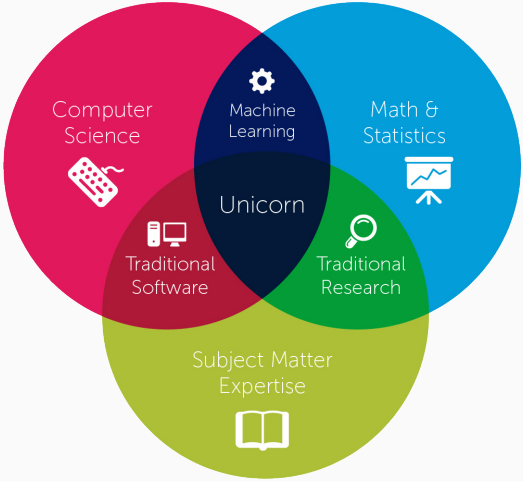
Introduction to Data Science

Gianluca Campanella

Contents

What is Data Science?

What is Data Science?



From S. Geringer (originally from D. Conway)

How's it different from…

Statistics

- Predates computers
- Understand why something happens in the face of uncertainty

How's it different from…

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

How's it different from…

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

Deep Learning

- Less structured inputs
- Computers can learn structure without explicit programming

Data-driven decision-making

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

What does Data Science deal with?

Problems!

Can we improve…

- The quality of offers we send to our customers?
- Road safety?
- How we identify people at high risk of cancer?

What does Data Science deal with?

Predictions?

How likely…

- Is a customer to respond to some offer?
- Are traffic accidents to occur in a certain area?
- Is a person to develop cancer in the next 10 years?

What does Data Science deal with?

Mechanisms?

Why…

- Does a customer decide to respond to some offer?
- Do traffic accidents occur regularly in certain areas?
- Do people develop cancer?

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

Recap

Data Science is…

- Evidence-based problem solving and decision-making
- Multidisciplinary but domain-centric
- Analysis-focused or building-focused

What can Data Science do?

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?

How much/many?

Examples

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

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

Supervised vs unsupervised algorithms

Supervised algorithms

- Are trained on existing data
- Can be compared according to some ‘goodness’ metric

Unsupervised algorithms

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

Data Science solutions

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