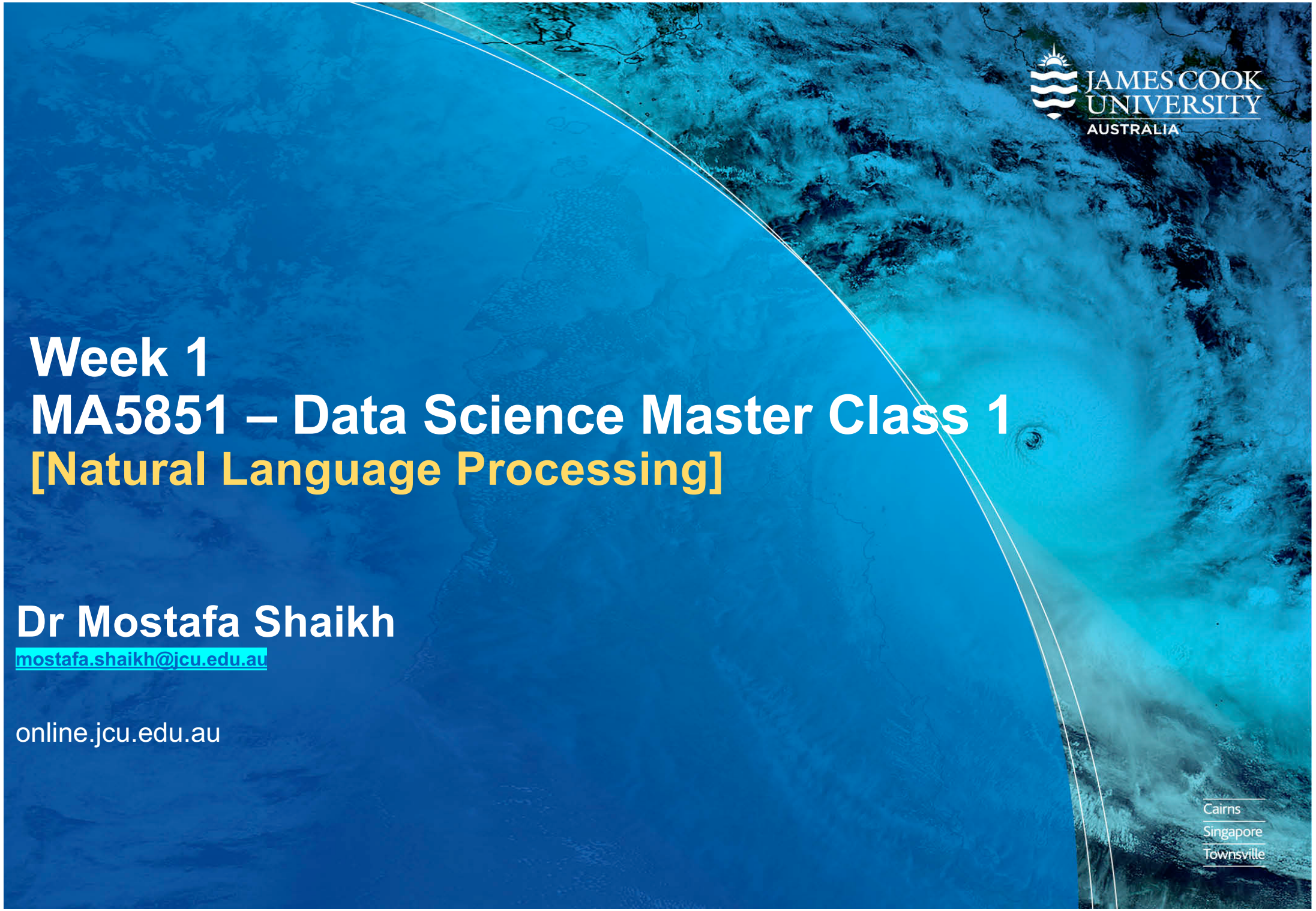

JAMES COOK UNIVERSITY
AUSTRALIA
Week 1
MA5851 – Data Science Master Class 1
[Natural Language Processing]
Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au
online.jcu.edu.au
Cairns
Singapore
Townsville

Agenda

- Tutor intro, announcement
- Brief outline of the course
- Brief outline of assessments
- Introduction to NLP
- Week 1 – recap
- Week 1 - SLP

JAMES COOK UNIVERSITY AUSTRALIA

Tutors

- Mostafa (mostafa.shaikh@jcu.edu.au)
- Mateen (mateen.moiz@jcu.edu.au)
- Shah (shah.ehoque@gmail.com)

About me:

- PhD in AI (Affect Sensing from Text), MS in CS
- Software Engineer/Senior Software Engineer (6 years)
- Data Engineer/Data Analyst/Data Scientist (8 years)
- Senior Lecturer/Lecturer (5 years)

Announcement

- Quiz 1 new due date 21 March 2021
- Completion of SLP
- One collaboration session per week

JAMES COOK UNIVERSITY AUSTRALIA

Outline of MA5851 – Master Class 1 (NLP)

- Week 1
  - General idea of NLP – text as data, some tools, sentiment analysis
  - Revisit Python
- Week 2
  - Information extraction, td-idf, quantification of text, metadata
  - Reflection on search engines, features
- Week 3
  - CFG, Parsing, NLTK
- Week 4
  - Chunking, NER, Relation extraction, Web scrapping
- Week 5
  - Apache SPARK, RDD
- Week 6
  - Model development, Code management, GIT

Assessments of MA5851

- A1 (10%, due on 21 March)
  - Quiz
- A2 (40%, due on 04 April)
  - Reading list, recommender system
- A3 (50%, due on 21 April)
  - Making your web scrapper to build text/data corpus
  - Apply NLP techniques on your own dataset
  - Report writing (insight)

Machine learning synopsis

JAMES COOK UNIVERSITY AUSTRALIA

Purpose	Problem Space	ML Technique
Anomaly Detection	more features, aggressive boundary	One-class SVM
	less features, fast training	PCA-based anomaly detection
Prediction	Linear model, fast training	Linear regression
	Linear model, small dataset	Bayesian linear regression
	Accuracy, long training time	Neural network regression
	Accuracy, fast training	Decision forest regression
	Predict event counts	Poisson regression
	Accuracy, fast training, large memory	Boosted decision tree regression
Discovering structure	Clustering	K-means
Classification (two class, multi-class)	Fast training, linear model	Logistic regression
	Accuracy, long training time	Neural network
	Accuracy, fast training	Decision forest, Decision jungle
	More features	Deep SVM
Recommendation	What you may also like	Association rules, matchbox
Text Analytics	NER, Sentiment Analysis	Rule based, SVM
Computer Vision	Image recognition	CNN, OpenCV Library

JAMES COOK UNIVERSITY AUSTRALIA

Microsoft Azure Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.

What do you want to do?

Extract information from text

Text Analytics

Derives high-quality information from text

Answers questions like: What info is in this text?

- Extract N-Gram Features from Text ← Creates a dictionary of n-grams from a column of free text
- Feature Hashing ← Converts text data to integer encoded features using the Vowpal Wabbit library
- Preprocess Text ← Performs cleaning operations on text, like removal of stop-words, case normalization
- Word2Vector ← Converts words to values for use in NLP tasks, like recommender, named entity recognition, machine translation

Predict values

Regression

Makes forecasts by estimating the relationship between values

Answers questions like: How much or how many?

- Fast Forest Quantile Regression ← Predicts a distribution
- Poisson Regression ← Predicts event counts
- Linear Regression ← Fast training, linear model
- Bayesian Linear Regression ← Linear model, small data sets
- Decision Forest Regression ← Accurate, fast training times
- Neural Network Regression ← Accurate, long training times
- Boosted Decision Tree Regression ← Accurate, fast training times, large memory footprint

Generate recommendations

Recommenders

Predicts what someone will be interested in

Answers the question: What will they be interested in?

- SVD Recommender ← Collaborative filtering, better performance with lower cost by reducing dimensionality

Discover structure

Clustering

Separates similar data points into intuitive groups

Answers questions like: How is this organized?

- K-Means ← Unsupervised learning

Find unusual occurrences

Anomaly Detection

Identifies and predicts rare or unusual data points

Answers the question: Is this weird?

- One Class SVM ← Under 100 features, aggressive boundary
- PCA-Based Anomaly Detection ← Fast training times

Predict between several categories

Multiclass Classification

Answers complex questions with multiple possible answers

Answers questions like: Is this A or B or C or D?

- Multiclass Logistic Regression ← Fast training times, linear model
- Multiclass Neural Network ← Accuracy, long training times
- Multiclass Decision Forest ← Accuracy, fast training times
- One-vs-All Multiclass ← Depends on the two-class classifier
- Multiclass Boosted Decision Tree ← Non-parametric, fast training times and scalable

Predict between two categories

Two-Class Classification

Answers simple two-choice questions, like yes or no, true or false

Answers questions like: Is this A or B?

- Two-Class Support Vector Machine ← Under 100 features, linear model
- Two-Class Averaged Perceptron ← Fast training, linear model
- Two-Class Decision Forest ← Accurate, fast training
- Two-Class Logistic Regression ← Fast training, linear model
- Two-Class Boosted Decision Tree ← Accurate, fast training, large memory footprint
- Two-Class Neural Network ← Accurate, long training times

Classify images

Image Classification

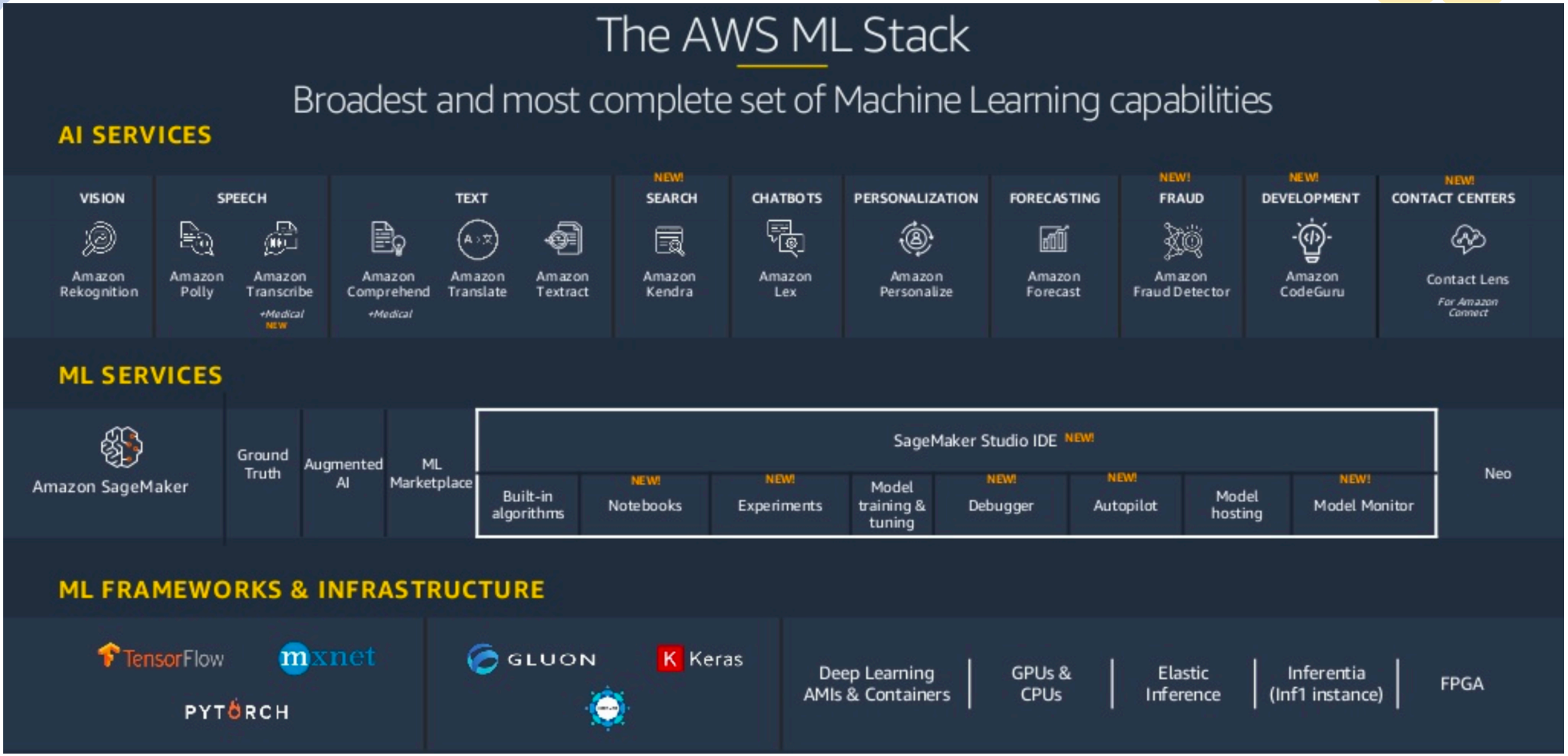
Classifies images with popular networks

Answers questions like: What does this image represent?

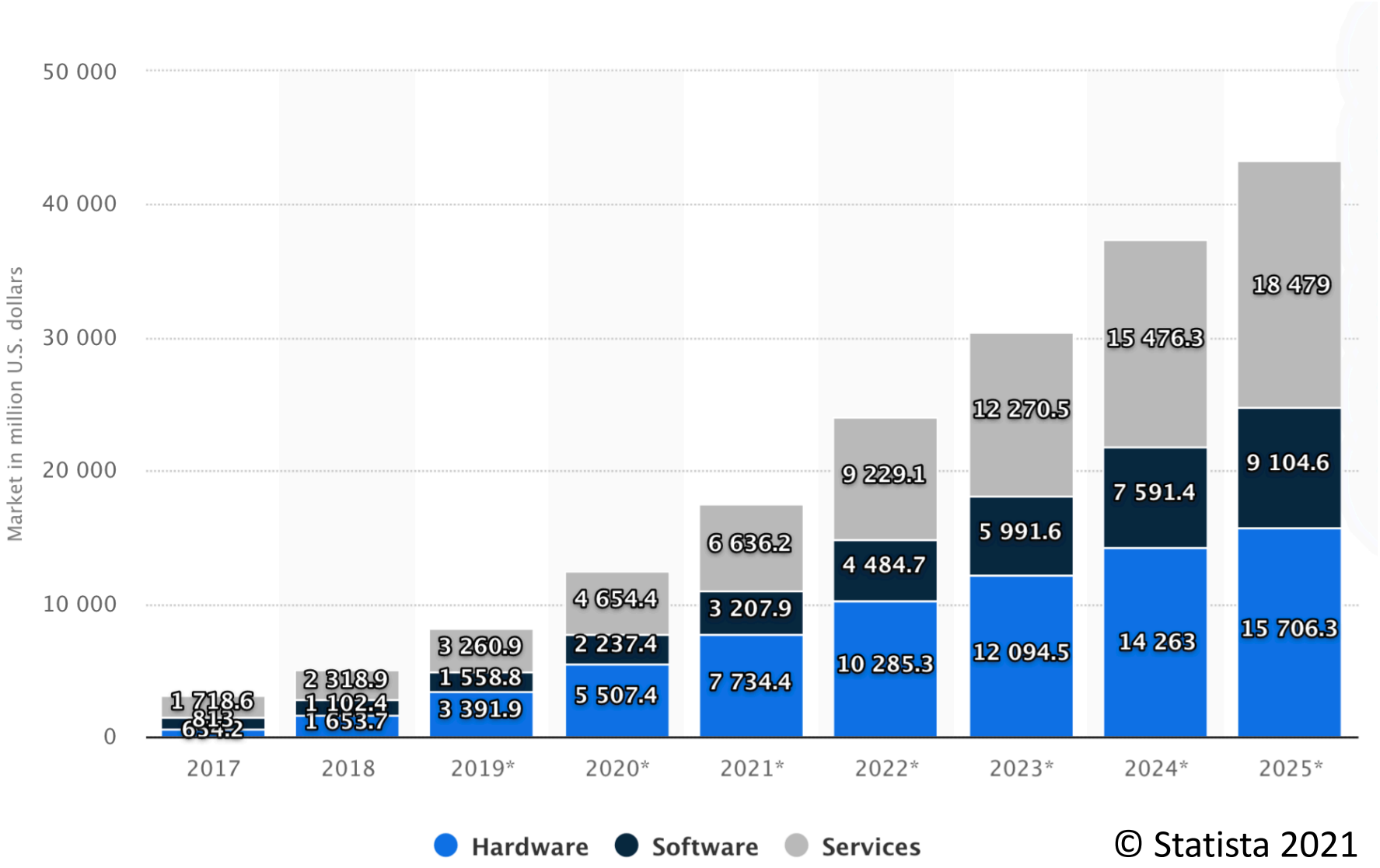
- DenseNet ← High accuracy, better efficiency

Share this poster: aka.ms/mlcheatsheet

Microsoft


The AWS ML Stack
Broadest and most complete set of Machine Learning capabilities
AI SERVICES
VISION
Amazon Rekognition
SPEECH
Amazon Polly
Amazon Transcribe
+Medical
NEW
TEXT
Amazon Comprehend
+Medical
Amazon Translate
Amazon Textract
NEW!
SEARCH
Amazon Kendra
CHATBOTS
Amazon Lex
PERSONALIZATION
Amazon Personalize
FORECASTING
Amazon Forecast
NEW!
FRAUD
Amazon Fraud Detector
NEW!
DEVELOPMENT
Amazon CodeGuru
NEW!
CONTACT CENTERS
Contact Lens
For Amazon Connect
ML SERVICES
Amazon SageMaker
Ground Truth
Augmented AI
ML Marketplace
SageMaker Studio IDE NEW!
Built-in algorithms
NEW! Notebooks
NEW! Experiments
Model training & tuning
NEW! Debugger
NEW! Autopilot
Model hosting
NEW! Model Monitor
Neo
ML FRAMEWORKS & INFRASTRUCTURE
TensorFlow
mxnet
PYTORCH
GLUON
Keras
Deep Learning AMIs & Containers
GPUs & CPUs
Elastic Inference
Inferentia (Inf1 instance)
FPGA

NLP – Revenues from NLP market worldwide from 2017 to 2025, by segment



NLP in healthcare

JAMES COOK UNIVERSITY AUSTRALIA

Doc2Vec: vector representations of clinical documents	Compare and detect changes in clinical guidelines and lab reports
Named entity recognition: leveraging unified medical language system (UMLS)	Extract clinical concepts (e.g., diagnoses, procedures, and symptoms) from electronic medical records, patient discharge summaries, and lab reports
Sequence to sequence: using stacks of long short-term memory (LSTM) networks	Map clinical concepts and diagnoses with codified clinical guidelines
Deep reinforcement learning and contextual bandits: using deep Q-networks[2]	Develop human-to-machine natural language instructions (e.g., robot-assisted surgery guided by human instructions, search-oriented conversations)

Source: McKinsey

Illustrative example for identifying ICD-10 code "H40.1121"

XX ICD-10 mapping

Disease category (H40) ← **Glaucoma** is a chronic condition in which fluid buildup causes increased pressure in the **eye**. This increased pressure can affect the optic nerve, potentially causing structural damage to the optic nerve fiber and visual field loss. The most common form of glaucoma is called **open-angle glaucoma**. Glaucoma can result in visual impairment when left untreated. Intraocular pressure (IOP) is the only risk factor for glaucoma that is currently **treatable**. Research has shown that lowering IOP can reduce the progression of loss of vision. → Body part, (.002)

Etiology (.11) ←

→ Extension (.0001)

Text from clinical guidance extract	ICD-10 nomenclature	ICD-10
Glaucoma	Disease category	H40
Open angle	Etiology	0.11
Eye	Body part	0.002
Treatable	Extension	0.0001

= H40.1121 (Glaucoma/Primary open-angle/ Left eye/Mild stage)

SOURCE: Multiple public sources on clinical guidelines; International Classification of Diseases, (ICD-10); International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM); expert interviews

NLP in healthcare

JAMES COOK UNIVERSITY AUSTRALIA

- Speech Recognition, Speech to Text
- Computer-assisted Coding
- Improvement in Clinical Documentation
- Automated Registry Reporting
- Data Mining Research
- Clinical Trial Matching
- Prior Authorisation
- Clinical Decision Support
- Risk Adjustment and Hierarchical Condition Categories
- Ambient Virtual Scribe
- Computational Phenotyping and Biomarker Discovery
- Population Health Management & Analysis

NLP – synopsis

JAMES COOK UNIVERSITY AUSTRALIA

- Stemming
- Lemmatisation
- POS tagging
- WSD
- NER
- Information Retrieval
- Sentiment Analysis
- Semantic Text Similarity
- Language Identification
- Text Summarisation
- Automatic Translation
- NLU
- NLG
- TTS, STT

JAMES COOK UNIVERSITY AUSTRALIA

Week 1 - Topics

- Why NLP is important
- What is NLP
- AI and NLP
- Supervised and Unsupervised Learning in NLP
- Power of NLP – sentiment analysis
- Constructing an NLP project
- Document separation – BoW
- Setting up jupyter notebook, running with examples

Why NLP is important

JAMES COOK UNIVERSITY AUSTRALIA

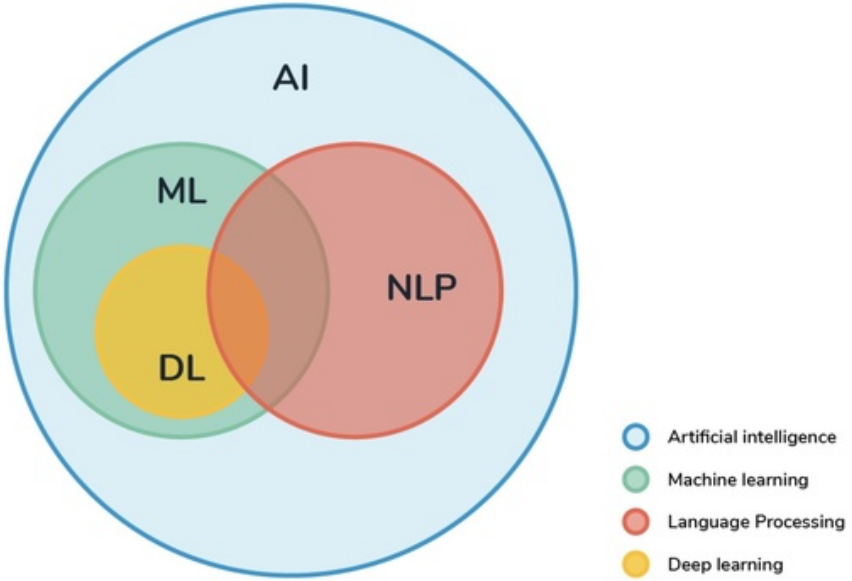
- Interest in NLP began in 1950 when Alan Turing published his paper entitled "Computing Machinery and Intelligence"

- Language is highly variable over time/place and genre/purpose as well as complex
  - Sparsity
  - Abstraction
  - Ambiguity
  - Context and world knowledge

- Why NLP?
  - Better interaction to obtain useful information from computing systems

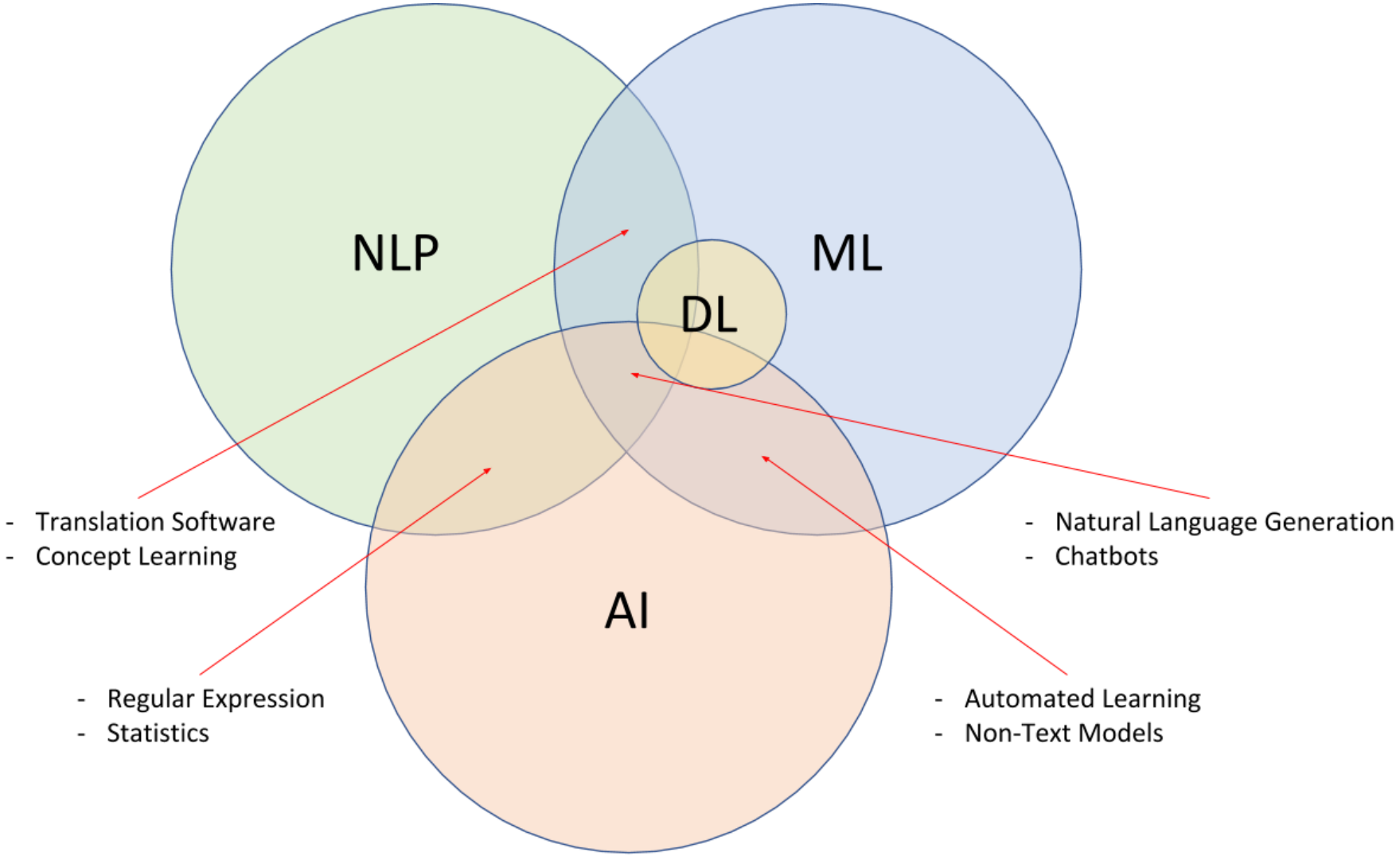
"A computer could be considered intelligent if it could carry on a conversation with a human being without the human realizing they were talking to a machine." -Alan Turing

Artificial intelligence and NLP



Generative Pre-trained Transformer 3 (GPT-3)?

What is NLP



deals with **analysing**, **understanding** and **generating** the languages that humans use naturally in order to interface with computers in both written and spoken contexts using **natural** human languages instead of computer languages.

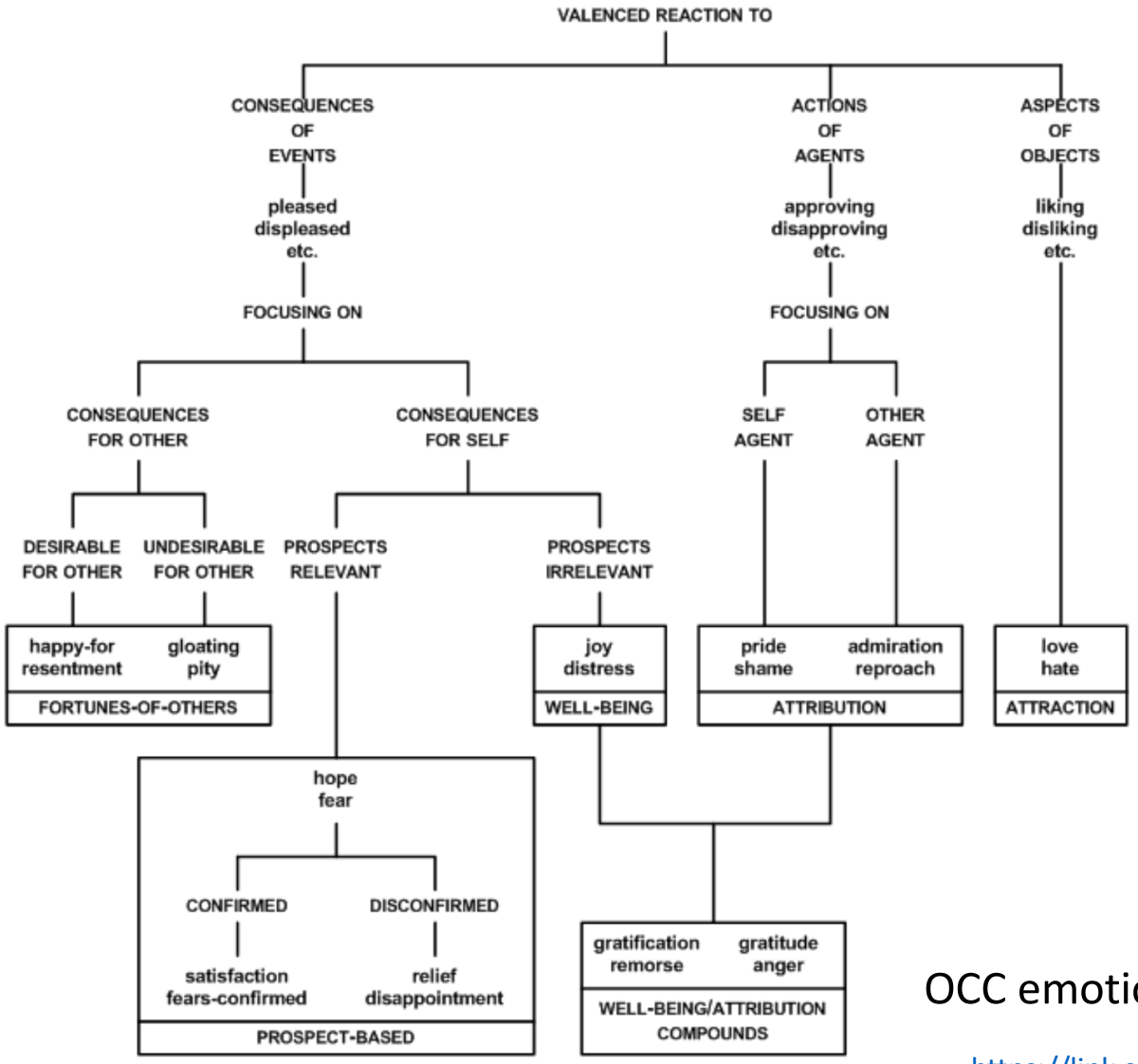
Unsupervised and supervised learning in NLP

JAMES COOK UNIVERSITY AUSTRALIA

- Clustering
  - Information retrieval: Finding documents with relevant content
  - Document categorisation: extracting themes
  - Anomaly detection: unusual (negative sentiment), naive Bayes
- Classification
  - Churn propensity models that include customer centre notes, website forms, emails and twitter messages
  - Hospital admission prediction models incorporating medical records notes as a new source of information
  - Insurance fraud modelling using adjustor notes
  - Sentiment categorisation from customer comments
  - Stylometry or forensic applications that identify the author of a particular writing sample

Power of NLP – sentiment analysis

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.



OCC emotion model provides rationale for emotion

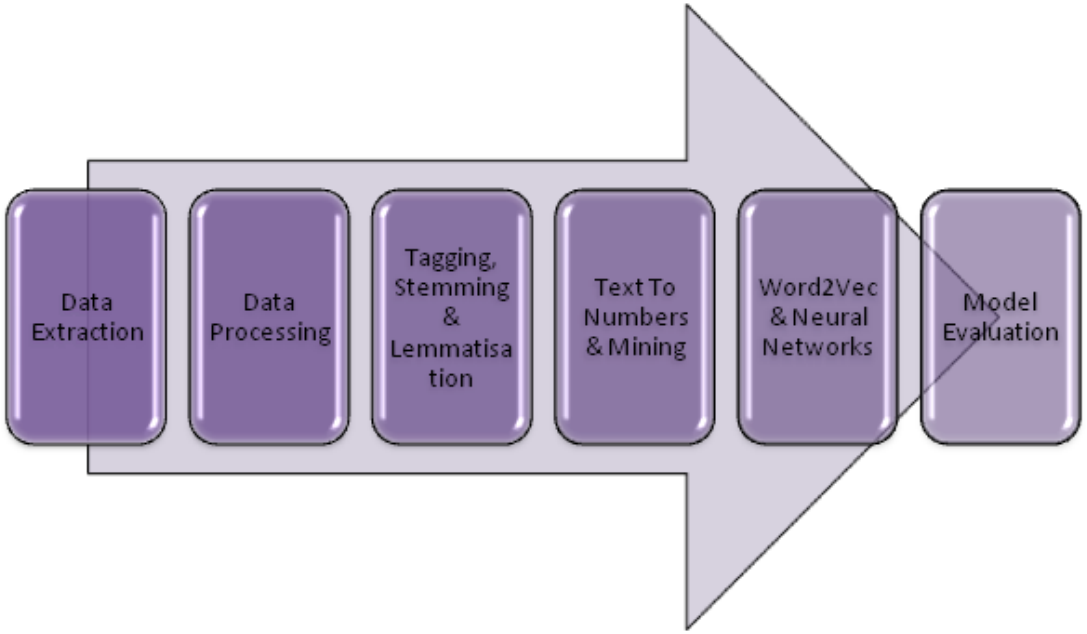
https://link.springer.com/chapter/10.1007/978-1-84800-306-4_4

Constructing an NLP project

JAMES COOK UNIVERSITY AUSTRALIA

- Applications you are exposed to every day that use natural language processing (NLP) and how these might have been built using an NLP style project
  - Identifying different cohorts of users/customers (e.g. predicting churn, lifetime value, product preferences)
  - Accurately detecting and extracting different categories of feedback (positive and negative reviews/opinions, mentions of particular attributes such as clothing size/fit…)
  - Classifying text according to intent (e.g. request for basic help, urgent problem)
  - Language Generation: New words, Lyrics, summary

- **Action 1. Define your corpus**
- **Action 2. Define your dictionaries**
- **Action 3. Establish synonym tables**
- **Action 4. Establish multi-word term tables**
- **Action 5. Establish topic tables**



Document separation – BoW

One-hot encoding

a way of extracting features from text for use in modelling

N gram analysis → document classifier
TF-IDF → score → document classifier

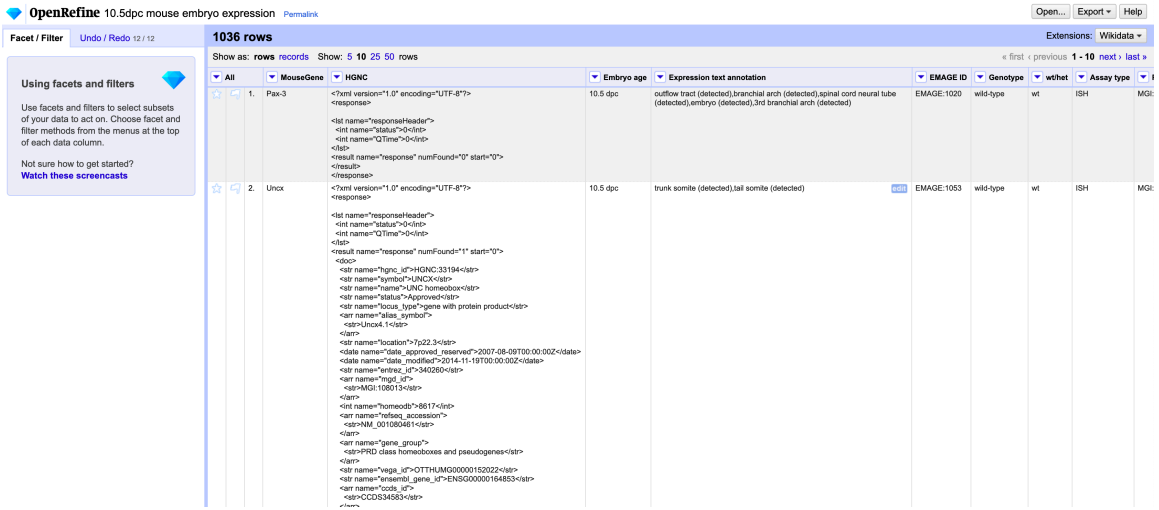
JAMES COOK UNIVERSITY AUSTRALIA

Setting up jupyter notebook and self-learning proc

- Python basics, list and dictionary
- Sentiment Analysis: Political parties, D Trump tweets

SLP - Week 1

- MA5851 Week1 Open Refine Introduction
- MA5851 Week1_SLP_OMIM



- MA5851 Week 1 Self Learning Practical 2
- MA5851 Week 1 SLP_3_
- Bonus: Tweet Analysis