A Glimpse of NLP in Industry

Bo HAN (bo.a.han@accenture.com.au)
24/05/2021

Outline

- My Journey & motivations (5 mins)
- Use Case: Geolocation Prediction (20 mins)
- Academia and Industry comparisons (5 mins)
- NLP landscape in industry applications (10 mins)
- Mindset for Industry (10 mins)
- Questions and Answers (10 mins)

My Journey with NLP

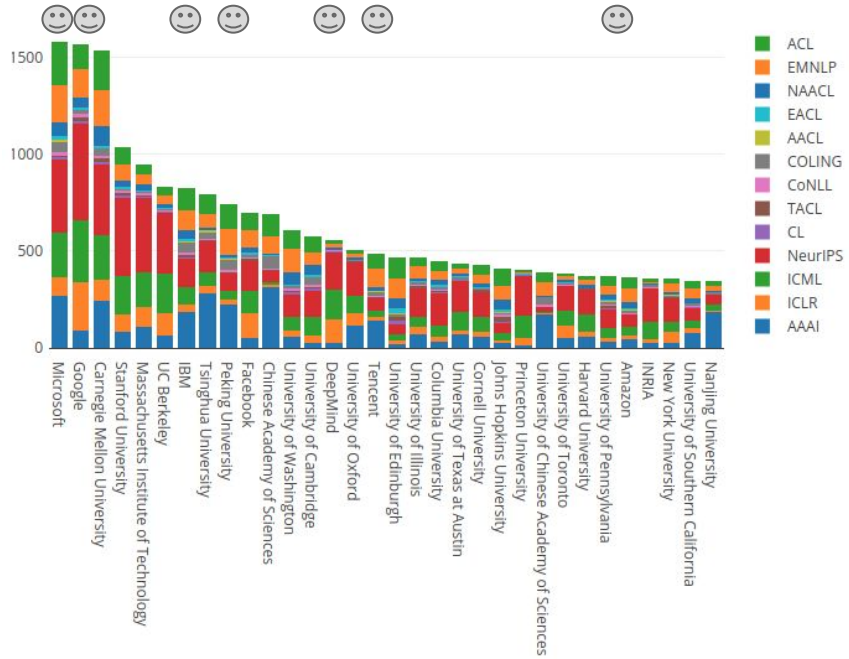
Industry Research Institutions: Microsoft Research Asia (2007-2009), IBM Research Australia (2014-2016)

Universities: University of Melbourne/NICTA (2010-2014)

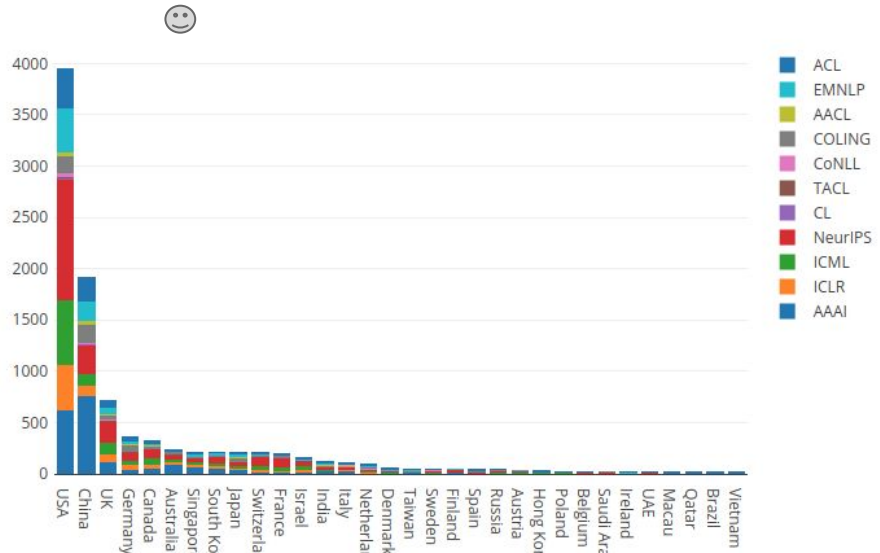
Professional Firms: Start-up (2016-2017), Kaplan (2017-2018), Accenture (2018-now)

Why should I care NLP/ML in industry?

Papers per organisations (2012-2020)

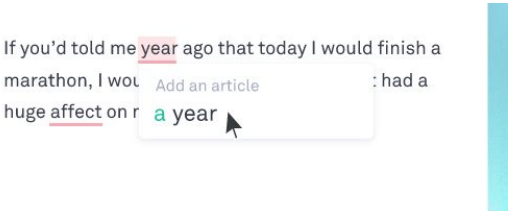


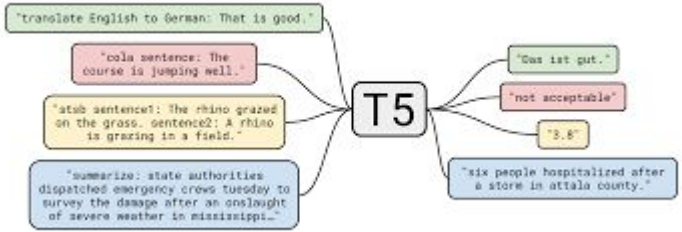
Papers per country/region (2020)
(Australia ranked 6th)

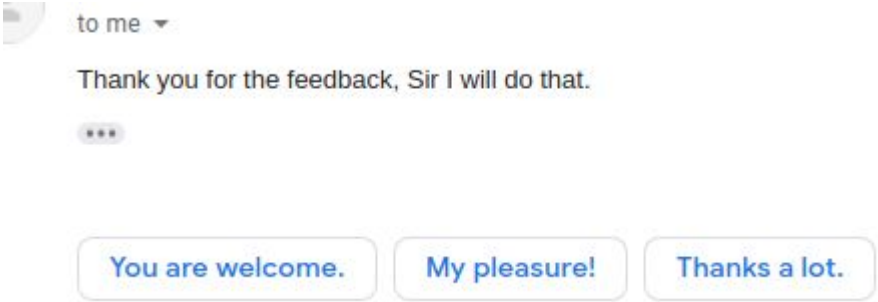


http://www.marekrei.com/blog/ml-and-nlp-publications-in-2020/

Why should I care NLP/ML in industry?







Pictures are from Google Image Search with URL embedded

Case study: Geolocation Prediction

Game time: Can you guess the city?

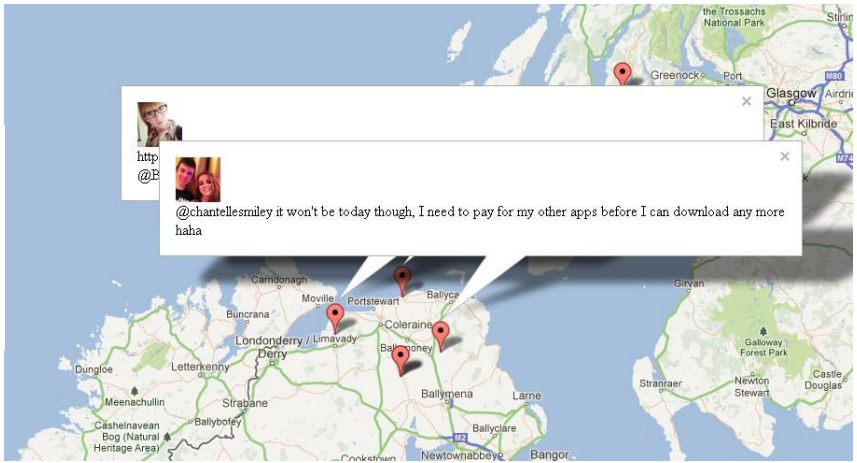
Text-based Geolocation Prediction

Assign a unambiguous geographical location to a piece of text

Input: text data, e.g. an English tweet

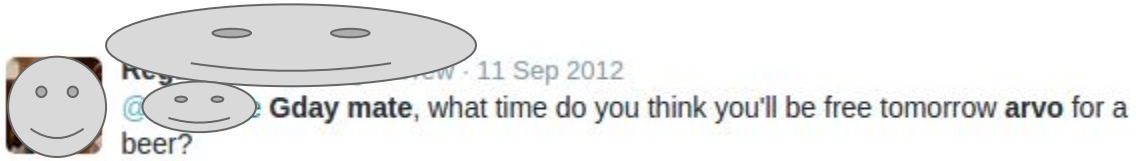
Output: one of metro cities across the world, e.g. London, Sydney, New York

Task: A multi-class classification task

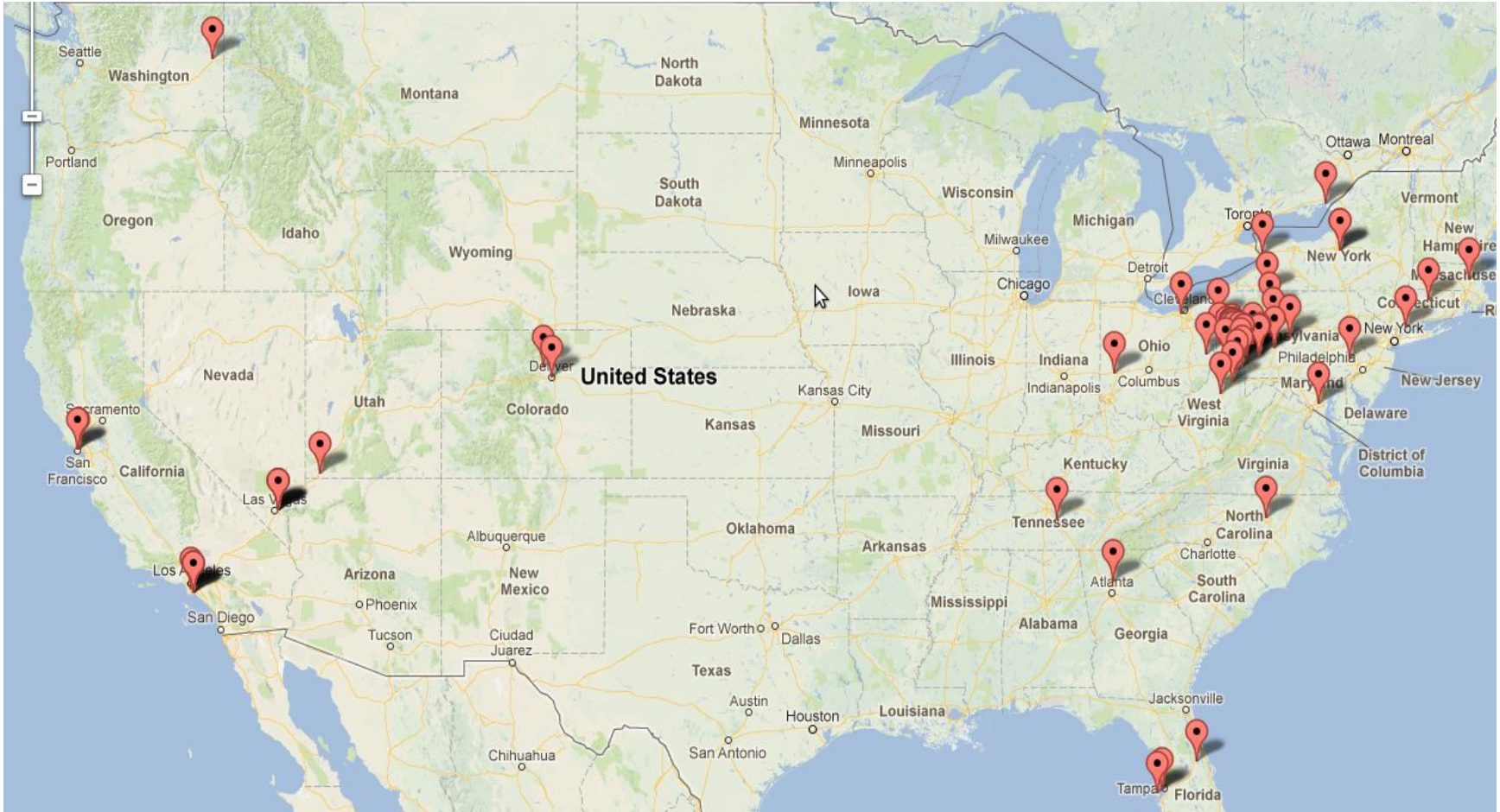


Hypothesis: Words carry varying amount of geolocation information

- Gazetted terms: Australia, Canada, London, Seattle,
- Local sports: hockey, footy, cricket
- Dialectal words: arvo, yinz, howdy
- Geo entities: tube, tram, skyscraper, ferry



Local Words: yinz

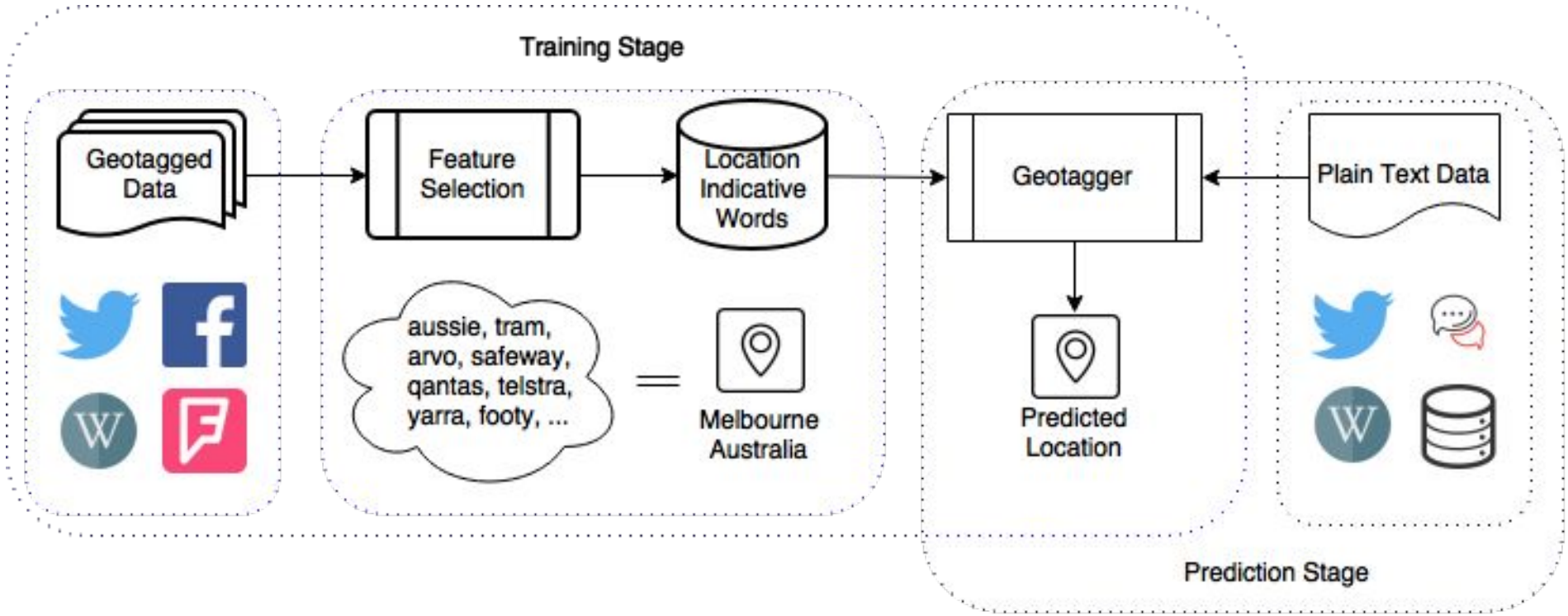


Somewhat Local Words: ferry

Common Words: today

Geolocation Prediction from Academia View

A Text-based Geo Prediction Framework



Text-based Geo Prediction (Academia)

Q: How to find Location Indicative Words? **(LIW)**

Q: How to measure model prediction accuracy? **(Evaluation)**

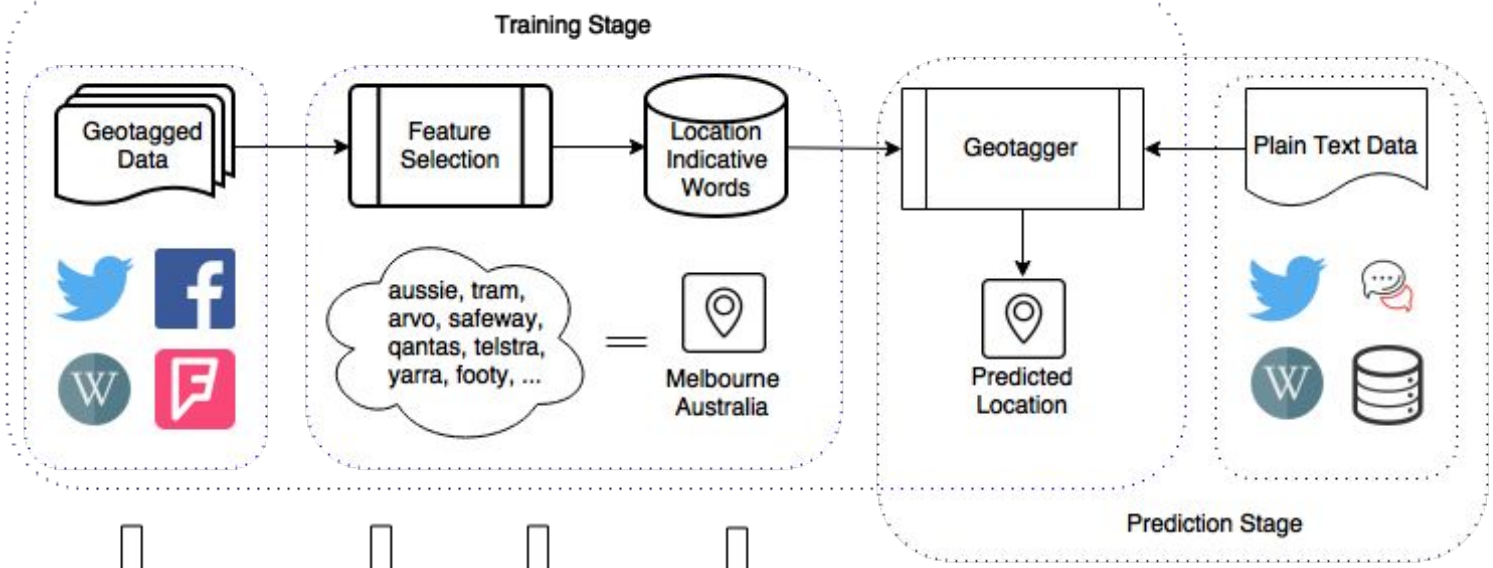
Q: What are suitable classifiers for this multi-classification? **(ML Model)**

Q: How does input size (i.e. amount of text data) affect the accuracy? **(Data)**

Q: Will my prediction model accuracy decrease over time? **(Generalisation)**

Q: Will language, metadata, text-derived network relations affect model accuracy? **(NLP)**

...



Data

Data

Geo-tweets vs. Geo-users

Features

(NLP)

English only vs. All Languages

Metadata

Evaluation

Data Partitions

City vs. Earth grids vs. Hierarchical

Feature Selection

LIW

Statistical Methods
Heuristic Methods
Information theory-based Methods

Multi-Classifiers

ML Model

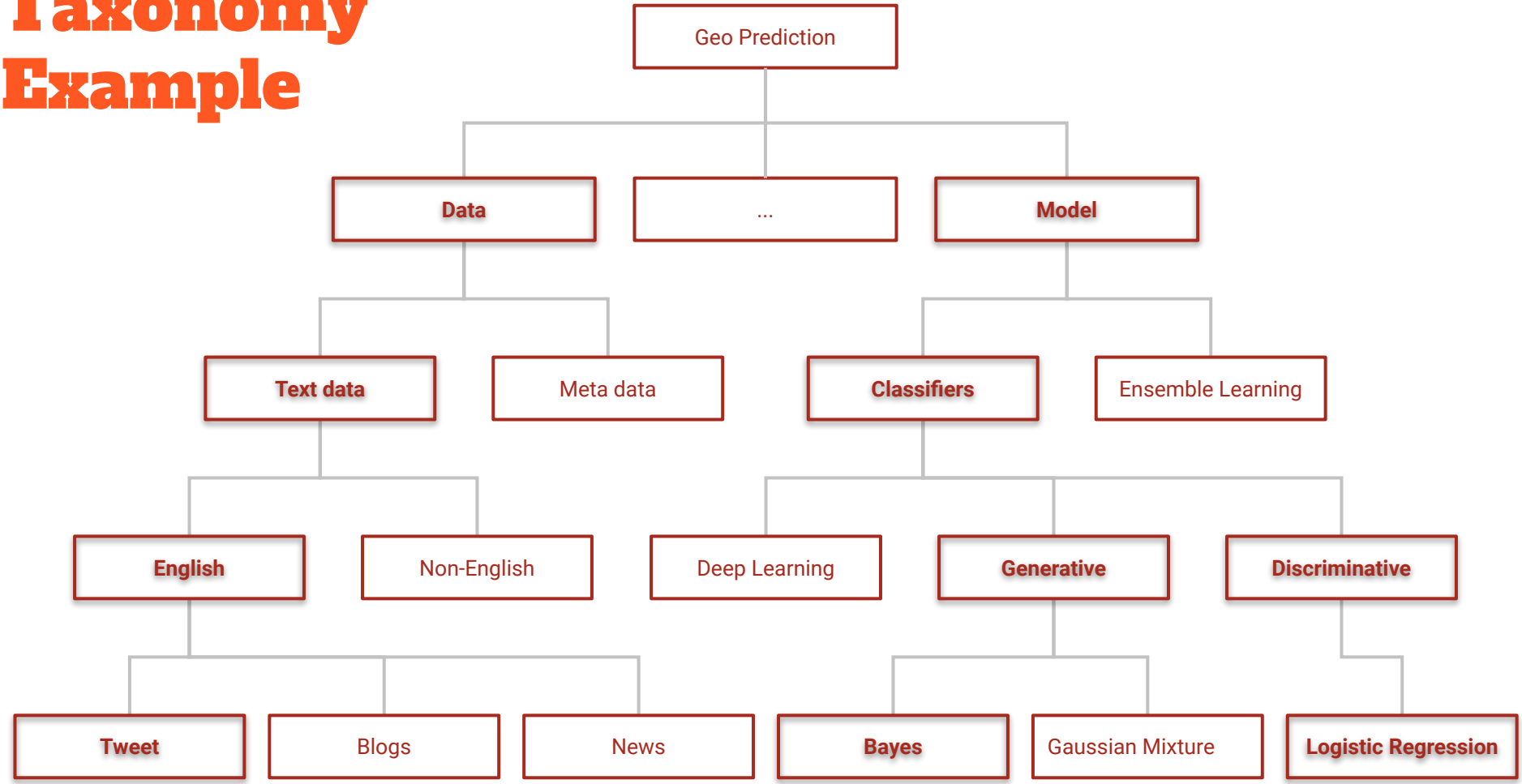
Simple (Naive Bayes) vs. Advanced (MaxEnt)
Generative (Naive Bayes) vs. Discriminative (LR)
Ensemble learning

Temporal Variance

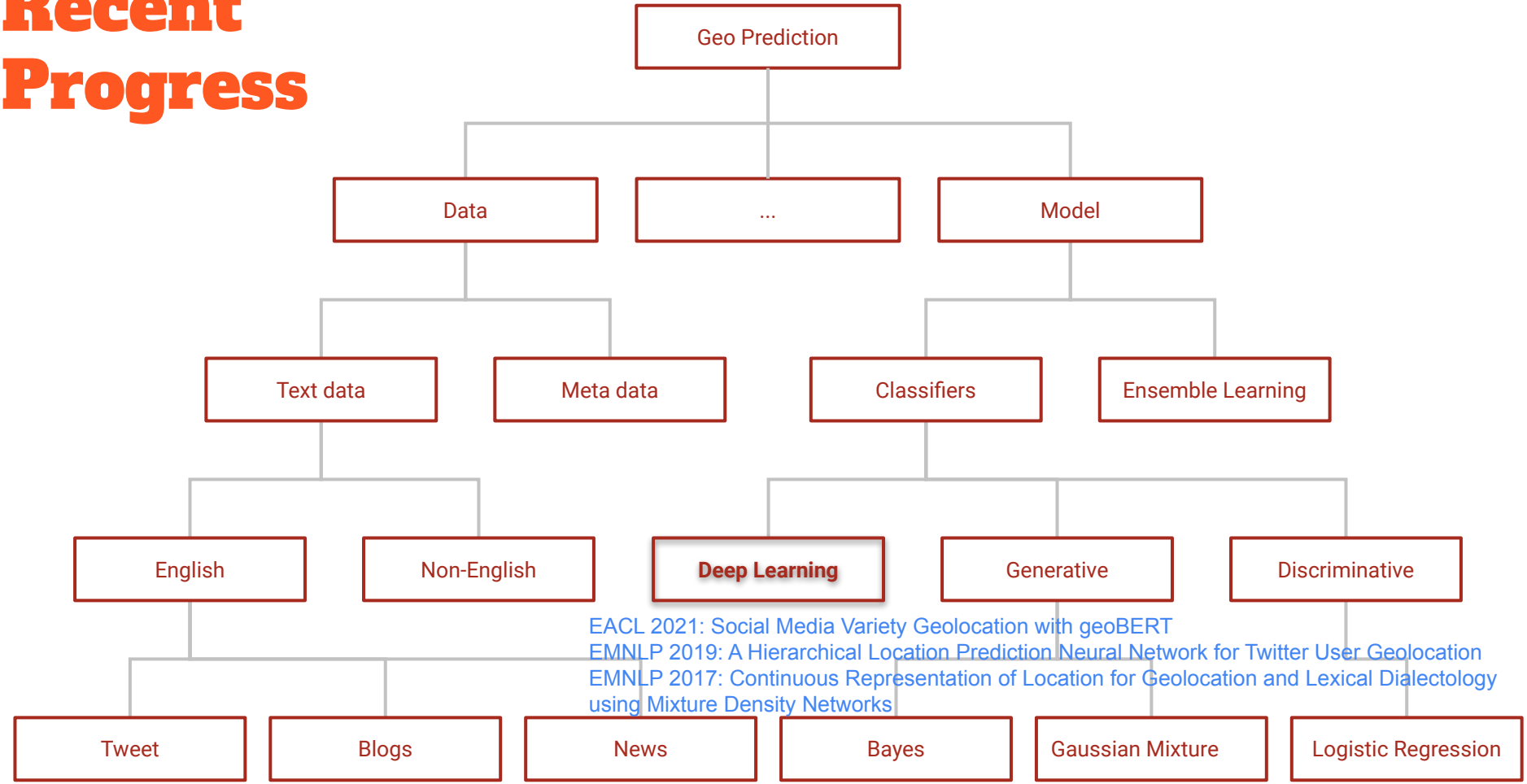
Will trained classifer be less effective over time?

Generalisation

Taxonomy Example

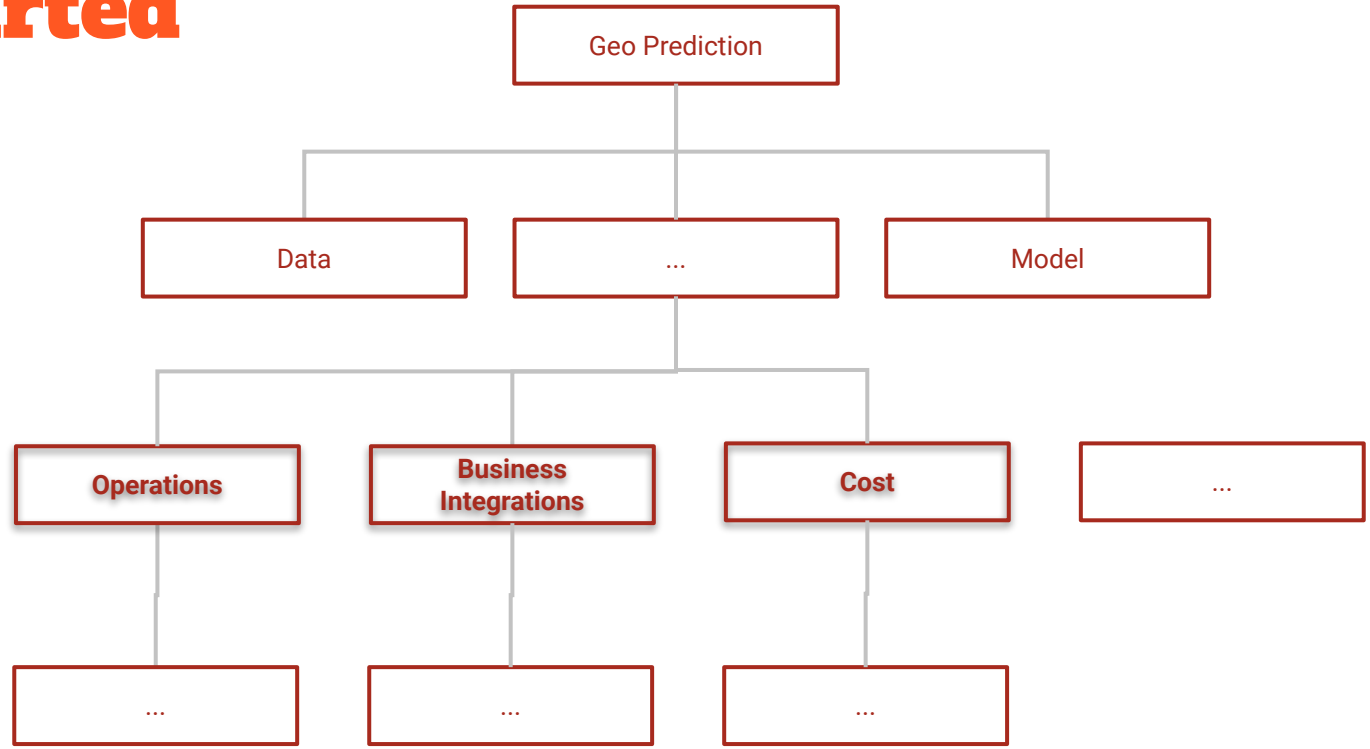


Recent Progress



EACL 2021: Social Media Variety Geolocation with geoBERT
EMNLP 2019: A Hierarchical Location Prediction Neural Network for Twitter User Geolocation
EMNLP 2017: Continuous Representation of Location for Geolocation and Lexical Dialectology using Mixture Density Networks

Uncharted


Geo Prediction
Data
...
Model
Operations
Business Integrations
Cost
...
...
...
...

Geolocation Prediction from Industry View

Text-based Geo Prediction (Industry App)

~~Q: How to find Location Indicative Words? **(LIW)**~~

Q: How to measure model prediction accuracy? **(Evaluation)**

~~Q: What are suitable classifiers for this multi-classification? **(ML Model)**~~

~~Q: How does input size (i.e. amount of text data) affect the accuracy? **(Data)**~~

Q: Will my prediction model accuracy decrease over time? **(Generalisation)**

~~Q: Will language, metadata, text-derived network relations affect model accuracy? **(NLP)**~~

...

Text-based Geo Prediction (Industry App)

Q: How to measure model prediction accuracy? **(Evaluation)**

Q: Will my prediction model accuracy decrease over time? **(Generalisation)**

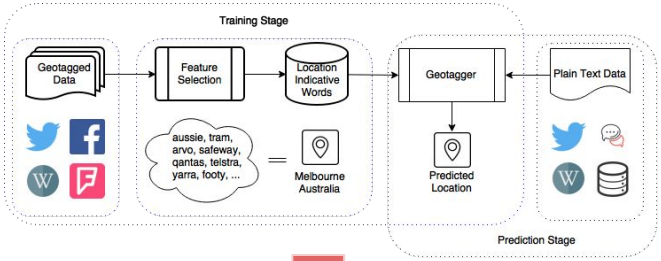
Q: What business service/product can leverage this service? (Utility)

Q: What is the throughput of this deployed service? (Performance)

Q: What are ethics/data privacy/... risks? (Risk)

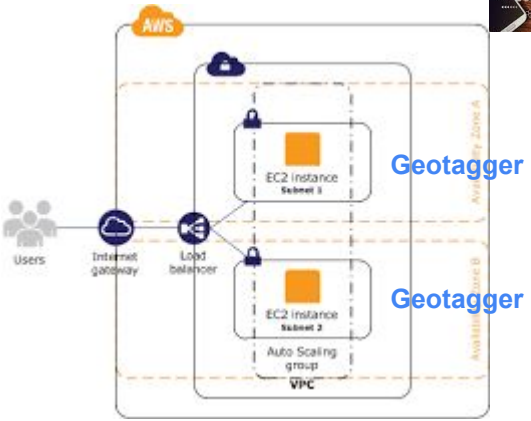
Q: Should we apply a patent or keep it as a business secret? (IP)

...



Regulations

High Availability

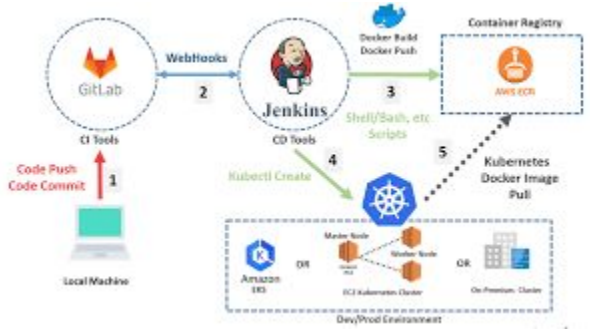


Geotagger

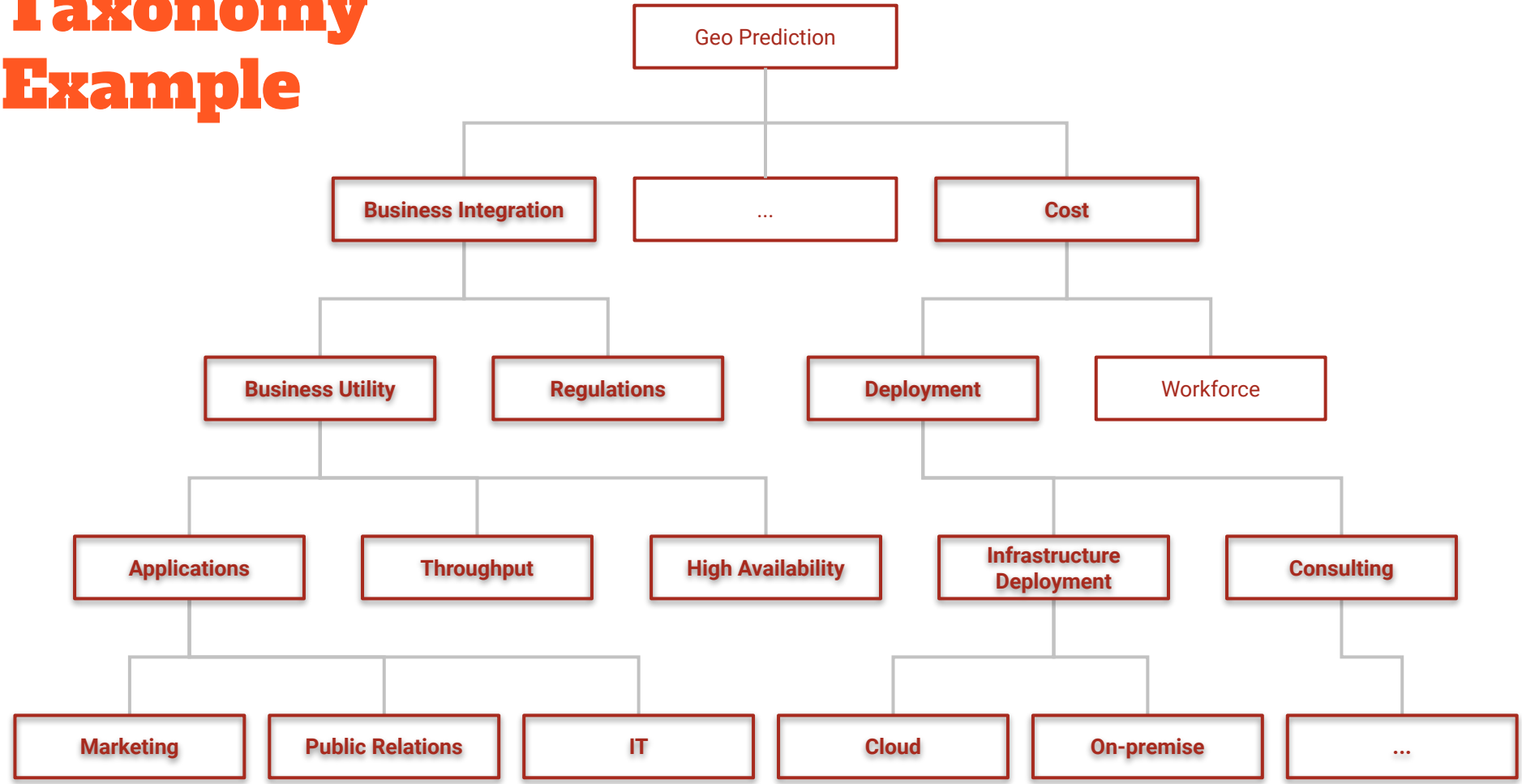
Geotagger

Data Lake

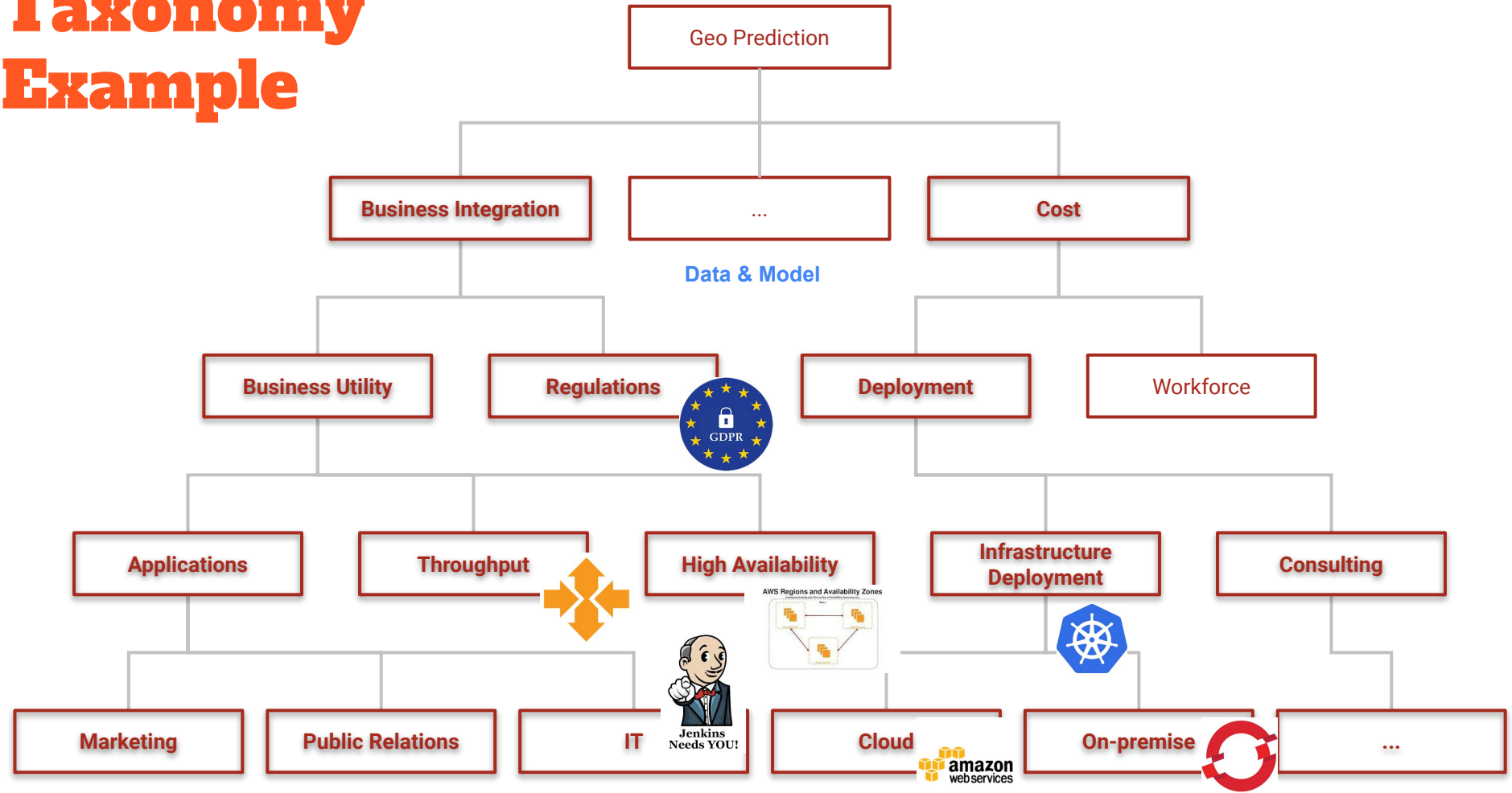
DevOps:
Version Control: Git/Bitbucket
CICD: Jenkins/Bamboo
Project Management: JIRA/Trello
Containerisation: Docker/K8S
Full Stack: ...



Taxonomy Example



Taxonomy Example



A Pilot Comparison

Academia: 

- Broaden the **human knowledge boundaries**, e.g., *improve accuracy from X% to Y% where Y > X and the result is statistically significant*
- It is typically driven by **research questions**
- Work output: **publications**
- Typical activities:
  - Literature review (required)
  - Experiments (required)
  - Publish papers (required)
  - Understand relevant work (required)
  - ...
  - **A working demo website (optional)**

Industry: 

- Mostly about **applications**, e.g., *apply sentiment analysis to collect customer feedback and improve our products.*
- It is typically driven by **business needs**
- Work output: **business application**
- Typical activities:
  - A working PoC demo (required)
  - Deployment (required)
  - Cost estimation (required)
  - Information security (required)
  - Regulation requirements (required)
  - ...
  - **Utilise state-of-the-art result from academia (required)**
  - **Papers (optional) and other IPs (required)**

Benefit from Mutuals

Benefit from mutuals (Industry -> Academia)

Academia: 

- **Business need** is a good (but not the only) source for your research topic

(Hypothetical) business need:
A small cafe short staffed

Automated Speech Recognition (ASR)
Text to Speech (TTS)
Neural networks
...

https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Benefit from mutuals (Industry -> Academia)

Academia: 

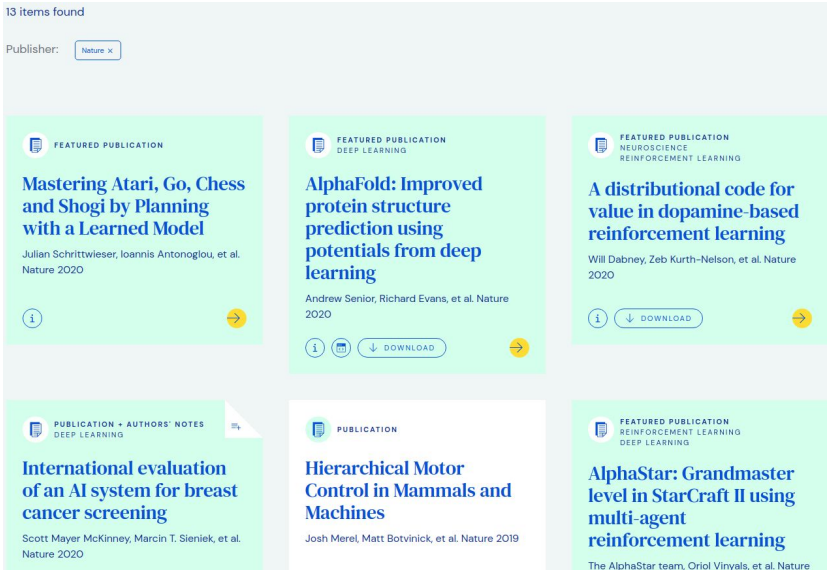
- Research with clear or potential **business applications** may get more funding
- Yahoo! Key Scientific Challenges Program
- Microsoft Faculty Fellowship
- Google Faculty Research Awards in NLP and other fields
- ...

https://www.netflix.com/ and https://bit.ly/3oaQVF9

Benefit from mutuals (Industry -> Academia)

Academia: 

- An increasing number of **key research papers** are from industry research labs



https://deepmind.com/research?filters_and=%7B%22publisher%22:%5B%22Nature%22%5D%7D

Benefit from mutuals (Academia -> Industry)

Industry: 

- Obtain **state-of-the-art algorithms and models** from academia
  - **LSTM**: Sepp Hochreiter; Jürgen Schmidhuber (21 August 1995), Long Short Term Memory
  - **Expectation-maximization algorithm**: Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". Journal of the Royal Statistical Society, Series B. 39 (1): 1–38. JSTOR 2984875. MR 0501537.
  - **Viterbi algorithm**: Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory. April 1967, 13 (2): 260–269
  - ...

Benefit from mutuals (Academia -> Industry)

Industry:

- Software, data and other resource **free to use for commercials**

	Apache	BSD	MIT	GPL	LGPL	AGPL
Type	Permissive	Permissive	Permissive	Copyleft	Copyleft	Copyleft
Provides copyright protection	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Can be used in commercial applications	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Provides an explicit patent license	✓ TRUE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE
Can be used in proprietary (closed source) projects	✓ TRUE	✓ TRUE	✓ TRUE	✗ FALSE	✗ FALSE partially	✗ FALSE for web
Popular open-source and free projects	Kubernetes Swift Firebase	Django React Flutter	Angular.js JQuery, .NET Core Laravel	Joomla Notepad++ MySQL	Qt SharpDevelop	SugarCRM Launchpad

https://moqod.com/understanding-open-source-and-free-software-licensing/ and https://www.wikipedia.org/

NLP Landscape in Industry

Two Key Factors

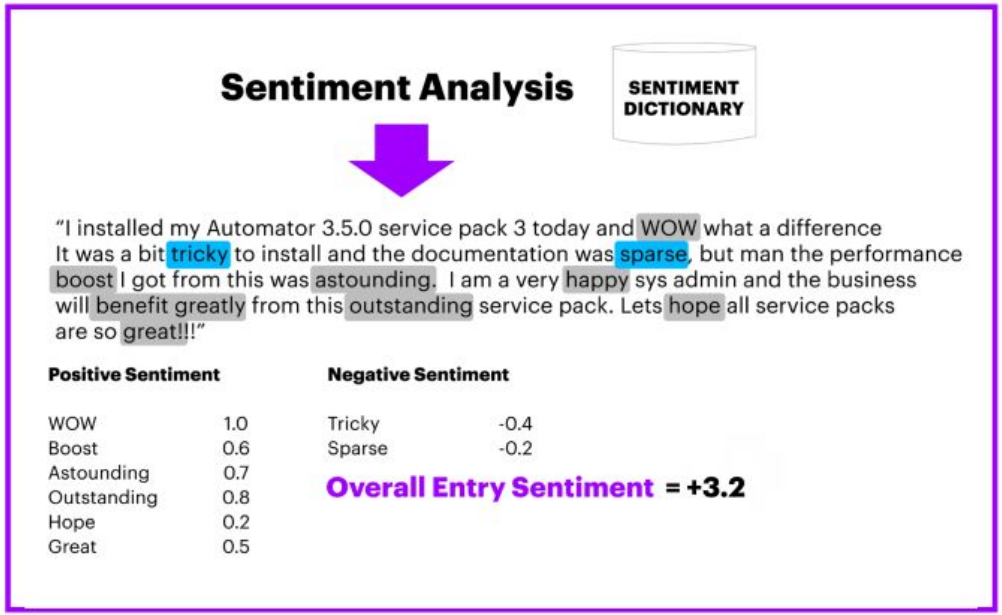
Cost

Revenue

NLP Applications in Industry

Sentiment Analysis to identify people's opinions or feelings towards a product/service to collect customer feedback and unlock potential actions

- Provide marketing and competitive intelligence
- Enhance product development
- Improve customer retention
- Analyze the impact of an event (e.g. a product launch or redesign)

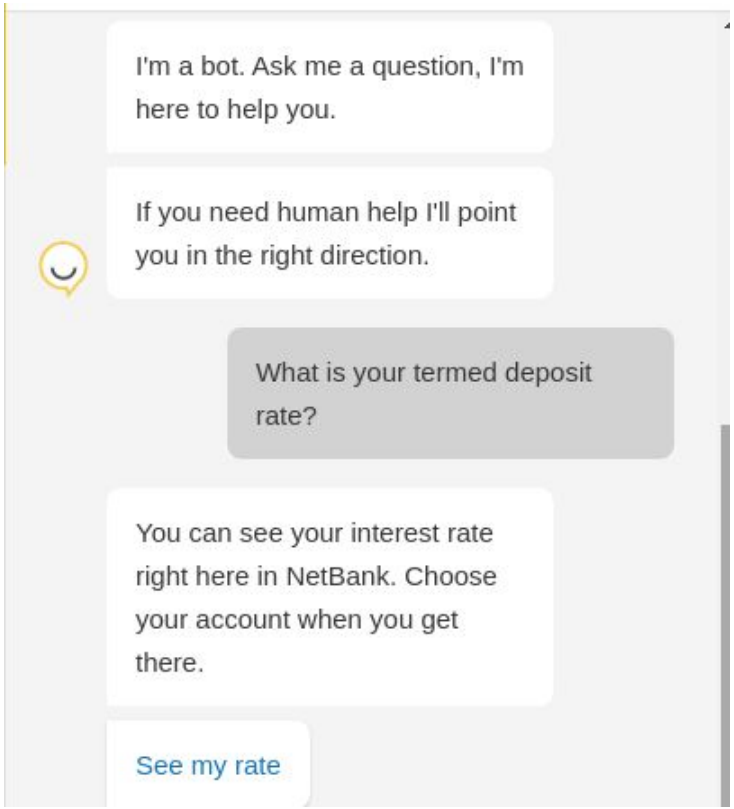


Ref: Top Natural Language Processing Applications in Business (Accenture)

NLP Applications in Industry

Chatbots (Virtual Assist) enable conversations between computers and customers to help customers seek relevant information or perform a specific task.

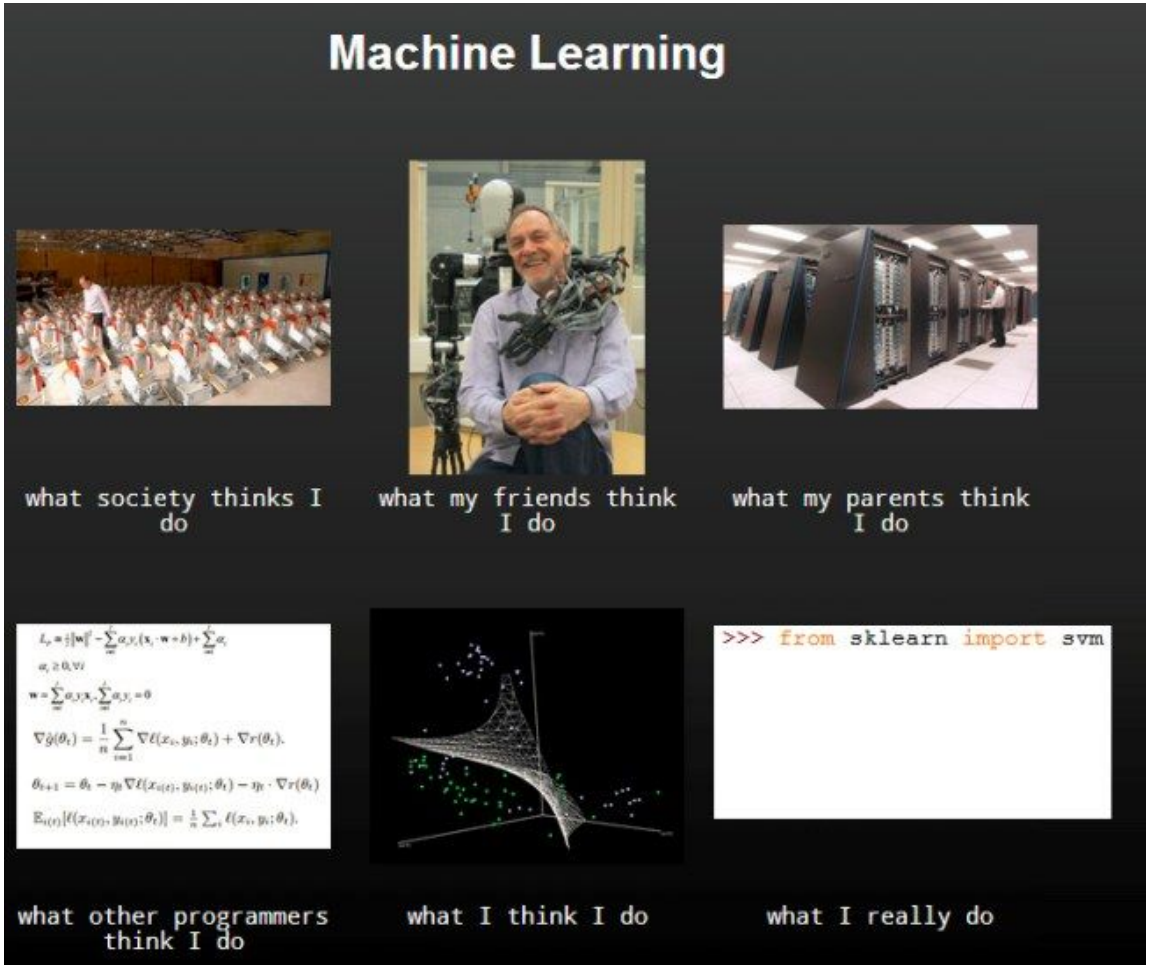
- Improve business processes and reduce support costs
- Enhance search and knowledge-seeking experiences
- Human-in-the-loop to compensate bad experience



Ref: Top Natural Language Processing Applications in Business (Accenture)

Mindset for Industry

NLP/ML Jobs in Industry (application)



https://www.kdnuggets.com/2017/04/cartoon-machine-learning-what-they-think.html

Example: Lower Customer Churn

Customer Service: Hi XXX, you recently cancelled the contract with us, I have a good deal for you

Customer Churn:
A customer leaves a company

https://miro.medium.com/max/1600/0*dzmm3qresODlScte and Analytical Skills for AI and Data Science

Lower Customer Churn Step 1

Business Question

1. Question: Can I lower the churn rate in my company?
2. Motivation:
   a. Customer churn will impact our revenue
   b. It will affect our long term growth and eventually our leader position in the market
   c. ...

Lower Customer Churn Step 2

Analysis

1. How many customers are we losing?
2. Who are they?
3. Are all customers the same?
4. Can I collect information that characterise customers
5. ...

Lower Customer Churn Step 2

Analysis

1. How many customers are we losing? **5% in a month**
2. Who are they? **New joiners during previous promotions**
3. Are all customers the same? **No**
4. Can I collect information that characterise customers? **Service, usage statistics, ...**
5. ...

Lower Customer Churn Step 3

Data Science Prediction

Background work:

- Data ETL (data collection, cleansing, validation, loading)
- Data modelling (a classification or a regression task)
- ...

Delivery model:

- Input: a customer's information
- Output: when this customer will leave the company

Lower Customer Churn Step 4

Actionable Insight

If those customer are going to leave,

- What retention policies should I use?
- How should I assign to them?
- Can we further segment those customers into subgroups for different policies?
- Based on your retention model, what would be the long term profits (after subtracting the retention cost)?

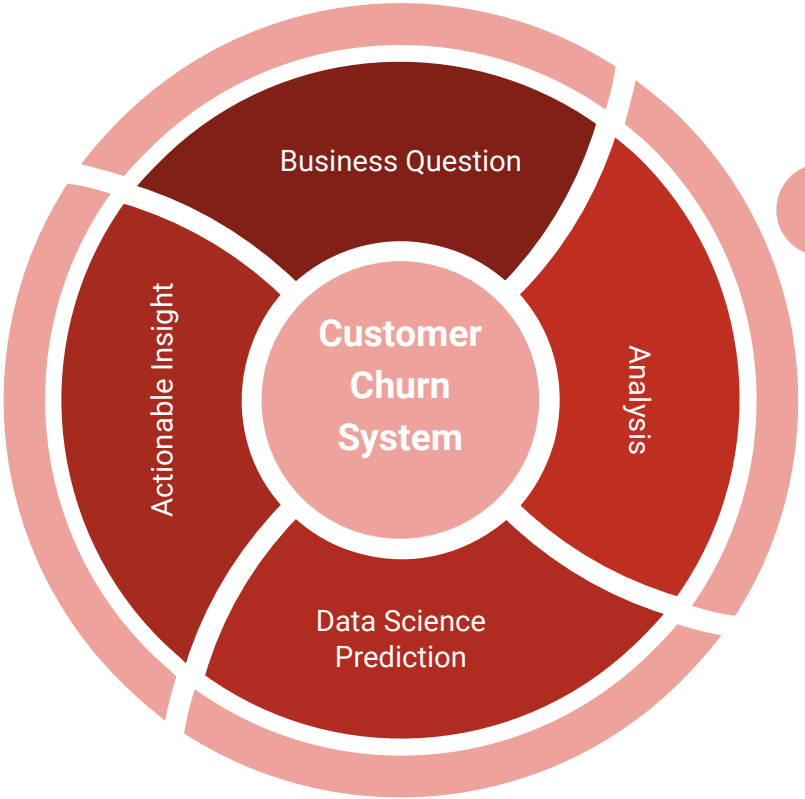
Lower Customer Churn Step 4

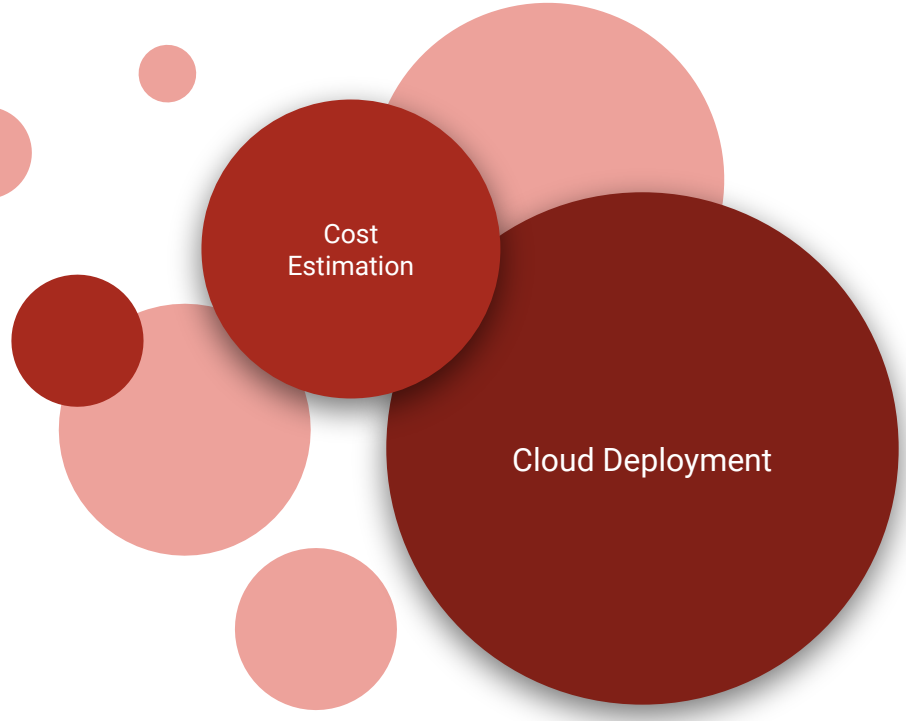
Actionable Insight

If those customer are going to leave,

- What retention policies should I use? **One month free, bonus gift card, ...**
- How should I assign to them? **Emails, mails, ...**
- Can we further segment those customers into subgroups for different policies? **Yes, based on their usage plan, we can ...**
- Based on your retention model, what would be the long term profits (after subtracting the retention cost)? **1M AUD this year**

Lower Customer Churn Loop





Recommended Practise

- Practise 1: Fast Food Store Locations
  - Given budget X, where should I select the location for my new store to maximum my profits?
- Practise 2: Who Should I Hire?
  - I need to fill a positions with X, Y, Z requirements, who should I hire?
- Guess techniques:
  - How would you implement an App that has ML/NLP components in your mobile phone?
  - Company X just released service Y, what are the underlying techniques they need to deliver and operate that service?
- Guess applications:
  - Where can AlphaGo and its variations algorithms apply?

A few more words to say

- Ask Alumni Service: https://www.unimelb.edu.au/alumni/get-involved/volunteer/ask-alumni
- Github, personal website or other public presence of your work
- Tech Meetups (a mixture of industry practitioners, researchers, hobbyist)
- Online Course: Coursera, Udacity, O'Reilly...
- Beginner class for cloud computing: AWS Cloud Practitioner
- ....