CORE Skills

Delivering Data Science
In Resources & Energy

Zero to Data Science in a Day: Data Culture, Data Science & Data Projects

DAY 3

15-Day Data Science Springboard

Dr Jeremy Mitchell
Data Mettle

Dr Ying Yap
Data Mettle

Program partners

_dm

The University of Western Australia

Curtin University

A CORE partnership

Program Timeline

DAY 3: Introduction to Data Projects

CORE Skills

Prerequisite		Introduction to Data Projects	Data Analysis			Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time - series Data	Special Data Types - Natural Language Processing and Text Mining	Special Data Types - Spatial Data	Capstone Project Development & Presentation	Capstone Propeller
Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Thurs 30 Jan 2020	Fri 31 Jan 2020	Thurs 6 Feb 2020	Thurs 13 Feb 2020	Thurs 20 Feb 2020	Thurs 27 Feb 2020	Fri 28 Feb 2020	Thurs 5 Mar 2020	Thurs 12 Mar 2020	Fri 13 Mar 2020	Thurs 19 Mar 2020	Thurs 26 Mar 2020	Thurs 2 Apr 2020	Fri 3 Apr 2020	Thurs 2 July 2020
Introduction to the program tools - Where Data Science comes from	Introduction to the program tools - The Why	Zero to Data Science in a day	Getting to know the program tools - data munging and exploratory data analysis	Simple predictions - regression and statistical model building	Multivariate analysis and model building	Effective data storytelling - communicating results to non-technical audiences	Pros and cons of commonly used statistical and machine learning techniques I	Pros and cons of commonly used statistical and machine learning techniques II	Sandbox - Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Pitching Capstone Projects	Project Review Day

Q&A, Issues & Announcements

Before we Get Started

- Resources & Tasks in **Github**
- Reflections, questions and issues on set up
- We'll start talking about **projects** a bit this afternoon, and help you start setting up yours

Schedule

DAY 3

AWST	AEST	Agenda	
07:30	**09:30**	Q&A, Issues & Announcements	Educator
07:45	09:45	**Creating a Data Culture**	
09:15	11:15	*Morning Tea*	Jeremy
09:30	11:30	**Getting From Here to There**	
11:00	13:00	***Lunch***	Jeremy
11:45	13:45	**Data Science Projects**	
13:15	15:15	*Afternoon Tea*	Jeremy
13:30	15:30	**Data Exploration, Modelling and Reporting**	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	Close	

Aims & Learning Outcomes

DAY 3

CORE Skills

Aims

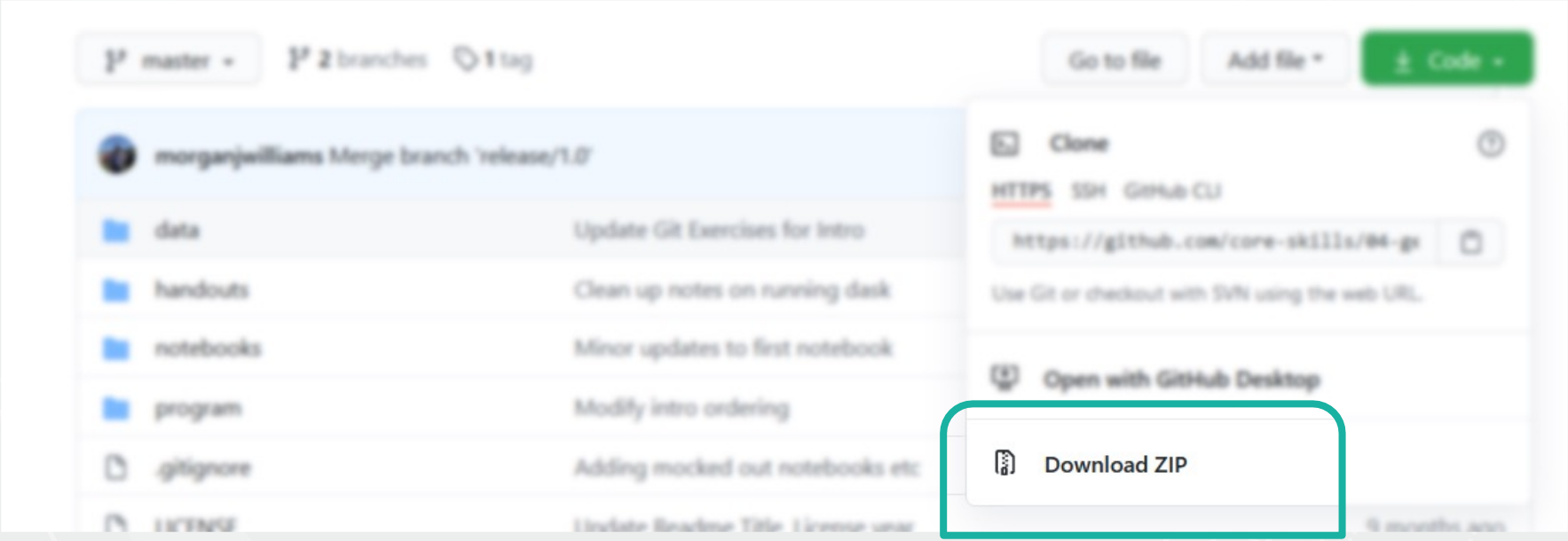
- Provide an overview of a 'typical' data science workflow.
- Go through setting up data science projects.

Learning Outcomes

- To appreciate what data science is.
- Understand the stages of a data science project and define a mental model of it
- Analyse the opportunity and potential value of data science in your organisation
- Practise with git, Python and Jupyter notebooks.

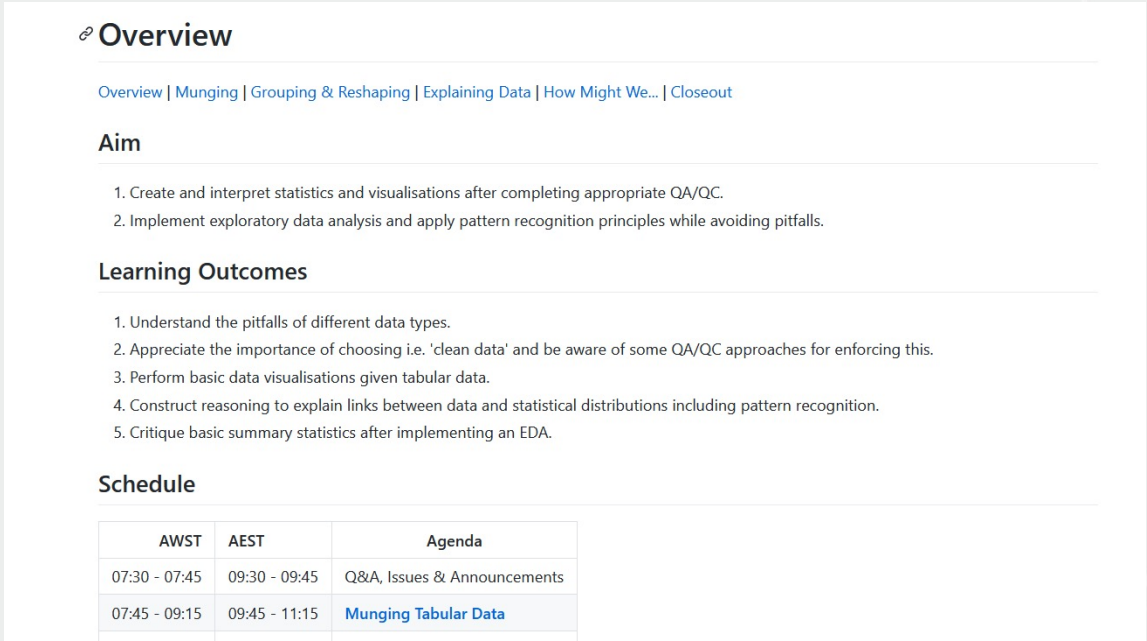
GitHub Content for Today

github.com / core-skills / 04-getting-to-know-the-tools



GitHub – Program Notes

[github.com / core-skills / 04-getting-to-know-the-tools / program / 00_overview.md]



Overview

Overview | Munging | Grouping & Reshaping | Explaining Data | How Might We... | Closeout

Aim

1. Create and interpret statistics and visualisations after completing appropriate QA/QC.
2. Implement exploratory data analysis and apply pattern recognition principles while avoiding pitfalls.

Learning Outcomes

1. Understand the pitfalls of different data types.
2. Appreciate the importance of choosing i.e. 'clean data' and be aware of some QA/QC approaches for enforcing this.
3. Perform basic data visualisations given tabular data.
4. Construct reasoning to explain links between data and statistical distributions including pattern recognition.
5. Critique basic summary statistics after implementing an EDA.

Schedule

AWST	AEST	Agenda
07:30 - 07:45	09:30 - 09:45	Q&A, Issues & Announcements
07:45 - 09:15	09:45 - 11:15	Munging Tabular Data

Environment

- Open an Anaconda Prompt
- Navigate to where you have the unzipped repository material

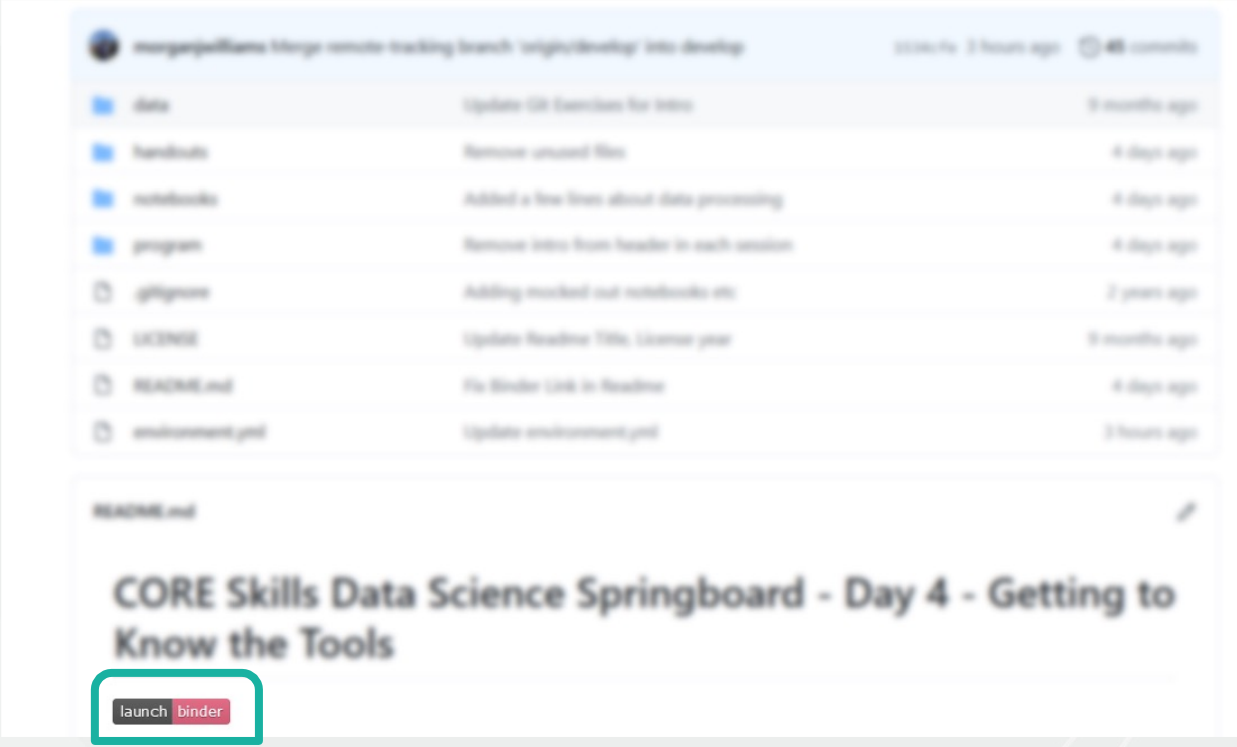
```
conda env create -f environment.yml

conda activate core04

python -m ipykernel install --user --name=core04

jupyter lab
```

Binder Backup



High-level Takeaways From Today

- **Different types of data culture** *Why are you doing data?*
- **Data literacy** *Who are you talking to?*
- **Data strategy** *How will you do it?*
- **Data science teams** *Who will you work with?*
- **Basic data science workflow** *What will yours involve?*
- **Delivering incremental change** *What does it look like?*

Fostering a Data Culture

This morning we'll talk about people.

What is Data Culture?

Beyond the buzz words, what does this mean?

Data Therapy suggests that the following list characterises businesses which have a data culture:

- Leadership prioritizes and invests in data collection, management, and analytics
- Leadership prioritizes **creative data literacy** for the whole organization
- Staff are encouraged and supported to access, combine and derive insight from the organization's data
- Staff recognize useful data when they see it.
  They offer creative ways to use the organizations data to solve problems, make decisions and tell stories

What is Data Culture?

Taking some of this apart:

- Data is treated as an asset
- Organisations create capability to realize value from this data.
  Data alone is not valuable, it can be leveraged to create value.
  The better your data management and data enrichment, the longer your data will remain useful.
- There are many opportunities on the leadership side, but in the end employees need to feel empowered to work with their data.
- Working across silos to leverage a workforce with common purpose, skills and technology.

Why build a data culture?

Some of the more common reasons for developing a data culture are to:

Optimize operations

- Measuring performance, efficiency and impact.

  This is one of the more common purposes, especially in a data rich environment.

Spread a message

- Using data to tell a story and communicate impact.

  Communicating information is part of our daily routines, and a data culture might help us do it better, especially when we're talking about the bigger-picture.

Bring people together

- How can you use data to break down silos and strengthen partnerships across organization?

  This is one which is less often a core aim, but is an important part of this program.

Why build a data culture?

What's the difference between developing a broad organisational data culture and just hiring some data scientists?

What will outcomes look like?
Are there key risks in doing it this way?

5 min

Data Literacy

Data literacy is a key aspect of developing data culture.

It's the ability to:

- **Read** data
- **Work** with data
- **Analyse** data
- **Argue** with data

Why does this matter?

- You can't develop a data-driven business if you have low levels of data literacy.
- You might be able to generate data, but turning it into value will be arduous.
- These are key components of data science.

Discussion

What happens when data-focused business activities are contained within a single part of the organisation?

3 min

Roadblocks to Data Culture

There's plenty of things which can get in the way of a data culture.

- Especially on the scale of entire organisations.

- Data Therapy lists three key aspects which tend to get in the way: Confidence, Technology & Process

Confidence	Organizations are not confident that they can work with data.	Build with small examples.
Technology	Appears daunting, expensive and requires technical expertise.	Start with simple tech.
Process	No laid-out process for working with data.	Step-by-step approach.

- These manifest on smaller scales too.

Discussion

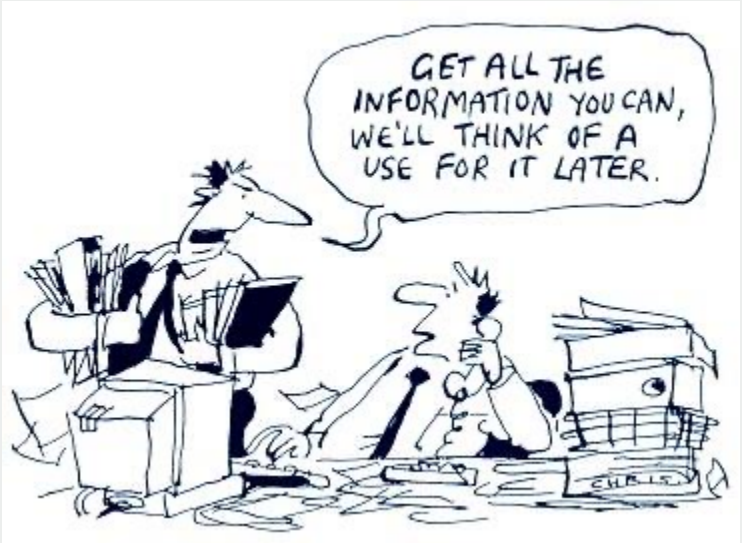
What else might stop you getting started with data science problems?

Are there simple solutions to some of these?

5 min

Styles of Data-oriented Businesses

There are different styles of data-oriented businesses, each with different uses for data:



Slane Cartoons Ltd

Styles of Data-oriented Businesses

There are different styles of data-oriented businesses, each with different uses for data:

- **Data-centric**: bring people together around data as the central driver to help make decisions
- **Data-informed**: take the data and its context as inputs to your conversation and decision-making process
- **Data-driven:** use data to make decisions directly - removing the human from the loop

Recognize that it's not either-or; all organizations can use a combination of all three approaches.

Where are each of these approaches useful?

What do you need for each of them?

Data-centric, Data-informed, Data-driven

5 min

High-Level Data Science Strategies

A strategy defines how you win.

At a high level, a data and analytics strategy must answer:

- What data?
- For what purposes?
- By whom?

At a slightly lower level, a strategy must:

- Identify and guide the allocation of critical resources
- Define how to measure success
- Adapt to changing circumstances dynamically
- Act both proactively and reactively

Identifying, Measuring and Demonstrating **Value** is Key

Baseline & Measuring Added Value

Start Where you Are

- From a baseline "as is" state you can plot a clear path to financial and business objectives.
- Your data and analytics incentives and activities should be linked to generating value.
- Decide on what you need to measure to demonstrate change, and to link this to increasing value.
- Remember that **people measure what is easy** to measure, not what is important to measure.

Baseline & Measuring Added Value

Consider how you might add value:

Information Value

- You learn something new, or have greater certainty in your information ("insight").
  Knowledge gain or sharing within the business.

Business Value

- You do better business: improve business processes with data and analytics.

Stakeholder Value

- What the data and analytics mean for stakeholders, such as assets, partners, shareholders and society at large.

Baseline & Measuring Added Value

Questions to dig into the value of data:

- What does this data help us do?
- What is missing from this data?
- Are we sharing data in the most effective ways?
- How do we get different silos to align?

What is my existing data culture?

What's your existing data culture?

How is data managed, used, and communicated?

10 min

What is my existing data culture?

Data Management

- Data are spread out over the organisation with no central point to get data from
- Data are pooled and accessible to all staff with descriptive formats
- Data lifecycles are managed with old data retired. Code and pipelines are managed.
- Code, pipelines and servers are reused across the organization to add value across silos

What is my existing data culture?

Data Science – Using Data

- Data are columns in spreadsheets and not connected to business decisions
- Data are connected to decisions through descriptive statistics, visualization and inferential statistics?
- Modelling process – coming up with your own models for the data making predictions of future states from current states
- Creating data products – making data speak for others through services and automated predictions

What is my existing data culture?

Communicating Data

- Data are not recognised
- Basic metrics are available and used to drive decisions routinely in other business processes
- Data are used to tell stories and show impact
- Data are used to argue for particular courses of action

Discussion

Do you feel empowered to make some data-oriented changes in your workplace?

Schedule

DAY 3

AWST	AEST	Agenda	
07:30	**09:30**	Q&A, Issues & Announcements	Educator
07:45	09:45	**Creating a Data Culture**	
09:15	11:15	*Morning Tea*	Jeremy
09:30	11:30	**Getting From Here to There**	
11:00	13:00	***Lunch***	Jeremy
11:45	13:45	**Data Science Projects**	
13:15	15:15	*Afternoon Tea*	Jeremy
13:30	15:30	**Data Exploration, Modelling and Reporting**	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	Close	

Getting from Here to There (90 min)

What needs to change?

Understand what already exists within your team before you try to change a bunch of things:

- Look for and highlight examples of good practise
- Look for internal advocates
- Build external relationships

What needs to change?

Other things to consider:

- Are there data champions already using data in good ways that you can celebrate as good role models?
- Are there roles in your organization aligned with your data needs?
- Is there a central person or team setting policies and best practices when it comes to your data related work?
- Are there things that are going to break when I introduce a new process?

Baby steps are fine.

Roadblocks

Do any of your potential changes help address some of the potential roadblocks?

- **Confidence**
  How do we build confidence in our skills?

- **Technology**
  What platforms should we use and what investments made to get ourselves to the next level?

- **Process**
  What can my organization do to make it easier to get to the next level on the hierarchy?

Making Changes

Some questions to consider:

What might you change in your workplace?

What are *important* changes, and what are *easy* changes?

For which changes can you easily demonstrate added value?

Who might be a good advocate for implementing some of these?

What are the processes which will be difficult to change?

Are there good practises to highlight and popularise?

Where might pushback come from?

Dealing with Roadblocks

Confusion
Use communication methods and language to meet people where they are
You will need to teach others to interpret your outputs for themselves

Not knowing your own data
How will you keep track of data and assets?
How do you identify useful datasets?

Organisational Silos
Acknowledge walls but demonstrate value when broken down.

IT-centric thinking
Infrastructure specific thinking (if I just have ___, my analytics problems will be solved)
Data is for everybody and Excel is just as much a useful tool as the best biggest data platform – invite people to work with data

Irrelevance
If staff don't understand the utility of the data they are collecting you will get bad data
You need staff to understand the significance, and it needs to be relevant

Boredom
Make data something that can be fun for people – creative times to generate new things from data – tell stories about your data share these – data are more than pivot tables!

HiPPOs (Highest Paid Person's Opinions)
When someone more powerful than you just wants to 'trust their gut' – importance of experimentation and establishing baselines!

Data Science in Context (20 min)

Defining Data Science

What is data science?

5 mins

Defining Data Science

What does data science encompass?

- Data analytics, machine learning, visualisation, communication and domain knowledge.
- Often creative, and involves 'building'
- **Building predictive models**
- Statistics 2.0?

Data Science is Different

What does data science encompass?

- Data analytics, machine learning, visualisation, communication and domain knowledge.
- Often creative, and involves 'building'
- **Building predictive models**
- Statistics 2.0?

- *Data science borrows some core scientific concepts, but differs to the way in which work is done.*
- *In particular, data science contrasts with the relatively linear workflows and timeframes of months to years we experience in typical research and development workflows.*

How is it different?

- Nonlinear Workflows
- Contrasting Timeframes
- **Speed & Iterative Improvement**

Key Differentiator: Speed & Iterative Improvement

The Approach

Data science workflows involve continuously assessing and improving what you're doing.

This cycle of improvement could be characterised by a series of stages, the latter of which involves a loop:

- Ideate
- Consolidate
- Implement:
- Draft, Refine, Optimise

Key Differentiator: Speed & Iterative Improvement

The Approach

Data science workflows involve continuously assessing and improving what you're doing.

When it comes to getting answers with data science, it's better to get an approximate answer quickly. You can use this as a feedback to quickly improve your models.

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise”

John Tukey

Data Science Projects Iterate Towards a Solution

Data Science Projects

- **Start with the definition of a problem**
- Data-driven investigation centred around building models
- Explorative, iterative testing of ideas
- Targeted communication + visualisation
- Vary in scope, but **not outcome driven**

Data Science Projects Iterate Towards a Solution

Data Science Projects

- **Start with the definition of a problem**
- Data-driven investigation centred around building models
- Explorative, iterative testing of ideas
- Targeted communication + visualisation
- Vary in scope, but **not outcome driven**

Iterative Development & Prototyping

Get an approximate answer quickly, then continuously assess and improve.

Prototype:
Build something which works, quickly.

Focus your time where you can add value.

Data Science Teams (20 min)

Parts of an Effective Data Team

Who fits into a ‘data science team’?

- Data teams can encompass a variety of different roles, given the breadth of work undertaken to complete data science projects.
- What does a data scientist do?
- How is this different to other data-focused roles?

Parts of an Effective Data Team

Who fits into a ‘data science team’?

- Data teams can encompass a variety of different roles, given the breadth of work undertaken to complete data science projects.
- What does a data scientist do?
- How is this different to other data-focused roles?

How does the work of data scientists overlap with and differ to:

- Data engineers?
- Database managers?
- Software developers?
- Typical researchers?
- Web developers?
- Data analysts?

Parts of an Effective Data Team

Who fits into a ‘data science team’?

“It’s important that our data team wasn’t comprised solely of mathematicians and other "data people". It’s a fully integrated product group that includes people working in design, web development, engineering, product marketing, and operations. They all understand and work with data, and I consider them all data scientists. We intentionally kept the distinction between different roles in the group blurry. Often, an engineer can have the insight that makes it clear how the product’s design should work, or vice-versa - a designer can have the insight that helps the engineers understand how to better use the data. Or it may take someone from marketing to understand what a customer really wants to accomplish.”

- DJ Patil

Ex-LinkedIn Chief Scientist, Chief Security Officer and Head of Analytics and Data Teams

Parts of an Effective Data Team

Map some of these people onto your workflow. Who fits in where?

Where do you fit on this spectrum of people?

Where would you like to be?

What are some of the more important skills which align with where you would like to be?

Where do you see gaps in your organisation with respect to these roles?

How might you start bridging those gaps?

Mapping Interfaces in your Business

Map out the interfaces between your team members considering how you work.

- Who manages these interfaces?
- What (data/code/information) needs to flow across them?
- What does effective leadership look like in a data-focused team?

Mapping Interfaces in your Business

Map out the interfaces between your team members considering how you work.

- Who manages these interfaces?
- What (data/code/information) needs to flow across them?
- What does effective leadership look like in a data-focused team?

Map interfaces from your data team to the rest of the business.

- What needs to be supported once you've built something?
- Who provides that support?
- Who funds your work?
- How do you measure ongoing value?
- How do you know when things go wrong?
- Who manages the new datasets that you're creating?
- Who helps others get confident with the new processes you've introduced?

Content developed between CORE Skills & Dr Morgan Williams | CSIRO

Mapping Interfaces in your Business

These interfaces are all potential communication barriers.

Schedule

DAY 3

AWST	AEST	Agenda	
07:30	**09:30**	Q&A, Issues & Announcements	Educator
07:45	09:45	**Creating a Data Culture**	
09:15	11:15	*Morning Tea*	Jeremy
09:30	11:30	**Getting From Here to There**	
11:00	13:00	***Lunch***	Jeremy
11:45	13:45	**Data Science Projects**	
13:15	15:15	*Afternoon Tea*	Jeremy
13:30	15:30	**Data Exploration, Modelling and Reporting**	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	Close	

Data Science Workflow (20 min)

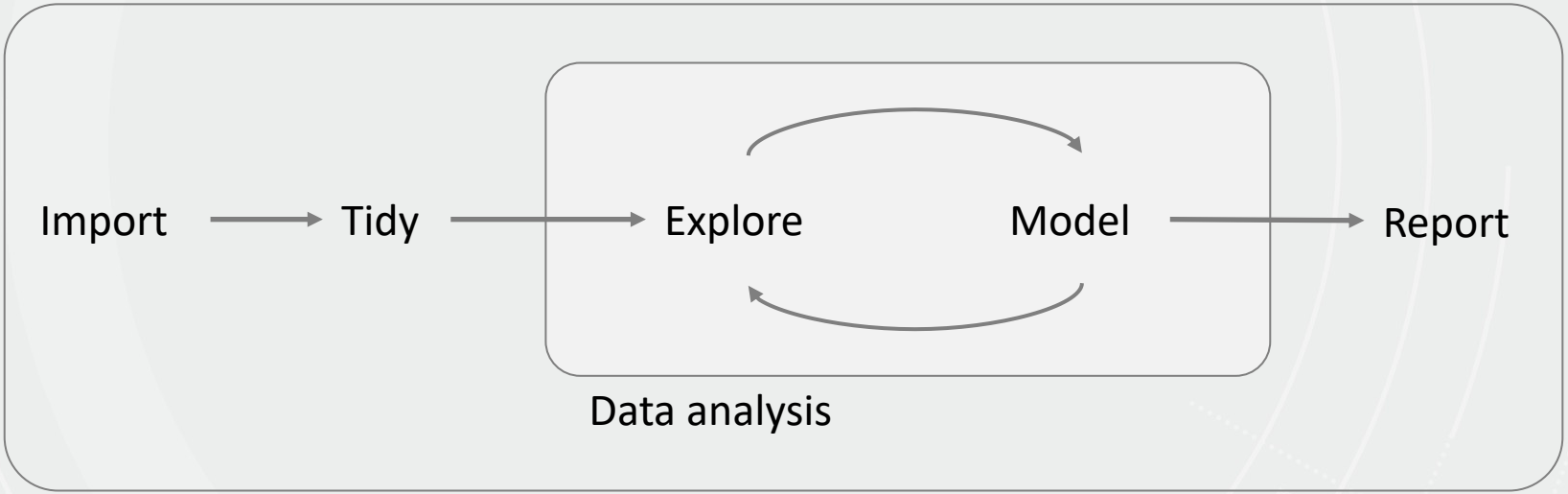
The Workflow

Data science is more iterative.

What does this look like as a workflow?

The Workflow

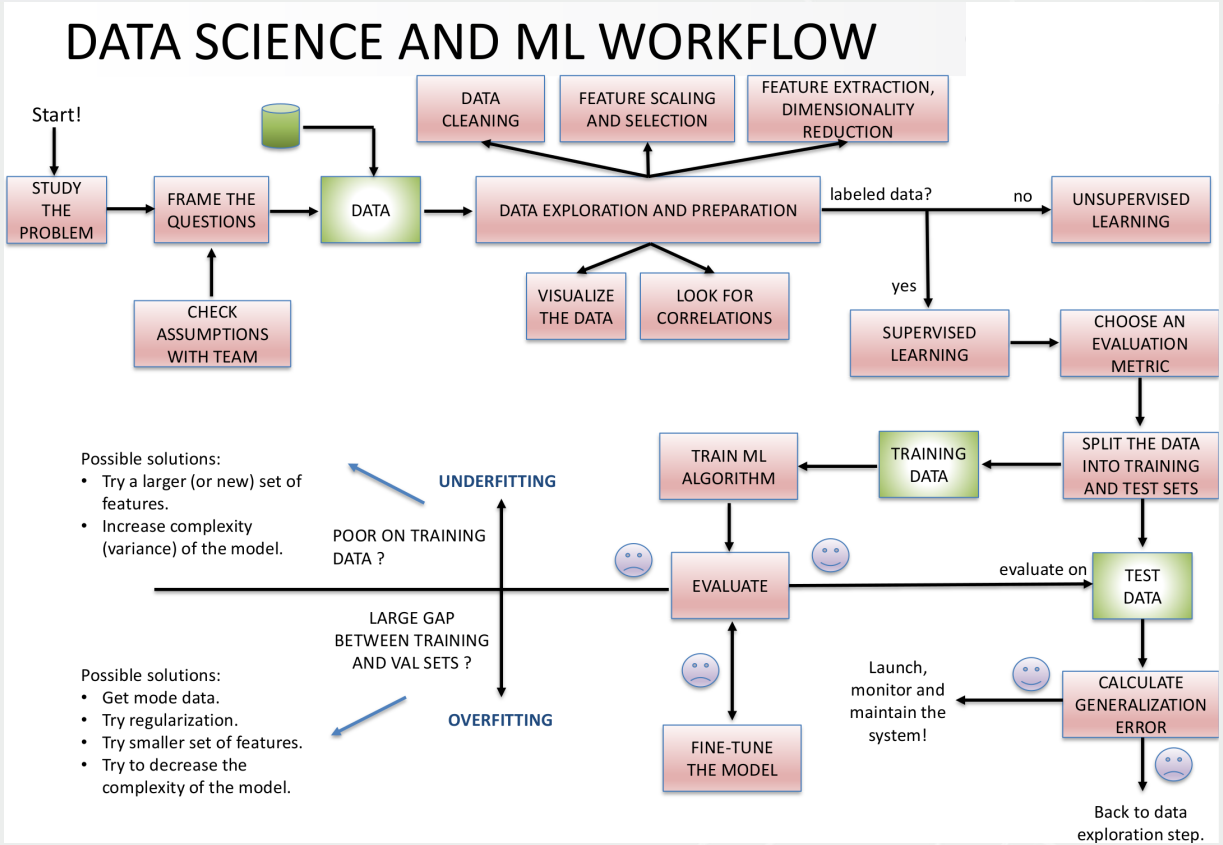
We Can Generalise Some of This for Today



The Workflow

Here's an example flow from the pilot.

- Starts with defining a problem and how you'll measure success
- Branches and loops
- Some details for evaluating models



The Workflow

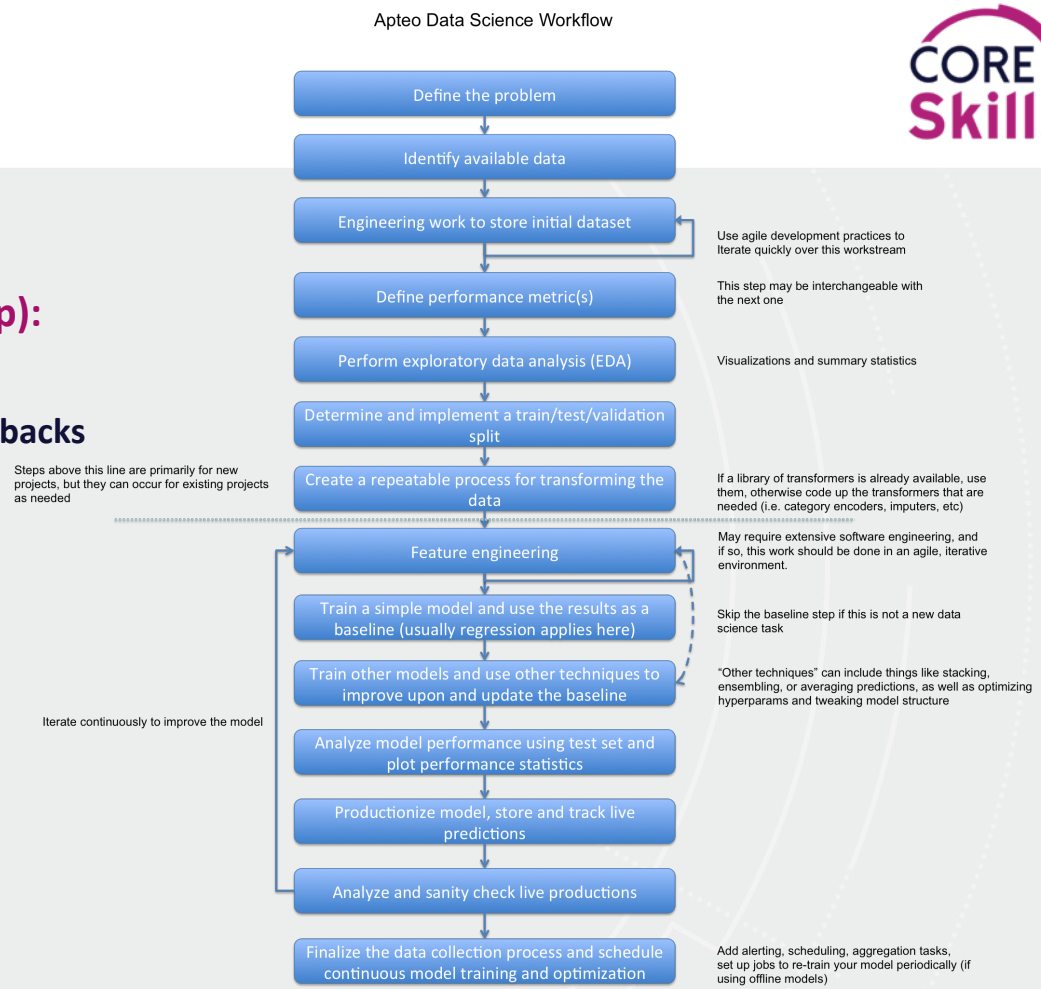
Here's one from Apteo (fintech startup):

- This looks pretty linear, with a few **feedbacks**
- Starts with **defining the problem** and **how you'll measure success**

Evaluation for both:

i) Model performance.

ii) How it works in the wild.



The Workflow

We'll work through parts of the front-end of this at a high level this afternoon.

The Workflow

Some things to think about regarding your projects:

- What makes something a good problem to address?
- What kinds of questions can you ask with this approach?
- What would be an acceptable outcome from this process?
- How might data/processing issues affect the outcome? (e.g. bias, missing data)
- Does your project easily break down to parts of a workflow?

Schedule

DAY 3

AWST	AEST	Agenda	
07:30	**09:30**	Q&A, Issues & Announcements	Educator
07:45	09:45	**Creating a Data Culture**	
09:15	11:15	*Morning Tea*	Jeremy
09:30	11:30	**Getting From Here to There**	
11:00	13:00	***Lunch***	Jeremy
11:45	13:45	**Data Science Projects**	
13:15	15:15	*Afternoon Tea*	Jeremy
13:30	15:30	**Data Exploration, Modelling and Reporting**	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Structuring & Prototyping Data Projects

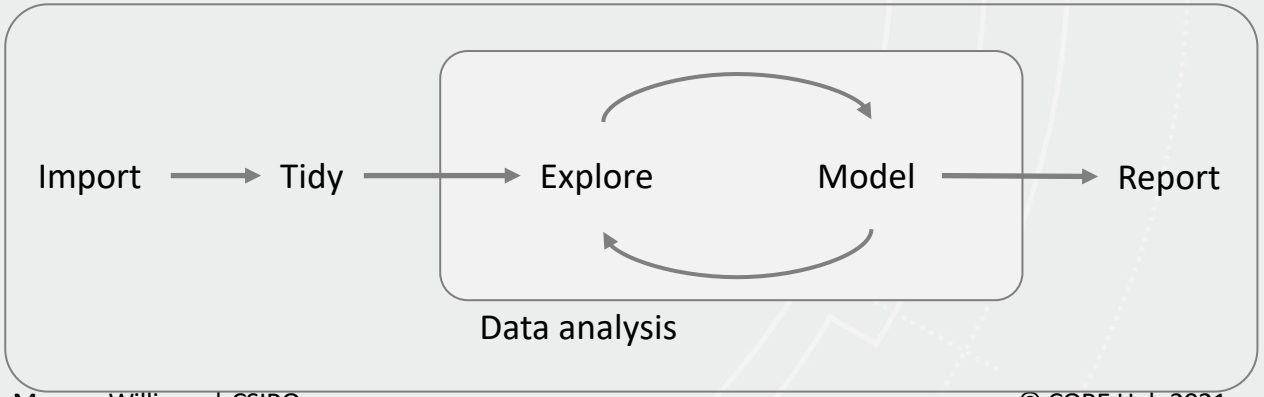
(35 min)

Structuring Your Data Project: Assembling the Parts

The different components of your project might resemble your workflow:

Scripts/modules to facilitate:

- A data import and processing pipeline
- Transforming your data ready for ML
- Building models
- Exploratory Visualisations
- Output visualisations to aid interpretation



Build a Workflow in (Independent) Parts

Design components to be independent:

- Can work on each separately
- **Focus effort** where greatest improvement can be made
- Separation of concerns
- **Sharing and Reusability**

Structuring Your Data Project

Segmentation/modularisation of your data science project allows you to develop parts independently.

Build Something which Works, Quickly

Prototyping

Start With a Solid Foundation, Then Iterate

- The first stages of building a project are a bit like plumbing.
- Your focus should be on where you'll get the most value.
- Avoid spending time making *all* things 'ideal'.
- Avoid the Ikea Effect
- You might be breaking things a lot. Fail fast.

What Goes into a Project?

Inputs

- Data
- Documentation & references

Code for:

- Data Processing
- Building Models
- Visualisation
- Automated reporting

Outputs

- Models
- Figures/plots/visualisations
- Logging and automated reporting

Anything else?

Structuring Your Data Project

Could you make an end-to-end data science project in a Jupyter notebook? Sure.

Why might this be a bad idea?

5 min

Structuring Your Data Project: On Disk

Beyond having some nice file structure, why is this important?

- You're typically **working with others**:
  knowing where to find & put things thanks to standardisation

When combined with other good coding practises (environments, documentation and version control):

- **Repeatability**
  Being easily able to regenerate your data, model and analysis (e.g. with new data).
- **Reproducibility**
  Get the same results after regenerating a model with the same parameters and data.

Structuring Your Data Project: On Disk

There's A Lot of Moving Parts

What's a good way to organise them?

Project Templates

Getting a Head Start – Some Opinionated Templates for Project Repositories

- File structure templates (how does your analytics team do it?)
- Project repository templates – e.g. GitHub
  - Including basic configuration files
- `cookiecutter`
  - Generate (versioned) project templates programmatically
- `cookiecutter-data-science`
  - "A logical, reasonably standardized, but flexible project structure for doing and sharing data science work."

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model s
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for orderi
│                         the creator's initials, and a short `-` delimited descriptio
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materi
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environme
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- Make this project pip installable with `pip install -e`
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
```

Structuring Your Data Project

CORE Skills

Other related things to watch out for:

- Keep raw data and processed data separate
- Keep track of models which you save to disk (‘which version is this?’)
- Don’t add data or images to a git repository (`.gitignore`)
- Don’t add passwords, access codes or *secrets* to a git repository

Let’s Clean up a Notebook

From a ‘sandbox’ environment to create some reusable code, and a component of our data workflow.

What to outsource?

- Rapid Deployment: Cloud and managed services *can* be your friend.
- When it comes to infrastructure, start small and try something.
- Focus on what differentiates your business.
  You're not a database admin, and you don't need to be like google.
- If this is a core part of your business, figure out how to do it yourself.

Schedule

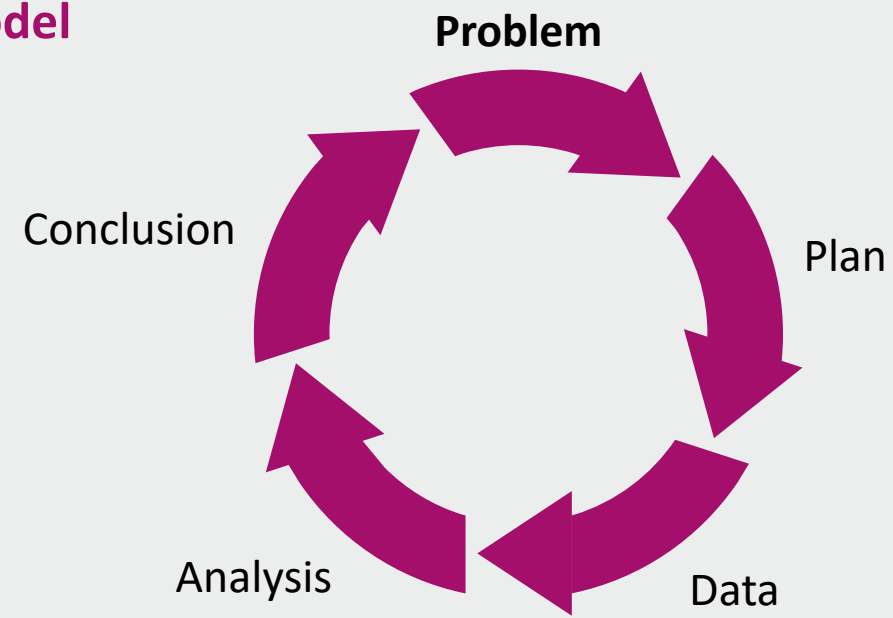
DAY 3

AWST	AEST	Agenda	
07:30	**09:30**	Q&A, Issues & Announcements	Educator
07:45	09:45	**Creating a Data Culture**	
09:15	11:15	*Morning Tea*	Jeremy
09:30	11:30	**Getting From Here to There**	
11:00	13:00	***Lunch***	Jeremy
11:45	13:45	**Data Science Projects**	
13:15	15:15	*Afternoon Tea*	Jeremy
13:30	15:30	**Data Exploration, Modelling and Reporting**	
14:45	16:45	**Closeout** – Reflections, Takeaways	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	Close	


Data Exploration, Modelling and Reporting
(90 mins)

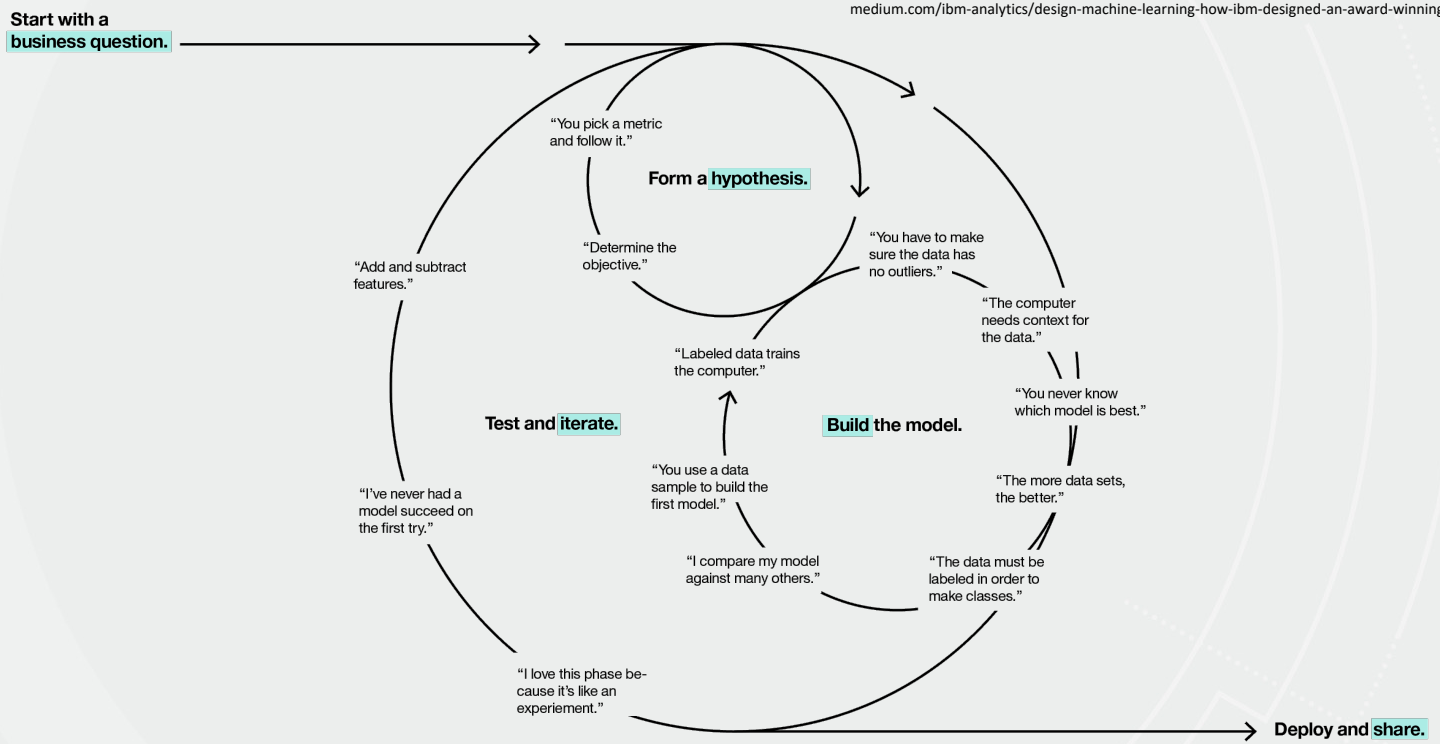
Define your PROBLEM first

The PPDAC model



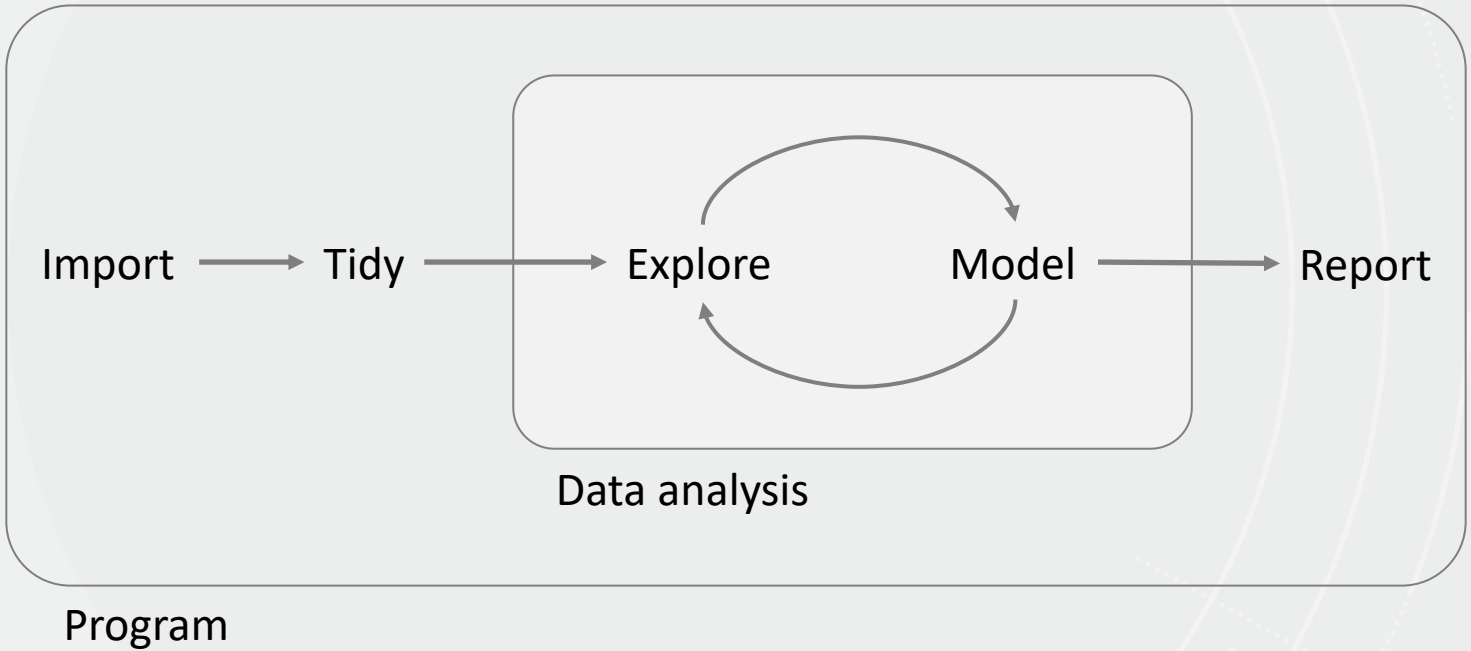
Then it's all about iterations

From IBM Data and ai:
medium.com/ibm-analytics/design-machine-learning-how-ibm-designed-an-award-winning-data-platform-ae6034063ffe



Seeking patterns in data for predictive uses

Basic workflow of data analysis



Take a step back from your data

- Who collected it? Why?
- Can it answer your problem?
- How quickly can you get there?
- How can you demonstrate the value of this work?

Take a step back from your data

- Who is responsible for Quality Assurance and Control?
- Why and when can QA/QC fail?

FAIR data

Findable
Accessible
Interoperable
Reusable

Make your data useful

FAIR data

Findable
Accessible
Interoperable
Reusable

1. Data
2. Metadata
3. Infrastructure

Make your data useful

FAIR data

- Who do you have to deal with to access data?
- Who needs access to the data?
- How do their needs differ?
- How can we get space to make changes?
- What do we do with cleaned data?
- How durable is the storage?

FAIR data

Goal = **Maximize value**

Tidy data

Concept:

- 1 column = 1 variable
- 1 row = 1 observation

Explore your data

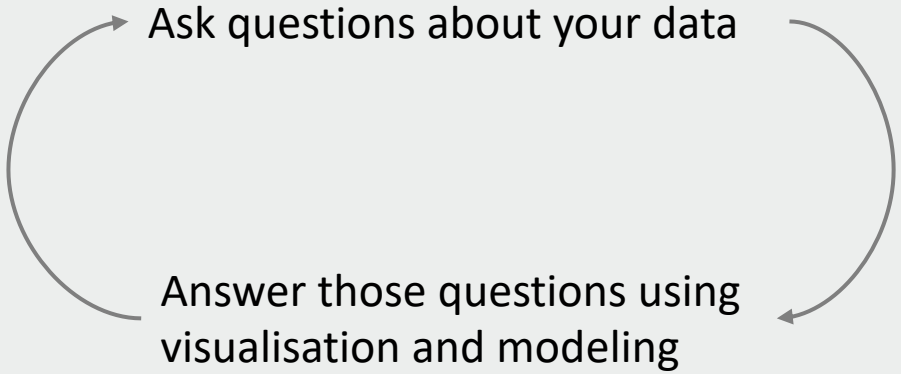
Goals

- Uncovering the structure of the data
- Finding the most important variables
- Detecting outliers
- Detecting anomalies

Explore your data

- What is the first thing you do?
- How does it influence your next steps?

Explore your data



Explore your data

Goal = **Maximize insight**

Explore your data

- How can you do it in practice?
- What techniques can you use?

Explore your data

Let’s have a look at

pm-data-exploration.ipynb

Step 1

What is a model?

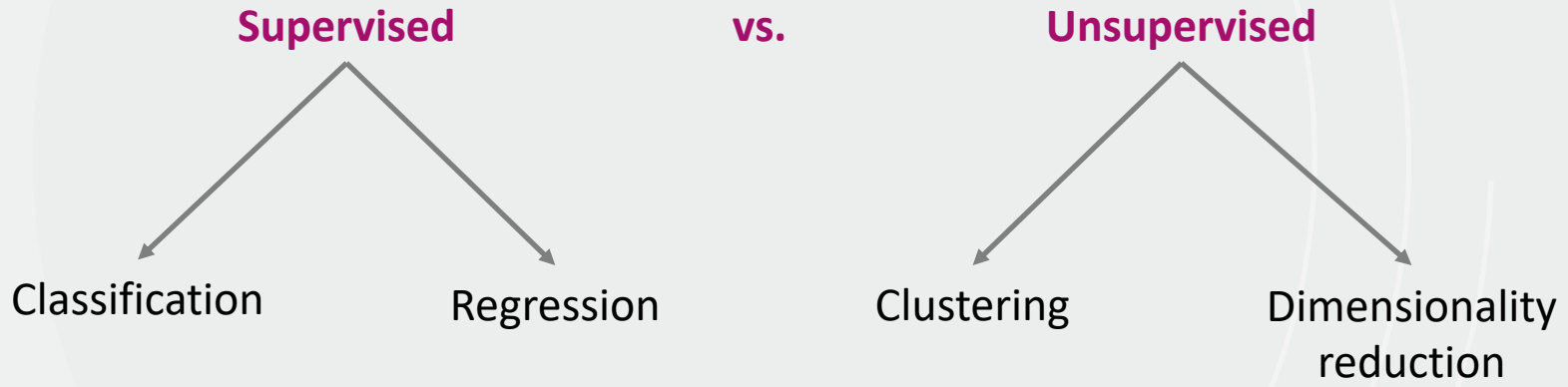
What is a model?

$$y = f(x)$$

Model your data

- What do we want to predict?
- How to we check success?
- Do we need to understand the model?

Select a model



Select a model

Simpler is better

Fit the model to the data

The high-level view…

$$y = f(x)$$

… The truth: hyperparameters

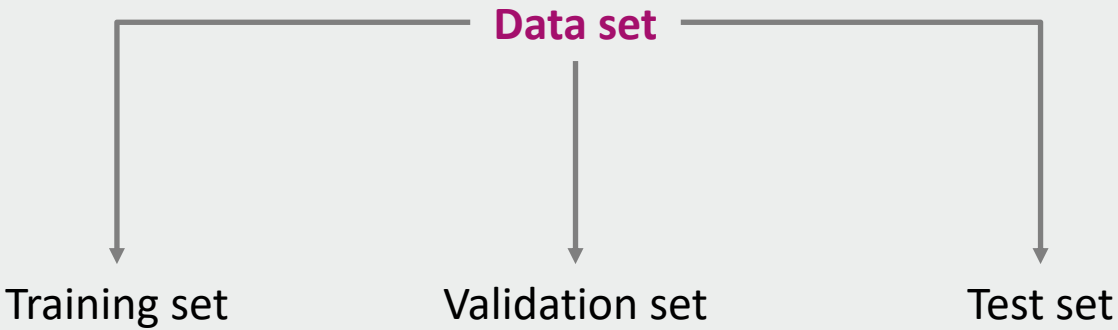
$$y = f(x,a,b,\ldots)$$

Fit the model to the data

Optimisation…

… Objective function

Prepare your data for modeling



Visualise the model

Results have two components:

Predictions = What the model has **captured**

Residuals = What the model has **missed**

Visualise the model

Goal = **Maximize insight**

Visualise the model

Let’s have a look at

pm-data-exploration.ipynb

Step 2

Don’t forget good programing habits

- How to organise your code?
- How to name variables, functions, classes?
- Where to add comments, and which ones?

Don’t forget good programing habits

CORE Skills

Programming ***LANGUAGE***

Don’t forget good programing habits

Goal = **Maximize value**

Report your findings

Two main goals:

Get **approval**

Get **feedback**

Know your audience

- What is their data literacy?
- What do they expect from my report?
- What do I expect from them?
- Which features of my analysis matter?

Je suis Charlie

Visualisation is fundamental

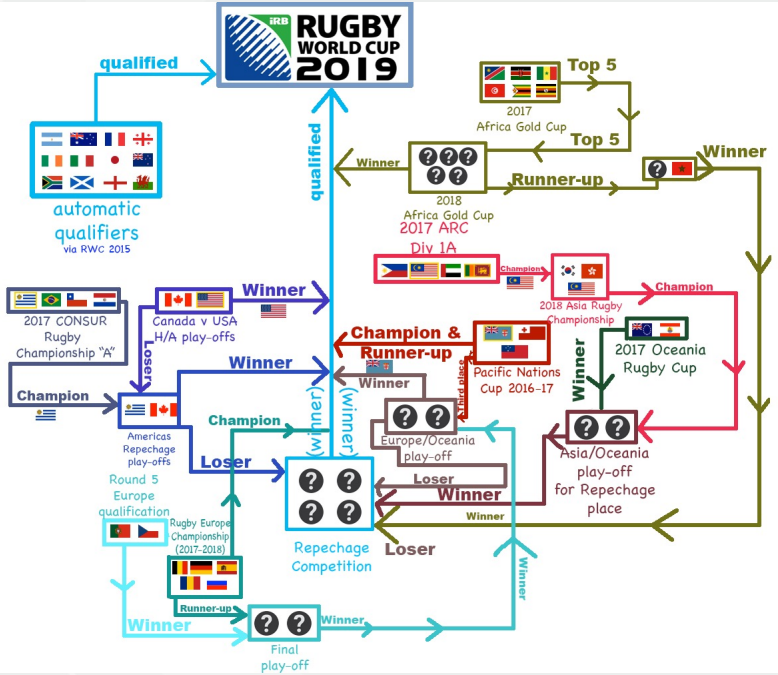
Visualisation is fundamental

- What are you trying to say?
- Is it clear from your design?
- Have you eliminated any non-informative content?

Visualisation is fundamental

From Kaiser Fung/Junk Charts:

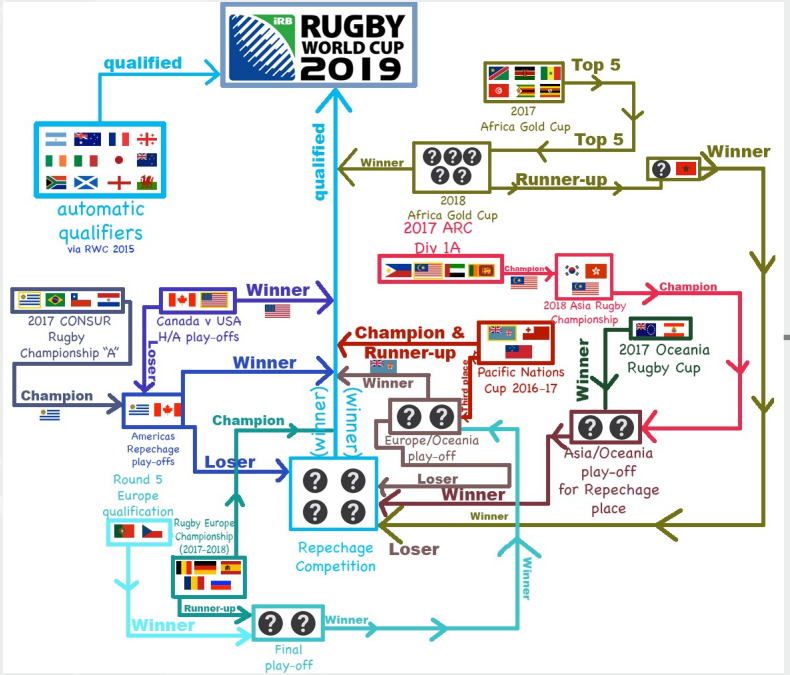
junkcharts.typepad.com/junk_charts/2019/09/the-windy-path-to-the-rugby-world-cup.html

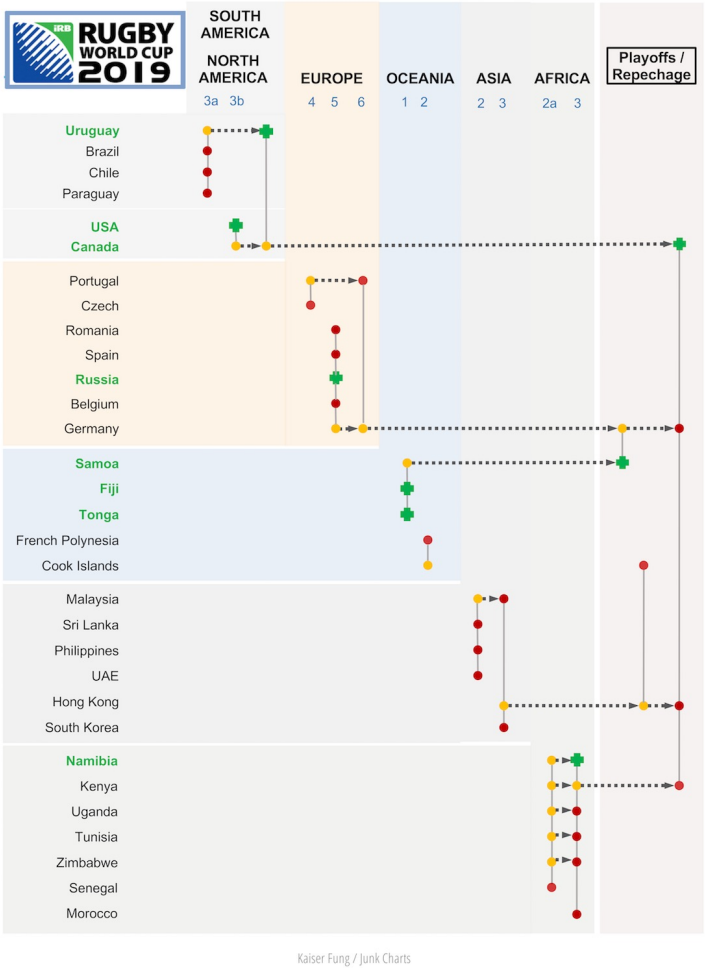


Visualisation is fundamental

From Kaiser Fung/Junk Charts:

junkcharts.typepad.com/junk_charts/2019/09/the-windy-path-to-the-rugby-world-cup.html





Content developed between CORE Skills & Dr Morgan Williams | CSIRO

Takeaways & Closeout

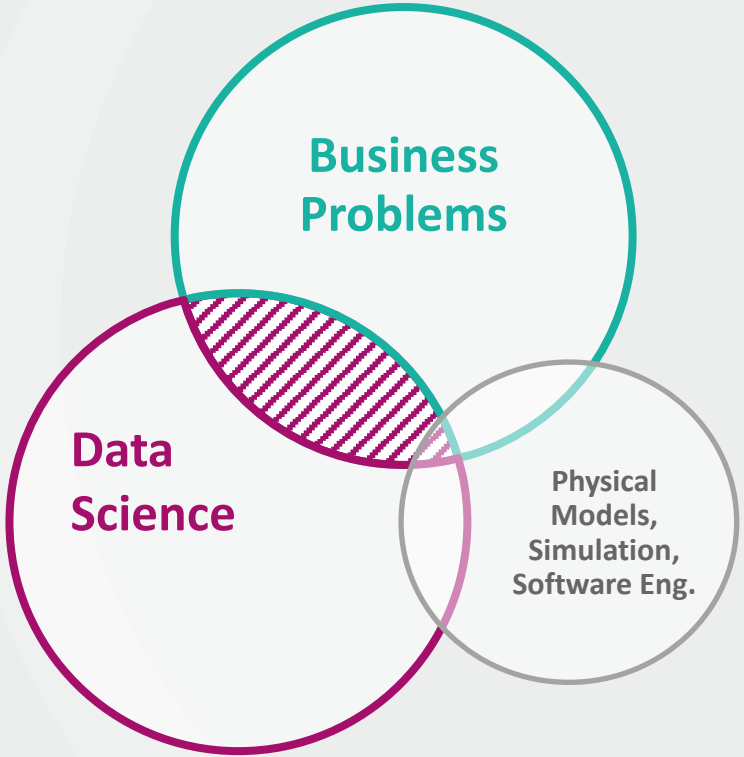
High-level Takeaways From Today

CORE Skills

- **Different types of data culture** *Why are you doing data?*
- **Data literacy** *Who are you talking to?*
- **Data strategy** *How will you do it?*
- **Data science teams** *Who will you work with?*
- **Basic data science workflow** *What will yours involve?*
- **Delivering incremental change** *What does it look like?*

Capstones: Assessing Fit, Timeframes and Planning (20 min)

When is Data Science the Right Approach to a Problem?



- ✓ You want to support **decision making**
- ✓ **Timeliness** of estimates or communications is key (feedbacks, early warnings)
- ✓ **A large amount of human time is involved** making simple decisions
- ✓ **Subjective decisions are made** based on a limited perspective (or limited data)
- ✓ **Uncertainty and variability** are important

Identifying Capstone Fit

CORE Skills

Some things to think about regarding your projects:

- What makes something a good problem to address?
- What kinds of questions can you ask with this approach?
- What would be an acceptable outcome from this process?
- How might data/processing issues affect the outcome? (e.g. bias, missing data)
- Does your project easily break down to parts of a workflow?

Capstones – Examples

Deploying a machine learning regression model that predicts B4 wagon weights and volumes, preventing ≈500ktpa or $45Mpa in lost rail capacity.

- Along-line delay between loading and weighing induces delay in feedback loop, and risk of both overloading (performance/maintenance issues) and underloading (lost capacity)
- Trained on historical loading data to identify factors contributing to adverse scenarios
- Can be deployed as an online model tightening the feedback loop and wagon weight variation

Capstones – Examples

Increasing impact of machine learning models in geometallurgy ($260M of NPV at Koodaideri to date) by distinguishing upgradeable and non-upgradable low grade ore, turning Resource Evaluation drilling into Metallurgical drilling; a paradigm shift for metallurgy.

- Adapting existing data assets to provide value along the processing chain

Capstones – Examples

Detecting water level anomalies with machine learning models, for more accurate data and rapid maintenance decision-making.

- Reduce need for human interaction, allow time to be focused on key flagged issues and otherwise freeing up time to be used elsewhere.

- Actively increases future data quality, and can be retrospectively applied to flag poor quality data

These Projects are Not Completed in Three Months

Learning from Prior Capstones

Start Small
One task or process within one person's role

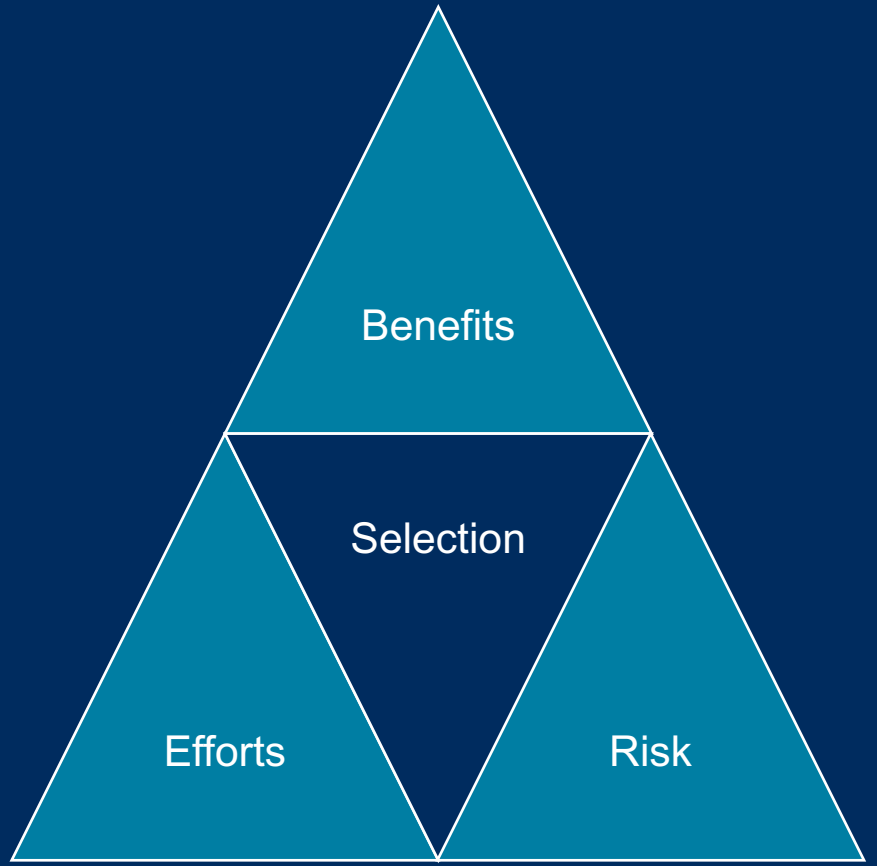
Plan for Proof of Concept
Previous projects are yet to be deployed

Key Suggestions at this Stage

1. Communicate the key problem in simple terms
2. Take the time to design
   - Identify the key question/problem
   - Break it down
3. Keep in Contact

Select Projects

(slides courtesy Matthew Ley)



Project Effort

1 Scoping	2 Data Assessment	3 Proof of Concept	4 Minimum Viable Product	5 Deployment	6 Transition
Key Activities					
• Conduct workshop to understand business problem • Develop understanding of value, feasibility and effort for project • Request data samples and data catalogue	• Mobilize project team • Hold workshops for data and process mapping • Send data request and get access to systems • Set up data processing environment • Create data dictionary • Conduct data assessment • Assess current site infrastructure • Develop understanding of end user and business value	• Pre-process and merge data sets • Conduct exploratory data analysis • Start feature engineering • Develop model using historical data (offline) • Develop understanding of model integration • Design initial solution architecture • Assess required infrastructure • Validate financial value	• Iterate model development and testing • Build data pipeline and solution architecture • Develop user interface prototype • Deploy model (online) connected to infrastructure and user interface • Conduct tests and demos of solution • Start change management initiatives	• Finalize model, solution architecture and user interface • Set up model performance tracking • Develop product documentation • Final testing of solution • Train users • Deploy solution	• Handover solution to support organization (IS&T, Analytics Franchise, PACE) • Define required support model for solution lifecycle • Retrain and recalibrate model (if required) • Continuously improve data pipeline and user interface • Deploy solution at scale and replicate
Key Deliverables					
• Problem statement • Project feasibility • High-level value	• Process map • Extracted data • Data dictionary • Data readiness report	• Offline model • Solution design • Model integration plan • Validated value	• Online model • Data pipeline • User Interface • Go-live of MVP	• Solution finalization • Product documentation • User training • GO-live of final solution	• Product handover • Product support model • Improvement plan • Scale up and replication
Timeline Estimates					
• 2-3 weeks	• 3-4 weeks	• 8-12 weeks	• 10-14 weeks	• 4-6 weeks	• 4+ weeks
Cumulative	• 5-7 weeks	• 13-19 weeks	• 23-33 weeks	• 27-39 weeks	• 31-43+ weeks

RioTinto

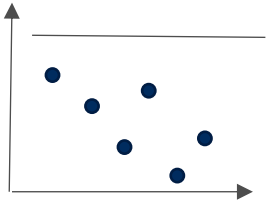
Some project risks are much greater for data science projects

1 Scoping	2 Data Assessment	3 Proof of Concept	4 Minimum Viable Product	5 Deployment	6 Transition
Key Risks					
	• Data unavailable • Data access difficult • Data lacks variation • Data quality poor	• Hypothesis incorrect • Solution architecture approval delays		• Turnover	• Turnover
Key Controls					
• Phases and gates	• Phases and gates • Seek early approvals	• Phases and gates • Seek early approvals	• Phases and gates	• Phases and gates	• Phases and gates

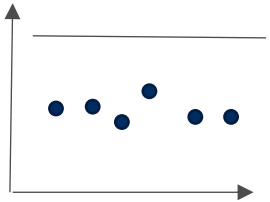
RioTinto

Benefits

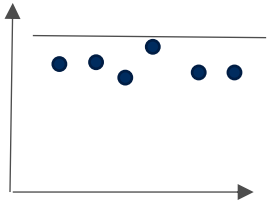
Run closer to target



Yarwun precipitation

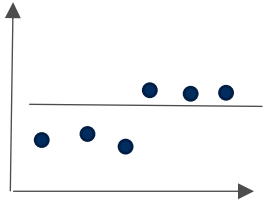


More predictable size

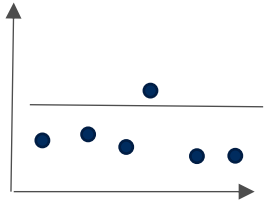


Improve yield

Respond quicker

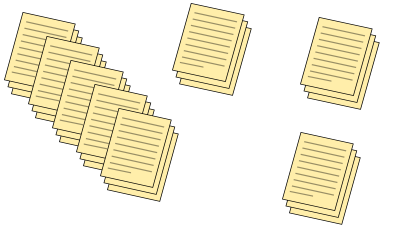


Bauxite grade prediction

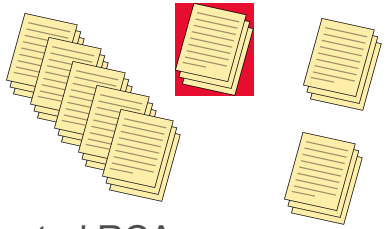


Respond to deviations quicker

Prevent losses



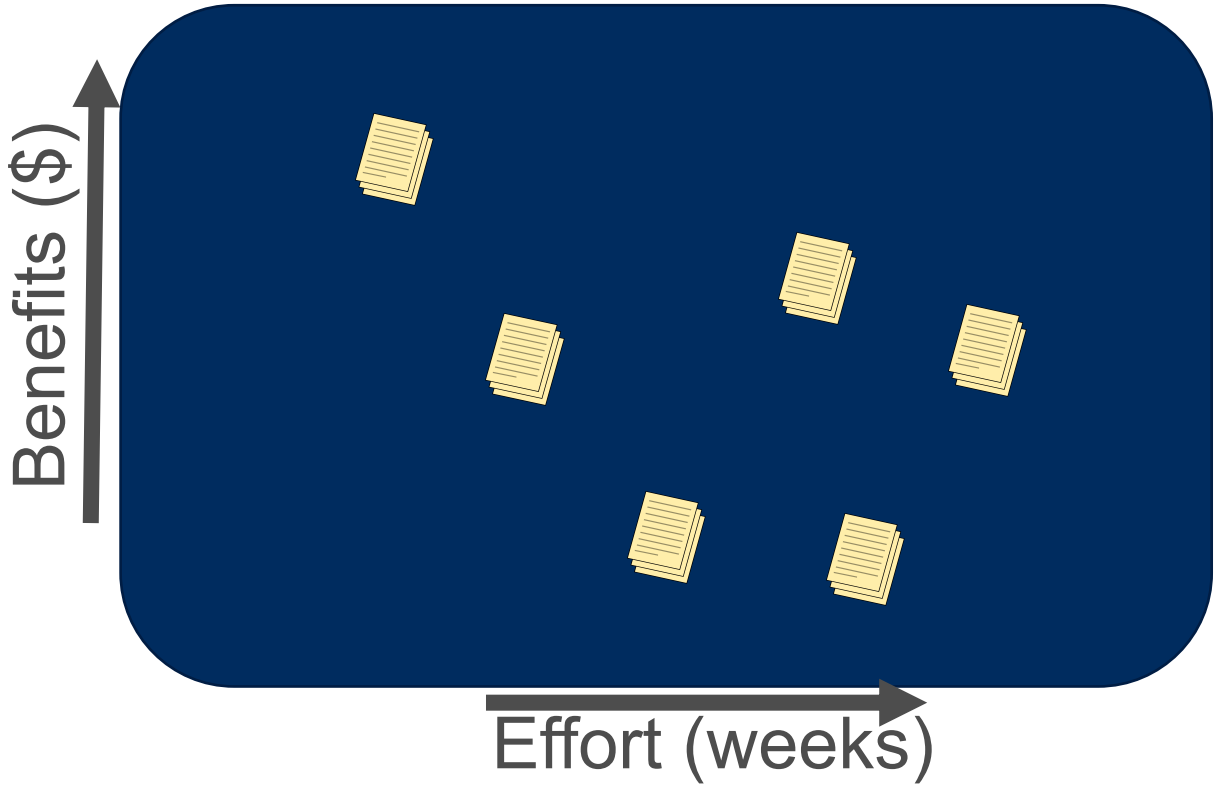
Finance anomaly detection



Targeted RCA

RioTinto

Springboard projects on a Benefits and Efforts grid



RioTinto

Milestones – Linking Capstones to Momentum

Data Scientist	Idea	Benefits	Effort	Risk
Hadley Wickham	Geo tech automation	$19M	1 week	L
Hadley Wickham	GAN for OBK	$9M	2 weeks	M
Hadley Wickham	Engine efficiency	$8M	21 weeks	H

Complete week 2
Circulate to sponsors
Enter L0

Complete week 5
Group review with sponsors
Enter L1

RioTinto

A3 - Grade Prediction.Docx

Objective

Current State

Problem

Future State

Solution

Value

Scope

In Scope	Out of Scope
•	•
•	•

Data sources

Existing	New
•	•
•	•

Stakeholders

Plan

•	•
•	•

KPIs

	Metric Type	Baseline	Target

Work and Deliverables

- •
- •

Critical Issues

- •
- •
- •

Work Schedule

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Project phases								
BO'B	Data assessment	Baseline	Concept	Proof of concept	Minimally viable product	Deployment	Industrialisation	Replicate
DEM								
AB	Process map	Data inputs		Model build	Model validation	Model application		
CU model								
MBL	Ship loading	Ship unloading		Ship circuit				
Data integration								
JB	Assess data source	Design integration layer		Modify				

Plan

Sponsor:

Team

Initial Tasks

Agree date	SPA	Duration	Date

RioTinto

Momentum
Initiative Stage Gates
Quick Reference Guide

Initiative Stage Gates

- **L0 – Idea Generated**
  A high level idea is described, including a high level value estimate.
- **L1 – Captured**
  The draft business case is developed.
- **L2 – Validated**
  The detailed business case with action plan, milestones and risk assessment is developed.
- **L3 – Planned**
  The initiative is implemented with the action plan fully executed.
- **L4 – Implemented**
  The benefits from the initiative are realised due to implementation.
- **L5 – Verified**

What is required in L0?

Description

Once an idea is generated the Initiative Owner begins to prepare for Stage Gate 1. There is no approval required to enter an idea into L0. However, in preparation to pass through Stage Gate 1, an Initiative Owner must submit a high level idea that has potential to add value to Rio Tinto.

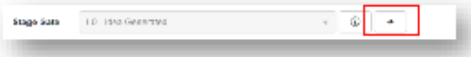
When the Initiative Owner is able to provide sufficient information and a rough value estimate, the initiative can be submitted for Stage Gate 1 approval. The Workstream Lead is required to determine if the initiative should progress from L0 to L1.

Required steps to complete in L0

1. Initiative Owner inputs an identified idea into Momentum.

2. Select the Workstream that will be delivering the work, and provide a description and rough value estimate to demonstrate the value it will deliver to Rio Tinto.

3. Select 'Continue' or 'Save & Close' to log an idea. Information can be updated over time before the initiative is submitted for Stage Gate 1 approval.
4. When the Initiative Owner is satisfied that the initiative has enough information to progress to Stage Gate 1, they select the '→' button to submit the initiative to the Workstream Lead for approval.



5. Workstream Lead assesses initiative, and determines if the initiative should progress through the stage gate to L1.

Required information

Problem

- Initiative Name
- Description

Plan

- Workstream
- Initiative Owner
- Initiative Owner's Leader

- Estimated annual value

L1 Approver (Stage Gate 1)

Workstream Lead

*Approval groups may differ by PG/BU.

RioTinto

Momentum
Initiative Stage Gates
Quick Reference Guide

Initiative Stage Gates

- L0 – Idea Generated
  A high level idea is described, including a high level value estimate.
- L1 – Captured
  The draft business case is developed.
- L2 – Validated
  The detailed business case with action plan, milestones and risk assessment is developed.
- L3 – Planned
  The initiative is implemented with the action plan fully executed.
- L4 – Implemented
  The benefits from the initiative are realised due to implementation.
- L5 – Verified

What is required in L1?

Description

Once an initiative has passed through Stage Gate 1 the initiative has been captured. Workstream Lead approval is required for the Initiative to have passed through Stage Gate 1.

In preparation to pass through Stage Gate 2, the Initiative Owner commences work to analyse the viability of the initiative and the value the initiative will bring to the business. Key stage gate dates and additional information about the initiative can now be captured, such as effort, complexity, confidence, value driver, etc.

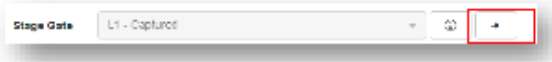
Progression to L2 occurs when an initiative is validated and has a targeted rough business case with documentation and high-level estimates of benefits. Approval is required by the Workstream Lead.

Required steps to complete in L1

1. Attach the initiative's high-level business case.

2. Enter date for L2 and L4.

3. Update any additional information that is required so that the approver can determine if the idea is worth pursuing and will provide value to the business. When the Initiative Owner is satisfied that the initiative has enough information to progress to Stage Gate 2, they select the '→' button to submit the initiative to the Workstream Lead for approval.



4. Workstream Lead assesses initiative, and determines if the initiative should progress through the Stage Gate to L2.

Required information

- L2 and L4 Dates
- Attach high-level business case

Plan

L2 Approver (Stage Gate 2)

Workstream Lead

*Approval groups may differ by PG/BU.

RioTinto

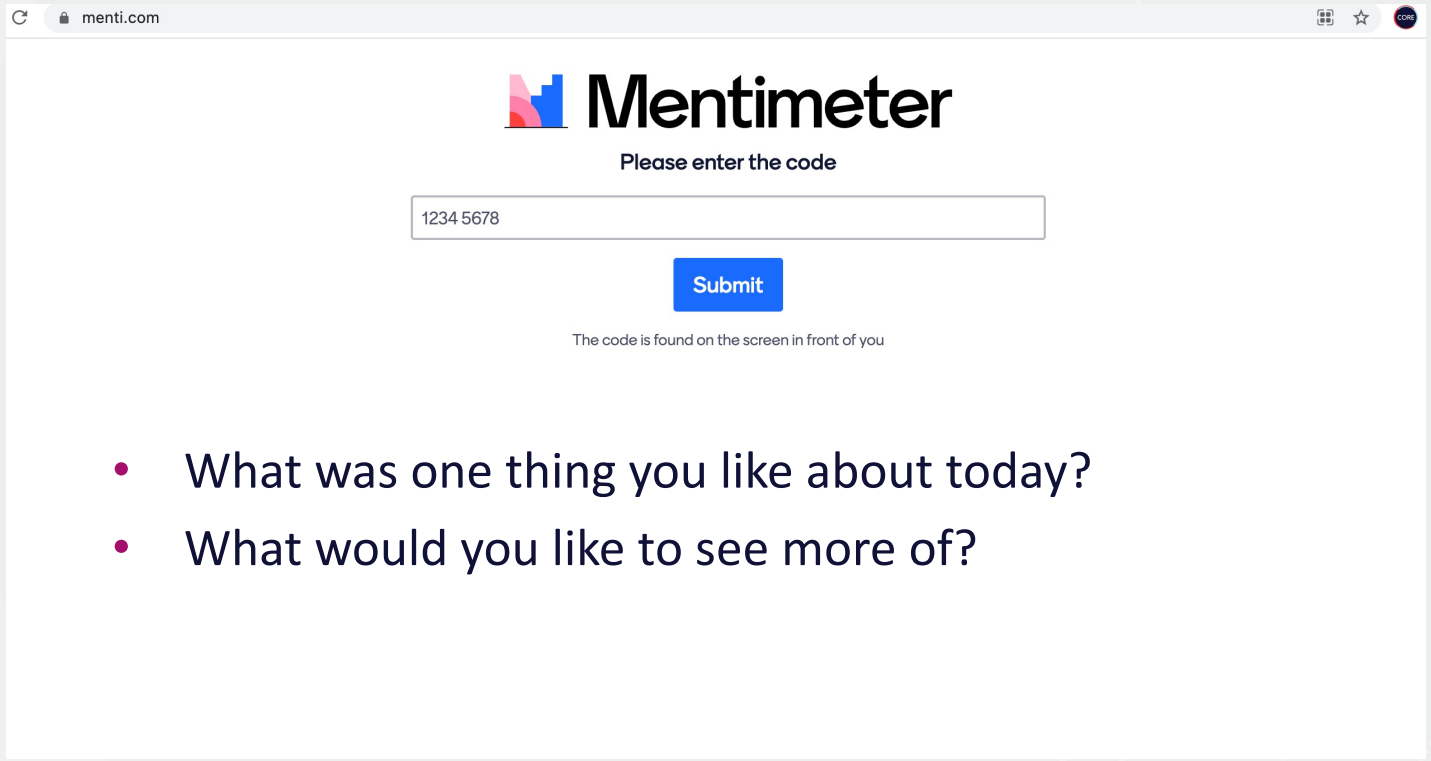
Capstone – First Milestone

Milestone #1:
Enter Current Capstone Ideas Into Spreadsheet

Morgan Williams & Guillaume Rongier | CSIRO

Daily Feedback

Menti



- What was one thing you like about today?
- What would you like to see more of?

CORE Skills

A CORE partnership

COREHUB.COM.AU/SKILLS