23 July 2018

HOW TO APPROACH THE DATATHON

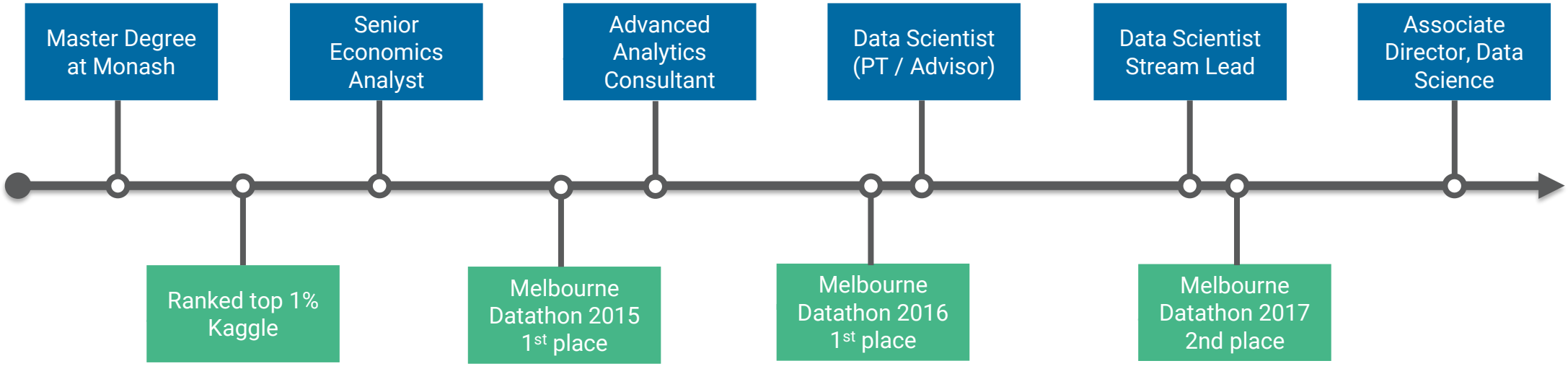
A lifecycle of data science product development

MY DATA SCIENCE JOURNEY

Hello I am, Ivan Liu,

Associate Director, Data Science @ ANZ

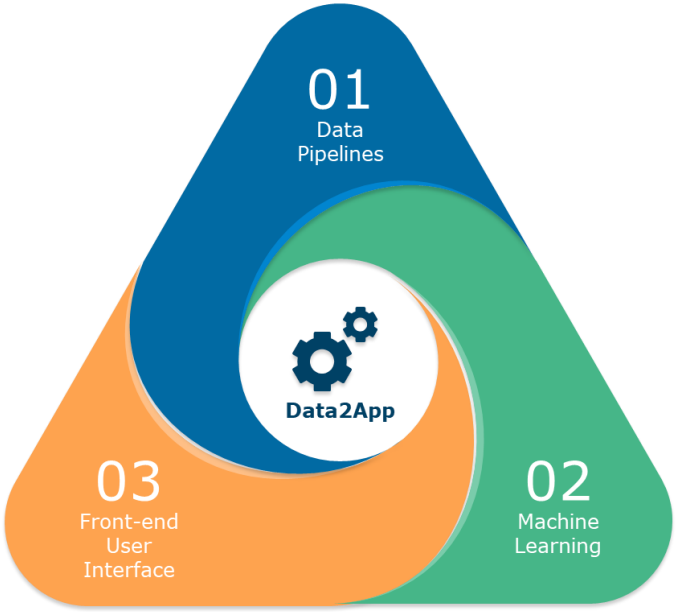
https://github.com/ivanliu1989
https://www.linkedin.com/in/ivanliu1989/
Ivan.Liu@anz.com



DATATHON - DATA2APP CATEGORY

Use the data provided (maybe a summarised version) and your data science and developer skills to build a world-class **application** with a beautiful, user-friendly interface, powered by some serious machine learning and analytics under the hood to bring your insights to life.

- Is the data presented in a way that tells a good story, highlighting key insights?
- Are those insights actionable – i.e. could a stakeholder make a practical business decision based on the data?
- Does the application effectively make use of current technology?
- Is the application scalable – could it easily be transitioned from a prototype to a production app on industry-sized data.



LAST WEEK...

Big data and **data science** have rocketed to the top of the corporate agenda.

Data Science has transformed the way organisations do their business and **data scientists** become one of the sexiest jobs in 21st century.

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

NETFLIX

Sep 2009, an algorithm competition with prize of **US$1,000,000** bested Netflix's own algorithm for predicting ratings by **10.06%**.

(100,480,507 ratings that 480,189 users gave to 17,770 movies)

kaggle

A platform for data science competitions founded in 2010 now has over **536,000** data scientists across **194** countries.

Solve real-world problems from top companies and research institutions.

LAST WEEK...

NETFLIX

Netflix never used its $1 million algorithm

Due to engineering costs and has no plan to use it in the future

Netflix used a suboptimal solution

For an 8.43% improvements (versus 10% improvements for the winner solution)

kaggle

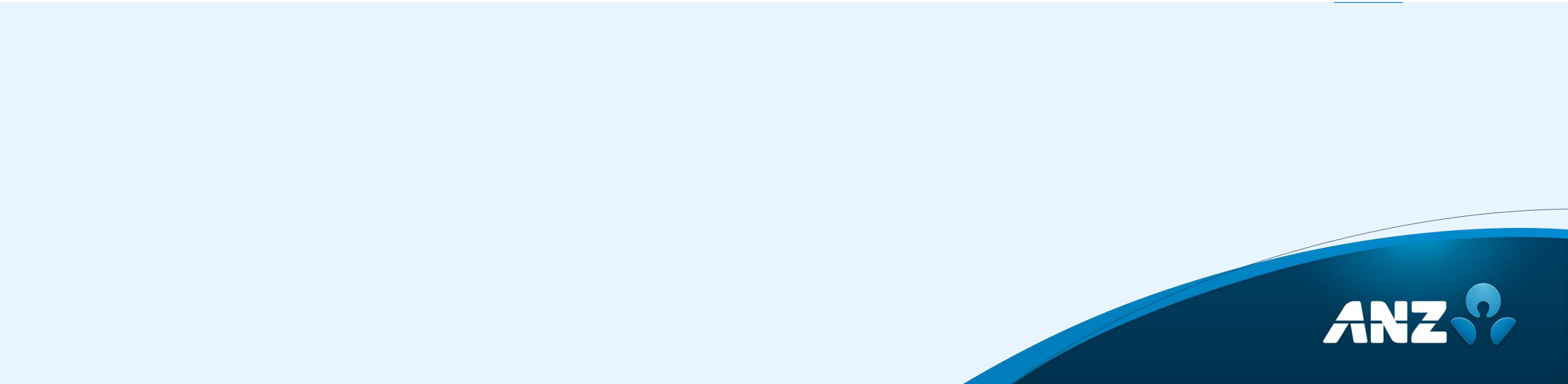
Most Kaggle winning solutions are impractical to be implemented

Kaggle solutions are aiming at model accuracies (complexities) rather than scalabilities.

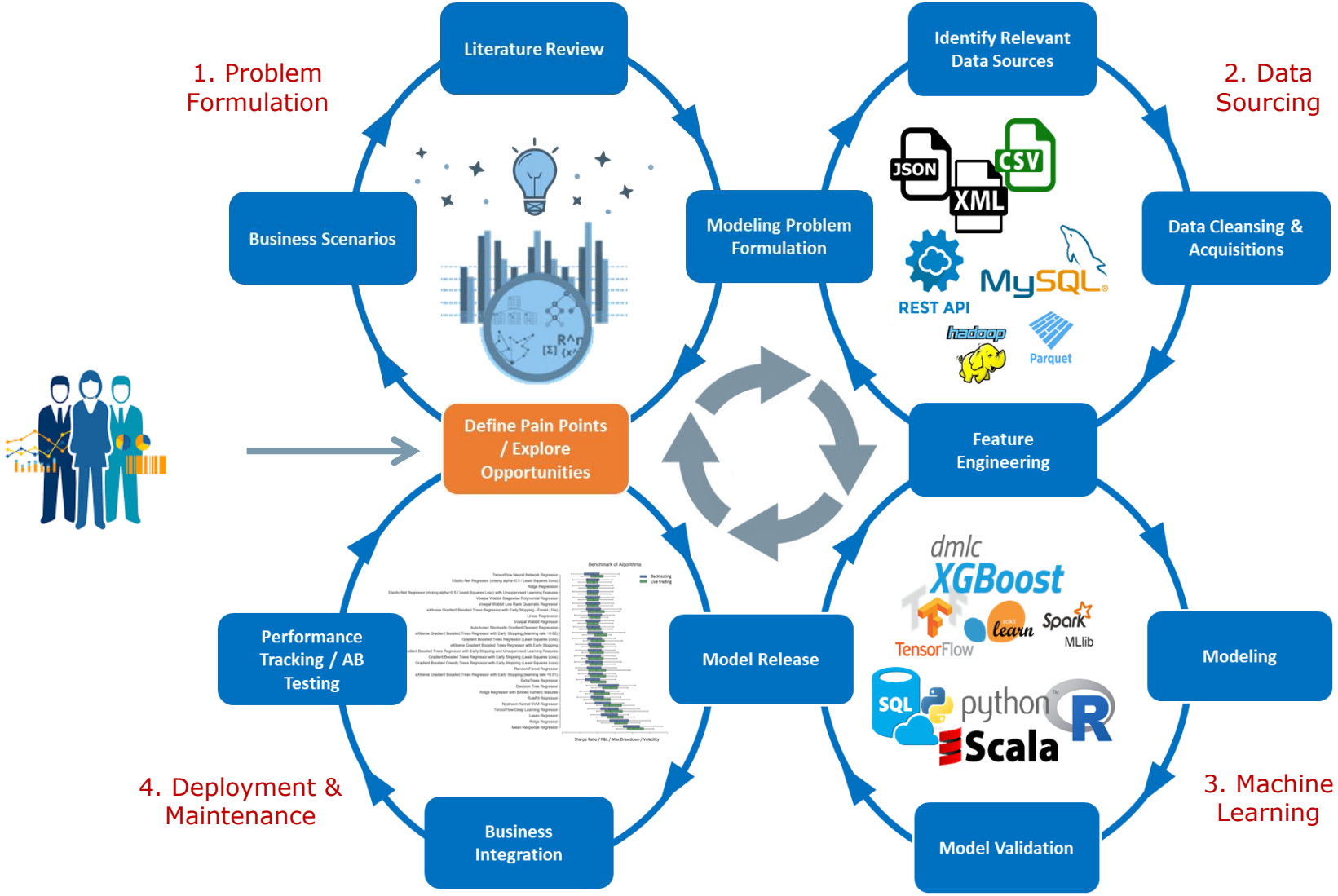
suboptimal solutions are more likely to be used

Simplified version of the winning solutions are more likely to be implement

DATA SCIENCE LIFECYCLE

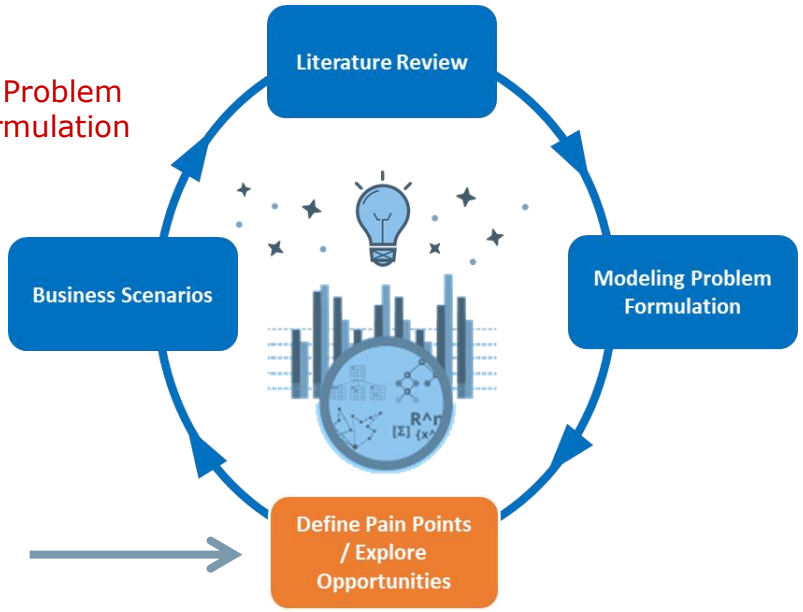


DATA SCIENCE LIFECYCLE



DATA SCIENCE LIFECYCLE

1. Problem Formulation



Opportunities

Is data science able to help to increase our sales?

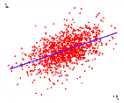
Use Cases

When?

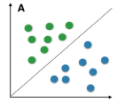
What?

Literature Review

Modelling

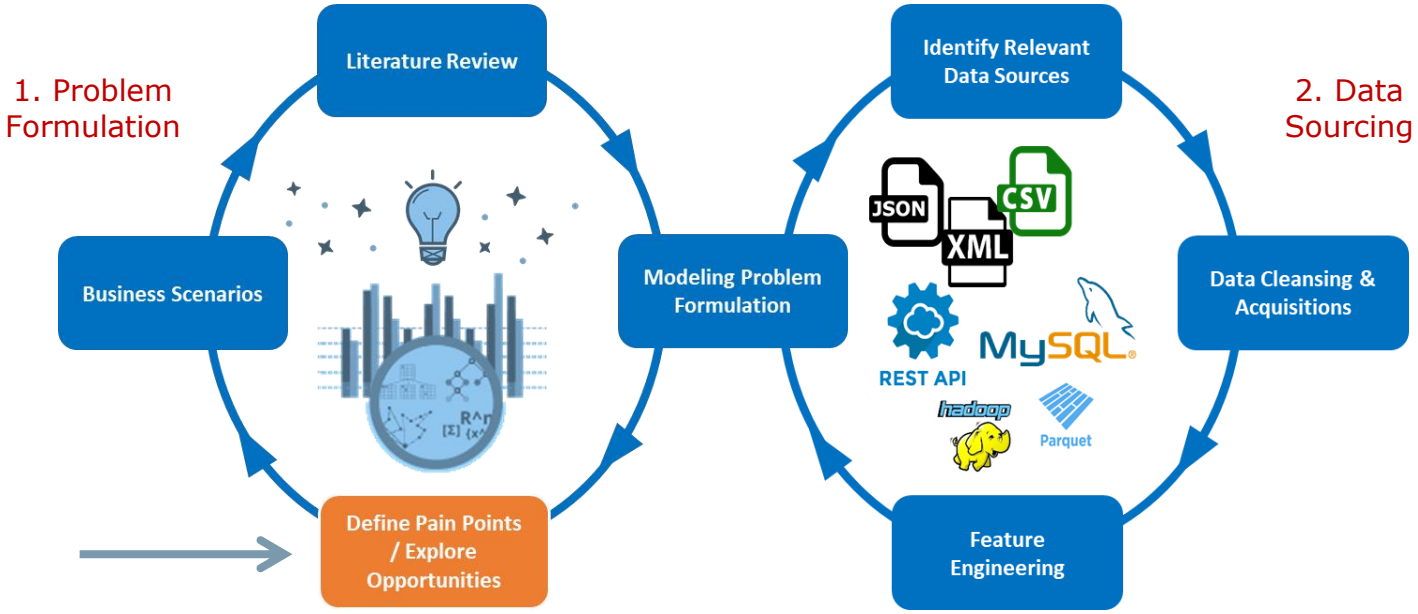


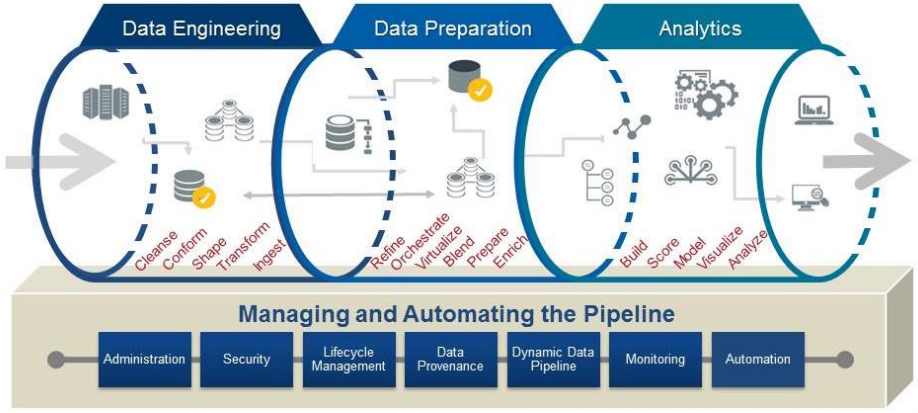
Regression model for Inter-Purchase-Interval



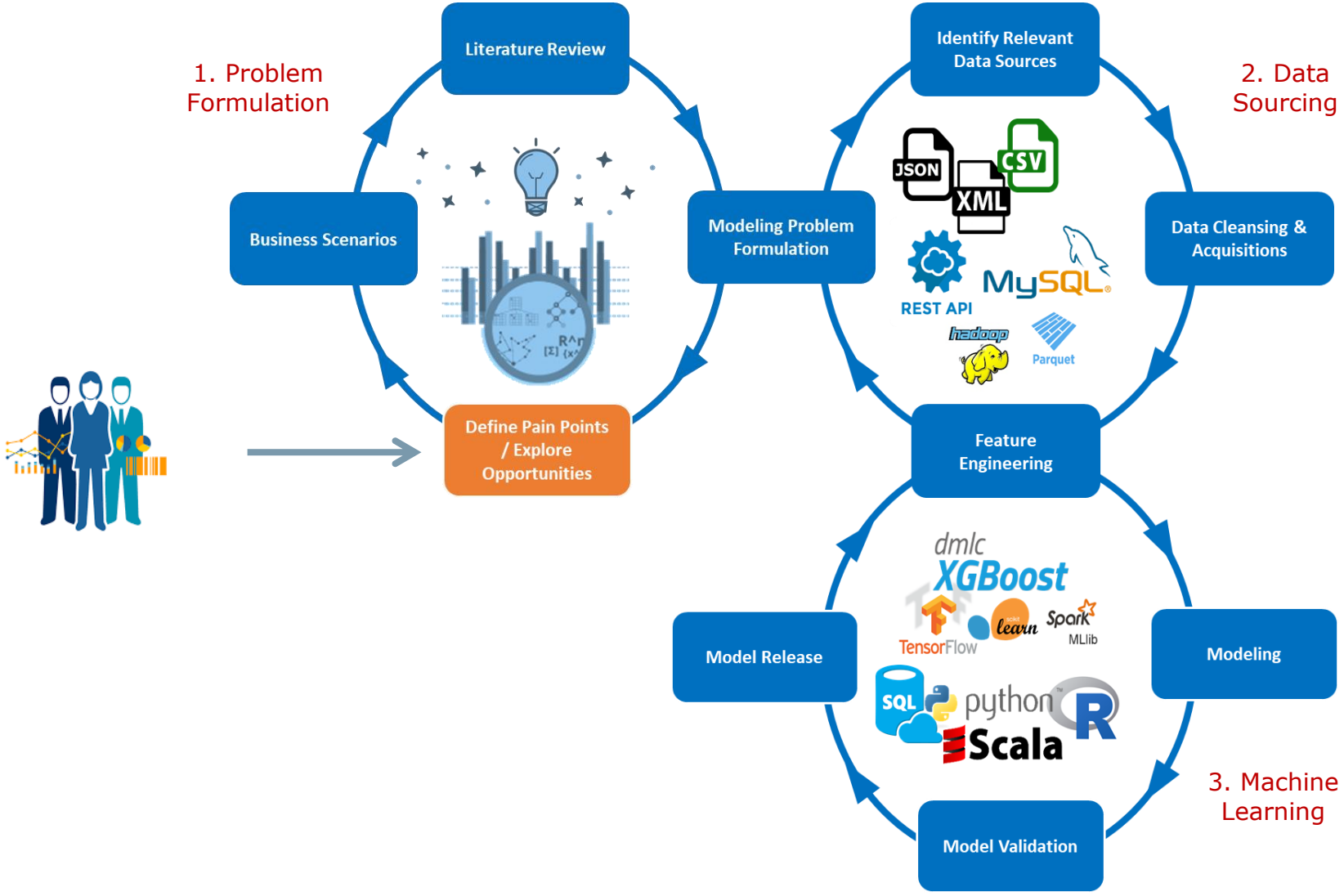
Classification model for product recommender

DATA SCIENCE LIFECYCLE

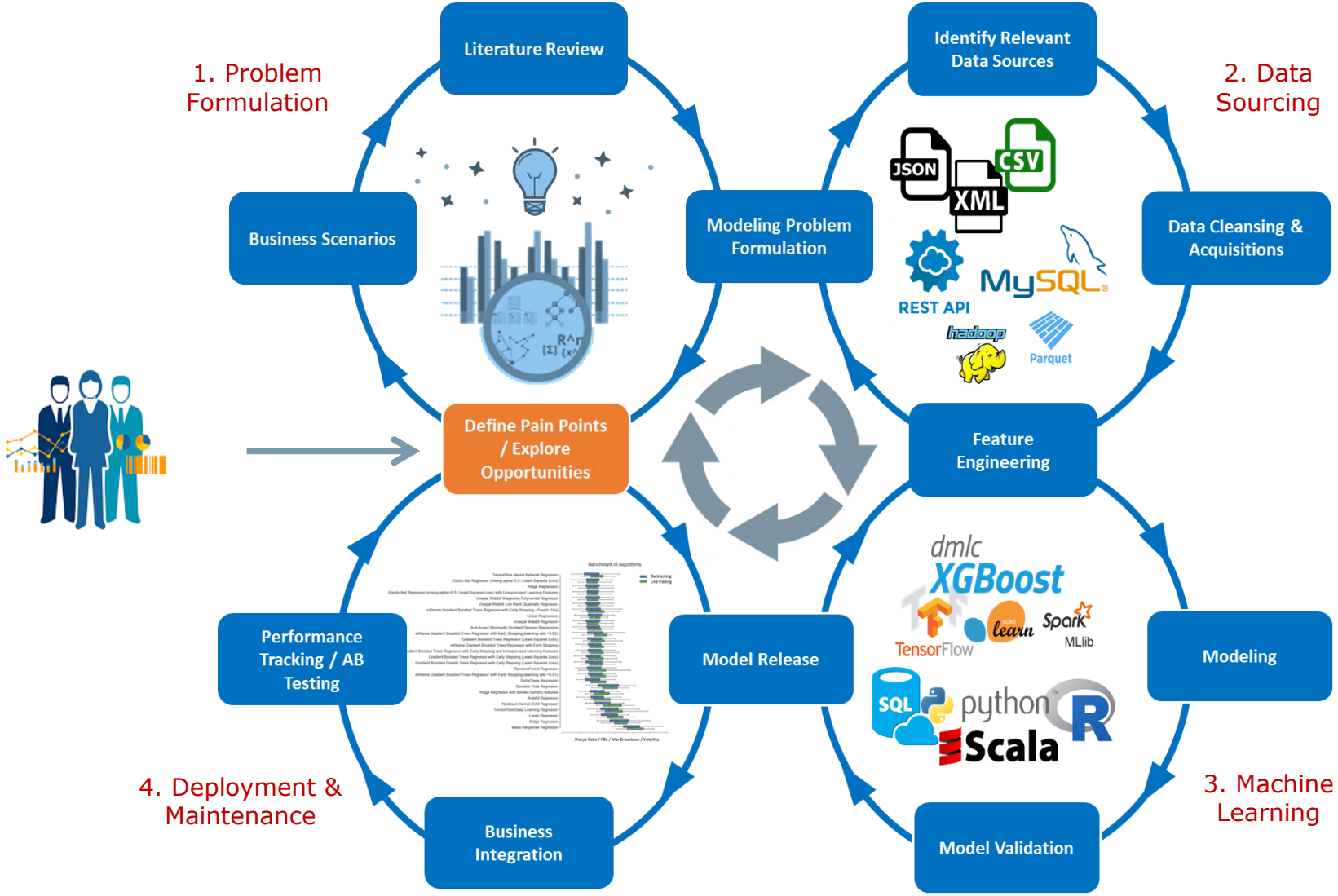




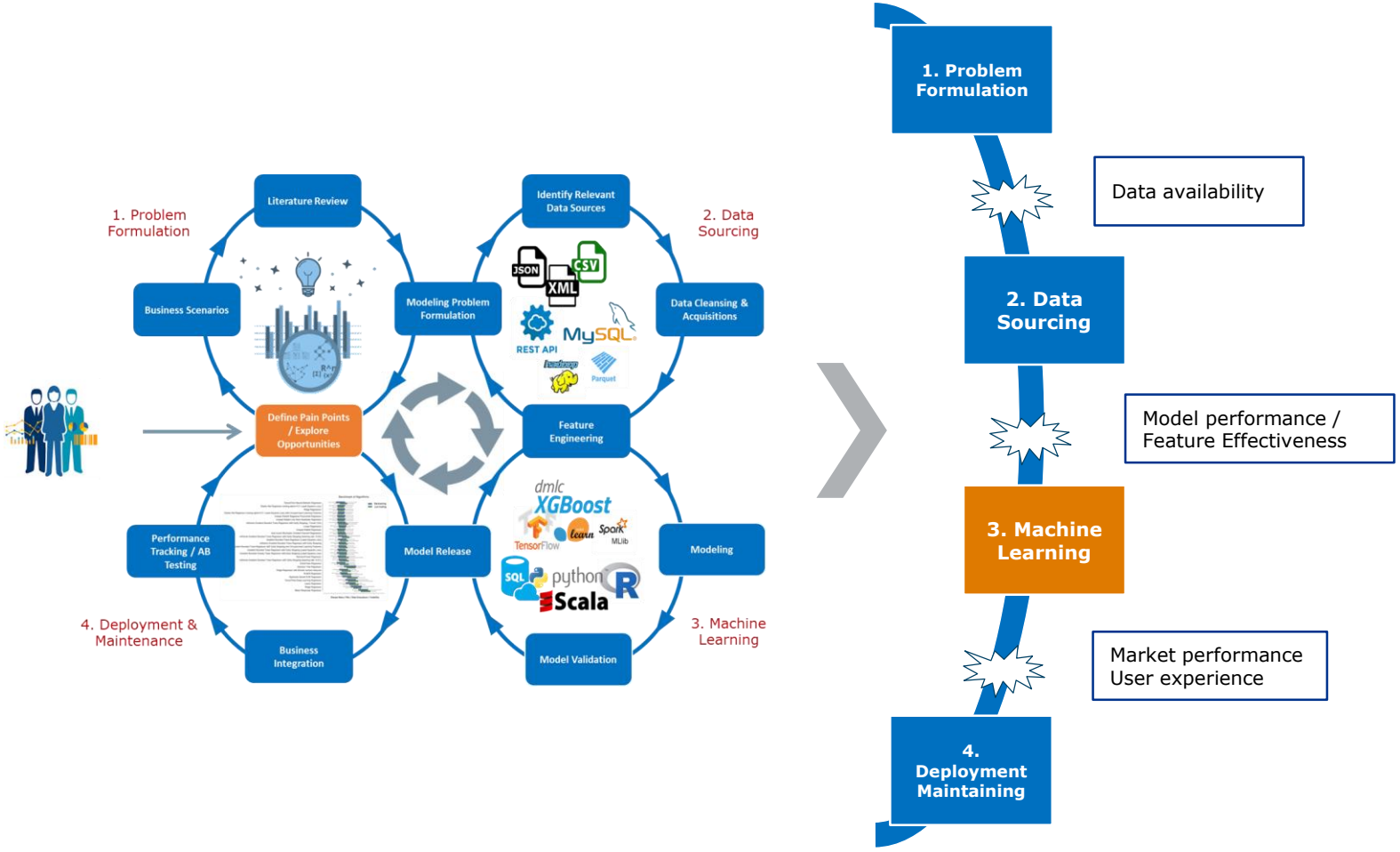
DATA SCIENCE LIFECYCLE



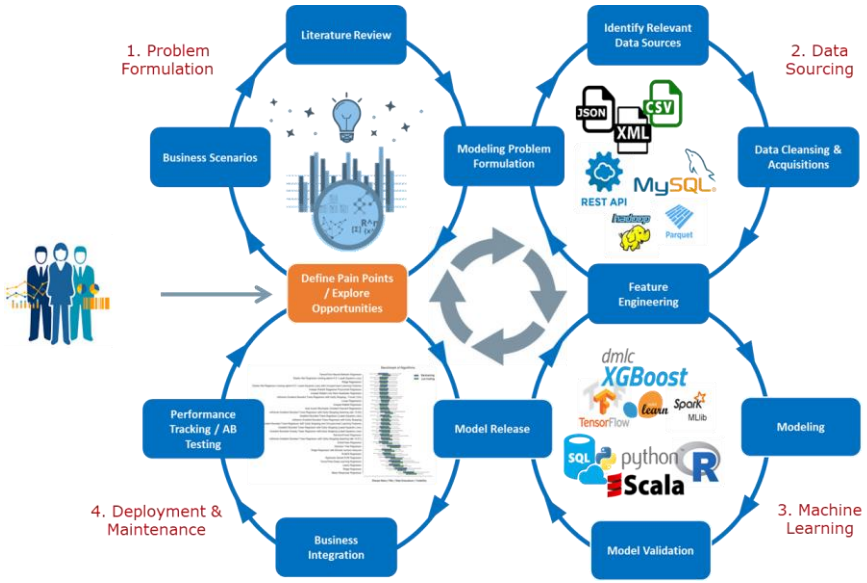
DATA SCIENCE LIFECYCLE

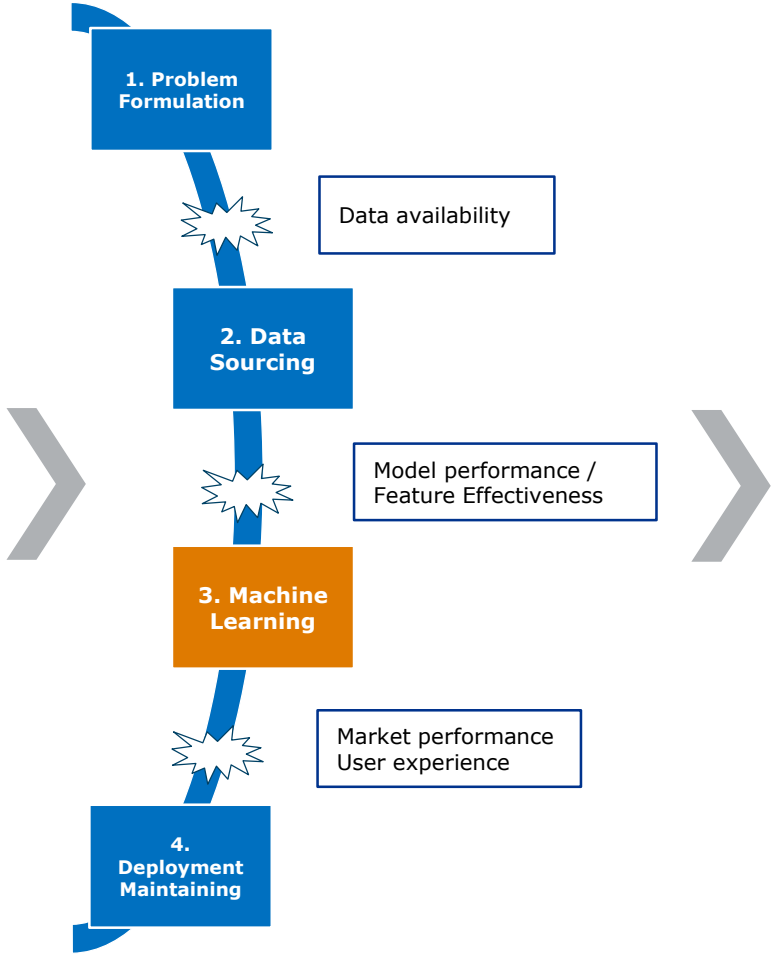


DATA SCIENCE LIFECYCLE



DATA SCIENCE LIFECYCLE





We tend to go for the **most complicated** solutions first which may result in wasting months doing the wrong thing.

We tend to engage heavily with stakeholders in very early stage to define the "**best**" outcomes.

We try to solve problems where a solution is not obvious and success is not guaranteed – **extreme uncertainty**

We have limited resources and the resources are **expensive**.

…

Data Scientists
are
Researchers

Data Science
are
R&D activities

Applied Data Scientists
are
Start-up Founders

Data Science
are
Start-ups

THE LEAN START-UP

START-UP

A start-up is a human institution designed to create a new product or service under conditions of extreme uncertainty.

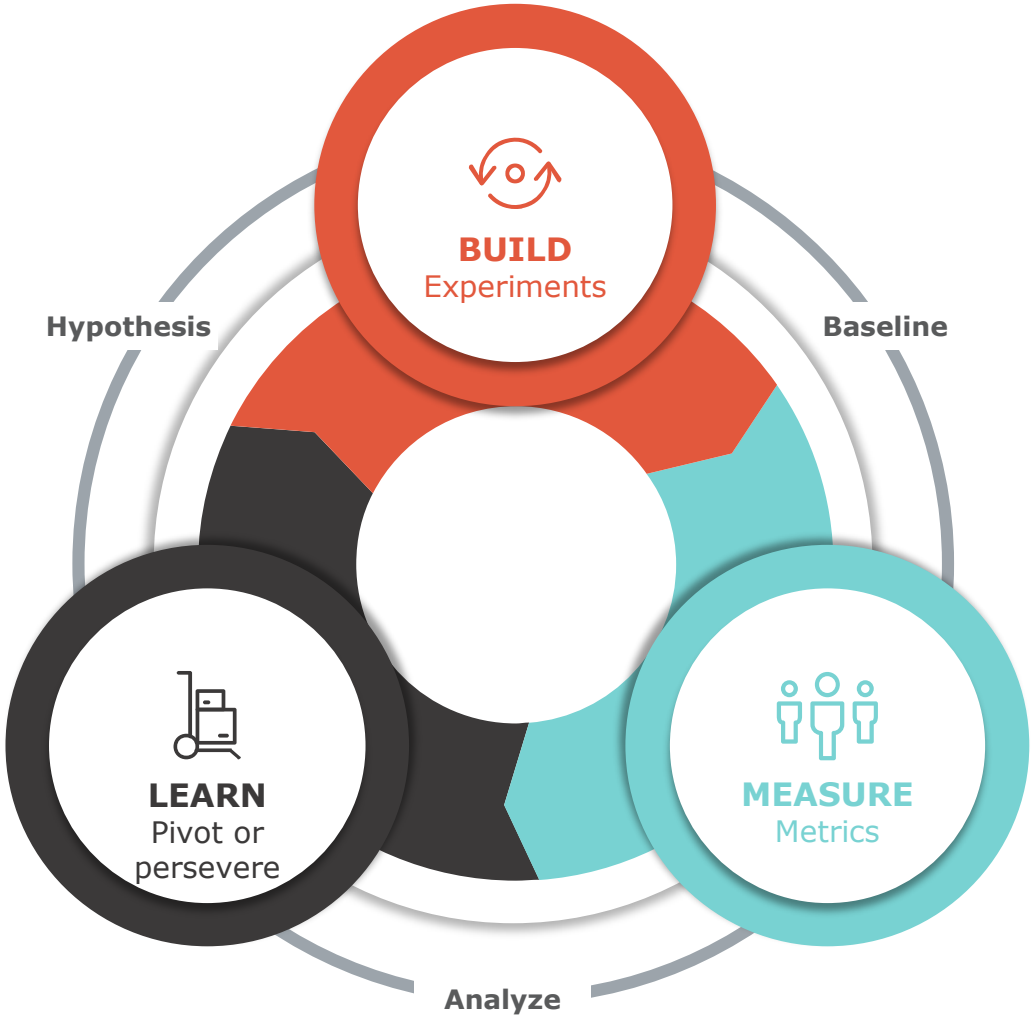
—Eric Reis The Lean Startup

DATATHON / DATA2APP

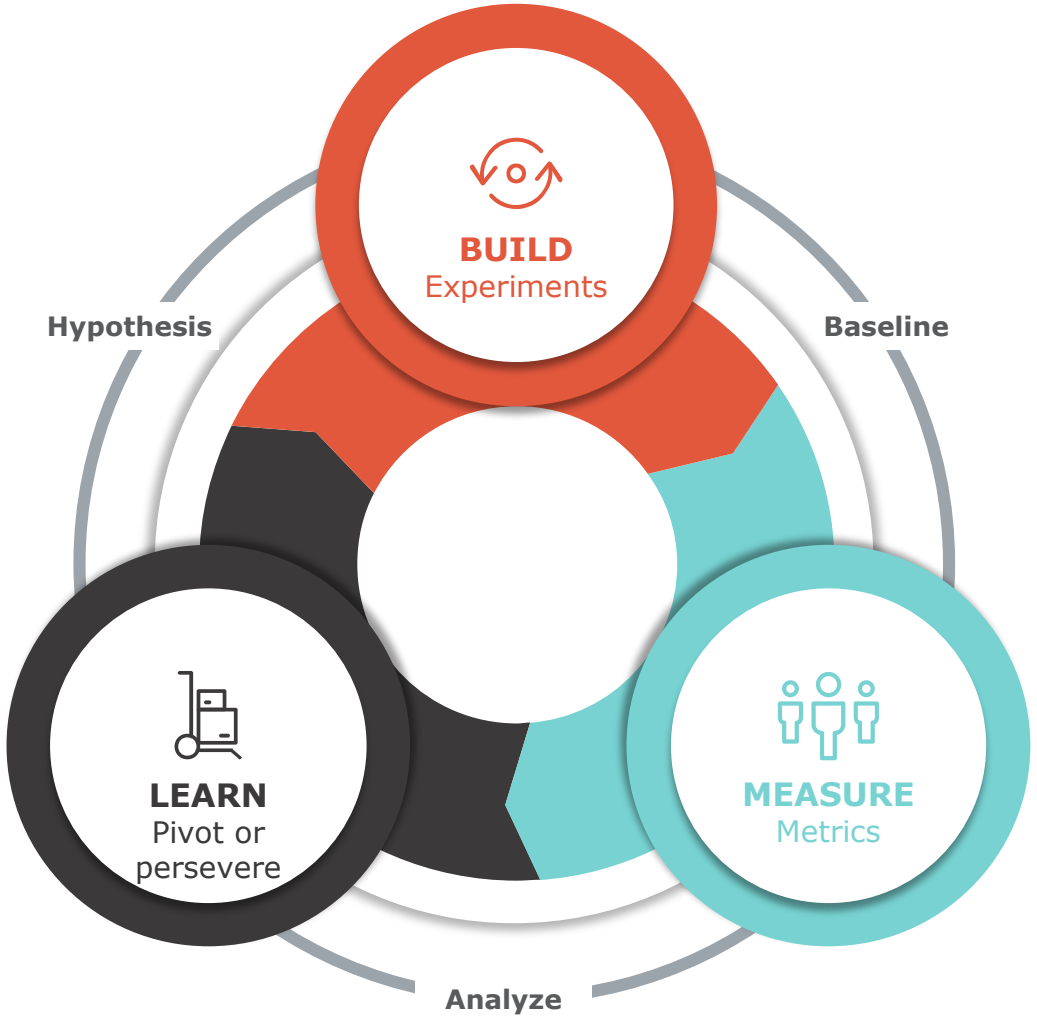
Data Science / Data2App is like building a new product or service at start-up.

—ANZ Insto Data Science

THE LEAN START-UP



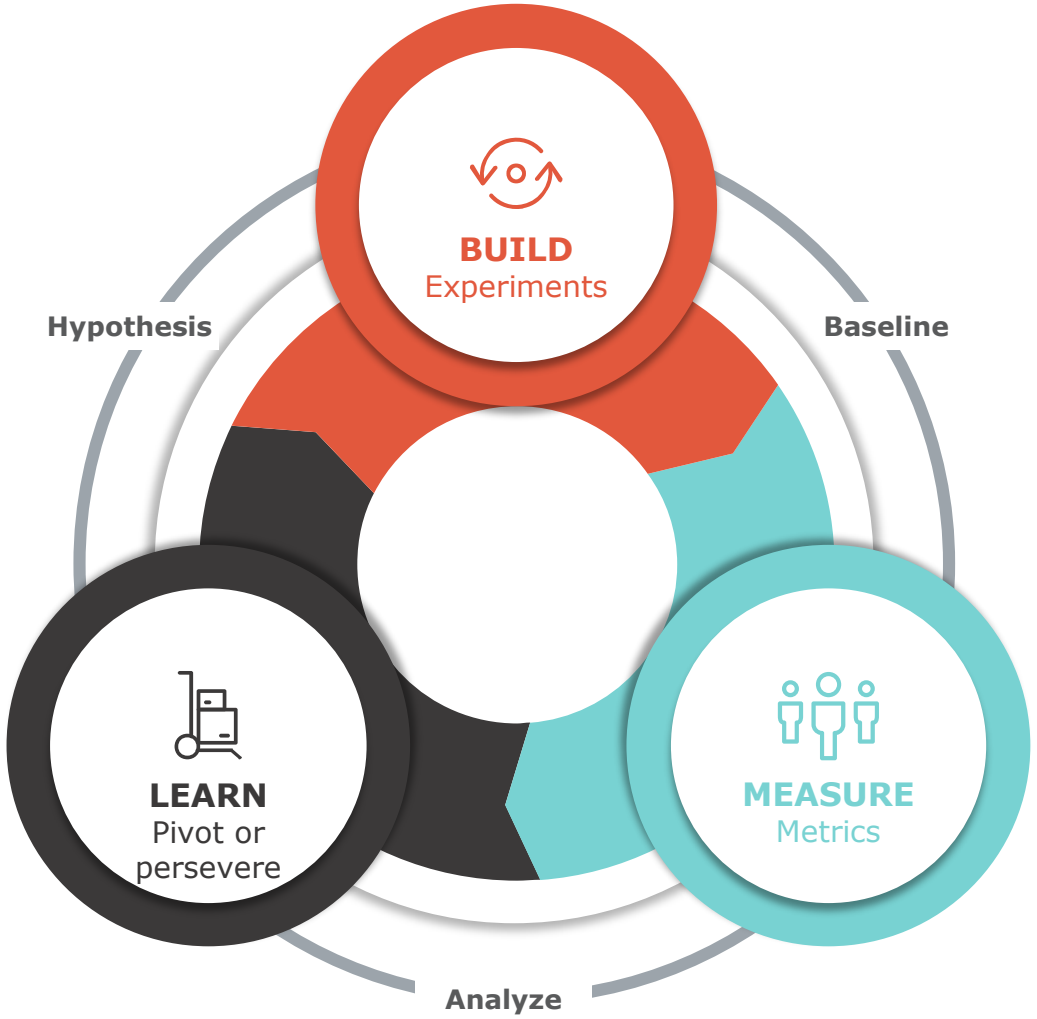
THE LEAN START-UP: HYPOTHESIS



Leap-of-faith Assumption

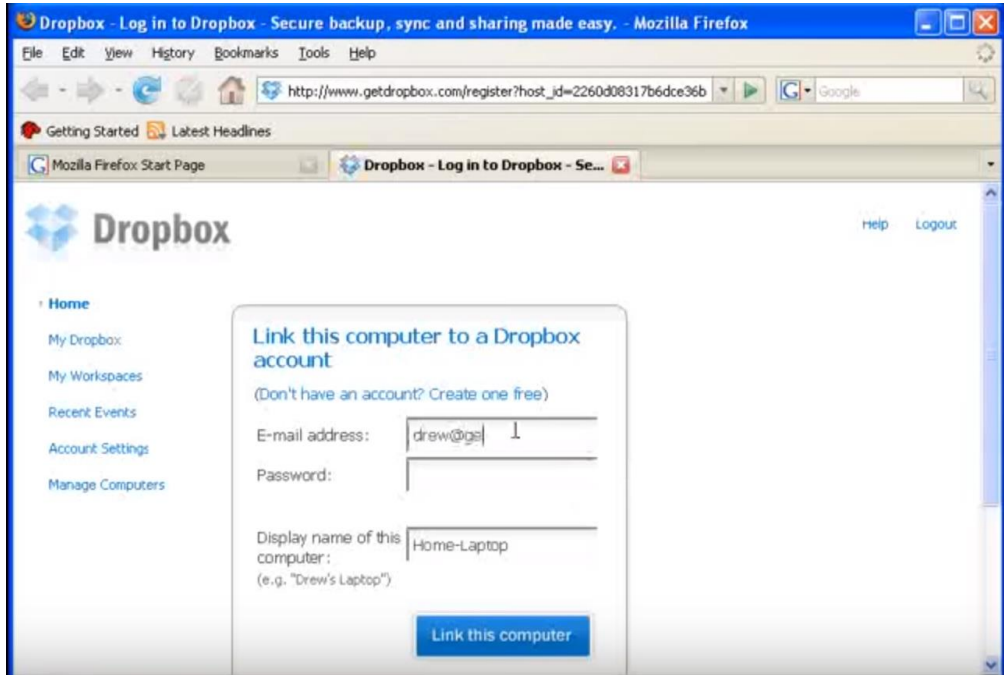
- Value hypothesis
- Growth hypothesis

THE LEAN START-UP: BUILD



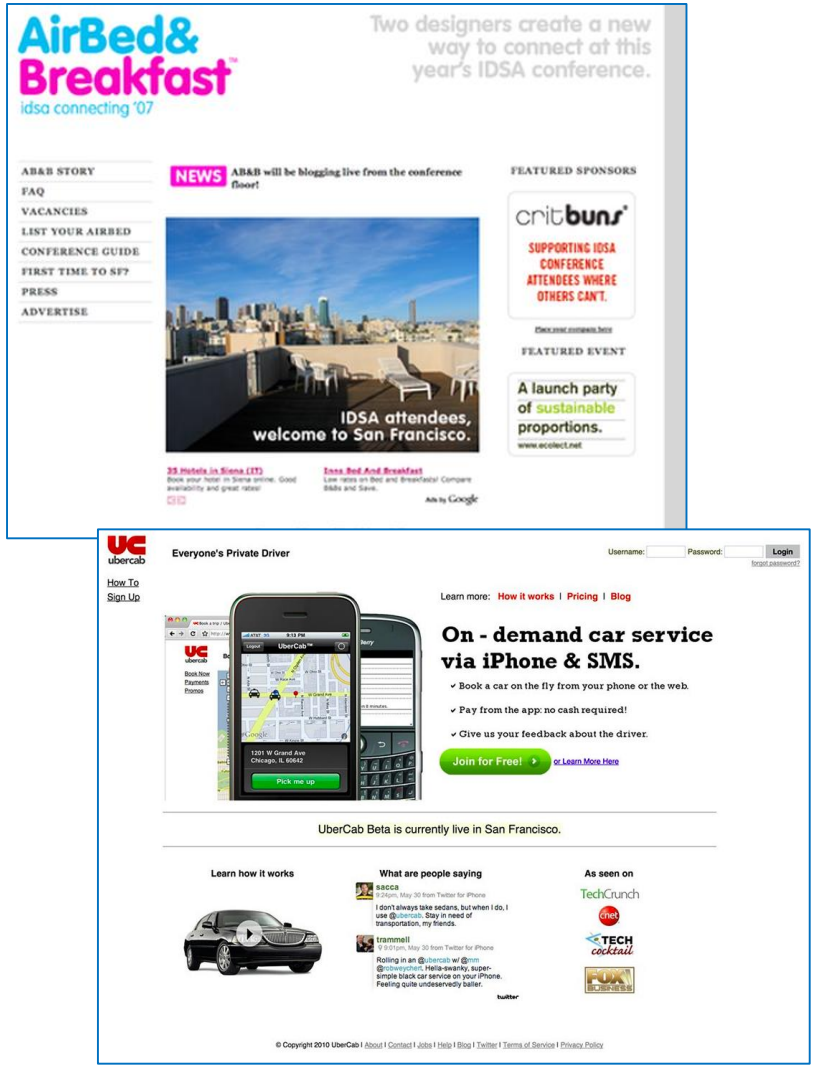
Build

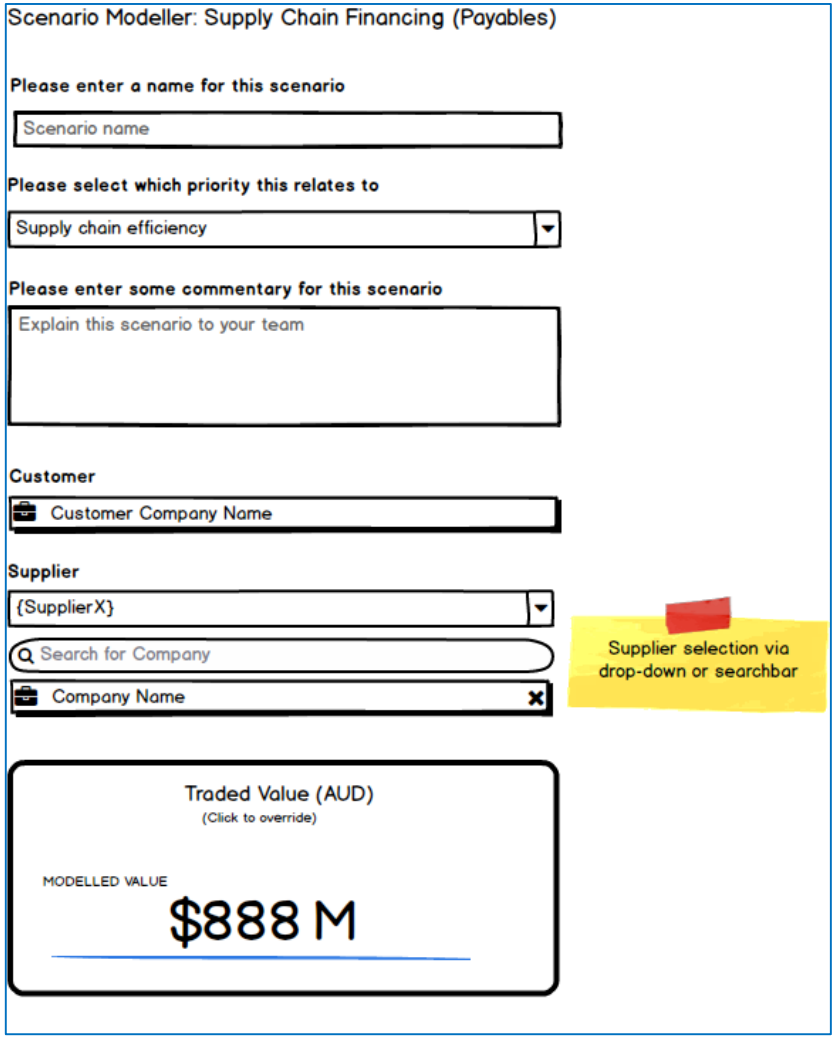
- Concierge Minimum Viable Product



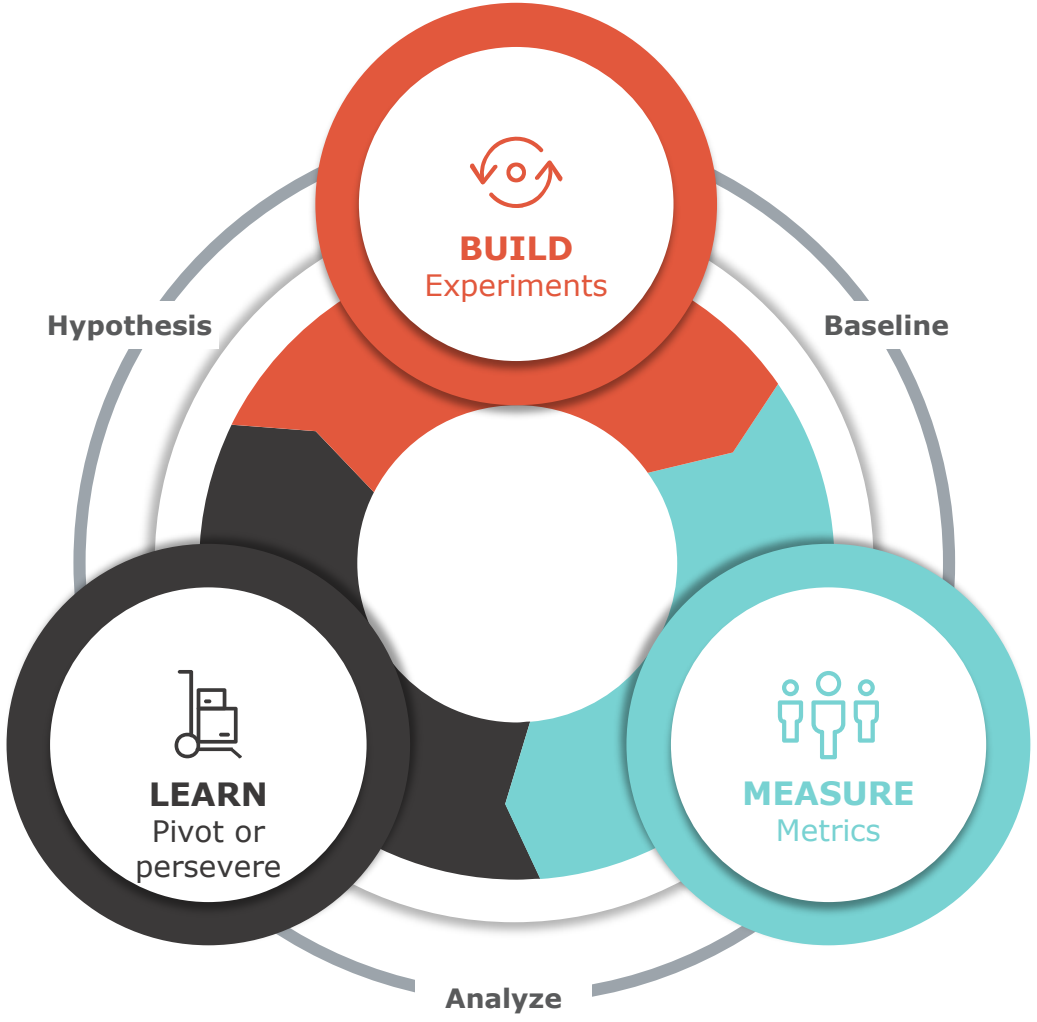
DROPBOX MINIMAL VIABLE PRODUCT

CONCIERGE MINIMAL VIABLE PRODUCT



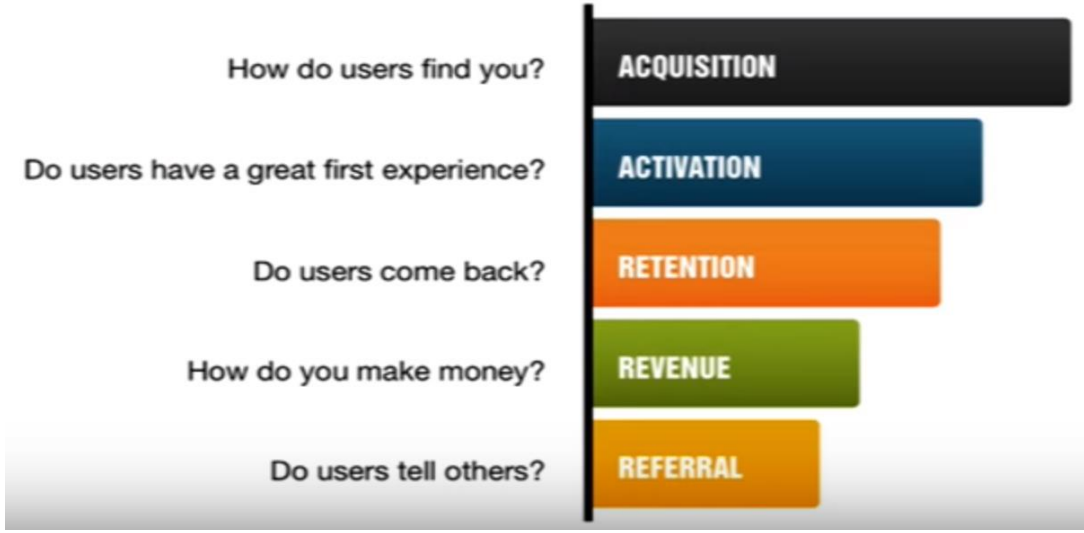


THE LEAN START-UP: MEASURE

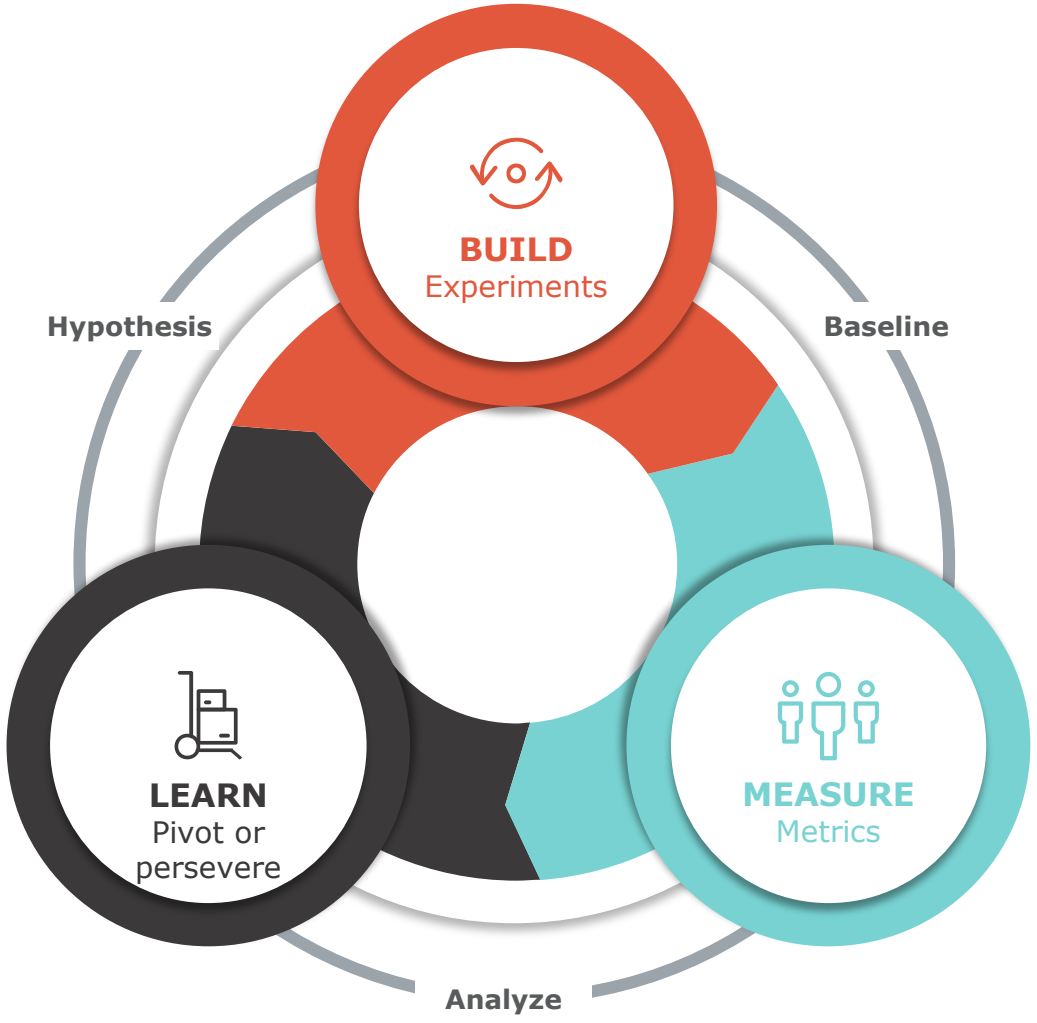


Measure

- Innovation accounting
- Actional Metrics

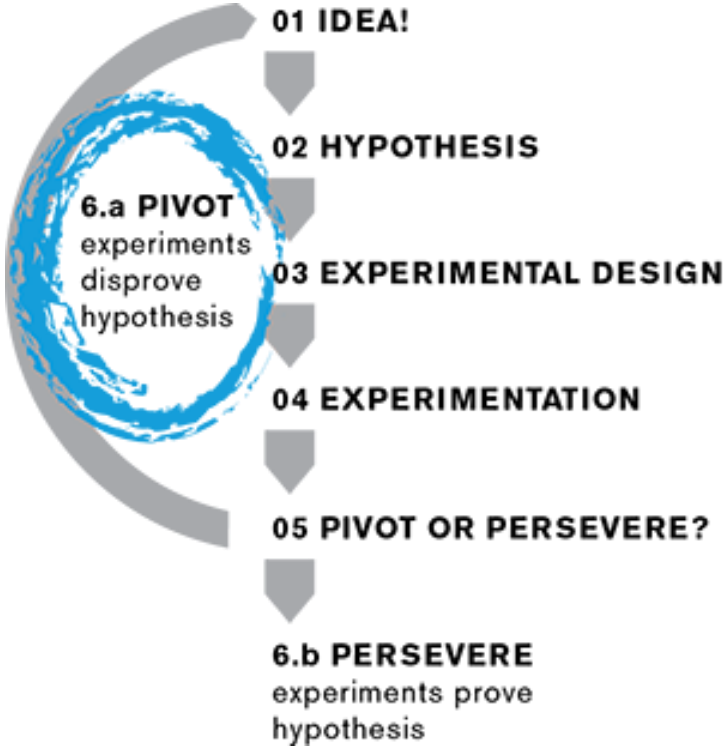


THE LEAN START-UP: LEARN



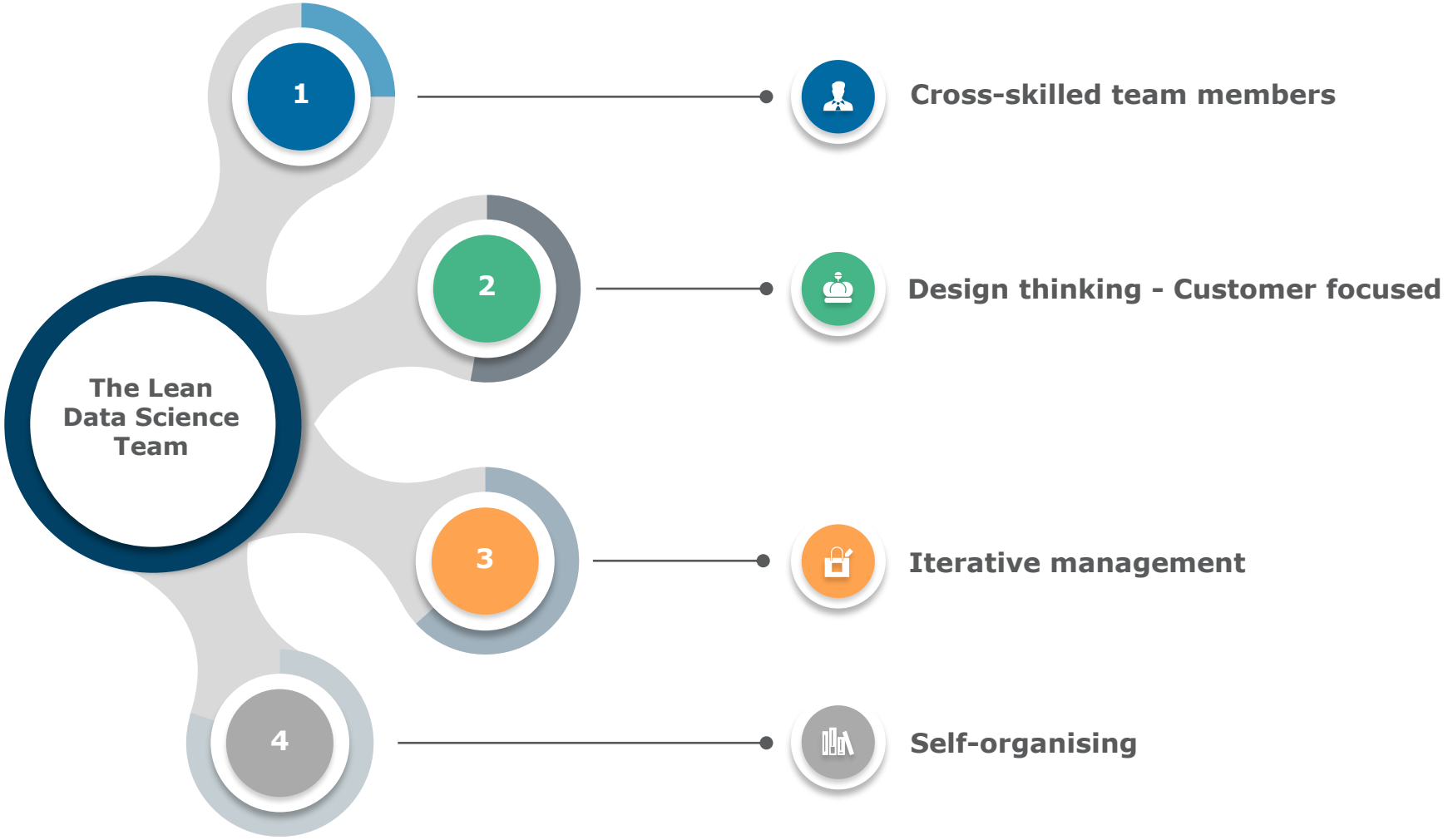
Learn

- Pivot
- Persevere

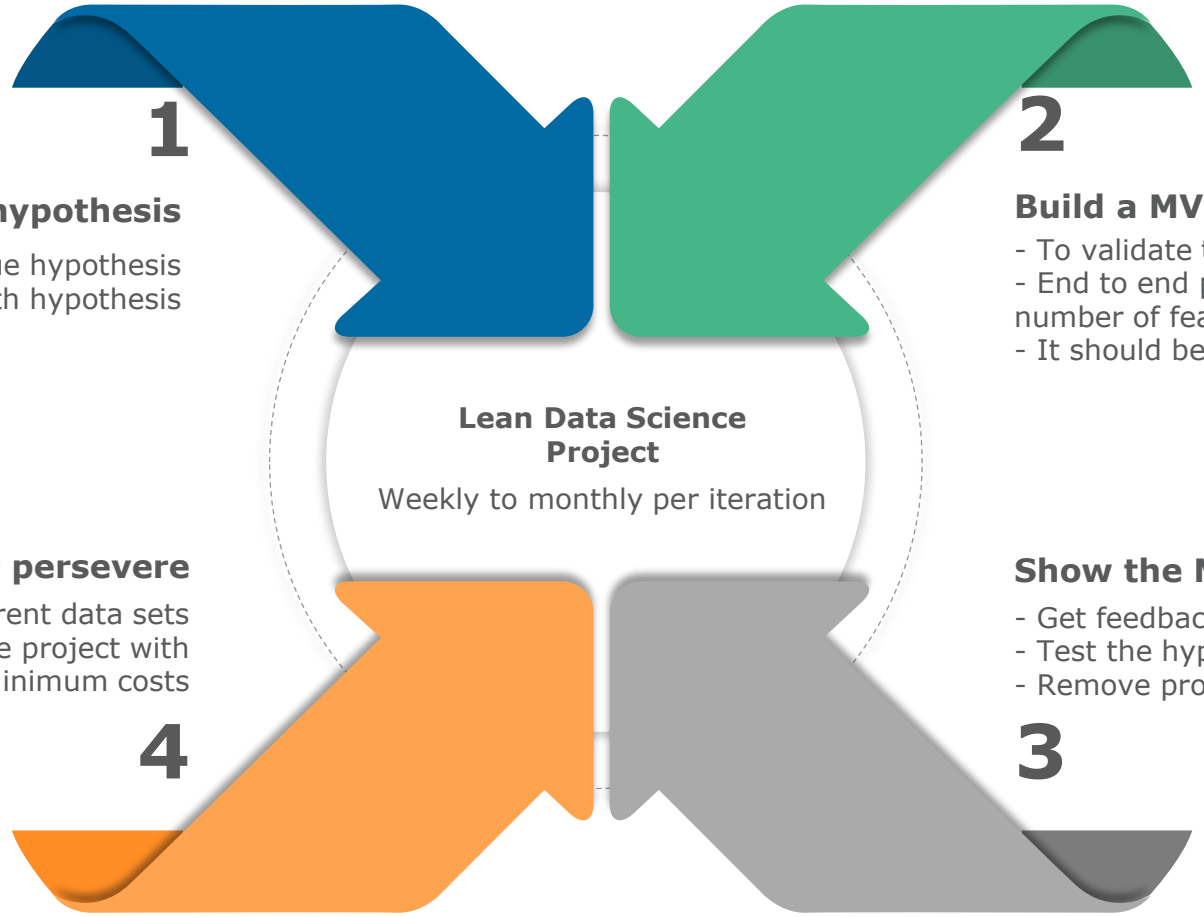


HOW TO APPROACH THE DATATHON

THE LEAN DATA SCIENCE TEAM



THE LEAN DATATHON LIFECYCLE



1

Set the hypothesis

- Value hypothesis
- Growth hypothesis

2

Build a MVP

- To validate the hypothesis.
- End to end product with the minimum number of features.
- It should be terrible.

Lean Data Science Project

Weekly to monthly per iteration

Pivot or persevere

Killing a project, bringing in different data sets or completely changing the project with minimum costs

4

Show the MVP to the customer

- Get feedback from customer
- Test the hypothesis
- Remove project risks

3

THE LEAN DATA SCIENCE TOOLS

Engineering

Cloud Computing

Customer Analytics

Lean Modelling

Visualisation

Work with people

CASE STUDY

CASE STUDY

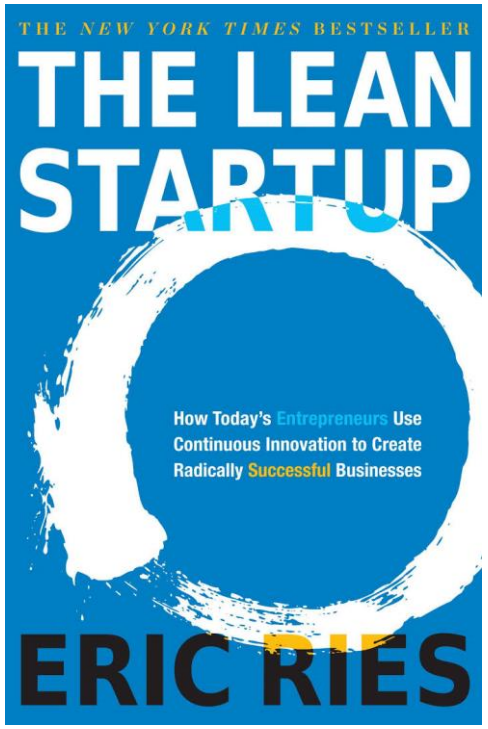
https://www.kaggle.com/c/home-credit-default-risk

Hypothesis:

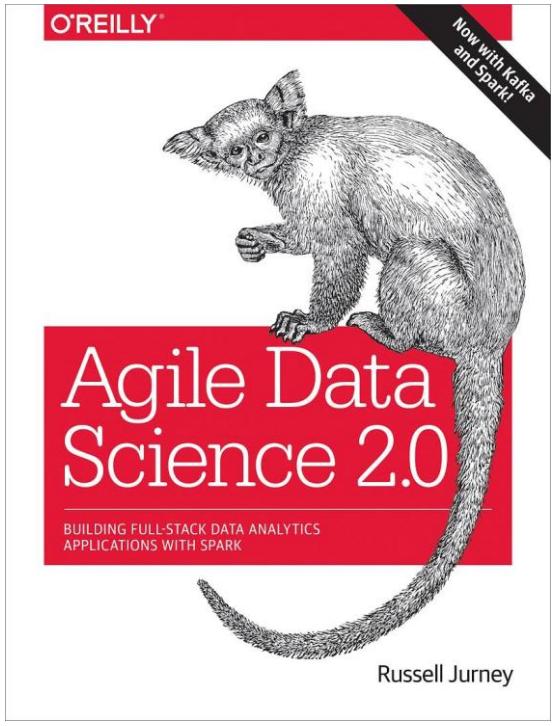
MVP:

Measure:

LEARNING MATERIALS



MELBOURNE DATATHON



THANK YOU

Tianxiang (Ivan) Liu

https://www.linkedin.com/in/ivanliu1989/
Ivan.Liu@anz.com