"Storytelling with Data — Prudent data analysis to solve real-world problems"

Anurag Sahay
Nagarro

nagarro

nagarro

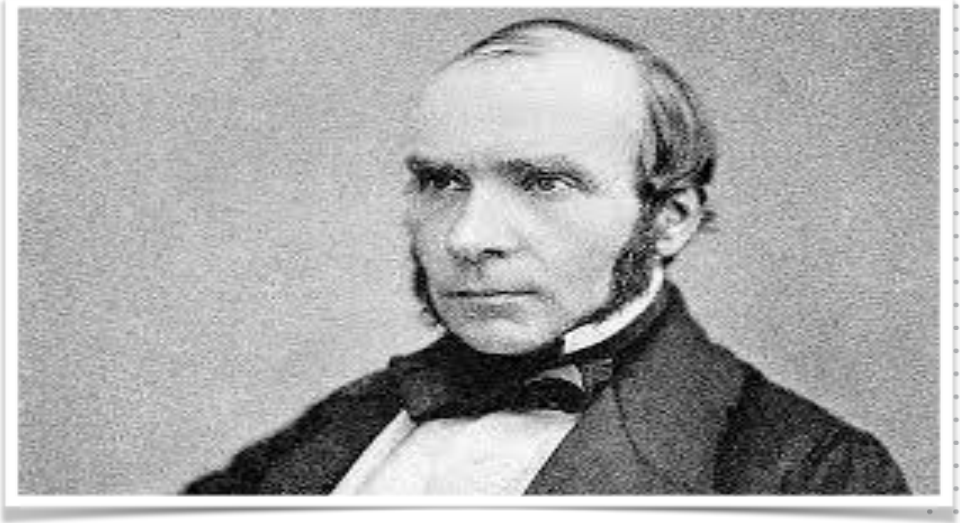
THE 3 DOORS TO AI

a thinking & design framework

Door #1

through insight

“focusing on insight” to find the solution to the problem



London in the 1850s

1	1–A few farmers
50	50–100
140	45–60,000
300	10–20,000
800	10–12,000
1000	20–25,000
1100	10–20,000
1200	20–25,000
1300	80–100,000
1350	25–50,000
1500	50–100,000
1550	120,000
1600	200,000
1650	350,000-400,000
1700	550,000-600,000
1750	700,000
1801	959,300
1831	1,655,000
1851	2,363,000

a population that grew very fast between 1801 and 1850

Thames was an open sewer and the local sewage system was very sparse.

1801	1,096,784	1,764	959,310
1811	1,303,564	2,097	1,139,355
1821	1,573,210	2,530	1,379,543
1831	1,878,229	3,021	1,655,582
1841	2,207,653	3,551	1,949,277
1851	2,651,939	4,266	2,363,341

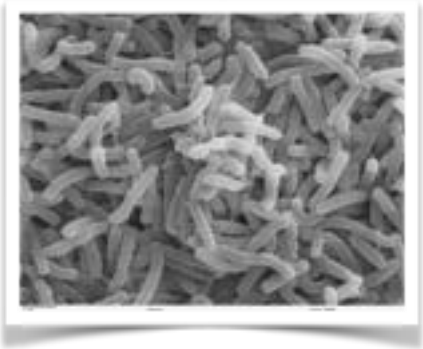
The density of population in inner London was approximately 5 people to a room

nagarro

a cholera epidemic

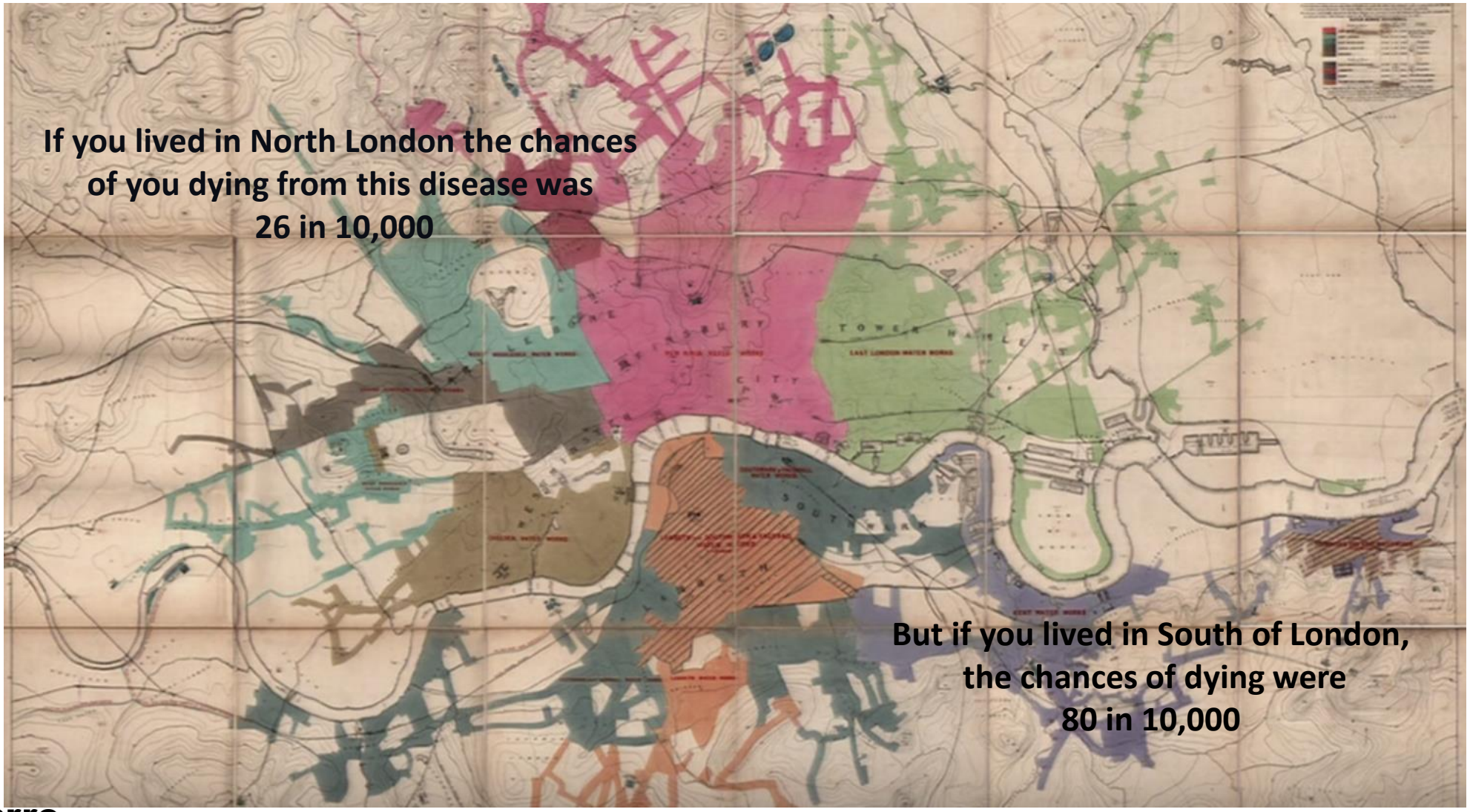
1831 — first recorded cholera death in London
1832-1833. — 20,000 dead
1840s - 50,000 dead

There was no germ theory of medicine. In those days, people believed that ‘bad air (miasma)’ was the cause of the disease.



vibrio cholera

evidence (aggregate statistics) but inconclusive



nagarro

evidence (aggregate statistics) but inconclusive

"[They] furnish no proof whatever of the correctness of his views."

-London Medical Gazette, 1849

A experiment was suggested and it fatefully happened…

"The *experimentum crucis* would be, that the water conveyed to a distant locality, … produced the disease in all who used it, while those who did not use it, escaped."

-London Medical Gazette, 1849

28th august 1954

Francis Lewis (5 year old) contracted cholera. Her mother while waiting for the Doctor, washed her diaper and threw away the dirty water in a cess-pit outside the door on 40th Broadstreet

The cess pit was extremely close to a water well. The cess pit was not properly maintained and it very soon spoiled the water well

In the next 3 days hundreds of people fell sick. In the next 2 days about 300 of them died.

On the 5th day, John Snow collected water from 5 wells in that neighborhood.

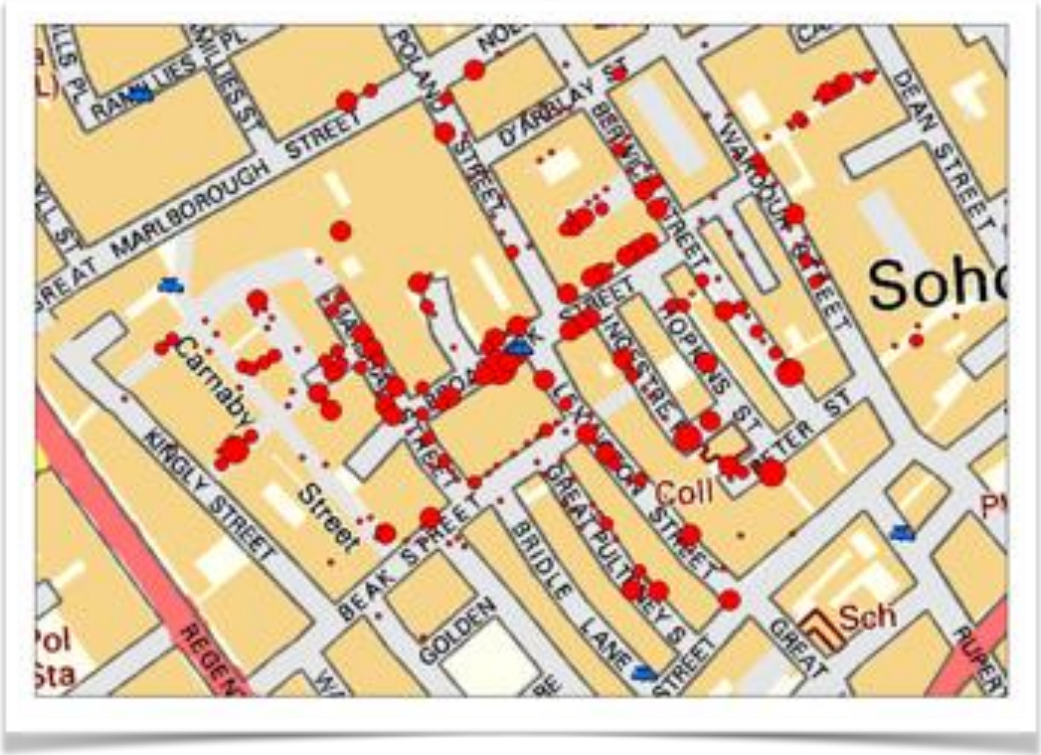
He tried to see the water under the microscope but the microscope was not powerful enough to reveal anything significant

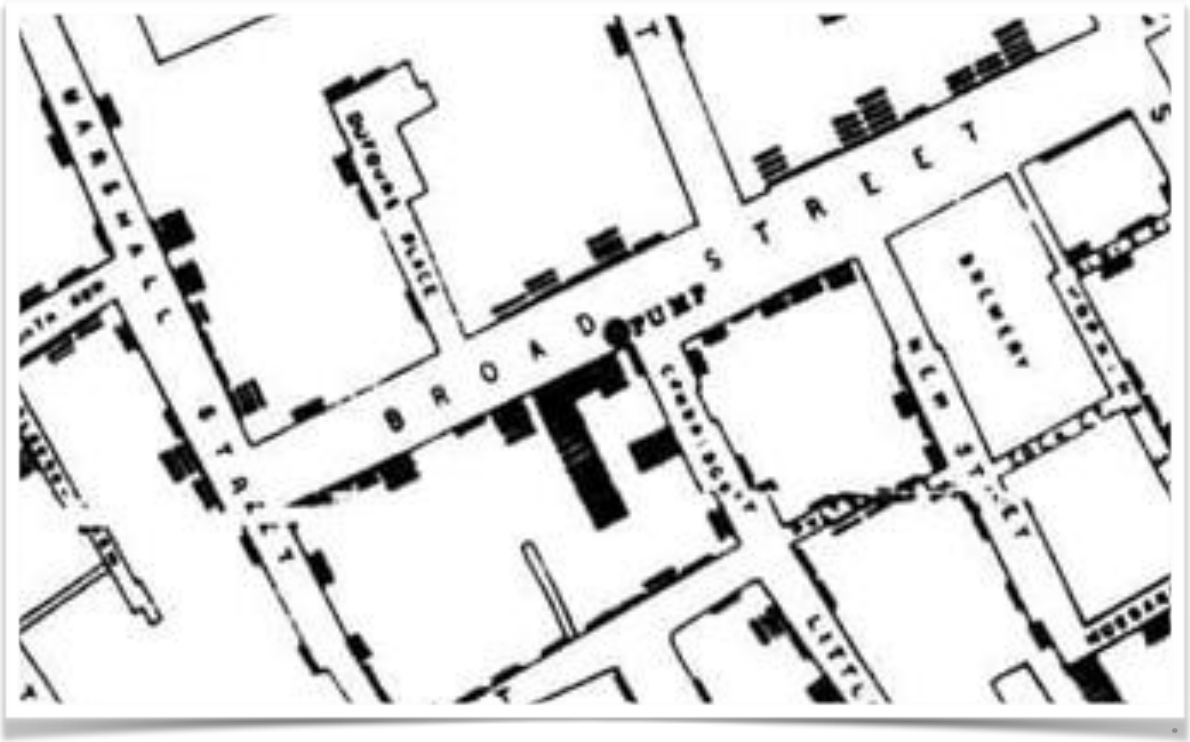
He walked to the 'Registrar of Deaths' and collected the residential information of all the deceased people.

And he plotted the deaths on top of the city map.

nagarro

Regions of high death (data distribution)





Analyzing the exceptions…

There was a neighborhood where only 3 children had died. They all had all gone to a school in the Broad Street district and the school water supply was infected

Deep Dive
Into the Data

There was a brewery in Broad Street area where no one had died. The staff in the brewery were allowed to drink beer by the management.

There was a woman who had died in an area far way from the Broad street region. This woman had grown up in the Broad street area and preferred the taste of water from that region. As a result, her sons had arranged for a regular supply of water to be picked up from the broadsheet region.

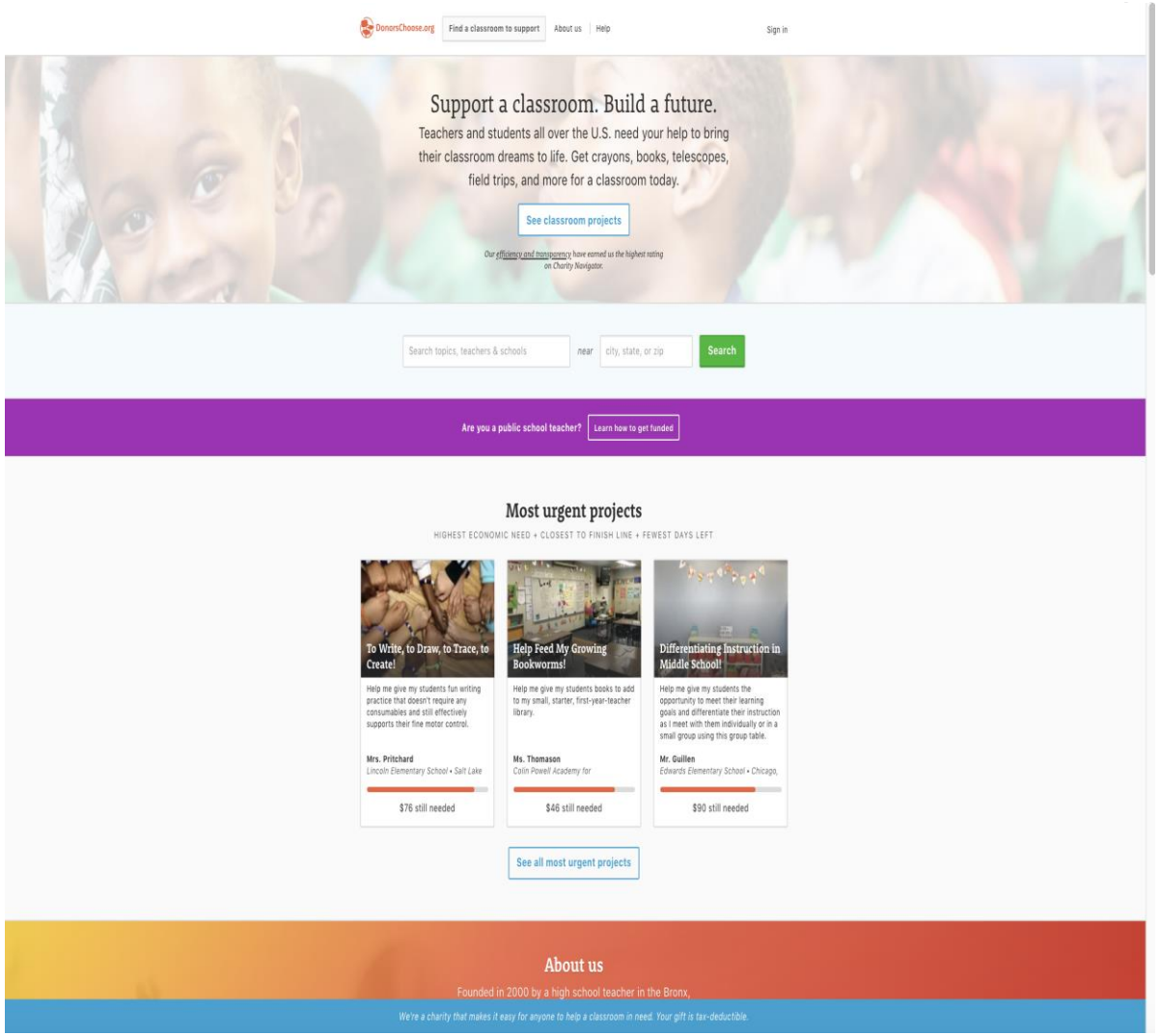
the infamous pump and the famous pub

With data comes data quality…

Door #2

through efficiency

“focusing on efficiency” to find the solution to the problem



nagarro

the needs of a classroom…

We were started by a history teacher. In 2000, Charles Best, a teacher at a Bronx public high school, wanted his students to read *Little House on the Prairie*. As he was making photocopies of the one book he could procure, Charles thought about all the money he and his colleagues were spending on books, art supplies, and other materials. And he figured there were people out there who'd want to help — if they could see where their money was going. Charles sketched out a website where teachers could post classroom project requests, and donors could choose the ones they wanted to support. His colleagues posted the first 11 requests. Then it spread. Today, we're open to every public school in America.

complex decision making & speedy execution

Most Needy

How do you choose a which proposals to fund and which to not

How do you do it in ‘reasonable time’

Most Impactful

Most Actionable

Topic Bias

People (Volunteers) have biases. The chances of a proposal getting approved/rejected depends significantly on these biases.

What do you do about that

Regional Bias

Impact Bias

nagarro

build a virtuous circle...

★ Modelling

Use Supervised Machine Learning to build a model that is trained on this curated Dataset. The model learns the important 'features' to focus on to make a decision

★ Dataset

Build a curated set of examples where experts take a proposal and accept/reject it based on effortful, unbiased decision-making.

★ Consumption

Use the model for all proposals. The model is very confident on some Proposals and not on some.

Use Volunteers to help with the Proposals where the model Is not very confident

nagarro

designing a good dataset…

Person/Teacher making the submission

Title of the Person (Mr, Mrs. Ms,..)

School state/district

Date Time of submission

Project for what Grade (Pre KG,..)

Category of Project (Music, Arts, Science..)

Sub Category of Project (Piano, Algebra..)

Project Resource Summary

Previously posted Projects by the same Teacher

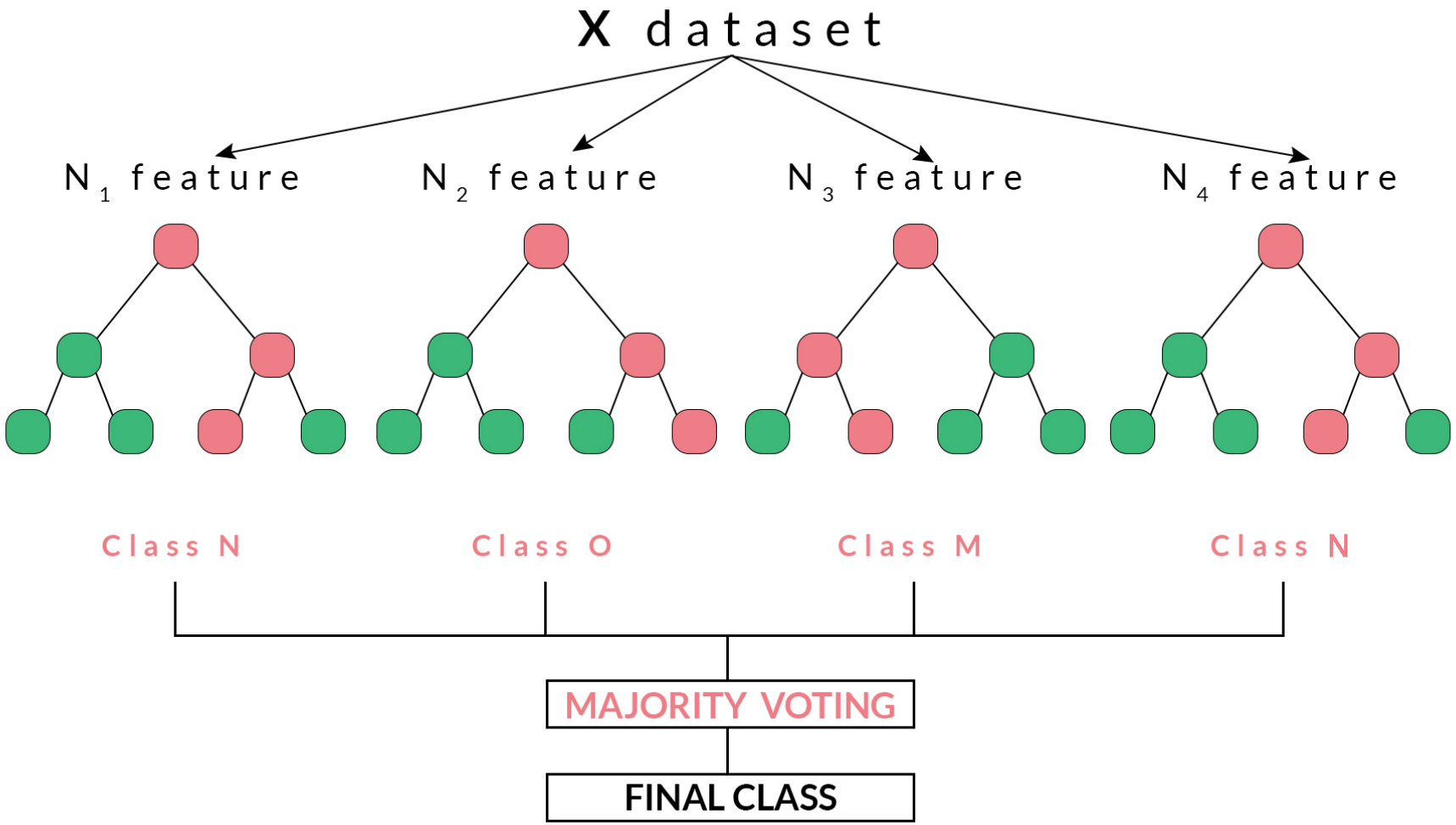
Essay 1
Introduce us to your classroom

Essay 2
Tell us about your students

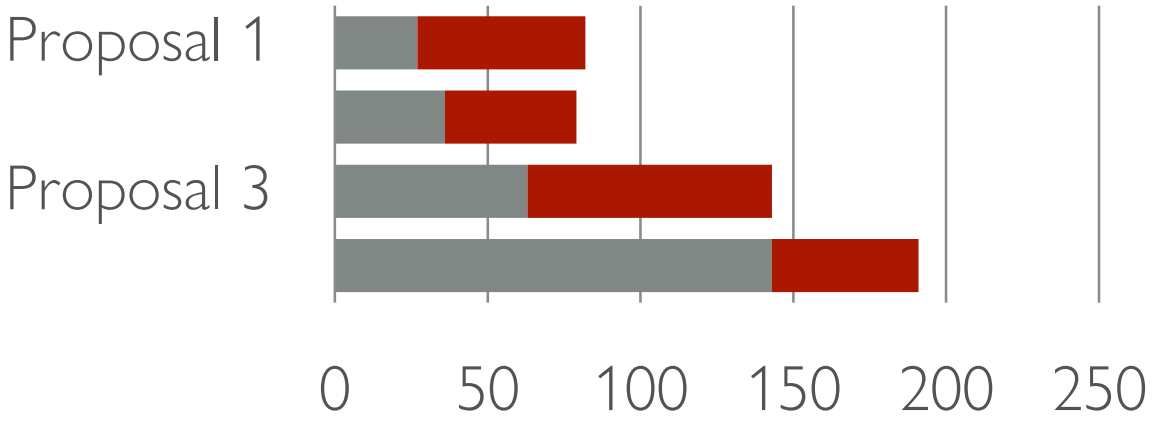
Essay 3
Describe how your students will use the resources requested

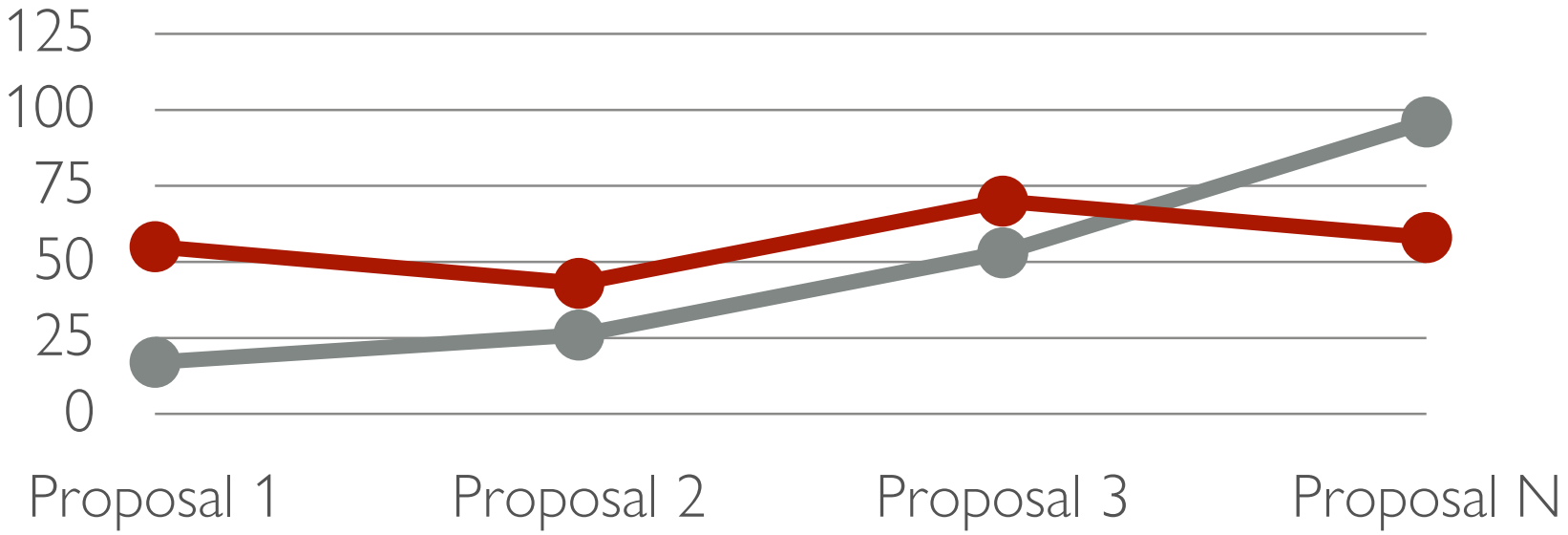
Essay 4
How will your project make a difference

building a good model for the dataset…



transparent and explainable…





nagarro

scale and efficiency…

$874,656,099
dollars raised

1,502,445
projects funded

35,920,275
students helped

82,785
schools participating

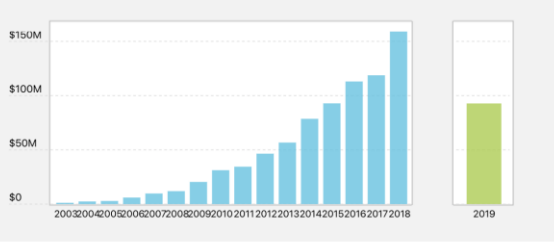
533,408
teachers funded

3,987,426
supporters

***94% of teachers** said their funded projects increased their effectiveness in the classroom.*

The majority of these supporters are first-time givers to public schools.

Project funding per calendar year



83%
of all the public schools in America have at least one teacher who has posted a project on DonorsChoose.org

84%
of schools where half or more of students are from low-income households have participated in DonorsChoose.org

63%
of projects are fully funded.

76%
of funded projects are from schools where half or more of students are from low-income households

Average donation size for new donors:
$53

Average cost of a funded project:
$551

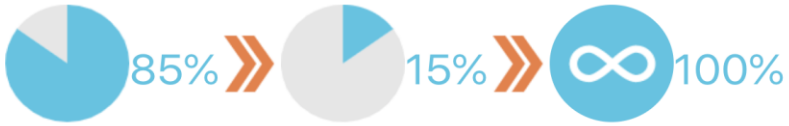
Average wait time of a fully funded project:
27 days

Project requests currently seeking funding:
95,474

Number of staff members:
110

Number of volunteers:
176

85% of donors allocate 15% of their donations to support DonorsChoose.org, making us 100% self-sustaining!



last updated September 10, 2019

nagarro

with modelling comes outcome…

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

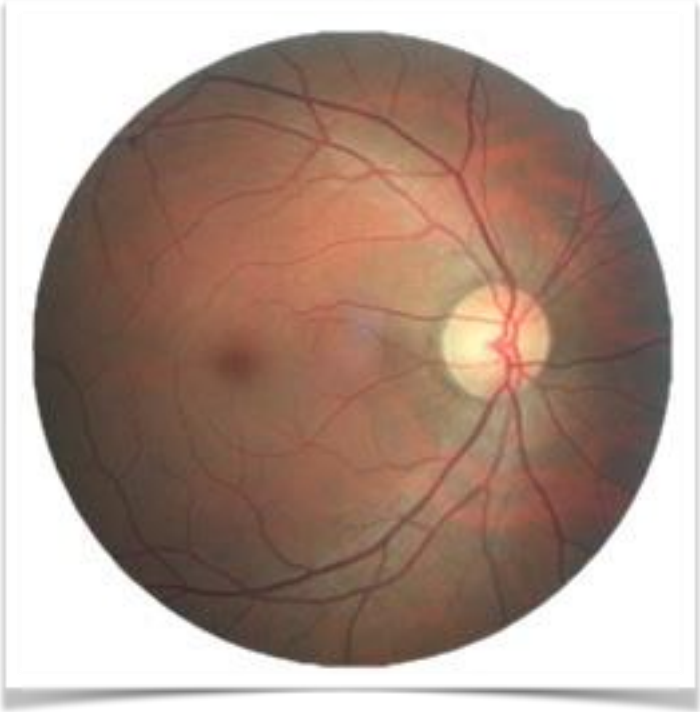
$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$

Predicted \ Actual	Pos	Neg
Pos	TP	FP
Neg	FN	TN

Door #3

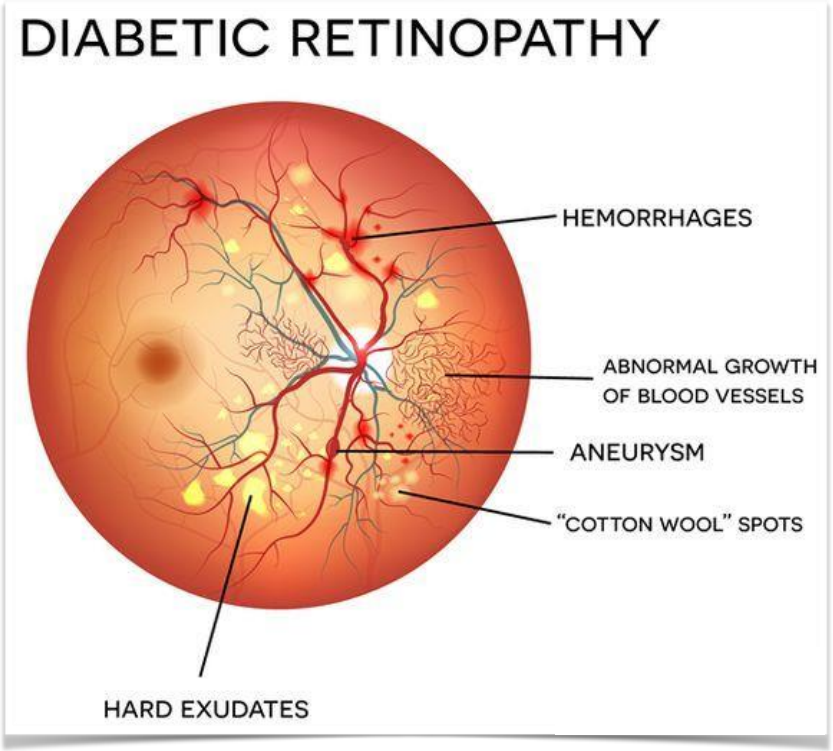
through disruption

“focusing on paradigm” to find the solution to the problem



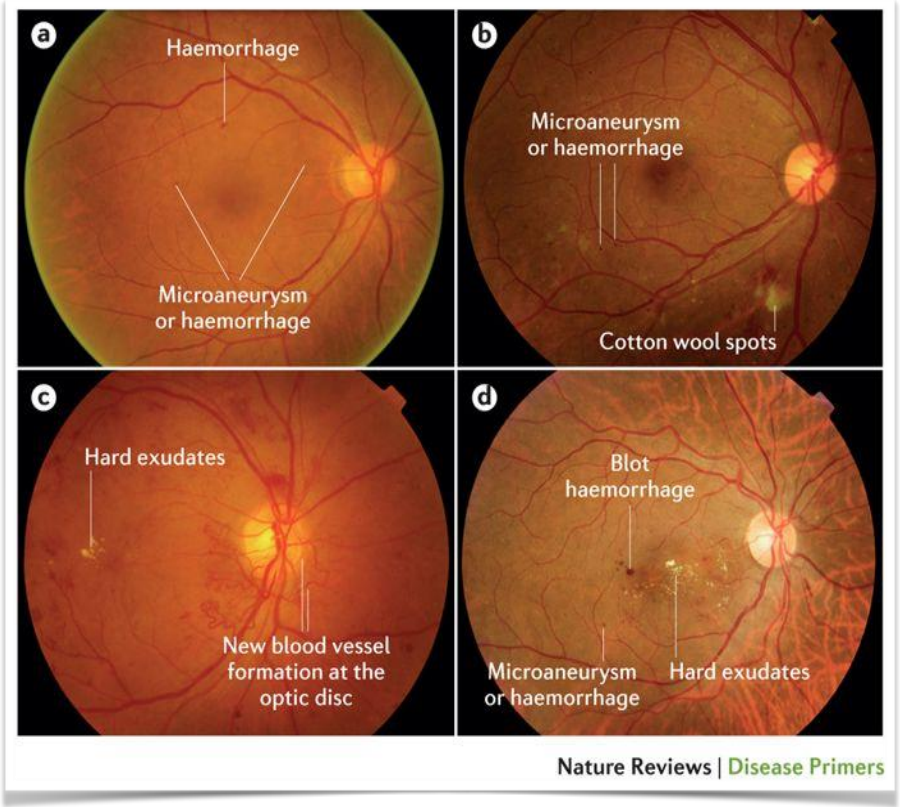
It’s a situation

1. Leading cause of blindness in the working population in the world
2. Affects 93 million people worldwide
3. About 40% - 45% of people with diabetes have this disease
4. Diagnosing DR is a time consuming and a manual intensive task which requires a trained clinician to examine and evaluate a fungus photograph of the retina
5. In a country like India - there is one Ophthalmologist for every 90,000 people



micro and sub-micro features…diagnosing difficulty

1. Abnormal blood vessels
2. swelling, blood or fatty deposits in the retina
3. Growth of new blood vessels and scar tissue
4. Bleeding in the clear, jelly like substance that fills the canter of the eye
5. Retinal detachment
6. Abnormalities in the optic nerve



In 2015 – an annotated dataset appeared…

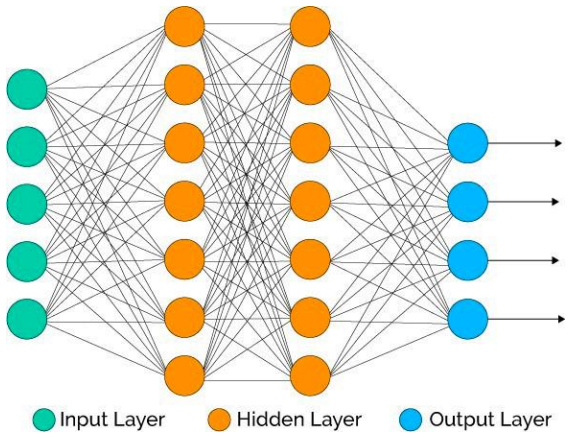
CALIFORNIA HEALTHCARE FOUNDATION

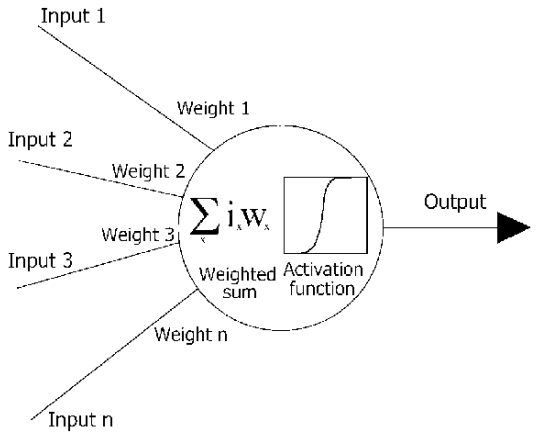
SUPPORTING IDEAS & INNOVATIONS TO IMPROVE HEALTH CARE FOR ALL CALIFORNIANS

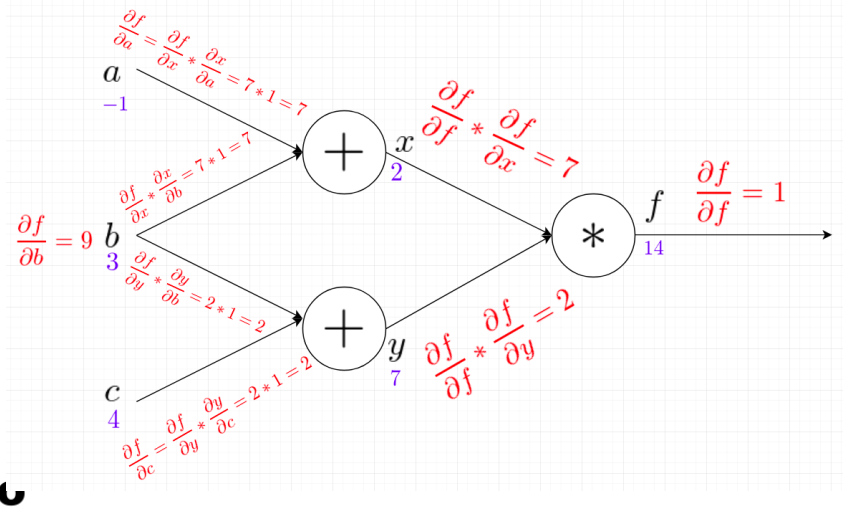
it was made available around the world

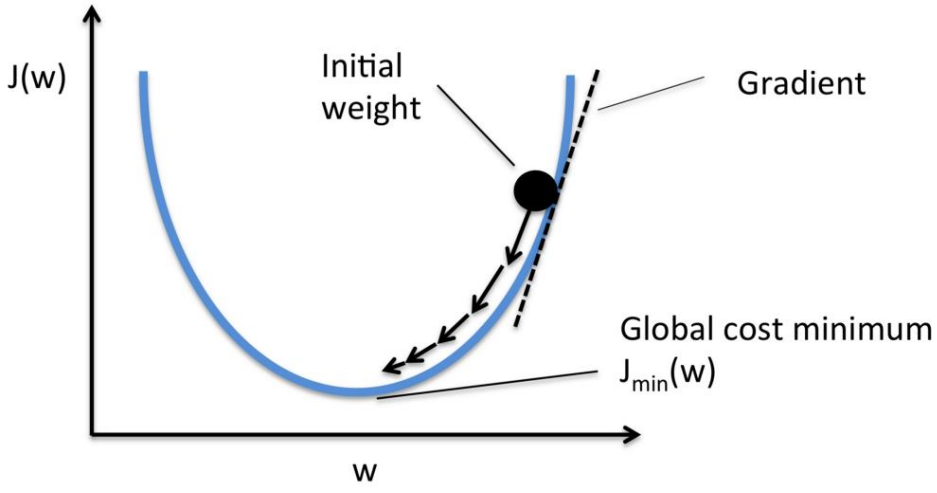
(data democratization)

In 2012 – neural networks made their way back…









In 2016, we released a model with 87% sensitivity…

Actual Severity: 2
Predicted 00 (48.9%): ****
Predicted 01 (30.4%): ***
Predicted 02 (20.3%): **
Predicted 03 (00.3%):
Predicted 04 (00.1%):

Actual Severity: 1
Predicted 00 (82.9%): ********
Predicted 02 (09.9%):
Predicted 01 (06.9%):
Predicted 04 (00.2%):
Predicted 03 (00.1%):

Actual Severity: 3
Predicted 02 (56.7%): *****
Predicted 00 (30.2%): ***
Predicted 01 (11.5%): *
Predicted 03 (00.8%):
Predicted 04 (00.8%):

Actual Severity: 0
Predicted 00 (61.0%): ******
Predicted 02 (21.6%): **
Predicted 01 (17.1%): *
Predicted 03 (00.1%):
Predicted 04 (00.1%):

Actual Severity: 0
Predicted 00 (80.4%): ********
Predicted 02 (10.9%): *
Predicted 01 (08.3%):
Predicted 04 (00.3%):
Predicted 03 (00.0%):

Actual Severity: 0
Predicted 00 (69.7%): ******
Predicted 01 (22.5%): **
Predicted 02 (07.7%):
Predicted 04 (00.1%):
Predicted 03 (00.0%):

Actual Severity: 0
Predicted 00 (72.7%): *******
Predicted 01 (14.0%): *
Predicted 02 (13.1%): *
Predicted 04 (00.2%):
Predicted 03 (00.1%):

Actual Severity: 0
Predicted 00 (82.6%): ********
Predicted 01 (09.6%):
Predicted 02 (07.5%):
Predicted 04 (00.1%):
Predicted 03 (00.1%):

With deployment comes responsibility…

today, across eye camps in poor areas, thousands of people are being screened for this disease. And there are no ophthalmologists that are present to do this diagnosis.

The three doors to AI…

Problem Solving through Insight

Problem Solving with Intelligent Automation

Problem Solving with Intelligent Disruption

nagarro

START YOUR DATA JOURNEY NOW