COMP5310: Principles of Data Science

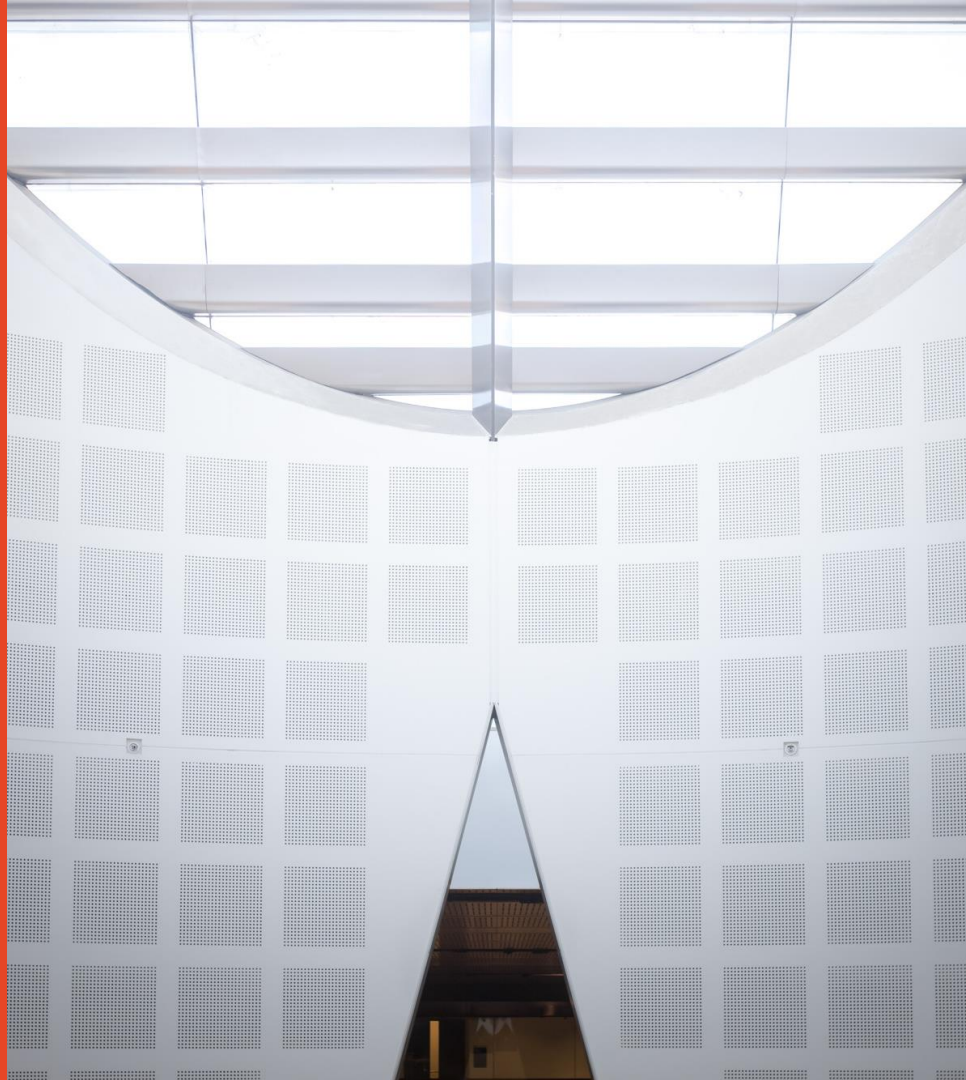
W12: Product Thinking & Ethics

School of Computer Science

Based on slides by previous lecturers of this unit of study

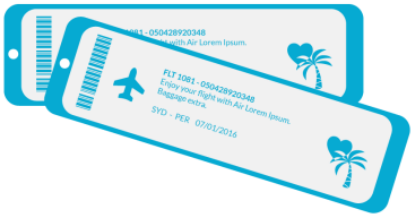
The University of Sydney

THE UNIVERSITY OF SYDNEY



Product Thinking

A product is a good or service offered to customers



A product can be:

- **A good:** a tangible or virtual item that provides certain benefits for a customer.
  A delicious cup of coffee is a product.
- **A service:** a single or package of activities performed to fulfill benefits for the customer.
  A professional haircut is a service.

https://roadmunk.com/blog/what-is-a-product/

Define the problem before the solution

First define the problem...

User problem: What problem do we solve?
Target audience: For whom are we doing this?

Vision: Why are we doing this?
Strategy: How are we doing this?
Goals: What do we want to achieve?

Only then does it make sense to think about the solution

Relationship of data scientist to product

Model 1: Data scientist as an ***owner***

Model 2: Data scientist as a ***service***

Model 3: Data scientist as a ***partner***

Adapted from: http://fututorial.weebly.com/

Video: https://www.youtube.com/watch?v=v2MJypl51zk

Data Scientist as Owner of Product

Operates in a “hacky way” (early stage companies)

Key steps in business rely heavily on data

- Recommender
- Relevance
- Matching
- Scoring

Mostly backend, relatively stand-alone features

Data Scientist as a Service

Engagement is “on-demand”, project-based

Examples:

- some of the BI roles
- strategy roles (consultancy)
- data API for product
- modeling for specific purposes: propensity to {x}, where x: {buy, attrite, convert, etc.}

Data Scientist as a Partner

Plays active role in every stage of Product Life Cycle

Shares the ultimate goal of product success

Often requires an embedded engagement model

Possibilities and impact

Data sense: What is possible?

Product sense: What is valuable?

Different roles require different balance of abilities

All data science should deliver on value proposition

The Ethical Data Scientist

THE UNIVERSITY OF SYDNEY

What are ethics?

"Ethics are the moral principles that govern a person's behaviour or the conducting of an activity"

http://www.oxforddictionaries.com/definition/english/ethics

Why is it a data scientist's job?

Consider:

- User behaviour data forms the foundation of data products
- Products assist users but may also influence their behaviour
- e.g. ranking algorithms, recommendation systems, friend suggestions.

Models/algorithms not only predict but affect the future

- This is both incredibly exciting and absolutely terrifying

Examples: Preventative Policing & User Tweaking

Stopping crime before it happens

- Chicago police use predictive modelling to create a heat list
- Pre-emptively approach people likely to commit violent crime
- Discuss risks. What features should/not be used?

Inducing emotional states

- A 2014 study explored whether user mood is contagious on Facebook
- Manipulated feeds to include fewer positive or negative posts
- Discuss risks. How should users be protected?

Questions of ethics: Preventative policing

How avoid perpetrating potentially unfair or damaging stereotypes/profiling present in the data?

How use information positively and manage potential prediction mistakes?

Baldridge. Machine learning and human bias: and uneasy pair. http://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an- uneasy-pair/

Questions of ethics: User Tweaking

This is an interesting study but emotional well-being should not be treated lightly.

How avoid distress and support users who may be negatively impacted?

Who will take responsibility for data ethics inside companies if not data scientists?

Baldridge. Emotional contagion: contextualizing the controversy. http://go.peoplepattern.com/blog/emotional-contagion-one/

Subprime Mortgages

What does modelling have to do with it?

Increase in risky lending
– Home loans close to the actual value of the property
– Aggressive sales of expensive, complex products

Many institutions and investors exposed through

Credit default swaps (CDS)
Mortgage-backed securities (MBS)
Collateralised debt obligations (CDOs)

Much exposure due to poor understanding of bad risk models

Poor assumptions underlying crisis

1. Housing prices would not fall dramatically;
2. Free and open financial markets supported by sophisticated financial engineering would most effectively support market efficiency and stability, directing funds to the most profitable and productive uses;
3. Concepts embedded in mathematics and physics could be directly adapted to markets, in the form of various financial models used to evaluate credit risk;
4. Economic imbalances, such as large trade deficits and low savings rates indicative of over-consumption, were sustainable;
5. Stronger regulation of the shadow banking system and derivatives markets was not needed.

Statistics is a shallow description

"Over-reliance on probability and statistics is a severe limitation. Statistics is shallow description, quite unlike the deeper cause and effect of physics, and can't easily capture the complex dynamics of default."

- Emanuel Derman and Paul Wilmott, *The Financial Modelers' Manifesto*

It is our models that are simple, not our world

"Our experience in the financial arena has taught us to be very humble in applying mathematics to markets, and to be extremely wary of ambitious theories, which are in the end trying to model human behaviour. We like simplicity, but we like to remember that it is our models that are simple, not the world"

- Emanuel Derman and Paul Wilmott, *The Financial Modelers' Manifesto*

“All models sweep dirt under the rug. A good model makes the absence of the dirt visible.”

- Emanuel Derman and Paul Wilmott, *The Financial Modelers’ Manifesto*

Machine Learning and Human Bias

What we teach machines

Tay.ai is a great example of how a machine can adopt human bias.

What if it was making decisions instead of chat.

We are responsible not just for the algorithms, but also for the influences we provide

Scenario: Optimise services to Homeless Families

Imagine our goal is to match homeless families with the most appropriate services

We have historical data with various characteristics: number and age of children and parents, zip code, number and length of previous stays in homeless services, race.

Which characteristics should we use?

What about race?

Now imagine we find that including race makes the model more accurate.

Should we use it?

No

Remember algorithm output will be used to help pair families with services.

What about race?

Using historical data means that we are “training our model” on data that is surely biased, given a history of racism

An algorithm cannot see the difference between patterns that are based on injustice and patterns that are based on traffic

So choosing race as a characteristic in our model would be unethical

What about race?

This example is taken from a real project carried out by New York City Health and Human Services

> "Looking at old data, we might have seen that families with a black head of household was less likely to get a job, and that might have ended up meaning less job counseling for current black homeless families. In the end, we didn't include race." –Cathy O'Neil

http://www.slate.com/articles/technology/future_tense/2016/02/how_to_bring_better_ethics_to_data_science.html

END