MAKE WITH H2O.ai

H2O.ai

Day in the Life of a **Data Scientist in Industry**

Tianchu Zhao

Agenda

H2O.ai

- Introduction (about me)
  - Work Experience
  - What is H2O.ai?
- Day in the Life of a Data Scientist in Industry
  - Company journey
  - The type/stage of companies
  - The projects
  - The roles
  - The tools
- How to get a data science job?
- Q&A

(Are the things we learn in this course helpful in a future job)

Remember to put down your question in Zoom

Introduction

Tianchu Zhao

H2O.ai

H2O.ai

Machine Learning Engineer

Commonwealth Bank Australia

Machine Learning Engineer

Argo Project

Software Engineer

WooliesX (Woolworths)

Data Scientist

What is H2O.ai?

Democratize AI with H2O.ai

33

Kaggle Grandmasters

World's #1, #2, #5, and #9

200K

Community & Companies

Founded in Silicon Valley, 2012

Investors (Series E): Goldman Sachs, Ping An, Wells Fargo, NVIDIA, Capital One, Nexus Ventures, Commonwealth Bank Australia

222 OF THE FORTUNE 500

❤ **H2O**

8 OF THE TOP 10 **BANKS**

7 OF THE TOP 10 **INSURANCE COMPANIES**

4 OF THE TOP 10 **MANUFACTURING COMPANIES**

AIaaS : AI as a Service using H2O AI

Datasets

AI Engines

H2O DRIVERLESSAI

H2O3

H2O DOCUMENT AI

H2O HYDROGEN TORCH

H2O LABEL GENIE

AI Engine Manager

Deployment

H2O FEATURE STORE

MLOps

Consumption

wave

AppStore

End Users

H2O LLMs Ecosystem

h2oGPT

H2O LLMSTUDIO

LLM DataStudio

LLM AppStudio

LLM EvalStudio

Doc-QA

Helium

Successful Customers Across Industries and Use Cases

Financial Services

Wholesale / Commercial Banking
- Know Your Customers (KYC)
- Anti-Money Laundering (AML)

Card / Payments Business
- Transaction frauds
- Collusion fraud
- Real-time targeting
- Credit risk scoring
- In-context promotion

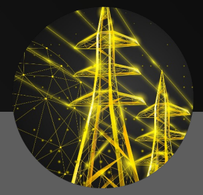
Retail Banking
- Deposit fraud
- Customer churn prediction
- Auto-loan

Commonwealth Bank · Wells Fargo · Bank of America Merrill Lynch

Healthcare and Life Science

- Early cancer detection
- Product recommendations
- Personalized prescription matching
- Medical claim fraud detection
- Flu season prediction
- Drug discovery
- ER and hospital management
- Remote patient monitoring
- Medical test predictions

UCSF Health · Kaiser Permanente · Bon Secours Mercy Health



Telecom

- Predictive maintenance
- Avoidable truck-rolls
- Customer churn prediction
- Improved customer viewing experience
- Master data management
- In-context promotions
- Intelligent ad placements
- Personalized program recommendations

AT&T · Cisco · Reliance Industries Limited

Marketing and Retail

- Funnel predictions
- Personalized ads
- Fraud detection
- Next best offer
- Next best action
- Customer segmentation
- Customer churn
- Customer recommendations
- Ad predictions and fraud

ebay · Walgreens · ADT

Democratize AI and Accelerate AI Results

Long AI Projects that Add No Value

09
MONTHS

Step 1

Data Scientist Has an Idea or is Told to work on a Project

Step 2

Toil on a notebook creating features and optimizing to build a “perfect” model

Step 3

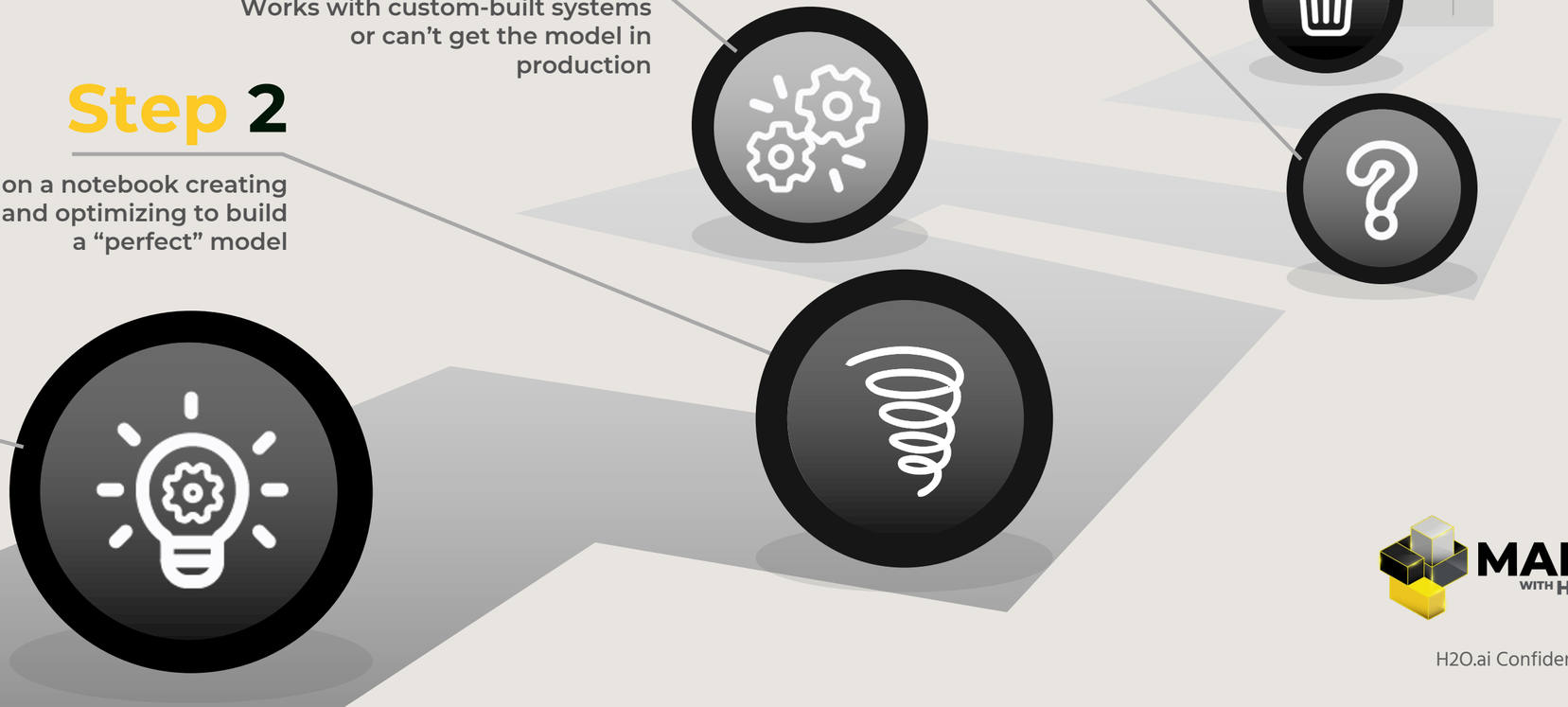
Works with custom-built systems or can’t get the model in production

Step 4

Hands predictions to the business user, who doesn’t understand the model or why the work was done in the first place

Step 5

Not Used



MAKE WITH H2O.ai

Fast And Successful AI Projects

02
WEEKS

Step 1

Data Scientist, Business Analyst, Developer, or Data Engineer Works with Business Owner on the Problem

Step 2

Uses No Code or AutoML services and builds highly accurate models in days or hours

Step 3

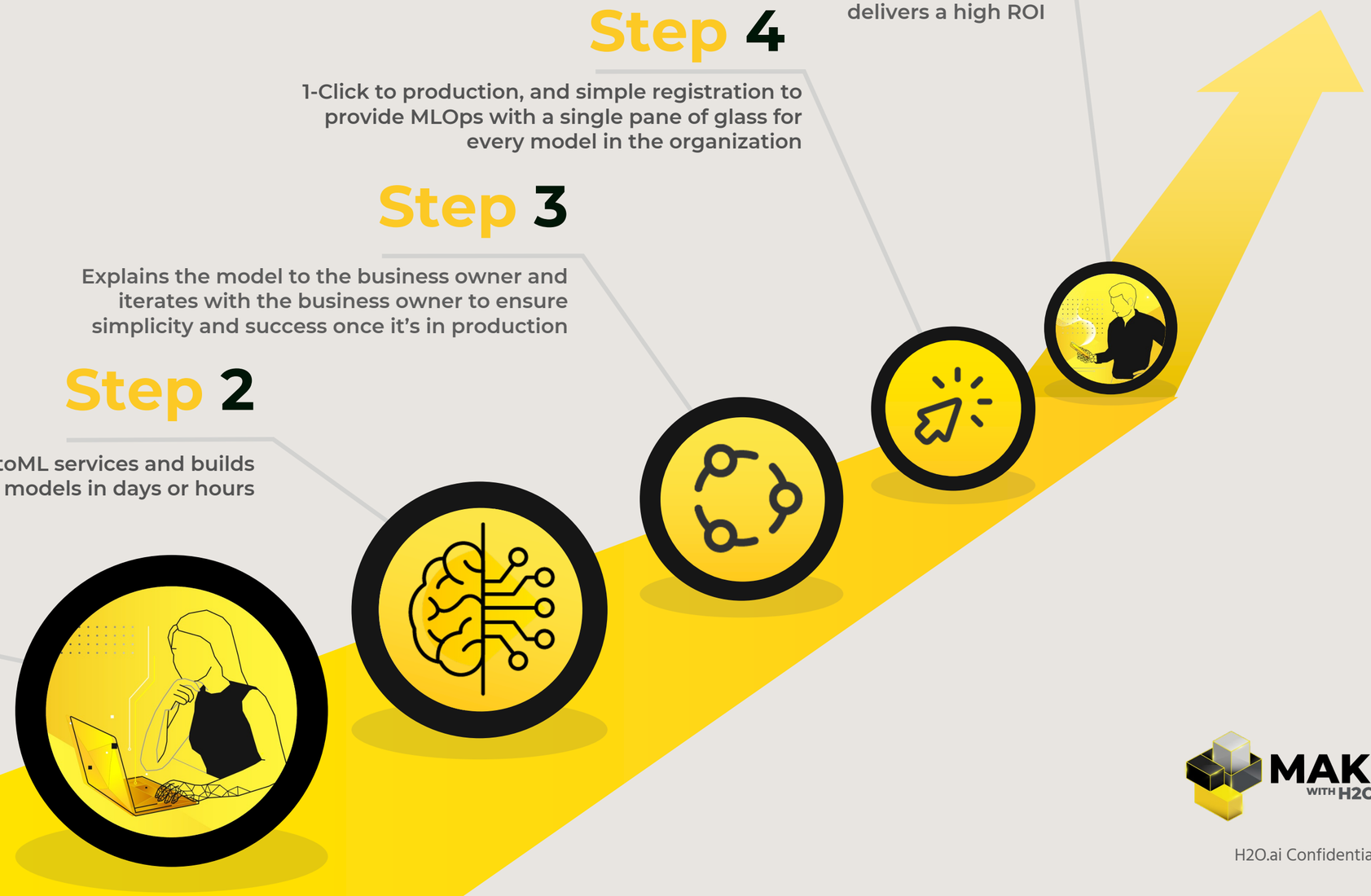
Explains the model to the business owner and iterates with the business owner to ensure simplicity and success once it's in production

Step 4

1-Click to production, and simple registration to provide MLOps with a single pane of glass for every model in the organization

Step 5

Low-code departmental app is built, or AI is integrated into existing apps or databases. It is used by business owners, and delivers a high ROI



MAKE WITH H2O.ai

Go To: DAI Documentation

H2O.ai Experiment 59f644

1.0.9

TRAINING DATA

DATASET
BNPParibas-train.csv

ROWS	COLUMNS	DROPPED COLS	TEST DATASET
114K	133	--	Yes

TARGET COLUMN
target

FOLD COLUMN
--

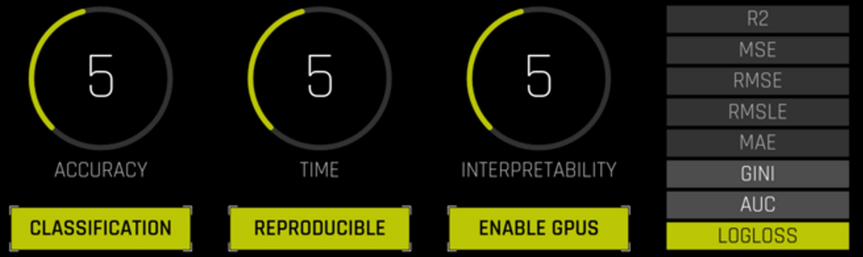
WEIGHT COLUMN
--

TIME COLUMN
--

TYPE	COUNT	UNIQUE	FREQ
int	114321	2	27300

SCORED 217/434 MODELS ON 3572 FEATURES

EXPERIMENT SETTINGS



CPU / MEMORY

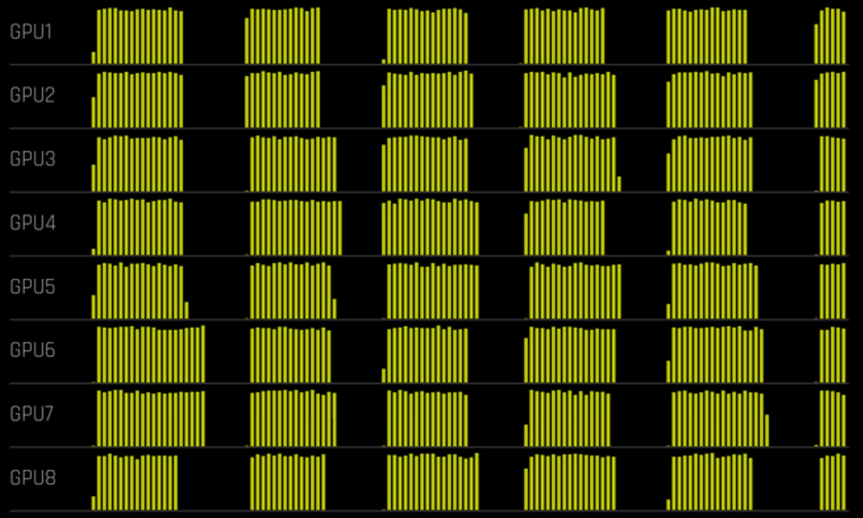
Trace

ITERATION SCORES - INTERNAL VALIDATION

VARIABLE IMPORTANCE

89_v50	1.00
139_NumCatTE_v3_v31_v5_v50_v56_v6_v66_v7_v79_0	0.53
16_CV_TE_v66_0	0.27
149_WoE_v113_v22_v30_v72_v75_0	0.25
14_CV_TE_v56_0	0.15
3_CV_TE_v113_0	0.13
147_NumCatTE_v3_v31_v5_v50_v6_v66_0	0.13
10_CV_TE_v31_0	0.10
21_CV_TE_v79_0	0.09
18_CV_TE_v72_0	0.09
150_WoE_v22_v30_v56_v66_v75_0	0.08
141_NumCatTE_v5_v56_0	0.08
146_NumCatTE_v2_v22_v3_v31_v5_v50_0	0.07
72_v34	0.06

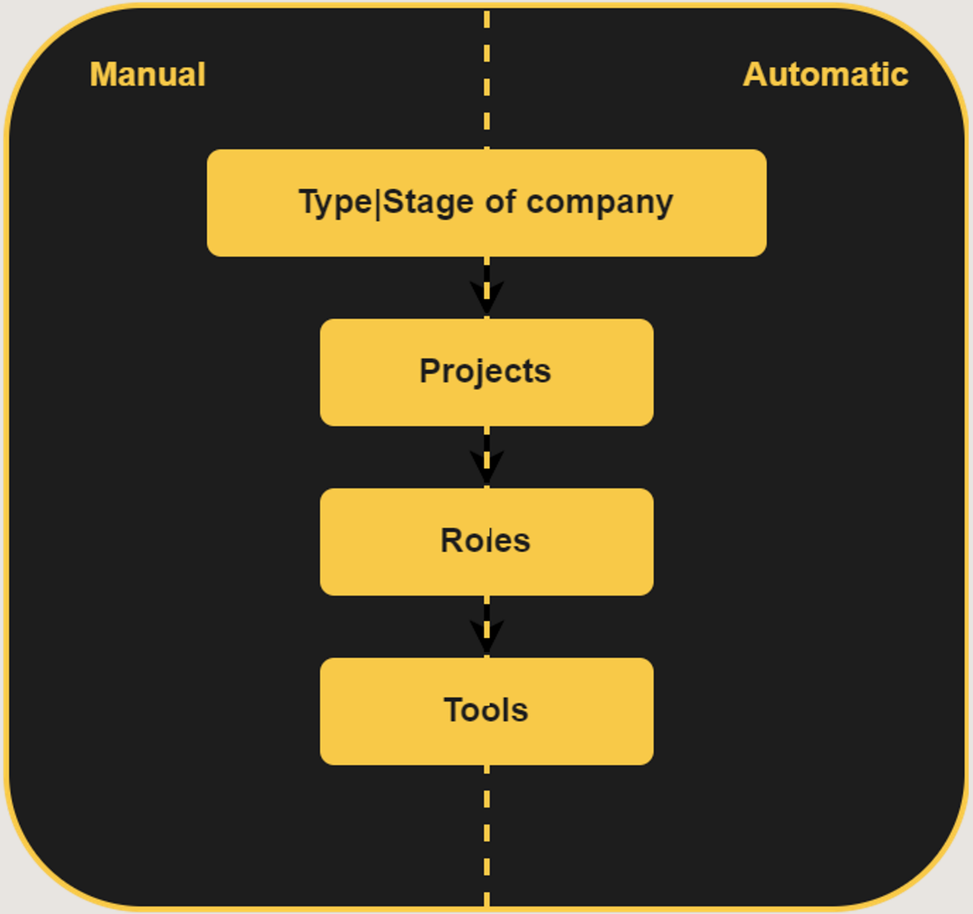
GPU USAGE



Day in the Life of a **Data Scientist in Industry**

Perspective

- Data science journey of company
- The Type/Stage of company
- The Projects
- The Roles
- The Tools

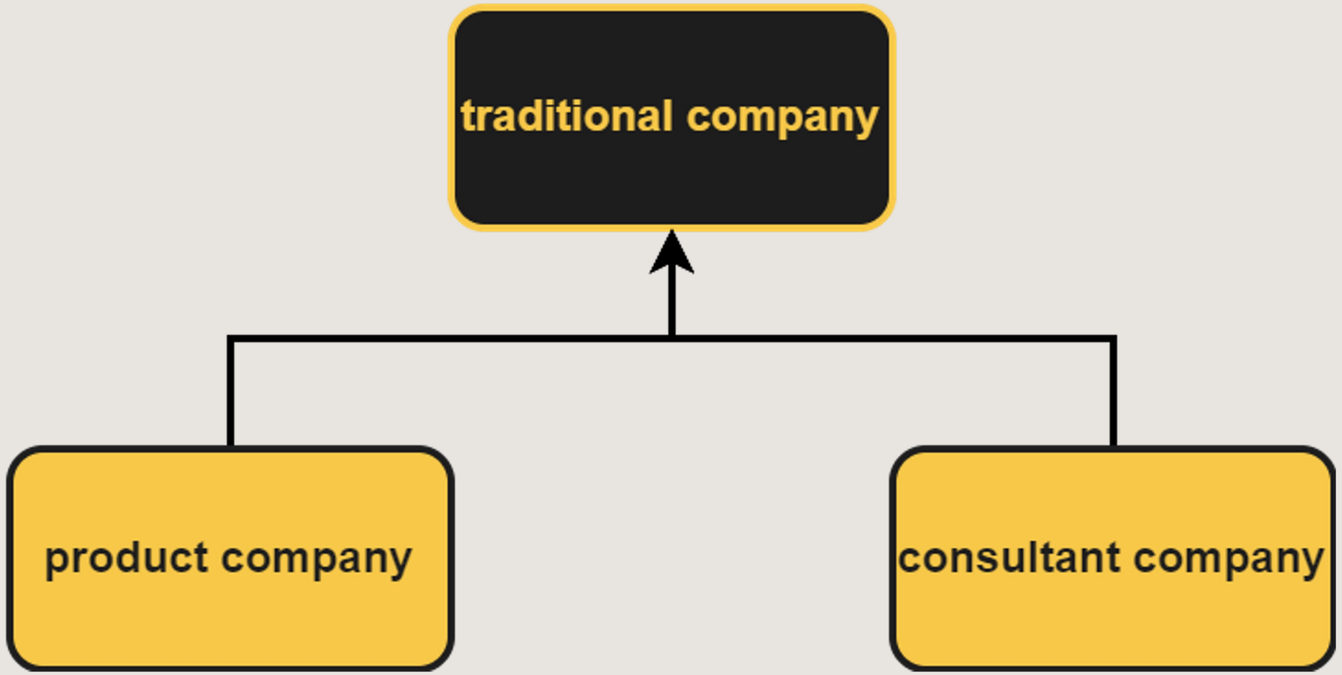


The type/stage of company

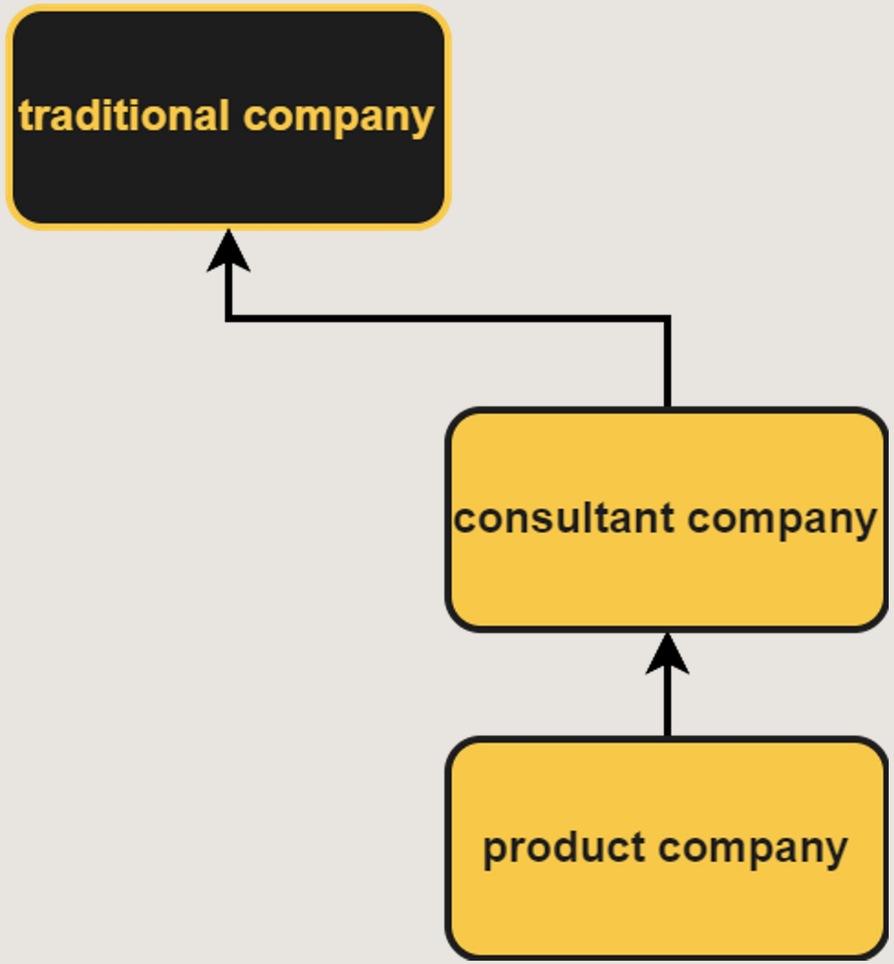
The Type of company

- **Traditional company**
  - Traditional company with its own revenue stream that doesn’t rely on data science
  - e.g. Woolworths, CBA
- **Product company**
  - Build product then license/sell to other companies (the normal company)
  - e.g. H2O.ai, Snowflake, Databrick
- **Consultant company**
  - Offer data science consultancy, usually works at client (the normal company) site/project
  - e.g. Quantium

The Type of company



The Type of company



The Stage of (traditional) company

Company journey

- **0-1 Company**
  - Company yet to have established data science capability
  - e.g. they are building data science model, aiming to achieve better __
  - e.g. they have data science project proof of concept but yet to productionise it
- **1-1+ Company**
  - Company with productionised data science project
  - e.g. They have a simple but effective logistic model but looking to advance it further
  - e.g. more data science projects, centralised metric monitoring……

The Projects

Projects

- **0-1 company (proving data science works)**
  - data Refactoring
  - new data science pipeline
    - business problem -> curate data -> modelling -> production
- **1-1+ company**
  - data refactoring
  - new data science pipeline
    - business problem -> curate data -> modelling -> production
  - refine/maintain existing data science pipeline

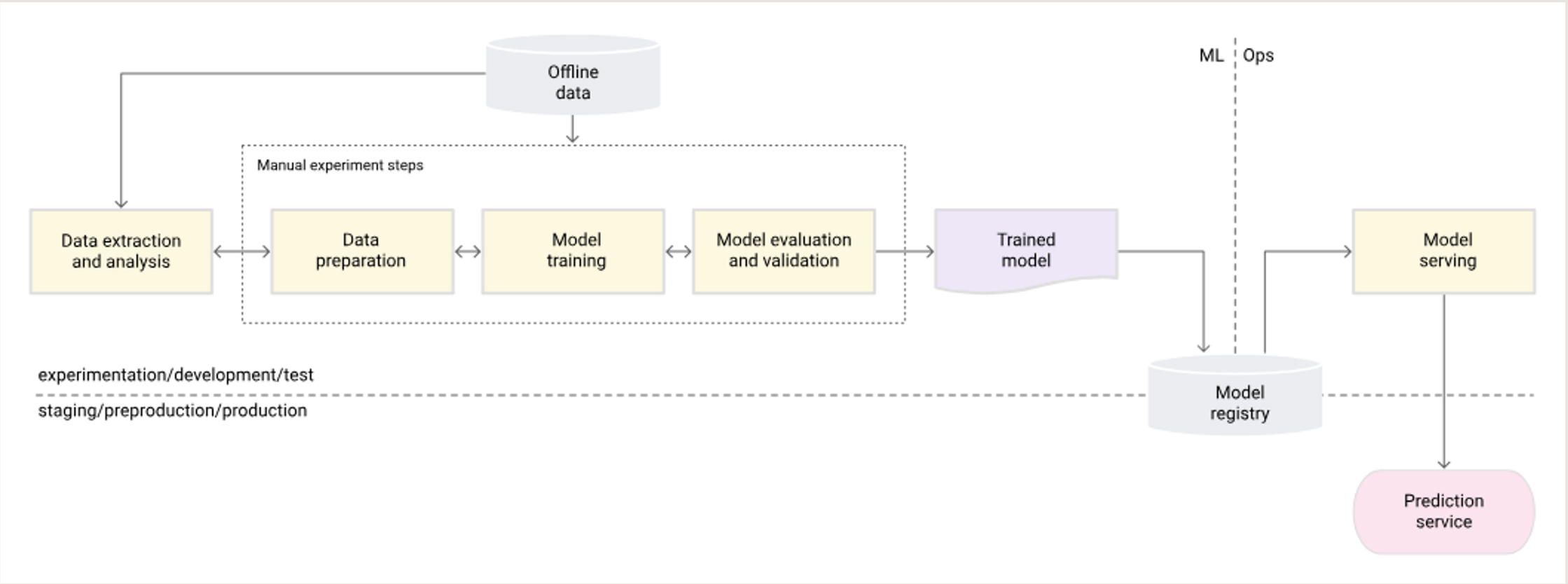
The Data Science Journey of a Company

Data Science Journey of A Company (or personal project)

- scope of project
- different tools
- different roles

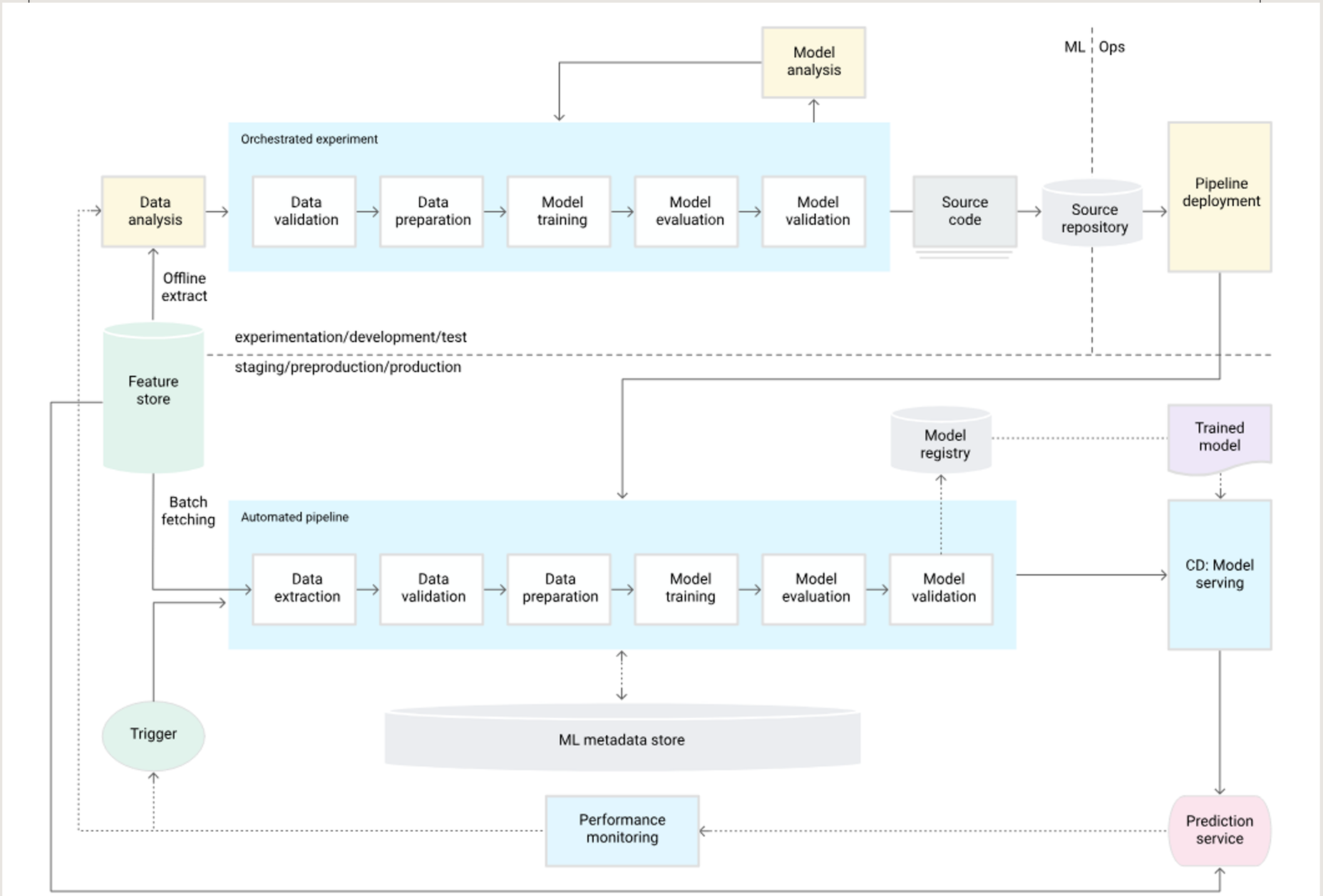
- Journey: Manual -> Automation -> More Automation

Manual process



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

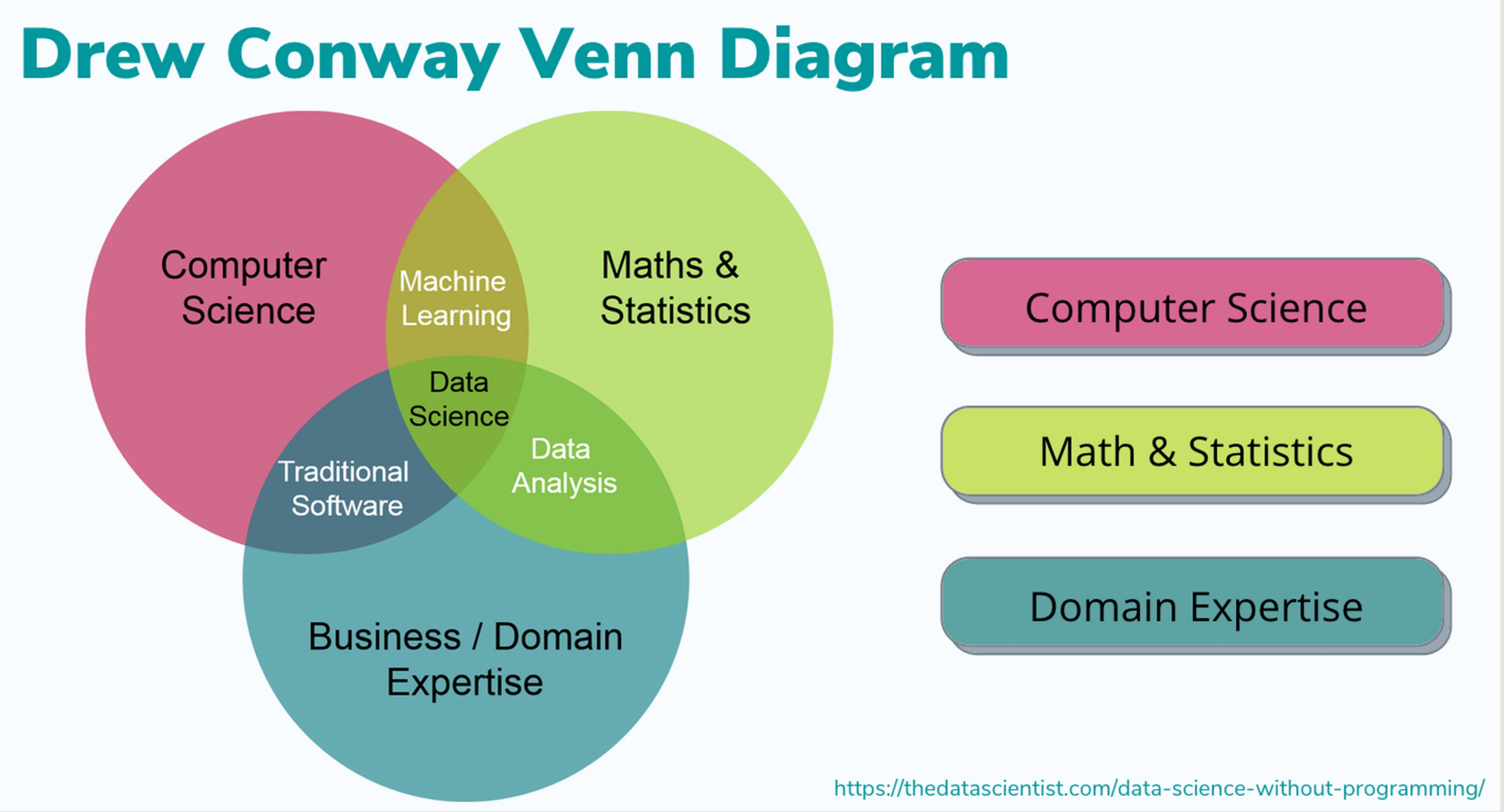
ML pipeline automation



https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

The Roles

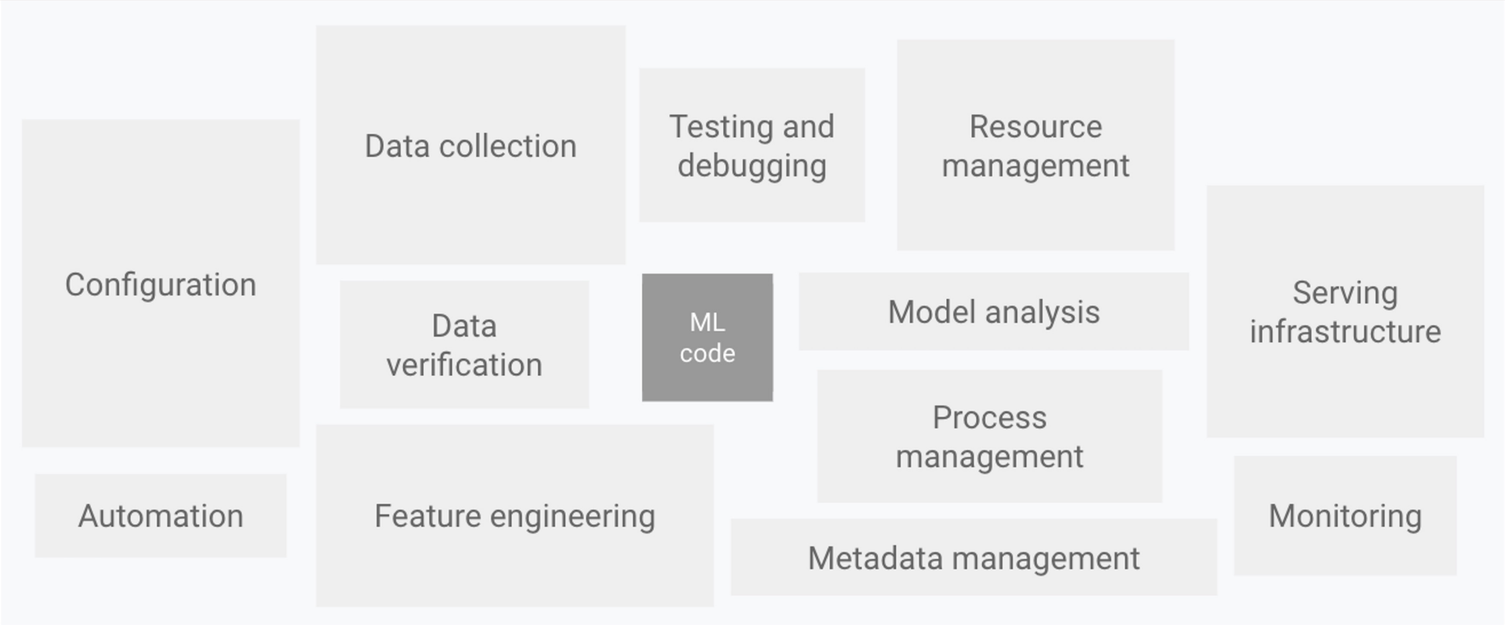
Roles


Drew Conway Venn Diagram
Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise
Computer Science
Math & Statistics
Domain Expertise
https://thedatascientist.com/data-science-without-programming/

Roles

- **Data Scientist**
  - could mean: data analyst/data engineer/machine learning engineer
  - could be jack of all trade
  - depending on the company
- **Data Analyst**
  - answer business question through data exploration
- **Data Engineer**
  - building data pipeline to clean/format data into usable format
- **Machine Learning Engineer**
  - generally anything to do with machine learning model, building/productionising

Roles



Hidden Technical Debt in Machine Learning Systems

Roles

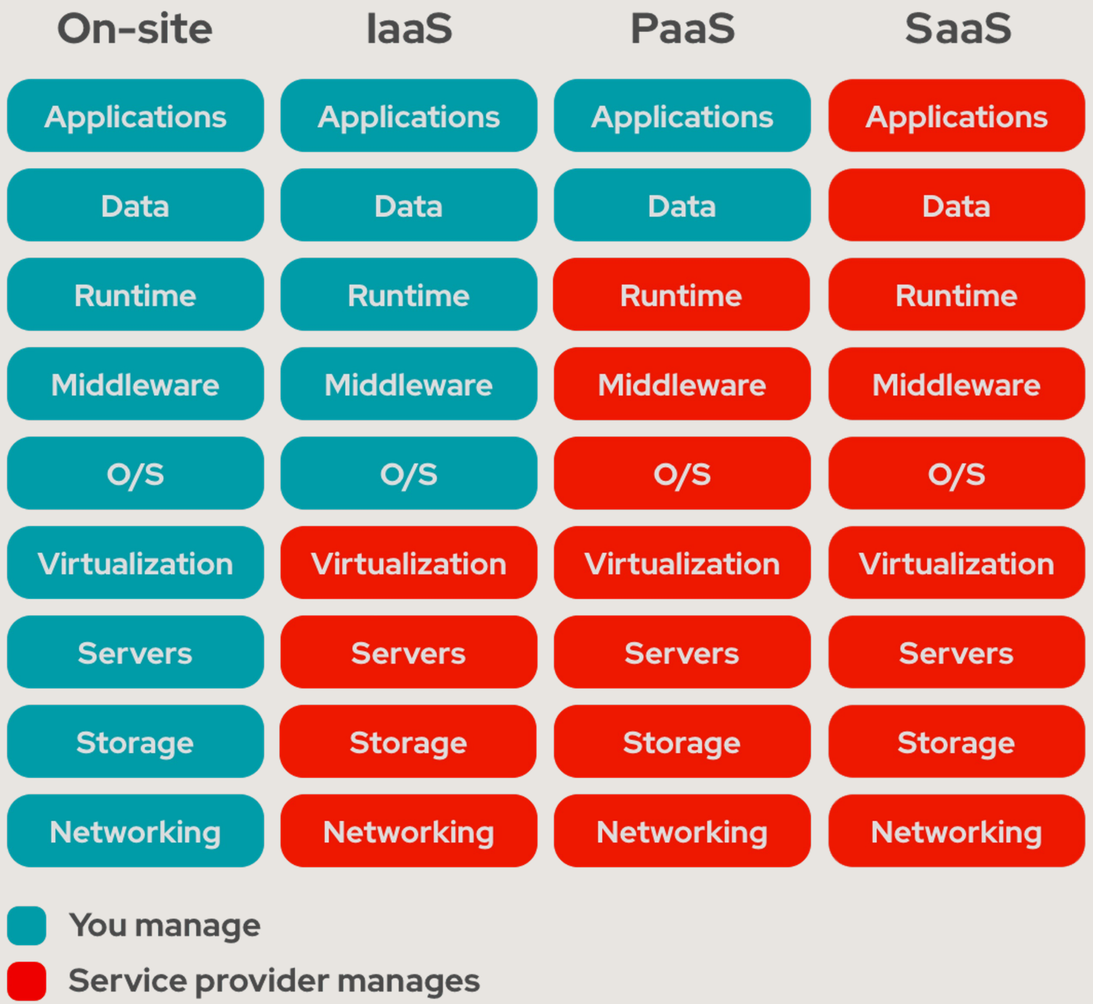
- **Data Scientist**
  - could be
  - more interacting with business understanding requirement
  - more coding to explore data
  - more coding to building the data pipeline
  - more coding to productionising the model
  - more reading to understand state of the art technique
  - more presentation to showcase the project outcome
  - ……

The Tools

Tools

- **Python, Jupyter, Pandas, Scikit-Learn**

Tools



https://www.redhat.com/en/topics/cloud-computing/iaas-vs-paas-vs-saas

Tools

- The project may have a way of working using
  - Python, Jupyter, Pandas, Scikit-Learn
  - **Write codes using packages**
- The project may have a way of working using
  - Home made software or SAAS (e.g. H2O DriverlessAI)
  - **Configure, Click and Run**
- Either way you need the **expertise in interpreting the result**

Tools

- With your foundation knowledge (e.g. getting comfortable at a specific tool)
- You could **pick up any new tool** to use it (Transferred knowledge)
- Workflow:
    - I **know what I want to achieve** -> **figure out how** to achieve it using the tools available

How to get a data science job?

How to get a data science job?

- You will find your way
- Put your effort in
- Be patient

Tianchu Zhao

H2O.ai

H2O.ai

Machine Learning Engineer

Commonwealth Bank Australia

Machine Learning Engineer

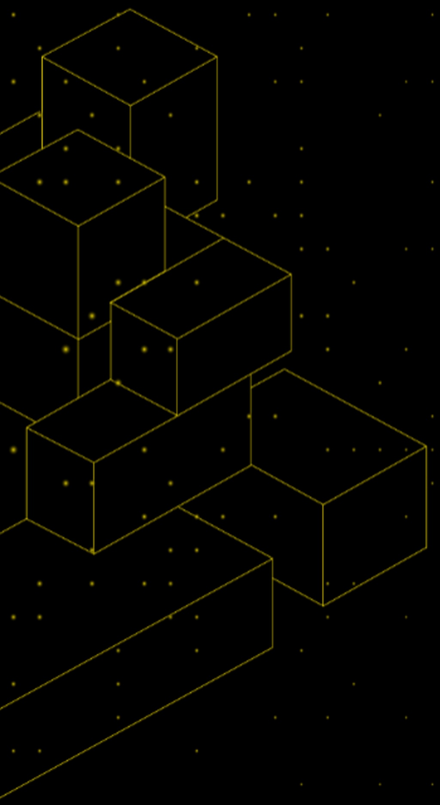
Argo Project

Software Engineer

WooliesX (Woolworths)

Data Scientist

Q&A



Thank you