Knowledge Discovery

Javier Béjar

URL - 2024 Spring Term

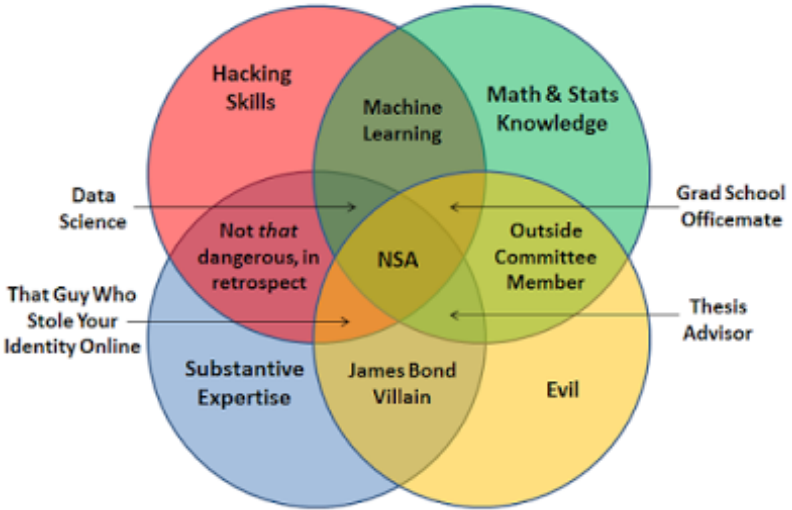
CS - MIA

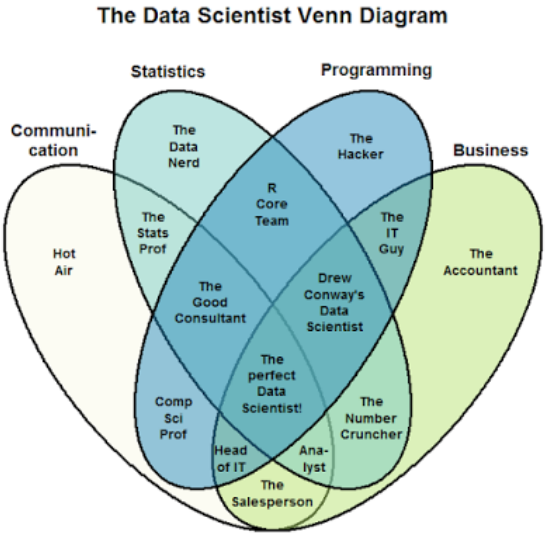
Knowledge Discovery (KDD)

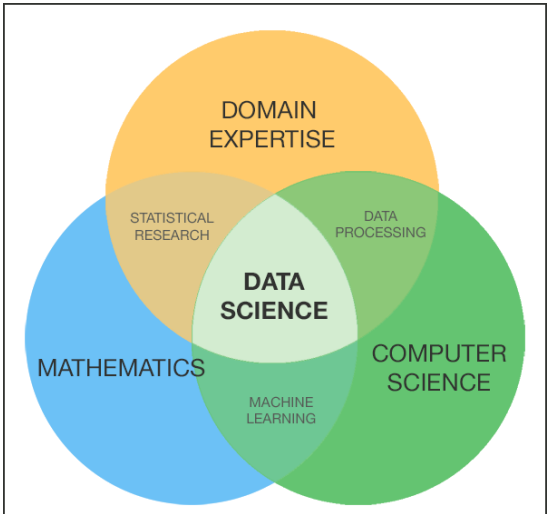
- Practical application of the methodologies from machine learning/statistics to large amounts of data
- **Problem:** The impossible task of manually analysing all the data we are systematically collecting
- Useful for automating/helping the process of analysis/discovery
- **Final goal:** To extract (semi)automatically actionable/useful knowledge

"We are drowning in information and starving for knowledge"

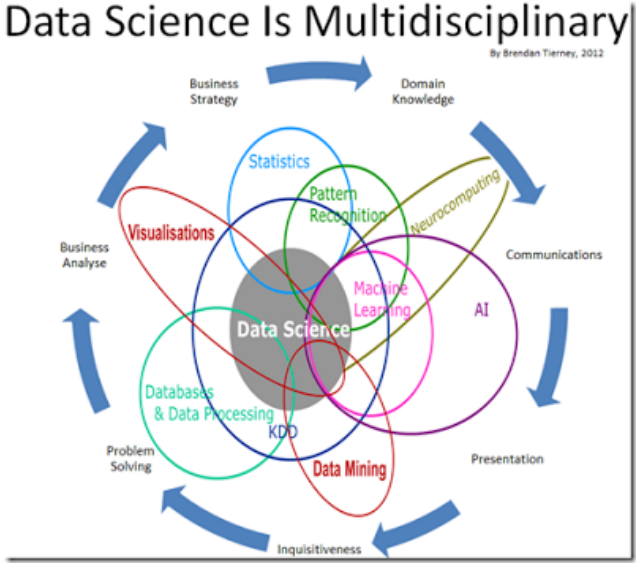
- The high point of KDD starts around late 1990s
- Many companies show their interest in obtaining the (possibly) valuable information stored in their databases (purchase transactions, e-commerce, web data...)
- The area has moved/integrated/transmuted several times to include several sometimes interchangeable terms: Business Intelligence, Business Analytic, Predictive Analytics, *Data Science*, Big Data...
- The Venn Diagram Wars

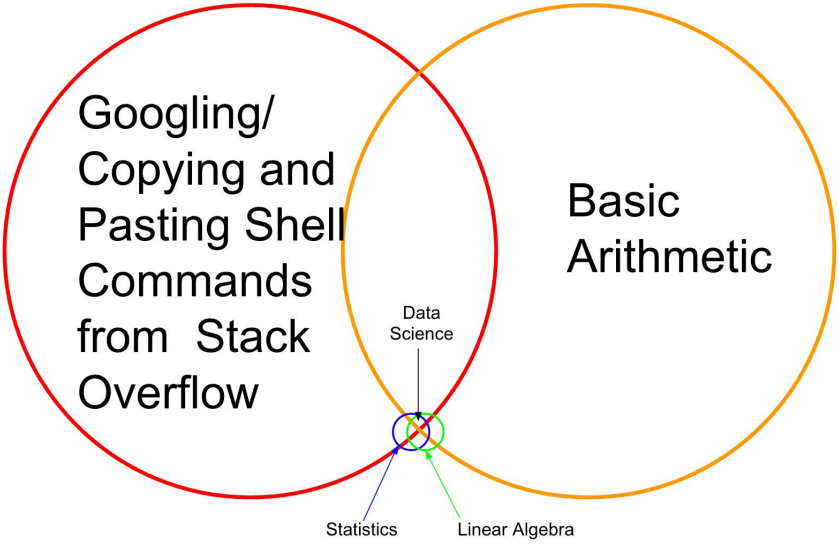

Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil


The Data Scientist Venn Diagram
Statistics
Programming
Communi-
cation
Business
The
Data
Nerd
The
Hacker
R
Core
Team
The
Stats
Prof
The
IT
Guy
Hot
Air
The
Accountant
The
Good
Consultant
Drew
Conway's
Data
Scientist
The
perfect
Data
Scientist!
Comp
Sci
Prof
The
Number
Cruncher
Head
of IT
Ana-
lyst
The
Salesperson



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.


Data Science Is Multidisciplinary
By Brendan Tierney, 2012
Business Strategy
Domain Knowledge
Communications
Presentation
Inquisitiveness
Problem Solving
Business Analyse
Statistics
Pattern Recognition
Neurocomputing
Visualisations
Machine Learning
AI
Data Science
Databases & Data Processing
KDD
Data Mining


Googling/
Copying and
Pasting Shell
Commands
from Stack
Overflow
Basic
Arithmetic
Data
Science
Statistics
Linear Algebra

"It is the search of valuable information in great volumes of data"

"It is the explorations and analysis, by automatic or semi-automatic tools, of great volumes of data in order to discover patterns and rules"

"It is the ***nontrivial*** *process of identifying* ***valid***, ***novel***, *potentially* ***useful***, *and ultimately* ***understandable patterns*** *in data"*

Pattern: Any representation formalism capable to describe the common characteristics of data

Valid: A pattern is valid if it is able to predict the behaviour of new information with a degree of *certainty*

Novelty: It is novel any knowledge that it is not know respect the domain knowledge and any previous discovered knowledge

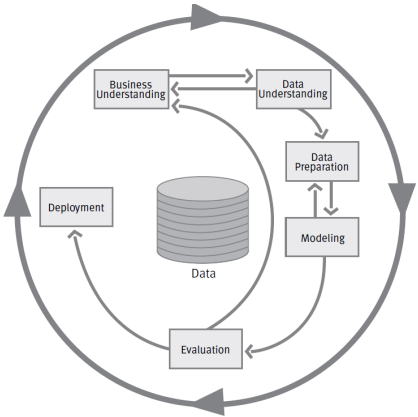
Useful: New knowledge is useful if it allows performing actions that yield some benefit given a established criterion

Understandable: The knowledge discovered must be analysed by an expert in the domain, in consequence the interpretability of the result is important

The KDD process

- The actual discovery of patterns is only one part of a more complex process
- Raw data in not always ready for processing (80/20 project effort)
- Some general methodologies have been defined for the whole process (CRISP-DM or SEMMA)
- These methodologies address KDD as an engineering process, despite being business oriented are general enough to be applied on any data discovery domain

- Steps of the Knowledge Discovery in DB process
  1. Domain study
  2. Creating the dataset
  3. Data preprocessing
  4. Dimensionality reduction
  5. Selection of the discovery goal
  6. Selection of the adequate methodologies
  7. Data Mining
  8. Result assessment and interpretation
  9. Using the knowledge



There are different goals that can be pursued as the result of the discovery process, among them:

Classification We need models that allow to discriminate instances that belong to a previously known set of groups (the model could or could not be interpretable)

Clustering/Partitioning/Segmentation We need to discover models that clusters the data into groups with common characteristics (a characterizations of the groups is desirable)

Regression We look for models that predicts the behaviour of continuous variables as a function of others

Summarization We look for a compact description that summarizes the characteristics of the data

Causal dependence We need models that reveal the causal dependence among the variables and assess the strength of this dependence

Structure dependence We need models that reveal patterns among the relations that describe the structure of the data

Change We need models that discover patterns in data that has temporal or spatial dependence

Where unsupervised learning fits?

- Preprocessing
  - Cleaning: Outliers, missing values
  - Transforming: Normalization, continuous to discrete
  - Dimensionality reduction
- Feature engineering
  - Feature extraction
  - Embeddings
  - Disentanglement
- Modelling:
  - Partitioning (for understanding/summarization)
  - Attribute relationships (frequent paterns discovery, causal rules)
  - Probabilistic generative modelling (for understanding/generating new data)