SCC.460 Data Science Fundamentals

What is Data Science

What is Data Science?

- Wikipedia: “Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data”
  - Could be used to increase efficiency, save costs, etc.
    - Insights to help detect fraudulent insurance claims
  - Support public health
    - Look for correlations relating to obesity
      - Gain Insights into when and where to use targeted advertising
  - Is this new?
    - *Cf,* H.P. Luhn, IBM Journal, October 1958
  - If not, what is?

What is data?

- Data are evidence of events.
- This could be indicative of an underlying chronicle (narrative).
  - e.g. Evidence of how I go about regulating the temperature of my office…
  - e.g. Evidence to support classifying (likely) fraudulent insurance claims based on social media activity
    - See for example:
      - https://insurancedatascience.org/project/2019_zurich/
  - e.g. Evidence to support correlation between obesity and deprivation?
- Digging into data to find the narrative and learning from that *is* data science.

So what is Data Science?

- Collect all the data!!
  - Different sources: Sensors/IoT, social media, etc.
- Automation
  - Takes away much of the heavy lifting
- Interpretation media
  - Richer communication, wider audiences, larger impact

The Data Science Process

- Jeff Hammerbacher’s Model:
  - Identify problem
  - Instrument data sources
  - Collect data
  - Prepare data (transform, clean, filter, integrate, aggregate)
  - Build model
  - Evaluate model
  - Communicate results

The Data Science Process

- O’Neil and Schutt
- Start with real world problems.
- Feed back into the real world.

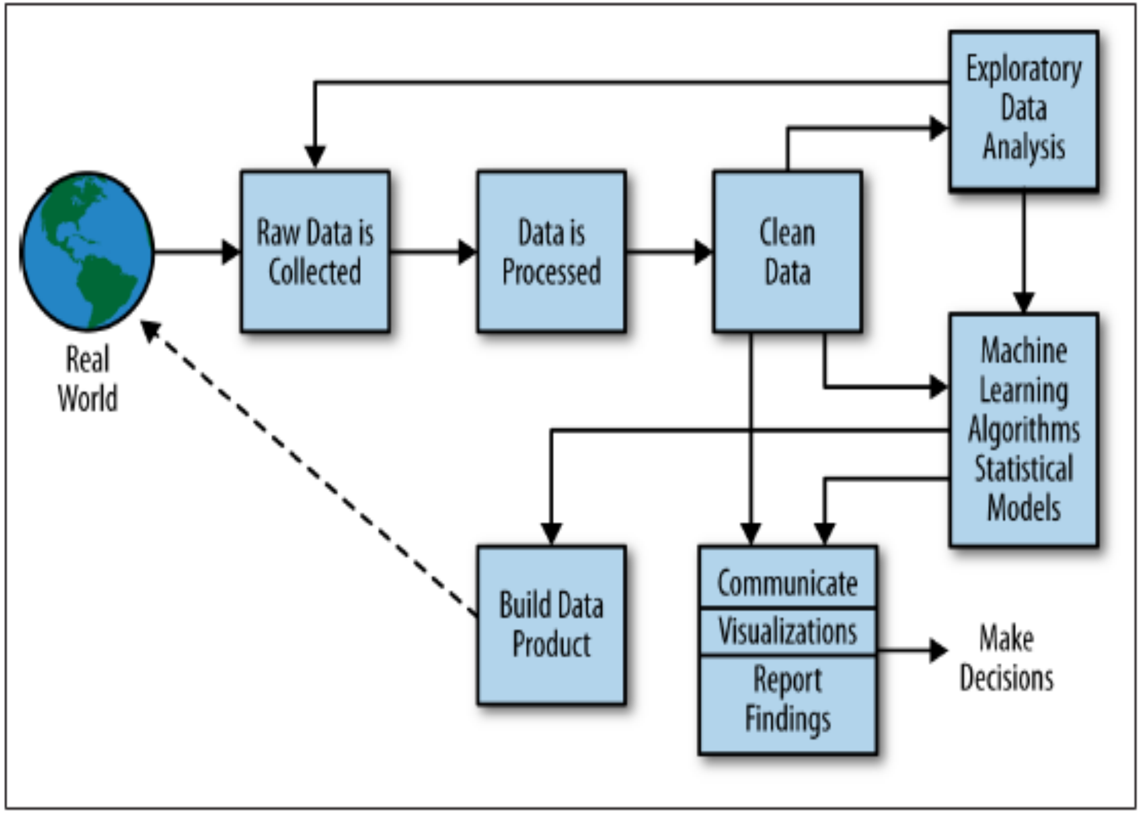


Figure 2-2. The data science process

The Data Science Process in your MSc

- SCC460
- SCC403
  - Data Mining
- SCC461
  - Prog’ing for Data Scientists
- CFAS440/CFAS406
  - Statistical Methods and Modelling
  - Statistical Inference
- SCC401/SCC411/SCC413

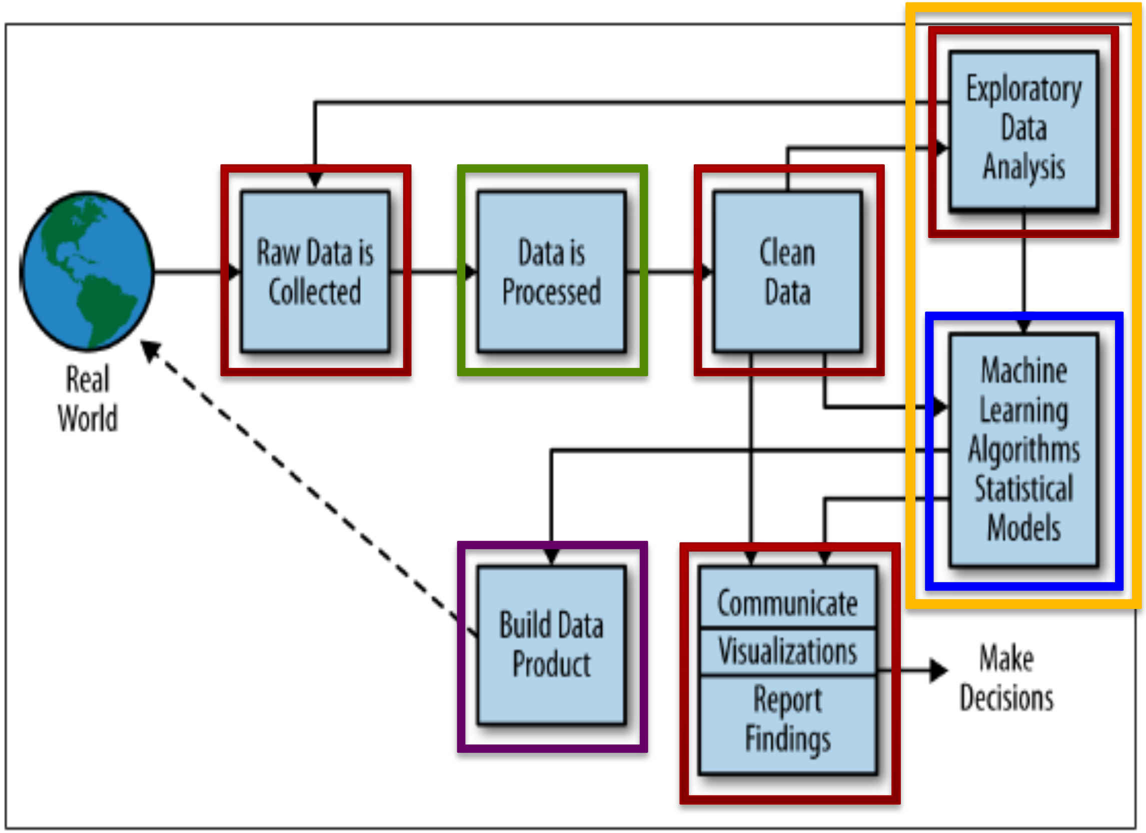
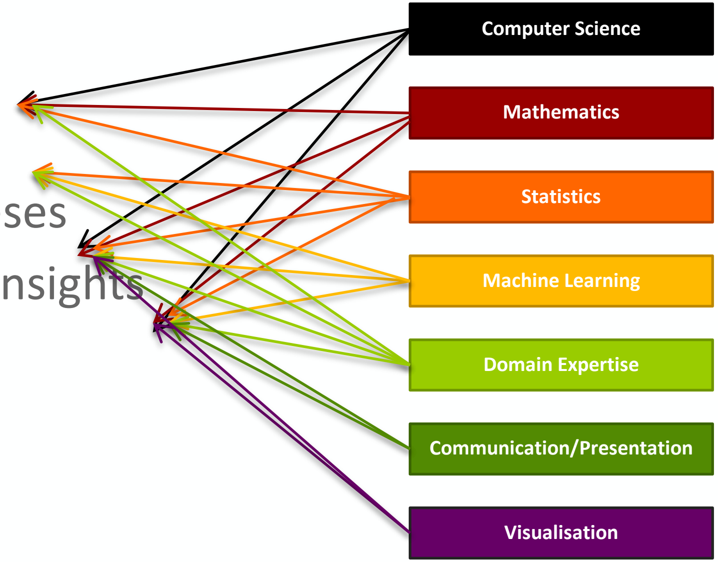

Real World
Raw Data is Collected
Data is Processed
Clean Data
Exploratory Data Analysis
Machine Learning Algorithms Statistical Models
Build Data Product
Communicate Visualizations Report Findings
Make Decisions

Figure 2-2. The data science process

What is Data Science

- A myriad of processes; some by human, others automated.

- Main human processes:
  - Data cleaning (at least initially)
  - Forming hypotheses
  - Designing tests for the hypotheses
  - Communication of results and insights



What is Data Science

- “An *academic* data scientist is a scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem.” -- O’Neill and Schutt

What is a Data Scientist

- Someone who knows how to:
  - identify problems, formulate hypotheses (domain expertise)
  - collect, clean, and transform data (programming, statistics)
  - extract meaning from data (statistics, mathematics, machine learning)
  - interpret data, forming hypotheses (domain expertise, machine learning)
  - communicate results (domain expertise, visualisation, communication)
- Many things:
  - highly technical knowledge
  - strong domain expertise
- You’ll find that good data science happens in **multidisciplinary teams**.
  - We will do that