Buzzwords surrounding Data Science

Eiko Yoneki
eiko.yoneki@cl.cam.ac.uk
http://www.cl.cam.ac.uk/~ey204

Systems Research Group
University of Cambridge Dept. Computer Science and Technology
Computer Laboratory

Outline

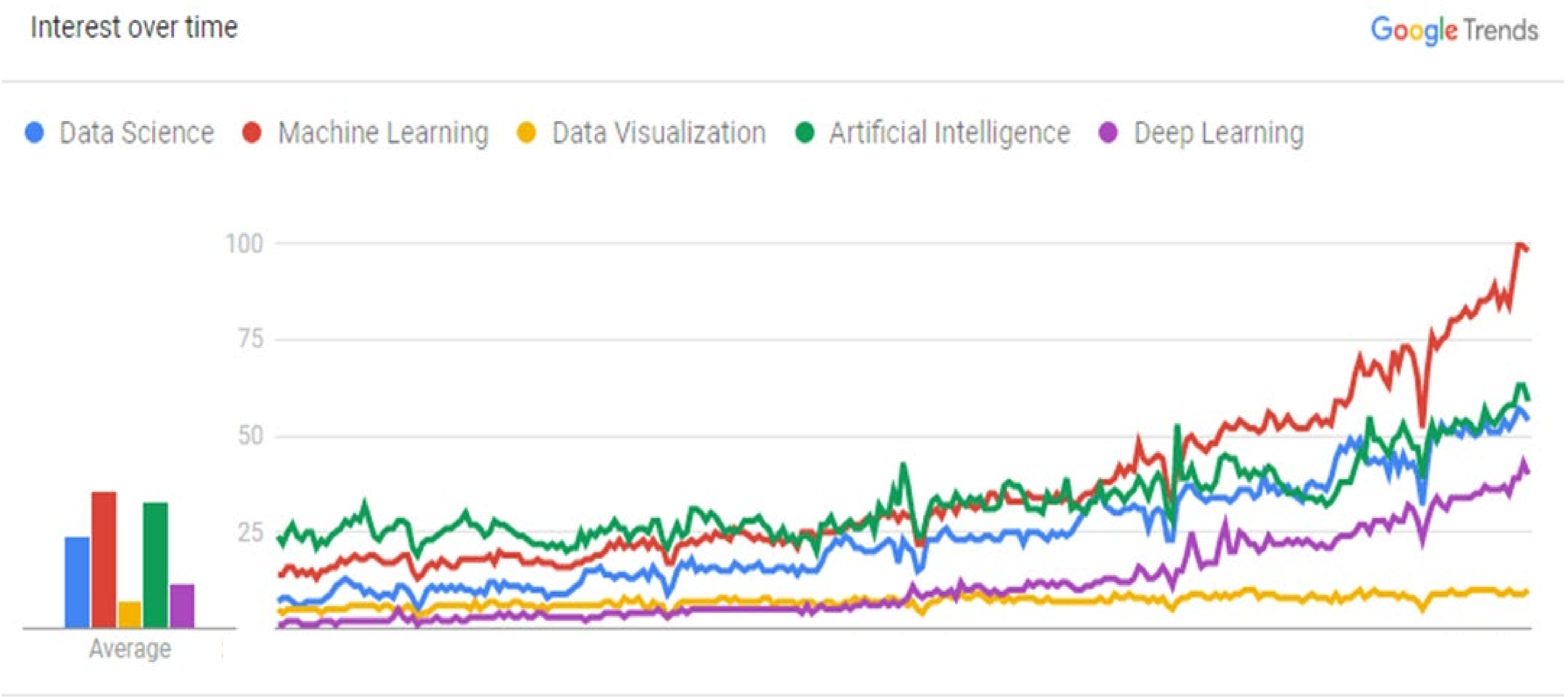
- Data Science Community is still Growing
- Where did Data Science come from?
- Becoming Data Scientist?

- Many Buzzwords?
- My Pick of Interesting Topics/Buzzwords

- Data Science is broad: Pick a right topic and its scope

→ Aim at RESEARCH in Data Science!

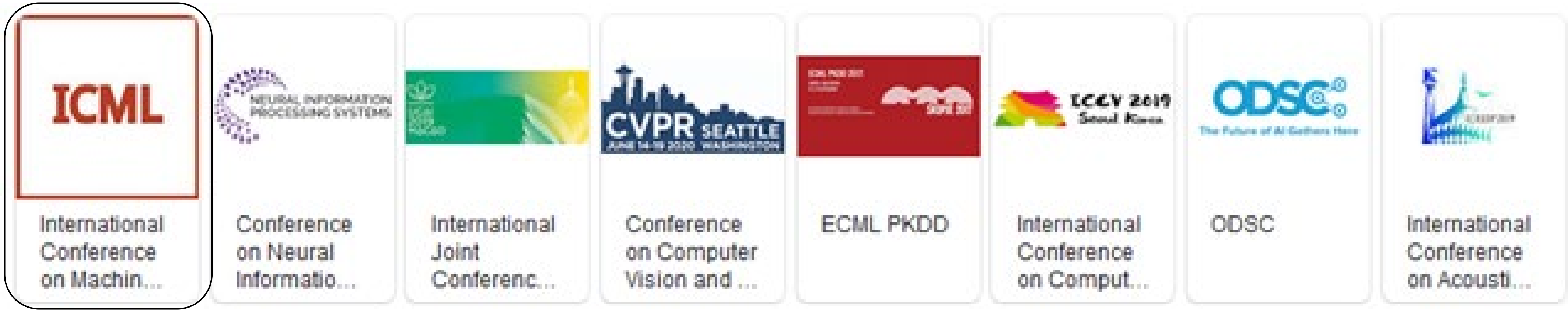
Rise of Data Science



Worldwide. 02/06/2012 - 02/06/2017.

Machine Learning Conferences

Neural Information Processing Systems



International Conference on Machin...

Conference on Neural Informatio...

International Joint Conferenc...

Conference on Computer Vision and ...

ECML PKDD

International Conference on Comput...

ODSC

International Conference on Acousti...

International Conference on Learning Representations

European Conference on Artificial...

ICAART

ACM SIGK... International Conferenc...

Strata Data Conference

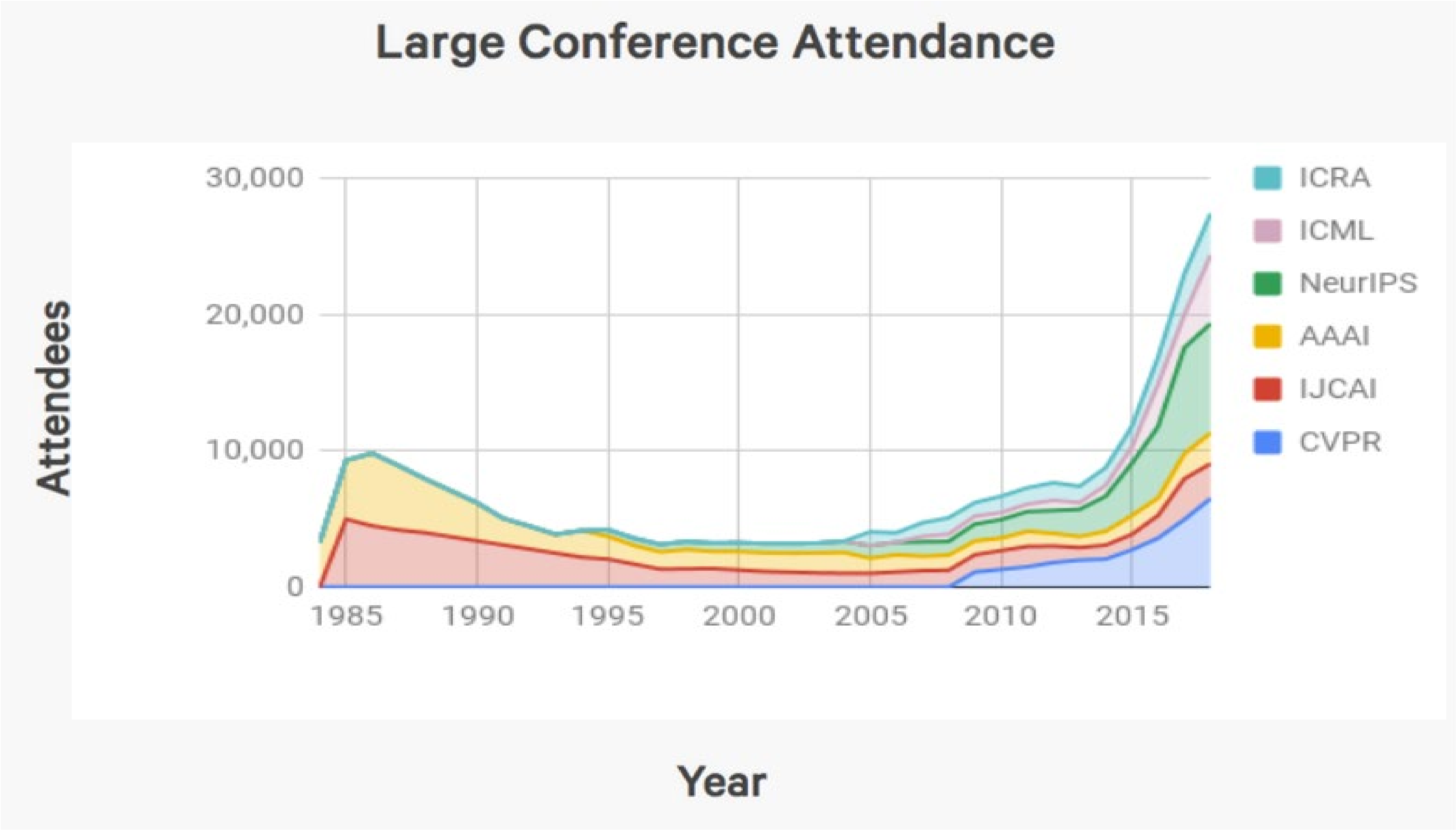
ICPRAM

International Semantic Web Confe...

Mexican International Conferenc...

International Conference on Comput...

Scale of Community Size in ML/AI



MLSys Conference spawn in 2019

- MLSys is a conference targeting research at the intersection of systems and machine learning

  https://mlsys.org
- Aims to elicit new connections amongst these fields, including identifying best practices and design principles for learning systems, as well as developing novel learning methods and theory tailored to practical machine learning workflows

Steering Committee

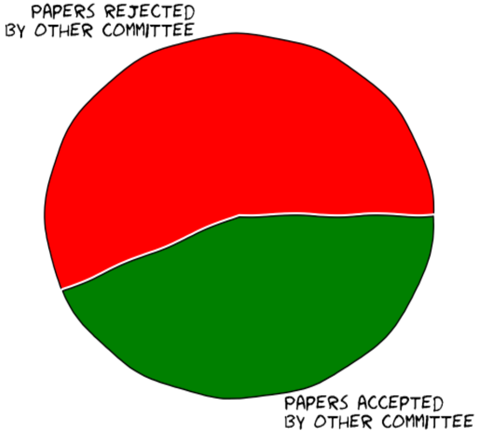
Jennifer Chayes
Bill Dally
Jeff Dean
Michael I. Jordan
Yann LeCun
Fei-Fei Li
Alex Smola
Dawn Song
Eric Xing

NIPS/NEURIPS: 8000 Attendees in 2017

Randomness of Paper acceptance?

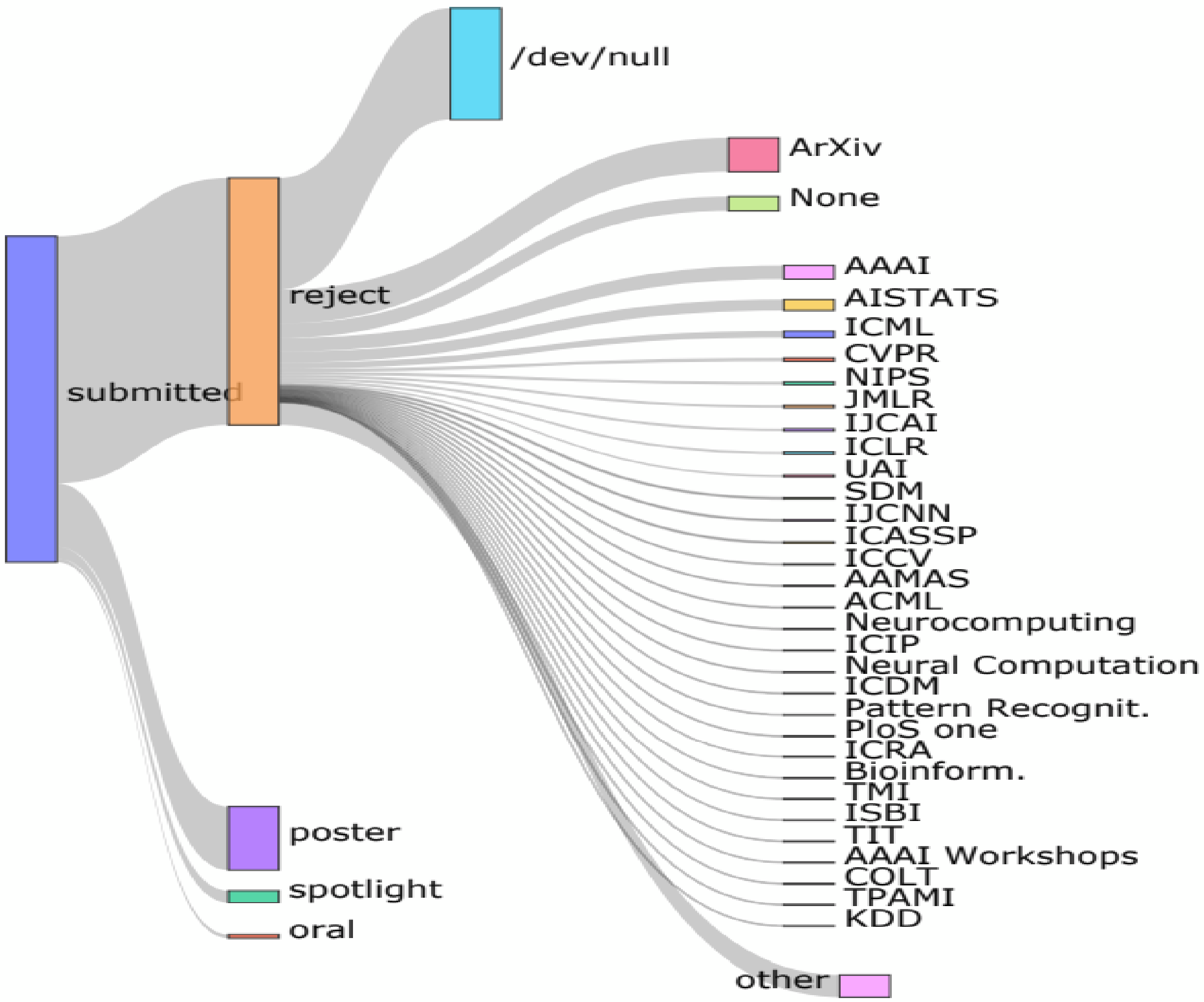
- 2016: 2,406 submissions and 568 acceptance (24% acceptance rate)
- 2017: 3,240 submissions and 679 acceptance (21% acceptance rate)
- 2020: 9,467 submissions and 1,990 acceptance (20% acceptance rate)
- In 2014, Corinna Cortes and Neil Lawrence ran the NIPS experiment where 1/10th of papers submitted to NIPS went through the NIPS review process twice, and then the accept/reject decision was compared. (http://blog.mrtz.org/2014/12/15/the-nips-experiment.html)
- In particular, about 57% of the papers accepted by the first committee were rejected by the second one and vice versa. In other words, most papers at NIPS would be rejected if one reran the conference review process (with a 95% confidence interval of 40-75%).
- 2021: Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment (https://arxiv.org/abs/2109.09774)

RESULTS IN 2ND COMMITTEE OF THE PAPERS ACCEPTED BY THE 1ST COMMITTEE

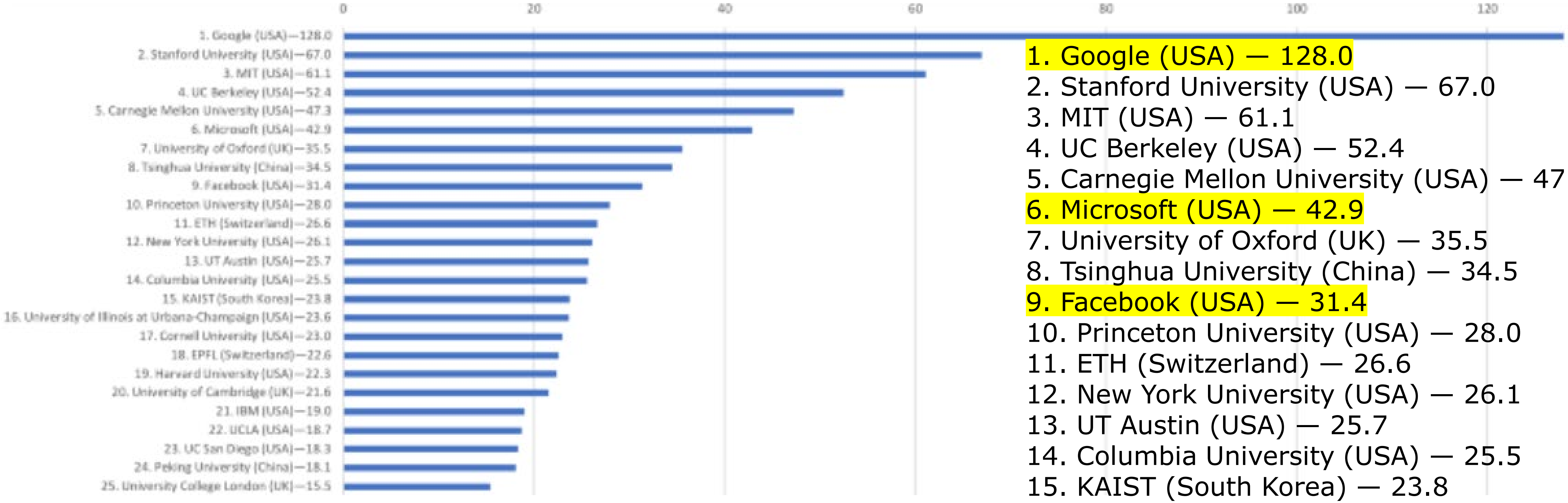


UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Where those Rejected Papers were published

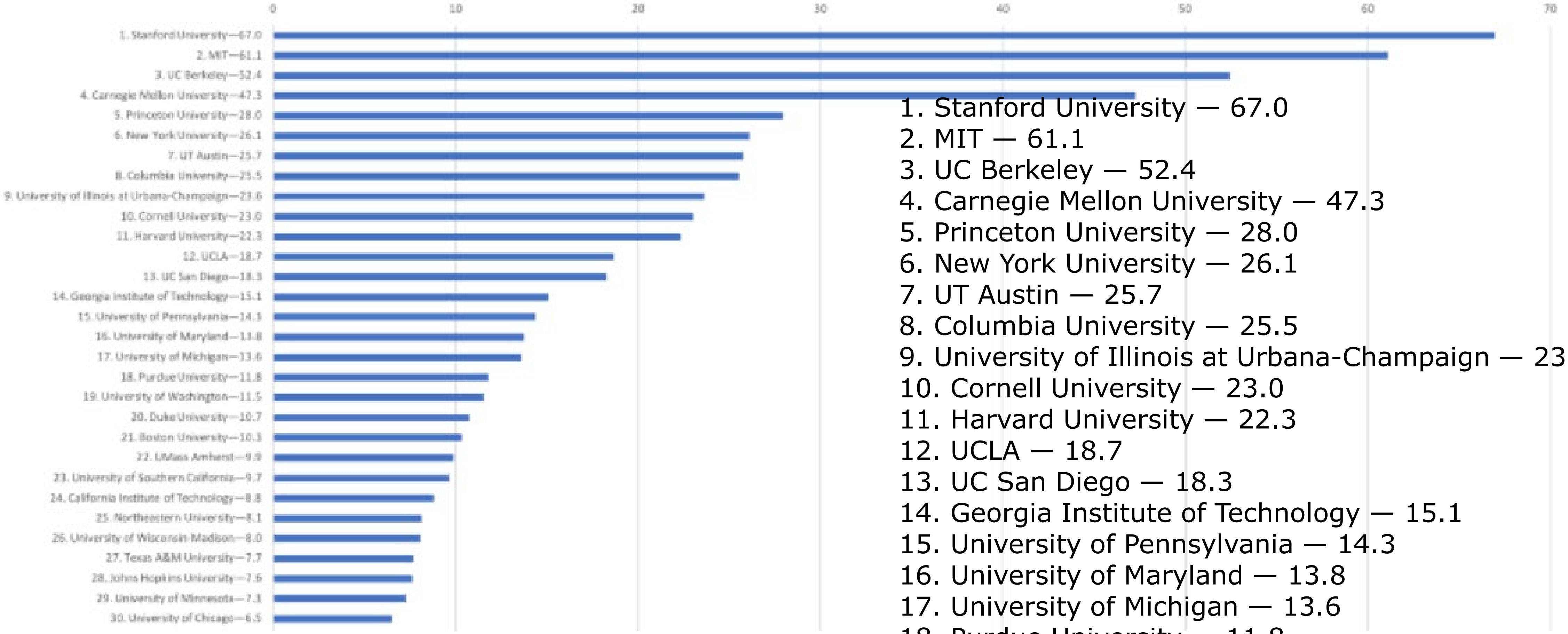


NeurIPS 2020 Publishing



1. Google (USA) — 128.0
2. Stanford University (USA) — 67.0
3. MIT (USA) — 61.1
4. UC Berkeley (USA) — 52.4
5. Carnegie Mellon University (USA) — 47.3
6. Microsoft (USA) — 42.9
7. University of Oxford (UK) — 35.5
8. Tsinghua University (China) — 34.5
9. Facebook (USA) — 31.4
10. Princeton University (USA) — 28.0
11. ETH (Switzerland) — 26.6
12. New York University (USA) — 26.1
13. UT Austin (USA) — 25.7
14. Columbia University (USA) — 25.5
15. KAIST (South Korea) — 23.8
16. Univ. Illinois at Urbana-Champaign (USA) — 23.6
17. Cornell University (USA) — 23.0
18. EPFL (Switzerland) — 22.6
19. Harvard University (USA) — 22.3
20. University of Cambridge (UK) — 21.6
21. IBM (USA) — 19.0
22. UCLA (USA) — 18.7
23. UC San Diego (USA) — 18.3
24. Peking University (China) — 18.1
25. University College London (UK) — 15.5

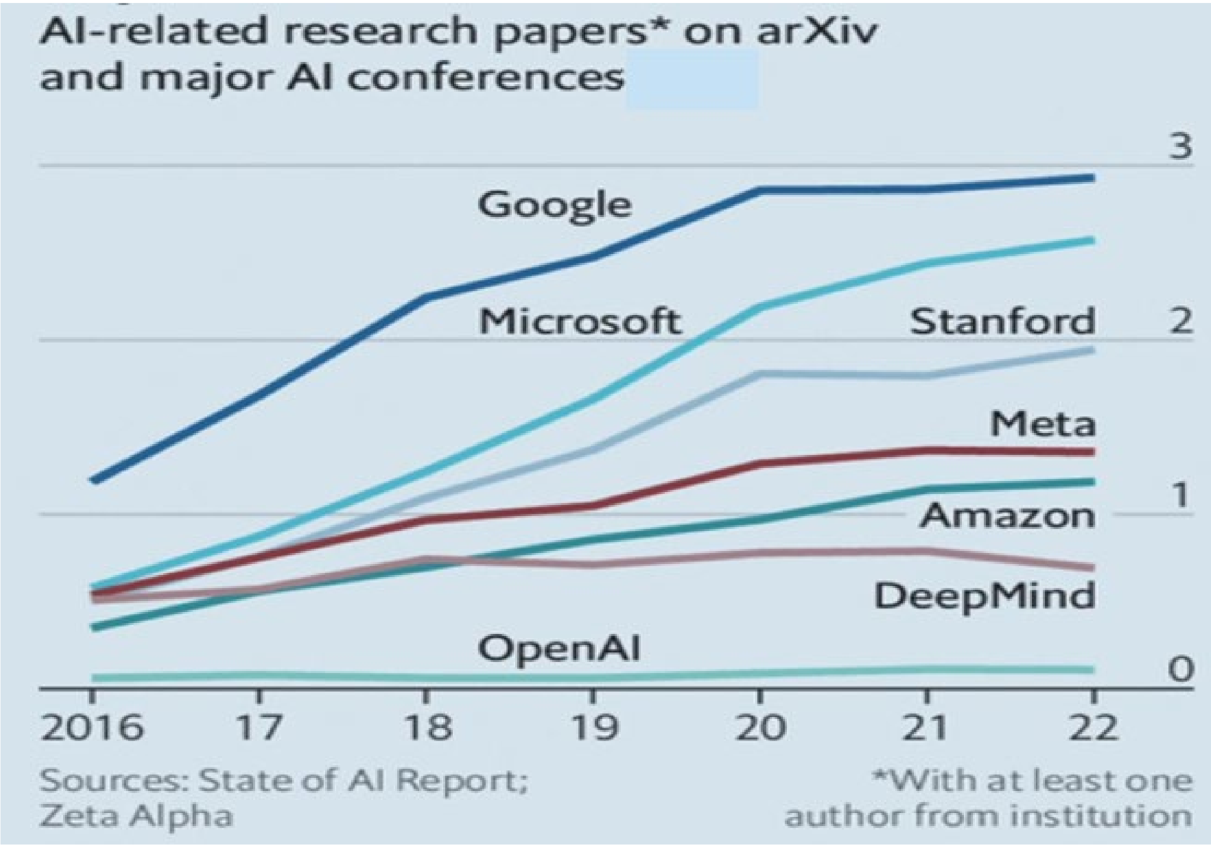
NeurIPS 2020 Publishing (American Universities)



1. Stanford University — 67.0
2. MIT — 61.1
3. UC Berkeley — 52.4
4. Carnegie Mellon University — 47.3
5. Princeton University — 28.0
6. New York University — 26.1
7. UT Austin — 25.7
8. Columbia University — 25.5
9. University of Illinois at Urbana-Champaign — 23.6
10. Cornell University — 23.0
11. Harvard University — 22.3
12. UCLA — 18.7
13. UC San Diego — 18.3
14. Georgia Institute of Technology — 15.1
15. University of Pennsylvania — 14.3
16. University of Maryland — 13.8
17. University of Michigan — 13.6
18. Purdue University — 11.8
19. University of Washington — 11.5
20. Duke University — 10.7
21. Boston University — 10.3
22. UMass Amherst — 9.9

Intellectual contribution to AI

- Some of organizations publish knowledge and open-source code for the entire world to use. Others just consume it.
- AI research output from prolific institutions:
  - Google
  - Microsoft
  - Stanford
  - Meta
  - Amazon
  - DeepMind
  - OpenAI



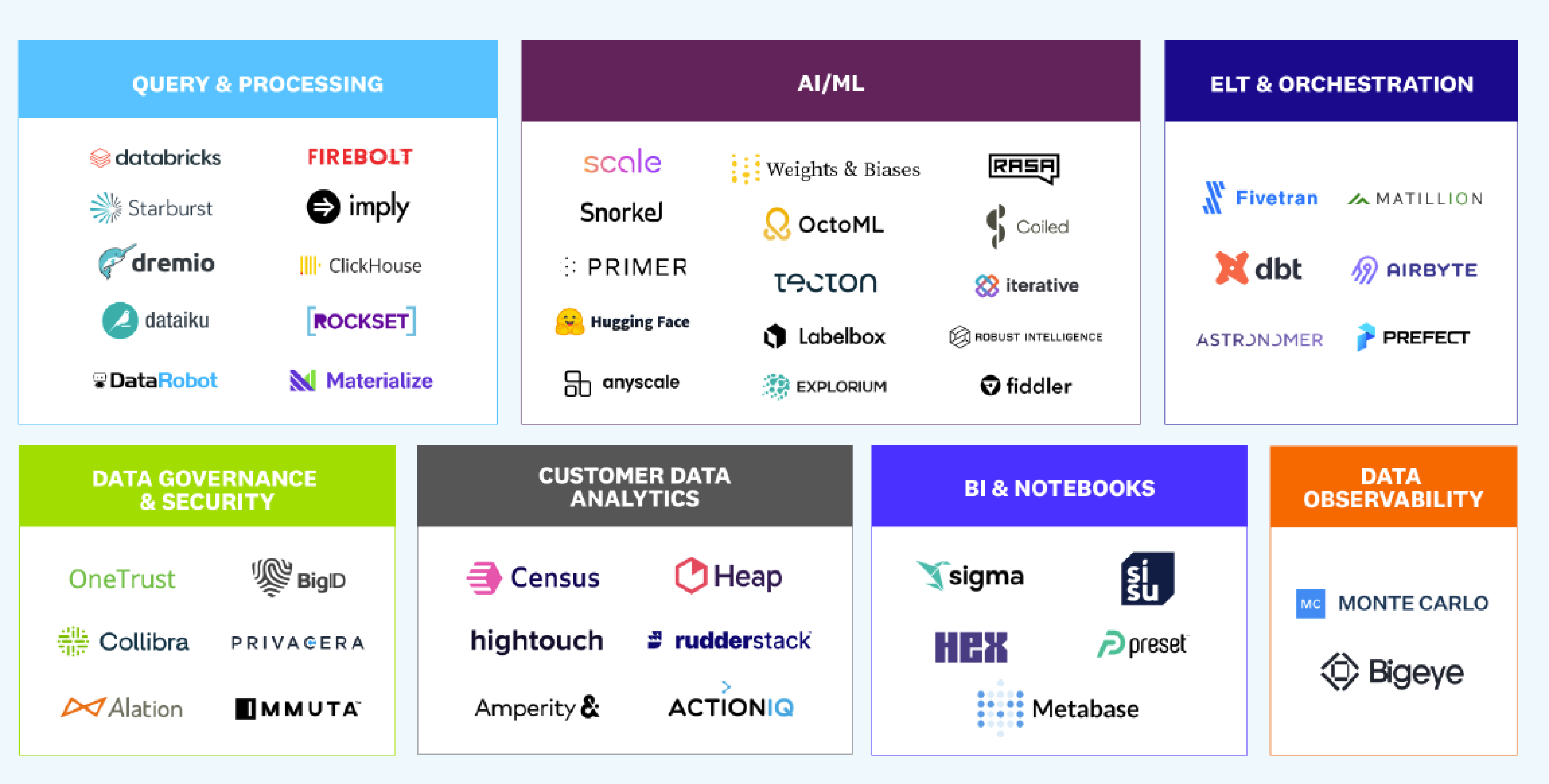
AI-related research papers* on arXiv and major AI conferences

Sources: State of AI Report; Zeta Alpha

*With at least one author from institution

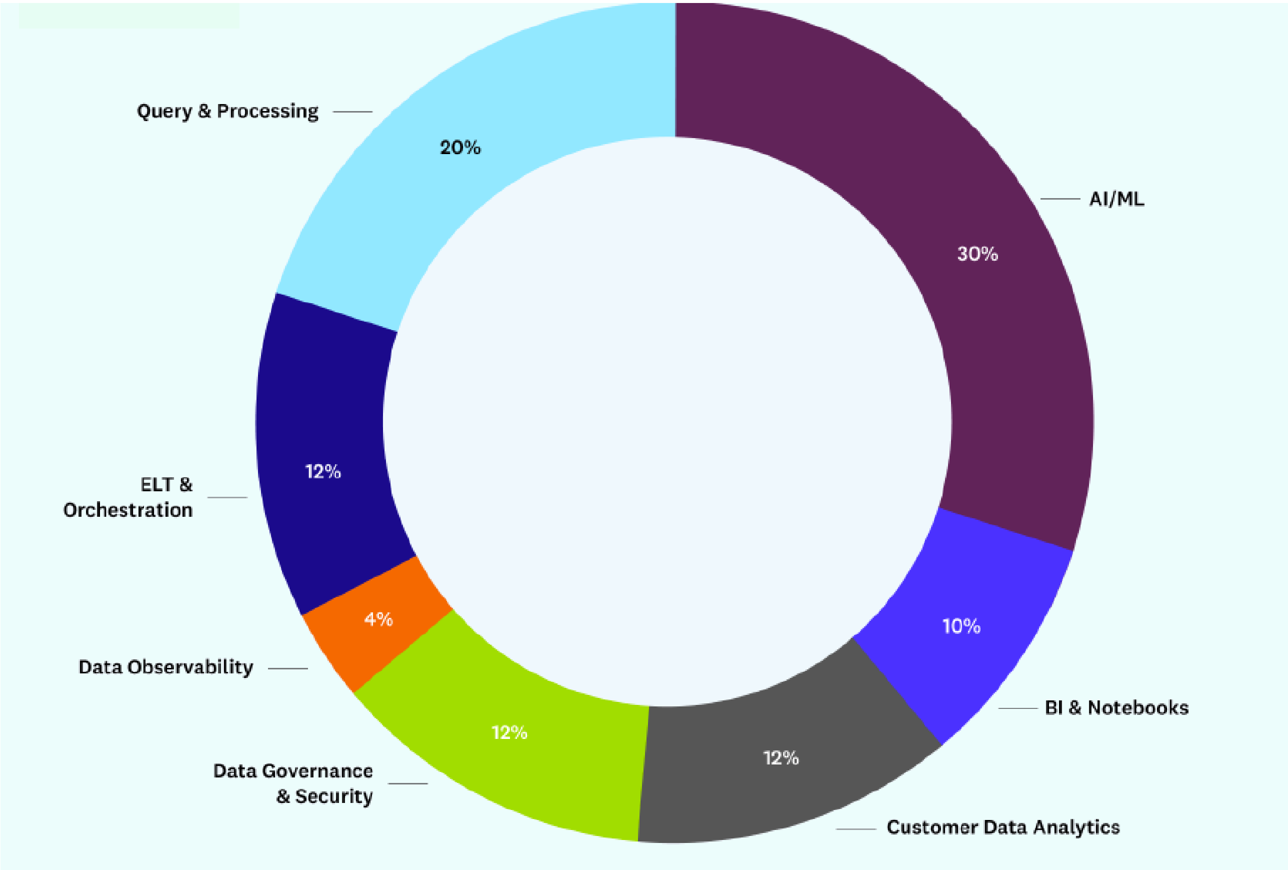
World's Top 50 Startups in AI/ML

- In aggregate, these 50 companies are valued at more than $100B and have raised approximately $14.5B in total capital.



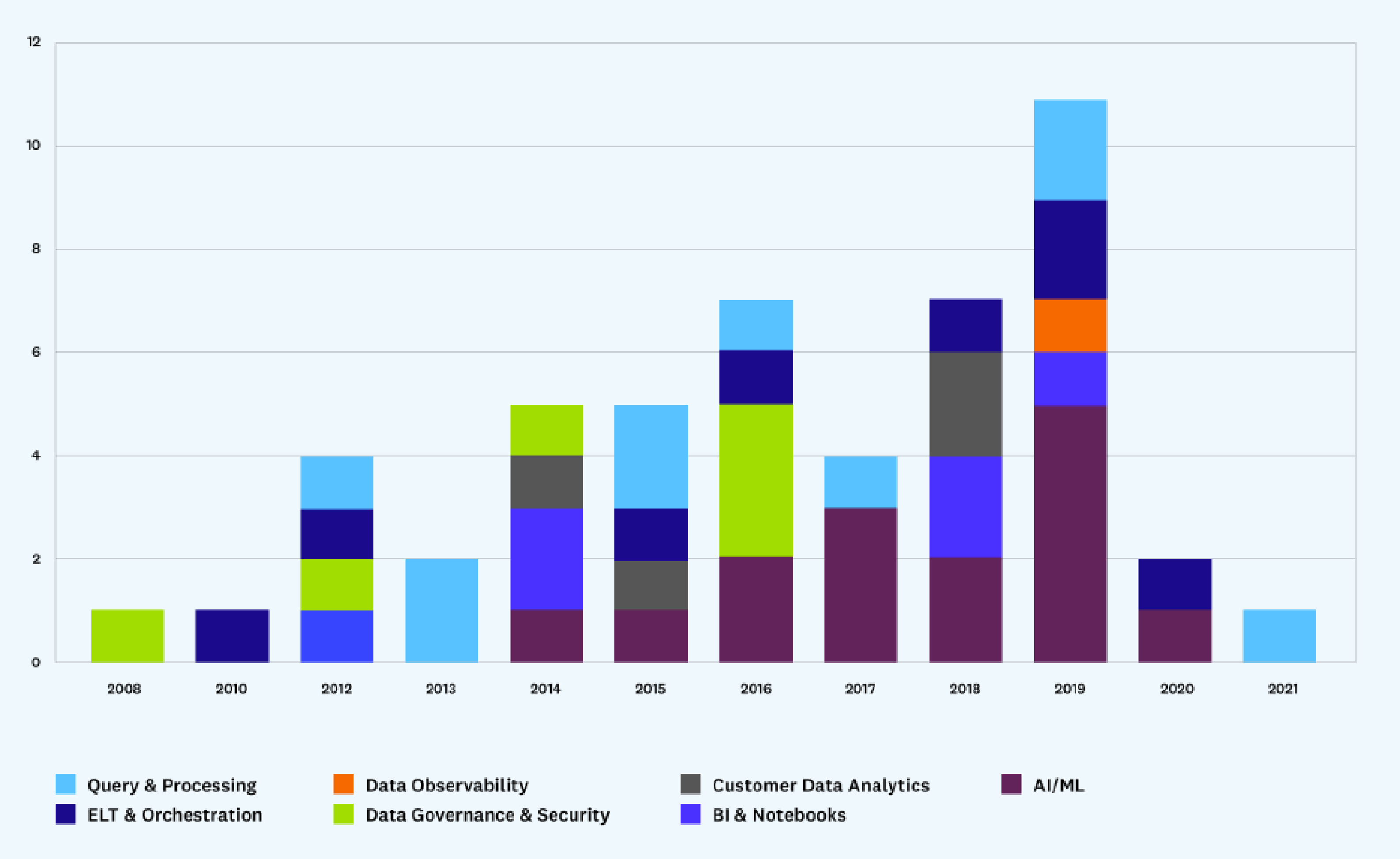
7 Subcategories

- AI/ML is the biggest category by the number of companies, largely because the space is still evolving and requires a new separate set of tools to train, measure, and production size models.



Per Category Investment

- AI/ML companies are picking up more investor interest than ever.



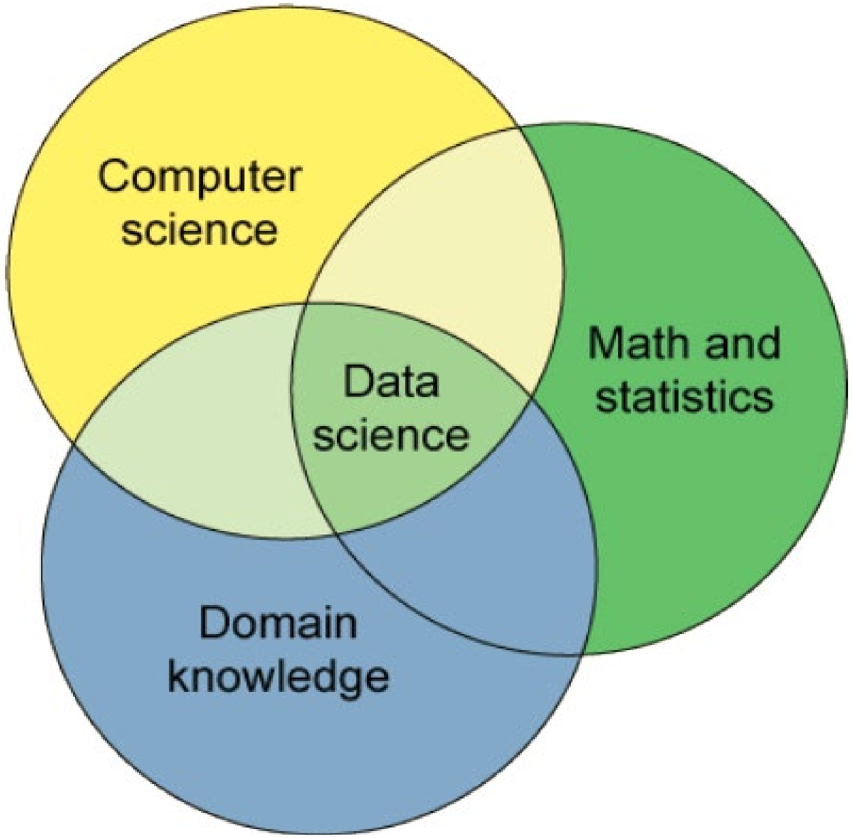
Data Science: Any new intellectual content?

- What does it mean to Computer Science?

- 1970’s: EE + Math → Computer Science
- 2010’s: CS + Stats/ML + Domain → Data Science

- Is something fundamental emerging here?

- Data Science is a very broad discipline
- Data Science PhD?
  - PhD normally with a narrow field with depth…



ـــed on Drew Conway, NYU

Attendance question!

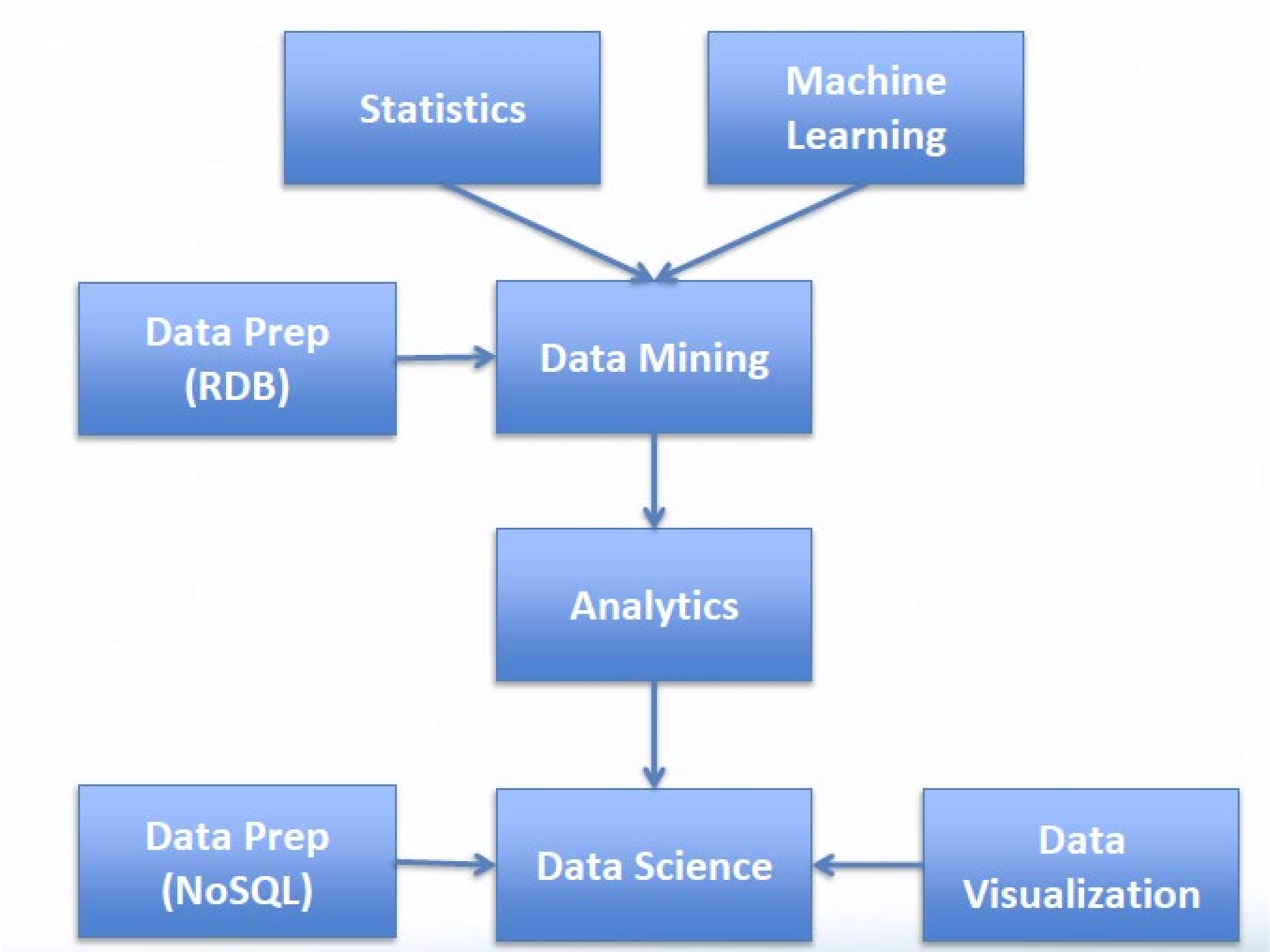
What is data science?

→ Computer Science

+ Statistics/Machine Leaning

+ Domain

Elements of Data Science



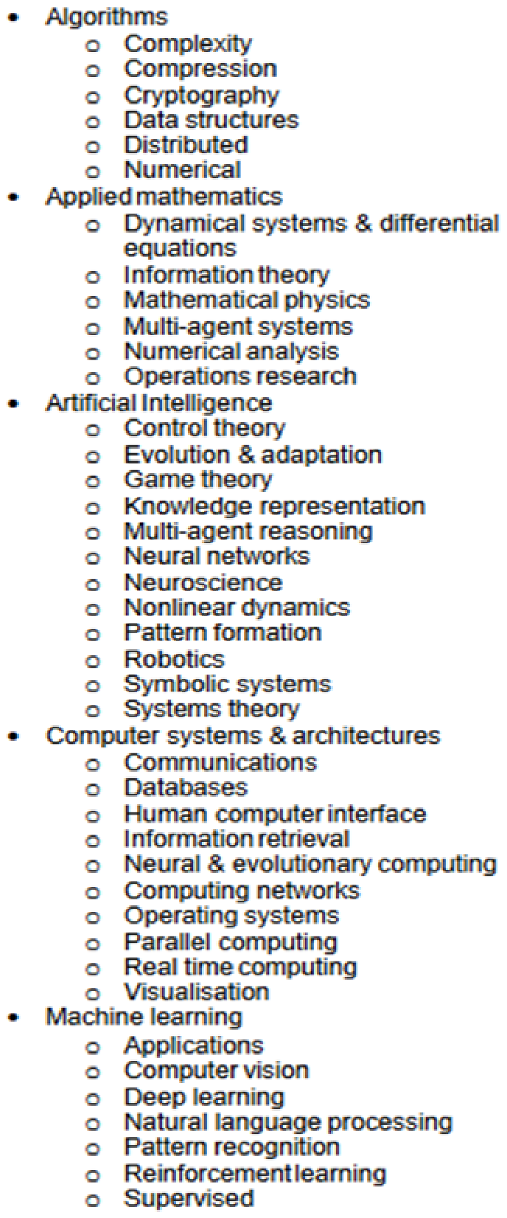
Buzzwords in Data Science

Many guidance and definitions in Web.

e.g. https://datascience.foundation/downloadpdf/142/whitepaper

- Artificial Intelligence (AI)
- Learning System or Algorithm
- How to decide which Learning Algorithm to use
- Machine Learning
- Machine Learning Algorithms
- Deep Learning
- Artificial Neural Network (ANN)
- Machine Learning Vs Deep Learning
- AI Vs ML Vs DL
- Deep Learning Vs Neural Network
- Data Science
- Data Science Flow Chart
- Why Deep Learning and why not SVM?
- What is deep learning? Why is this a growing trend in machine learning? Why not use SVMs?
- Five main reasons why deep learning is so popular
- Explainable AI (XAI)
- …

ATI Theme List



- Algorithms
  - Complexity
  - Compression
  - Cryptography
  - Data structures
  - Distributed
  - Numerical
- Applied mathematics
  - Dynamical systems & differential equations
  - Information theory
  - Mathematical physics
  - Multi-agent systems
  - Numerical analysis
  - Operations research
- Artificial Intelligence
  - Control theory
  - Evolution & adaptation
  - Game theory
  - Knowledge representation
  - Multi-agent reasoning
  - Neural Networks
  - Neuroscience
  - Nonlinear dynamics
  - Pattern formation
  - Robotics
  - Symbolic systems
  - Systems theory
- Computer systems & architectures
  - Communications
  - Databases
  - Human computer interface
  - Information retrieval
  - Neural & evolutionary computing
  - Computing networks
  - Operating systems
  - Parallel computing
  - Real time computing
  - Visualisation
- Machine learning
  - Applications
  - Computer vision
  - Deep learning
  - Natural language processing
  - Pattern recognition
  - Reinforcement learning
  - Supervised

2020 Buzzword highlights

Some Buzzwords shaped 2020

- **Data engineer** – as data science continues to evolve, the importance of data is becoming more clear → a new emphasis on engineers who could work directly getting the most out of datasets.

- **AutoML –** more than just machine learning (ML) → the process automated and delivered to the public.

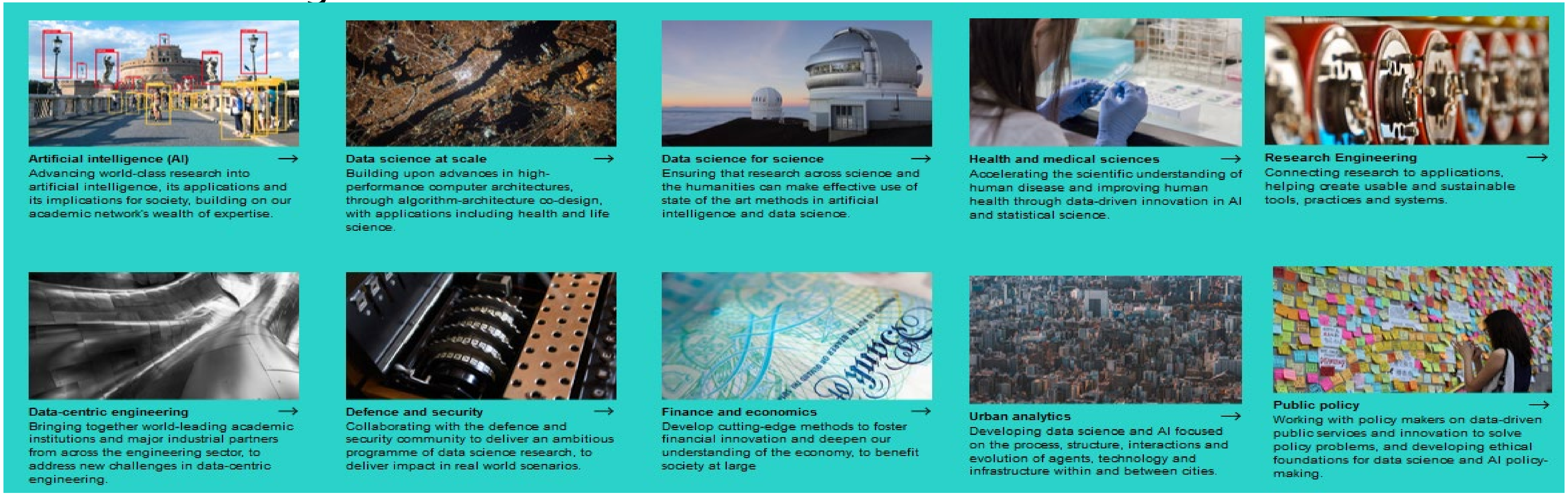
- **Explainability –** Building complex models is great, but without the visibility into how they work, they're not as good as they can be → explainability became a major goal of data scientists looking to understand their models and make them better.

from Explorium

Alan Turing Institute (ATI)

- Established in 2015 in London as a National Institute for Data Science
- >£20M Capital Investment from Government
- Translating output into practice https://www.turing.ac.uk

Research Programmes:



Artificial intelligence (AI) →
Advancing world-class research into artificial intelligence, its applications and its implications for society, building on our academic network's wealth of expertise.

Data science at scale →
Building upon advances in high-performance computer architectures, through algorithm-architecture co-design, with applications including health and life science.

Data science for science →
Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

Health and medical sciences →
Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

Research Engineering →
Connecting research to applications, helping create usable and sustainable tools, practices and systems.

Data-centric engineering →
Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

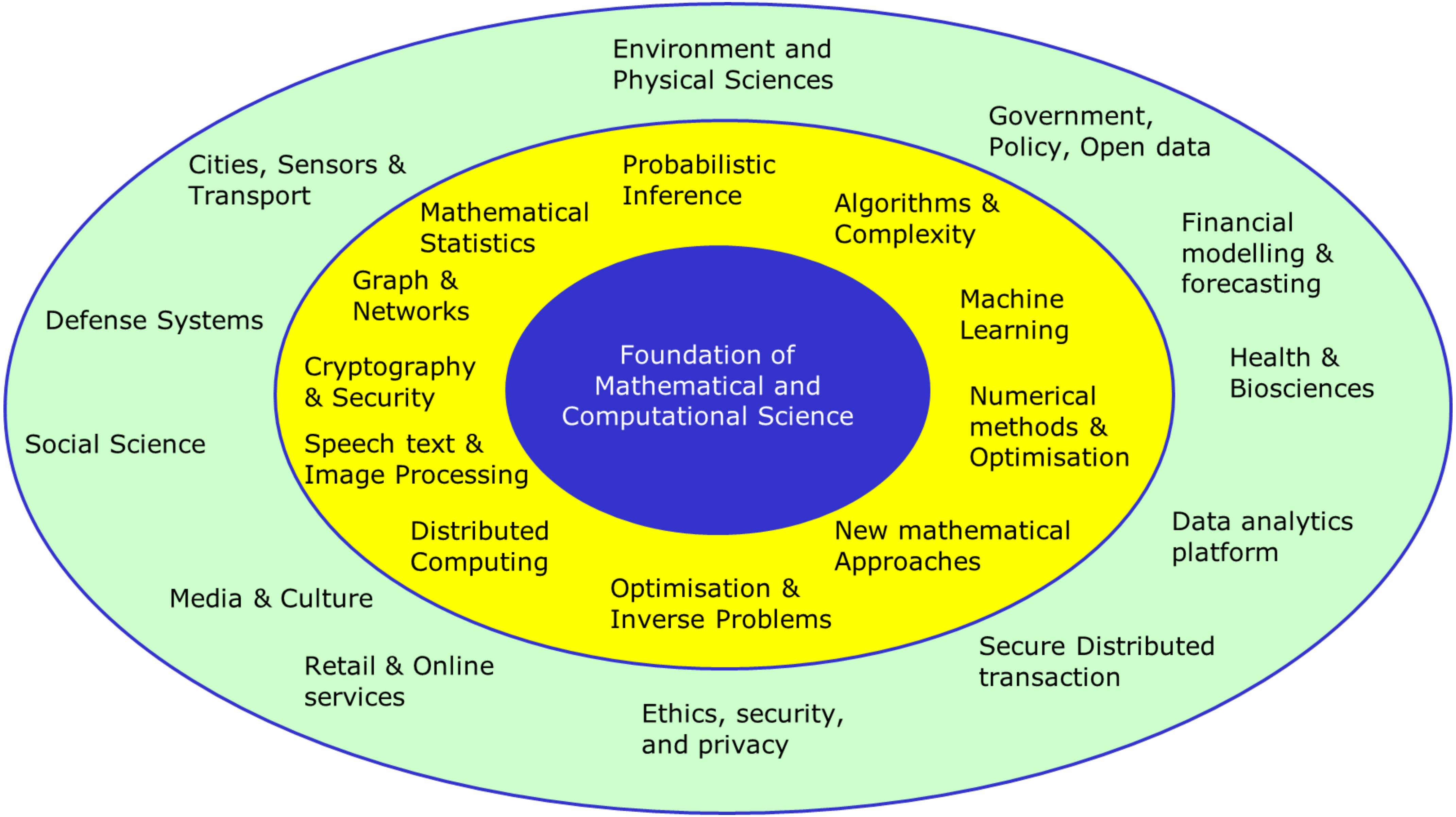
Defence and security →
Collaborating with the defence and security community to deliver an ambitious programme of data science research, to deliver impact in real world scenarios.

Finance and economics →
Develop cutting-edge methods to foster financial innovation and deepen our understanding of the economy, to benefit society at large

Urban analytics →
Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Public policy →
Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Landscape of Data Science (~=Buzzwords Map)



Foundation of Mathematical and Computational Science

Probabilistic Inference

Algorithms & Complexity

Machine Learning

Numerical methods & Optimisation

New mathematical Approaches

Optimisation & Inverse Problems

Distributed Computing

Speech text & Image Processing

Cryptography & Security

Graph & Networks

Mathematical Statistics

Environment and Physical Sciences

Government, Policy, Open data

Financial modelling & forecasting

Health & Biosciences

Data analytics platform

Secure Distributed transaction

Ethics, security, and privacy

Retail & Online services

Media & Culture

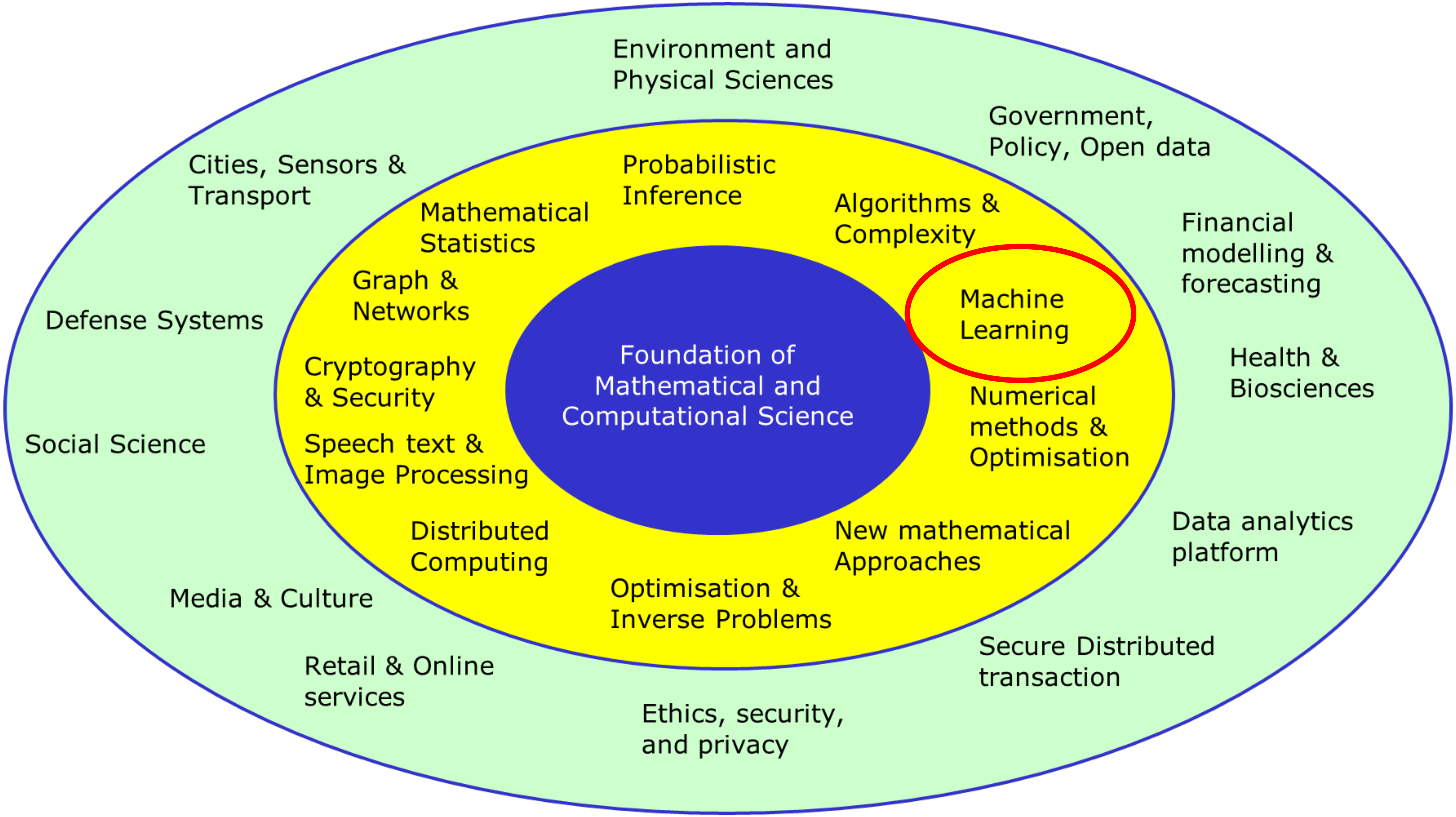
Social Science

Defense Systems

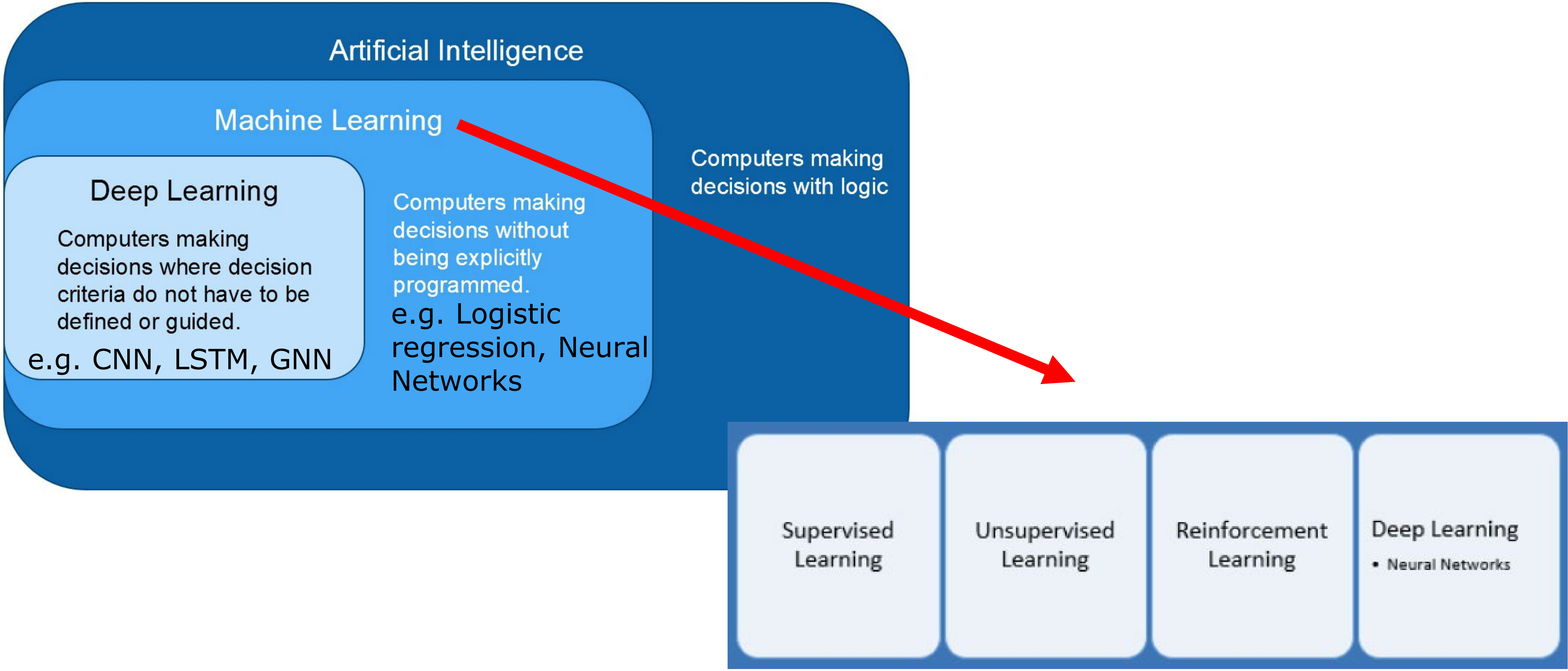
Cities, Sensors & Transport

ATI

AI, Machine Learning, Deep Learning


Environment and Physical Sciences
Government, Policy, Open data
Cities, Sensors & Transport
Financial modelling & forecasting
Defense Systems
Health & Biosciences
Social Science
Data analytics platform
Media & Culture
Secure Distributed transaction
Retail & Online services
Ethics, security, and privacy
Probabilistic Inference
Mathematical Statistics
Algorithms & Complexity
Graph & Networks
Machine Learning
Cryptography & Security
Numerical methods & Optimisation
Speech text & Image Processing
Distributed Computing
New mathematical Approaches
Optimisation & Inverse Problems
Foundation of Mathematical and Computational Science

Deep Learning, Machine Learning, and AI…



Artificial Intelligence

Computers making decisions with logic

Machine Learning

Computers making decisions without being explicitly programmed.

e.g. Logistic regression, Neural Networks

Deep Learning

Computers making decisions where decision criteria do not have to be defined or guided.

e.g. CNN, LSTM, GNN

Supervised Learning

Unsupervised Learning

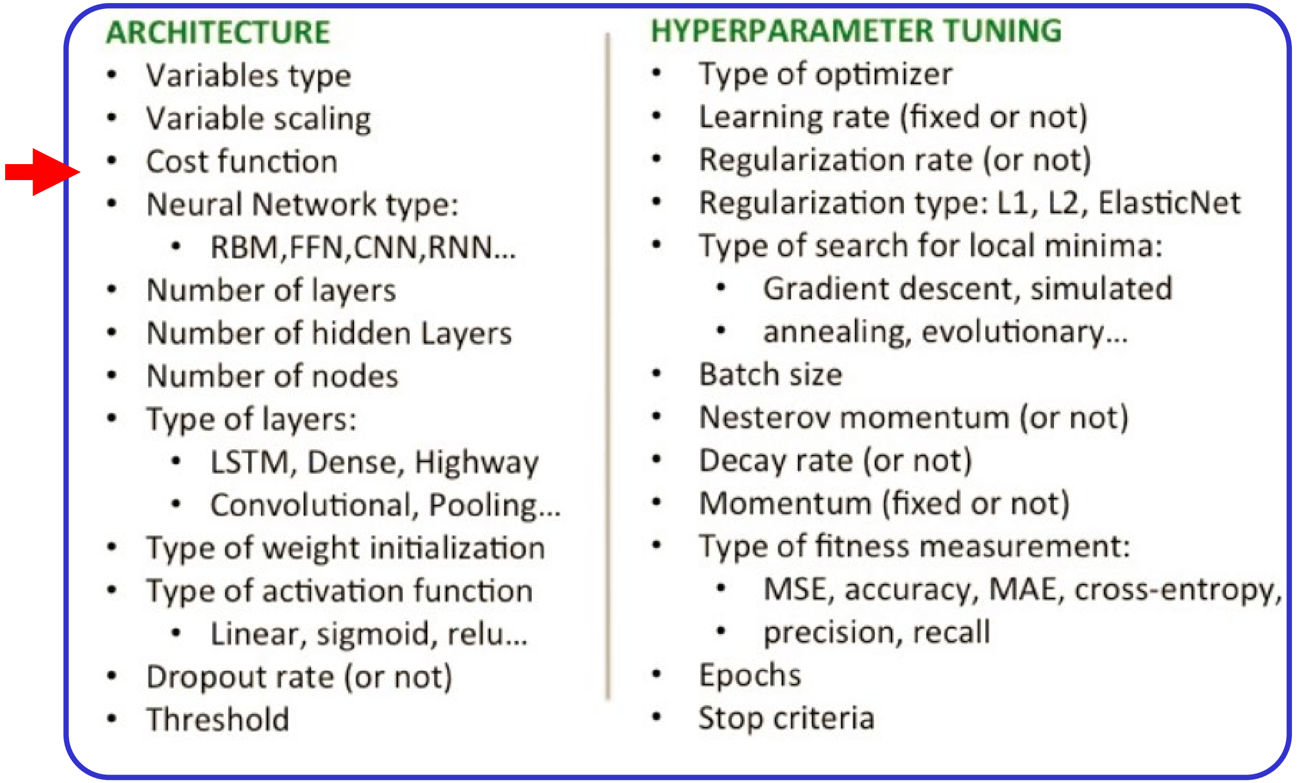
Reinforcement Learning

Deep Learning

- Neural Networks

Four Pictures Illustrating ML Concepts

- Neural Networks: The Backpropagation algorithm
- Cheat Sheet on Probability
- 24 Neural Network Adjustments →
- Matrix Multiplication in NN



ARCHITECTURE

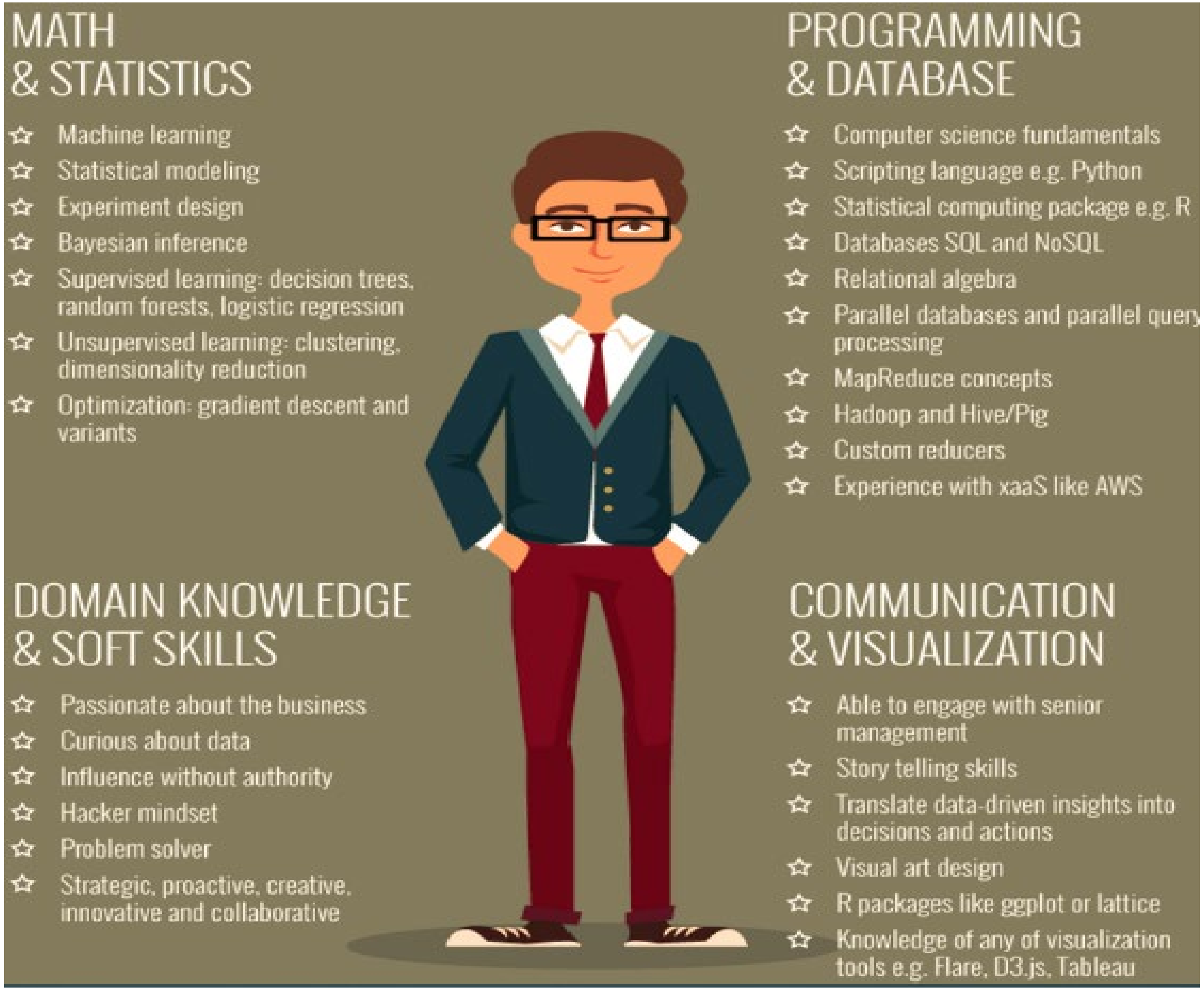
- Variables type
- Variable scaling
- Cost function
- Neural Network type:
  - RBM,FFN,CNN,RNN…
- Number of layers
- Number of hidden Layers
- Number of nodes
- Type of layers:
  - LSTM, Dense, Highway
  - Convolutional, Pooling…
- Type of weight initialization
- Type of activation function
  - Linear, sigmoid, relu…
- Dropout rate (or not)
- Threshold

HYPERPARAMETER TUNING

- Type of optimizer
- Learning rate (fixed or not)
- Regularization rate (or not)
- Regularization type: L1, L2, ElasticNet
- Type of search for local minima:
  - Gradient descent, simulated
  - annealing, evolutionary…
- Batch size
- Nesterov momentum (or not)
- Decay rate (or not)
- Momentum (fixed or not)
- Type of fitness measurement:
  - MSE, accuracy, MAE, cross-entropy,
  - precision, recall
- Epochs
- Stop criteria

See https://www.datasciencecentral.com/profiles/blogs/four-great-pictures-illustrating-machine-learning-concepts

Modern Data Scientist: The sexiest job of 21th century



MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

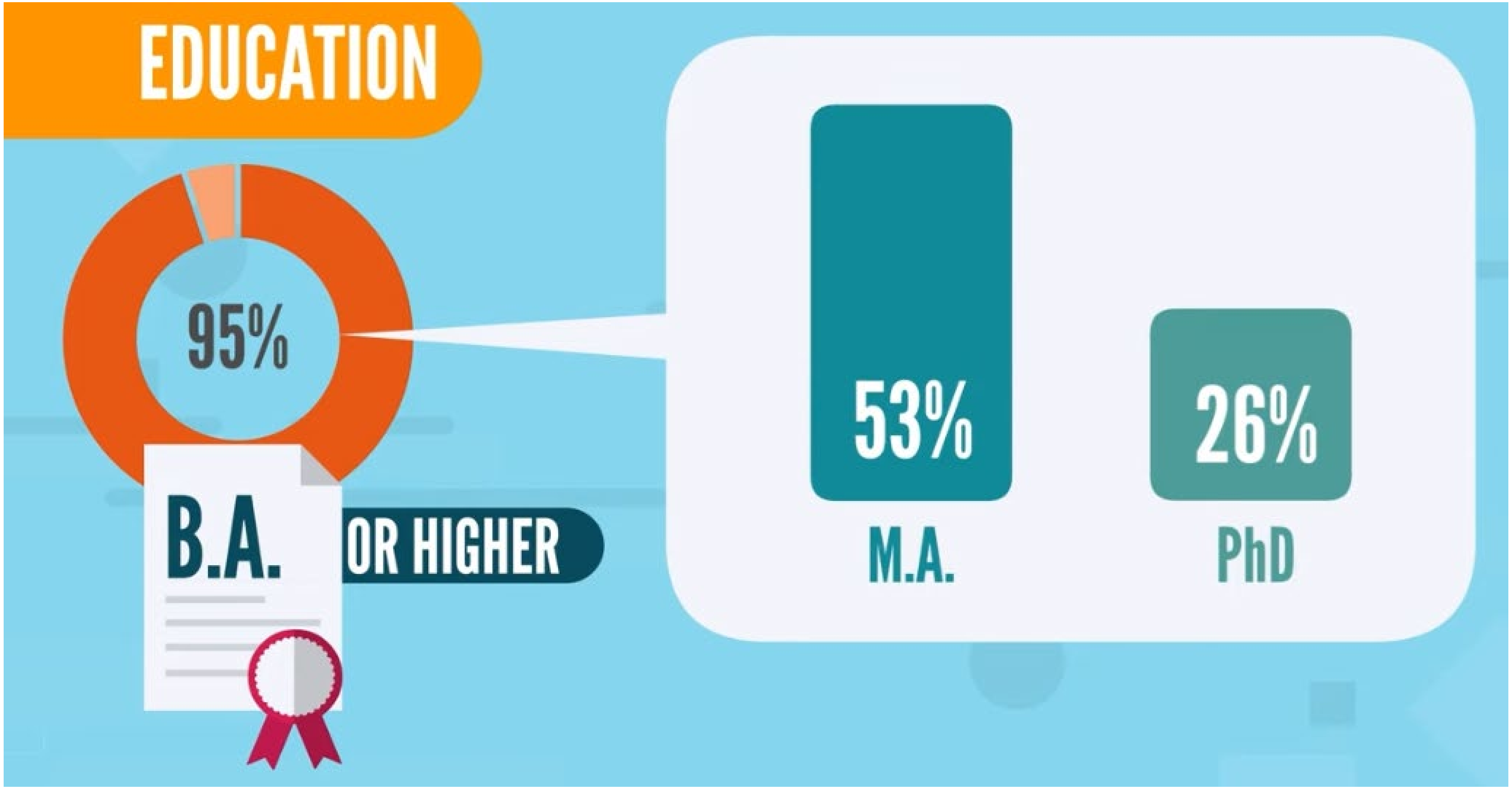
COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

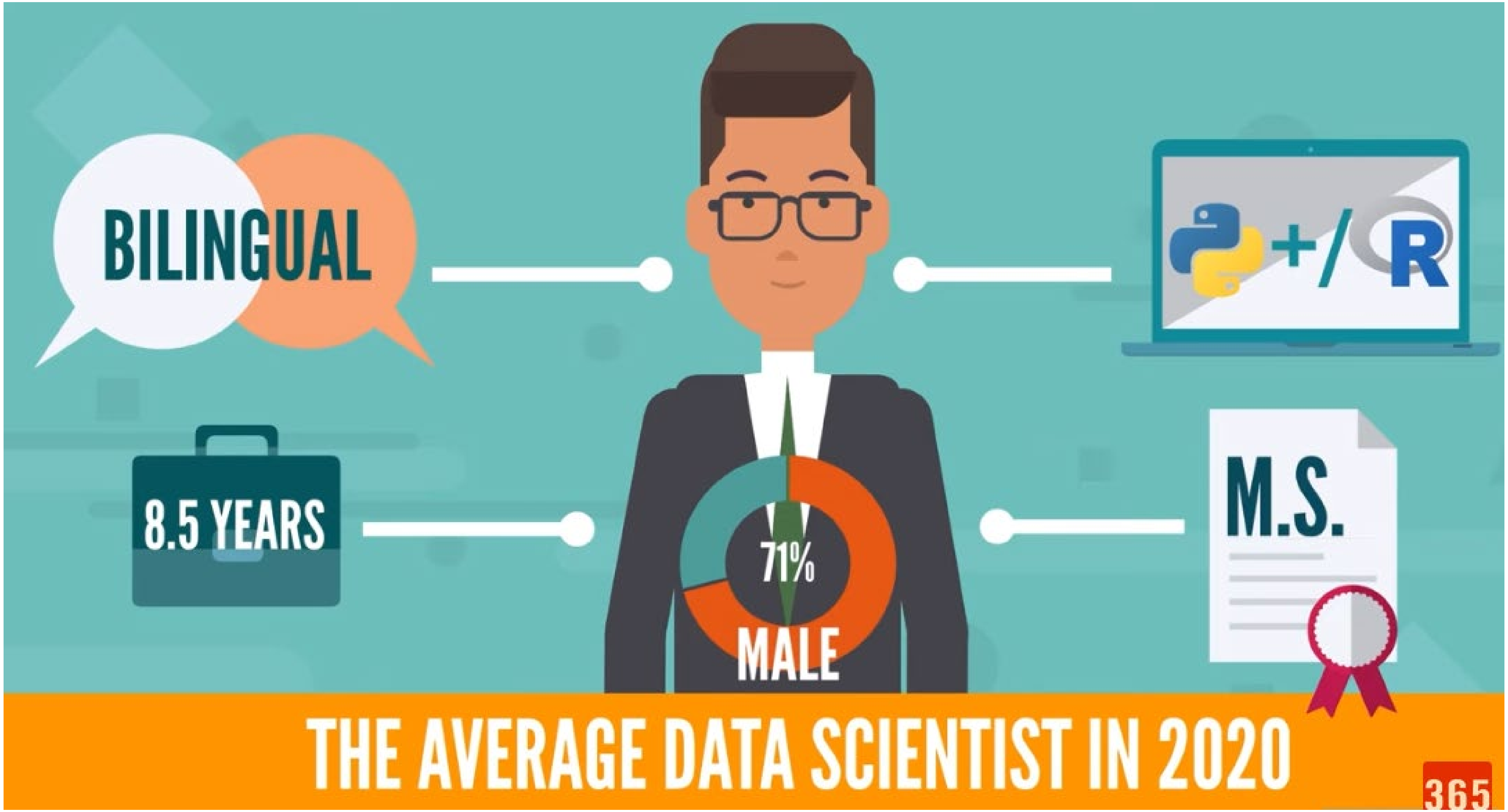
"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning"

Hilary Mason, chief scientist at bit.ly

Educational Background

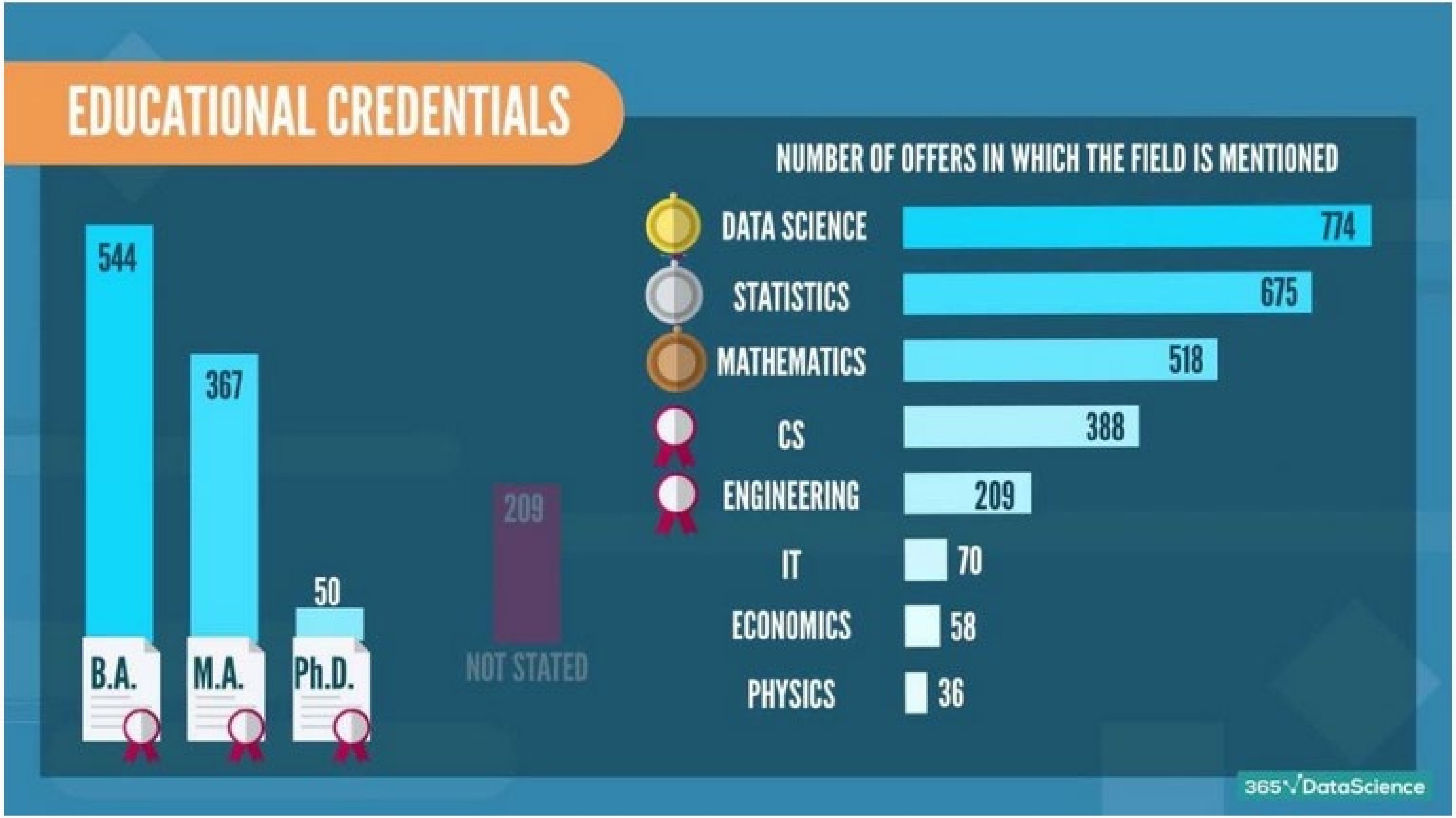


Typical Data Scientist in 2020



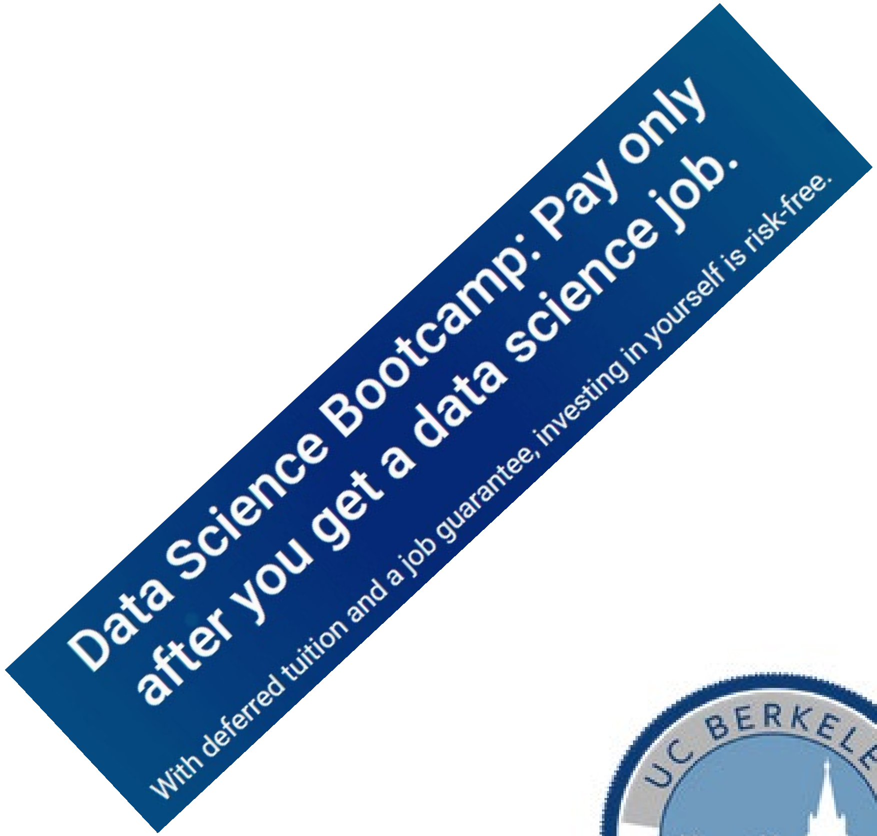
From 365:Data Science

Fields of study preferred



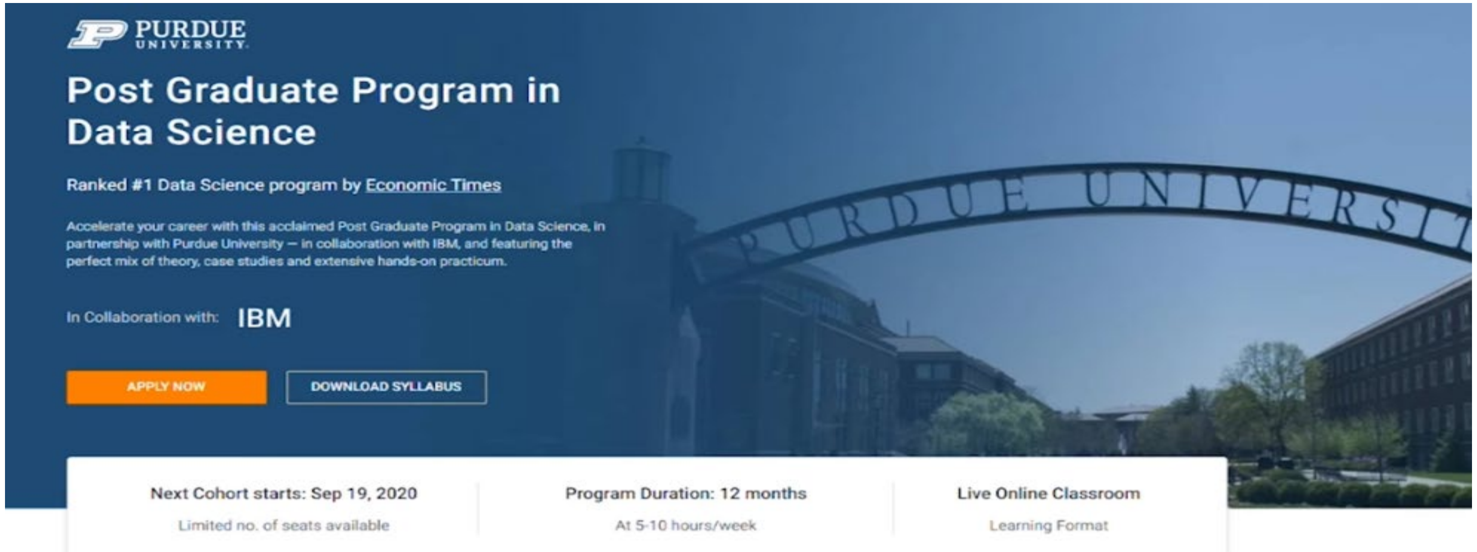
Medium

Many Courses offered: e.g. Master Certificate



Data Science Bootcamp: Pay only after you get a data science job.

With deferred tuition and a job guarantee, investing in yourself is risk-free.



PURDUE UNIVERSITY

Post Graduate Program in Data Science

Ranked #1 Data Science program by Economic Times

Accelerate your career with this acclaimed Post Graduate Program in Data Science, in partnership with Purdue University – in collaboration with IBM, and featuring the perfect mix of theory, case studies and extensive hands-on practicum.

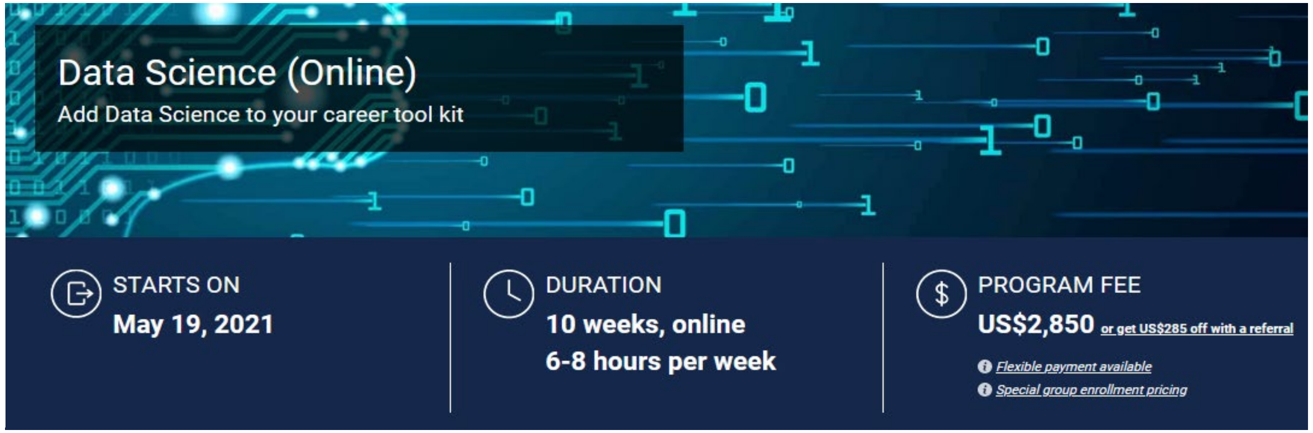
In Collaboration with: IBM

APPLY NOW

DOWNLOAD SYLLABUS

Next Cohort starts: Sep 19, 2020	Program Duration: 12 months	Live Online Classroom
Limited no. of seats available	At 5-10 hours/week	Learning Format

UC BERKELEY Executive ★ Education



Data Science (Online)

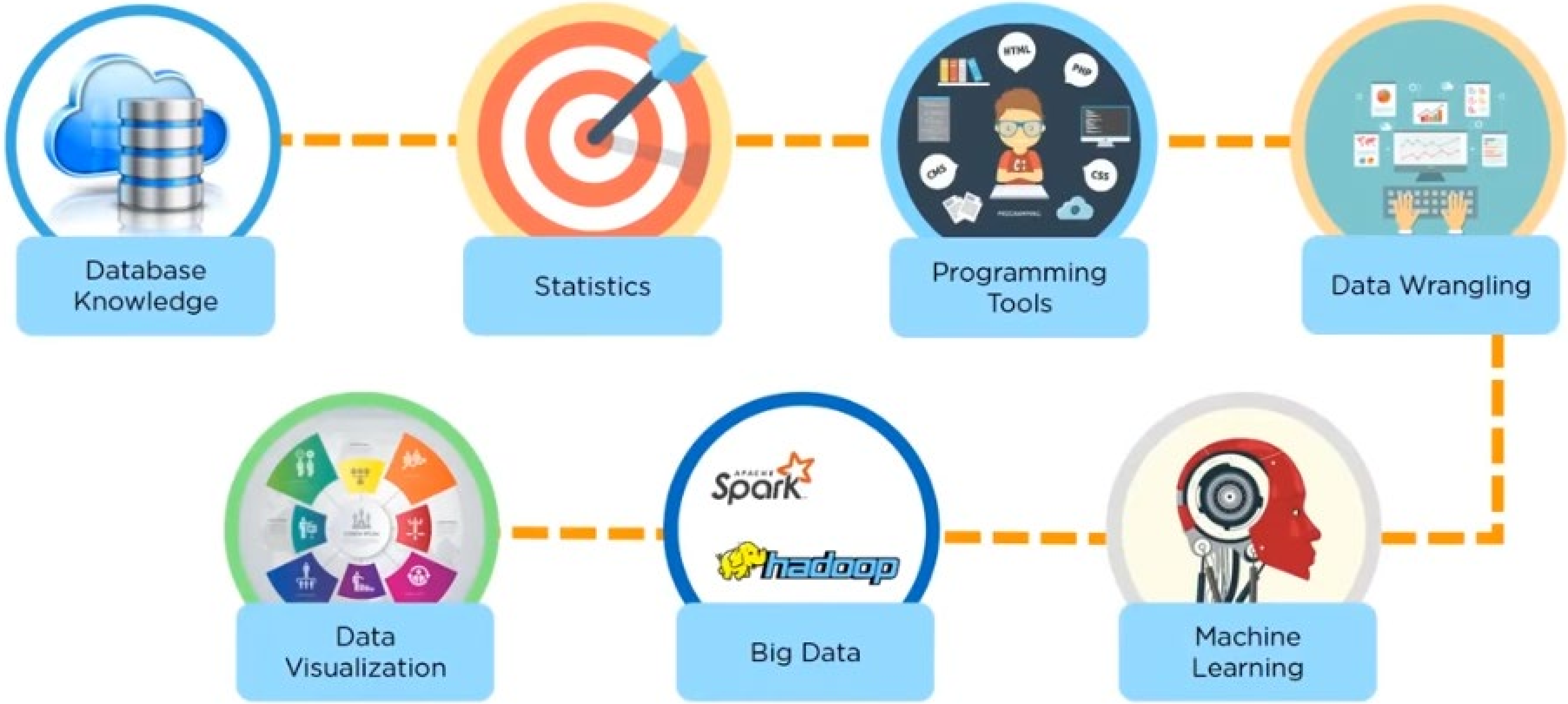
Add Data Science to your career tool kit

STARTS ON
May 19, 2021

DURATION
10 weeks, online
6-8 hours per week

PROGRAM FEE
US$2,850 or get US$285 off with a referral
Flexible payment available
Special group enrollment pricing

Skills required in Data Science



Database Knowledge

Statistics

Programming Tools

Data Wrangling

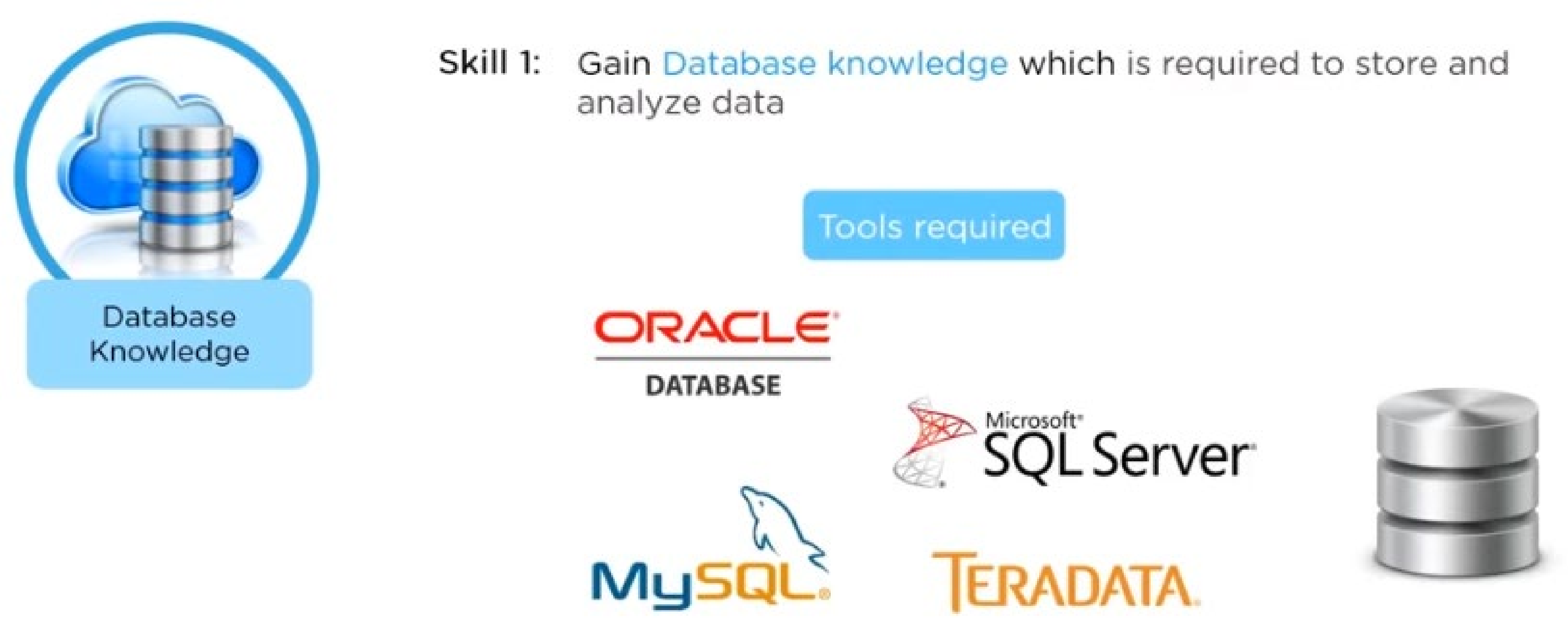
Machine Learning

Big Data

Data Visualization

from Simlilearn

Skill 1: Database Knowledge

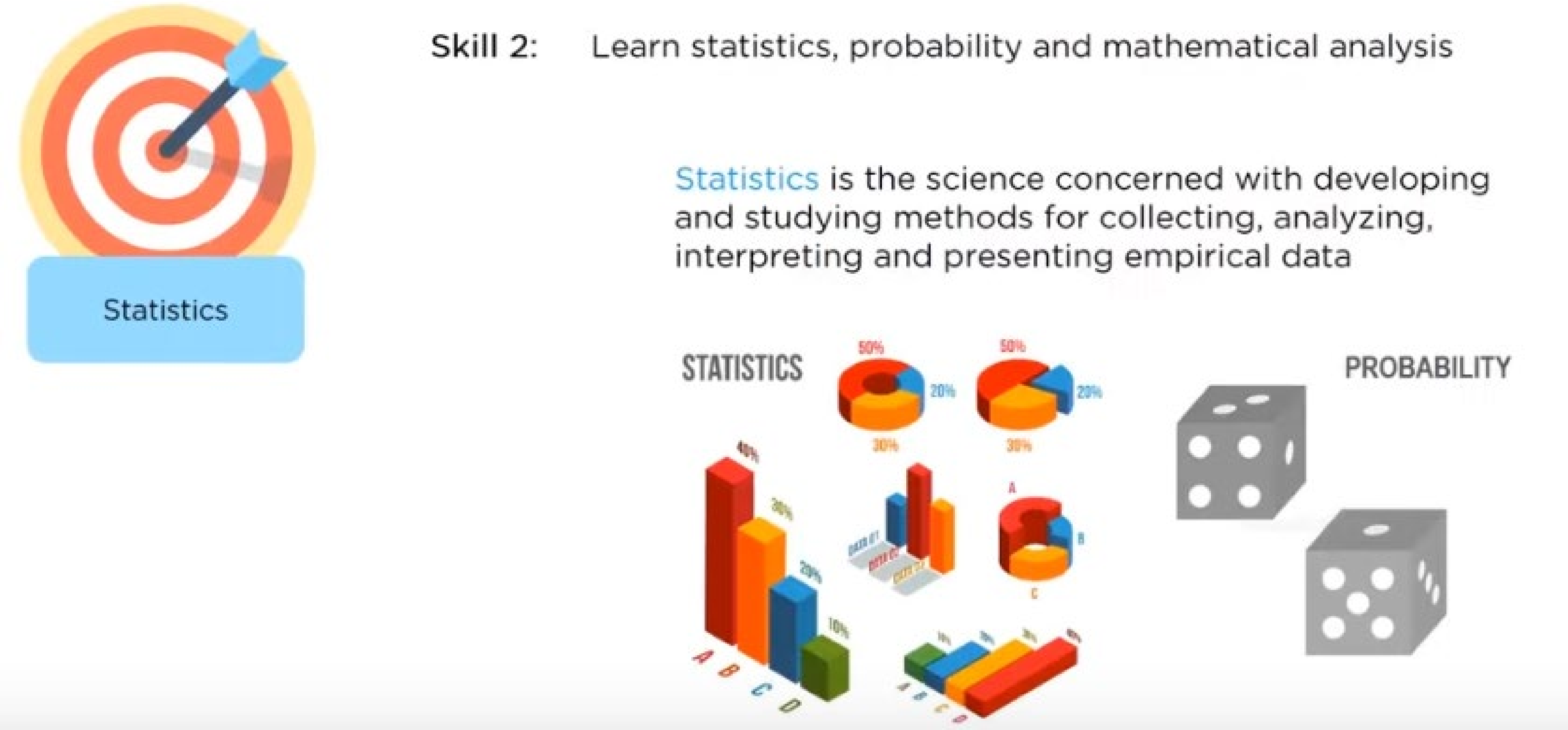


Skill 1: Gain Database knowledge which is required to store and analyze data

Tools required

from Simlilearn

Skill 2: Statistics

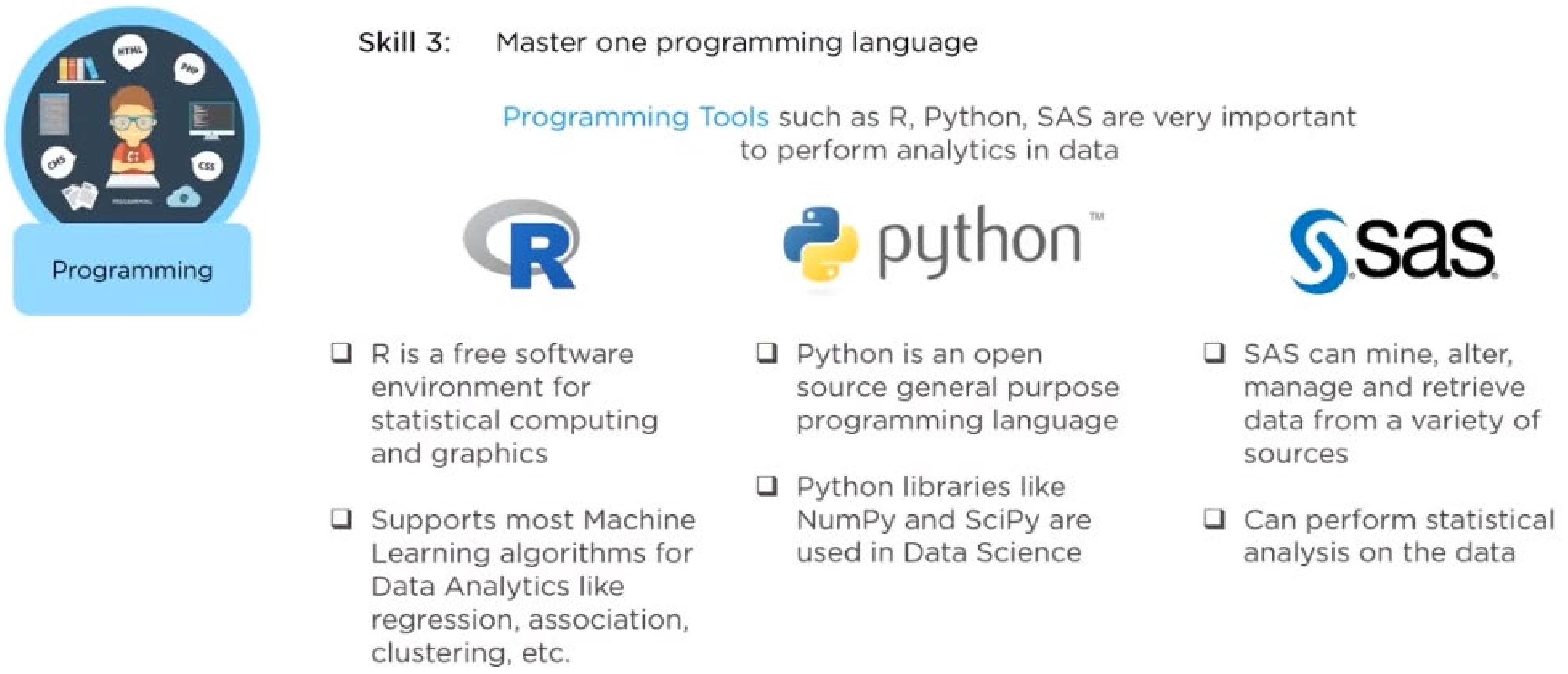


Skill 2: Learn statistics, probability and mathematical analysis

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data

from Simlilearn

Skill 3: Programming



Programming

Skill 3: Master one programming language

Programming Tools such as R, Python, SAS are very important to perform analytics in data

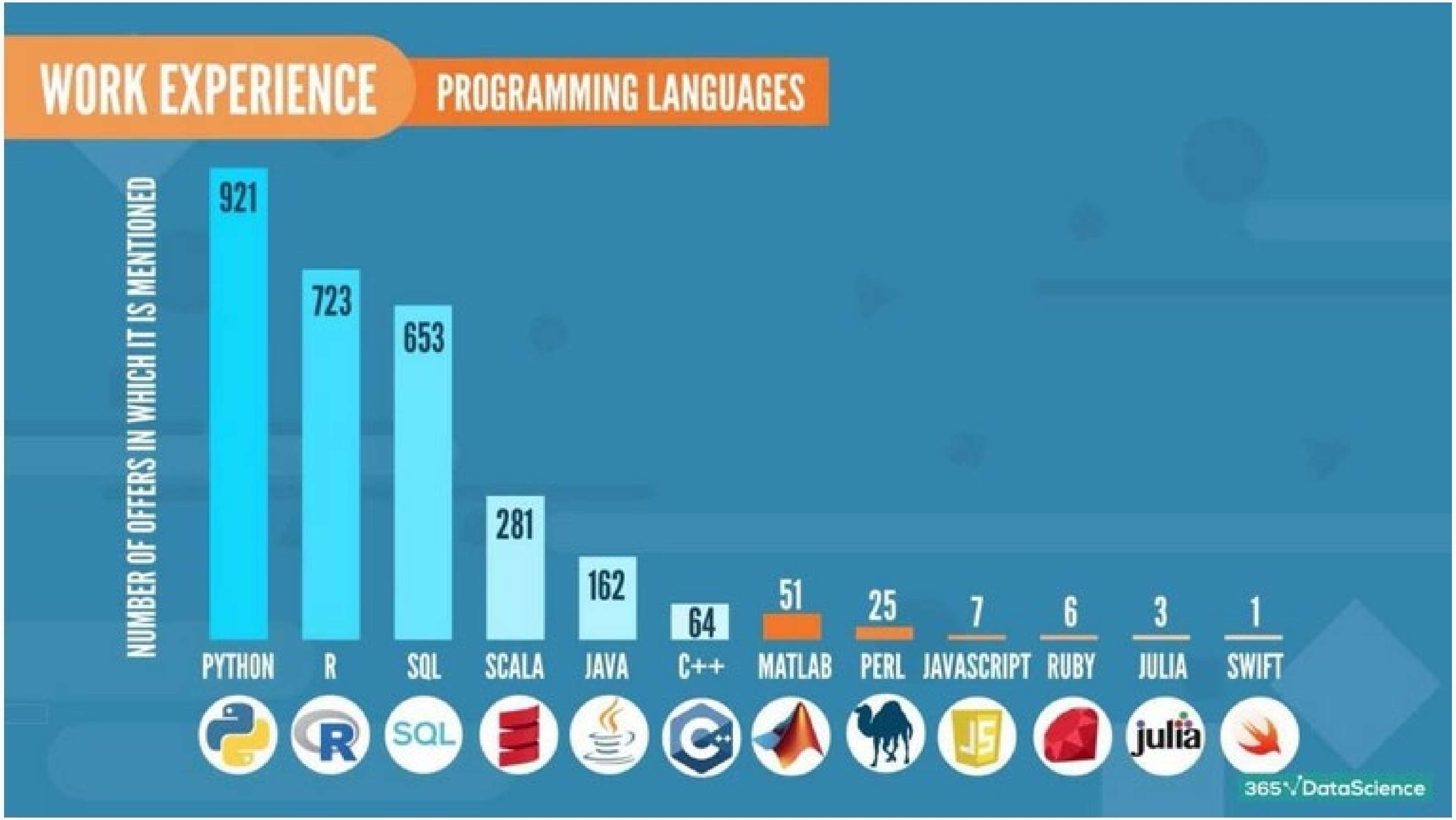
- ❑ R is a free software environment for statistical computing and graphics
- ❑ Supports most Machine Learning algorithms for Data Analytics like regression, association, clustering, etc.

- ❑ Python is an open source general purpose programming language
- ❑ Python libraries like NumPy and SciPy are used in Data Science

- ❑ SAS can mine, alter, manage and retrieve data from a variety of sources
- ❑ Can perform statistical analysis on the data

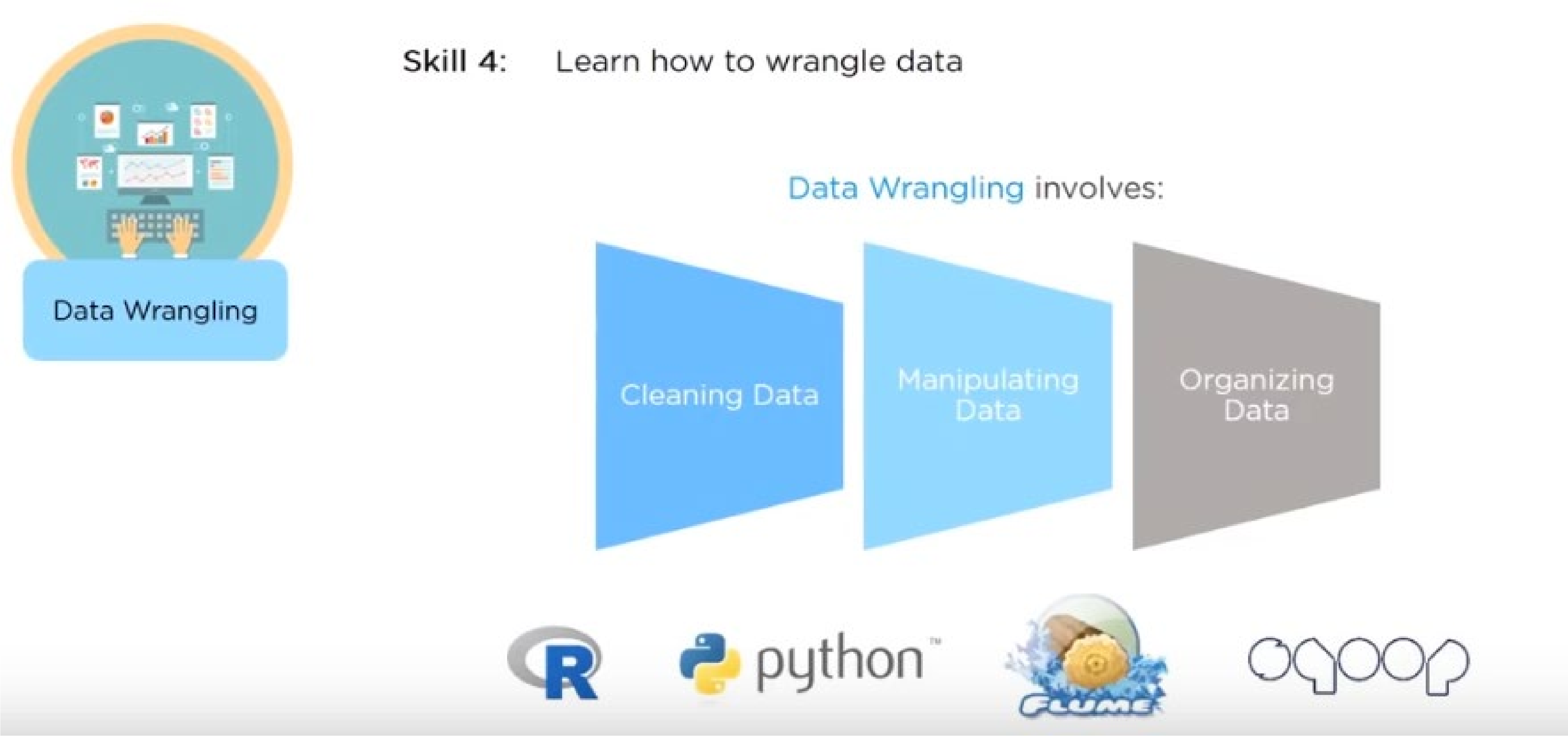
from Simlilearn

What Programming skills do you need?



From Medium

Skill 4: Data Wrangling



Data Wrangling

Skill 4: Learn how to wrangle data

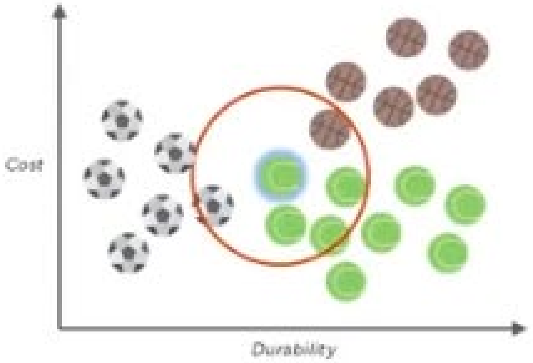
Data Wrangling involves:

- Cleaning Data
- Manipulating Data
- Organizing Data

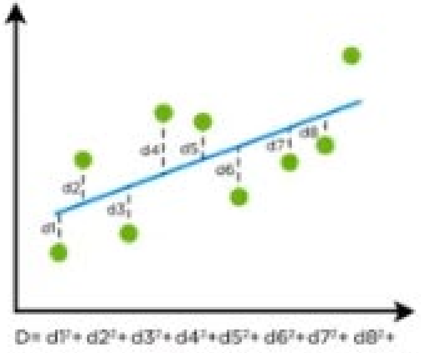
from Simlilearn

Skill 5: Machine Learning

Machine Learning can be achieved through various algorithms such as Regression, Naive Bayes, SVM, K Means Clustering, KNN and Decision Tree algorithms to name a few



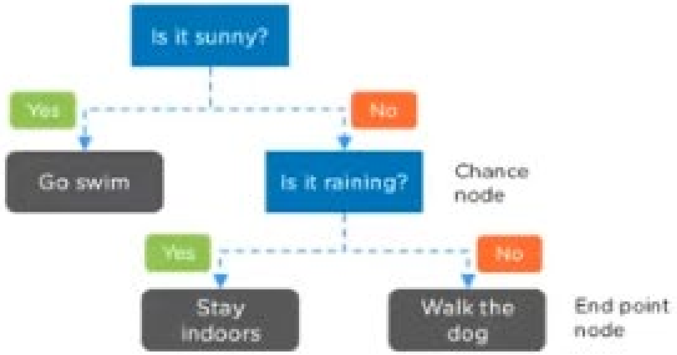
KNN



D= d1²+ d2²+ d3²+ d4²+d5²+ d6²+d7²+ d8²+

The regression line (blue) has the least value of D

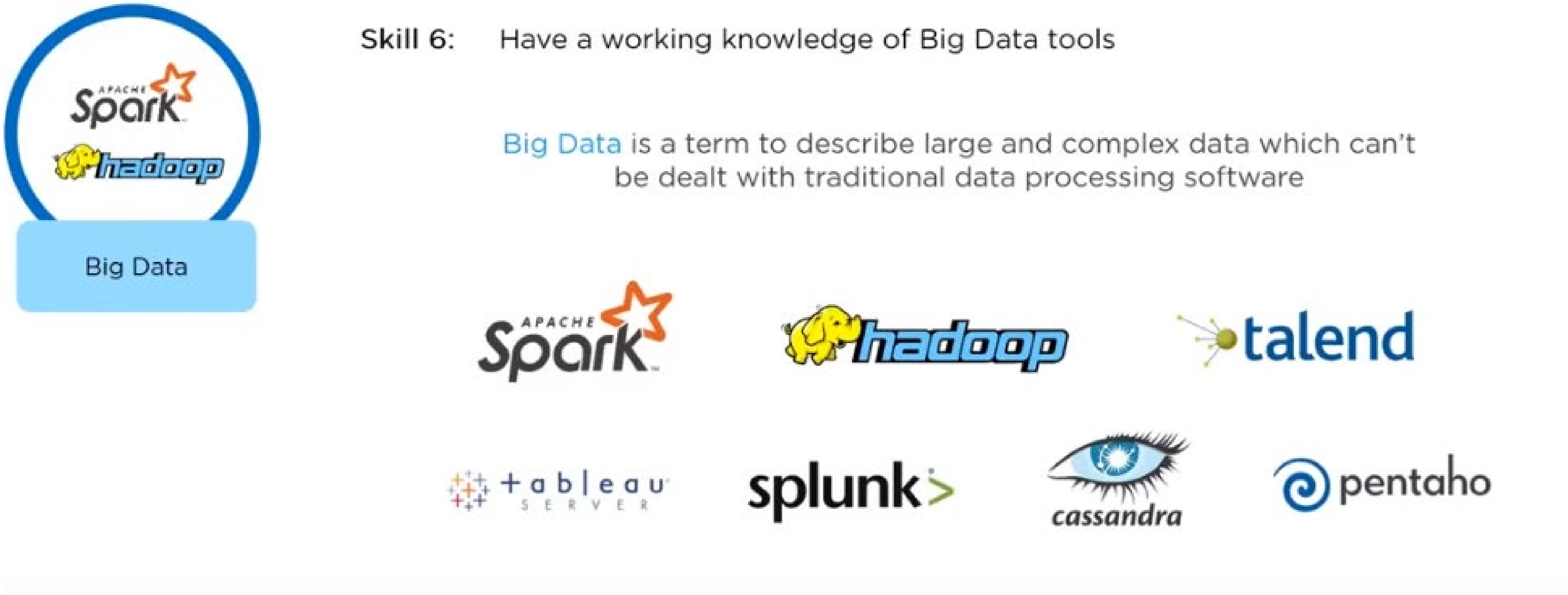
Linear Regression



Decision Tree

from Simlilearn

Skill 6: Big Data



Skill 6: Have a working knowledge of Big Data tools

Big Data is a term to describe large and complex data which can't be dealt with traditional data processing software

Big Data

from Simlilearn

Skill 7: Data Visualisation

Data Visualization

Skill 7: Develop the ability to visualize results

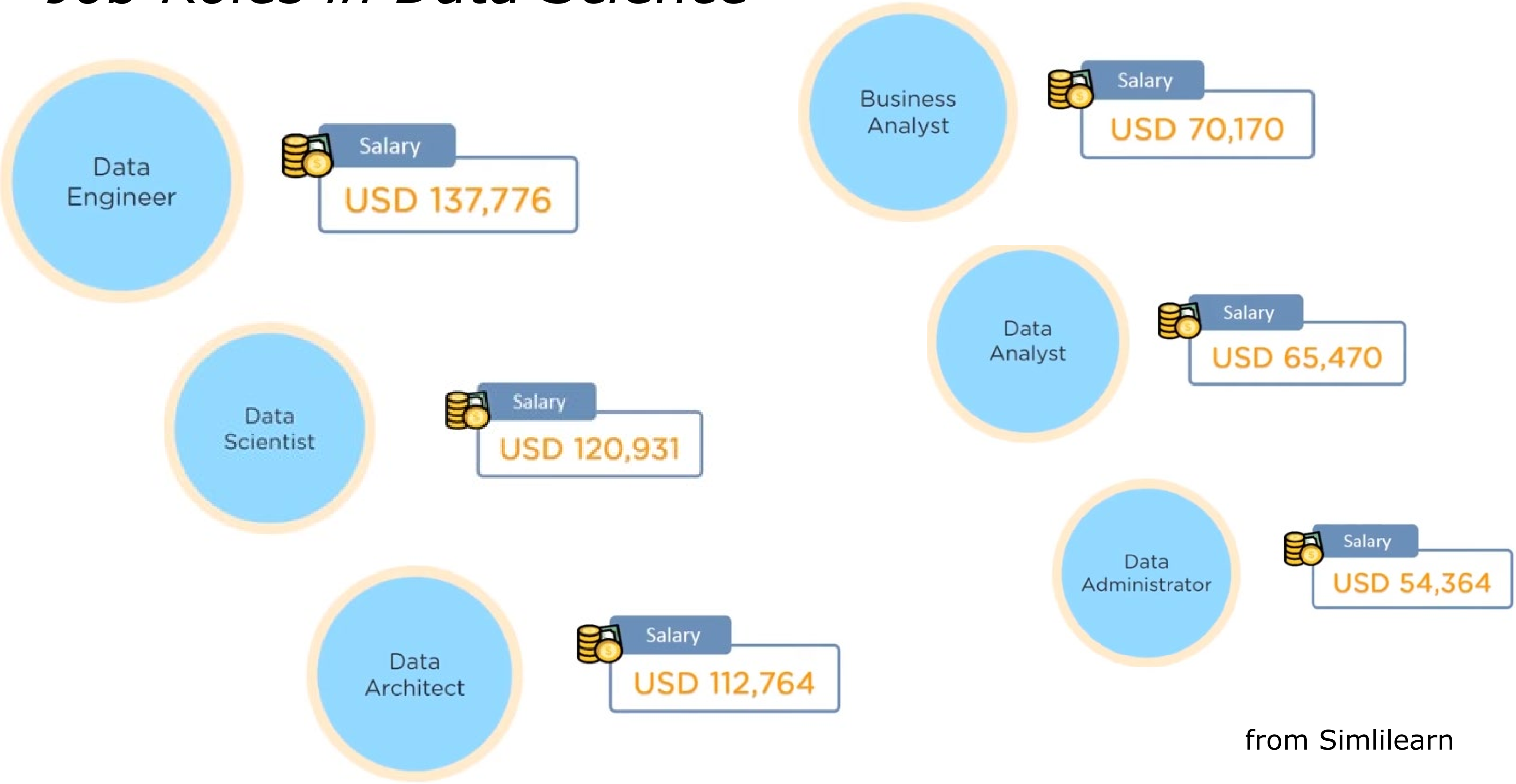
Data Visualization involves integrating different datasets, analyzing models and visualizing them in the form of diagrams, charts and graphs

tableau QlikView

Power BI Google Data Studio

from Simlilearn

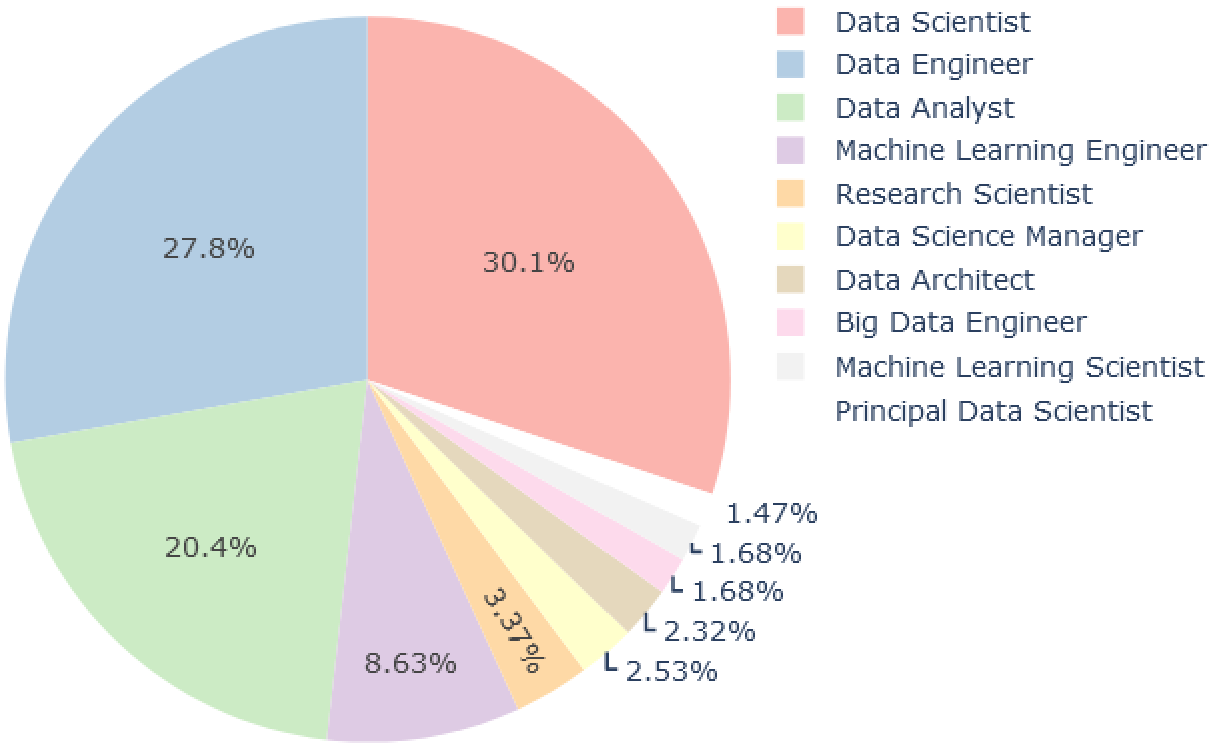
Job Roles in Data Science



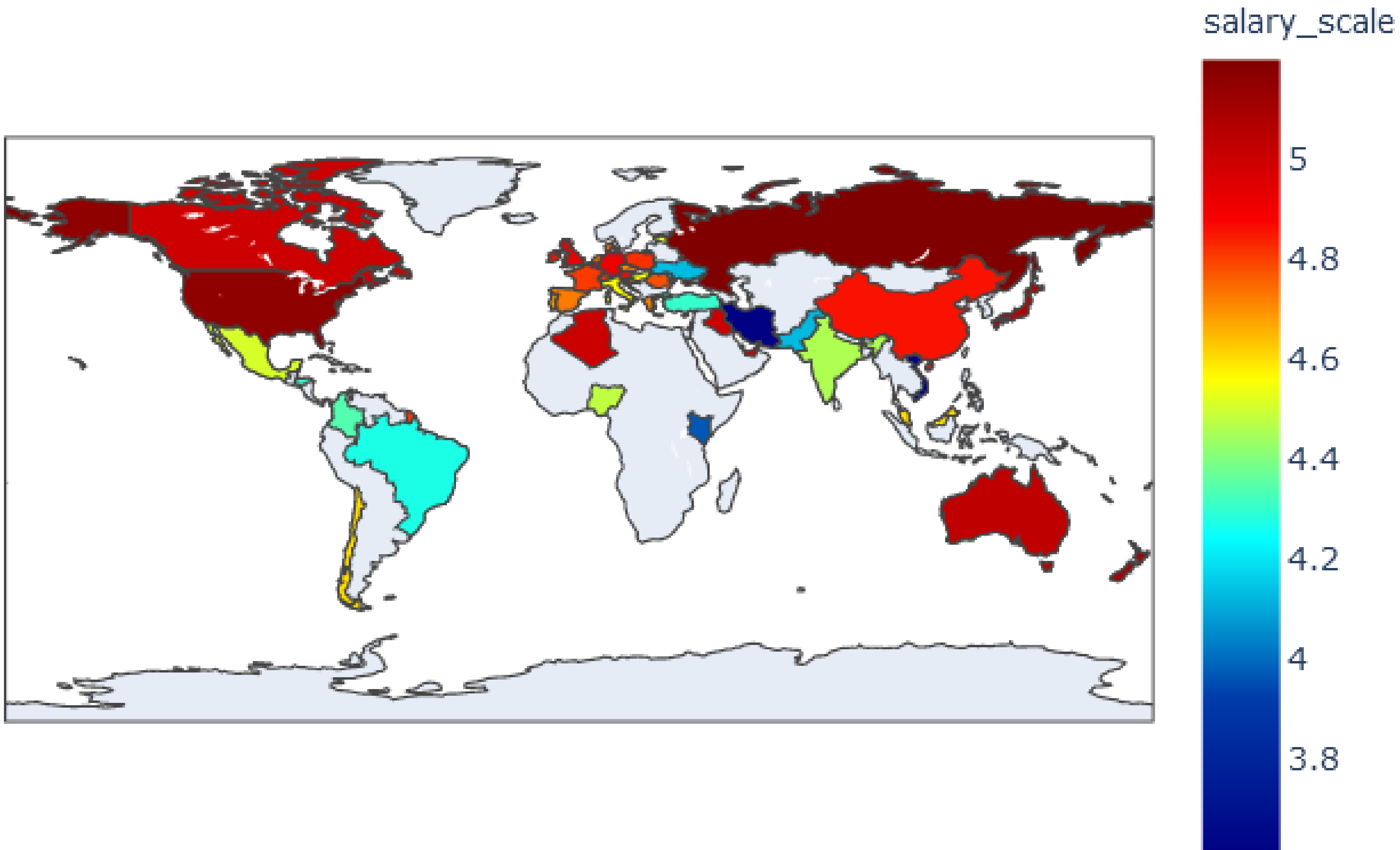
	Salary
Data Engineer	USD 137,776
Data Scientist	USD 120,931
Data Architect	USD 112,764
Business Analyst	USD 70,170
Data Analyst	USD 65,470
Data Administrator	USD 54,364

from Simlilearn

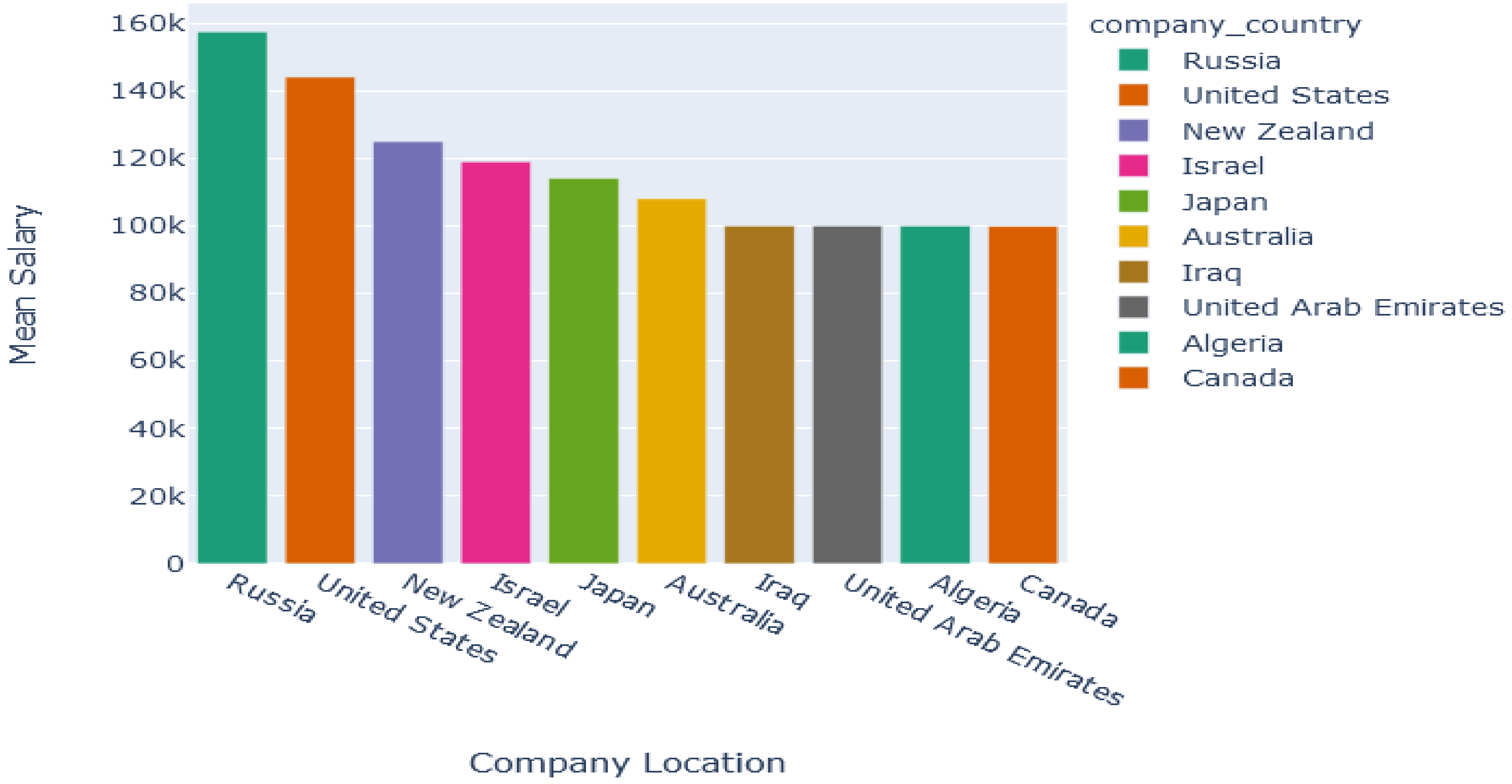
Distribution of Job Positions



Mean Salaries across the World



Average Salary as a Function of Company Location



Data Engineer

It's possible to get a job as a data science engineer by simply knowing Python, Amazon S3, and Postgres. Many of tools are also interchangeable.

- Programming - SQL, Python, Scala, C/C++, Java, Javascript
- Relational Databases - Oracle, PostgreSQL, MySQL
- Non-relational Databases - MongoDB, Google Firebase, Apache Cassandra
- Data Warehouses - Snowflake, Amazon Redshift, Google Bigquery
- Data Lakes - Amazon S3, Google Cloud Storage
- Distributed Computing Frameworks - Apache Spark, Apache Hadoop
- DevOps tools - Git, Docker, Kubernetes, Airflow, Amazon Lambda, Jenkins, JIRA

It seems Software Engineer – not specific to data science ?

Interview Query Blog

ML Interview Basic Questions

Q1. Explain different types of machine learning with their pros and cons?

Q2. What is difference between Regression and Classification ?

Q3. What is difference between Structure & Unstructured Data ?

Q4. What are assumption of Linear Regression (LR) ?

Q5. What is multicollinearity and why it is a problem in LR?

Q6. In LR, what is the value of the sum of the residuals for a given dataset?

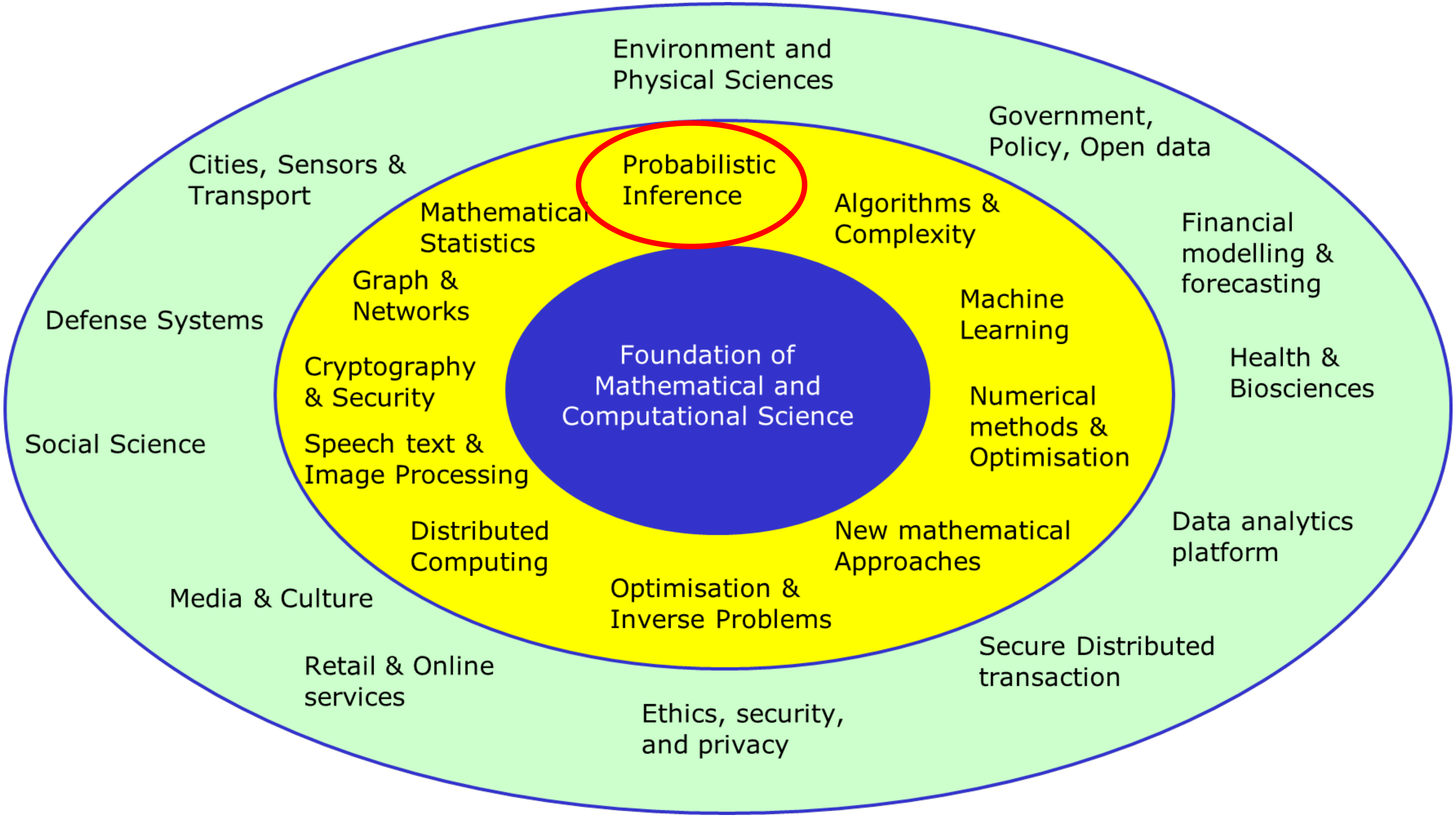
Q7. What is homoscedasticity & Heteroscedasticity ?

Q8. What are the reasons and effect of Heteroscedasticity on model ?

Q9. How to handle problem of Heteroscedasticity on model ?

See https://medium.com/@dishitaneve/ml-interview-questions-1-1-2c14c4dcdf18
for the answers

My Pick of Buzzword 1: Probabilistic Programming

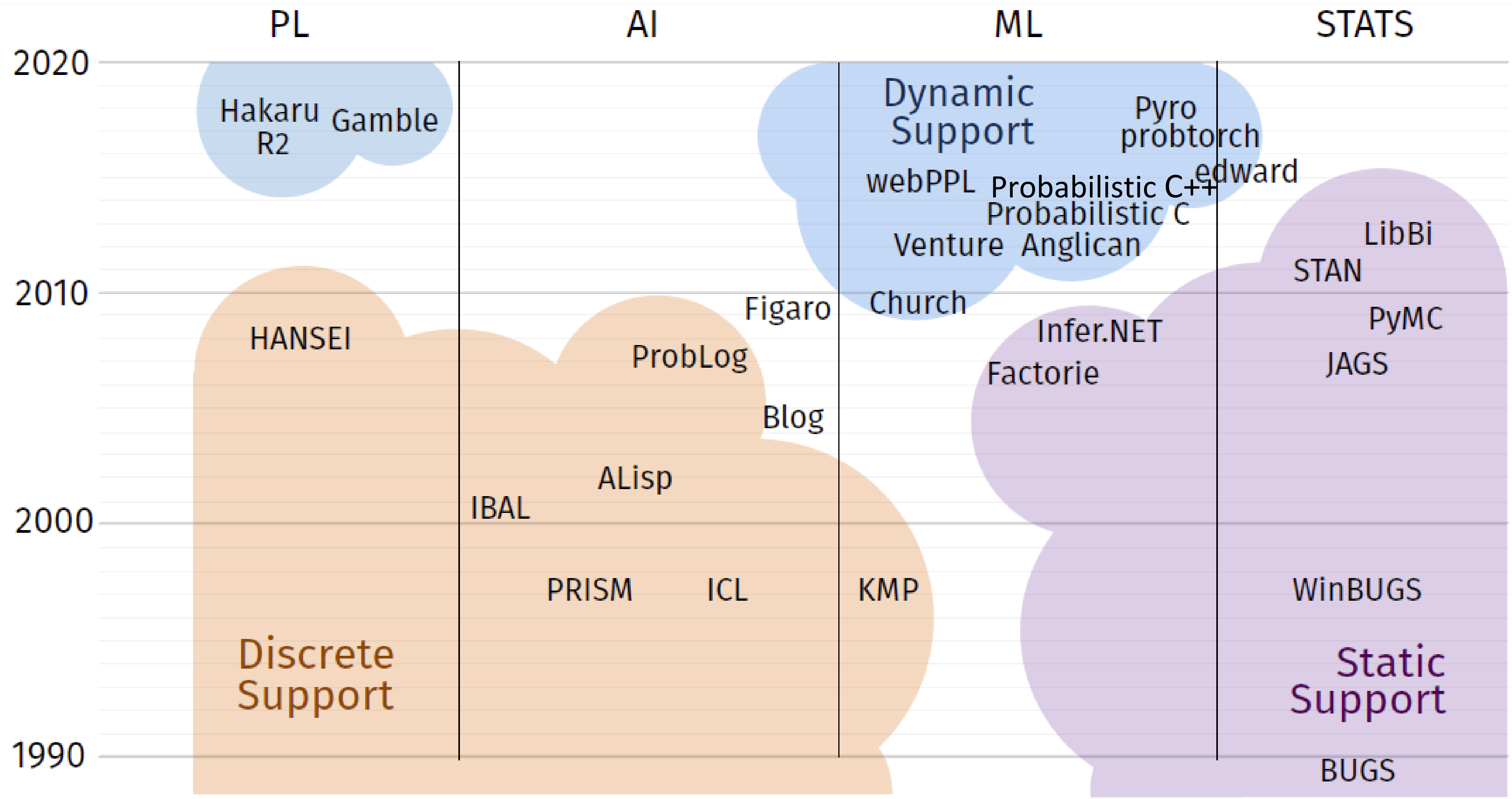

Environment and Physical Sciences
Government, Policy, Open data
Cities, Sensors & Transport
Probabilistic Inference
Mathematical Statistics
Algorithms & Complexity
Financial modelling & forecasting
Graph & Networks
Machine Learning
Defense Systems
Cryptography & Security
Foundation of Mathematical and Computational Science
Health & Biosciences
Numerical methods & Optimisation
Social Science
Speech text & Image Processing
Data analytics platform
Distributed Computing
New mathematical Approaches
Media & Culture
Optimisation & Inverse Problems
Secure Distributed transaction
Retail & Online services
Ethics, security, and privacy

Probabilistic Model

- Probabilistic models incorporate random variables and probability distributions into the model
  - Deterministic model gives a single possible outcome
  - Probabilistic model gives a probability distribution
- Used for various probabilistic logic inference (e.g. MCMC-based inference, Bayesian inference…)

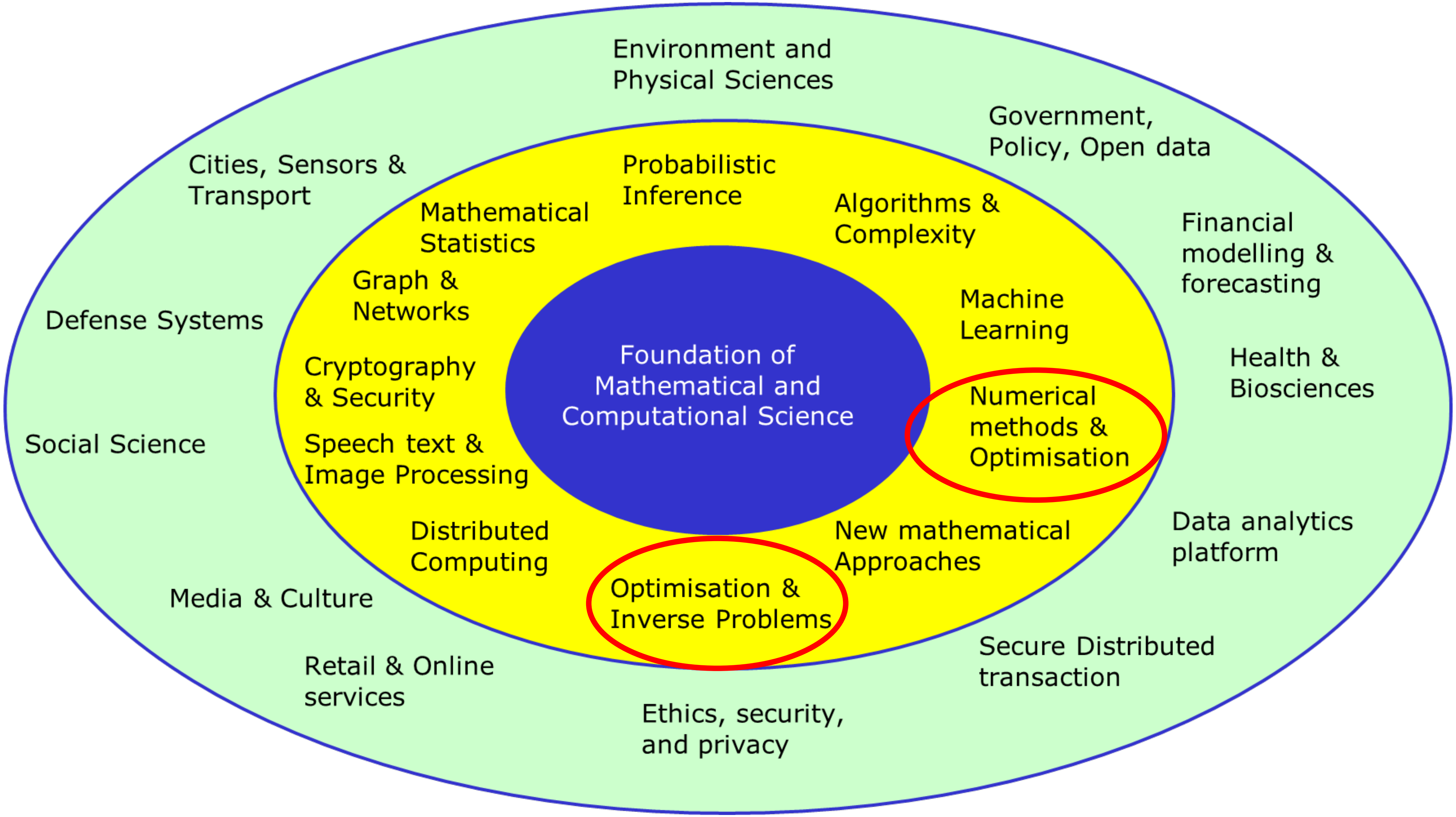
Tutorial:
https://www.cl.cam.ac.uk/~ey204/teaching/ACS/R244_2022_2023/guestlecture/slides/S6/probprog-cambridge-2021.pdf

Probabilistic Programming



B. Paige

My Pick of Buzzword 2: Optimisation


Environment and Physical Sciences
Government, Policy, Open data
Cities, Sensors & Transport
Financial modelling & forecasting
Defense Systems
Health & Biosciences
Social Science
Data analytics platform
Media & Culture
Secure Distributed transaction
Retail & Online services
Ethics, security, and privacy
Probabilistic Inference
Mathematical Statistics
Algorithms & Complexity
Graph & Networks
Machine Learning
Cryptography & Security
Numerical methods & Optimisation
Speech text & Image Processing
Distributed Computing
New mathematical Approaches
Optimisation & Inverse Problems
Foundation of Mathematical and Computational Science

Machine Learning and Optimisation

- Function Optimisation
  - Find the set of inputs to a target objective function that result in the minimum or maximum of the function
- Function Approximation:
  - Generalise from specific examples to a reusable mapping function for making predictions on new examples
  - ML can be described as function approximation as approximating the unknown underlying function that maps examples of inputs to outputs in order to make predictions on new data
  - Function approximation often uses function optimisation

- At the core of many ML algorithms is an optimisation algorithm!

Optimisation in Machine Learning

- Hyperparameter Tuning as Optimisation
  - Machine learning algorithms have hyperparameters that can be configured to tailor the algorithm to a specific dataset
  - Function inputs are algorithm hyperparameters, optimisation problems that require an iterative global search algorithm

- Model Selection as Optimisation
  - Function inputs are data transform, machine learning algorithm, and algorithm hyperparameters; optimisation problem that requires an iterative global search algorithm

Machine Learning and Optimisation

- Process of working through a predictive modeling involves optimisation at multiple steps:
  - Choosing hyperparameters of a model
  - Choosing transforms to apply to the data prior to modeling
  - Choosing modeling pipeline to use as the final model
- Learning as Optimisation
  - A numeric quantity must be predicted in the case of a regression problem, whereas a class label must be predicted in the case of a classification problem…

Optimisation: Iterative Operation

- Common to use an iterative global search algorithm for optimisation problem
- e.g. Bayesian optimisation algorithm that is capable of simultaneously approximating the target function that is being optimised while optimising it.

- Automated machine learning (AutoML) algorithms being used to choose an algorithm, an algorithm and hyperparameters, or data preparation, algorithm and hyperparameters, with very little user intervention

Search Parameter Space

Random search: No risk of ’getting stuck’
potentially many samples required

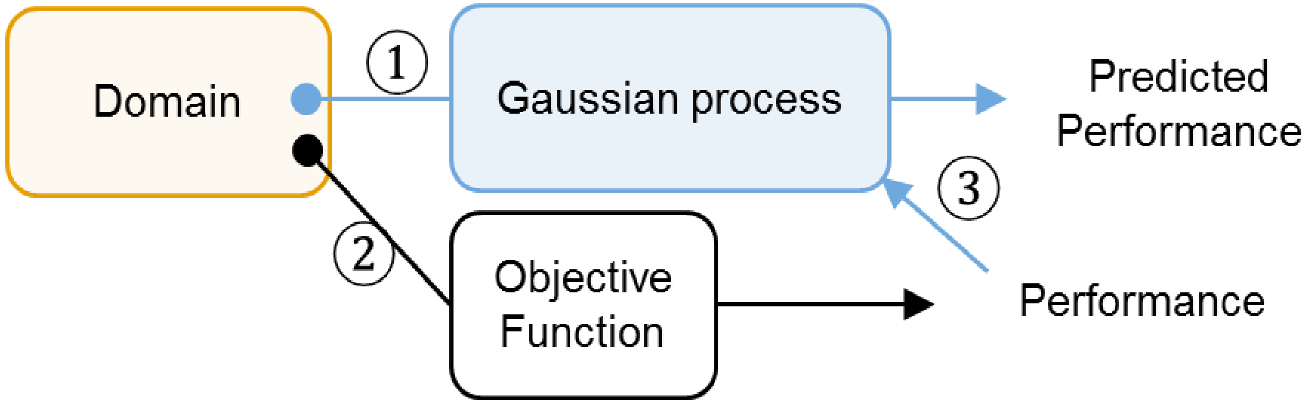
Evolution strategies: Evaluate
permutations against fitness function

Bayes Opt: Sample efficient, requires
continuous function, some configuration

Random Search	Genetic algorithm / Simulated annealing	Bayesian Optimisation
No overhead	Slight overhead	High overhead
High #evaluation	Medium-high #evaluation	Low #evaluation

Bayesian optimisation

Iteratively build a probabilistic model of objective function



① Find promising point (parameter values with high performance value in the model)

② Evaluate the objective function at that point

③ Update the model to reflect this new measurement

Pros:

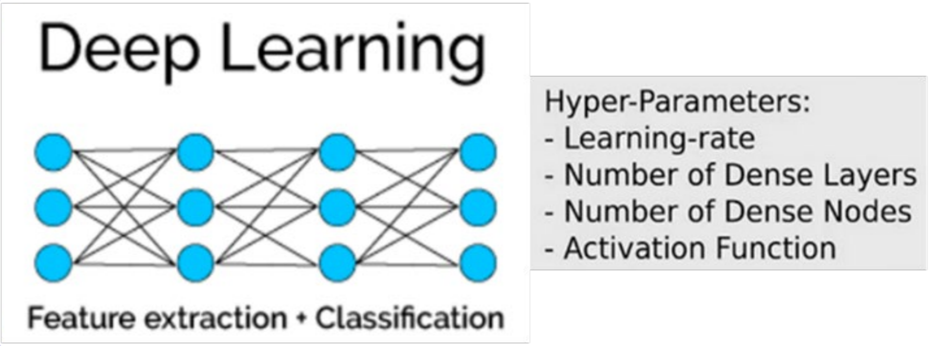
- ✓ Data efficient: converges in few iterations
- ✓ Able to deal with noisy observations

Cons:

- ✗ In many dimensions, model does not converge to the objective function

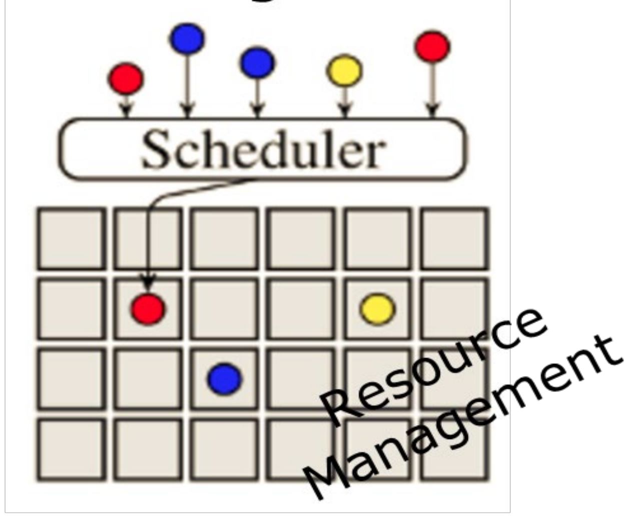
Use of ML based Optimisation Methods

- Increasing data volumes and high-dimension parameter space
- Expensive Objective Functions
- Hand-crafted solutions impractical, often left static or configured through extensive offline analysis

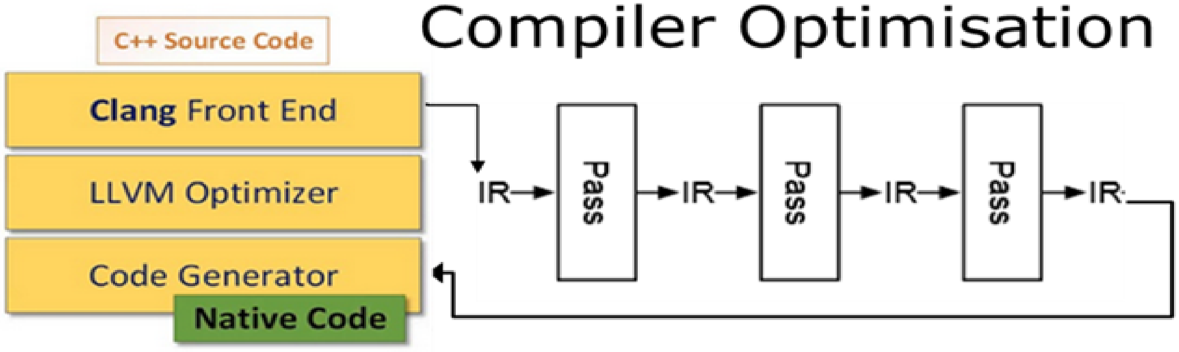


Hyper-Parameters:
- Learning-rate
- Number of Dense Layers
- Number of Dense Nodes
- Activation Function

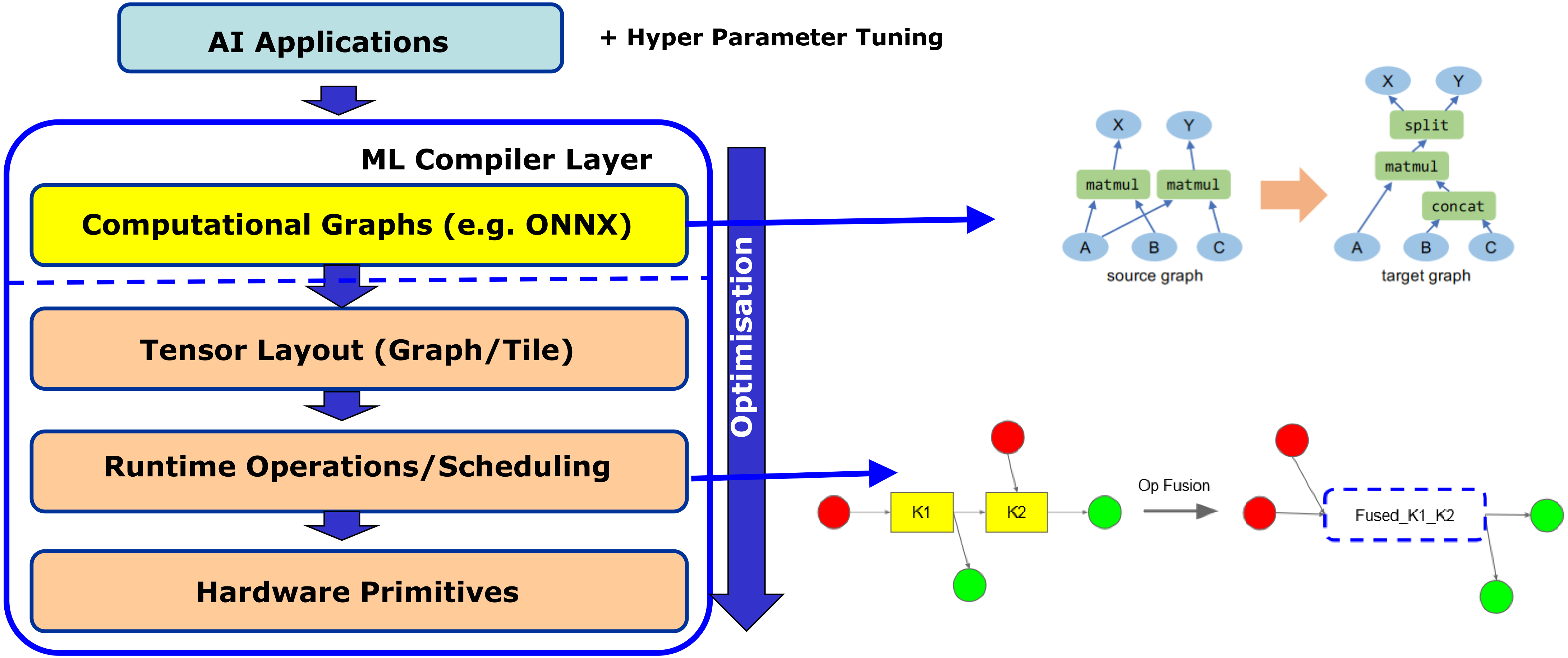
Cluster Workload Management



Compiler Optimisation



ML Compiler Optimisation



AI Applications

+ Hyper Parameter Tuning

ML Compiler Layer

Computational Graphs (e.g. ONNX)

Tensor Layout (Graph/Tile)

Runtime Operations/Scheduling

Hardware Primitives

Optimisation

TensorFlow, MXNet, and PyTorch

Reinforcement Learning for Optimisation

Many problems in systems are sequential decision making and/or combinatorial problems

- Compiler Optimisation
- Chip placement
- Datacentre resource allocation
- Network congestion control with multiple connections
- Wide range of signals to make decisions (e.g., VM allocation)
- Database: Query optimiser, Dynamic indexing…

RL in Computer Systems

What makes RL different from other ML paradigms?

- There is no supervisor, only a reward signal
- Feedback is delayed, not instantaneous
- Time really matters (sequential)
- Agent’s actions affect the subsequent data it receives

Practical Consideration:

- Action spaces do not scale (Systems problems often combinatorial
- Exploration in production system not a good idea
- Simulations can oversimplify problem (Expensive to build)
- Online steps take too long

A brief history of RL Tools

Gen (2014-16): Loose research scripts (e.g. DQN), high expertise required, only specific simulators

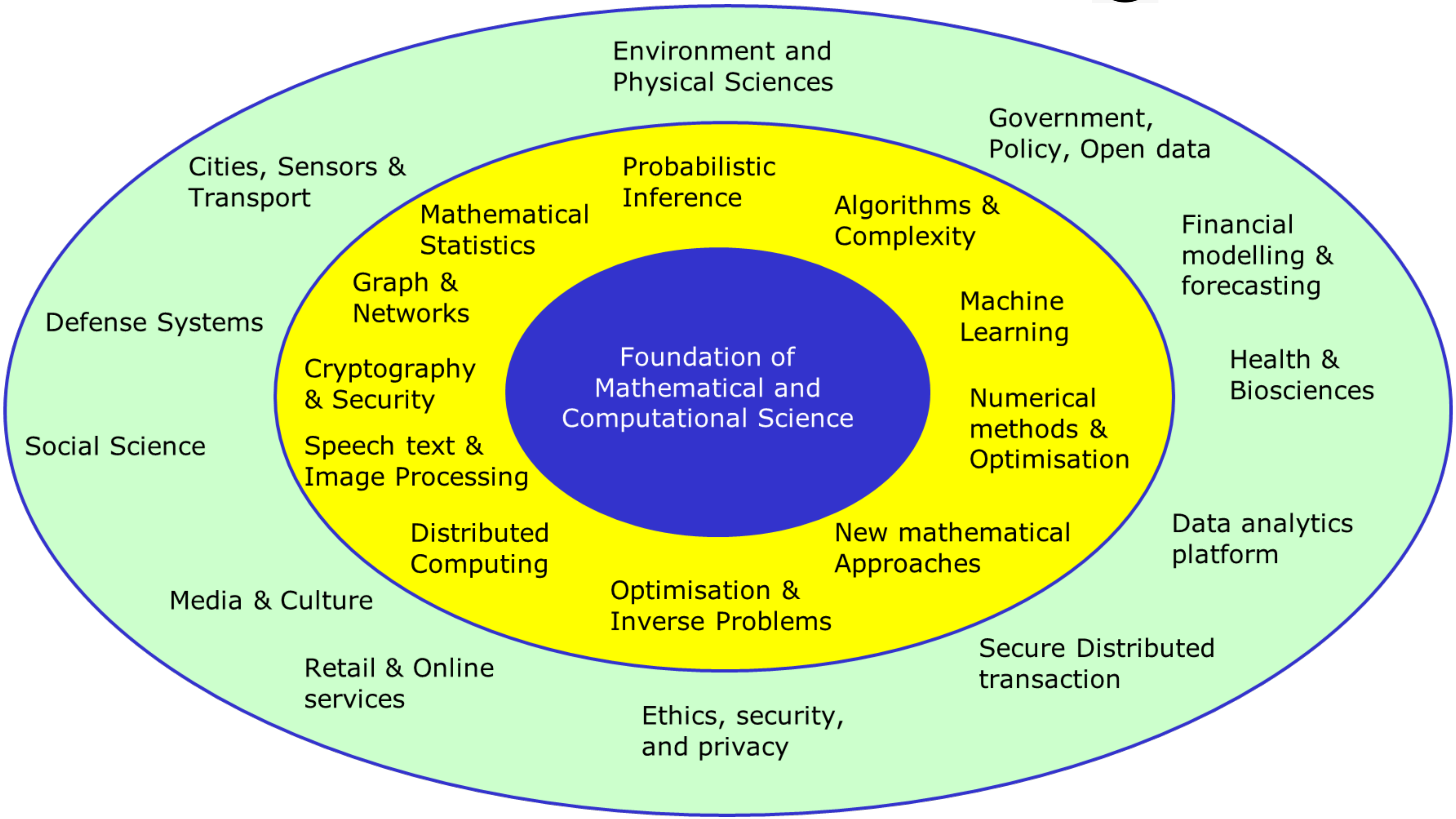
Gen (2016-17): OpenAI gym gives unified task interface, reference implementations

- Good results on some environments (e.g. game), difficult to retool to new domains and execution modes
- Abstractions/Libraries: not fully reusable, customised towards game simulators
- High implementation risk: lack of systematic testing, performance strongly impacted by noisy heuristics

Gen (2017-18): Generic declarative APIs, distributed abstractions (e.g.Ray Rllib, RLGraph), some standard *flavours* emerge

Still Problems... Tightly coupled execution/logic, testing, reuse...

Curve your Buzzwords – It's Open 😀



Foundation of Mathematical and Computational Science

Probabilistic Inference

Algorithms & Complexity

Machine Learning

Numerical methods & Optimisation

New mathematical Approaches

Optimisation & Inverse Problems

Distributed Computing

Speech text & Image Processing

Cryptography & Security

Graph & Networks

Mathematical Statistics

Environment and Physical Sciences

Government, Policy, Open data

Financial modelling & forecasting

Health & Biosciences

Data analytics platform

Secure Distributed transaction

Ethics, security, and privacy

Retail & Online services

Media & Culture

Social Science

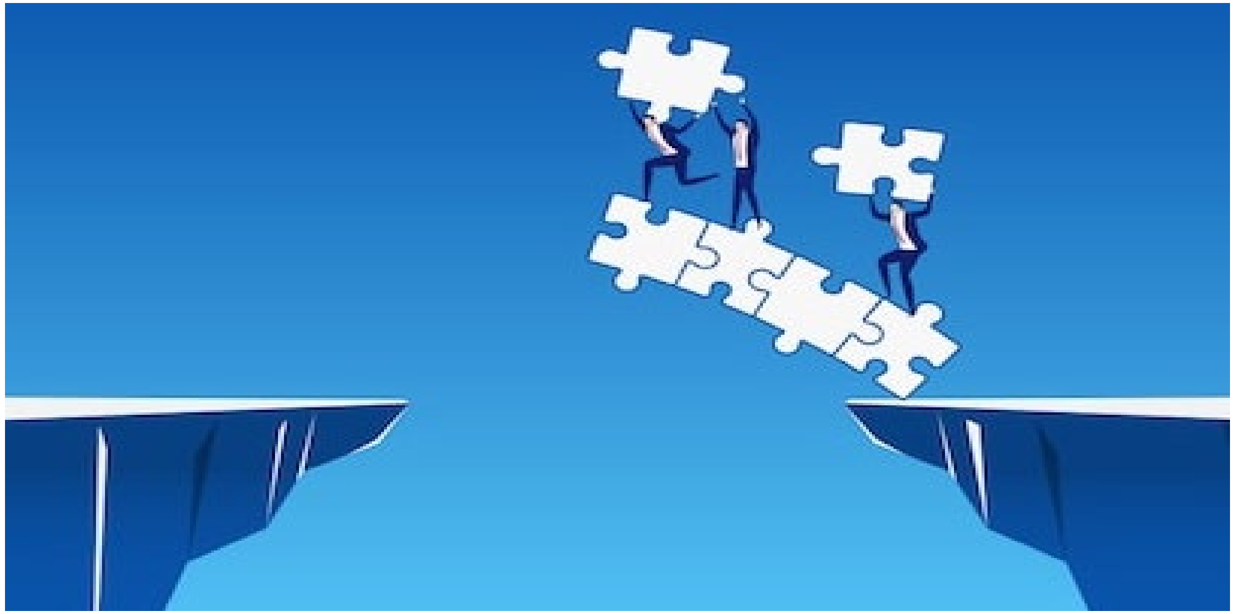
Defense Systems

Cities, Sensors & Transport

Pick your own Buzzwords and Dive in Research!

Data Science is broad: Pick a right topic and its scope

→ What do YOU want to RESEARCH in Data Science?



Email: eiko.Yoneki@cl.cam.ac.uk