6 August 2018

LEAN MODELING

Terrence Szymanski
Data Scientist, ANZ

SEXY DATA SCIENCE & DATA SCIENTIST

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

ANZ

AGILE / LEAN / ITERATIVE DATA SCIENCE

Iteratively Finding a Good Mac

Towards Data Science

Iteratively Finding

An Iterative Approach to Data

NYC DATA SCIENCE ACADEMY | Blog

BLOG HOME > DATA SCIENCE NEWS AND SHARING > AN ITERATIVE APPROACH TO DATA SC

An Iterative Approach to Data Science

Measure Once, Cut Twice: Mo

DataRobot

'Lean' Data Science

DOMINO

DATA SCIENCE CODE MACHINE LEARNING PRACTICAL TECHNIQUE

Agile Data Science (2.0!) | Mee

Terrence

Meetup

Start a new group Login Sign up

Past Meetup

Agile Data Science (2.0!)

This Meetup is past 300 people went

Hosted by Phil B.

From Data Science Melbourne

Public group

Thursday, 17 May 2018

5:30 pm to 8:00 pm

Zendesk Basement

395 Collins St, Melbourne · Melbourne

What we'll do

Join us after work from 5:30 for a beer and pizza.

Help yourself to pizza, and if you see the pizza man outside please give him a hand carrying them in. There should be plates and serviettes floating around.

Please assist in keeping the venue tidy during the evening - please put you

How Data Science Teams Can

Insights Summits OnDemand

Menu

innovation enterprise CHANNELS

Big Data.

How Data Science Teams Can Win By Behaving Like Startups

23 Mar

Ben Dias Royal Mail's Head of Data Science reveals a practical guide to acting like a startup

Analytics Artificial Intelligence Big Data

3141 VIEWS

There are many ways in which data science teams are like startups. By adopting a lean startup methodology, corporate data teams can navigate internal politics and deliver bigger and better results.

Speaking at Predictive Analytics Innovation Summit, Ben Dias shared for the first time in public how Eric Ries' seminal book on startups has helped him to achieve successes in building high-performing data science teams at Royal Mail and Tesco.

Taking an Agile Approach to D

ORACLE + DATASCIENCE.COM

Platform Solutions Resources Blog Customers Company

BEST PRACTICES, DEVOPS, IT MANAGEMENT

Taking an Agile Approach to Data Science

Brittany-Marie Swanson | 11.02.17

Data science is powering applications around the clock, from Netflix's powerful content recommendation engine to Amazon's virtual assistant Alexa. But not all companies are equipped to support large-scale, always-on data science projects; doing so requires an infrastructure that can handle the large volumes of traffic these models receive while in production, as well as a process for getting a finished model into a production environment.

Want to share your expertise on our blog?

WRITE FOR US

Fractal spiral (source: Pixabay)

doing Agile data

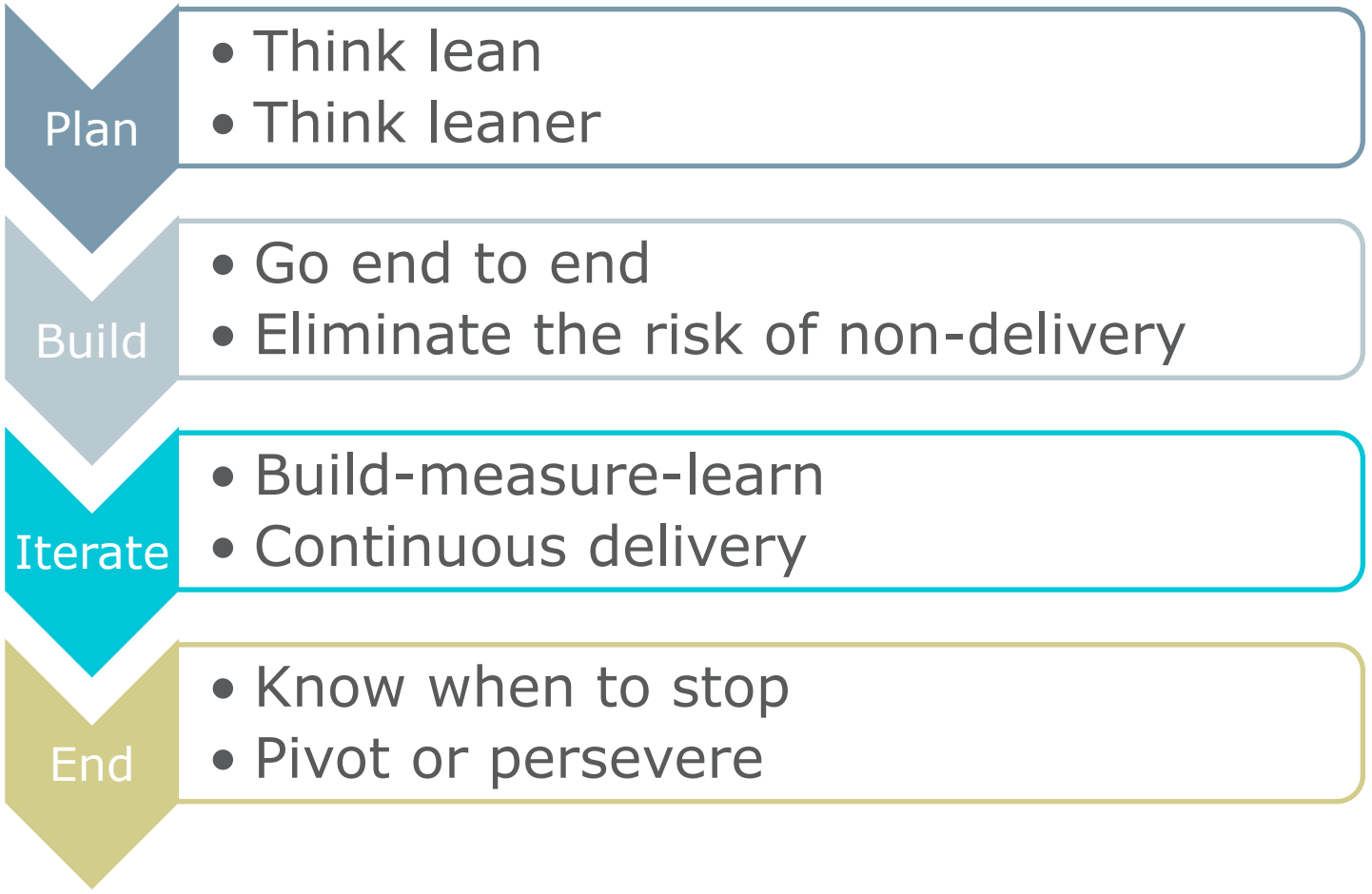
evolutionary

applications—for a

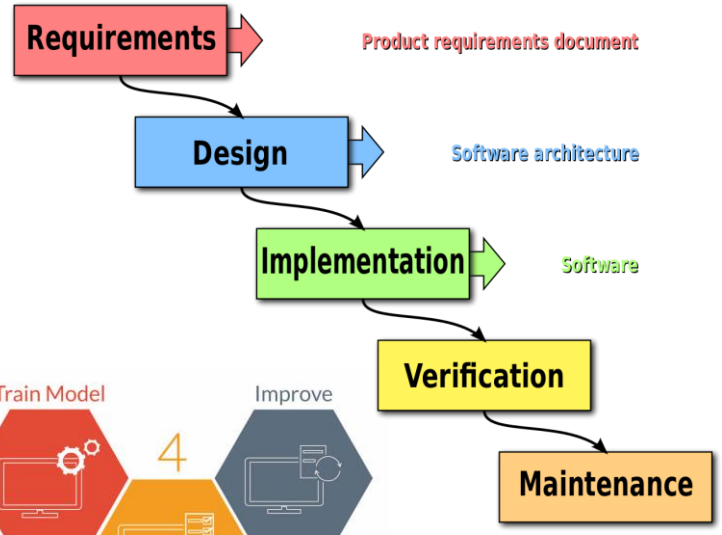
decade, since before I knew what to call it. Working as a lone full-stack developer, it was only natural to iteratively evolve the analytics software I built. When I joined a team, I expected things to work that way. They didn't.

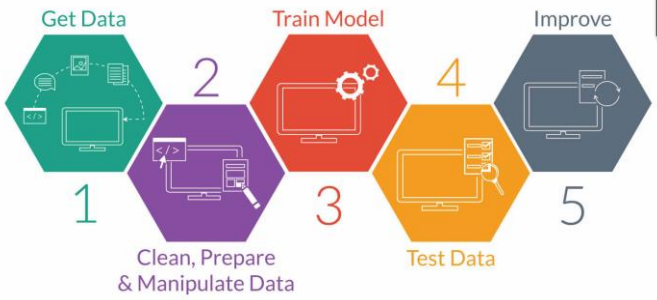
I had been exposed to Agile methods as a web developer, so I was surprised when I entered my first job as a data scientist to

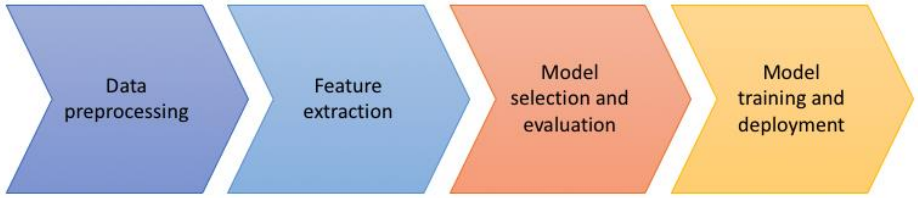
LEAN MODELING AND LEAN DECISION-MAKING

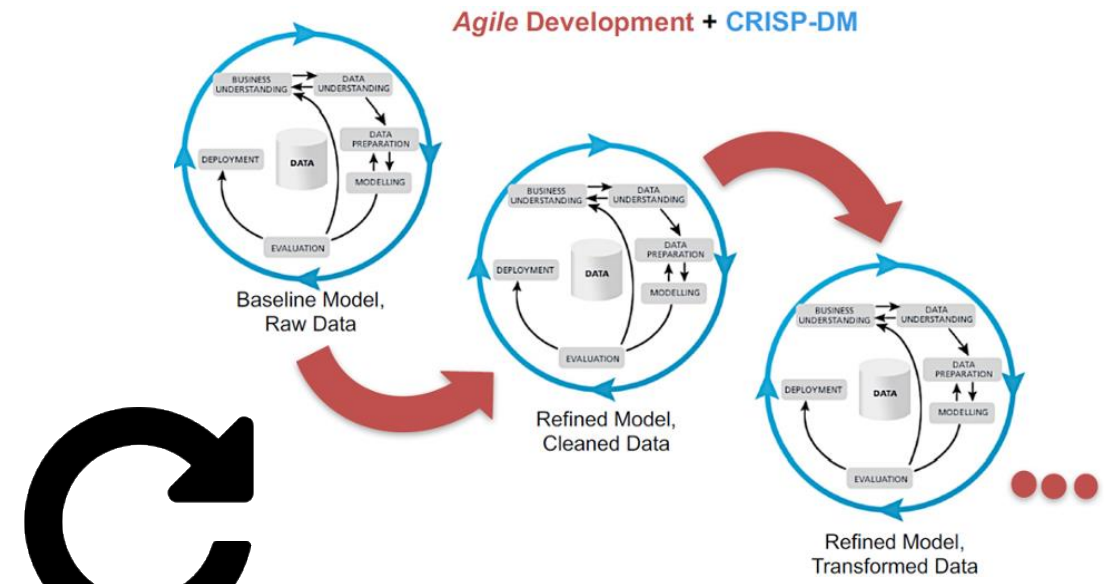


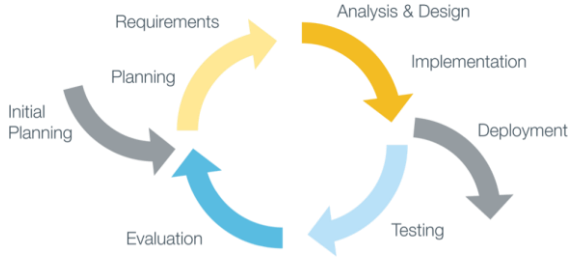
“WATERFALL” VS “AGILE”











WHAT IS LEAN? WHAT IS AGILE?

- The term “lean” gets thrown around a lot
- It comes from *The Lean Startup* by Eric Reiss
- Generally, it refers to maximizing learning while minimizing wasted time and effort.
- It does *not* mean building a bad or inferior product.
- “Agile” refers to a set of principles used for software development.
- They both involve iteration and metrics to achieve a desired result.

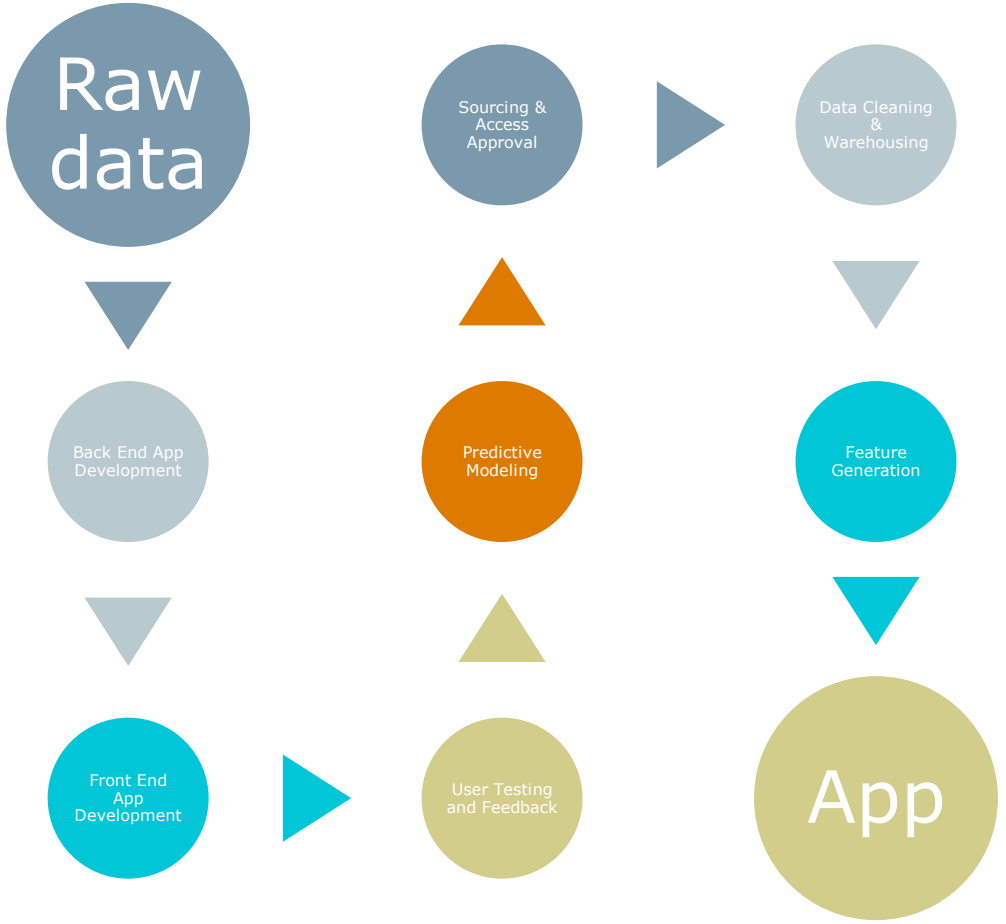
On our team, we sometimes talk about being **ruthlessly lean** – eliminating all non-essential waste and getting to the end state as quickly as possible.

- If I asked you to build a model to predict customer age, and I need it in two hours, what would you build?
- (Hint: go leaner)

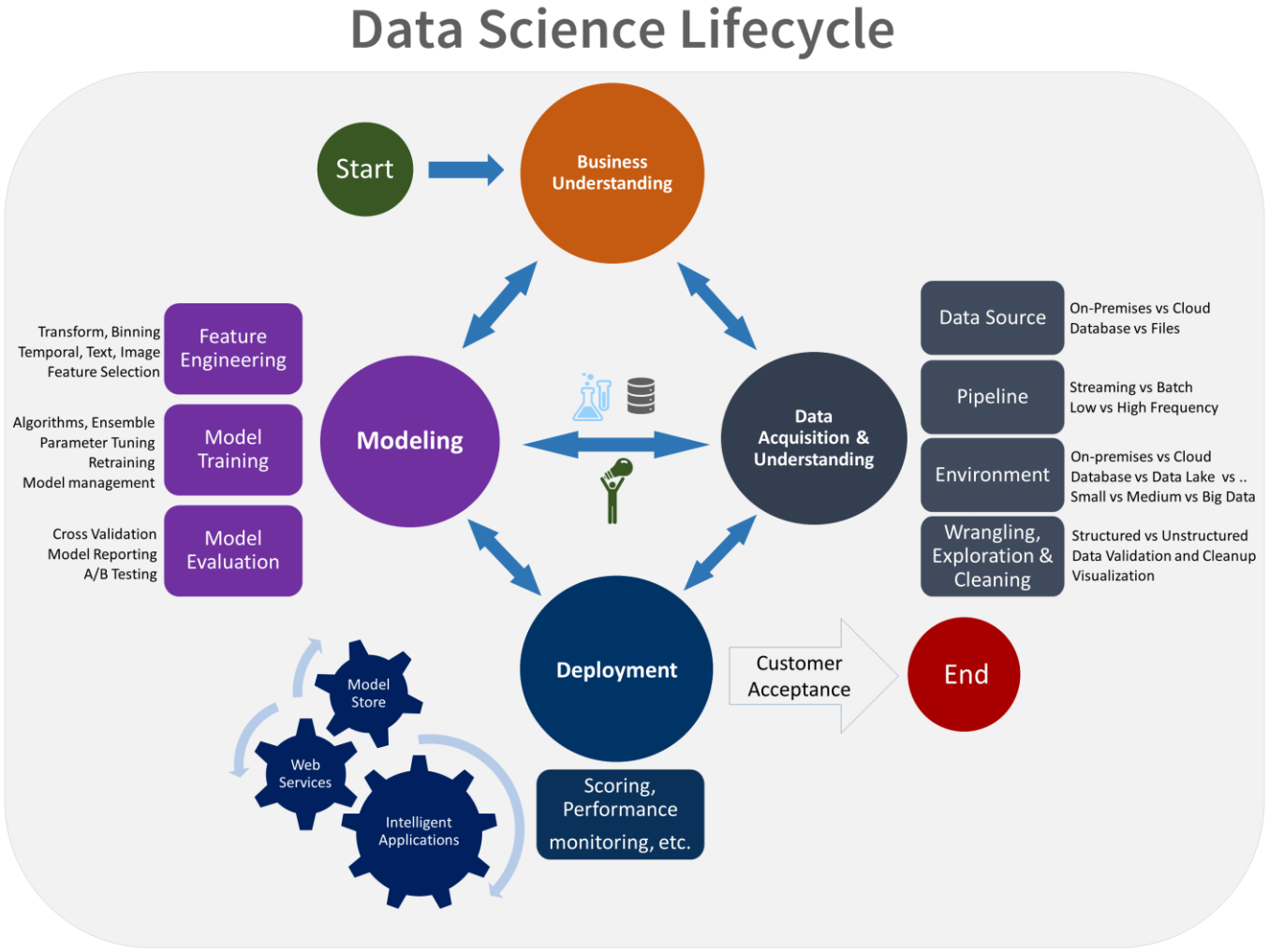
GOING END TO END

"Eliminate the risk of non-delivery"
- Tez

- If you have two months to complete a project, finish the first version in one week.
- Even if things go horribly wrong after that, you still have something to show.
- Identify any roadblocks as early as possible.
- Everything is easier the second time around



A DATA SCIENCE PRODUCT HAS MANY COMPONENTS



- All of these components are necessary and must work together.
- With coordination, they can be developed in parallel in a lean and agile way.

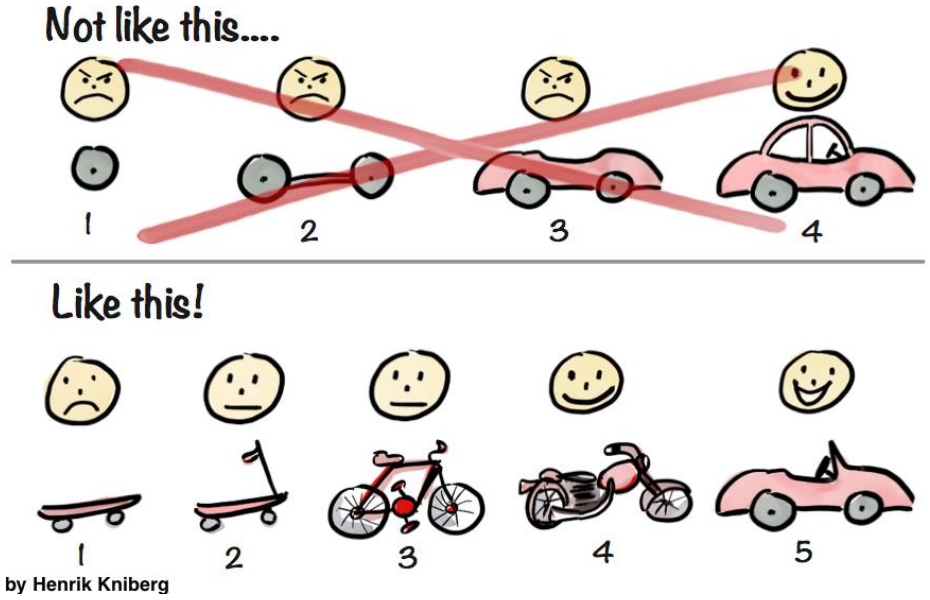
From https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

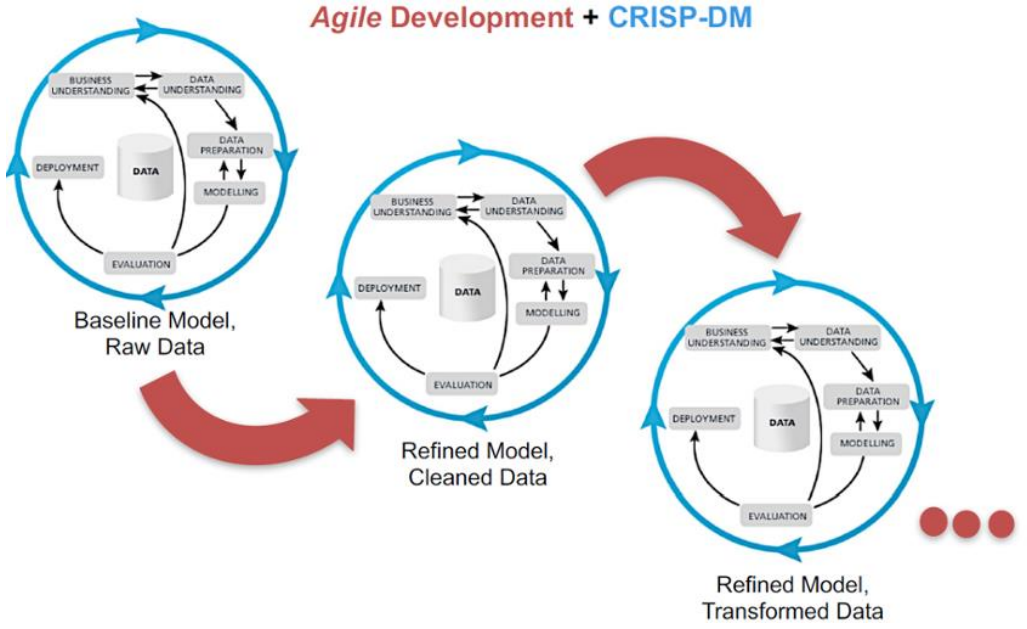
MINIMAL VIABLE PRODUCTS (MVP) FOR DATA MODELING

The role of a MVP is to test hypotheses

- I can obtain access and approval for the data I need
- A model can built with sufficient accuracy to meet the business needs
- The model provides a demonstrable value to the business
- My model can scale and be integrated in a production system

The MVP should lead to continuous delivery of working software





EVALUATING MODELS IN CONTEXT

In ML research and Kaggle competitions, the best model is the one that maximizes a specific accuracy metric.

In data science, the best model is the one that delivers the most impact to the business (powering applications or informing decision-making processes)

For any new model development, ask the question: Is the payoff worth the investment?

Intrinsic evaluation:

- Evaluate a model in isolation
- E.g. accuracy / F1 / AUC

Extrinsic evaluation

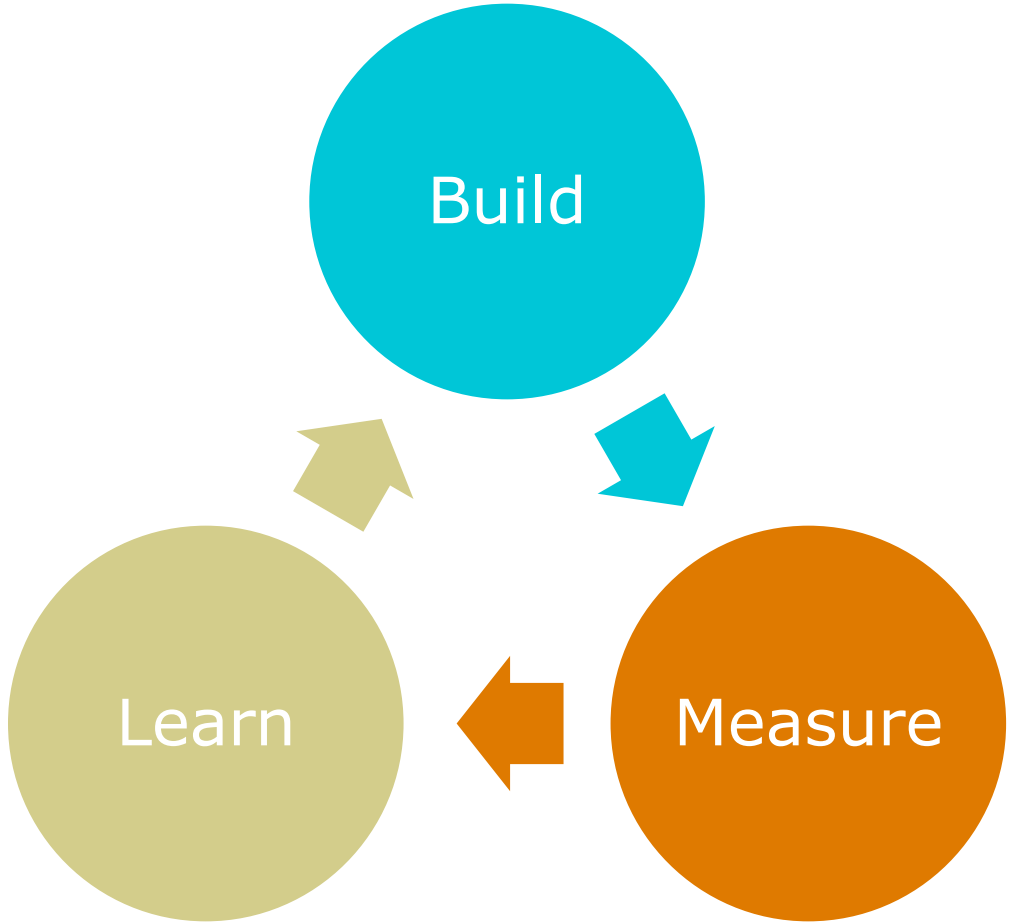
- Evaluate a model based on its contribution to a larger system
- E.g. A/B testing

THE BUILD-MEASURE-LEARN CYCLE

Iterate on successive models in a structured way.

Not just randomly trying different things.

Think about the metrics you measure.



WHAT IS A LEAN MODEL?

Lean:

- Quick to implement
  (this depends on your skillset)
- Efficient training and inference
- Interpretable
- Potential for scalability
- Can be integrated into broader system

Not lean:

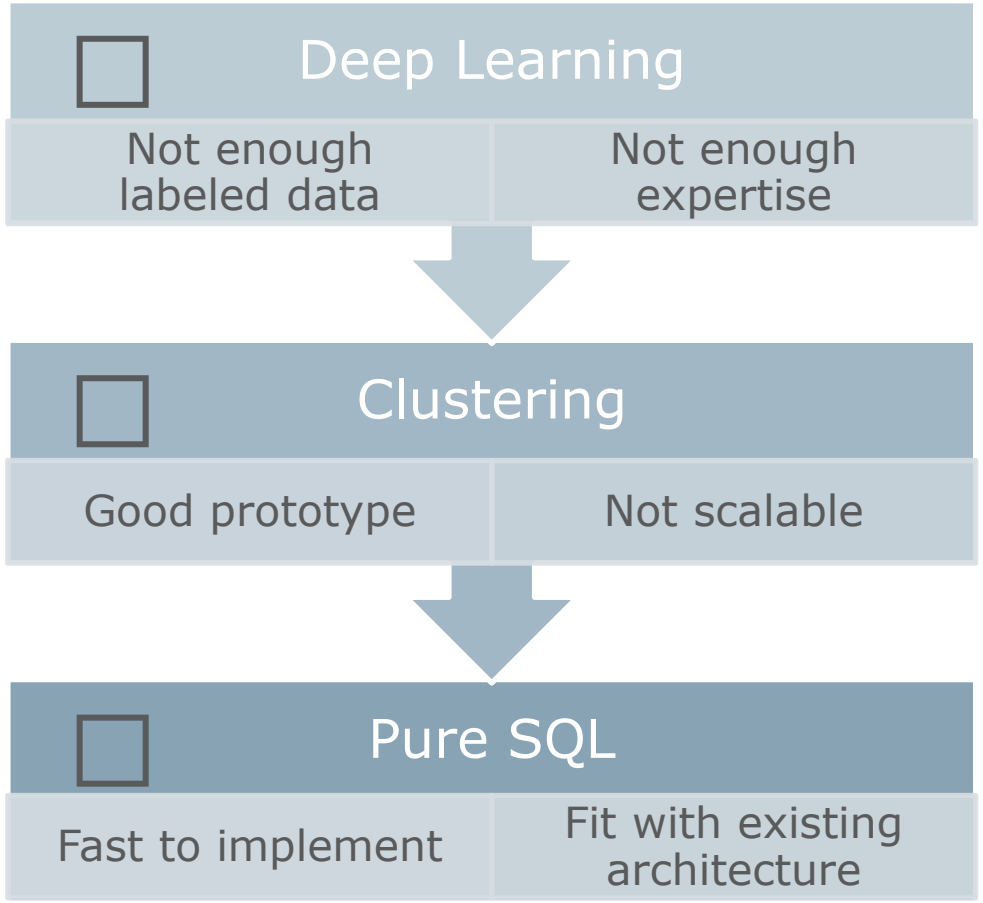
- The latest and greatest new library
- Lots of parameter tuning or feature engineering
- Opaque / black box
- Long training times or inference times
- Impractical to scale up

Are these lean techniques?

- Logistic regression
- Random Forests
- K-nearest neighbours
- Deep learning
- Rule-based systems

SOME PERSONAL EXAMPLES

Text Labeling



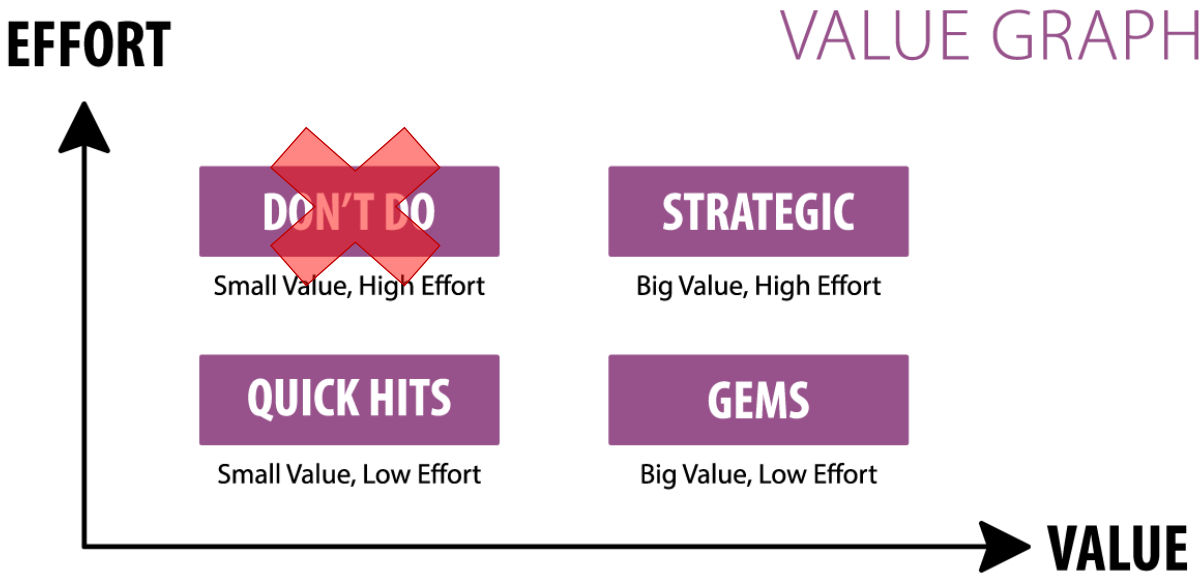
Report Generation

Need to aggregate data on features A, B, C and D.

1. Get data for features A and B from existing feature store
2. Generate feature C because it's easy
3. Create dummy values for feature D because it's hard
4. Write the aggregation and report generation code
    1. Deliver this to stakeholders for feedback
5. Write the logic for D

KNOWING WHEN TO STOP

- Sometimes, external factors will tell you when you've run out of time or money.
- Resist the urge to build the perfect model. Don't optimize unless that is your current primary objective.
- Remember the 80:20 rule.
- Always be evaluating your own level of productivity – is there something else you could be doing with your time that is more valuable?



POTENTIAL PITFALLS OF LEAN MODELING

Pitfalls

- Your project might die an early death because the results don’t meet expectations
- Clients / stakeholders might not be used to viewing work in progress
- Your MVP might end up being the production product!

…and how to avoid them

- Communicate clearly. Explain that this is not the final result and you are in a process of collecting feedback to learn and iterate
- Always remember there is a distinction between “lean” and “inferior”
- Stay disciplined and stick with the process

HOW TO TAKE A LEAN APPROACH TO THE DATATHON

Strategy

- Go end to end as fast as you can
  - First, build the leanest model you can think of
- Iterate, iterate, iterate
  - Use the value graph to prioritize work an identify easy wins
  - Put yourself in the shoes of stakeholders at each iteration

Teamwork

- Minimize / eliminate dependencies between team members
- Collaborate continuously – don’t wait until the end

THANK YOU