine learning algorithm to understand data. Once you have your final prediction, you can use that to interpolate or extrapolate to find the relationship between company size or cost.

Of course, you can go one better than that and expand from a few data to thousands of data points across multiple dimensions. The concepts are exactly the same. To learn how to take this further, these resources are great for explaining machine learning for a regular audience.

Let's look at what we've used to build our machine learning:

- In the cost function, we are using basic statistical values such as **means** and **errors** to understand what is happening with our data.
- The version of machine learning we have described here is an extension of linear regression, known as multivariate **linear regression.**
- The outcome of any machine learning algorithm can be thought of as a probability. Though that is not as obvious with this example, consider an algorithm that segments your customers. Your calculation will give you the **probability** that a customer is in one segment or another
- Each run of the cost function above is a test, where we test our **hypothesis**, h(theta) against the real answer.
- Finally, we are using a cost **function** to understand the error in our algorithm, and calculus to help us minimize that error.

Some of these ideas are tangential to statistics, some are core. But they should show you that as impressive as machine learning and other advanced computational techniques are today, they are still based on ideas that you can completely understand. If you can calculate means and errors in your business data, you are on your way to building AI that will provide completely new insights into your company for you and your customers.

THE STATISTICAL BASIS OF THE FUTURE

This is only a taster of what AI can do. We haven't hit on deep learning or neural networks, which is where the real promise of AI research is. But the algorithm described above is bonafide machine learning, and 99% of problems can be solved with this approach. Whenever you hear of a company using AI as their underlying tech, they are probably just running a lot of gradient descents on a lot of multivariate linear regressions with a lot of data.

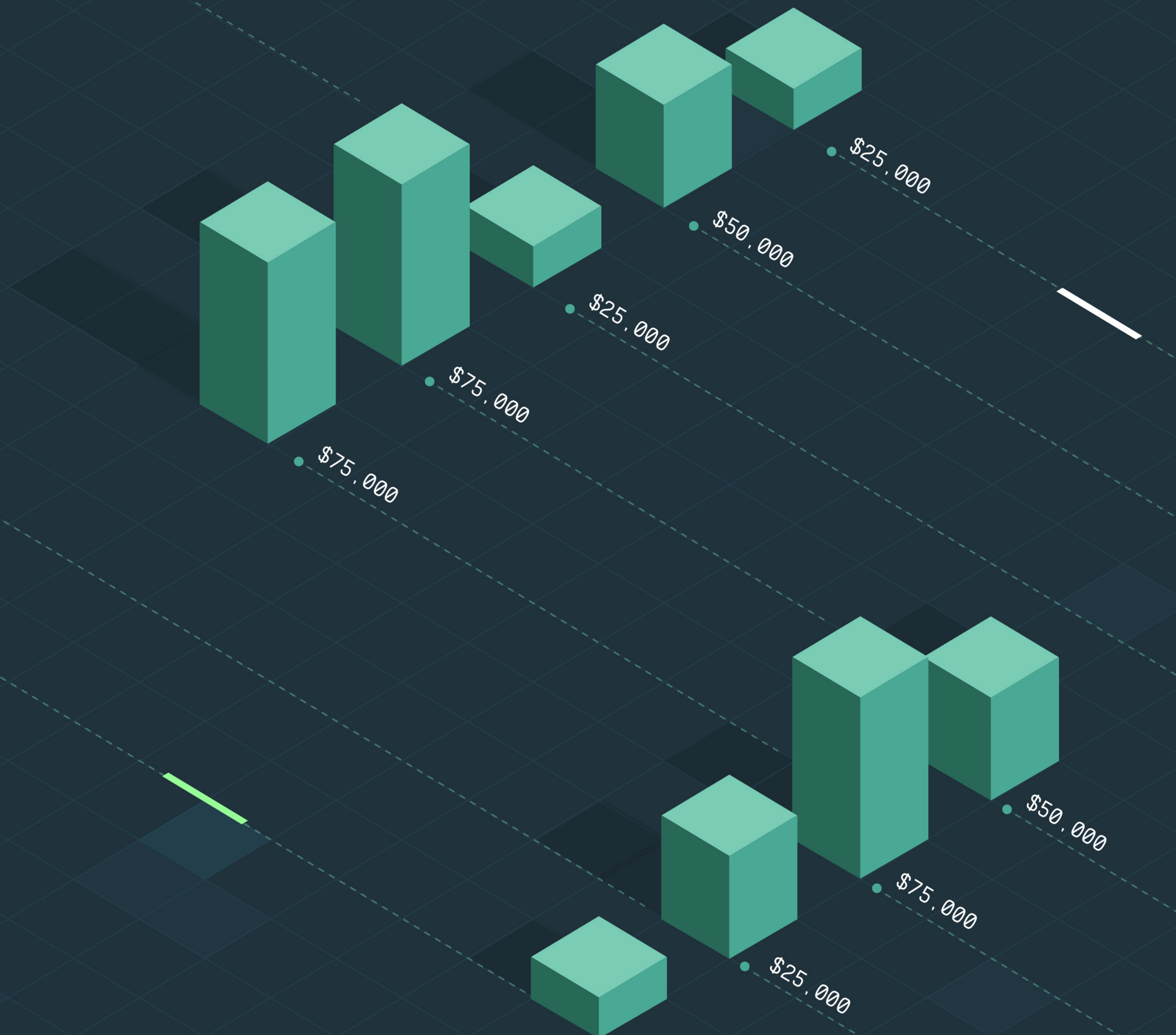
But even neural nets and deep learning are based on the same concepts—functions, probabilities, minimizing costs. Now that you have an understanding of these concepts, you have an understanding of AI.

But even better, you have a greater understanding of your business. From simple averages through complicated growth functions to ML-based data mining, when you understand math, you understand your world.

THE

COMPLETE GUIDE TO STATISTICS

FOR THE SAAS EXECUTIVE



ProfitWell™