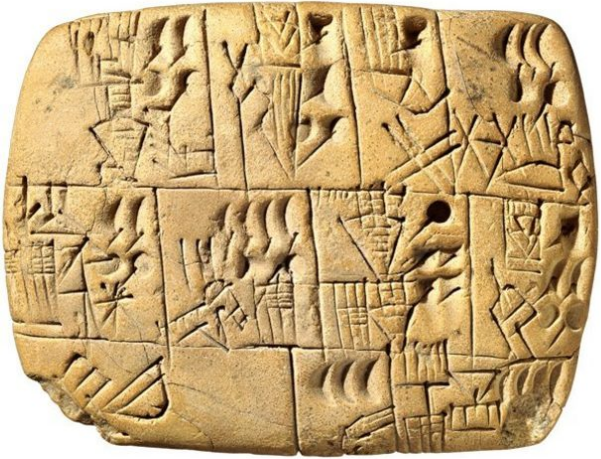
Data Science for Scientists

The what, why and how of Data Science

Gianluca Campanella
17th July 2018

What is Data Science?

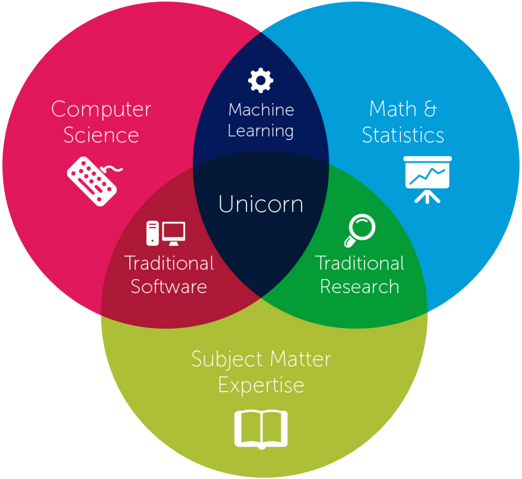
When did Data Science start?



From the British Museum collection

So…
What is Data Science?

What is Data Science?



From S. Geringer (originally from D. Conway)

How’s it different from…

- Applied Mathematics?
- Statistics?
- Operational Research?
- Business Intelligence?
- Predictive Analytics?
- Machine Learning?
- Data Mining?
- Knowledge Discovery?
- Deep Learning?
- Artificial Intelligence?

Data Science is…

Data-driven decision-making

- Focus is on the problem-solving process
- Multidisciplinary but domain-centric
- Tools are secondary!

Does this sound familiar?

Life in academia

The good…

- You figure out how things work
- You explain how things work to others (and to yourself)
- You build solutions to complex problems

Life in academia

…**and the bad**

- Few opportunities to move up the career ladder
- Fighting for research funds is fierce
- Work-life balance is nonexistent
- You’re constantly writing grants
- Did I mention Brexit?

Academia… or research?

The 'good' is actually quantitative research

- You're probably already doing it
- Companies like that! *

* Especially if you call it Data Science

What can Data Science do?

Two types of Data Science

Analysis-focused

- Maths and Statistics
- Business Intelligence

→ Assist human decision-making

Building-focused

- Machine Learning
- Software Engineering

→ Develop and deploy data-driven products

Statistics vs Machine Learning

Statistics

- Predates computers

→ **Understand why something happens** in the face of uncertainty

Statistics vs Machine Learning

Statistics

- Predates computers

→ **Understand why something happens** in the face of uncertainty

Machine Learning

- ‘Algorithmic modelling’ (L. Breiman)

→ Computers can **learn rules** without explicit programming

Who uses Data Science?

Opportunities

Domain	Applications
Finance	Financial forecasting Fraud and risk management
Marketing and sales	Churn analytics Dynamic pricing
Operations	Inventory optimisation Predictive maintenance Quality assurance
Workforce	HR analytics Resource planning

The five questions

1. How much/many?
2. Is this A or B?
3. How is this organised?
4. Is this weird?
5. What should I do next?

Supervised vs unsupervised learning

Supervised methods

- Learn from existing data
- Can be compared according to some ‘goodness’ metric

Unsupervised methods

- Don’t use examples with known outcomes
- Give clues, not ‘right answers’

How much/many?

Examples

- What will the temperature be next Sunday?
- What will total sales be next quarter?

$\downarrow$

Regression algorithms

Is this A or B?

Examples

- Which is more effective: a £10 voucher or a 10% discount?
- Will this machine fail in the next month?

↓

Classification algorithms

How is this organised?

Examples

- Which users like similar movies?
- Which items are frequently purchased together?

↓

Clustering algorithms

Is this weird?

Examples

- Is this transaction fraudulent?
- Is this blood pressure reading normal?

↓

Anomaly detection algorithms

What should I do next?

Examples

- Should the thermostat adjust the temperature?
- Where should the robot vacuum go next?

↓

Reinforcement learning algorithms

The five questions… revisited

Family	Class	Question
Supervised	Regression	How much/many?
	Classification	Is this A or B?
Unsupervised	Clustering	How is this organised?
	Anomaly detection	Is this weird?
Reinforcement learning		What should I do next?