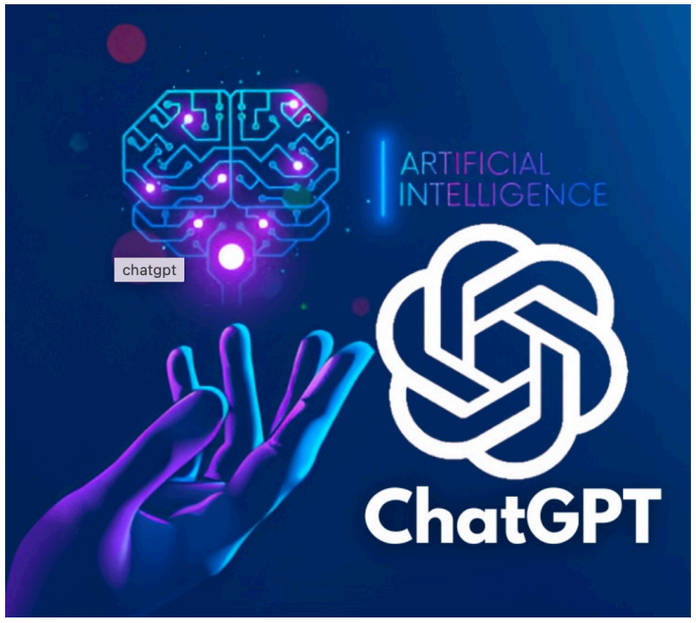
Day 1: Overview and Introduction to Data Science

ME314: Introduction to Data Science and Machine Learning
LSE Summer School
8 July 2024

Emerging trends

Technologies



Computing paradigm shifts

From calculation to delegation to personalisation

The Generative AI Revolution

Unsung hero: open-source

Voila!

Dot-product Attention

Attention Is All You Need

Stable Diffusion

midjourney

ChatGPT

Seq2Seq for NLP
(2014)

IMAGENET

Convinced by DL

TensorFlow

PyTorch

Software advances

GPT-1
(2018)

Microsoft-OpenAI
partnership

Scaling up

Generative AI
"boom" for public

2012

RNN Attention
(2013)

2015

2017

2019

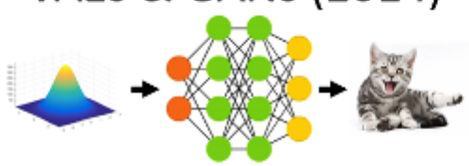
2022

2023

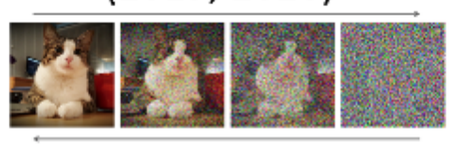
Network architecture evolution, hardware (GPU) advancements

NVIDIA

VAEs & GANs (2014)



Diffusion Models
(2019, 2020)



Concept of Data Science

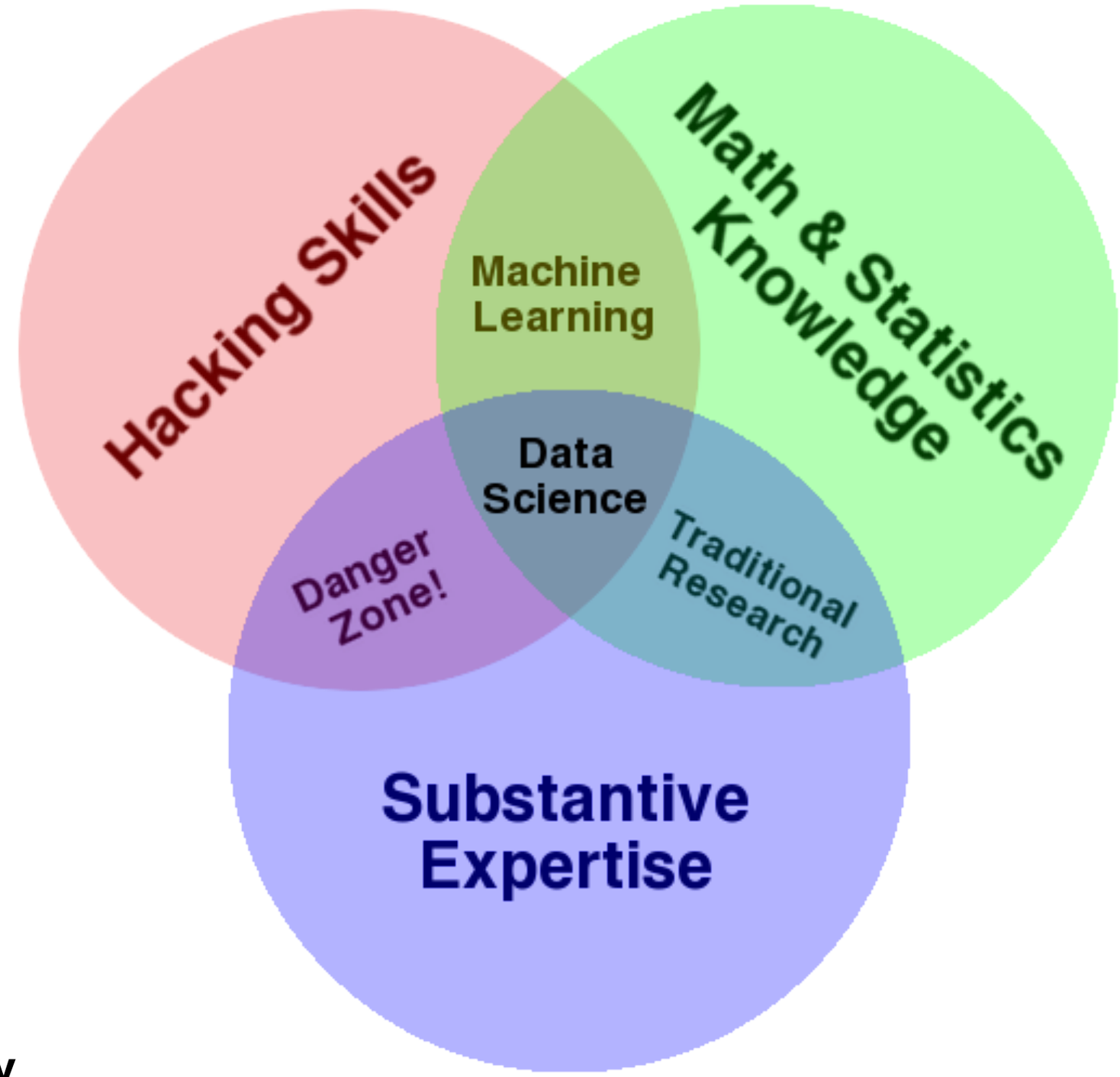
Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

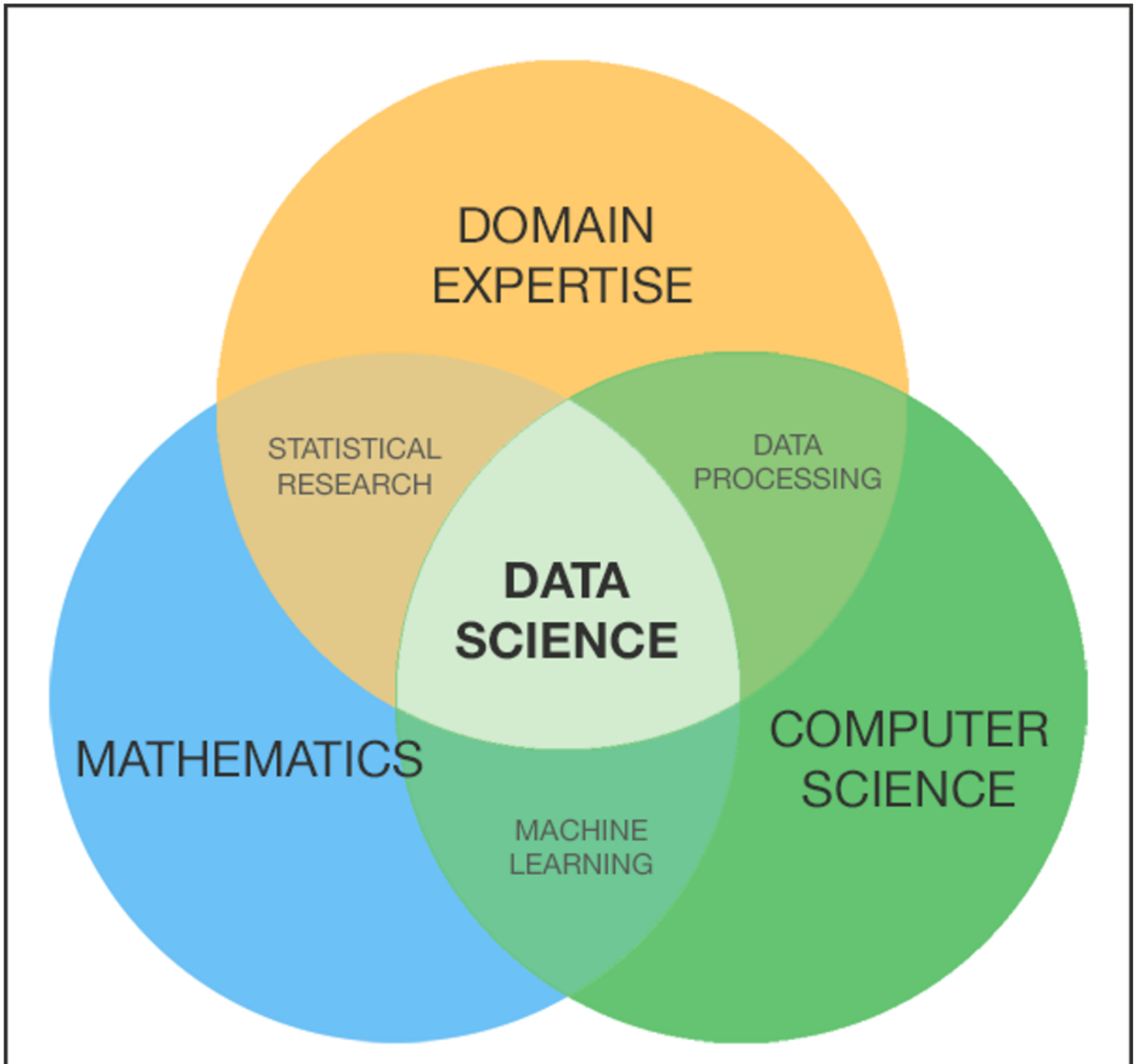
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

``I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?''
- Hal Varian (Chief Economist at Google, 2009).

What is Data Science?



Drew Conway



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

LOOKING BACKWARD AND FORWARD

FIRST THERE WAS BUSINESS INTELLIGENCE	NOW WE'VE ADDED DATA SCIENCE
Deductive Reasoning	Inductive and Deductive Reasoning
Backward Looking	Forward Looking
Slice and Dice Data	Interact with Data
Warehoused and Siloed Data	Distributed, Real Time Data
Analyze the Past, Guess the Future	Predict and Advise
Creates Reports	Creates Data Products
Analytic Output	Answer Questions and Create New Ones
	Actionable Answer

Source: Booz Allen Hamilton

Inductive and deductive reasoning

- Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning

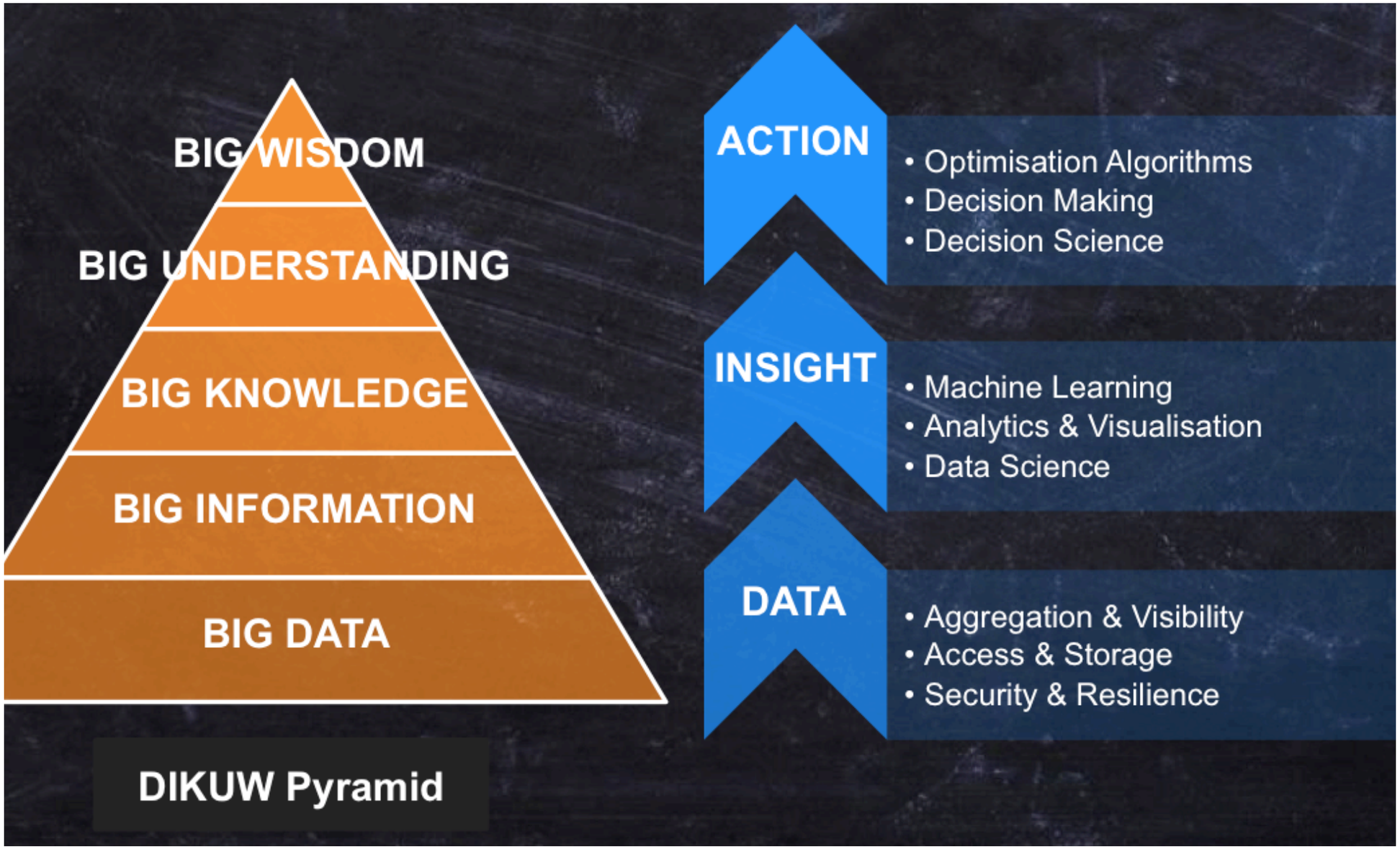
- This is a fundamental change from traditional analysis approaches.

- Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths.

- Models of reality no longer need to be static.

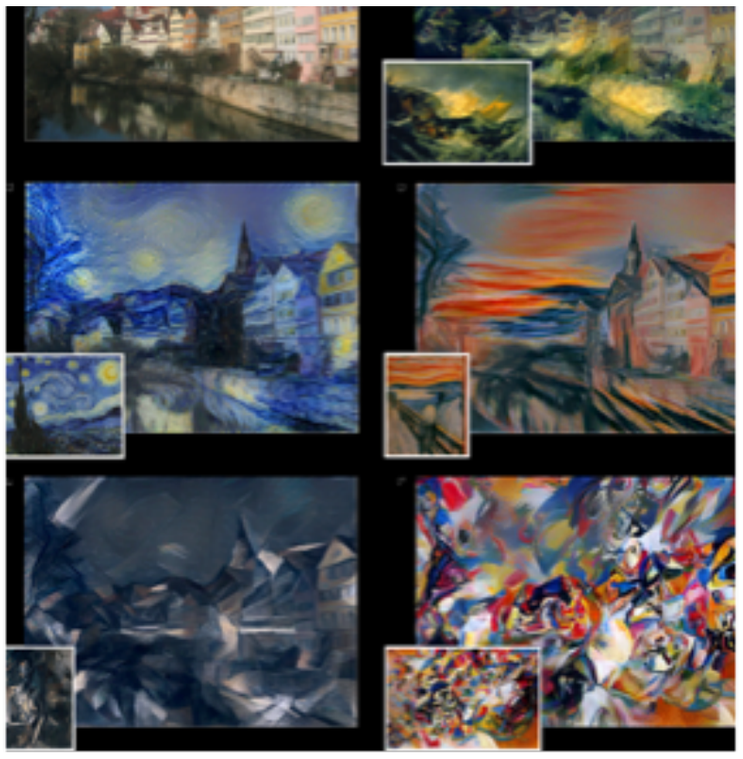
- They are constantly tested, updated and improved until better models are found.

From data to wisdom



Data Science principles

- Be willing to fail.
- Fail often and learn quickly.
- Keep the goal in mind.
- Dedication and focus lead to success.



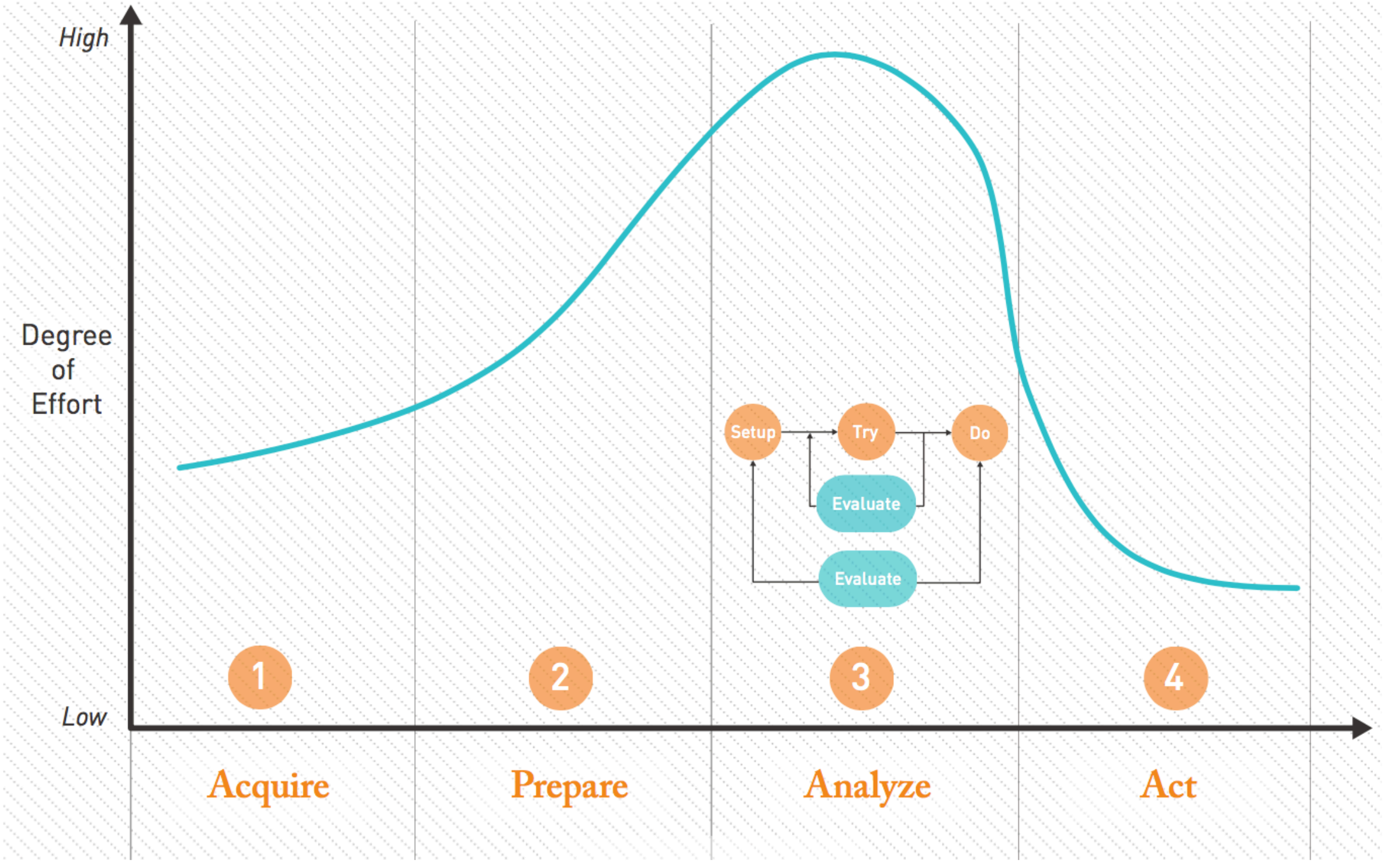
- Leon A. Gatys, Alexander S. Ecker, Matthias Bethge. ``A Neural Algorithm of Artistic Style.'' arXiv:1508.06576. September 2015.
- Prisma and Convolutional Neural Networks: June 2016.

What makes data science different

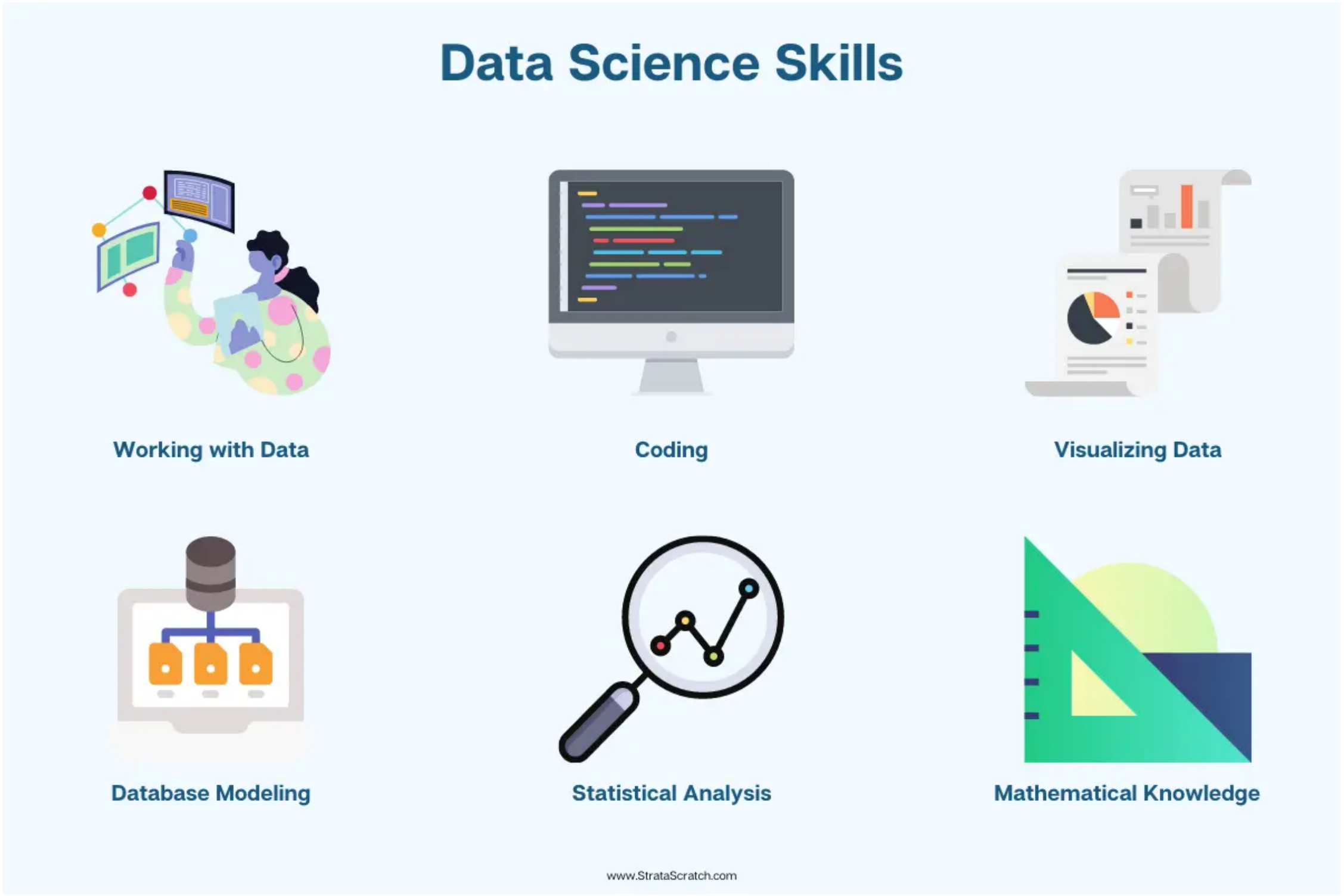
- Not just maths: programming is (just as, or even more) important
  - Common data science languages: Python, R
  - Development languages for numerical computing: C++
- Not just programming: databases and APIs also needed
  - APIs: Application programmer interface, how machines communicate data, especially across networks
  - Databases store and organize data, central to big data analysis
- Not just a by-product
  - Data is no longer a by-product of enterprise activities, but often its primary commodity
  - Requires data by design, not just incidentally
- Data is distinguished by volume, velocity, and variety (big data)

Practice of Data Science

Data science workflow



Different skill sets in the field of data science



Different skill sets in the field of data science

- “Data Scientist”
- Data Analyst
- Data Engineers
- Database Administrator
- Machine Learning Engineer
- Data Architect
- Statistician
- Business Analyst
- Data and Analytics Manager

Source:https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/

Four Categories of Essentials “Skills” for a Data Scientist

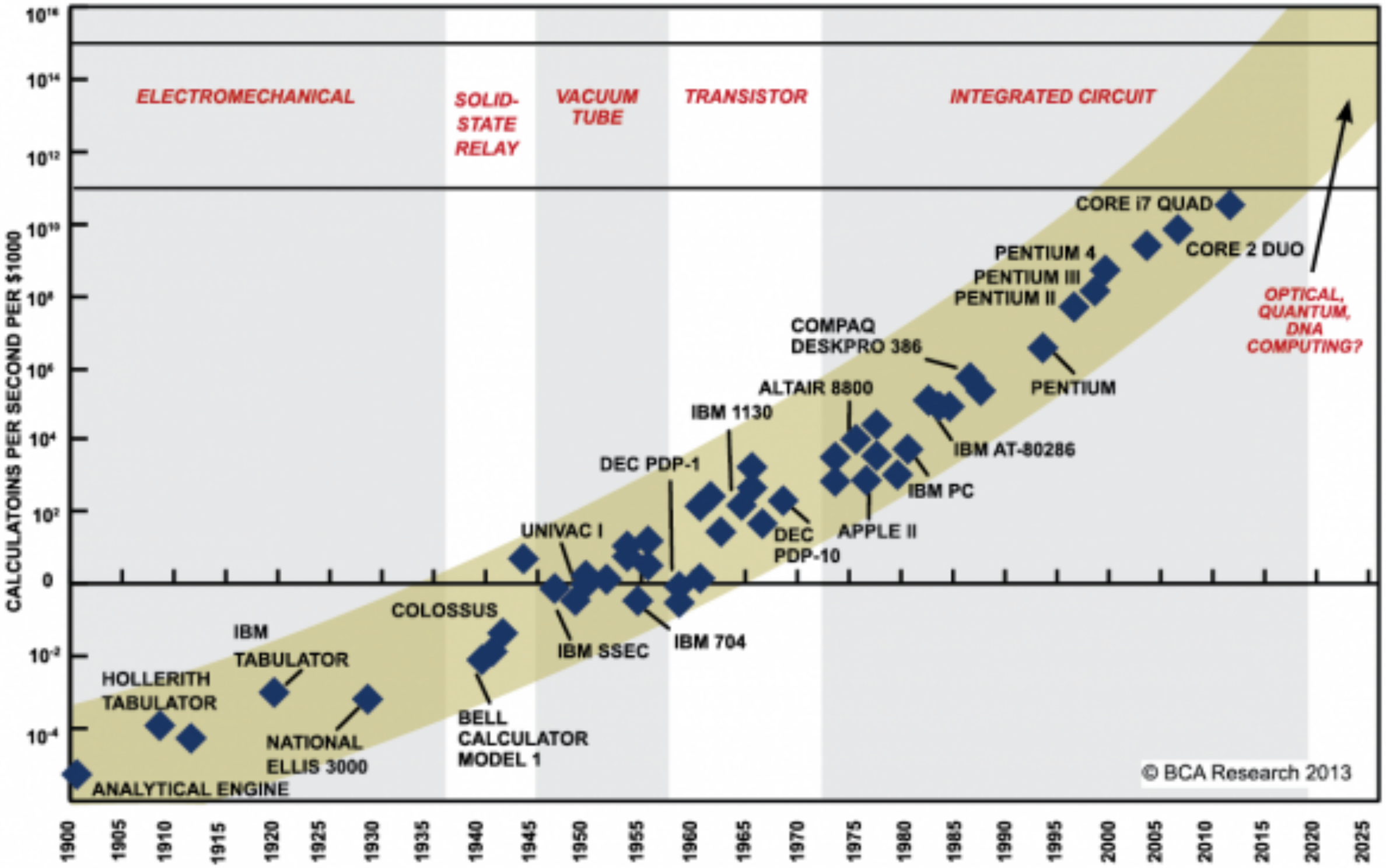
- **Tools**: The instruments you need to perform data science
- **Skills**: The ability to use the tools
- **Knowledge**: The understanding of how and when to apply the skills using the tools, and interpreting the results
- **Ethics**: Understanding how to use the above responsibly, legally, and morally for the good of society and avoiding harm

Source:https://www.mygreatlearning.com/blog/different-data-science-jobs-roles-industry/

Data Science and AI

The (Third) Coming of AI

- Birth
- Early years & realisations
- Expert Systems
- AI Winter
- The (big) come back: artificial neural networks



SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.

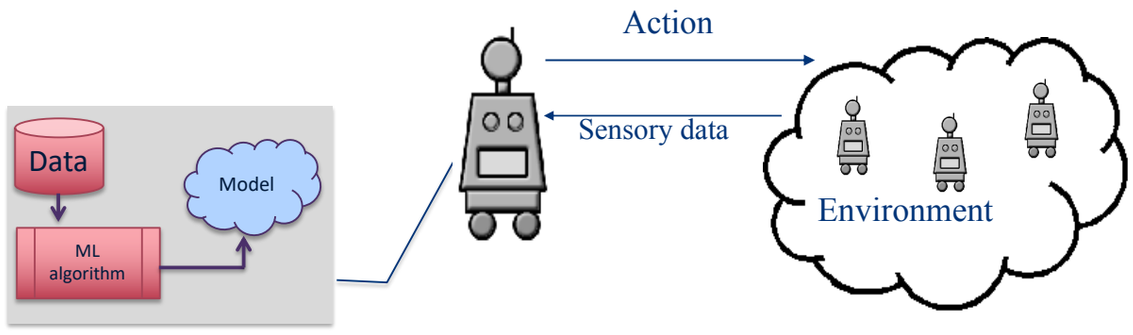
AI-Generated Cats

Source: https://www.tidio.com/blog/ai-test/

ML, statistical learning, and AI

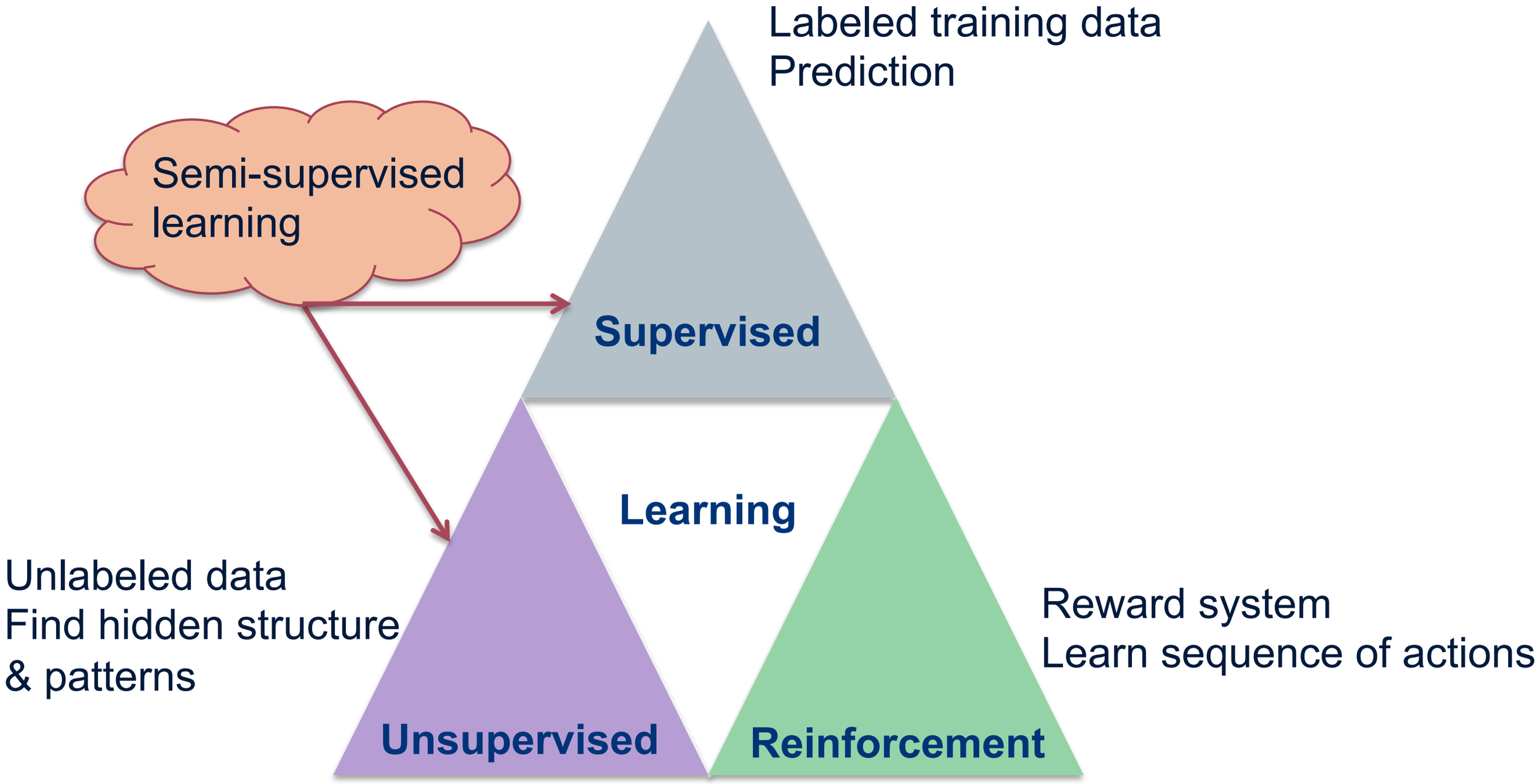
- Artificial Intelligence (AI)
  - The broad concept of machines being able to carry out tasks in a way that we would consider “smart”
  - Broad field that has changed a great deal since its inception
- Machine learning (ML)
  - A subset of AI in which machines learn specific applications from data for specific purposes
  - Statistical learning: a set of methods from the field statistics adapted to ML
- “Deep learning”
  - Special application of ML using “deep” artificial neural networks (or deep reinforcement learning)

An AI Perspective



- Truly intelligent systems need to adapt their behaviour
- Learning: the process of acquiring knowledge, skills, or attitudes through experience, imitation, or teaching, which then causes changes in behaviour
- Hence Machine Learning

Types of Learning



Why is Learning Important

- Impractical/impossible to specify systems correctly and completely at the time of design/implementation

- Implemented systems may not work as well as desired or expected when put in operation

- Knowledge about certain tasks may simply be too large to be explicitly encoded by humans

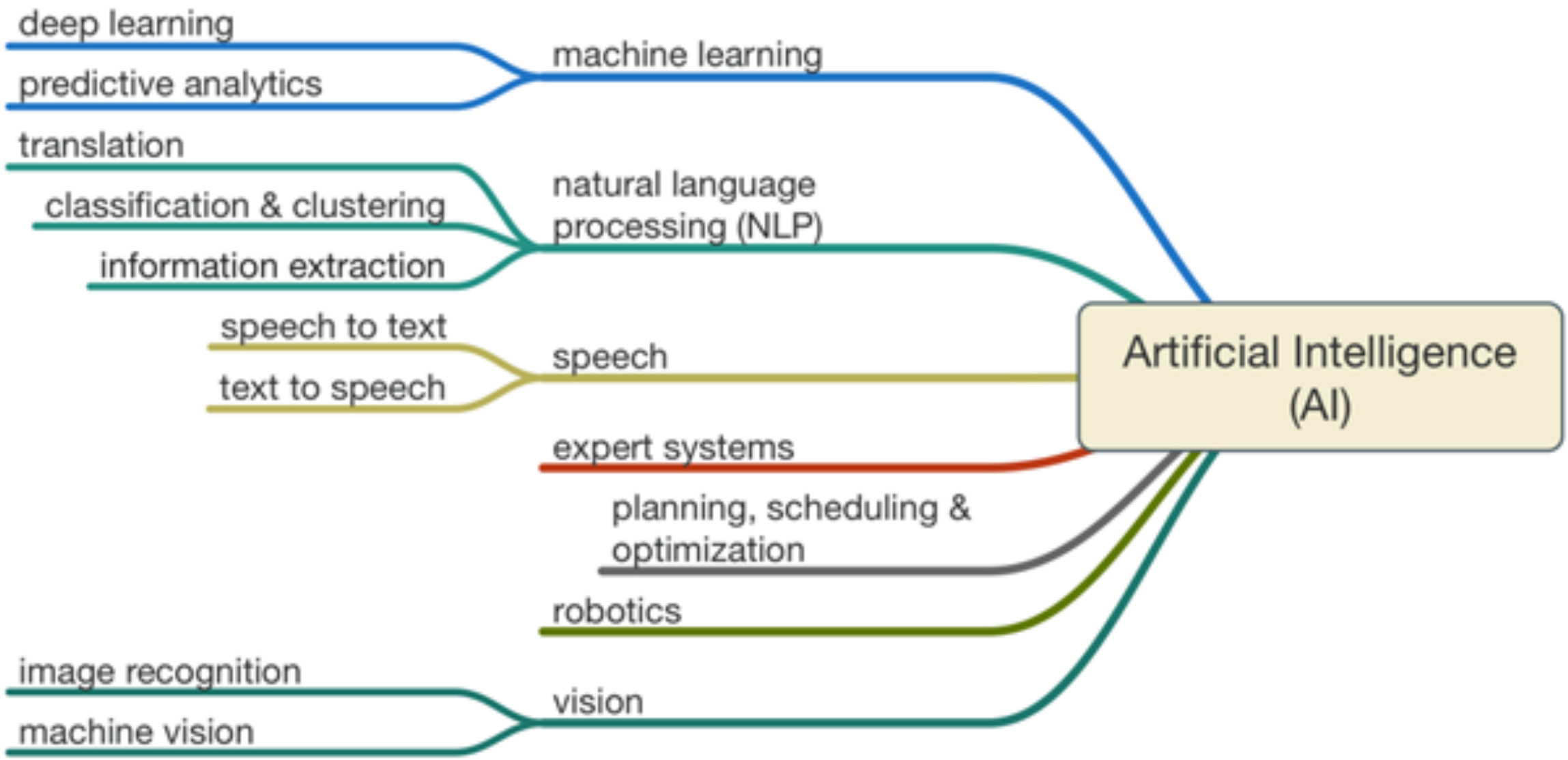
- The environment may change and hence the system’s goals need to be changed as well

- Hidden relationships and correlations among huge amounts of data

Why has ML become popular?

- Data explosion – Big Data!
    - Structured, unstructured, social media, labelled, unlabelled
    - Cost effective storage
- Computational power
- Faster processors, GPUs
- HPC, cloud computing, computing as a service
- Advances in algorithms and availability of toolkits

Main approaches in AI

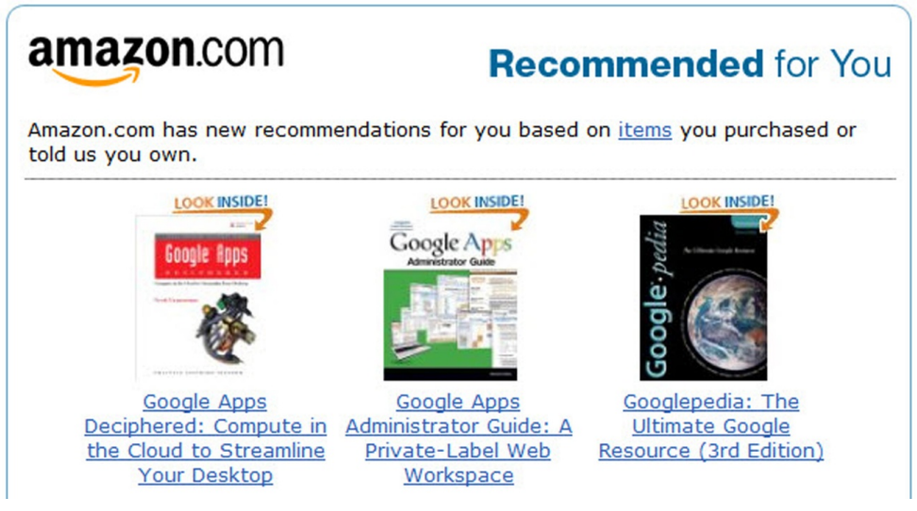


Neota Logic

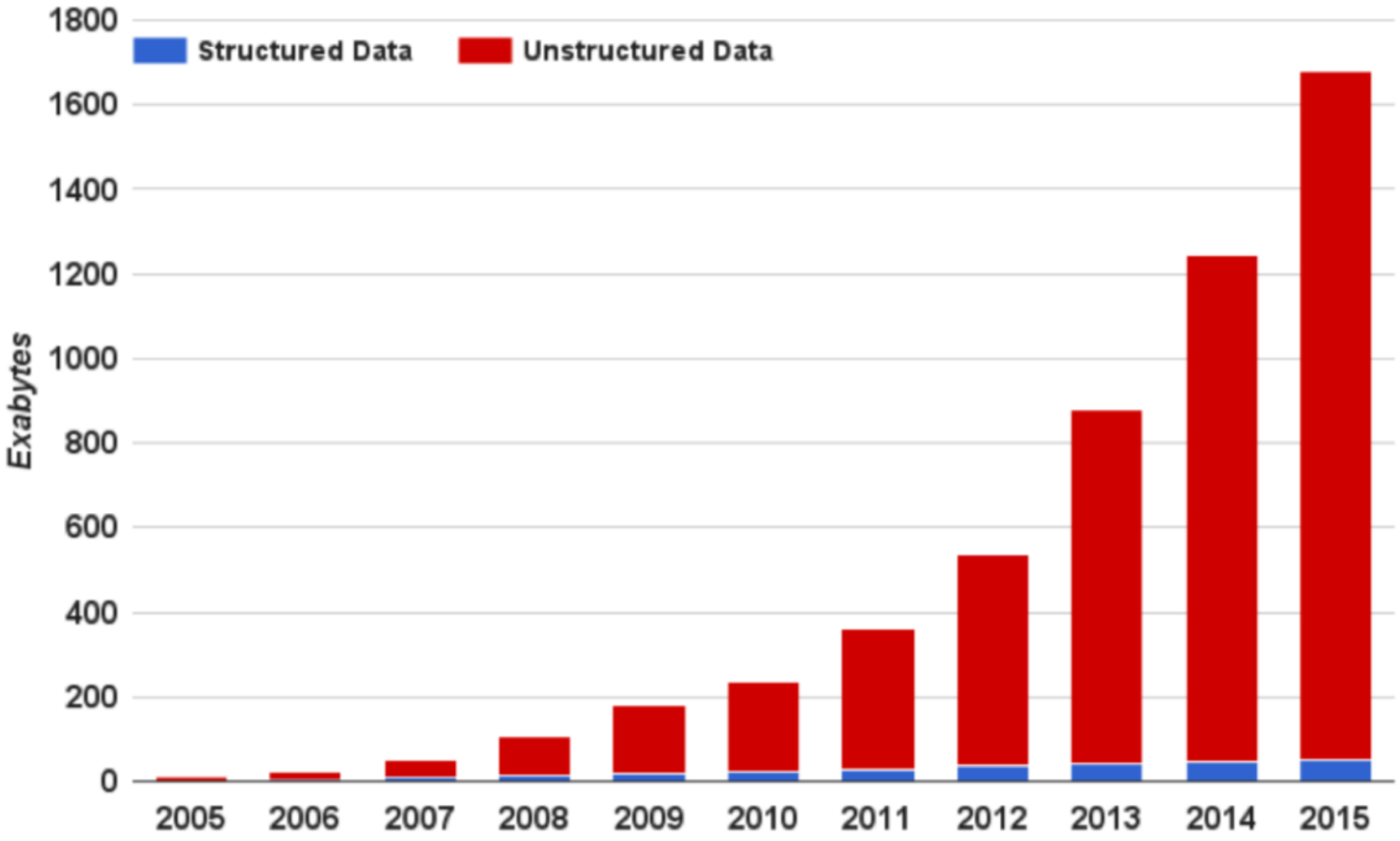
Data Science in the Wild

Personalisation

- What articles should be shown on the homepage of an online newspaper?
- What titles and images would attract the most clicks?
- Which product order would yield the highest profit?
- What is the best combination of drugs for patient?



Unstructured Data



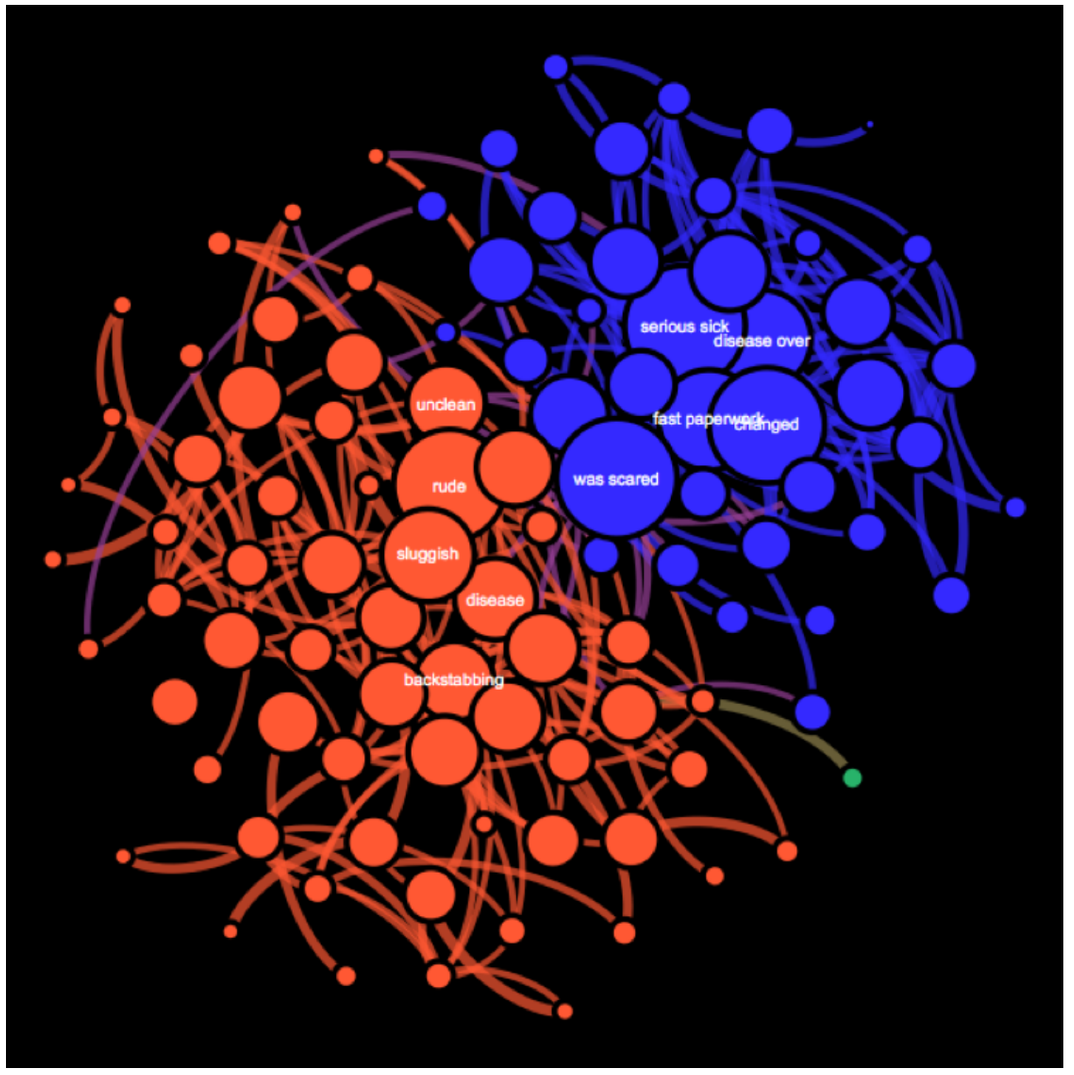
A.Nadkarni, N.Yezhkova, “Structured versus unstructured data: The balance of power continues to shift.” IDC (Industry Development and Models), March 2014.

Understanding Patients

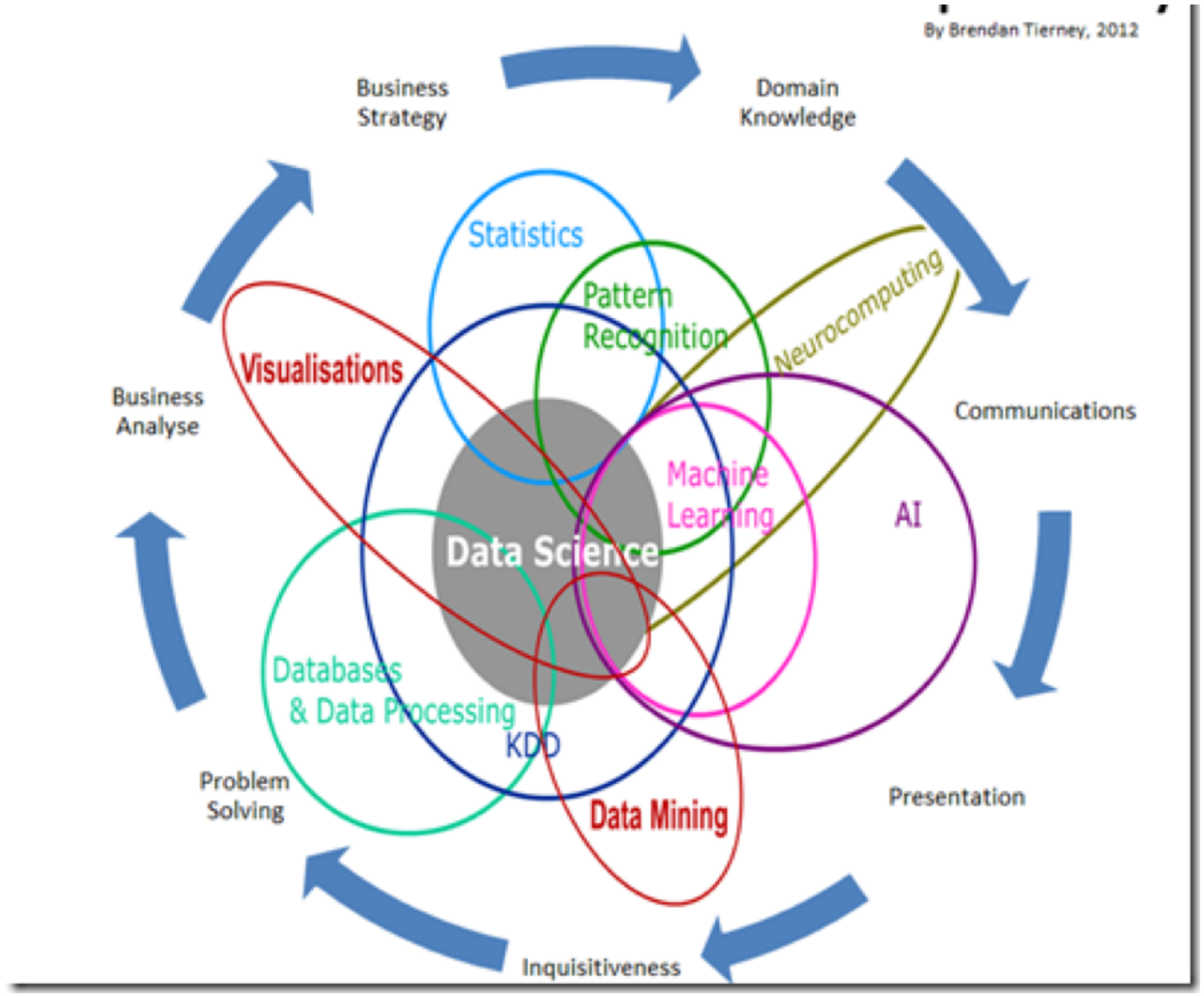
- Online reviews of primary care services (GPs) in England

- July 2013 - January 2017, 7.7K GP practices, 145K reviews

- ~ 3-5K reviews per month, 5-6 sentences long



Ideal Data Scientist



http://www.kdnuggets.com/2014/06/data-science-skills-business-problems.html

Image from http://www.mysticwish.co.uk/product/anne-stokes-forest-unicorn-fridge-magnet/

Benefits for graduates

glassdoor

data scientist | London, UK

Sign In

Data Scientist Salaries London, UK

For Employers | Post Jobs

Overview | **Salaries** | Interviews | Insights | Career Path

Data Scientist Salaries in London

Updated 7 Jul 2023

Very High Confidence

£62,312 /yr

Average Base Pay

3,071 salaries

Additional Cash Compensation

Average: **£6,616** Range: £1,658 - £26,401

The average salary for Data Scientist is £62,312 per year in the London. The average additional cash compensation for a Data Scientist in the London is £6,616, with a range from £3,193 - £13,711. Salaries estimates are based on 3071 salaries submitted anonymously to Glassdoor by Data Scientist employees in London.

What is the salary trajectory for a Data Scientist? in London

£62,312 /yr
Data Scientist

£83,184 /yr
Senior Data Scientist

£94,496 /yr
Lead Data Scientist

See Full Career Path

Download as data table

By the time we are finished…

Remarks

- We now have a reasonable machine learning armoury to draw from
- You can (semi-)automate the machine learning pipeline – business- and problem-dependent
- Lots of toolkits and software
- Generative AI is advancing **very** rapidly
- Hunting for unicorns…
- Explainability and ethical considerations
- Collaborate!