Data Science In the Wild: Our Role In Society

RMIT, 20 August 2018

Colette Marais

Outline

- My Journey
- The Current State of Data Science
- When Big Data Becomes Small
- Bias In = Bias Out
- The Devil Is In The Question
- Be The Change

My Journey*

*A journey based on N = 1

The Data Science Journey Begins

So, now what? Academia…?

Overview:

- How do mining companies make investment decisions
- Understand the impact of regulation on decision making
- Access to mining, production and pricing data from a mine
- Opportunity to apply many techniques
  - SVM for geological classification
  - Markov model of pricing
  - Optimisation strategy for dynamic programming

Benefits

- Self-directed
- Hard and interesting problems
- Time to explore latest techniques
- Attend conferences and meet top academics
- Unlimited access to university resources

Realities

- Self-directed
- Funding
- Real-world data is near-impossible to access
- Difficult to work across disciplines
- Take years for impact outside academia
- Scalability? Production quality?

Give me some variety!

Overview:

- Worked on applications in media, marketing, insurance and banking
- Focus on building scalable, end-to-end analytical solutions
- Close engagement with business stakeholders
- Exposure to large datasets and the challenges they pose
- Access to enterprise tools
- Colleagues from diverse backgrounds

Benefits

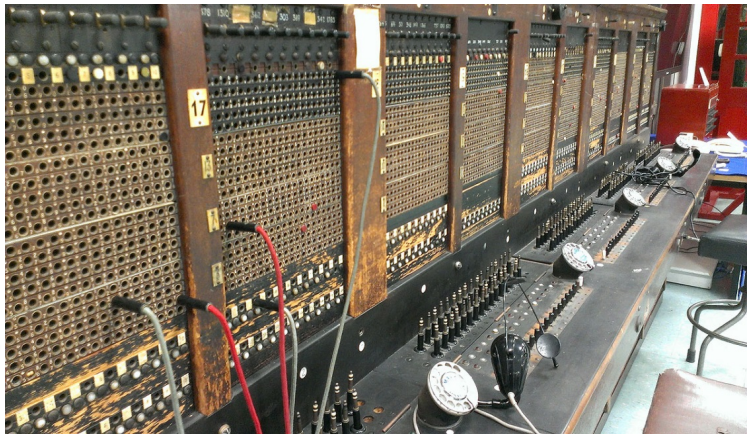
- Diverse set of problems
- Work across multiple industries
- Develop network
- Multi-disciplinary teams
- Best practice solution implementations

QUANTIUM Deloitte.

Realities

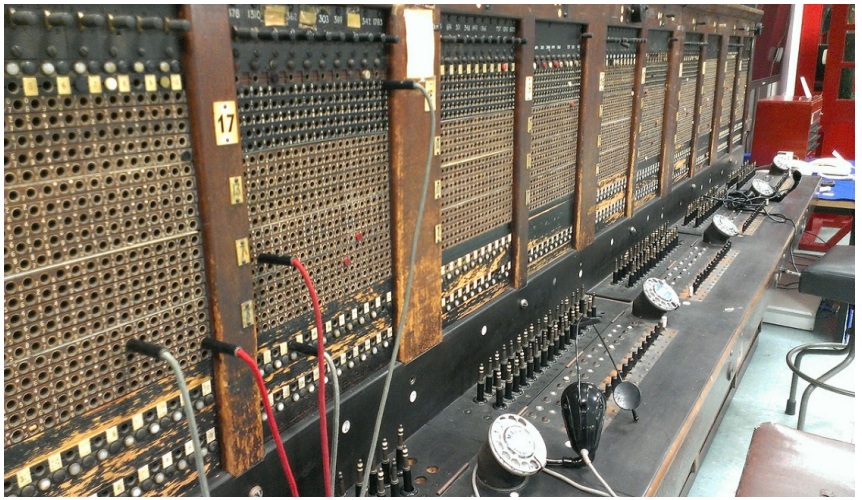
- Analytics problems are hard to sell
- Competing with in-house analytics teams for best problems
- Timelines can limit depth of work
- Limited control over the design or course of the project
- Do not always see the impact of work

Where did all the big problems go?



Overview:

- Biggest telco in Australia which serves 40% of population
- So much data!
- A whole new world of technology
- Dominant market player facing a changing marketplace
- Mandate to use analytics to transform the business



Benefits

- So much data!
- Guide a project from start to end
- See the impact of your work
- Many learning opportunities

Realities

- So much MESSY data!
- Immature technology stack
- Gaining the trust and understanding of your business stakeholders
- Get caught up in BAU
- Turning a large ship takes time

The Current State of Data Science

Current State of Data Science: The Exciting

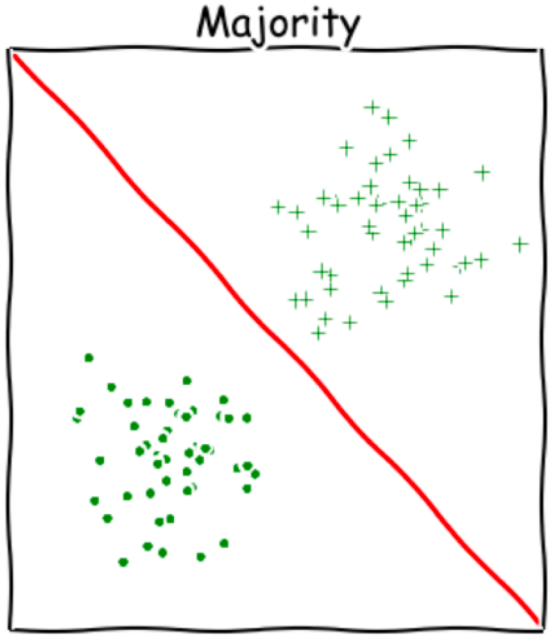
- The sexiest job of the 21st Century
- Number of data science programs grew by 7.5% between 2010 and 2015
- Salaries increased by 16% between 2012 and 2014
- McKinsey Global Institute estimates a shortfall of 250,000 data scientists by 2024
- Data-driven decision-making is becoming central to many organisations

Current State of Data Science: The Complex

- Our problems have evolved
- Datasets are not simply big, they are *granular*
- Moved away from domain expertise
- Tools, tools, tools!
- Is this really my problem?

When Big Data Becomes Small

When Big Data Becomes Small



Problem statement:

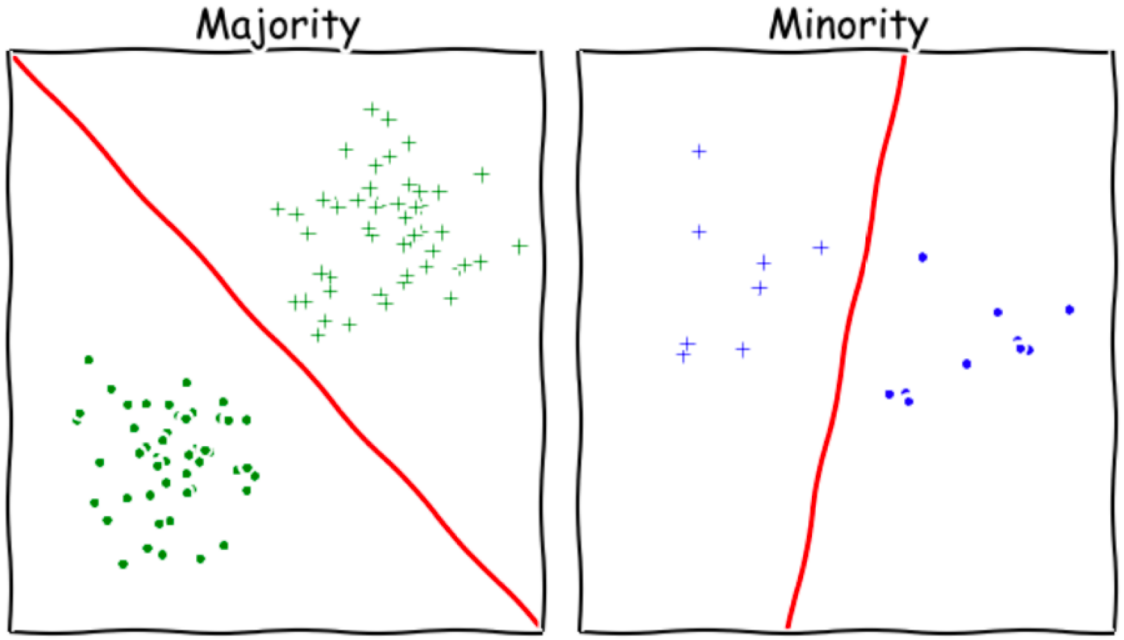
- Determine fraudulent profiles from user names

A classifier generally improves with the number of data points

- White male names are straightforward to classify

Source: Moritz Hardt, “How big data is unfair”

When Big Data Becomes Small



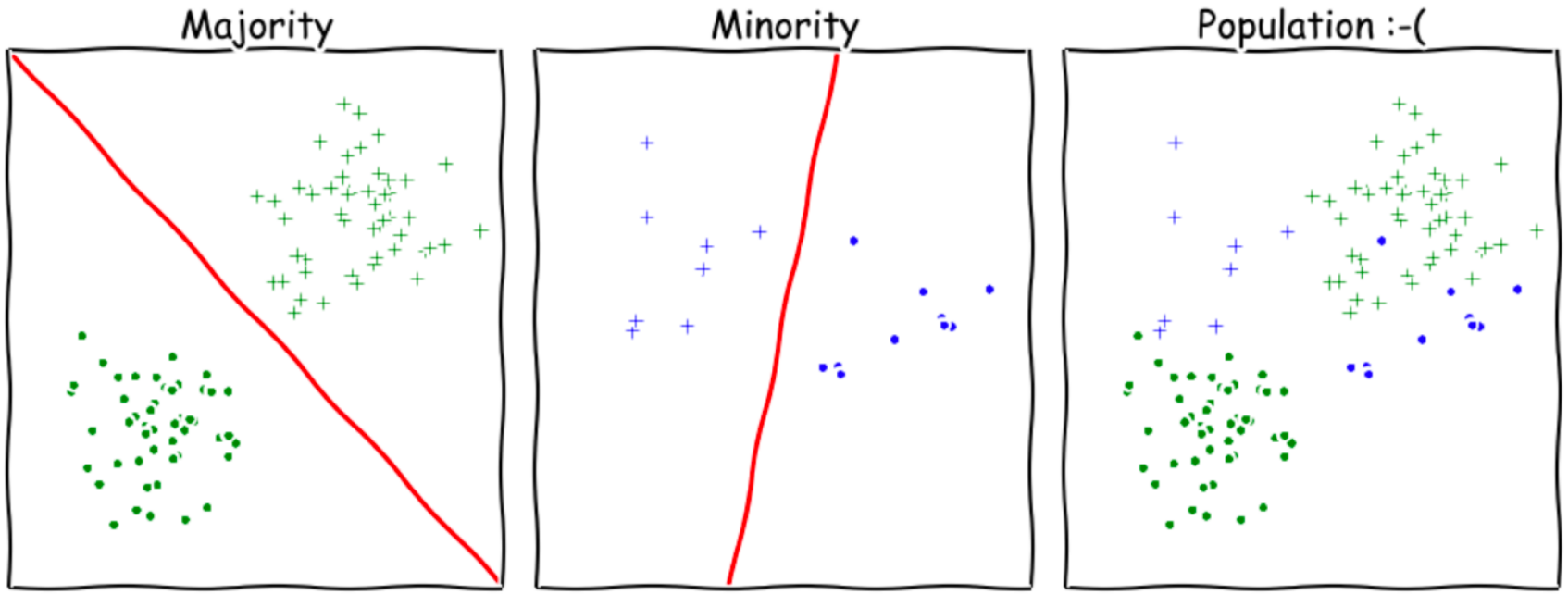
Statistical patterns that apply to the majority might be invalid within a minority group

Fewer people (if any) carry the same name in some ethnic groups

- **A typical sign of a 'fake' profile in Anglo societies**

Source: Moritz Hardt, "How big data is unfair"

When Big Data Becomes Small



Source: Moritz Hardt, “How big data is unfair”

What is My Responsibility?

Domain knowledge should not be ignored

Understand the data

Investigate where your model does not perform well

- Break into separate models
- Up-sample minorities
- More complex decision rules (use with caution)
- Feed back model performance

Bias In = Bias Out

Can Machine Learning Be Biased?

In the news

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours
Telegraph.co.uk - 5 hours ago
To chat with **Tay**, you can **tweet** or DM her by finding **@tayandyou** on **Twitter**, or add her as a ...

Microsoft Releases AI Twitter Bot That Immediately Learns How To Be Racist
Kotaku - 3 hours ago

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.
New York Times - 3 hours ago

Bias In = Bias Out

Problem statement:

- Match resumes to jobs to get the right candidates in the right roles

Where to start?

- Large text dataset
- Word embeddings from process such as word2vec

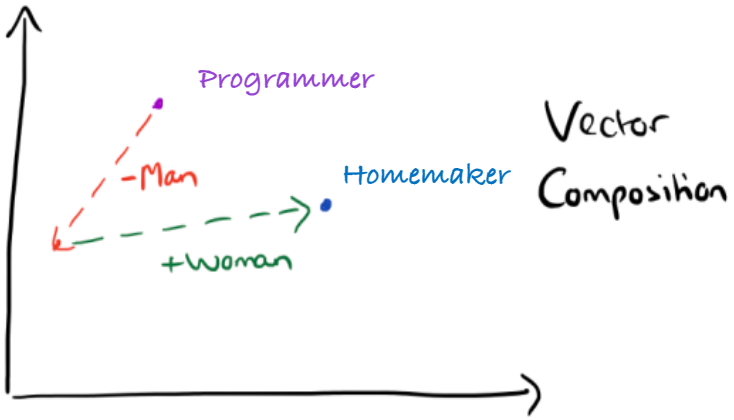
This could take a while…

- https://code.google.com/archive/p/word2vec/
- Pre-trained vectors based on Google News dataset (~ 100 billion words)
- 300-dimensional vectors for 3 million words

Let's get cracking!

Bias In = Bias Out

Reflects linguistic word relationships



Reflects societal biases

Source: http://www.wordbias.org/

What Is My Responsibility?

Awareness of how the model recommendations are implemented

Critical evaluation of model performance “in the wild”

Apply techniques to reduce/remove bias

- Additional data sources
- Sample dataset to remove bias – double-edged sword
- Transparency of techniques

The Devil Is In The Question

The Devil Is In The Question

The Devil Is In The Question

Problem statement:

- Improve user engagement as measured by clicks and shares

People like similarity

- Movie recommendations
- Search

Similarity drives user engagement

- More of what I like
- More of what my friends like
- More from the organisations I like

Success!

The Questions We Ask Matter

Media is driven by likes, clicks and recommender systems

Traditional goal of media has been unbiased content

- Point and Counter-Point
- How to differentiate ‘right to live’ from ‘right to choose’?

Fake news works when it is close to our beliefs

Are we asking the right question?

What Is My Responsibility?

Our *most important* job is to design the right question

Data and methods should fit the question

Our responsibility extends beyond the delivery of the model

Be The Change

The *Social* Data Scientist

- With great power comes great responsibility
- Understand the *context* and *impact* of our work
- Put questions and data above methods and tools
- Diversity in Data Science teams

What will our legacy be?