Institute
and Faculty
of Actuaries

Modular Framework of Machine Learning Pipeline

John Ng MA FIA BPharm

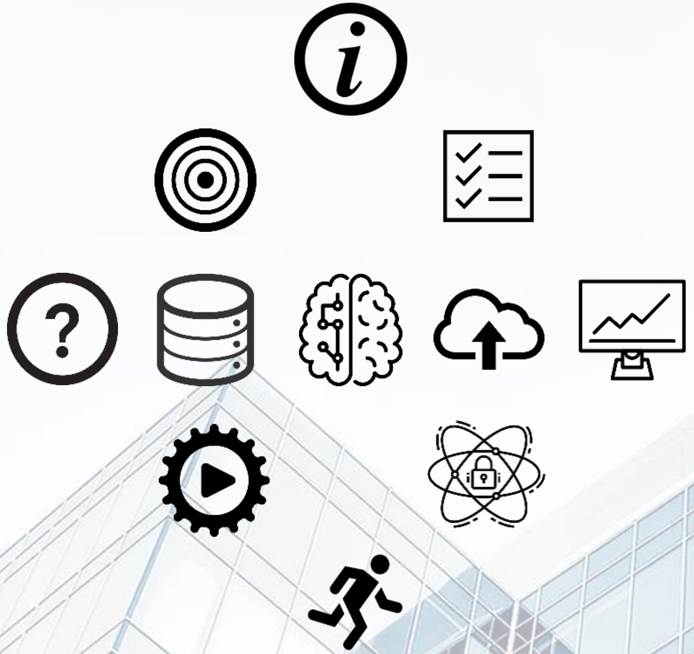
September 14, 2020

Who is Speaking?

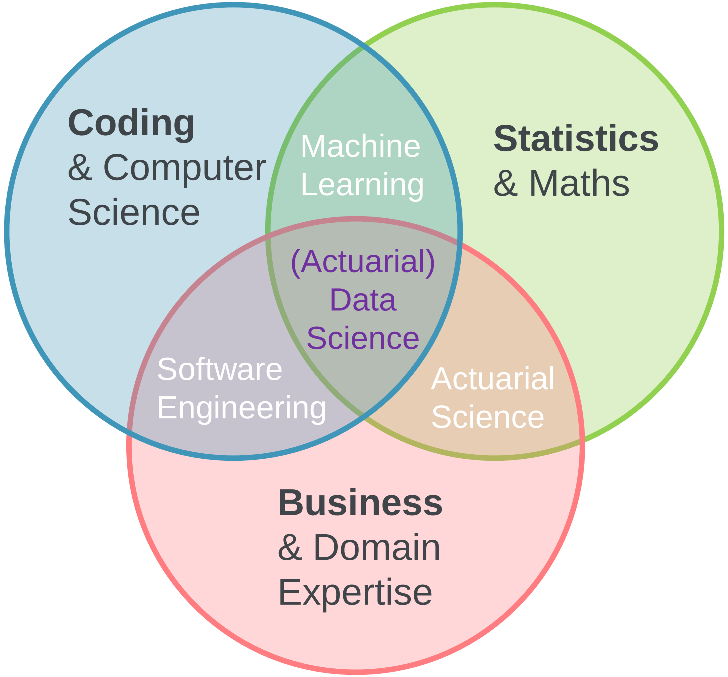
- John Ng, MA FIA BPharm
- Senior Data Scientist at Reinsurance Group of America (RGA), London
- Registered Pharmacist in Australia - worked in hospital and retail pharmacies
- Deputy Chair of IFoA Health and Care Research committee
- Chair of IFoA Data Science Research workstream
- Contact: LinkedIn

Agenda

- Introduction
- Strategy
- ML Pipeline
- Management
- Application



Actuarial science vs data science?

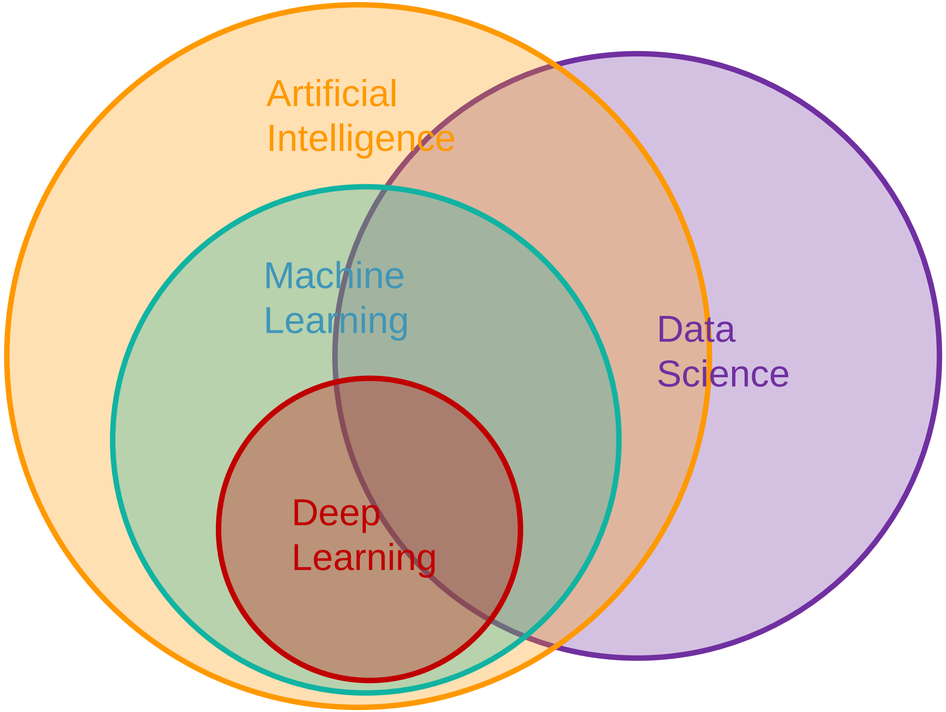


Demystifying AI, Machine Learning, Deep Learning

“Thinking” machines

Object Labeller

Multilayer
Neural Network



Understanding Data

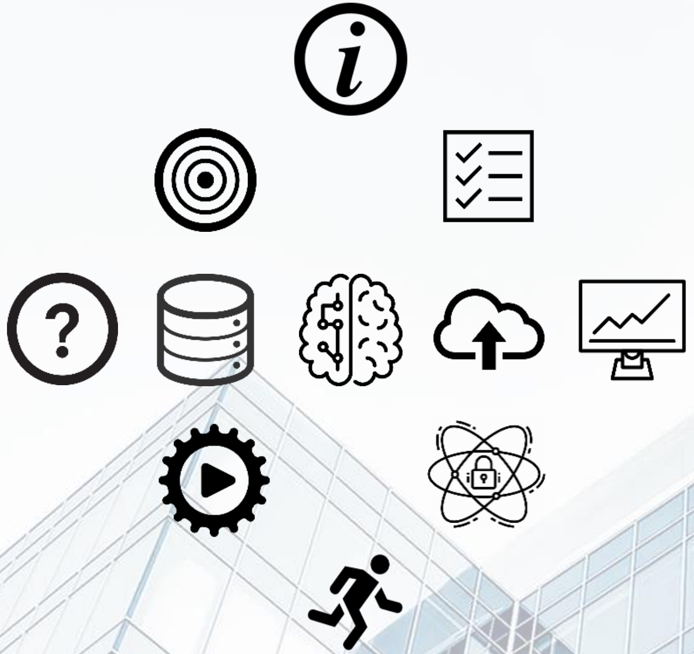
More ~~Jargons~~ Toolkits?

Data is the new LEGO

Source: Medium and Lego

Agenda

- Introduction
- **Strategy**
- ML Pipeline
- Management
- Application



Objectives of Machine Learning Pipeline

S	SPEED
P	PERFORMANCE
R	RISK MANAGEMENT
I	INTEGRATION
S	SCALABILITY

Data Science use cases in Insurance

- Chat-bots
- Robo-Advisors
- Customer Service prioritisation
- Paperwork automation
- Unstructured data

- Conversion
- Persistency / Renewal
- Churn / Lapse
- Cross-Selling
- Customer Segmentation
- Customer Life-Time-Value (LTV)
- Recommendation Engine
- Sentiment Analysis

- Pricing Accuracy
- Pricing Sensitivity & Elasticity
- Pricing Optimisation
- Dynamic Pricing
- Reserving
- Capital Modelling
- Mortality and Morbidity

- Claims management
- Risk Granularity
- Accelerated Underwriting
- Motor Telematics
- Healthcare analytics, Wearables
- Portfolio Analytics

Agenda

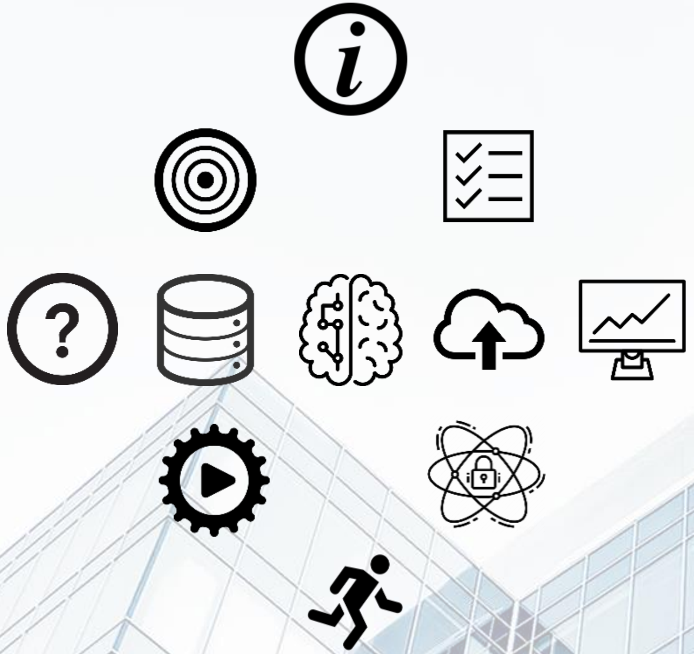
Introduction

Strategy

ML Pipeline

Management

Application



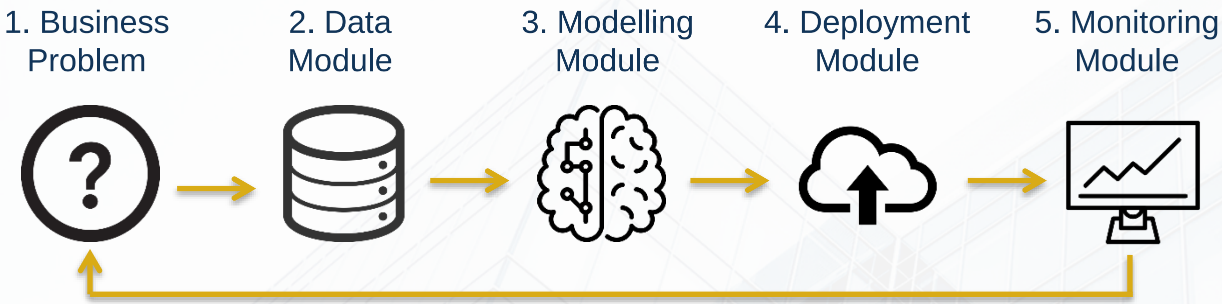
Modular Framework of Machine Learning Pipeline

1. Business Problem | 2. Data Module | 3. Modelling Module | 4. Deployment Module | 5. Monitoring Module

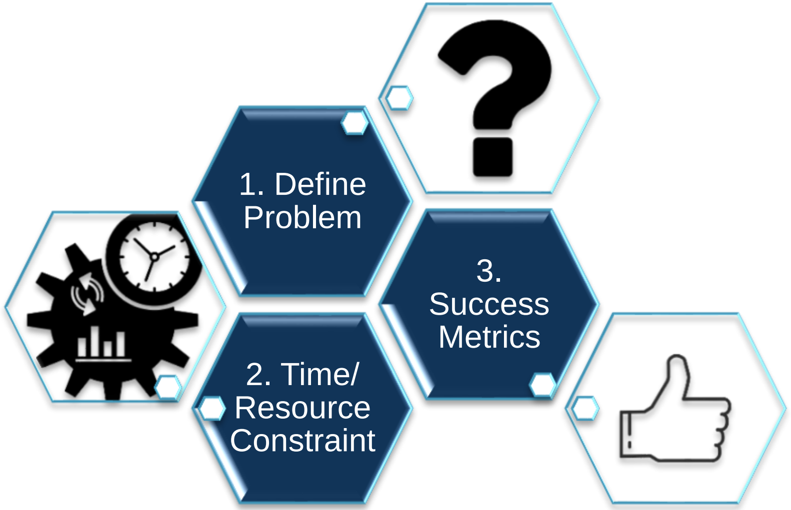
Opportunity | Information | Insight | Execution | Result

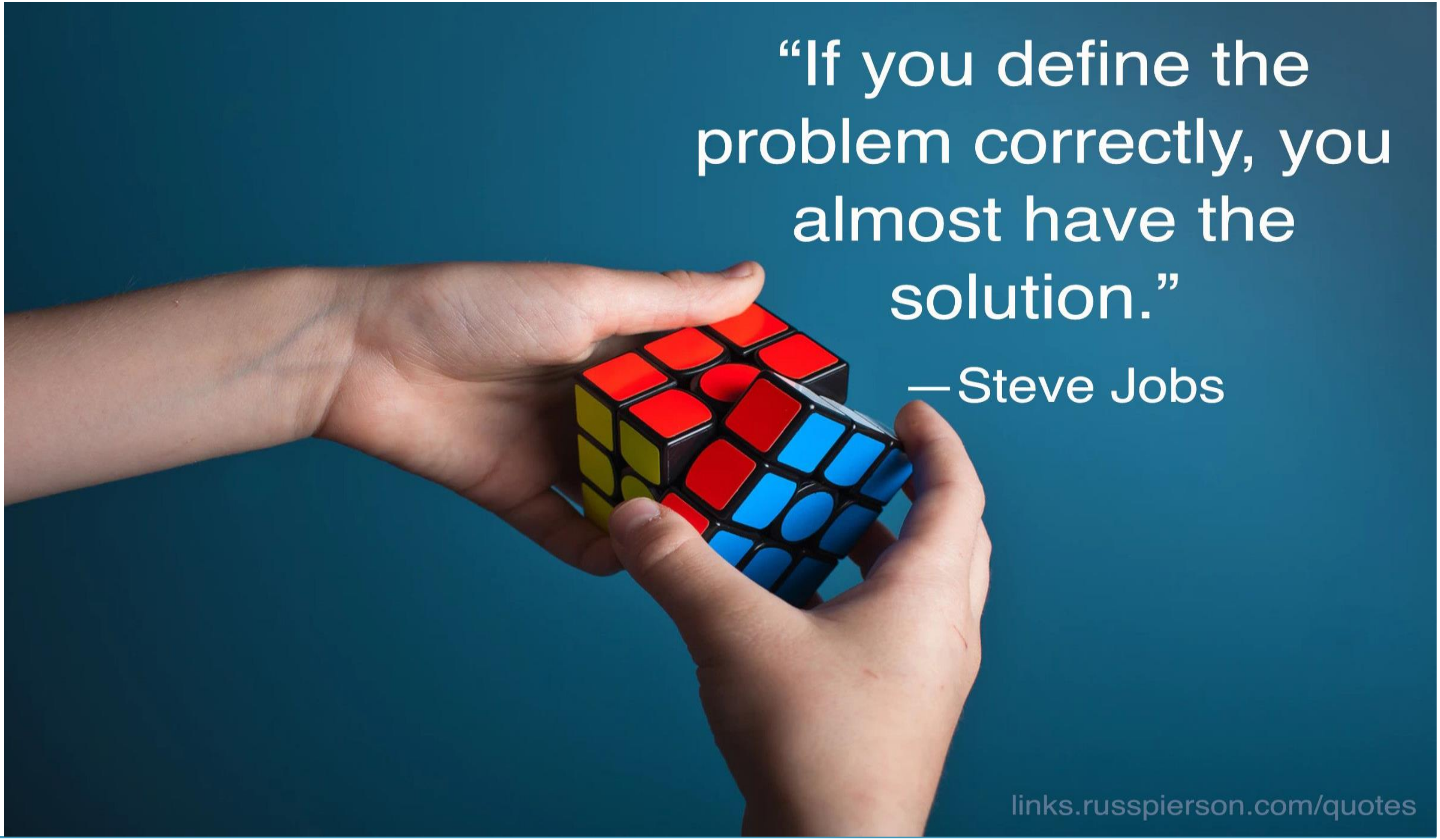
Actuarial Control Cycle

Actuarial Data Science Control Cycle



Business Problem




"If you define the problem correctly, you almost have the solution."
—Steve Jobs
links.russpierson.com/quotes

Data Module

Data Sourcing & Engineering

- Data Sources
- Data Connectivity
- Data Engineering
- Data Warehouse

Exploratory Data Analysis (EDA)

- Assess Quality
- Statistics
- Distributions
- Correlations
- Reporting & Visualisation

Data Cleaning & Preparation

- Identify errors
- Formatting
- Outliers
- Remove “post-event” information

Feature Engineering

Feature ..
1. Extraction
2. Transformation
3. Selection

- Expert Driven
- Automatic F.E.

Data Segregation

Split into Train, Validation & Test Set

- Random split
- Stratified sampling

Imputation

Impute missing values
- Fit learner on training set then transform train and test sets
- Fixed imputations (Mean, Median)
- Better approaches: MICE, KNN

Data Dictionary

Feature Store

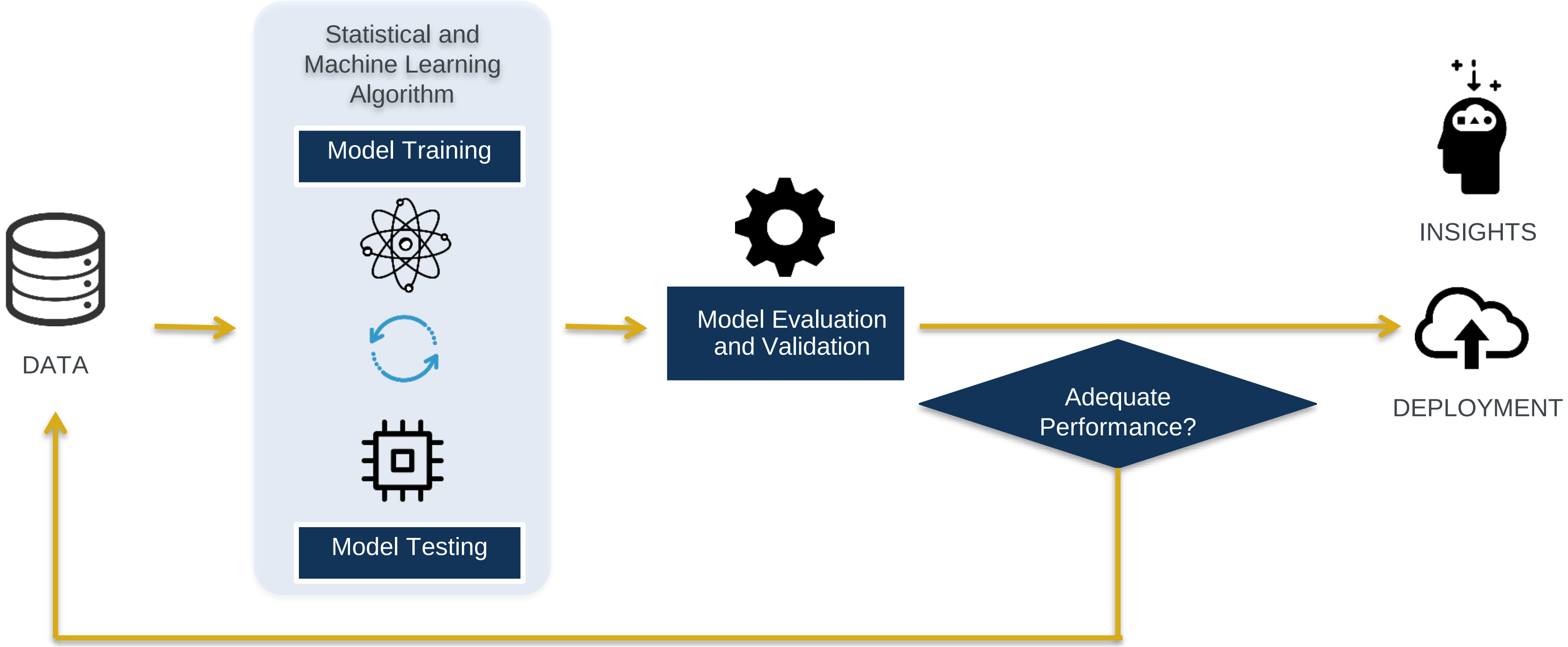
Feature Engineering

Response (output) — Features (input) — Additions from Feature Engineering (input)

Policy_ID	Claim	DriverAge	AreaCode	VehClass	VehVAL	Mileage	DriverAge[2]	DriverAge_int_VehClass
POL20190901001	1	50	E07	30	25000	36500	2500	1500
POL20190901002	1	23	E05	22	6500	80000	529	506
POL20190901003	0	43	E04	23	4300	33000	1849	989
POL20190901004	0	65	E01	8	10000	75000	4225	520

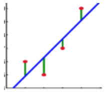
- Common feature engineering techniques include transformations (e.g. logarithm, powers), box-cox, interactions, splines, fractional polynomial, new ratios, one-hot encoding, binning, aggregation etc.
- Hand Crafted Feature Engineering is usually complicated and tedious, however encoding domain knowledge into the feature space could boost performance of predictive models
- Automatic Feature Engineering: representation learning such as PCA, the use of interactions from random forest, autoencoders in deep learning etc.
- **Feature Store** is a storage service for features to be registered, shared and used in ML pipelines
- Combination of automated and expert driven approaches

Modelling Module



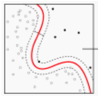
Modelling Module

Model Catalogue | Optimisation Metric | Hyperparameter Tuning

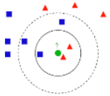


Linear Regression

GLM & Regularization

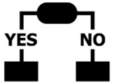


SVM

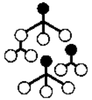


K-Nearest-Neighbour

Survival Modelling



Decision Tree

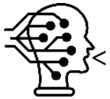


Random Forest

Gradient Boosted Machines (GBM)

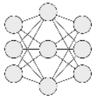
XGBoost

EXtreme Gradient Boosting



Natural Language Processing (NLP)

K-means clustering



Artificial Neural Network

+

Custom Model

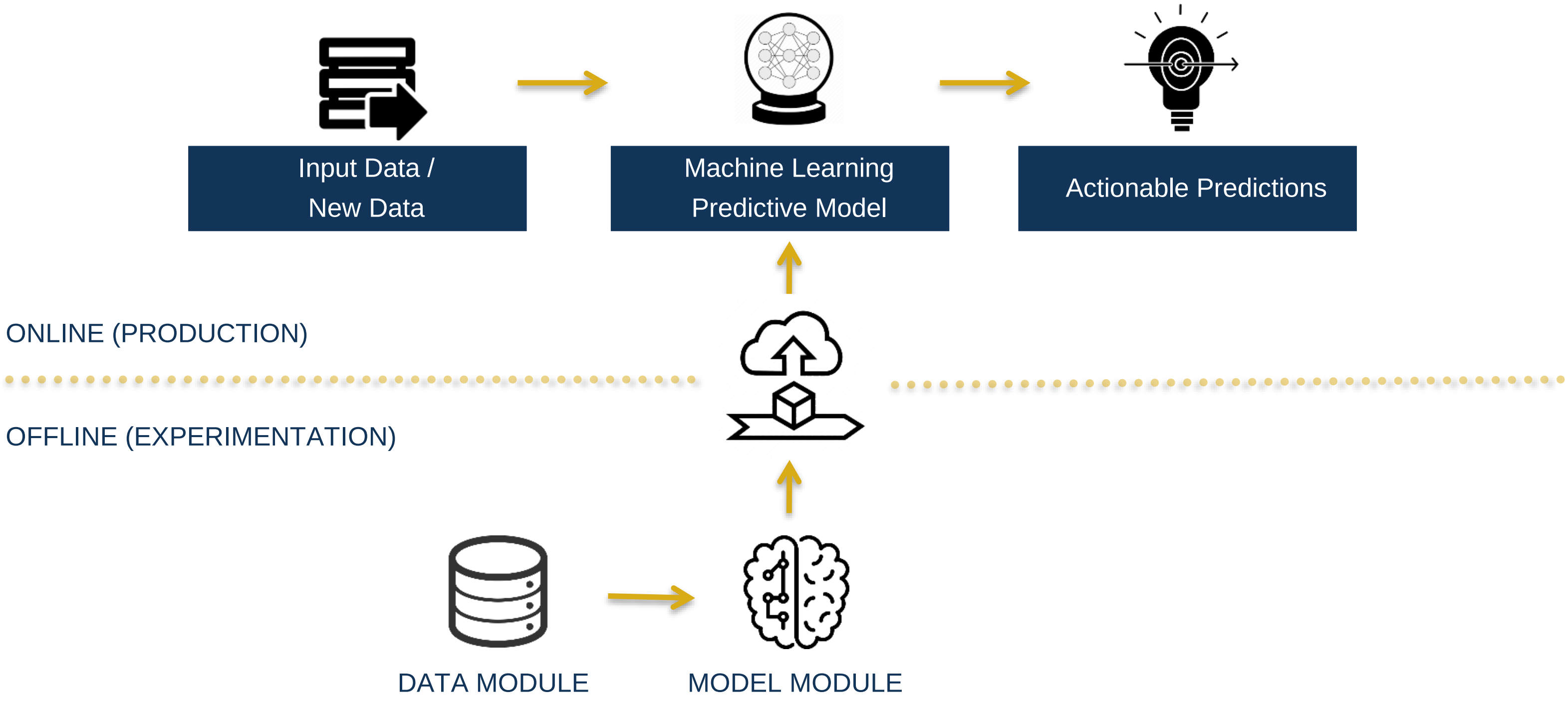
No Free Lunch Theorem

There's no such thing as a Free Lunch in supervised machine learning

In other words, there is no "super algorithm" that will work best for ALL datasets

See more discussion here

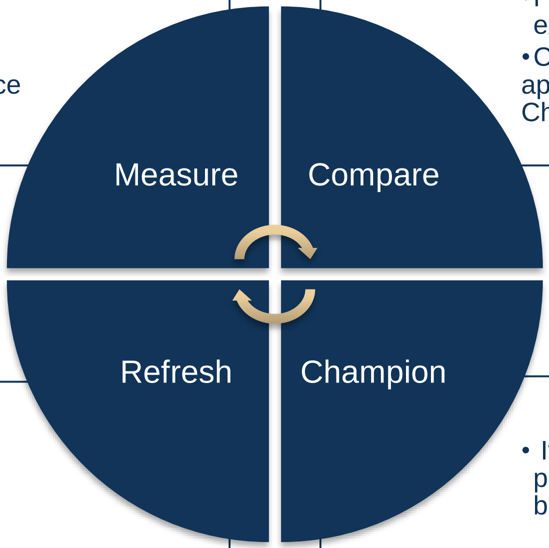
Deployment Module



Monitoring Module

- Constant monitoring of newly deployed model using agreed Performance Metrics
- Actual vs. Expected Performance

- Perform Champion Challenger experiment (a.k.a. A/B testing)
- Compare against incumbent best approach/ rule/ model (Current Champion)

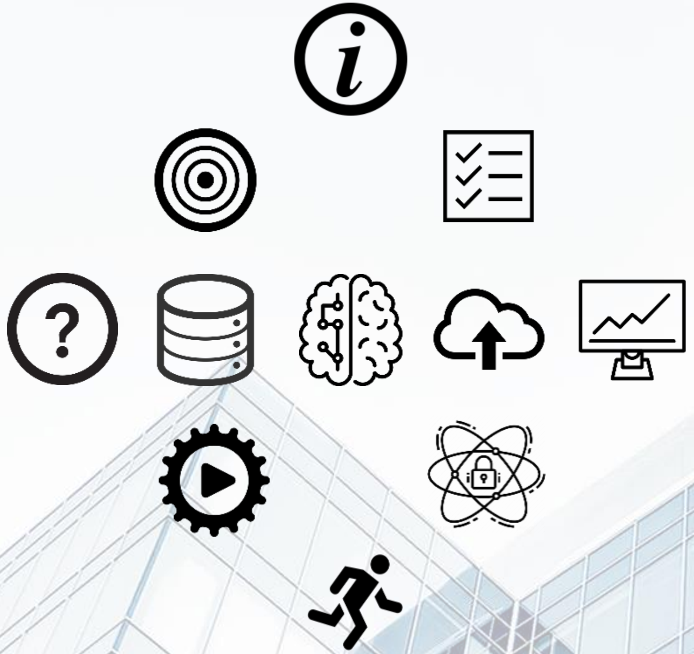


- Model will eventually degrade (change in data, market etc)
- Rebuild when metric dropped below a determined threshold

- If successful, promote the best performing challenger model to be the new Champion

Agenda

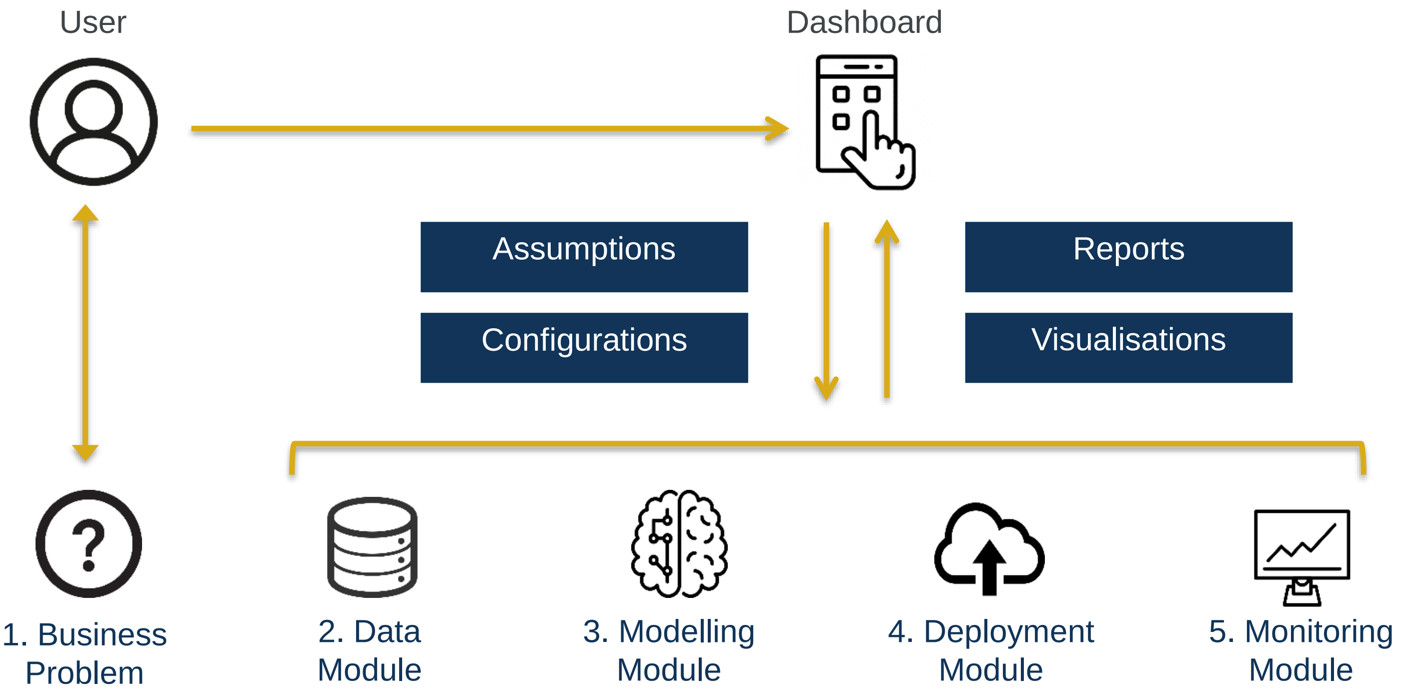
- Introduction
- Strategy
- ML Pipeline
- **Management**
- Application



Pipeline Operation and Automation

Speed	• Automation of Processes: Efficiency and Consistency • Simplify Machine Learning lifecycle development
Performance	• Best-in-class algorithms for better prediction accuracy • Leverage best practices in data across enterprise
Risk Management	• Automated Logging, Reporting, Audit Trail • Error Handling
Integration	• Integration into Enterprise • Common Platform for Business-As-Usual, R&D and Proof-Of-Concepts
Scalability	• Version Control (e.g. Git) • Scalability & Iterative Improvement

Pipeline Automation: Dashboard

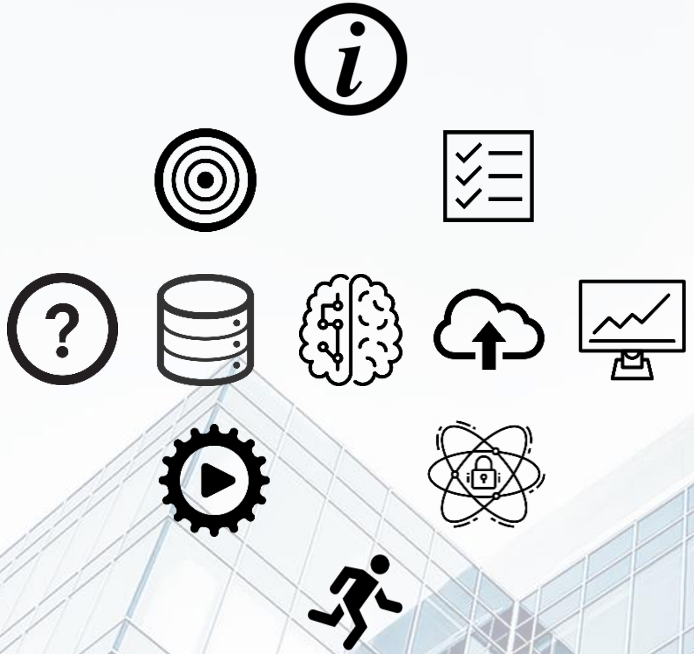


Pipeline Governance

- Ethics, Fairness
- Regulatory requirements
- Data Protection
- Data Lineage
- Model Explainability / Explainable AI (XAI)
    - SHAP, LIME, DeepLIFT, permutation feature importance
- Access Control and Security

Agenda

- Introduction
- Strategy
- ML Pipeline
- Management
- **Application**




Catch me …
if you can

Application 1: Fraud control

- Insurance fraud is estimated to cost insurers at least $40 billion per year in the US (FBI, non-health insurance) and £1.3 billion in the UK (ABI, 2016)
- Traditionally an agent investigate each case manually. This is time consuming and costly, increasing premium for honest customers. Pre-programmed rule-based systems are tedious too.
- Classification ML pipeline could help agents to detect fraud faster, and to detect as many as possible (true positive) while to not mistakenly flagging excessive amount of non-fraudulent claims (false positive)
- Challenges:
  - High imbalance data (due to very low fraud rate)
  - Optimising popular metrics such as “Accuracy” or “AUC” is not ideal. A better approach is to select threshold based on

    value metric = sum of saved claim amounts (from true positive) – wasted investigation cost (from false positive)
- Benefits of pipeline:
  - Integrating financial impact to the business; improve profitability
  - As fraudsters get more sophisticated and creative, an ML pipeline system is capable of monitoring and frequent model refresh

Example of fraud control implementation

1. Business Problem

- Motor, home or health insurance
- Engage claims managers and business experts
- Reduce Fraud
- Optimise resources
- Improve profitability

2. Data Module

- Claims history, frequencies, amounts
- Attributes of policyholder, policy, insured risk
- Fraudulent claims

3. Modelling Module

- Binary Classifier (supervised learning)
- Suitable performance metric
- Balancing of classes
- Example algorithms: Random Forest, XGBoost, Lasso, Neural Networks, NLP

4. Deployment Module

- Integrated into business as Recommender system

5. Monitoring Module

- A/B Test against incumbent
- Or A/B Test different approaches
- Monitor performance and economic value
- Monitor model degradation

"Fraud Control is a dynamic game"

Application 2: Risk modelling / Pricing

- Risk modelling involves predicting risk, claim cost or “technical price” as accurately as possible
- ML pipeline focuses on predictive accuracy and less dependent on assumptions on models
- Machine Learning pipeline could be used to
  - Run a “league” and select the best risk model (AutoML)
  - Estimate value of external data enrichment and assess performance of lift curves
  - Compare and measure different ways of building models (quick experimentation)
  - Automate like a production line
  - Automate reporting and audit trail
- Always measure the additional performance gained from a more complex model vs a simpler baseline model
- Free up more time to consider interpretability, potential biases, ethical issues in pricing

Five Models of Pricing Operation

Tariff
- Regulator has significant influence over the rates

Qualitative
- “Correct” pricing cannot be determined purely by numerical analysis and subjective factors play a significant role
- Data maybe incomplete or not exist

Cost Plus
- Statistically driven analysis
- Based on **expected cost of claims**, appropriately loaded for expenses, profit etc
- Typically single distribution channel

Distribution
- Price also allows for non cost elements such as **propensity to shop around, price elasticity**
- Pricing strategy for similar products being managed across multiple distribution channels

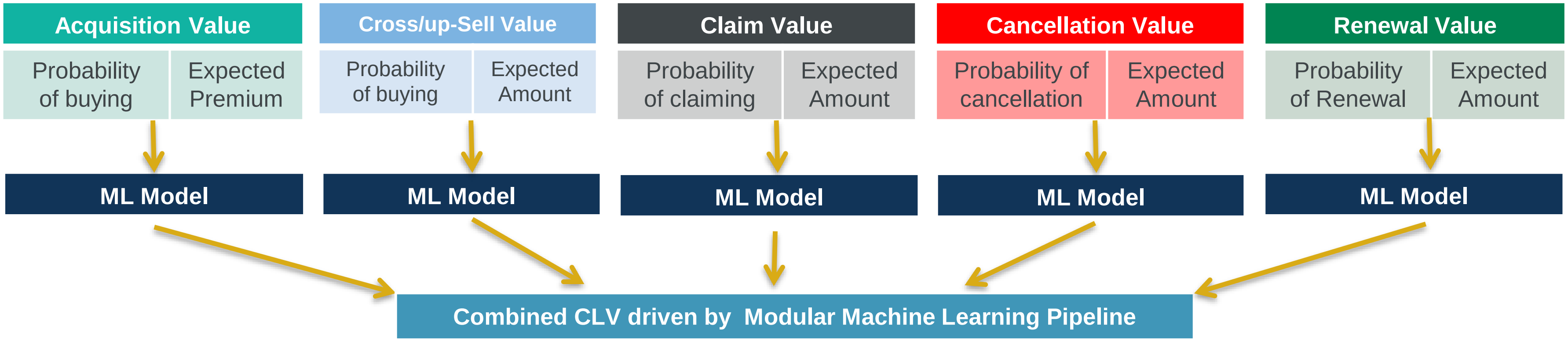
Industrial
- Typically domain of very large insurers
- multiple brands, channels, countries
- **Machine oriented approach**
- Focus on operating efficiency and economies of scale

↔ where Machine Learning Pipeline can add value

[Source: GRIP report]

Application 3: Customer Lifetime Value (CLV)

- Definition: *The net present value of a customer during entire relationship with the company*
- Customer Lifetime Value = Present value + Future Value
  - Present value = Premiums + cross/up-sell revenue – Claim costs – Activity-based costs (ABC)
  - Future value = (Premiums + cross/up-sell revenue – Claim costs – Activity-based costs (ABC) – Cancellation)/$(1+i)^t$



Application 3: Customer Lifetime Value Segmentation

CLV ML pipeline helps you to make smart decisions (decision science) and grow business

New Customers: Acquisition Lifetime Value

- Pricing
- Inform <u>marketing</u> target profiles
- Generate <u>sales</u> leads for new customers + prioritisation
- Manage customer <u>service</u> resources
- Cross sell and up sell
- Personalised products
- Product designs or features
- Channel optimisation (affinity partners, price comparison websites)

Existing Customers: Future Lifetime Value

High Value customers

- Cross sell and up sell
- Reduce churn and improve persistency
- Personalised servicing
- Selective discounting and offers

Low Value customers

- Termination or reduce cost of service

Application 3: Customer Lifetime Value Optimisation

CLV Machine Learning Pipeline

➕

Optimiser

Commercial Price associated with
Optimal portfolio CLV value
(subject to constraints)

- Pricing Optimisation
- Granular Price Elasticity
- Dynamic Pricing

Application 4: Mortality

- Data on event, exposure and other risk factors
- Approaches: Traditional, Poisson GLM, XGBoost, Random forest, cox, survival modelling, deep learning, deep survival analysis …
- GLM is rather commonly used (besides MS Excel), but struggles with speed for large volume of data, variable selection and non-linear predictive factors
- Machine Learning pipeline helps with:
  - Speed and Accuracy
  - Granular risk factors extraction and selection, underwriting and Claims management
  - Basis setting
  - Integrating and valuating new potential data sources such as wearables, genome sequencing, search engine, social media

Application 5: Unstructured Data

- In Natural Language Processing (NLP), the “Feature engineering” element in ML pipeline is also known as “Transformer”. Tokenizer is one type of transformer that maps the original values (words) with new ones (numbers), for example N-gram tokenization.
- Idea: “Unstructured” → “Structured”. Then run through ML pipeline.
- Applications:
  - Intelligent document analysis: Assist claim adjusters in analyzing large volume of reports and emails (for example those that involve bodily injury), set more accurate reserves by more consistent claims handling
  - Improve customer service interactions by lowering friction
  - Sentiment analysis, also known as opinion mining
- Example: See IFoA webinar on a recent end-to-end application of ML Pipeline on unstructured data: “Twitter Sentiment Analysis: What does social media tells us about coronavirus concerns in the UK?”
  - View slides

How to start applying this framework?

Once having the right team, technology and data:

1. Identify opportunities and the right questions with champions (strong stakeholders)
2. Aim for quick wins of high impact with relative low effort, then create business case
3. Build Minimum-Viable-Product (MVP) that is scalable; Modelling and Deployment
4. Communicate, Monitor and Review performance
5. Scaling and Maintenance

Actuaries, having business domain and statistical knowledge, could harness the strength of data science and champion data-driven advancements at organisational level.

Actuaries can become ***Revolutionary.***

Questions

Comments

The views expressed in this publication are those of invited contributors and not necessarily those of the IFoA. The IFoA do not endorse any of the views stated, nor any claims or representations made in this publication and accept no responsibility or liability to any person for loss or damage suffered as a consequence of their placing reliance upon any view, claim or representation made in this publication.

The information and expressions of opinion contained in this publication are not intended to be a comprehensive study, nor to provide actuarial advice or advice of any nature and should not be treated as a substitute for specific advice concerning individual situations. On no account may any part of this publication be reproduced without the written permission of the *authors.*