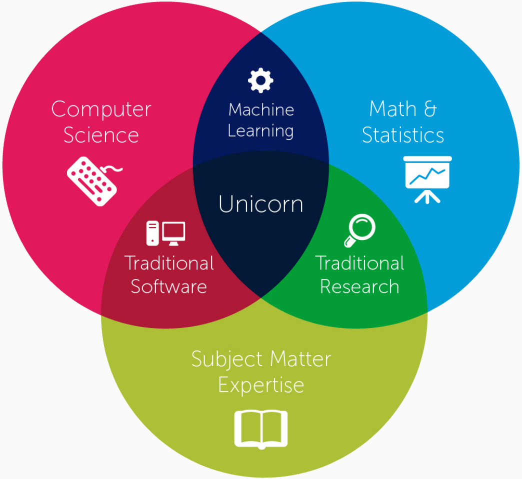
Demystifying Data Science

Gianluca Campanella

Contents

What is Data Science?

What is Data Science?



From S. Geringer (originally from D. Conway)

How’s it different from…

- 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!

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

What can Data Science do?

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?

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?

$\downarrow$

Classification algorithms

How is this organised?

Examples

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

$\downarrow$

Clustering algorithms

Is this weird?

Examples

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

$\downarrow$

Anomaly detection algorithms

What should I do next?

Examples

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

$\downarrow$

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	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?

How do you do Data Science?

High-level view

Business goal

↓

Testable hypothesis

↓

Experimentation and modelling

High-level view

Research question

$\downarrow$

Obtain $\longleftrightarrow$ Explore $\longleftrightarrow$ Model

$\downarrow$

Summarise / Operationalise

High-level view

This process is
non-linear and iterative

Define the research question

What to do

- Identify the problem and why it should be solved
- Frame it in the context of data collection

What to ask

- Which metrics do I need to improve?
- Which are possible actions to solve the problem?
- What is the benefit of solving the problem?

Obtain the data

What to do

- Measure the gap between ideal and available
- Think about assumptions and limitations

What to ask

- Are there enough data?
- Are they relevant to the research question?
- Can they be trusted?

Explore the data

What to do

- Data dictionary and any other documentation
- Descriptive statistics and visualisations

What to ask

- What kind of simple visualisations can I use?
- Which data types and distributions?
- Are there missing values or outliers?

Model the data

What to do

- Model selection and fitting
- Focus on inference and/or prediction

What to ask

- What is an appropriate model for the data?
- How can I evaluate model performance?
- Can the model be refined?

Summarise the findings

What to do

- Storytelling and visual aids to interpretation
- Communicate assumptions and limitations

What to ask

- How can I communicate results effectively?
- What format should I adopt?
- Who are my audience?

Operationalise

What to do

- System integration
- Monitoring and maintenance

What to ask

- What (visual) outputs do I care about?
- How often does the model need retraining?
- Do we need to think about scalability?

How do you implement it?

The secret is in the ingredients

Good Data Science requires:

- Tidy data
- Sharp questions
- A capable process
- Good people

How do you implement it?

Tidy data

Good data

Data Scientists

C onnected
A ccurate
R elevant
E nough

about data!

No data is better than bad data

Bad data < no data < good data < tidy data

Bad data

- Duplicate
- Missing
- Inaccurate or incorrect

Tidy data

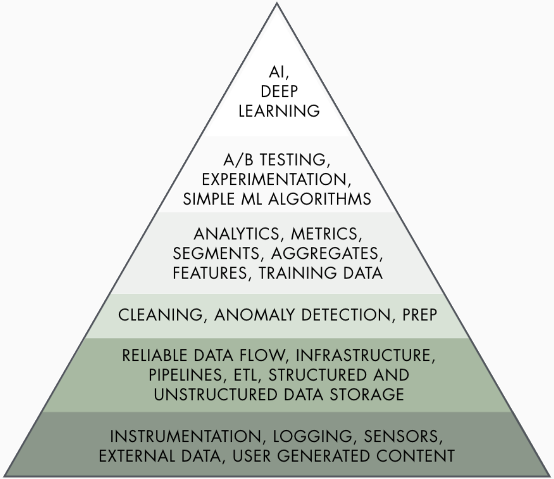
- Variables → columns
- Observation → rows
- Types → tables

How much data do I need?

- Appropriateness is normally more important
- However, there are certain statistical requirements…

Type of analysis	Sample size
Summary statistics	> 10
Parametric models	> 100
Most ML models	> 1,000
Deep Learning	> 100,000

Don’t try to run before you can walk



From M. Rogati

How do you implement it?

Sharp questions

Sharp questions

What people think of as the moment of discovery is really the discovery of the question.

— J. E. Salk

Sharp questions can be answered with data

→ Give clues as to which algorithms can answer them
→ Help identify the target data
→ Can be rephrased to give more useful answers

✗ What's going to happen with sales?
✓ What will total sales be next quarter?

Modelling misconceptions

Most well-executed data science projects don't...

- Use complicated tools
- Fit complicated models

Instead, they do...

- Focus on solving the problem
- Use appropriate — not necessarily big! — data
- Use relatively standard models

The 80—20 rule of modelling

- The first reasonable thing you can do goes 80% of the way
- Everything after that is to get the remaining 20%...
  often at additional cost!

The 80—20 rule of modelling

- The first reasonable thing you can do goes 80% of the way
- Everything after that is to get the remaining 20%…
  often at additional cost!

Is it worth it?

Know your domain

Domain knowledge allows you to...

- Understand possible data collection flaws
- Identify feature dependence and leakage
- Create new features (feature engineering)
- Interpret your results correctly
- Be understood by stakeholders

How do you implement it?

A capable process

A Data Scientist's dream

In an ideal world there are...

- Tidy data
- Sharp questions
- Resources and time to experiment and model

The sad reality…

What's a capable process?

Are you maximising…

Certainty	**Growth rate**
'Unchanging' truths	Changing truths
↓	↓
Science	Adversarial industries

ROI of Data Science projects

No one knows which projects will have the best ROI!

- Power law-like distribution of returns

→ Do several projects in short sprints

- Failure is always an option!

→ Learn when to cut losses

High-risk, high-reward innovation culture

Data strategy $\rightarrow$ Product roadmap

Data collection

Leadership buy-in

How do you implement it?

Good people

Good people

- Traditional analysts often focused on specific tools
- Many programmers don’t have business experience

Good people

- Traditional analysts often focused on specific tools
- Many programmers don’t have business experience

Successful Data Scientists are...

- Practical, impact-driven, dependable people
- Passionate about their domain
- Knowledgeable about research methods and statistics
- Coding ninjas

Good teams

Successful Data Science teams are...

- Flexible and open
- Diverse
- Collaborative