Swiss Re

General Public Release

The Data Scientists: Actuary 2.0?

Doan Le FIAA & Guan Wang

AAC2018 - 18 September 2018

AT WORK

Dust Off Your Math Skills: Actuary Is Best Job of 2013

By Lauren Weber

Apr 22, 2013 5:00 pm ET

THE WALL STREET JOURNAL.

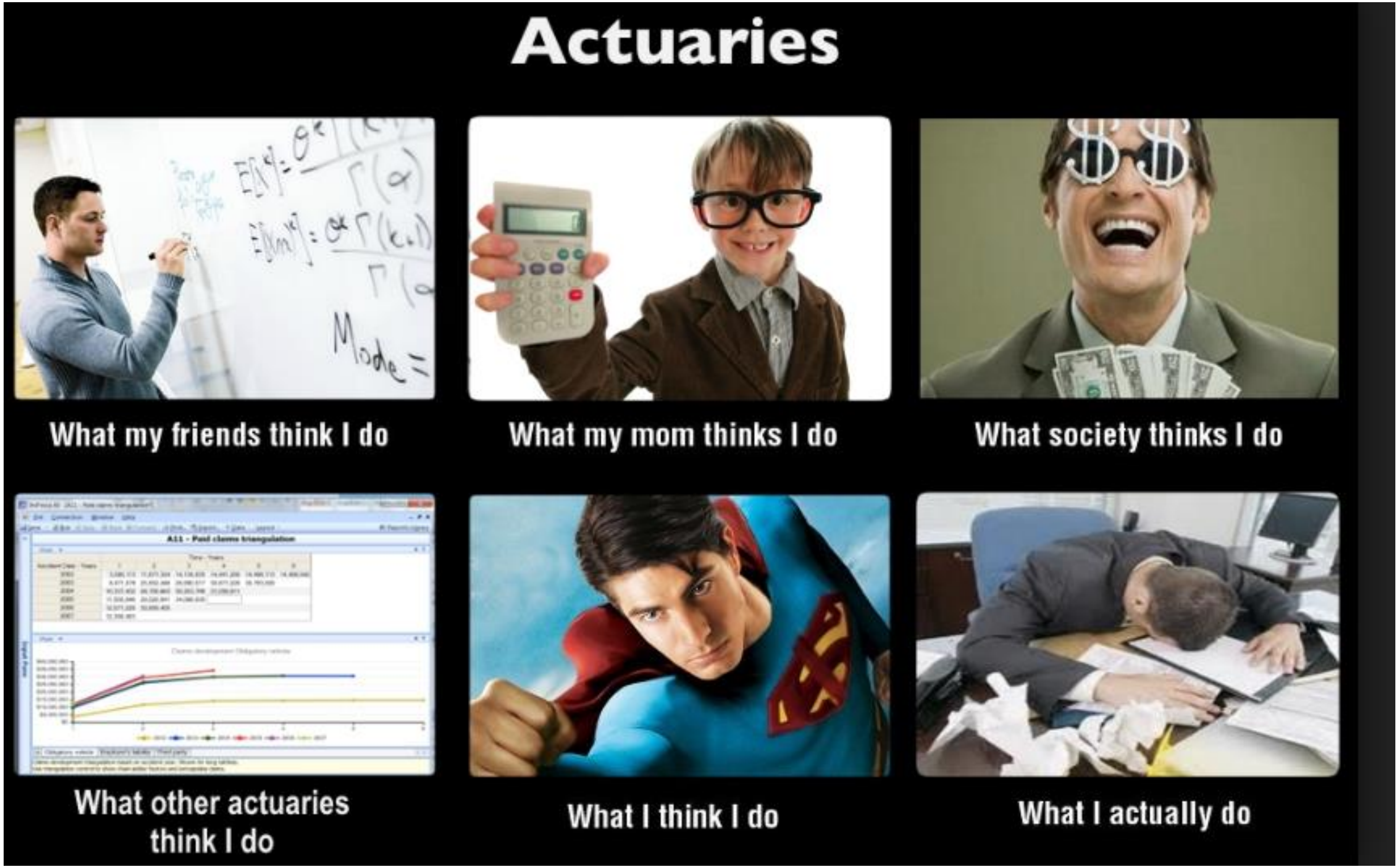
DATA

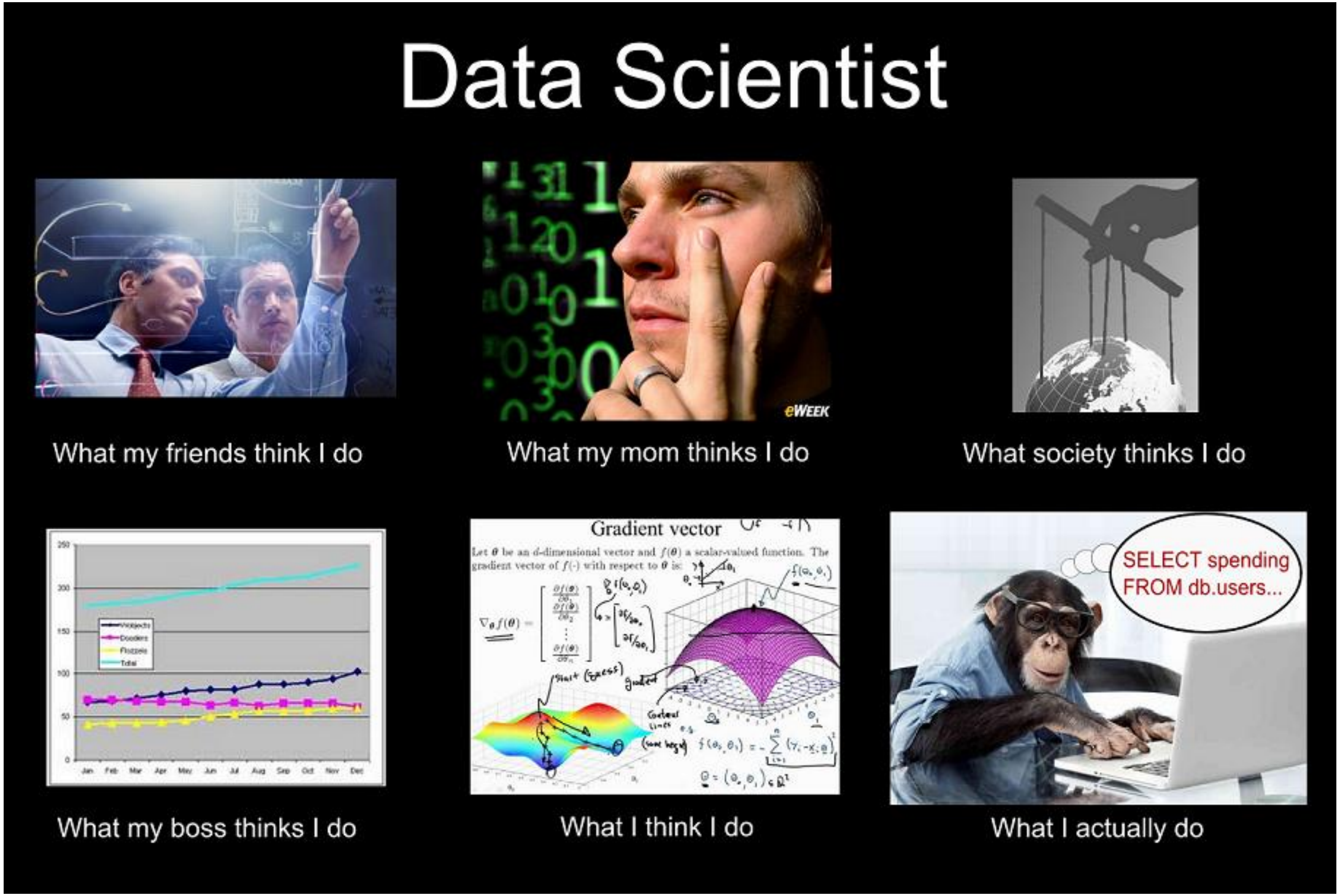
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

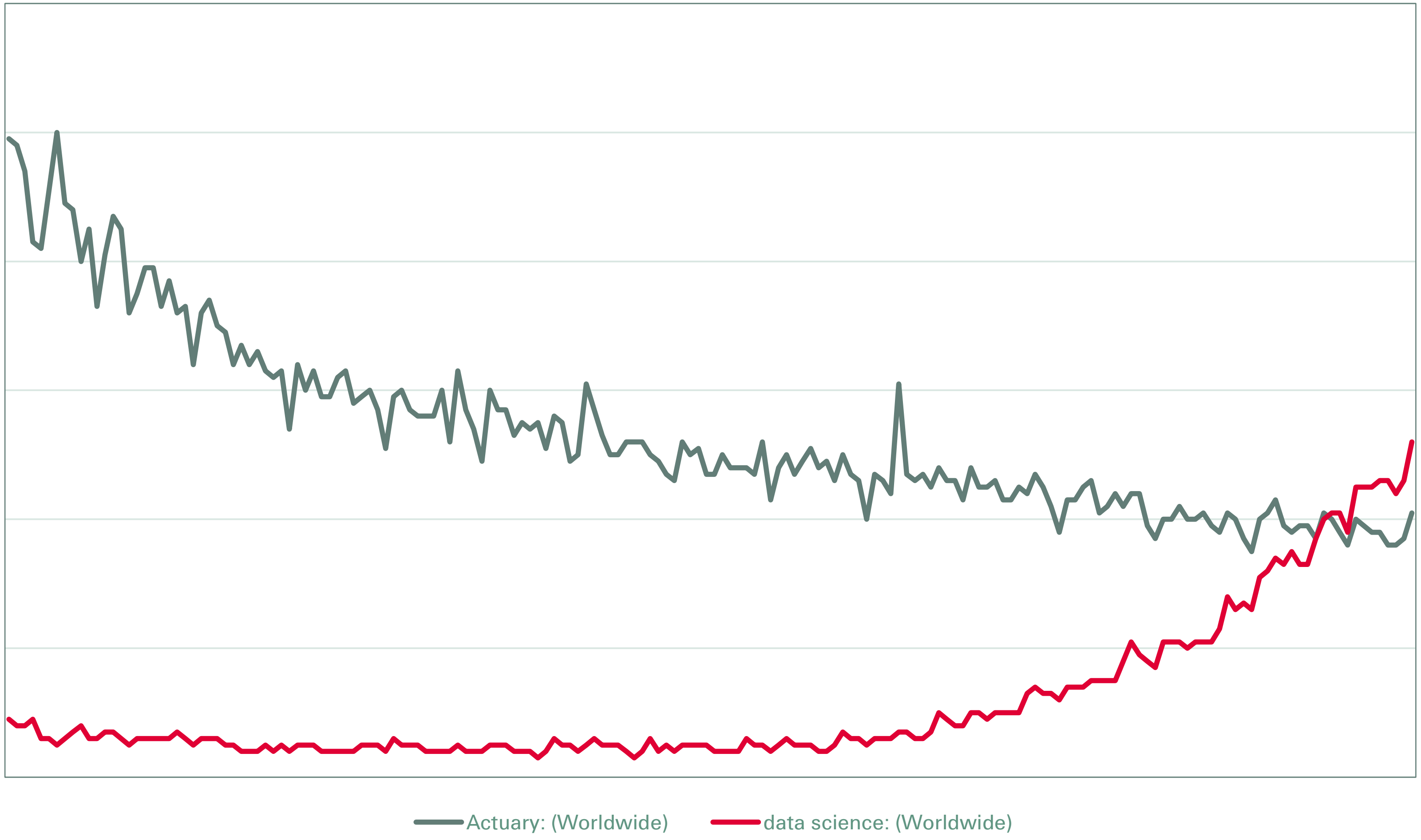
FROM THE OCTOBER 2012 ISSUE

Harvard Business Review

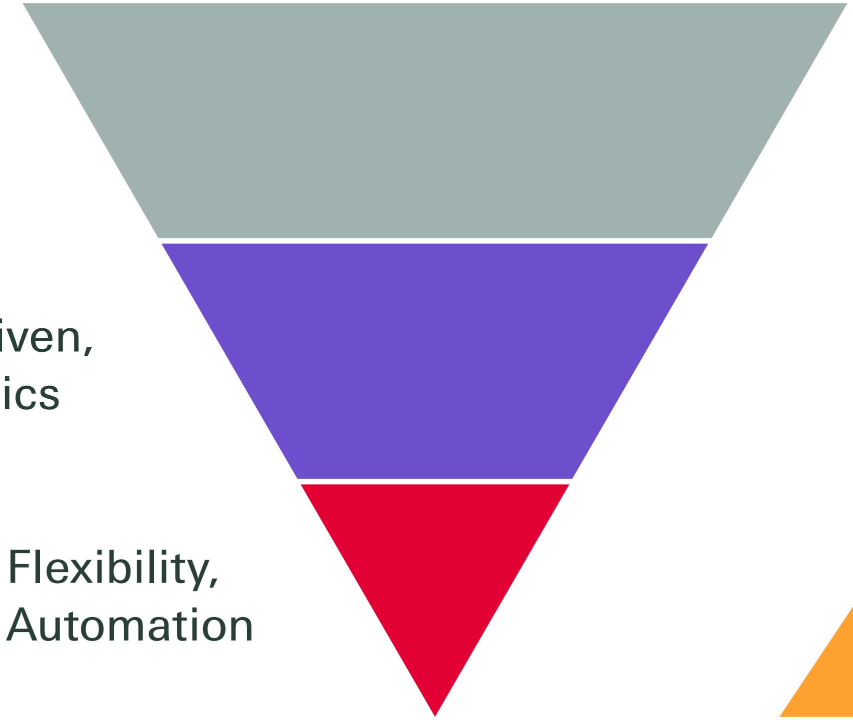

Actuaries
What my friends think I do
What my mom thinks I do
What society thinks I do
What other actuaries think I do
What I think I do
What I actually do


Data Scientist
eWEEK
What my friends think I do
What my mom thinks I do
What society thinks I do
Gradient vector
SELECT spending FROM db.users...
What my boss thinks I do
What I think I do
What I actually do

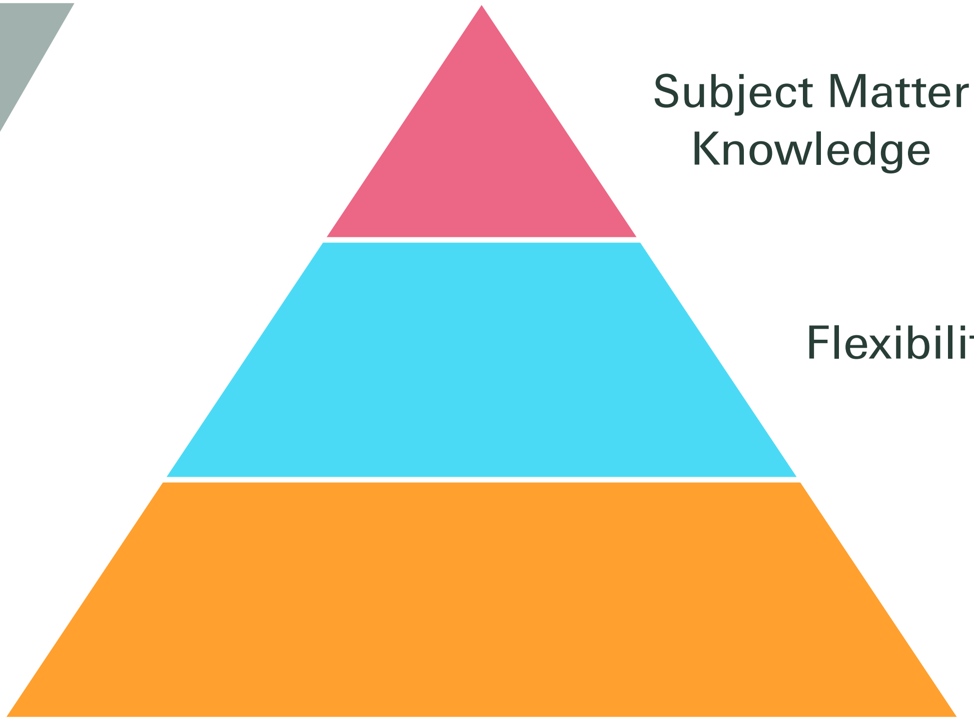
"Actuary" vs "Data Science" Google Searches (2004 – 2018)



Actuary



Data Scientist



	Actuary	Data Scientist
Underlying principle	• **Pooling** is a key concept • Interested in **cohort** experience of similar risk profiles. • Balance between **data insight vs. business knowledge**	• Data science generally looks at **individuals**. • A major goal is to achieve a good balance of **high accuracy** and **good generalization**
Application	• Actuarial modelling **combines** data analysis and subject matter expertise • Problems solved are almost always **financial** • **Risk**-oriented	• Data Science modelling's raw outcome is usually a **probabilistic score** • Involved in **wide** range of problems not just financial e.g. optimizing customer experience • **Opportunity**-focused
Data	• Actuaries predominantly focus on attributes directly **applicable** to problem statement • Reliant on data but has to fill in gaps with **expert knowledge**	• Data Science modelling is **exploratory** and usually looks at as many attributes as available • Structured, unstructured, labelled, unlabelled • **Completely dependent** on data
Modelling technique	• Actuarial modelling tends to rely on **established** methods and are more **explanatory** • A global **explanation** is important e.g. male is x% more likely to claim than female	• A **wider** selection of **modelling** techniques are used. Some are extremely complicated • Selection of model **depends** on data • Strong command of **programming** and **automation** skills

A COMMON ACTUARIAL PROBLEM

How much premium should we charge to cover policyholder death?

A COMMON ACTUARIAL PROBLEM

How much premium should we charge to cover policyholder death?

A COMMON ACTUARIAL PROBLEM

How much premium should we charge to cover policyholder death?

Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment


Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment

Data Sourcing: Why do we need label data?

Supervised

Learn a function by observing examples containing the input and the expected output.

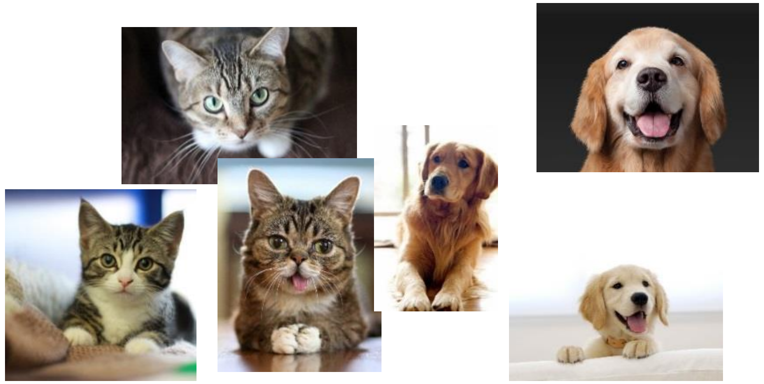
- Classification
- Regression

This looks like the dogs I saw before, it should be a dog too

Unsupervised

Find underlining relations in data by observing the raw data only (without the expected output)

- Clustering
- Dimensionality reduction

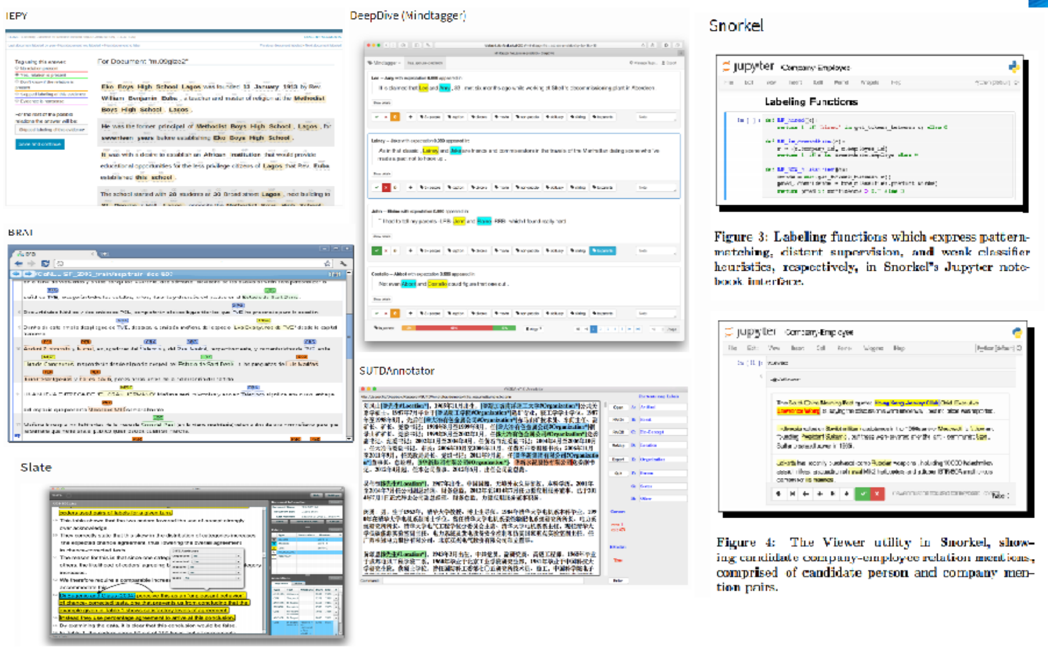


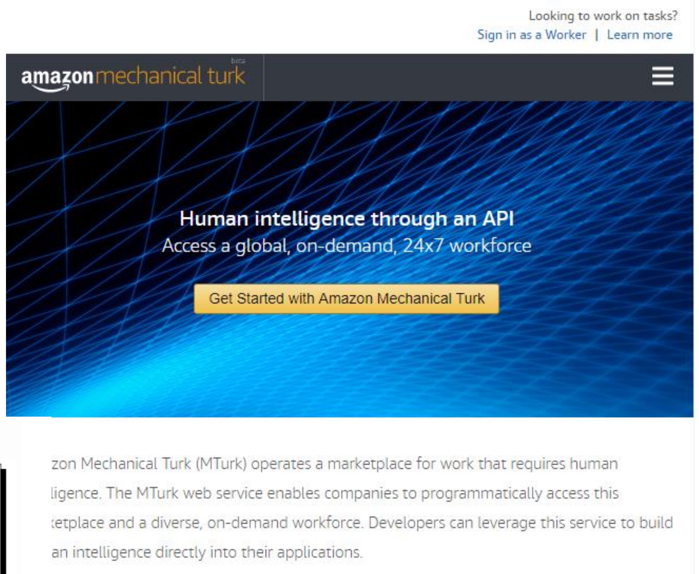
I can see two types of animals. But I don’t know what they are. Can someone tell me what they are?

Data Sourcing: How do we get label data?

Refine Problem → **Source Data** → Prepare Data → Modelling → Testing → Results & Deployment

- Most Machine Learning Applications are **Supervised** Learning
- If lucky: automatically generated
    - Click Google ad on a webpage
    - Add stuff in shopping cart on Amazon
    - Listen to music on Spotify
- Not so lucky, but there are smart tricks
    - Crawl from Internet
    - Distant Supervision
- Not lucky, no tricks (usually)
    - Label by hands





https://www.mturk.com/


Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment

Cleansing and Feature Engineering

Categorical Features

Product_Info, Employment_Info, Insurance_History, Medical_History, etc.

Numerical Features

Ins_Age, Ht, Wt, BMI etc.

Unstructured Features

Medical Texts, X-Ray Pictures etc.

- Ht_Ratio = (Ht-Minimum_Ht)/(Maximum_Ht – Minimum_Ht)

Feature Transformation

- BMI_Age = BMI * Age^2

Feature Combination

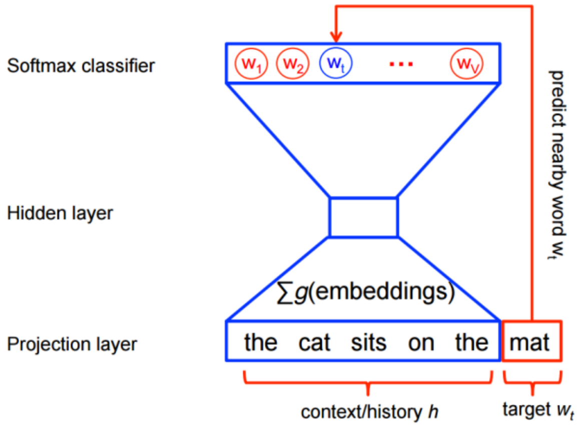
- Computer Vision: pictures to labels
- NLP: Text to Labels; Text to embeddings

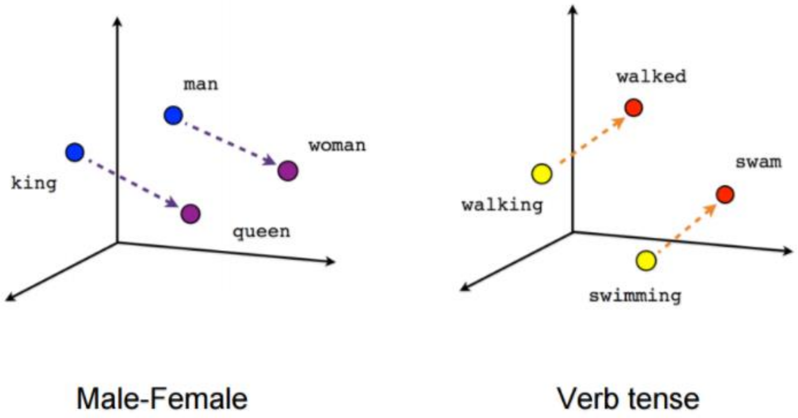
Unstructured Feature Conversion

Feature Engineering: Representation Learning

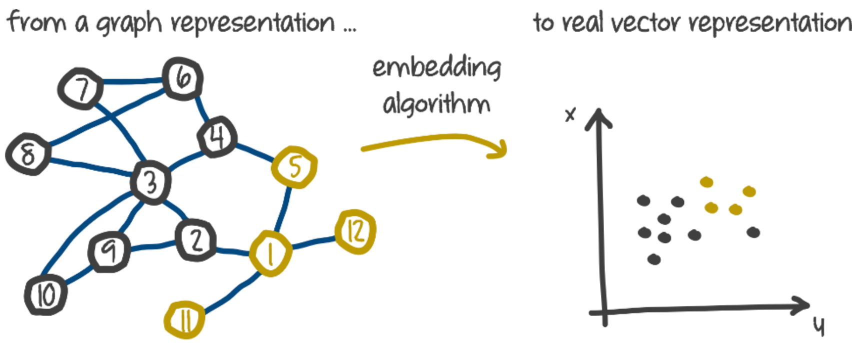
Word Embedding:
From text to vectors

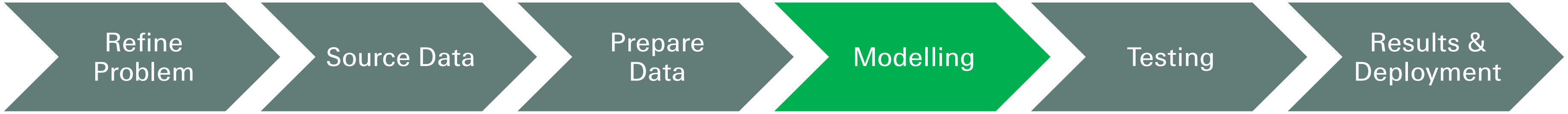
King - Man + Woman = Queen





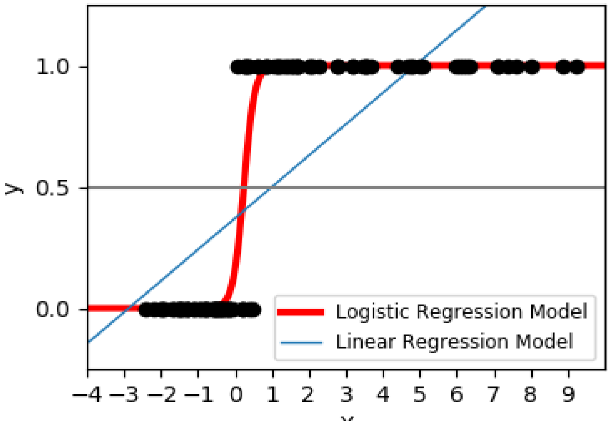
Graph Embedding:
From graphs to vectors



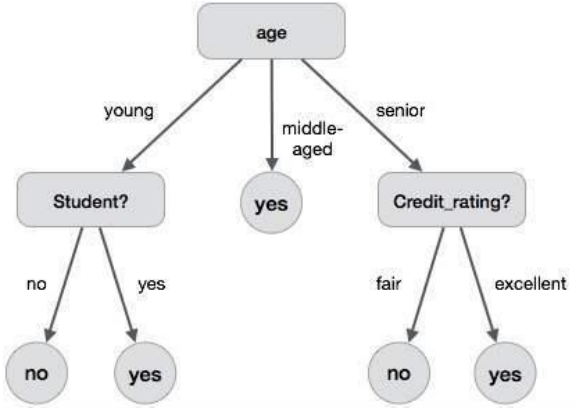

Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment

Modelling: Model Selection

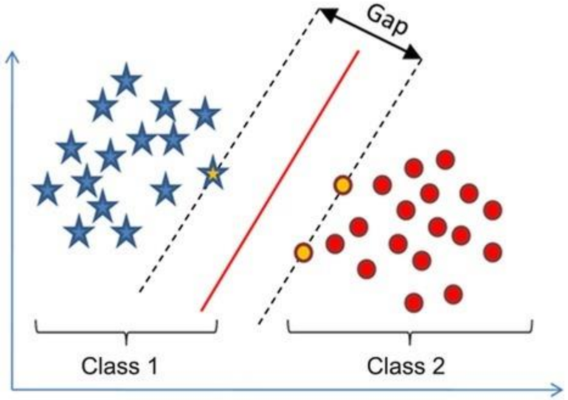
- Single Model Algorithms



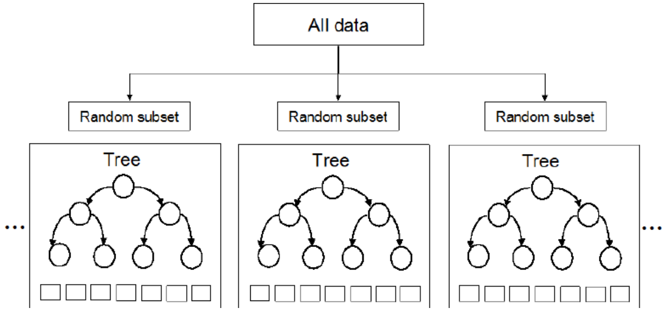
GLM



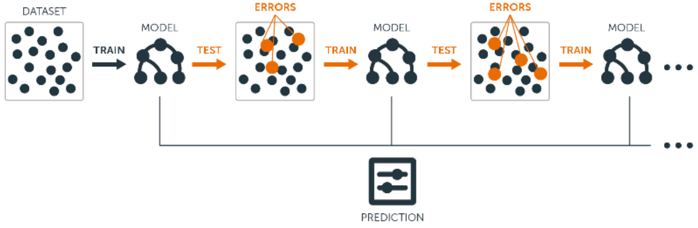
Decision tree



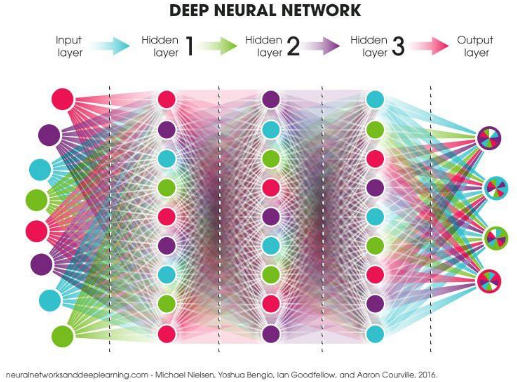
SVM



Random forest



Gradient boosting tree



Neural network

Modelling: Model Ensembling

- Ensemble single models to get better results (in the cost of more complex models)

XGBoost (Gradient boosting tree)
Random Forest
Logistic Regression
Neural Networks
...

Training to update internal parameters
Validation to update hyper parameters

Ensemble by Bagging or Stacking

Training to update internal parameters
Validation to update hyper parameters

Final Model

Testing to get final accuracy

Modelling: Data Splitting

- Data Splitting

Training set

Test set

Train and tune model (with cross-validation). Typically multiple models are trained

Don't use this until the very end. It is used to select the model with the best technical performance

In model build, our goal is to apply machine learning techniques to learn the pattern of labelled data, and this pattern is able to generalize well to unseen data.

The data is split into two sets, training and test set.

Training set is used to train and tune model (with cross-validation). The trained model contains the pattern of labelled data. The test set acts as unseen data, which is used to select the one with the best technical performance among the multiple models trained.

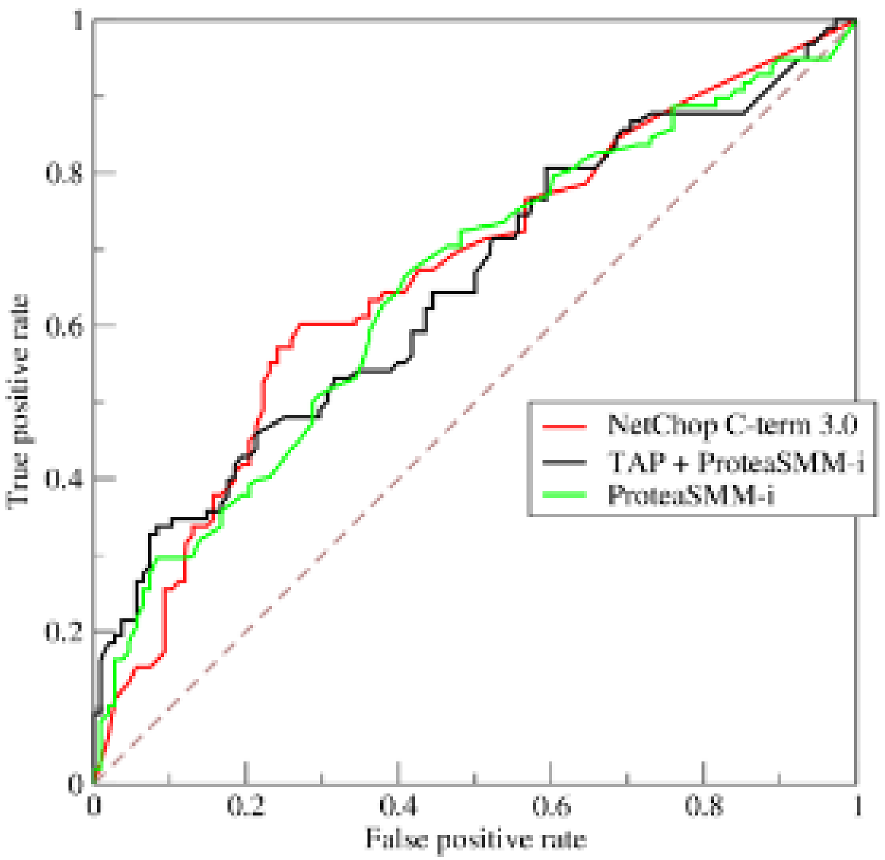

Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment

Modelling: Testing

– Confusion Matrix

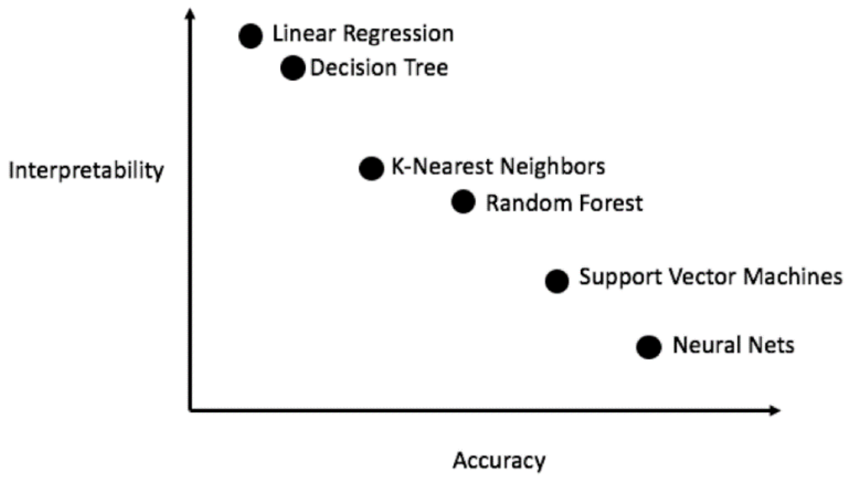
		True condition				
	Total population	Condition positive	Condition negative	Prevalence = Σ Condition positive/Σ Total population	Accuracy (ACC) = Σ True positive + Σ True negative/Σ Total population	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = Σ True positive/Σ Predicted condition positive	False discovery rate (FDR) = Σ False positive/Σ Predicted condition positive	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = Σ False negative/Σ Predicted condition negative	Negative predictive value (NPV) = Σ True negative/Σ Predicted condition negative	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = Σ True positive/Σ Condition positive	False positive rate (FPR), Fall-out, probability of false alarm = Σ False positive/Σ Condition negative	Positive likelihood ratio (LR+) = TPR/FPR	Diagnostic odds ratio (DOR) = LR+/LR−	F_1 score = 2/1/Recall + 1/Precision
		False negative rate (FNR), Miss rate = Σ False negative/Σ Condition positive	True negative rate (TNR), Specificity (SPC) = Σ True negative/Σ Condition negative	Negative likelihood ratio (LR−) = FNR/TNR		

– Receiver Operating Characteristic (ROC) and Area Under Curve (AUC)



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

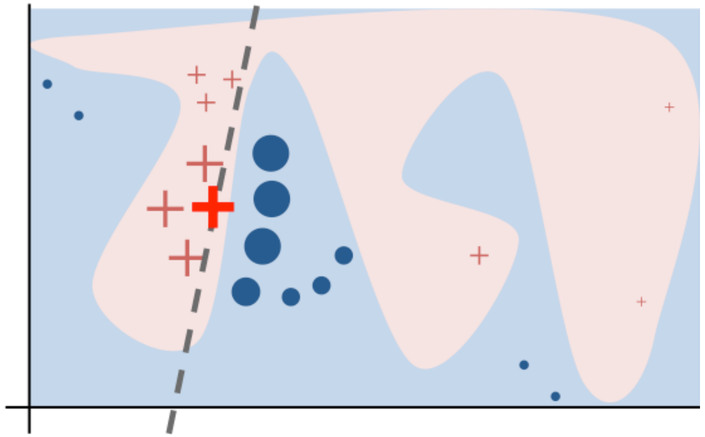
Modelling: Interpretation



https://medium.com/ansaro-blog/interpreting-machine-learning-models-1234d735d6c9

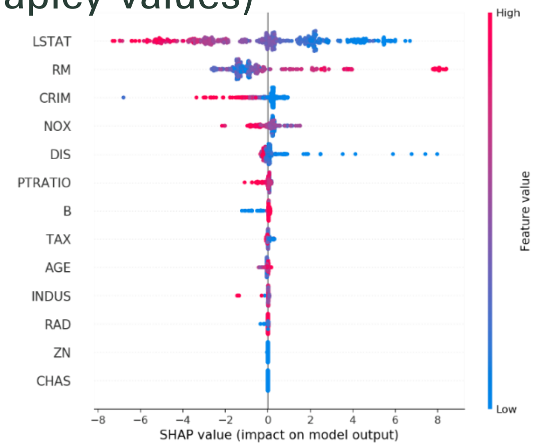
- Many ways to interpret the "Black Box"

  - Fit local prediction into a linear model (LIME)



https://github.com/marcotcr/lime

- Switch on and off feature combinations to get feature importance (Shapley Values)

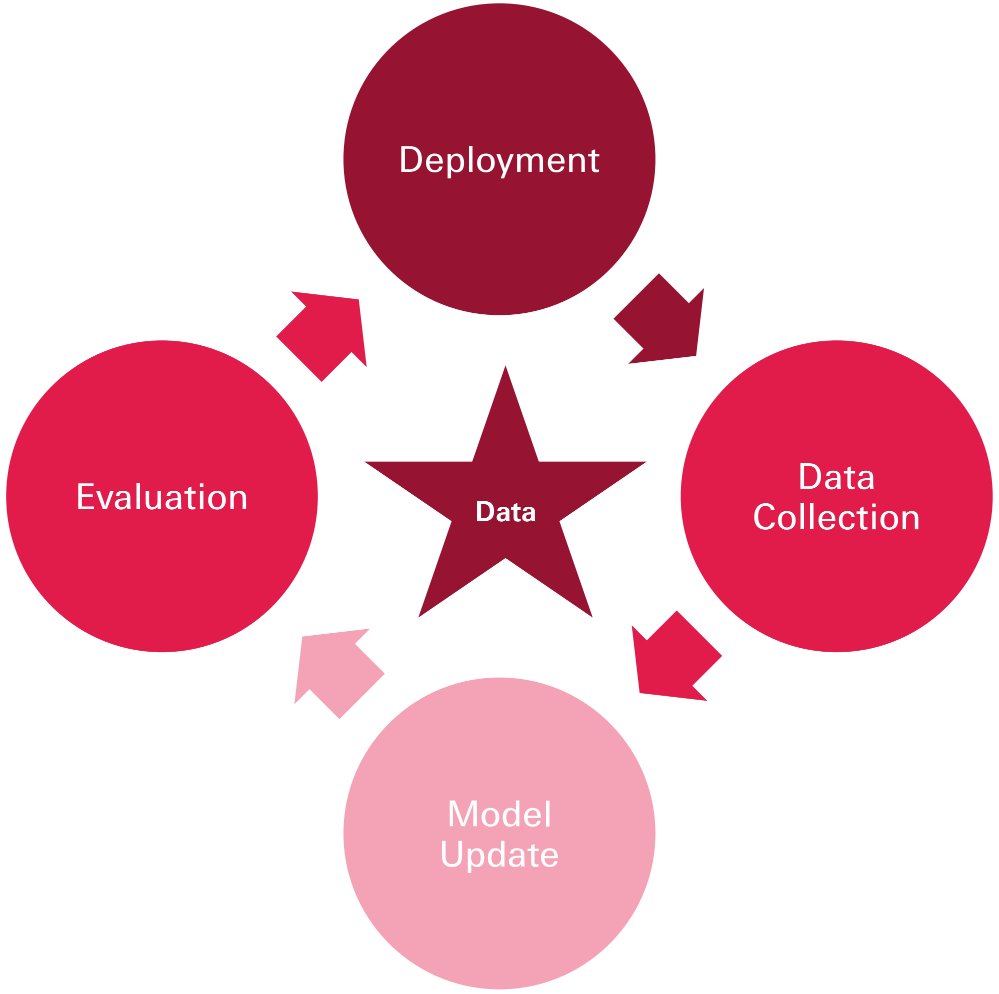


https://github.com/slundberg/shap


Refine Problem
Source Data
Prepare Data
Modelling
Testing
Results & Deployment

Model Deployment

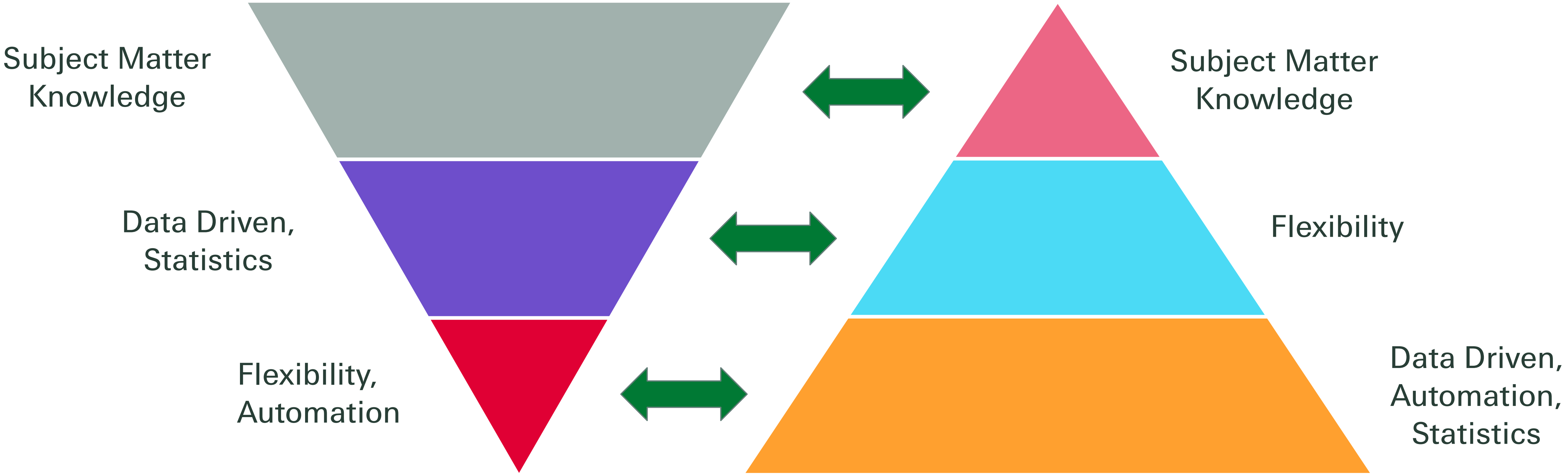
- Model access through API
- Easy integration into business process
- Constant Model Update
- A/B Testing
- Automation



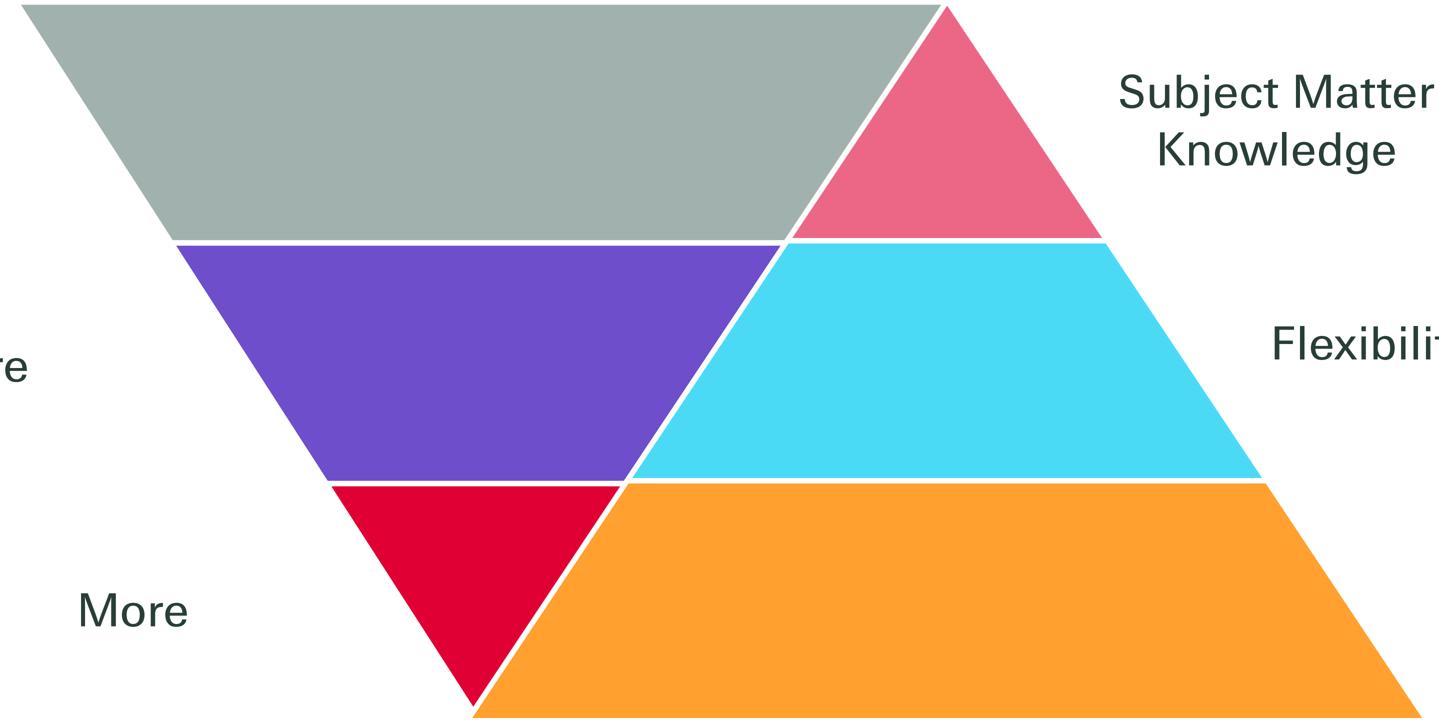
Summary

Actuary

Data Scientist



Actuaries together with Data Scientists



How would you build a system to provide personalized insurance product recommendations to potential customers, just like advertisement on Google, books on Amazon, movies on Netflix and music on Spotify?

Questions?

Legal notice