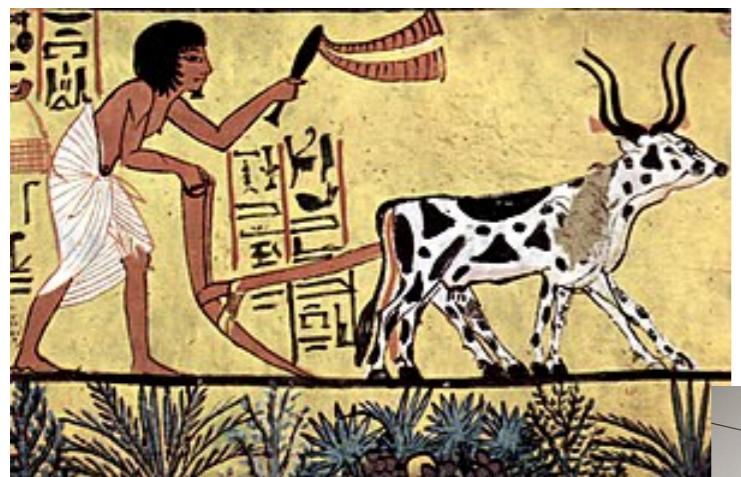


Day 1: Overview and Introduction to Data Science

**ME314: Introduction to Data Science and Machine Learning
LSE Methods Summer Programme
29 July 2019**

Emerging trends

Technologies



Computing paradigm shifts

From calculation to delegation to personalisation



Big changes...



St Peter's Square 2005

In too little a time...



St Peter's Square 2013

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

W

hen Jonathan Goldstein 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 making new connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. An exec 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."

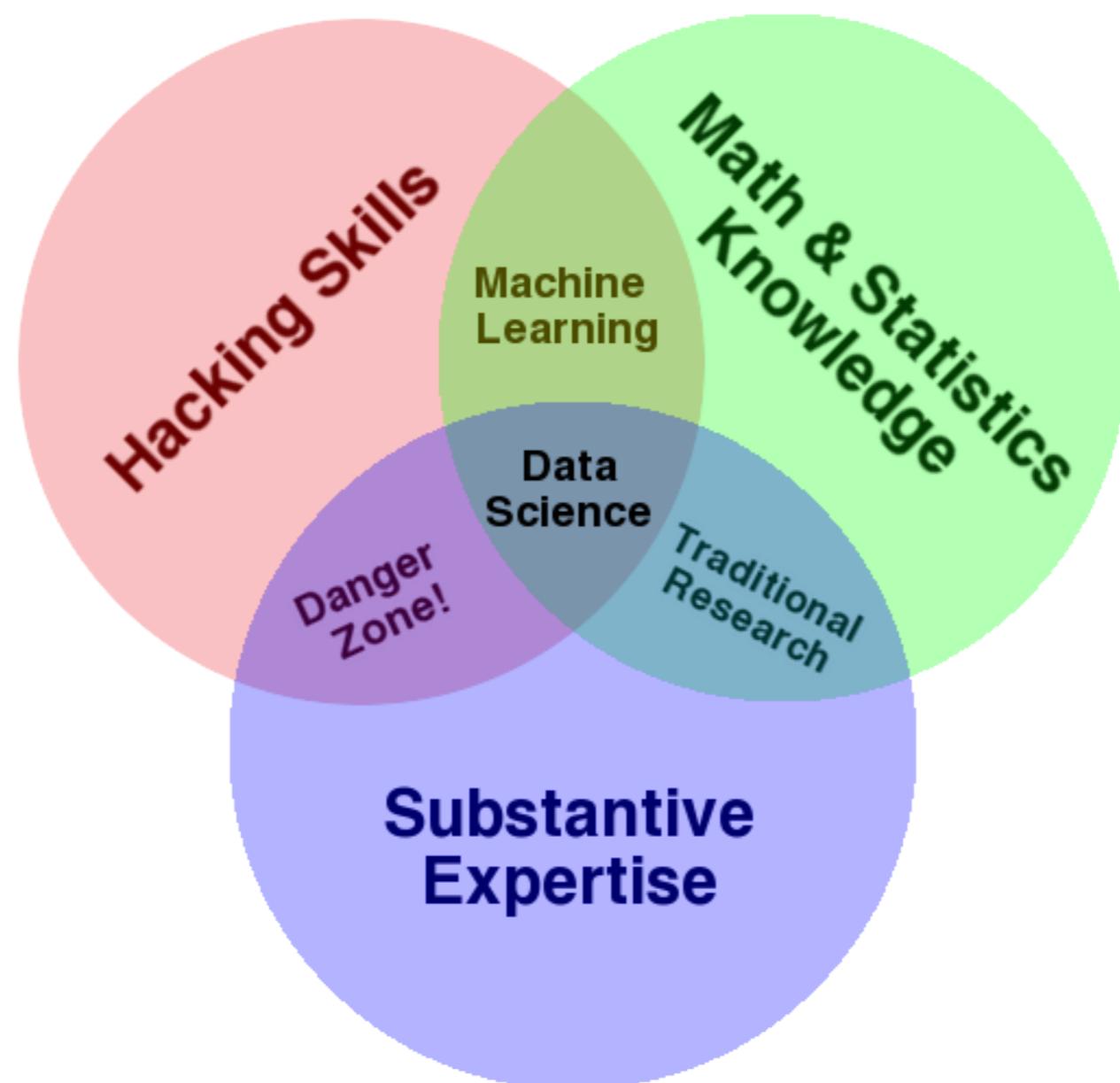


106 Harvard Business Review October 2010

“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

LOOKING BACKWARD AND FORWARD



FIRST THERE WAS BUSINESS INTELLIGENCE

Deductive Reasoning
Backward Looking
Slice and Dice Data
Warehoused and Siloed Data
Analyze the Past, Guess the Future
Creates Reports
Analytic Output

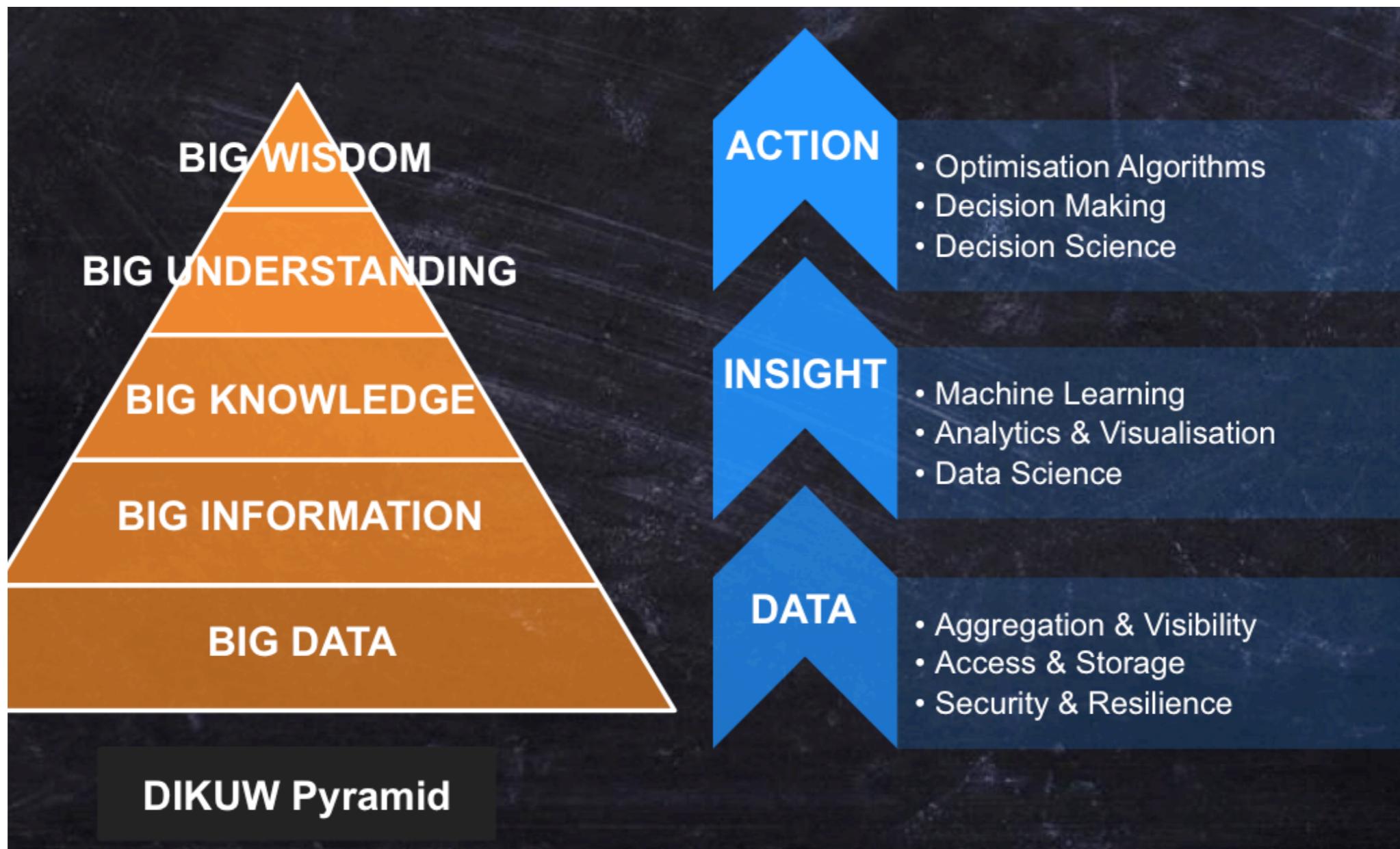
NOW WE'VE ADDED DATA SCIENCE

Inductive and Deductive Reasoning
Forward Looking
Interact with Data
Distributed, Real Time Data
Predict and Advise
Creates Data Products
Answer Questions and Create New Ones
Actionable Answer

Inductive and deductive reasoning

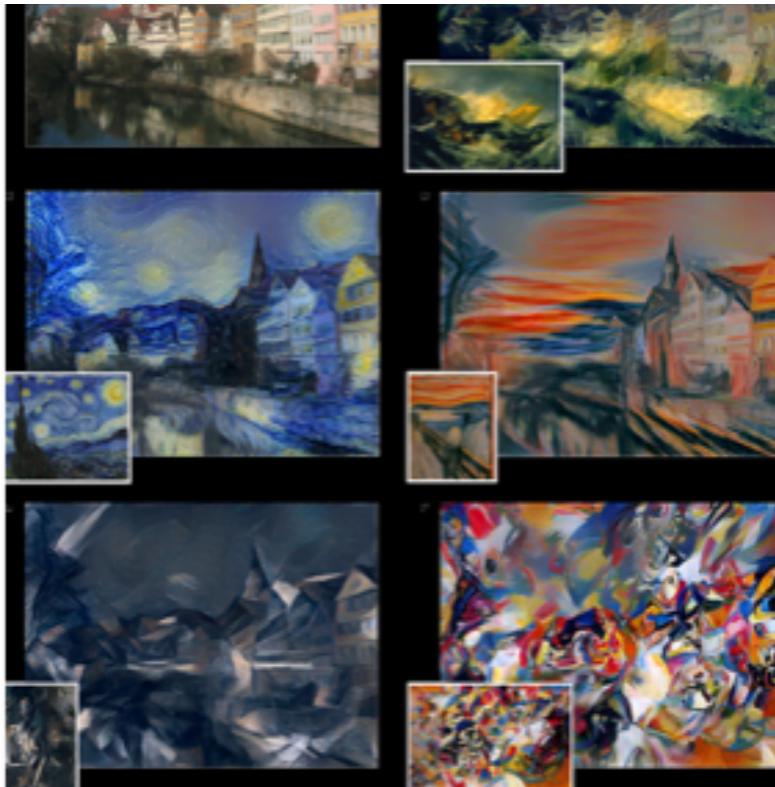
- Data Science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning
- This is a fundamental change from traditional analysis approaches.
- Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover 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.

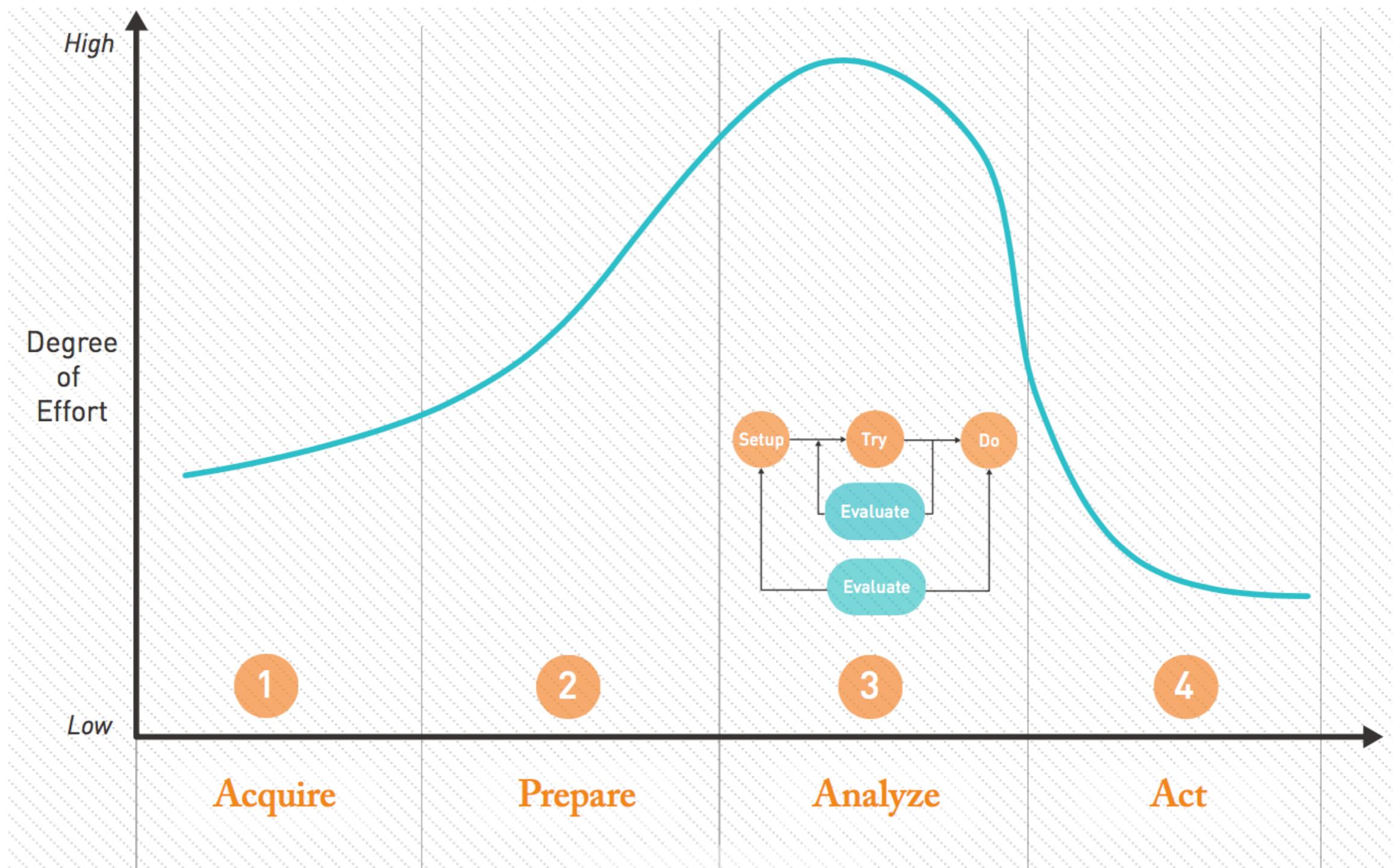


- Leon A. Gatys, Alexander S. Ecker, Matthias Bethge. "A Neural Algorithm of Artistic Style." arXiv:1508.06576. September 2015.
- Prisma and Convolutional Neural Networks: June 2016.



Practice of Data Science

Data science workflow



Data Science and AI

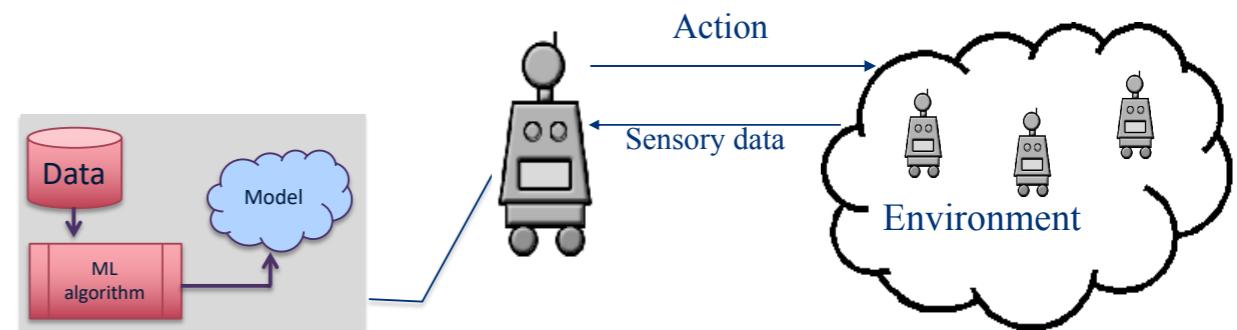
The (Third) Coming of AI

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

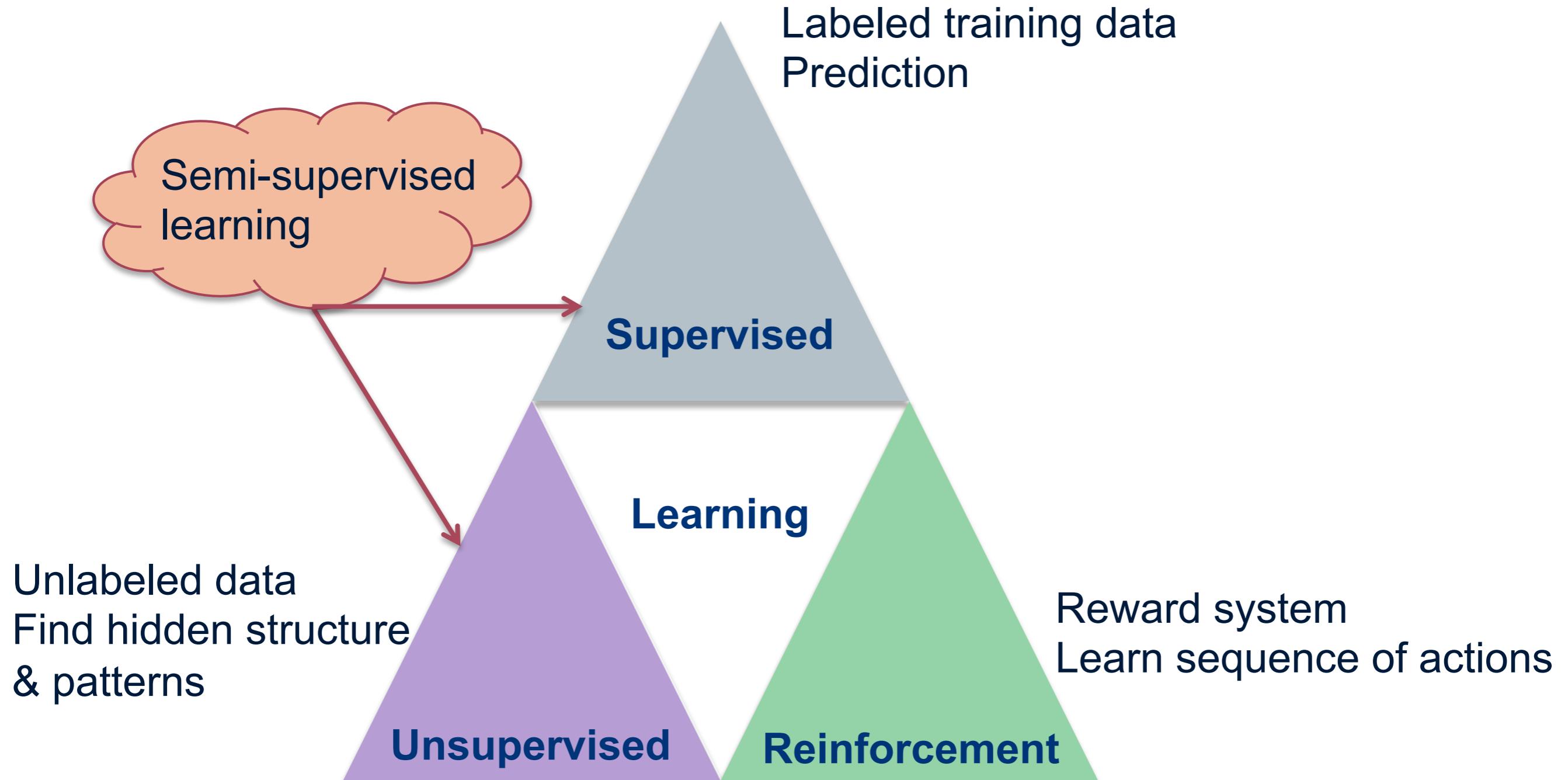


An AI Perspective

- 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



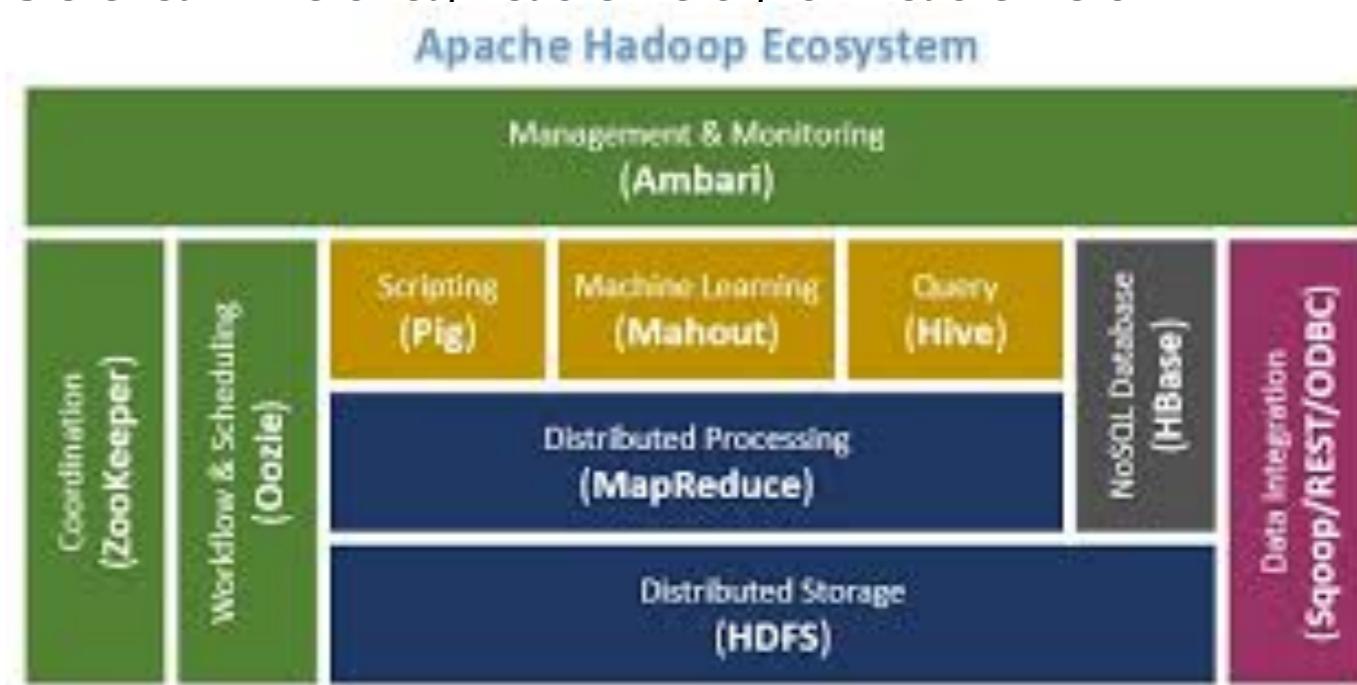
Why is Learning Important

- 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

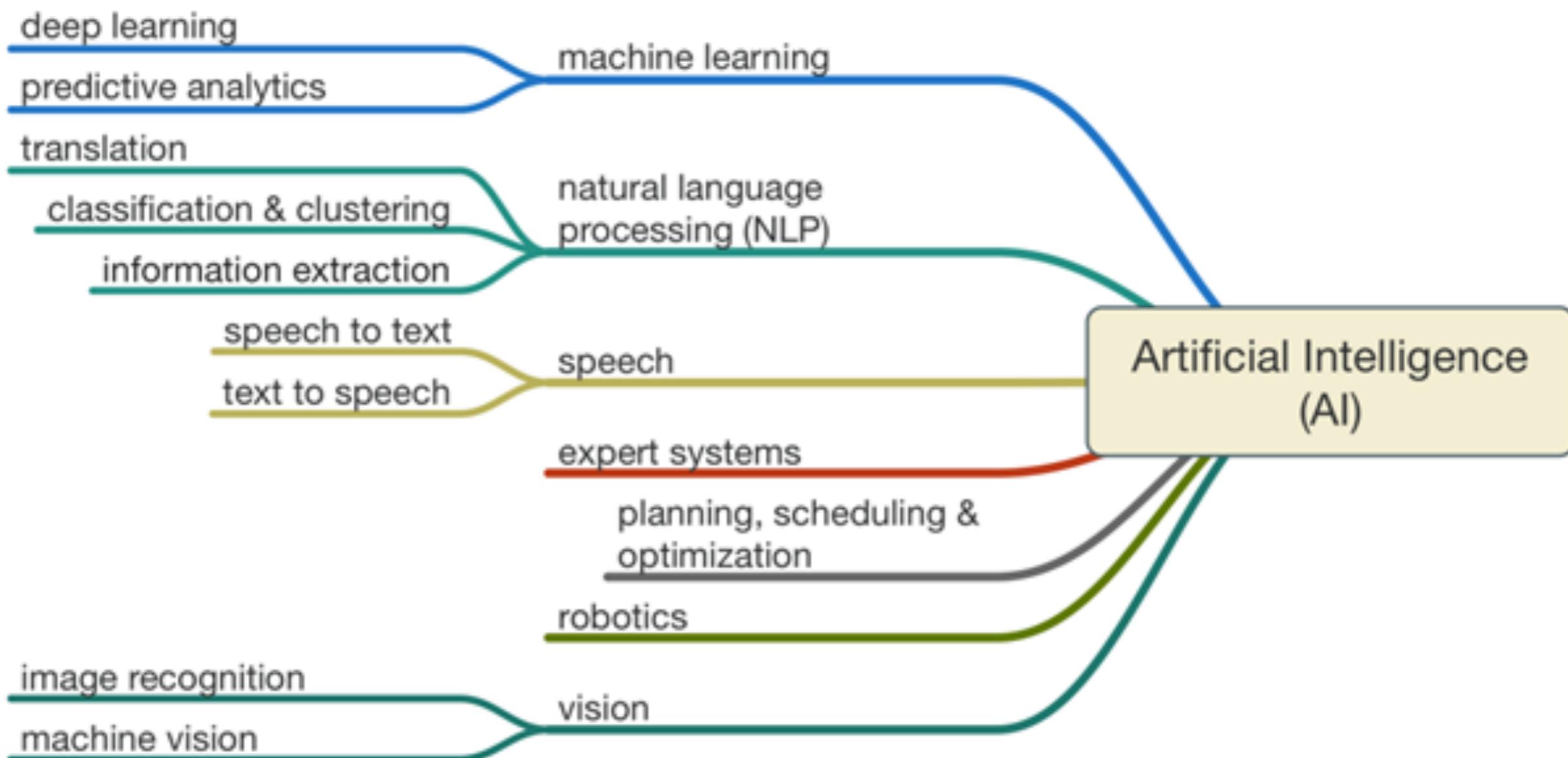


Why has ML become popular?

- 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



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?



amazon.com **Recommended for You**

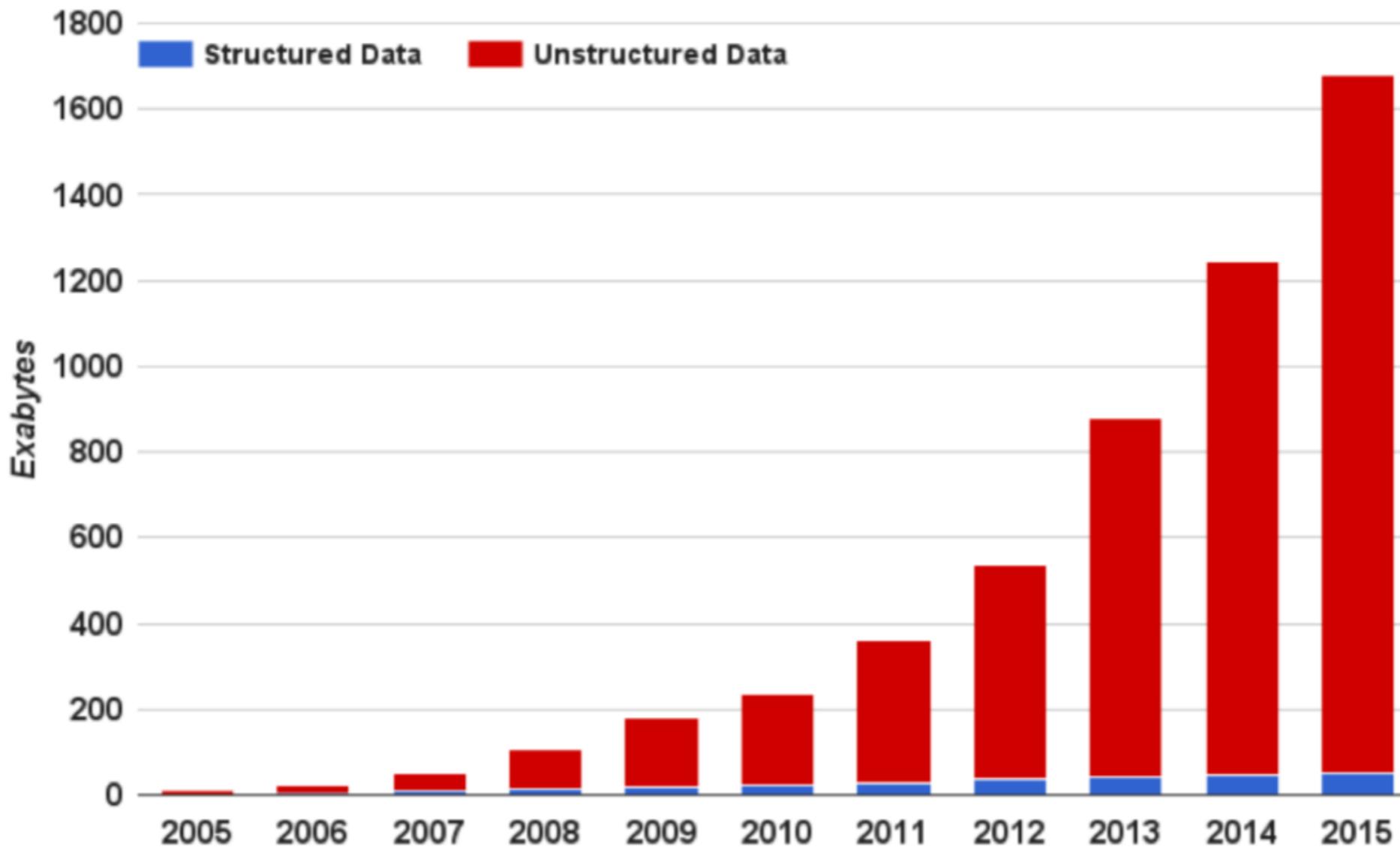
Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.

[Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop](#)

[Google Apps Administrator Guide: A Private-Label Web Workspace](#)

[Googlepedia: The Ultimate Google Resource \(3rd Edition\)](#)

Unstructured Data

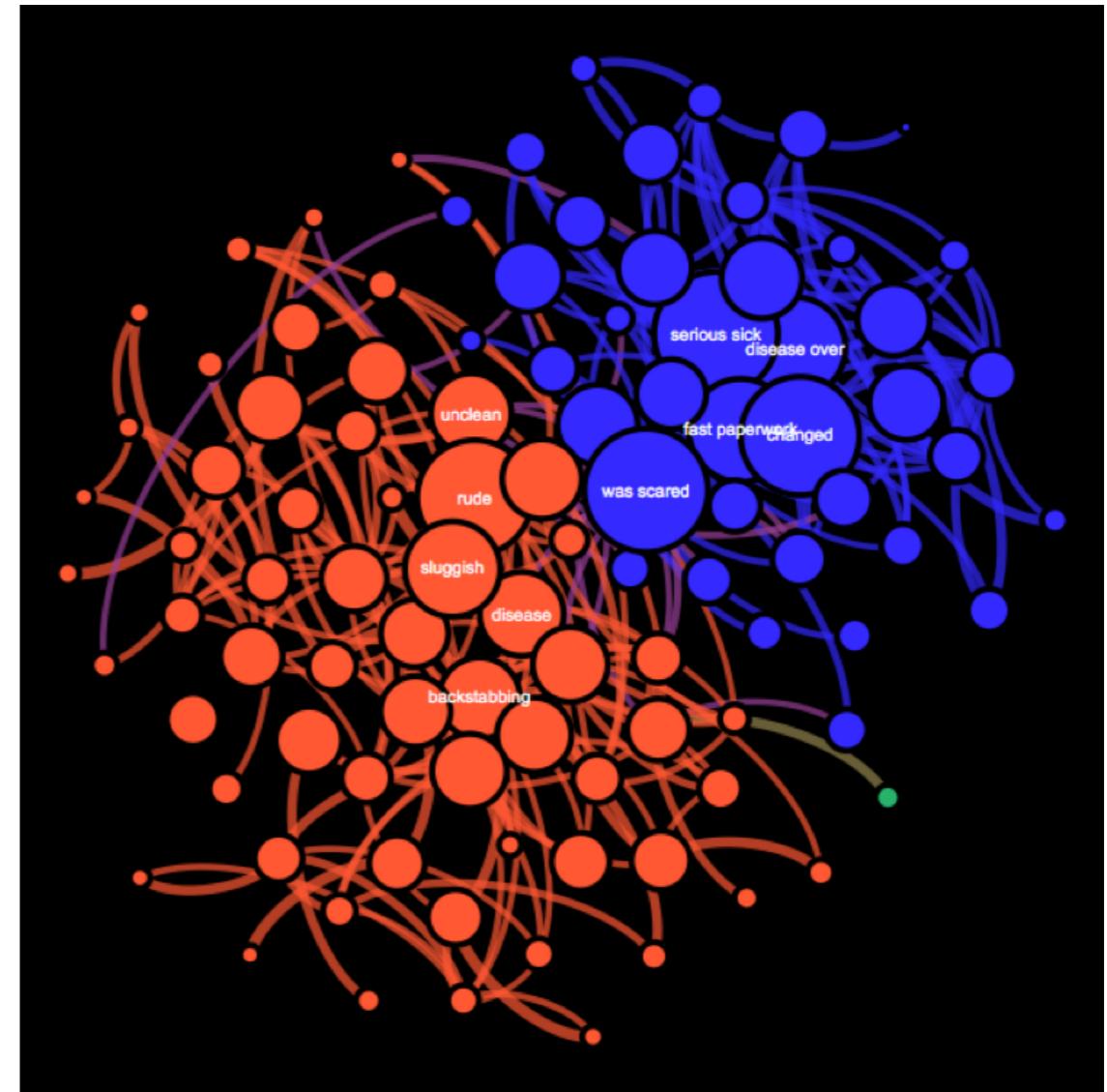


A.Nadkarni, N.Yezhkova, “Structured versus unstructured data: The balance of power continues to shift.” IDC (Industry Development and Models), March 2014.

Understanding Patients

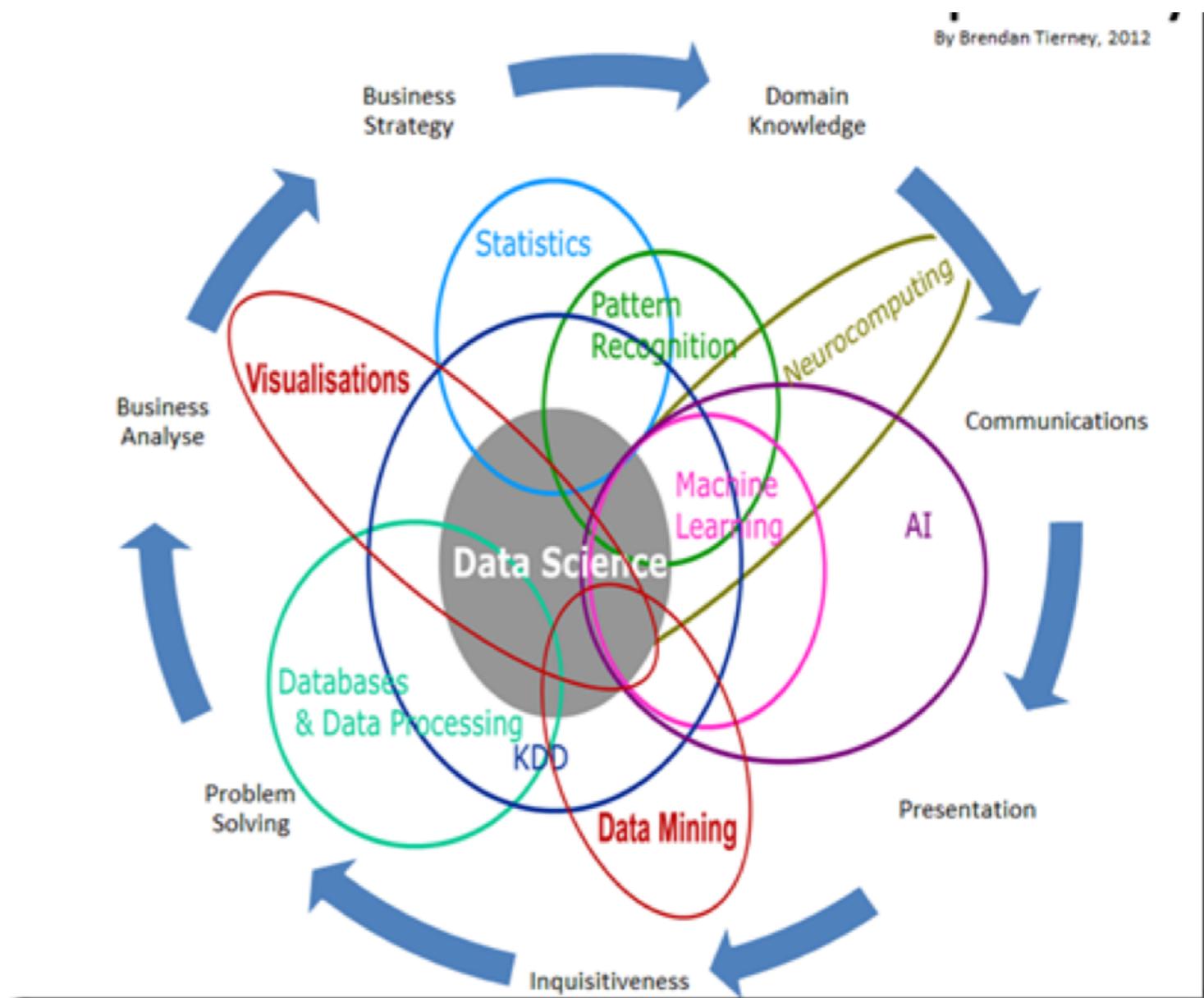


- Online reviews of primary care services (GPs) in England
- July 2013 - January 2017, 7.7K GP practices, 145K reviews
- ~ 3-5K reviews per month, 5-6 sentences long



Collaborate!

Ideal Data Scientist



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

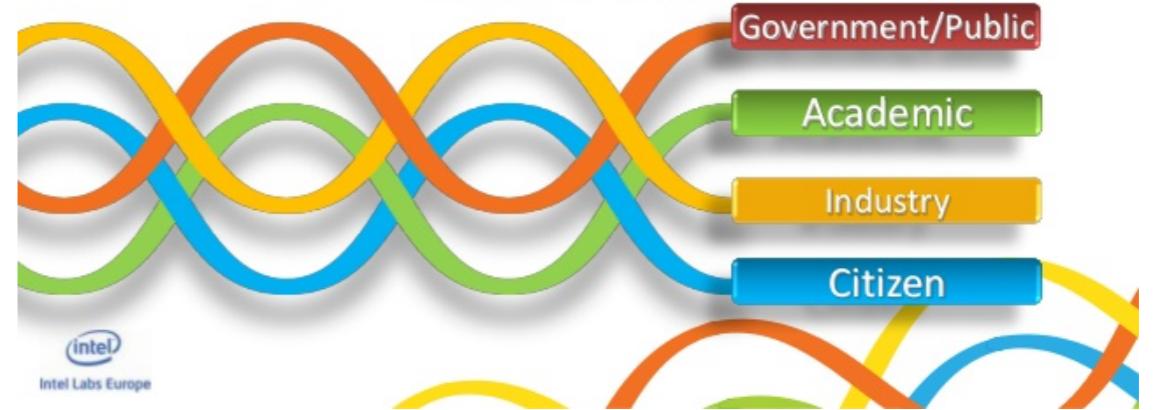
Image from <http://www.mysticwish.co.uk/product/anne-stokes-forest-unicorn-fridge-magnet/>

Wicked Problems Require a System Approach



Quadruple Helix Innovation

Government, Academia, Industry and Citizens collaborating together to drive structural changes far beyond the scope of any one organization could achieve on its own



“Research in Big Data should be grounded in the quadruple helix model where civil society joins with business, academia, and government sectors to drive changes far beyond the scope of what any organization can do on their own.”

*Intel Corp policy position paper on
Big Data*

Benefits for Academia



“In ML, where algorithms get published quickly and state-of-the-art frameworks are open-source, there isn't any first-mover advantage. Rather, competitive edge comes from data accumulation and infrastructure know-how. Which tends to benefit established large companies, rather than nimble upstarts with better tech.”

François Chollet, Deep learning at Google, Author of Keras, @fchollet

Collaboration

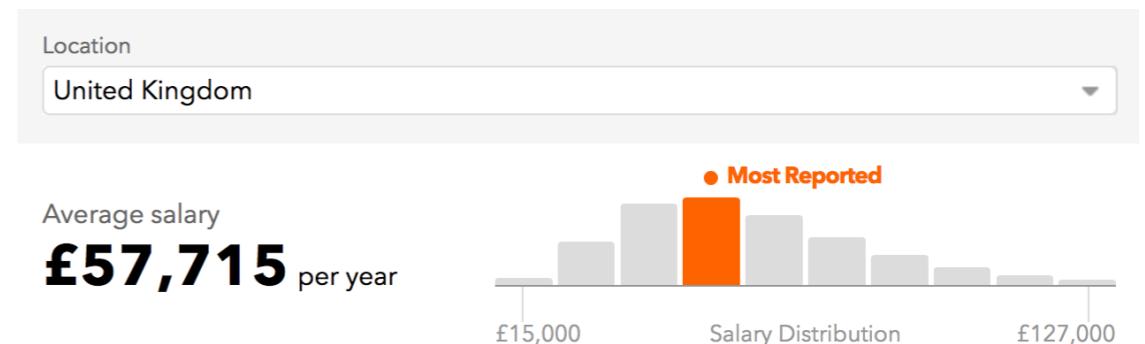
- Cost-Benefit rather than technical issues
- Unclear benefits of sharing data: vague and conceptual rather than tangible and related to business outcomes
- Cost of sharing (perceived privacy, security risks, resource costs) outweigh unclear benefits.



- Knowledge Transfer Partnerships
- Embedding
- Secondments
- Joint appointments

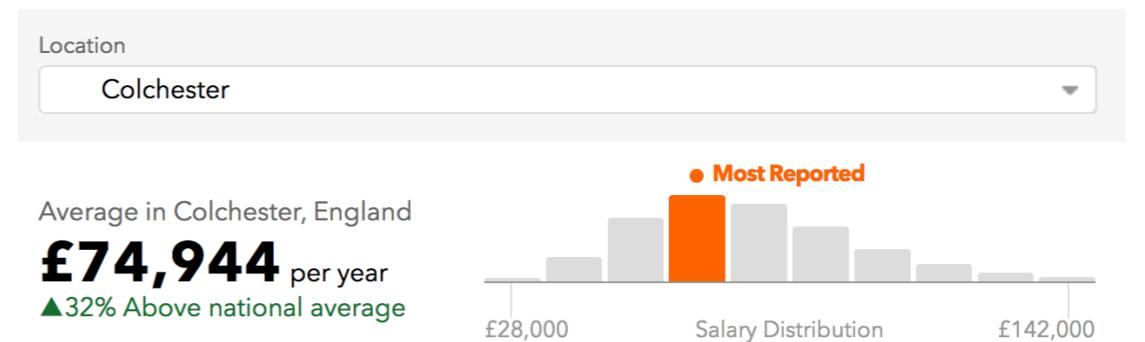
Data Scientist Salaries in the United Kingdom

Salary estimated from 14,248 employees, users, and past and present job advertisements on Indeed in the past 36 months. Last updated: 21 June 2018



Data Scientist Salaries in Colchester, England

Salary estimated from 24 employees, users, and past and present job advertisements on Indeed in the past 36 months. Last updated: 16 March 2018



Using Data Science in Policy

A report by the Behavioural Insights Team

- Children's social care
- Given the text of the initial referral and assessment, and structured data relating to the case, could we predict whether the case would be re-referred and escalated if it were closed?
- Machine learning and Natural Language Processing
- <http://bit.ly/2FIL0H3>

Figure 5: The inputs of the machine learning algorithm used to detect escalated closed cases.

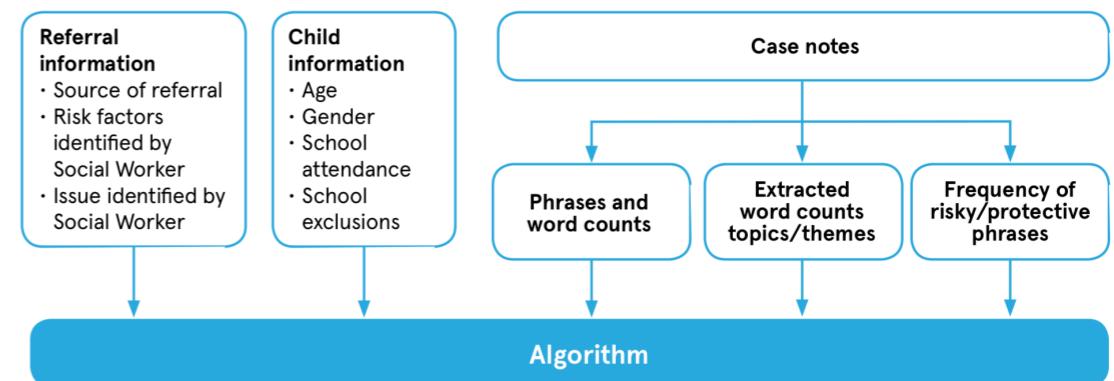
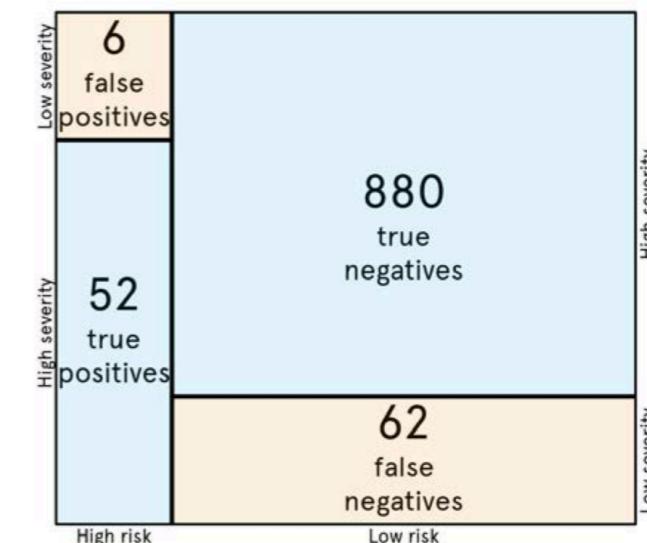


Figure 6: Expected distribution of true and false positives and negatives for 1,000 previously unclassified cases.



'Risk' refers to the model's prediction and 'severity' to the actual outcome. For example, a 'false negative' is a case that escalated but that the model categorised as 'low risk'.

EssexData: A risk profile for school readiness

- Can we predict whether children will be school ready on starting school?
- What factors are most likely to cause this early indicator of “best start in life”?
- 511 households at risk in the pilot area;
- Including 280 previously unknown to public services;
- 74% prediction accuracy rate.



**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
- AI is advancing fast (chatbots/agents)
- Hunting for unicorns...
- Explainability and ethical considerations
- Collaborate!