Intro to Data Science

Alistair Walsh

Alistair Walsh

Instructor

PhD candidate at The Melbourne Brain Centre (The Florey Institute)

Lead Instructor Melbourne Data Science Intensive course - GA

Fellow (by Training) of the Australasian College of Health Informatics

BSc. Hon. Cognitive Neuroscience

BSc. Psychology/Psychophysiology

Dip. Laboratory Technology (Biotechnology)

Introduction

What is Data Science?

What is a Data Scientist?

“Data Scientist’ is a Data Analyst who lives in California”

“A data scientist is someone who is better at statistics than any software

engineer and better at software engineering than any statistician."

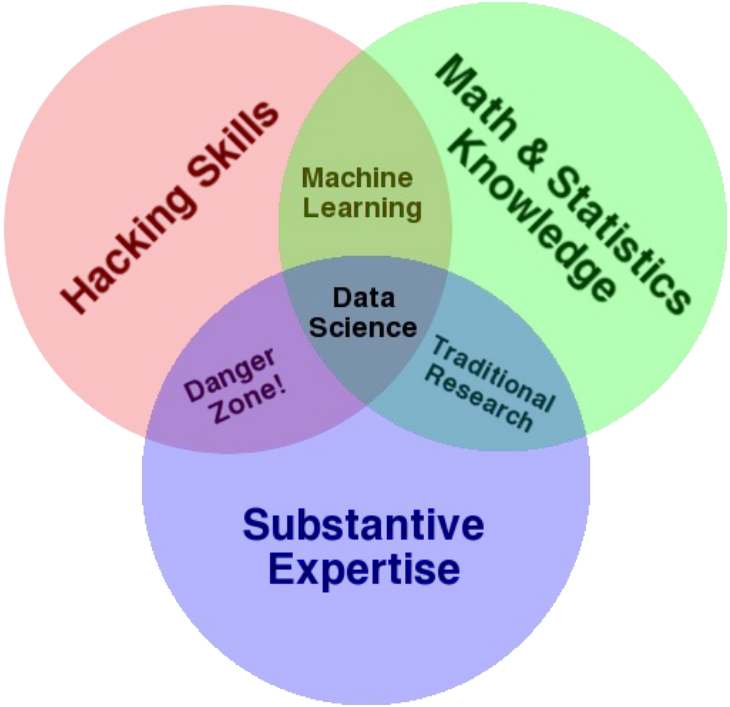
Someone who can collect, statistically explore and analyse data in an efficient and reproducible manner... but who can also translate from Dataese to Peoplese. Oh, and something something machine learning.

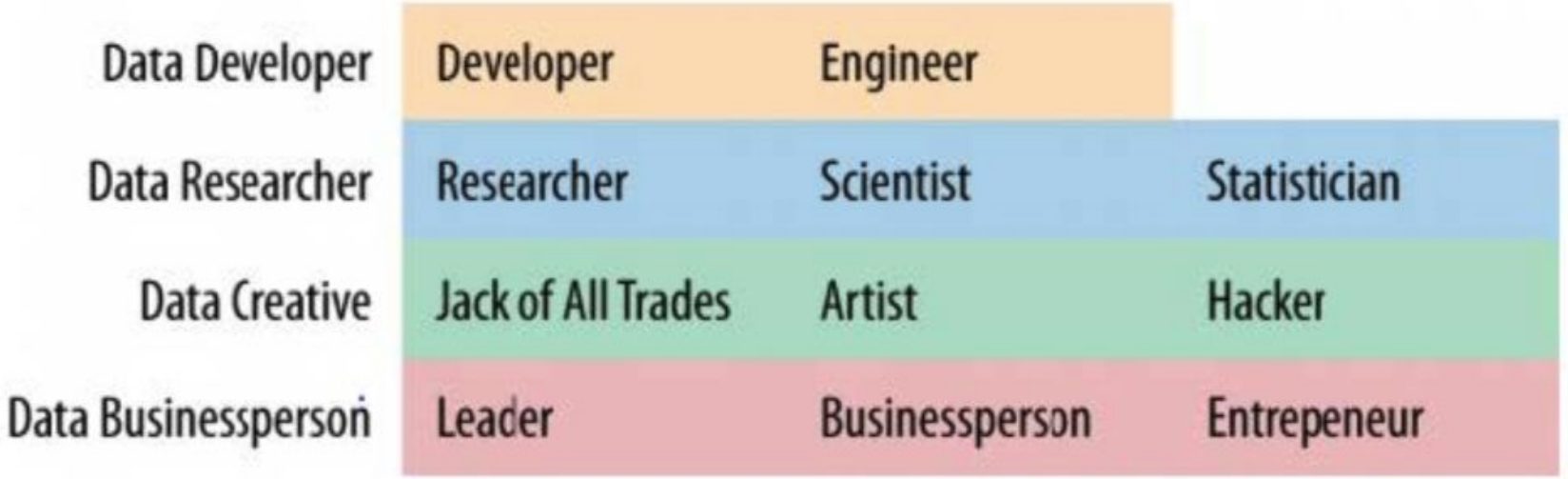
WHAT IS DATA SCIENCE?

A set of tools and techniques for data

Interdisciplinary problem-solving

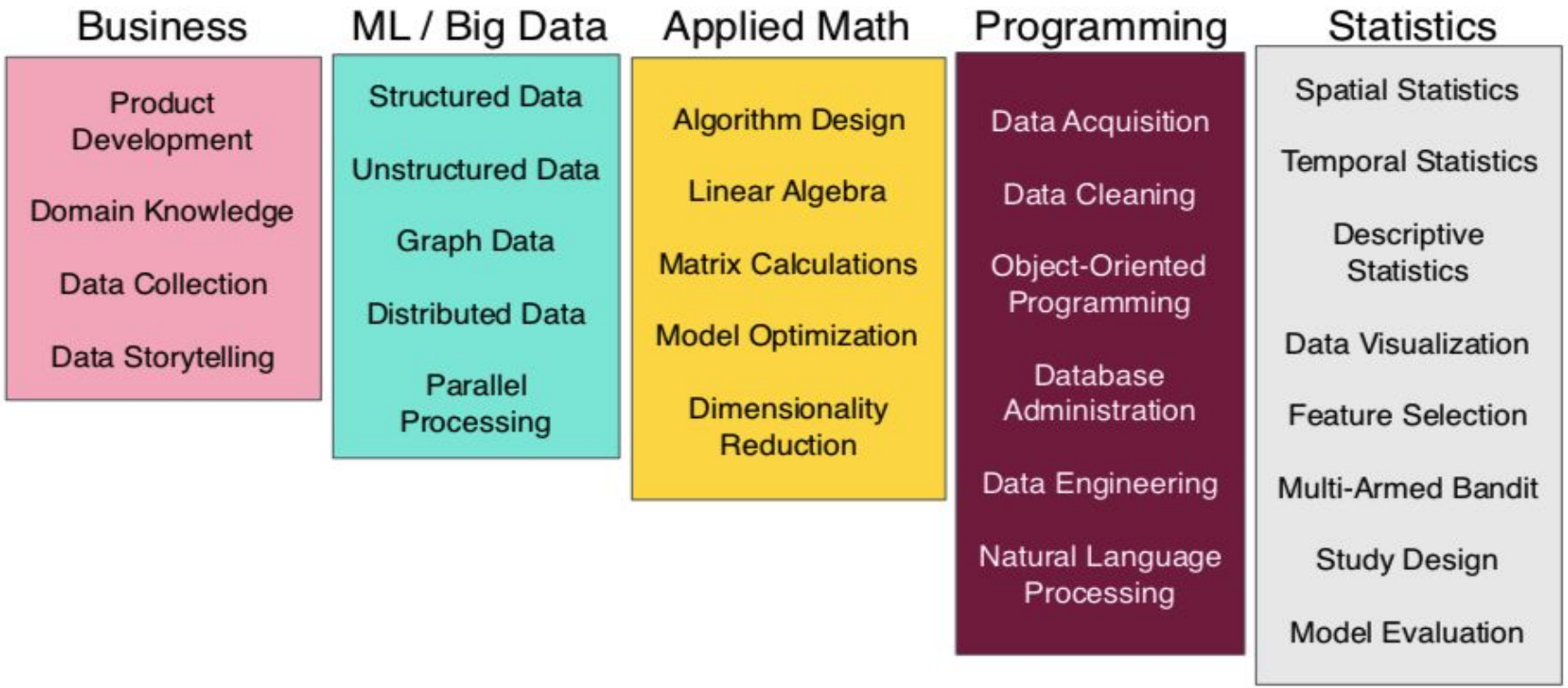
Application of scientific techniques to practical problems

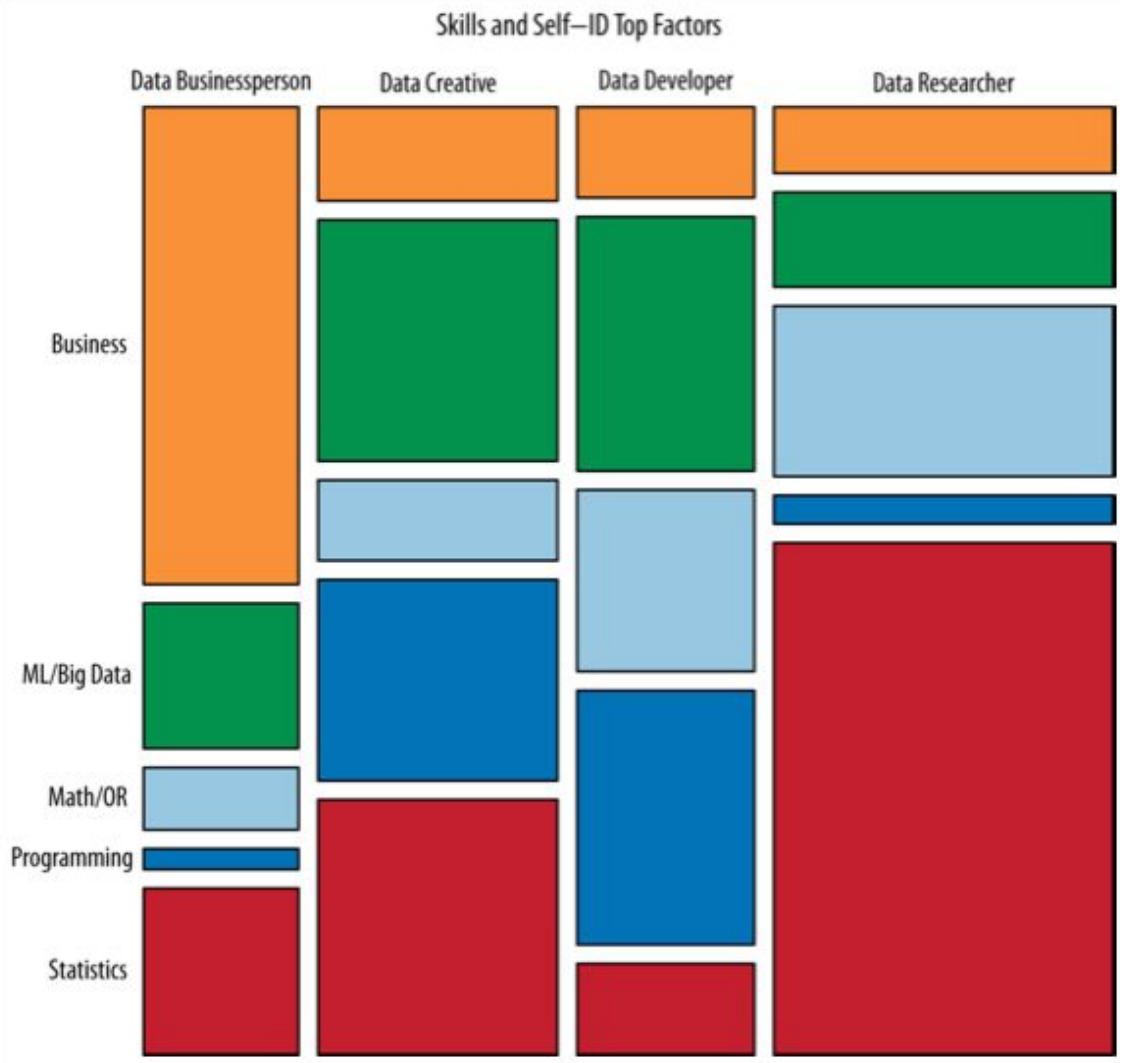




Data Science involves a variety of roles, not just one.

Data Science involves a variety of skill sets, not just one.





GENERAL ASSEMBLY

Roles in Data Science

These roles prioritize different skill sets.

However, all roles involve some part of each skillset.

Where are your strengths and weaknesses?

Why Now?

‘Big Data’

+ Algorithms

+ **Computing Power** = Data Science

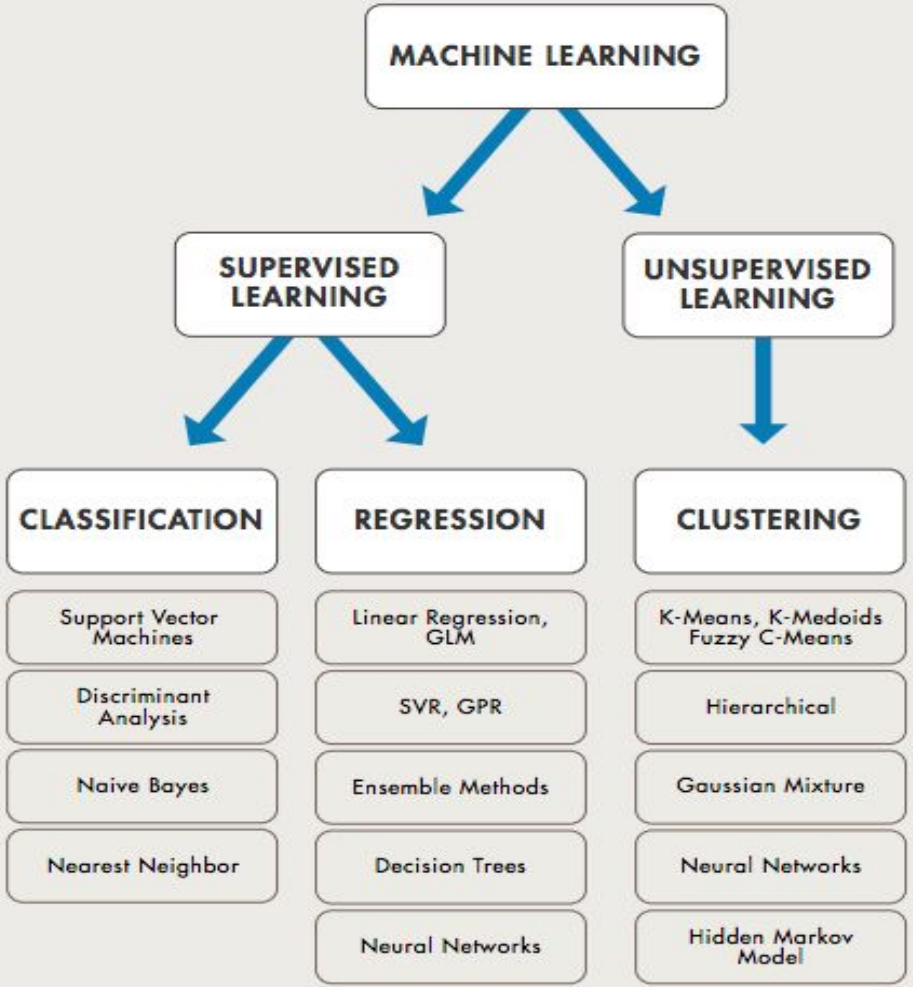
Problems We Solve

Understand	Unsupervised Learning	What has happened?
Predict	Supervised Learning	Given this, what will happen?
Influence	Feature Importance	What factors must I change to bring about what I want

Cambridge Analytica

https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

How we solve the problems we solve



What type of ML are we doing?

Classification or regression?
Supervised or Unsupervised?

I want to know what type it is based on the type of other examples - Classification

I want to predict a value based on past values of similar items - Regression

My dataset is unlabeled - Unsupervised/Clustering

The Data Science Workflow

GENERAL ASSEMBLY

The steps:

1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results

IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

ACQUIRE THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

PARSE THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

MINE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean, slice, and combine data in Python
- ☐ Create necessary derived columns from the data (new data)

REFINE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Document and transform data

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

Data Science Toolbox

SQL

Let’s Code!

GENERAL ASSEMBLY