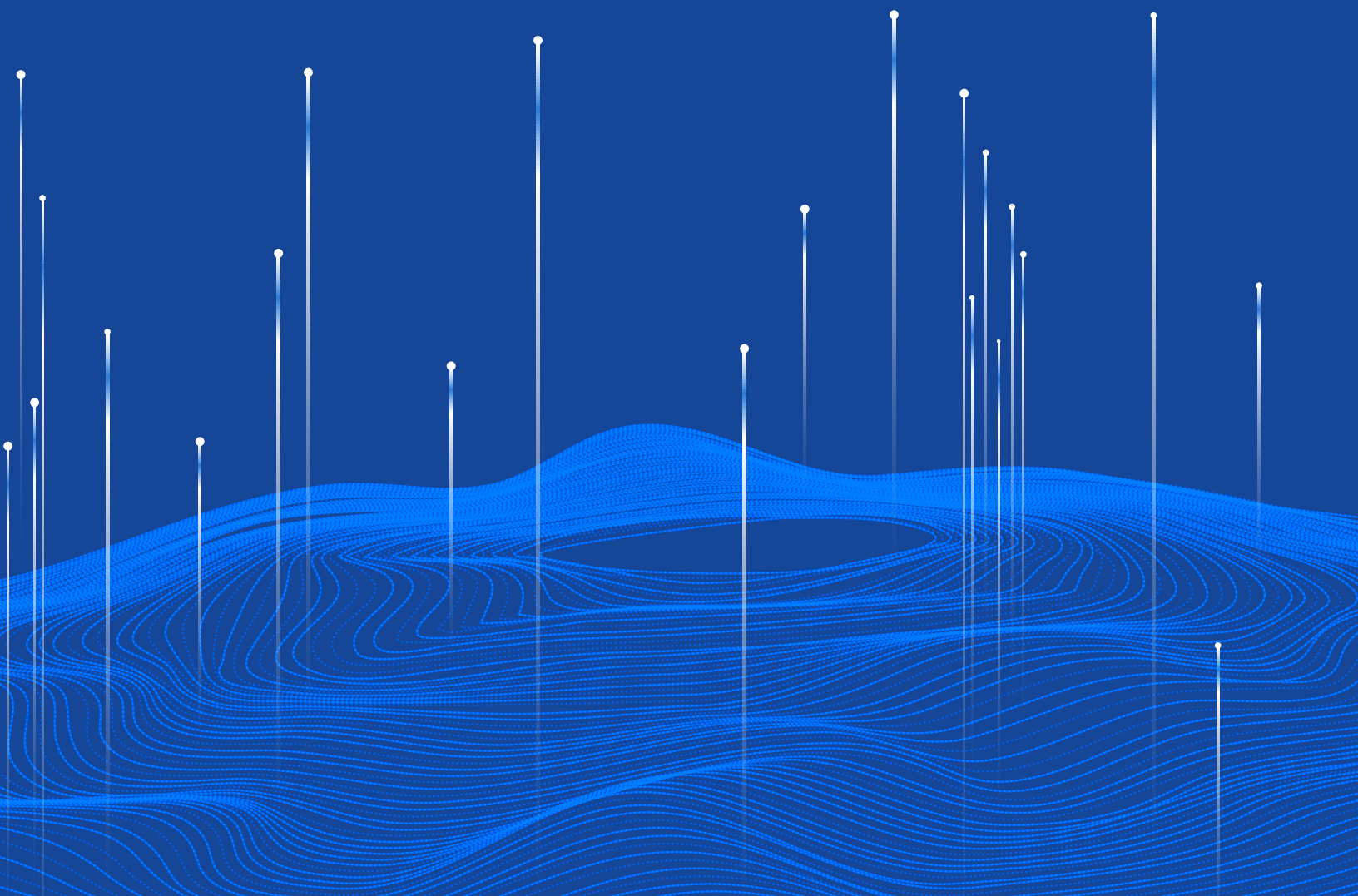
Trends in Data Science 2019/2020

ODSC



Foreword

Across six major conferences in Boston, NYC, Sao Paulo, Bangalore, London, and San Francisco, ODSC hosted an unprecedented 920+ speakers in 2019. It was another year of rapid advances and exciting developments in the fields of data science and artificial intelligence. The year was significant on several fronts.

Frameworks like PyTorch, Keras, and TensorFlow saw major releases that further cemented their position as the leading machine and deep learning tools and featured in many of our hands-on training sessions. However, NLP transformer architectures was one of our speakers' favorite topics in 2019 due to major advances such as OpenAI's GPT-2, Google's BERT, DiDI's Elmo, and FaceBook's RoBERTa. Hugging Face's transformer library which exposes an API to these transformers got a lot of traction also. These pretrained models, especially BERT, have been especially hailed by many as NLPs Imagenet moment. Employing new techniques like bi-directional sequencing and transformers these models are saving data scientists the time and expenses normally required to train NLP models thus marking a major development.

Our new MLOps and Data Engineering focus area tracks coincided with the massive ramp up of efforts to increase the percent of data science projects into production. Tools like MLFlow and Kubeflow continued to grow in popularity. Interest in AutoML grew exponentially in 2019, given its potential as a productivity tool in all stages of the machine learning life cycle. Sessions around Annotation, model interoperability, pipelines, deployment, and testing saw increased interest. This coupled with the fact that 2019 saw a significant drop in the cost of modeling helped accelerate deployment in production environments.

Increased model deployment in the real world raised the importance of Trusted and Responsible AI discussions. IBM's AI Fairness 360 Toolkit and Google's' Differential Privacy library were but two of many projects allowing teams to put responsible AI into practice.

We're incredibly grateful to all our speakers in 2019 for sharing their knowledge and insights with the community. Below are some of their thoughts looking back on 2019 and to the year ahead.

Sheamus McGovern
ODSC Founder

ODSC speakers are all subject-matter experts in their respective fields. From machine and deep learning, to NLP and neural networks, all of our past and current event speakers truly are thought leaders in their specialties. On top of their day-to-day work, they're all proactive in staying up-to-date with the latest advances in the fields of data science and AI, whether that be from reading recent research papers or trying out the latest tools themselves. Because of their demonstrated expertise, we reached out to them to get their thoughts on the biggest updates from 2019, and what they're most excited for in 2020. We asked the following two simple questions, and let them take it away:

What were the major developments, topics or trends in data science in 2019? What are you most excited to see in 2020?

Take your time reviewing their answers, learn about new frameworks, libraries, and languages, read some of their favorite research papers, and get a jump on what they're excited about in 2020.

ODSC

Speaker Commentary

Sihem Romdhani
Data Scientist,
Veeva Systems

"

2019 has been a big year in NLP. Many breakthroughs and developments are occurring at an unprecedented pace. From the Google's BERT and Transformer frameworks to FacebookAI's Multilingual Language Model, we have seen a big performance improvement in a variety of NLP tasks.

"

Self-supervised learning and transformer networks have shown a tremendous success in cross-lingual classification and machine translation, I am really excited to see, in 2020, how to translate this success to image, video and signal processing.

Natasha Latysheva
Machine Learning
Research Engineer,
Welocalize

"

2019 has been a standout year for natural language processing in particular – with new state-of-the-art models being released seemingly every few weeks, huge wins for transfer learning and pre-trained models, and a fun media frenzy over a "too dangerous to release" language model (OpenAI's GPT-2 model). We saw a fairly new neural network architecture (Transformers with self-attention) continue to beat RNN-based approaches across many language tasks – especially language modelling and machine translation.

"

Performance in NLP looks more promising than ever before, and I'm excited to see these advances trickle down into commercial language-based products. We can probably look forward to smarter voice recognition, better autocomplete, smoother translation, and more sensible chatbots – perhaps with more ambitious applications, like free-form dialogue with NPCs in video games.

Speaker Commentary

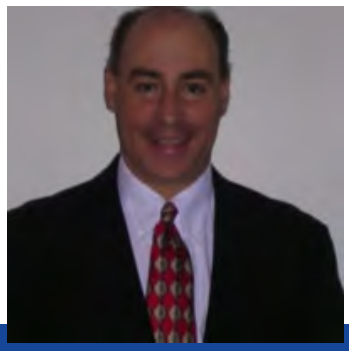
Sudha Subramanian
NLP Enthusiast & AI Solutions Architect, Sirius Computer Solutions

"

2019 has been a big year in NLP. Many breakthroughs and 2019 has seen tremendous growth across various facets of the AI world, including GANs, Reinforcement Learning and NLP. In the realm of NLP, newer and optimized algorithms such as ALBERT and RoBERTa showed to achieved state-of-the-art results on a range of NLP tasks and GPT-2 in the art of synthesizing text based on context.

"

Conversational AI is certainly making its presence in organizations worldwide. Another exciting area where we expect to see a lot of advancements in 2020 is in the field of quantum computing. With the massive speedup of AI tasks that it can provide, it can certainly prove to be a gamechanger.



Steve Geringer
Data Science Solutions Architect, IBM

"

2019 marked a significant turning point in our industry. AI / Data Science is no longer a novelty or science experiment.
AI/DS is now widely recognized as an extremely powerful force in society. It is now firmly implanted in the minds and psyches of the public at large.

"

Moving into this "AI Everywhere" future, we must put protections in place to ensure this tremendous apparatus of power cannot be commandeered by a psychopathic dictator and thereby enslaving the math geeks that created it along with everyone else.

Ben Weber
Distinguished Data Scientist, Zynga

"

2019 was the year of tools democratizing data science, with PySpark and TensorFlow improving the accessibility of building large-scale ML workflows.

"

It'll be interesting to see what techniques from deep learning transfer to other machine learning disciplines. In 2019 I started using deep feature synthesis for shallow learning problems, and I look forward to the broader impact of deep learning research.

Speaker Commentary

Amy Hodler
Graph Analytics, AI Program Mgr., Author, Neo4j

"

2019 was a tipping point of public, private, and government interest in creating guidelines for AI systems that better align to cultural values. In 2020, that momentum will increase and we'll see more frameworks published such as the EU Ethics Guidelines.

"

The data supply chain will become the way to evaluate responsible AI for two reasons: 1) Understanding our data, it's sources and details, is the only way we can build more ethical and unbiased AI. 2) Data lineage and protection against data manipulation is foundational for trustworthy AI.

Jennifer Redmon
Chief Data Evangelist, Cisco Systems, Inc.

"

2019 will be forever known as the year that scientists leveraged AI to image the first supermassive black hole. Although the image itself won't change most people's lives, it demonstrated the power that Narrow AI (computer vision in particular) has to aid humanity in the achievement of seemingly impossible feats.

"

I'm most excited about future developments in explainable AI. Until deep learning (black box) models can articulate how they make decisions, their opacity prevents the awe-inspiring power of deep learning from being leveraged in a multitude of problem spaces, especially those for which bias carries a high cost to individuals or society as a whole.

Juan Manuel Contreras
Data Science Manager, Uber

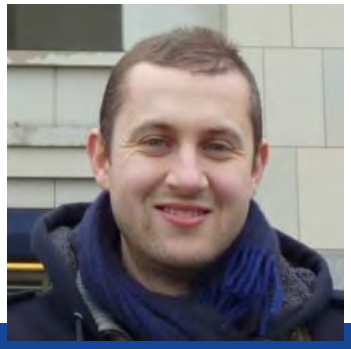
"

A persistent trend is the reckoning of data science-driven automation with its unintended consequences. Our profession continues to discover how machine learning can go bad and how we can work to mitigate these problems.

"

An increasingly healthy and productive alliance of data scientists and the public at large to thoughtfully define how we want to integrate data science-driven automation into our daily lives and what risks we are and are not willing to bear.

Speaker Commentary



Colin Gillespie
Senior Lecturer,
Newcastle University

"

AI is moving away from the mysterious and secretive, to the useful and practical.

"

As machine learning gains traction, more and more companies are using their data in an intelligent fashion. Hopefully reducing waste and cost!

Alex Ng
Senior Data Engineer,
Manifold

"

Along with various development in deep learning and autonomous AI, there has been significant development in the areas of explainable and fair AI. Given the rise in acknowledgement of ethics in AI, these two areas seems to take a front seat in coming years.

"

I am looking forward to new AI offerings by cloud providers. As a practicing data scientist, I would like to see more development in and around quantum computing too.

Pieter Gijsbers
PhD student, Eindhoven
University of Technology

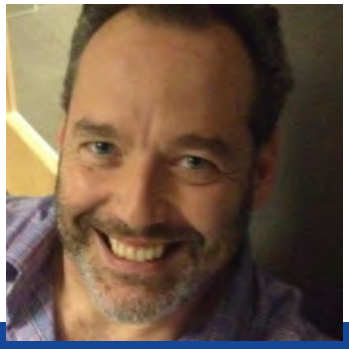
"

AutoML has been gaining popularity over the past years. In 2019 we saw this trend continue with the AutoML workshop being the 2nd biggest at ICML. Where previously research focused on pipeline optimization, there is increasingly more attention for Neural Architecture Search.

"

I'm excited to see how AutoML tools will develop further. This could include expanding to new problem types, working with dirty data, taking into account model interpretability and/or including a human in the loop (Semi-AutoML).

Speaker Commentary



Charles Martin
CEO, Calculation Consulting

"

BERT, ELMo, GPT2, etc. 2019 was the year of AI for NLP. Even Google announced the biggest change in Search Relevance, with BERT, since the Panda update nearly 10 years ago. NLP will never be the same.

"

NLP is where computer vision was 5-6 years ago. 2020 will be the start of the revolution that transforms Enterprise Search, Document Discovery, and numerous other NLP applications. I can't wait to help bring it to life.

Carl Gold
Chief Data Scientist, Zuora

"

Deflation of the hype bubble proclaiming General Artificial Intelligence is imminent, evidenced by wide scale rollback of promises for products like driverless cars and chatbots, and the parade of demonstrations deep neural networks' weakness against "adversarial" conditions that are trivial to humans.

"

Advances in tooling for deploying machine learning systems in production and supporting reproducible data science, like MLFlow.

Speaker Commentary

Diego Galar
Professor of Condition Monitoring, Luleå Univ. of Technology

"

Data science has become of vital importance for innovation and economic growth in the Digital and AI Economy. The major developements have been focused on human centered and explainabble data driving innovation towards new or human centred products, processes, methods or services. In summary data science as a service for humans bridging the gap of digitization and servitization.

"

Definitely I would love to see cognitive analytics deployed in many fields, from autonomous vehicles to manufacturing and many other scenarios. We have to be aware that while AI has augmented human thinking to solve complex problems. It focuses on accurately reflecting reality and providing optimum results. Cognitive analytics focuses on mimicking human behavior and reasoning to create new ways to solve problems that can potentially be better than humans.

Anjali Shah
Senior Data Scientist, IBM

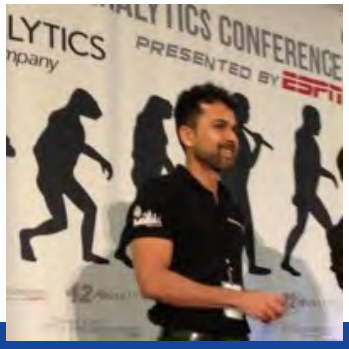
"

Generative models, natural language generation, deep reinforcement learning and bias problems with AI have been some of the most important data science topics in 2019. As AI dominance grows, we must come together and develop countermeasures to tackle bias, deepfakes and other potential AI threats without curtailing innovation.

"

I am excited to see more applications of reinforcement learning (RL) in 2020. RL is the most human-like learning we currently have and it would be fascinating to leverage NLP advances like transformers to RL networks to improve it's longer-term memory.

Speaker Commentary



George Williams
Director of Data Science, GSI Technology

"
Biometrics algorithms and data collection was a booming business in 2019... but not without controversy.

"
2020 will be the year of biometric hacks, fakes, bans, and GANs!

Pramod Singh
Chief Analytics Officer, VP Data Sci. & Analytics, Yodlee Envestnet

"
I believe there was and will be growth in areas of Automated Machine Learning for smarter and faster analytics. Also, Data-as-a-Service (DaaS) was be extremely popular as companies could seamlessly integrate these new technologies in their own data centers or on cloud platforms such as Amazon AWS, Microsoft Azure, and Google Cloud Platform. Natural Language Processing and Conversational AI was another area of interest. Analytics will no longer be about purity of data alone, but about making sure that the models and algorithms are useful and decisions based on them are understood, transparent and unbiased.

"
Artificial Intelligence will no longer be a black box as data scientists understand the impact, application and decision the algorithm is making. Models inherently lack transparency and explanation on what is made or why something can go wrong. XAI will be the exciting new trend in 2020. Its model agnostic nature allows it to be applied to answer some critical questions in data science in areas of fraud detection, preventive medical science and national security.

Speaker Commentary

Ben Vigoda
CEO, Gamalon

"Emerging focus on natural language as a key unsolved type of data, and the recognition that deep learning is fantastic for signal processing, but has limitations for more symbolic problems like natural language where explainability and re-usability are critical.

Rajiv Shah
Data Scientist,
DataRobot

"I thought of 2019 as the year where data science embraced and untwined interpretability, explainability, and fairness.

"I am excited in 2020 for data scientists to recognize that building a model is only the first step in a journey.

Mark Schindler
Managing Director,
GroupVisual.io

"I'm a designer, not a data scientist, so I probably bring a different perspective than most here. I am excited to see new developments that can help non-technical people get better context around data and analytics. A great example of this is some recent work in machine learning image classification models that are easily interpretable.

"I'm hoping to see more work like this in other areas of Deep Learning and AI, so people can more easily understand why a model came to the conclusions it did, and better judge whether and how they should trust it to help them.

Speaker Commentary

Robert Crowe
Developer Advocate,
TensorFlow, Google

"

In 2019 I've been really impressed with the advancements in weak supervision, including the maturing of Snorkel and the early work on Snuba. Labeling is always such a difficult, expensive, and time-consuming effort for any but the simplest supervised learning problems, and advancements in weak supervision really help, especially in domains where the ground truth changes quickly. The related growth of slice-based learning is very exciting also, especially for dealing with issues of bias and fairness.

"

In 2020 I'm looking forward to watching the development of some of the work to apply brain research directly in machine learning, especially in language processing. The recent paper from Carnegie Mellon on this is fascinating!

Vinod Bakthavachalam
Data Scientist, Coursera

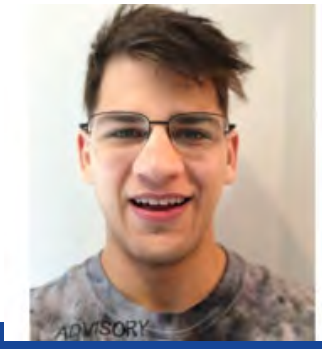
"

In 2019 we saw several developments in data science, but the biggest were probably the increases in computing power and software accessibility. The rise of GPUs and other infrastructure capabilities improved the ability to use deep learning, making it an increasingly popular go to ML algorithm, especially with its performance in tasks related to NLP and image recognition. Improvements in Python, R, and other software like TensorFlow are also making it increasingly easy to use Machine Learning, expanding the number of people who can leverage it in their work.

"

Most applications of machine learning today in companies still require data scientists or other technical employees to figure out how to leverage it for business value. In 2020 I am really excited to see advances in figuring out how to help technical and non-technical folks better work together to leverage machine learning. These work team improvements could dramatically increase productivity and increase the spread of machine learning within companies.

Speaker Commentary



Nick Acosta
Developer Advocate, IBM

"I have greatly enjoyed the increasing point of emphasis our community has made on identifying ways in which data science can be used for social good in 2019, and I am excited to see the outcomes that work produces in 2020 and the years to come.

Joris Cadow
Data Scientist, IBM Research

"Interpretability started to get some track. Advances in pipelines, deployment, logging around training, and inference got more options and became more convenient.

"Transformers continue to change the world.

Casey Fitzpatrick
Machine Learning Engineer, DrivenData

"In my mind, 2019 was the year of the transformer. The research has been there, but thanks to awesome open source projects like the Hugging Face Transformers library, using, comparing, and customizing the highest-performing language models has never been easier.

"There are a few groups tacking multimodal learning, e.g., leveraging audio, video, and text features simultaneously. For example, by using tensor fusion networks which basically implement the deep learning version of interaction terms. I'm excited to see how this research develops and where it proves useful in applications.

Speaker Commentary

Kerstin Frailey
Sr Data Scientist, Head Corporate Training Exec. Programs, Metis

"

In 2019 we finally saw legislation begin to tackle issues in data science and the collection of personal data. From GDPR to CCPA to the DETOUR ACT, legislation certainly isn't catching up with data science - but for the first time we saw it enter the race.

"

Ever since Obama's campaign famously brought big data to politics I've had a bit of a morbid fascination with how data science shapes elections. The 2020 US Presidential election is sure to be no exception.

Kirk Borne
Principal Data Scientist, Data Science Fellow, & Exec Advisor, Booz Allen Hamilton

"

In 2019, we saw strong growth in automated machine learning (AutoML) platforms for citizen data scientists from many vendors, plus an emergence of data tagging, labeling, and annotation services, which are critical for feeding massive amounts of clean data to AI algorithms and processes.

"

I am excited for 2020 to show renewed attention to data strategy, including more deliberate business-focused discussion and planning: what data are we collecting or should be collecting? why? who is using it? where is it being applied? how are we measuring the corresponding business outcomes?

Speaker Commentary

Nico Van de Bovenkamp
S.r Data Scientist, Instructor, Nielsen, General Assembly

"

In 2019, we saw an explosion of impressive ML model management and deployment frameworks. Every company has their own internal solution, but I'm really excited to see companies like Databricks and Netflix open sourcing products like MLFlow and Metaflow.

"

I'm eager to see progress in model interpretability and explainability in 2020. Every year, models become more complex and embedded in our day-to-day activities. Understanding a model's biases and edges will become increasingly critical to the sustainability of ML-based products.

Joy Payton
Supervisor, Data Education, Children's Hospital of Philadelphia

"

In 2019, increased focus was put on the delicate balance between data privacy and data utility, especially in the public sector. The United States Census Bureau has made differential privacy a key focus of its work, and the statistical methods they are developing and evaluating are under close scrutiny. It will be exciting to see how differential privacy starts to be implemented, not just in the 2020 census, but in big data efforts generally, in both the public and private sector.

"

Obviously I'm very interested in the 2020 census, but another area I'm excited about is the rise of citizen data science in the cloud. As data collection and interpretation becomes increasingly politicized, more and more individuals with basic training in free-tier cloud computing and open source data analytics tools will begin to draw and disseminate their own conclusions, just in time for an election year in the U.S.

Speaker Commentary

Jeff Clune
Senior Research Mgr,
Harris Associate
Professor, Uber AI Labs

"

One thing I predict is that we're going to have to as a community come up with a new set of grand challenges. We've been solving grand challenges at such a quick rate that all of the ones that were universally agreed upon no longer are challenges because they've been solved, which is quite interesting. A few years back if you asked people what are the open challenges in AI you would have had many listed by most people, namely Go, Starcraft, Montezuma's Revenge, and multiplayer poker, but all of those have fallen in the last year or recent years. It will be interesting to see if we as a community find consensus on a new set of specific challenges or instead simply start pursuing the many hard problems that remain in AI, but without clear, agreed upon, grand challenges like those in the above list.

Finally, I also think you'll see far more work with self-supervised and unsupervised learning, versus the traditional model of just getting a very large labeled data set and training on that. That has been predicted for a while but we're starting to be at the inflection point where as opposed to being a rare thing that looks promising it will become a default way to get state-of-the-art results. Yann LeCun recently tweeted out two new papers that do just that."

Gil Benghiat
Co-founder,
DataKitchen

"

Inspired by new methods and tools in data analytics, we pulled many ideas from industry thought leaders together and penned the DataOps Manifesto. We wrote 18 principles that describe a new approach to data analytics embodied by Agile, DevOps and lean manufacturing methods. To date, over 7,000 people have signed it and we get more every day.

"

The transformation of data analytics from manual to automated processes will be a big theme in 2020. Organizations that have dozens or hundreds of data sources cope with data errors on a daily or weekly basis. Automated tests and process controls will trap these anomalies before they corrupt charts and graphs. Better quality control will reduce unplanned work and improve credibility, leading to greater trust in analytics and improved data-driven decision-making.

Speaker Commentary

Violeta Misheva
Data Scientist, ABN AMRO Bank N.V.

"Machine learning explainability has been a prominent trend in 2019. Not only academics but also research and governmental institutions have started to push this work further and to think how to make it tangible.

"I am definitely excited to see more developments in XAI in 2020, as well as the way we deal with biases and ensure that we build fair machine learning models.

Matteo Manica
Research Staff, Cognitive Health Care & Life Sci, IBM Research Zürich

"It has been amazing to see how much attention has been given to the need for interpretable models and interpretability for machine/deep learning. A lot of research groups and industrial players are building amazing tools and methodologies, and I'm confident we are going in the right direction as a united community.

"What excites me the most about the near future, is the growing interest to extend currently established interpretable methodologies to new fields and domains. I think that the application of explainability techniques to the physical sciences for discovery purposes will be the next big thing.

Veysel Kocaman
Senior Data Scientist, John Snow Labs

"We have witnessed many exciting breakthroughs in NLP in 2019 and thanks to transformers, we have already surpassed human baselines in several NLU tasks. I believe that NLP will gain even more attention in the healthcare industry.

"The autonomous driving race is gearing up and even though we are not there yet, I believe that the hardware side of this technology will be more affordable in 2020 and it will introduce new competitors to the race.

Speaker Commentary

Faith Xu
Sr Program Mgr,
Machine Learning,
Microsoft

"

It's tremendous to see more and more applied uses of ML come to reality across a diverse range of industries – manufacturing, healthcare, automotive, consumer products, and so much more! Helping engineering teams build maintainable and scalable operational ML solutions will be key to help drive that innovation and realize all the magic coming out of AI research.

"

I'm really excited to see the community enthusiasm and growing adoption of ONNX. Having an open and interoperable model format standard will be critical to help us build streamlined solutions for operationalizing ML models in production. I'm also excited to see expanding performance improvements in ONNX Runtime for inferencing ONNX models and the support of a burgeoning hardware ecosystem, and hope to see it continue to evolve as an asset for the industry.

Alfredo Kalaitzis
Applied Research
Scientist, Element AI

"

We've seen more tools on the explainability of ML models. We now realize that user-centric design in AI systems is more important than powerful but opaque decision-making. All parts of a systems must be accountable, whether it be a human or machine in the loop.

"

Fairness in ML models is primarily a human problem, not an AI problem. AI is biased only because our data that we train it with is biased. I'm excited to see an increase in inter-discipline research that includes experts in policy, governance and ethics. Also, with the proliferation of satellite instruments, Earth observation imagery will become a new theme of applied research in machine learning.

Speaker Commentary

Alex Holub
CEO, Vidora

"

One of the exciting machine learning trends from 2019 is the increased prevalence of tools to enable data scientists to be more productive while simultaneously spreading machine learning to a broader set of constituents, all generally falling under the umbrage of 'AutoML'.

"

In 2020, we will see an increased focus on helping solve some of the main organizational roadblocks in deploying machine learning technology, in particular helping automate and make 'data wrangling' tasks easier.

Benn Stancil
Chief Analyst, Mode Analytics

"

The biggest development was a dream deferred: Automated data science and business decision making still isn't here, despite previous predictions that it's just around the corner. Data science teams are – and will continue to be – defined by the skills of the people on the team, not the technologies they use.

"

More investment in deeper strategic analysis. Now that executives are inundated with data, they're realizing that timely insights are more valuable than another dashboard. Companies that see this will get results and, importantly to me as a data scientist, have a much happier data team.

Scott Haines
Principal Software Engineer, Twilio

"

The idea of the centralized platform for data science, analysis, and engineering really took off this year. Companies across the globe continue to struggle getting the most out of their data science teams and reducing friction by making it easier to collaborate across teams has been a big win.

"

I am excited to see the field of active learning taking off. I think this year will bring more advances in the field, and help more companies and people take advantage of active learning for real-time systems.

Speaker Commentary

Gautam Tambay
CEO, Co-founder, Springboard

"

The continued democratization of AI due to the convergence of maturing deep learning frameworks (e.g. PyTorch), cloud-based managed ML services (e.g. Azure ML Studio), and automated tuning (e.g. AutoML). AI is now really accessible to all.

"

I'm excited about the convergence of data and design. Design tools enhanced by AI, and AI apps are becoming more human-centered. Conversational UX is a great example of a field that combines design with the increasing maturity of NLP technology.

Nathaniel Tucker
Lead Instructor, Data & Analytics, General Assembly

"

In 2020, I'm most excited to see more work on artist tools. There are many unsung heroes in AI that currently help artists by speeding up time consuming processes: smoothing and ray tracing. There are also big strides in generating art as well. Generating character movement for games or celebrity faces is only the start. I'm hoping we continue to see tools that democratize art and allow great artists to do even more!

Michael Sollami
Lead Data Scientist, Salesforce

"

Over the past year, we've seen some major methodological advances in deep learning, specifically improved attentional mechanisms and large-scale pretraining. But currently, our best networks fail to understand abstract representations and use very simplistic compositional reasoning.

"

I would like to see a few breakthroughs in causal models for computer vision and reinforcement learning. I'm also excited to see what happens with differentiable programming and learning outside of python, for example with Julia Zygote or Swift for TensorFlow.

Speaker Commentary

Anirudh Koul
Author, "Practical Deep Learning for Cloud, Mobile and Edge"

"

Privacy is already on everyone's mind and will help give federated learning (which involves training on the user's device and combining the learnings collectively from all users) a big splash in 2020. So far, only limited apps from Apple and Google have publicly been able to use this process primarily due to a lack of tooling, which is thankfully changing now. More management roles in AI will start appearing. Just like 'Head of Data Science' started appearing 4-5 years post 'Data Scientist' being introduced earlier this decade, more Directors and VPs of AI as well as Chief AI Officer roles will finally start opening.

Christopher Bergh
CEO, Head Chef, DataKitchen

"

We witnessed a sharp rise in DataOps interest. Data industry thought leaders wrote thousands of articles about DataOps in 2019, and inquiries at analytics firms about DataOps are up over 200%. We talk to people daily who face the challenge of slow, error-prone data analytics. People are beginning to understand that DataOps offers a proven method for reducing analytics development cycle time and virtually eliminating data errors. It's exciting to see these ideas take-off.

"

In 2019, enterprises grew disillusioned with significant investments in machine learning and AI that yielded little ROI. Now the discussion has matured to what processes and tools need to be in place, supporting the AI/ML models in order to make them successful. Ultimately, companies with the most robust and efficient DataOps workflows will gain the most competitive advantage from cutting-edge techniques like machine learning.

Speaker Commentary

Roman Yampolskiy
Distinguished Teaching Prof., Founding Director, Cyber Security Lab, Author, University of Louisville

"

Practical efforts to improve transparency of AI-black boxes were met with theoretical results, indicating that explainability and comprehensibility of such systems have limits- and we are already hitting hard against such limits.

"

I'm excited to see in 2020 the continuation of research on theoretical limits of intelligent systems, including unexplainability, incomprehensibility, and unpredictability so we can better understand how to make AI safer and more secure.

Alan Rutter
Founder, Fire Plus Algebra

"

The increasing focus on communication in data science. To drive meaningful change within organisations or the wider world, it's necessary to explain both the meaning of data and the role of the data scientists to lay audiences – colleagues from other disciplines, senior executives, customers, and members of the public.

"

The discussions over ethics and explainability in AI and machine learning. This is going to be crucial in maintaining trust and transparency, and data scientists will be at the heart of the debate.

Jon Krohn
Chief Data Scientist, Author, Deep Learning Illustrated, Untapt

"

The broadening influence of transformer architectures within deep learning models applied to range of natural language processing tasks. Most notably, the GPT-2 model released by researchers at OpenAI is capable of generating remarkably coherent lengthy sequences of text.

"

After many big names in machine learning- including Andrew Ng, Demis Hassabis, and Yoshua Bengio- contributed to a paper on tackling climate change with ML, I'm hopeful that we'll begin to see models applied more frequently and more meaningfully to tackling this issue as opposed to simply monitoring it.

Speaker Commentary

Ido Shlomo
Senior Data Science Manager, BlueVine

"

Significant headway has been made in developing platforms for managing and deploying ML models in production. Both sophistication and ease of use are on the rise, and correspondingly these platforms are being more widely adopted. Today's data scientists are actually able to independently deploy models at scale.

"

I'm excited to see how streamlined these new ML deployment platforms can become, and what kind of effect they will have on both on the data science profession and on the composition of data teams. So much is wasted/lost in transition between ML development and deployment, and these platforms are set to dramatically change this.

Matthew Kenney
Researcher, Duke University

"

The progress in NLP, from BERT-variations to GPT-2 and MultiFiT, blew me away. Accurate text classification across many languages is becoming common. Performance on other tasks, such as QA and NER, is improving. Text generation is quite impressive and has many implications for the world.

"

The work done by the NVIDIA RAPIDS team to improve efficiency (by 10x-100x) for many different data types (tabular, geospatial, graph, and cyber) will enable data scientists to develop solutions that were unthinkable just a year ago.

Byron Galbraith
Chief Data Scientist, Talla

"

The success and availability of large, pretrained language models based on the transformer architecture, e.g. BERT and its variants. Hugging Face, in particular with their Transformers PyTorch library, has made these advances widely available and easily accessible to a larger practitioner audience.

"

Two things: First, increased focus on real-world applications of reinforcement learning, especially around contextual bandits and imitation learning. Second, increasingly sophisticated applications and explorations of AI by artists and musicians.

Top 10 Blogs

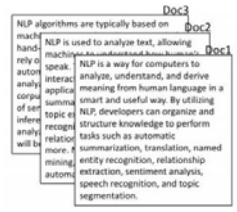
We published over 400 articles to OpenDataScience.com in 2019, discussing everything from machine and deep learning to use cases in countless industries. Here are the most-viewed articles from 2019.

25 Excellent Machine Learning Open Datasets | Elizabeth Wallace

Here are our top 25 picks for open source machine learning datasets. Each one offers clean data with neat columns and rows so that your training sets run more smoothly. Let's take a look.

Reviewing Amazon's Machine Learning University: Is it Worth All of the Hype? | Daniel Gutierrez

The same machine learning courses used to train Amazon employees are now available to all data scientists and data engineers through AWS for free. So, how is it?

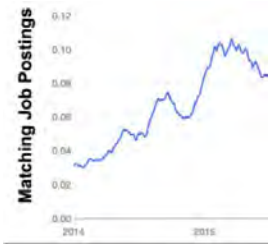


Deep Learning Finds Fake News with 97% Accuracy Lutz Hamel

Fake news is everywhere, and many data scientists are tackling the problem in various ways. Here's how one team created a deep learning model with amazing success.

Making Sense of Confusing Data Science Job Postings Elizabeth Wallace

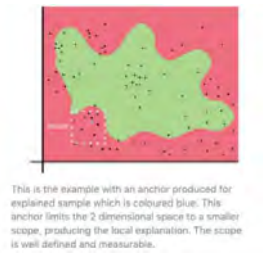
The data scientist career path is challenging enough without having to decipher ambiguous job postings. Here's how you can make sense of them.



Here's Why You Aren't Getting a Job in Data Science Daniel Gutierrez

Here, we discuss a few reasons why you may be having difficulty finding a job in the data science field - such as a lack of demonstrated experience - and how to overcome these obstacles.

Top 10 Blogs (*continued*)



This is the example with an anchor produced for explained sample which is coloured blue. This anchor limits the 2 dimensional space to a smaller scope, producing the local explanation. The scope is well defined and measurable.

Cracking the Box: Interpreting Black Box Machine Learning

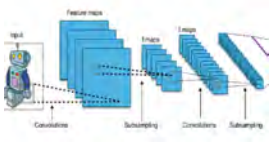
In this article, we let you in on major methods of tackling the interpretability of ML models using Python.

Strategies for Addressing Class Imbalance | Nate Jermain

Here, we demonstrate common techniques for addressing class imbalance including over-sampling, undersampling, and synthetic minority over-sampling technique (SMOTE) in Python.

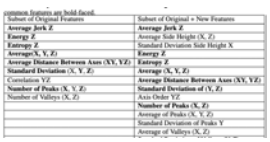
An Introduction to Natural Language Processing (NLP) Diego Lopez Yse

Today, NLP is booming thanks to huge improvements in the access to data and increases in computational power, leading to impressive results in countless fields. What exactly is NLP?



The Data Science Dictionary – Key Terms You Need to Know | Daniel Gutierrez

Familiarize yourself with twenty key terms you need to know in data science for 2019, all relating to machine learning, deep learning, AI, open data, and more.



common features are bold-faced.

Subset of Original Features	Subset of Original + New Features
Average Jerk Z	**Average Jerk Z**
Energy Z	Average Side Height (X, Z)
Entropy Z	Standard Deviation Side Height X
Averages(X, Y, Z)	**Energy Z**
Average Distance Between Axes (XY, YZ)	**Entropy Z**
Standard Deviation (X, Y, Z)	**Average (X, Y, Z)**
Correlation YZ	**Average Distance Between Axes (XY, YZ)**
Number of Peaks (X, Y, Z)	**Standard Deviation of (Y, Z)**
Number of Valleys (X, Z)	Axis Order YZ
	Number of Peaks (X, Z)
	Average of Peaks (X, Y, Z)
	Standard Deviation of Peaks Y
	Average of Valleys (X, Z)

10 Compelling Machine Learning Dissertations from PhD Students | Daniel Gutierrez

PhD candidates often work on some fascinating data science projects. Here are 10 standout machine learning dissertations that may interest you.

Top 10 Videos

In 2019, across all ODSC events, we held over 550 workshops, trainings, talks, and panels. Out of all of them, these are the ten highest-rated as voted by conference attendees.

State-Of-The-Art Text Classification With Ulmfit

Matthew Teschke, Director, Applied Machine Learning, Novetta

ODSC East 2019

Matt provides an overview of ULMFiT before walking through a use case in which he used Amazon SageMaker to train ULMFiT on hand-labelled quotes sourced from thousands of news articles. After demonstrating the success of this method, Matt discusses a novel approach that further improves upon ULMFiT by up to 10% by incorporating article-level metadata such as publication name, author name, and speaker. The session ends with a short discussion of how this model has been implemented to increase the efficiency of the analysts who are manually tagging quotes.

Programming Machine Learning with Weak Supervision

Alex Ratner, PhD Student, Stanford University

ODSC East 2019

In this talk, Alex describes Snorkel, our open-source system for training data labeling (snorkel.stanford.edu), that can reduce training data creation time from months to days; other recent work around data augmentation and multi-task supervision; and applications of this work in domains ranging from medical imaging to unstructured data extraction.

Data Art: Seeing the Future

Jane Adams, Data Visualization Artist, The Vermont Complex Systems Center

ODSC East 2019

Horizons in data science—big data, networks, machine learning—all necessitate innovation in the field of visualization. As the field of data science evolves, the field of "data art" is emerging and responding to this new landscape. There is growing potential to leverage artistic insight to capture the humanity of data; to engage with the maker movement through data physicalization; and to dream beyond two dimensions by involving artists of all mediums. This added perspective can aid researchers in better understanding their data, and can be a powerful way to engage customers and the public in discussions about data science.

Top 10 Videos *(continued)*

Adversarial Attacks On Deep Neural Networks

Sihem Romdhani, Data Scientist, Veeva Systems

ODSC East 2019

Through use cases and illustrative examples, this video discusses how adversarial attacks pose a real world security threat, how these attacks can be performed, what the different types of attacks are, different defense techniques, and how you can make a system more robust against these attacks.

Tools for High Performance Python

Ian Ozsvald, Principal Data Scientist, Mor Consulting

ODSC Europe 2019

Your tools and workflow govern how quickly you can deliver results on new challenges. Often we're constrained by slow algorithms, inefficient data pipelines, and suboptimal use of complex tools like Pandas. This talk looks at recent changes in the Python ecosystem enabling fast identification of slow code, simple compilation of CPU-bound numpy processing with Numba, efficient Pandas operations, and parallelised medium-data operations with Dask. You will walk away with new tools and process to take back to the office.

Sequence Modelling with Deep Learning

Dr. Natasha Latysheva, Machine Learning Research Engineer, Welocalize

ODSC East 2019

This presentation/tutorial will start from the basics and gradually build upon concepts in order to impart an understanding of the inner mechanics of sequence models – why do we need specific architectures for sequences at all, when you could use standard feed-forward networks? How do RNNs actually handle sequential information, and why do LSTM units help longer-term remembering of information? How can Transformers do such a good job at modelling sequences without any recurrence or convolutions? And whatever happened to Markov chains?

Making Data Useful

Cassie Kozyrkov, PhD, Chief Decision Scientist, Google

ODSC Europe 2019

Despite the rise of data engineering and data science functions in today's corporations, leaders report difficulty in extracting value from data. Many organizations aren't aware that they have a blindspot with respect to their lack of data effectiveness and hiring experts doesn't seem to help. Let's talk about how you can change that!

Top 10 Videos *(continued)*

Machine Learning Design Patterns

Valliappa Lakshmanan, PhD, Global Head, Data Analytics & AI Solutions, Google

ODSC West 2019

As the practice of machine learning gets formalized, the community learns best practices of setting up large scale training loops and moving from development to production. In the talk, Valliappa introduces some of these design patterns, explaining the problem and the code for these patterns in Keras and TensorFlow 2.0.

Project GaitNet: Ushering in The ImageNet Moment for Human Gait kinematics

Vinay Prabhu, PhD, Chief Scientist, UnifyID AI Labs

ODSC West 2019

In this talk, we will introduce the ImageNet moment for human gait analysis by presenting 'Project GaitNet', the largest ever planet-sized motion sensor based human bipedal gait dataset ever curated. We also present the associated state-of-the-art results in classifying humans harnessing novel deep neural architectures and the related success stories we have enjoyed in transfer-learning into disparate domains of human kinematics analysis.

Mapping Geographic Data in R

Joy Payton, Supervisor, Data Education, Children's Hospital of Philadelphia

ODSC West 2019

In this hands-on workshop, we use R to take public data from various sources and combine them to find statistically interesting patterns and display them in static and dynamic, web-ready maps. This session covers topics including geojson and shapefiles, how to munge Census Bureau data, geocoding street addresses, transforming latitude and longitude to the containing polygon, and data visualization principles.

Be a part of the ODSC Community

There are many ways that you can engage with the Open Data Science Community today.

ODSC Events

East 2020 | Boston
April 13-17, 2020

India 2020 | Bengaluru
September 9-12, 2020

Europe 2020 | Dublin
September 14-18, 2020

West 2020 | San Francisco
October 26-30, 2020

More Downloadable Guides:

Did you like this guide? We also have downloadable guides for machine learning and deep learning. Download them for free now!

Weekly Newsletter

Don't miss any future articles on data science and machine learning! Sign up for our weekly newsletter and get tutorials, insights, and the latest news sent to you directly.

Meetups

We hold meetups in 37 cities around the world, designed to convene data scientists for education, networking, and even a little fun. See upcoming events here.

Webinars

We offer free webinars several times a month, covering a variety of topics. Follow this page to learn more about upcoming webinars.

Becoming a Part of ODSC Events

Are you a technical or business expert in the world of data science and AI? Consider speaking at one of our events! Each event has its own speaker submission page. East 2020

... more coming soon!
We also offer partnership opportunities! Have your product, service, or research seen by thousands of data scientists at an event. Learn more here.

ODSC