Data Science and Agility

At Springer Nature

Charles Kubicek

Agile on the Beach 2024

@ckubicek
/in/ckubicek

Charles

- Software Engineer 20+ years
- 10 years agile and digital transformation
- Worked with data for a few years now

Talk Overview:

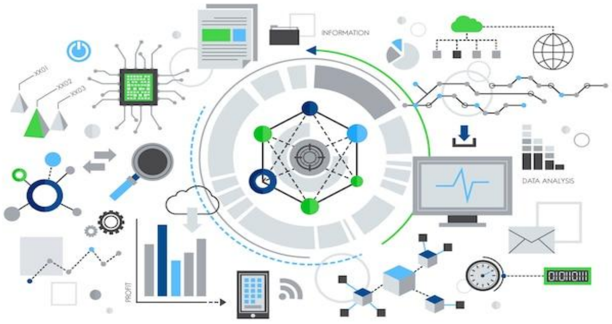
- Data Challenges
- Shift-Left
- Data-as-a-product
- Delivery

Scope:

- Data Science Output
  - An answer
  - A tool
  - A model

Scope:

- Data Science Output
  - An answer
  - A tool
  - A model



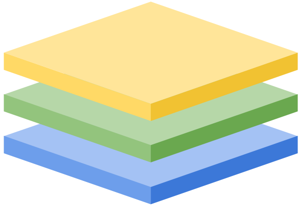
- A deliverable deployed into a production environment

1.

Challenges

Challenges

- Hidden raw data hand-off
- Data inconsistency
- Production hand-off



Challenge 1

Hidden data hand-off

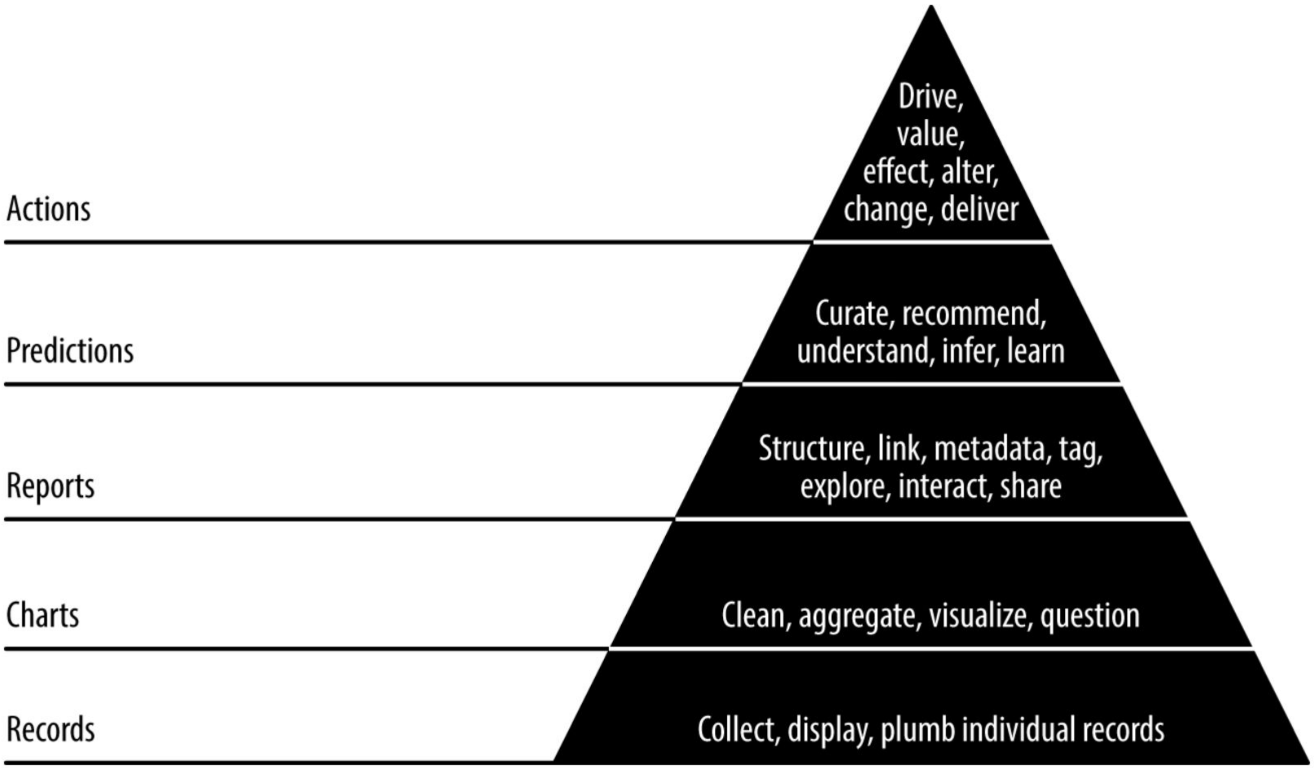


Figure 1. The data-value pyramid. Figure courtesy of Russell Jurney.

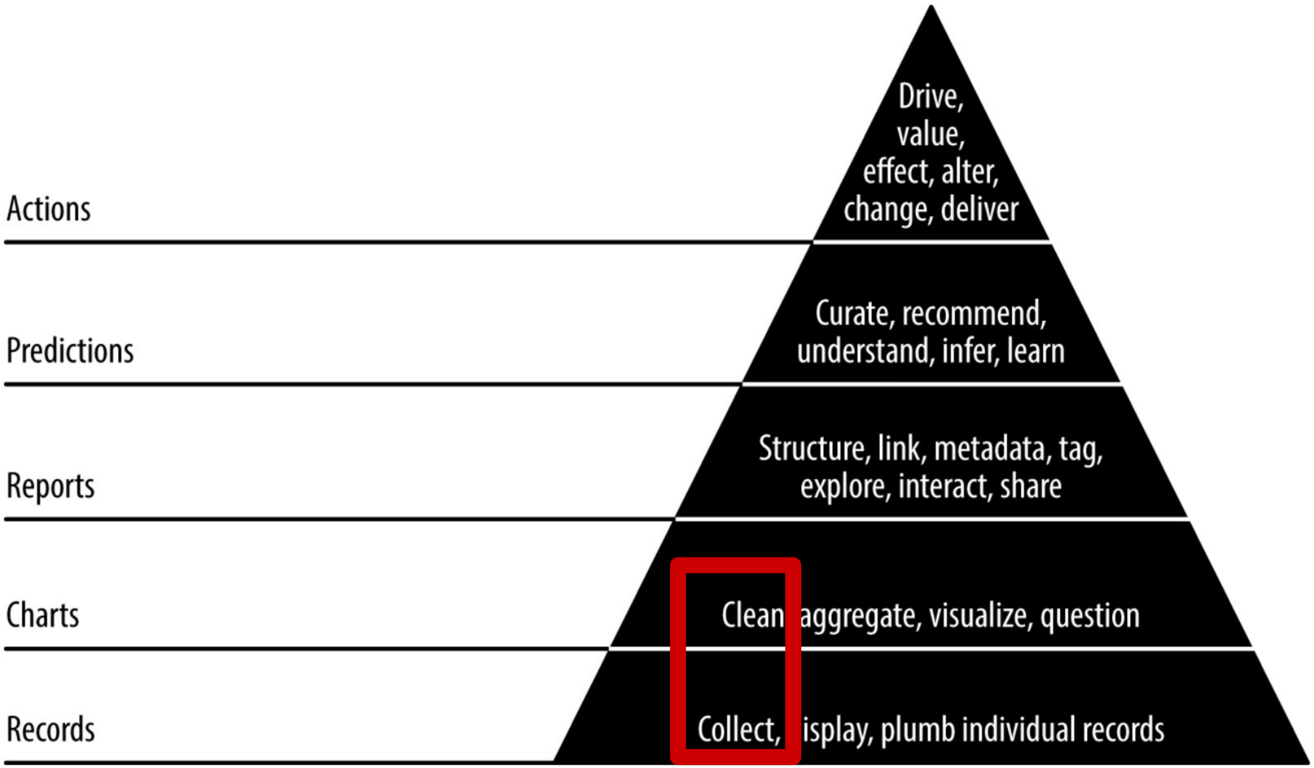


Figure 1. The data-value pyramid. Figure courtesy of Russell Jurney.

Collecting

Data Dump

Data Dump

- 0 effort data export, but **it is a hand-off**

Data Dump

- 0 effort data export, but **it is a hand-off**
- Clean rubbish?

Data Dump

- 0 effort data export, but **it is a hand-off**
- Clean rubbish?
- Fine for experimentation but not production data

Data Dump

- 0 effort data export, but **it is a hand-off**
- Clean rubbish?
- Fine for experimentation but not production data
- **Connect** to data instead

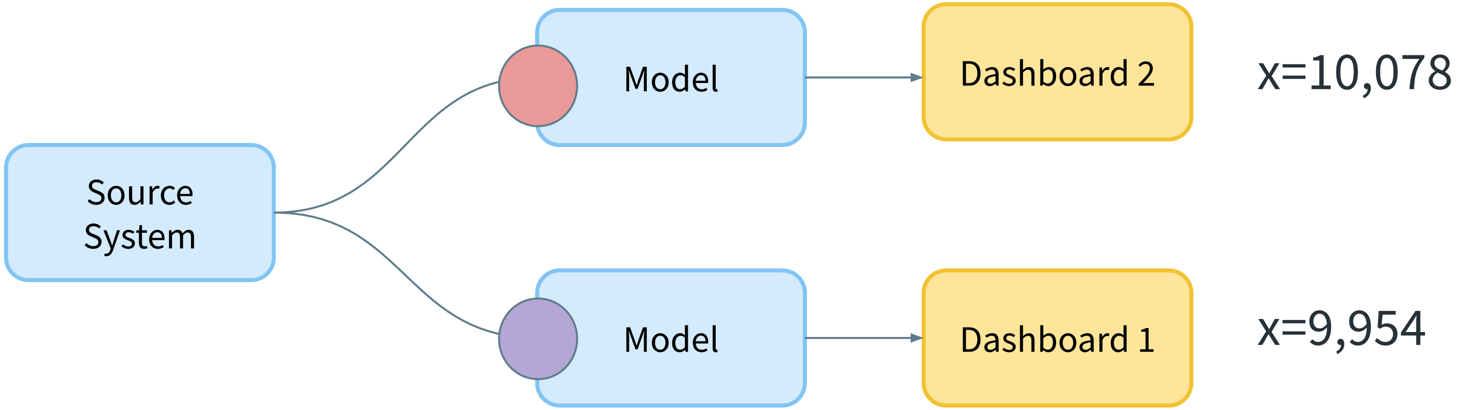
Challenge 2

Data Inconsistency

Cleaning

- Filtering
- Formatting
- Converting
- De-duplicating
- Removing outliers
- Correcting

Inconsistency



Challenge 3

Production hand-off

The Data science process is like scientists creating a vaccine in a lab, they ***hand-off*** *to manufacturing to scale up and deliver*


001

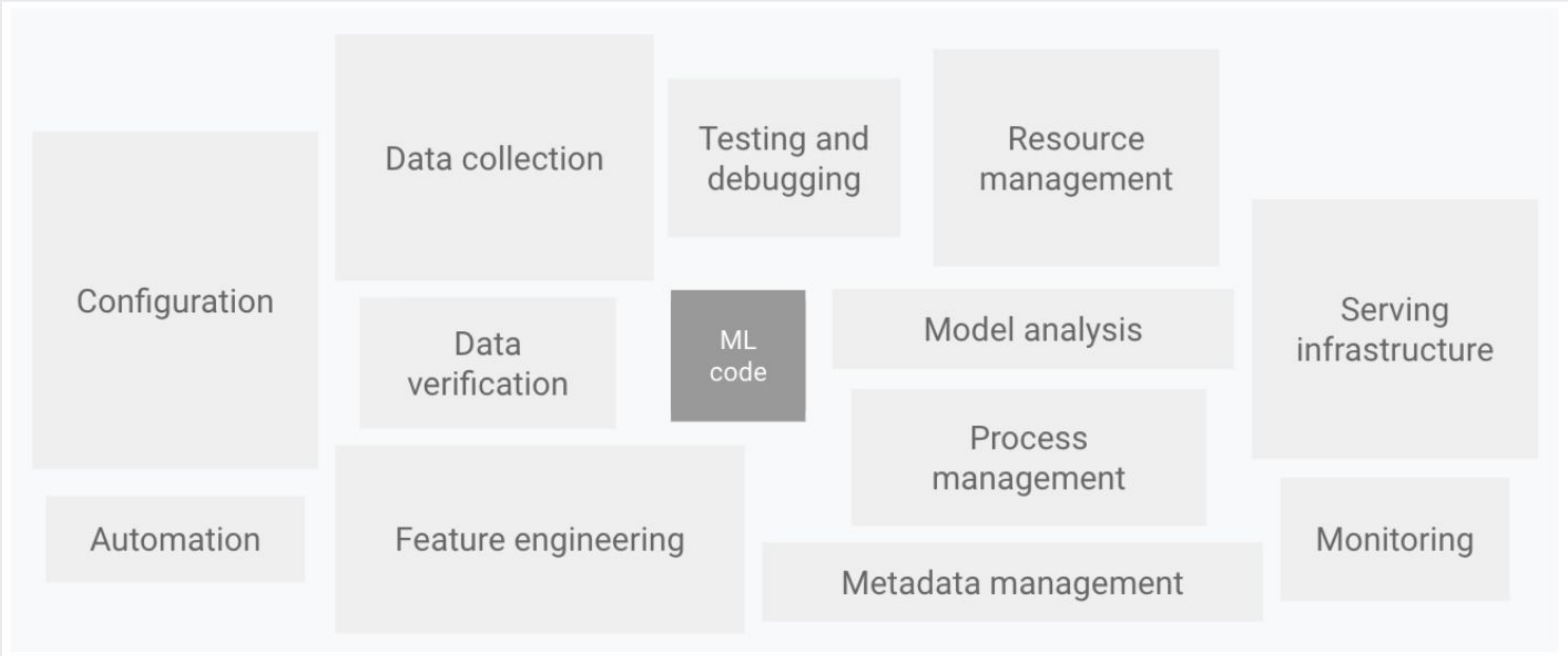
Triggers:

- Assumes software works as hardware

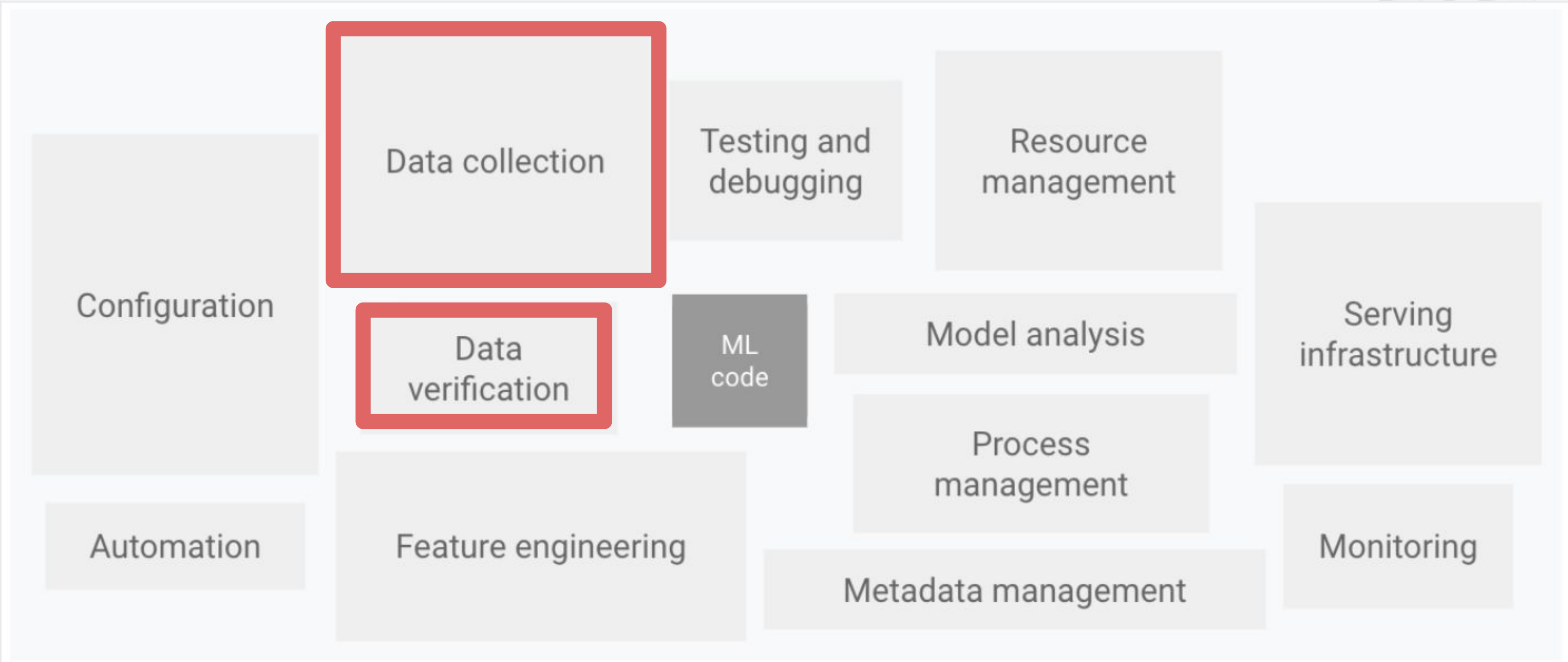
Triggers:

- Assumes software works as hardware
- Scaling ML models is as hard as developing them

Hidden technical debt in Machine learning systems



Hidden Technical Debt in Machine Learning Systems (D. Scully et al, 2015)

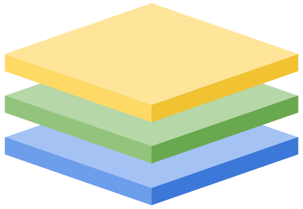

Data collection
Testing and debugging
Resource management
Configuration
Data verification
ML code
Model analysis
Serving infrastructure
Process management
Automation
Feature engineering
Metadata management
Monitoring

Production Hand off:

- Too soon in the development phase
- Not enough context
- Lots of data work still to be done

Challenges - Summary

- Hidden raw data hand-off
- Data inconsistency
- Production hand-off

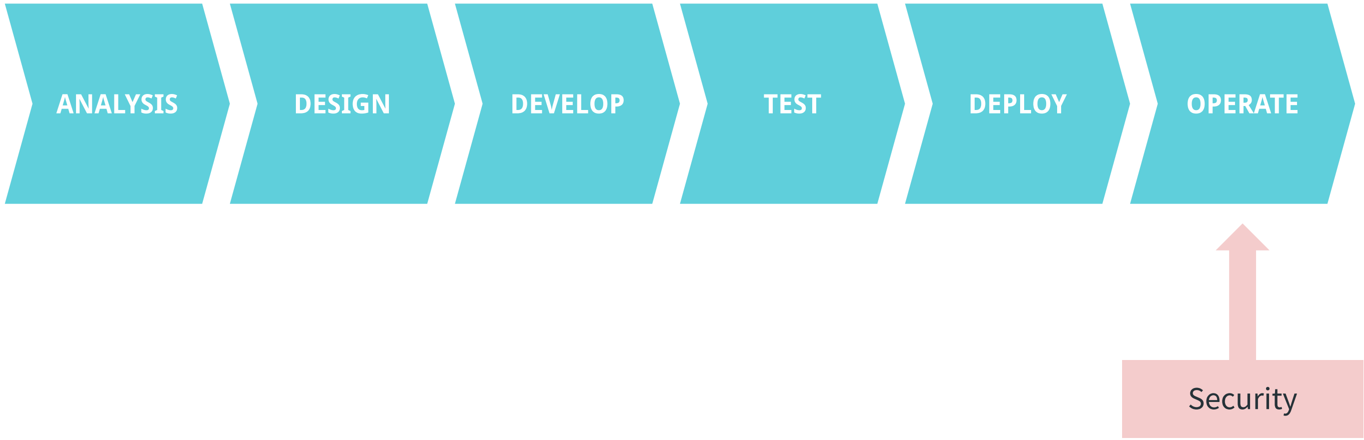


What might a software vaccine look like?

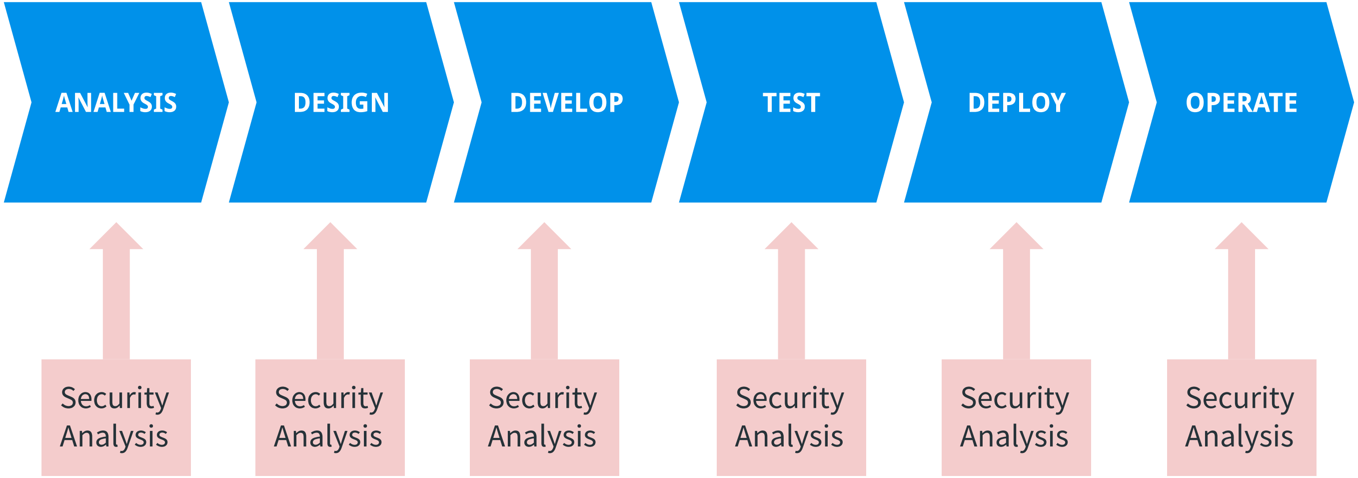
What might a software vaccine look like?

Protection against viruses?

Security V1



Dev-Sec-Ops


ANALYSIS
DESIGN
DEVELOP
TEST
DEPLOY
OPERATE
Security Analysis
Security Analysis
Security Analysis
Security Analysis
Security Analysis
Security Analysis

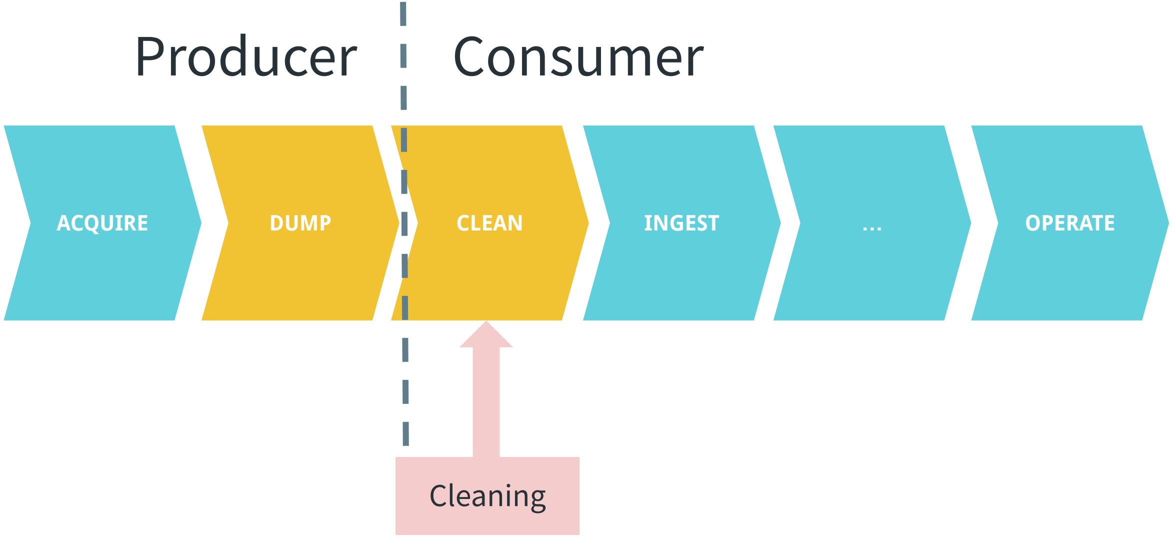
Shift-Left

implementing a process, or using a tool as early as possible in the development chain

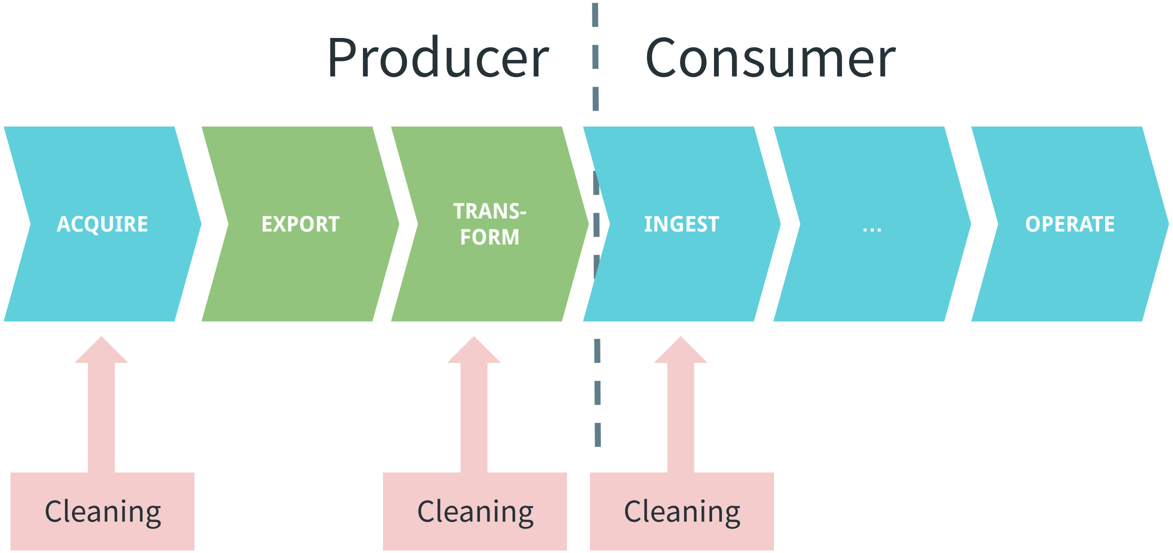
Shift Left

What does shift-left for data look like?

Data delivery



Data delivery



3.
Data as a Product

Data as a Product

Standalone data designed **for data consumers**

- Discoverable
- Well Described
- Interoperable
- Secure
- Trustworthy

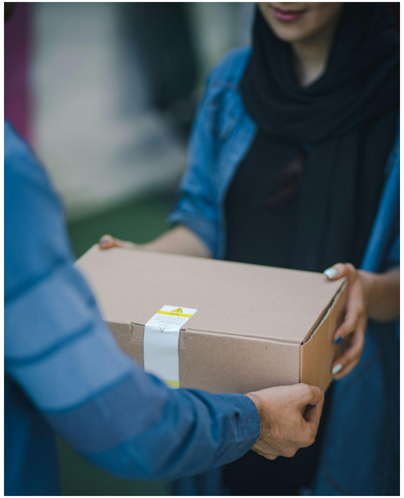
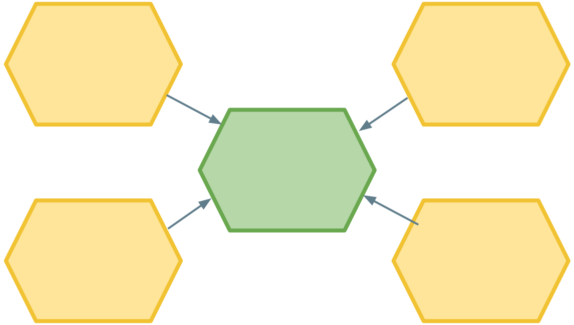


Photo by RoseBox رز باکس on Unsplash

Data as a Product

Published on an analytical data platform

- Owned appropriately
- Standards
- Data contracts
- Quality-observed
- Usage monitored



Data Products

The data-generating team* transforms data for use

Why? Because they:

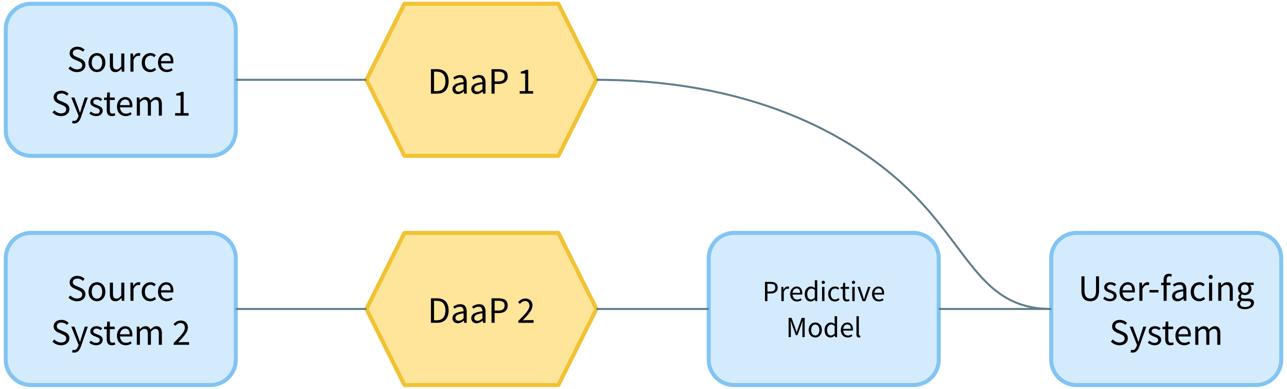
- Acquired the data
- Understand nuances
- Have existing processes
- Care how data is used elsewhere

Data Products - Network effects

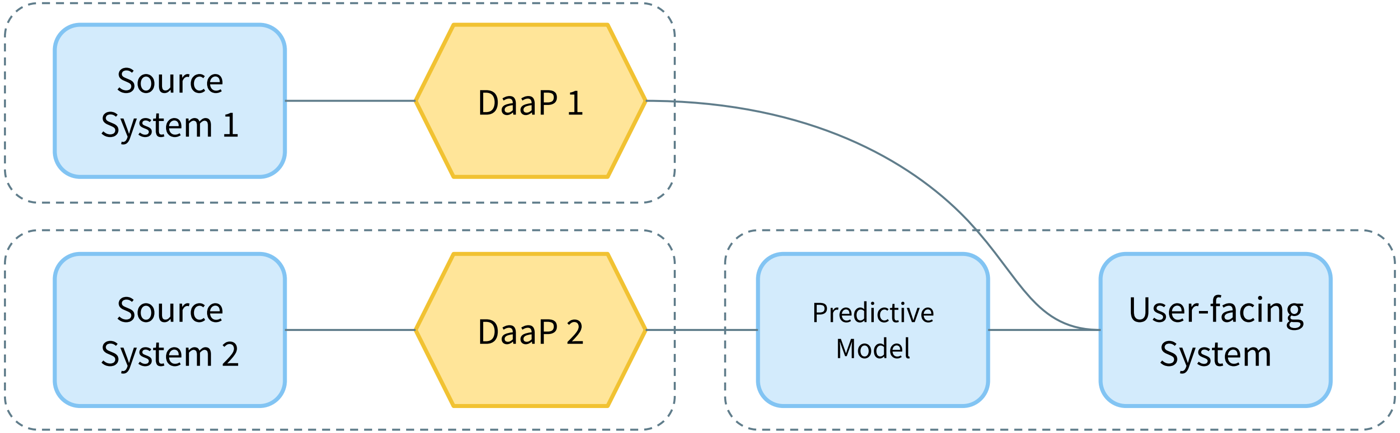
The more consumers;

- the better the data gets
- The more trustworthy the data becomes
- The more consumers use it
- Fewer duplicates exist
- More consumers there are

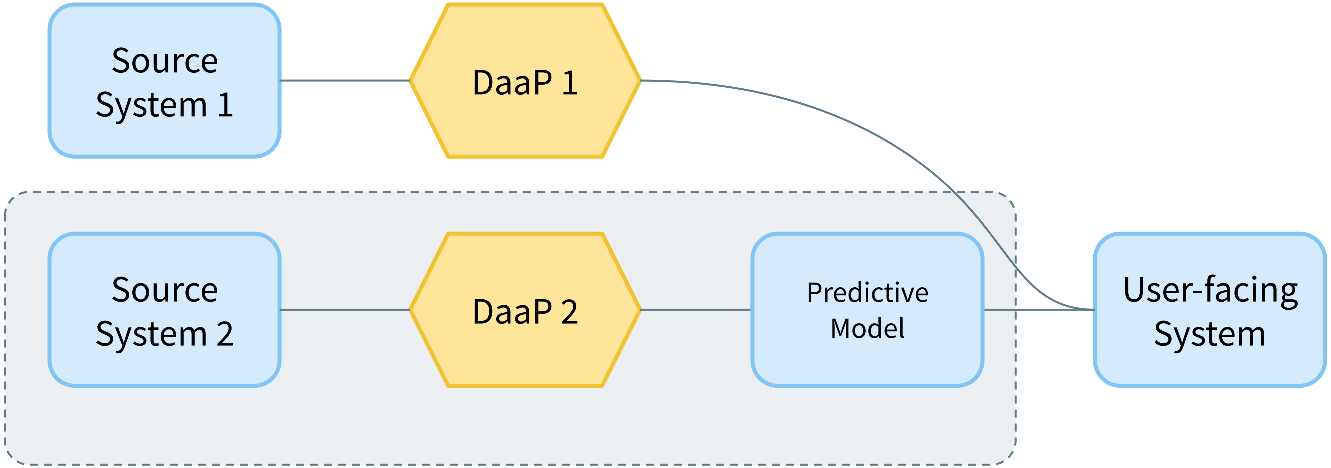
Data as a Product - Example



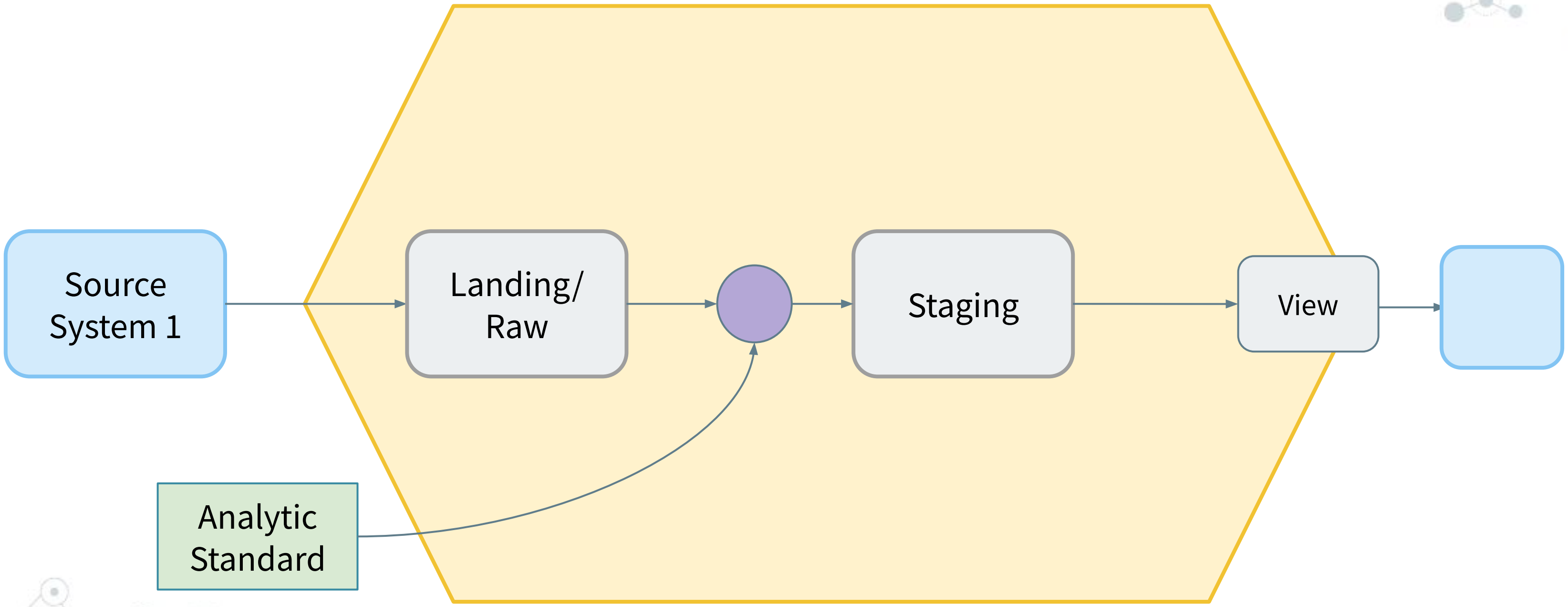
Data as a Product - Ownership Boundary



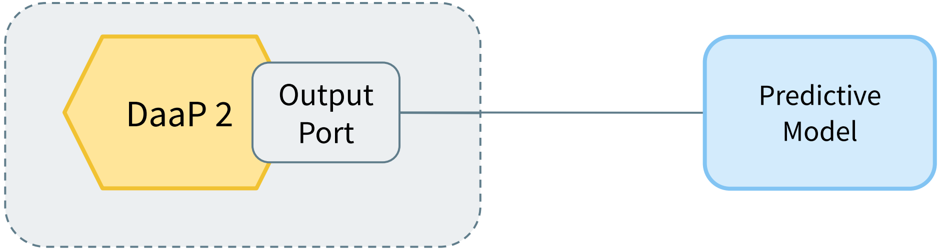
Data as a Product - Example



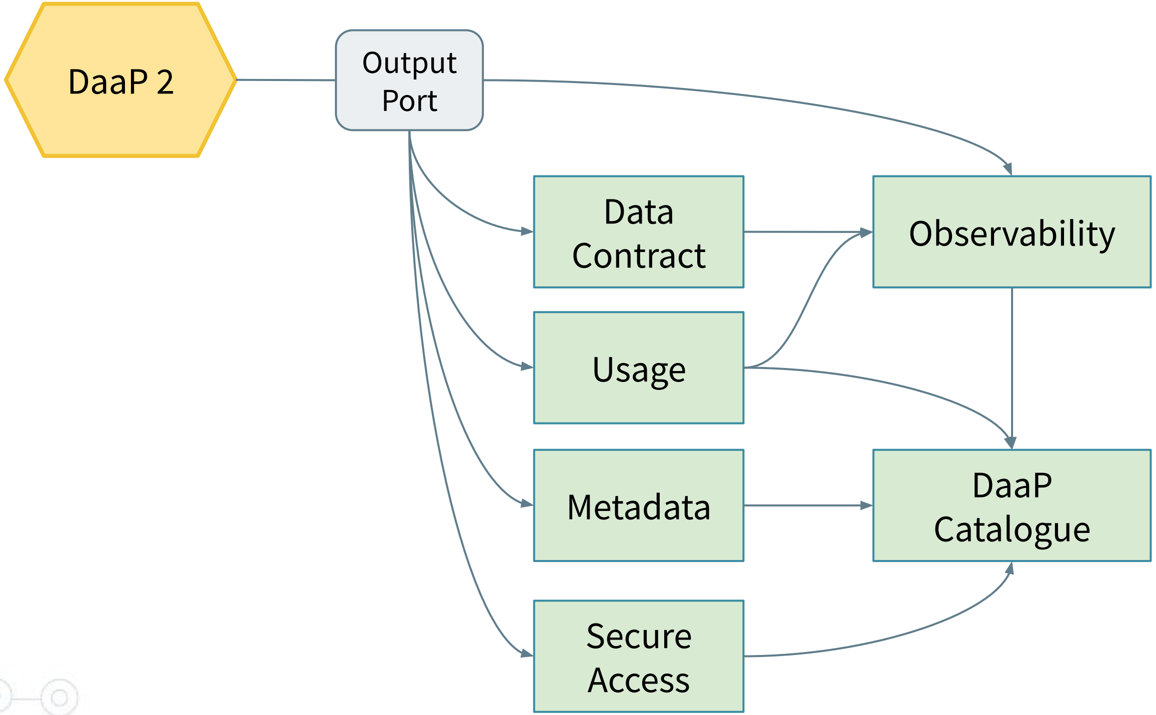
Data Products - Domain Specific



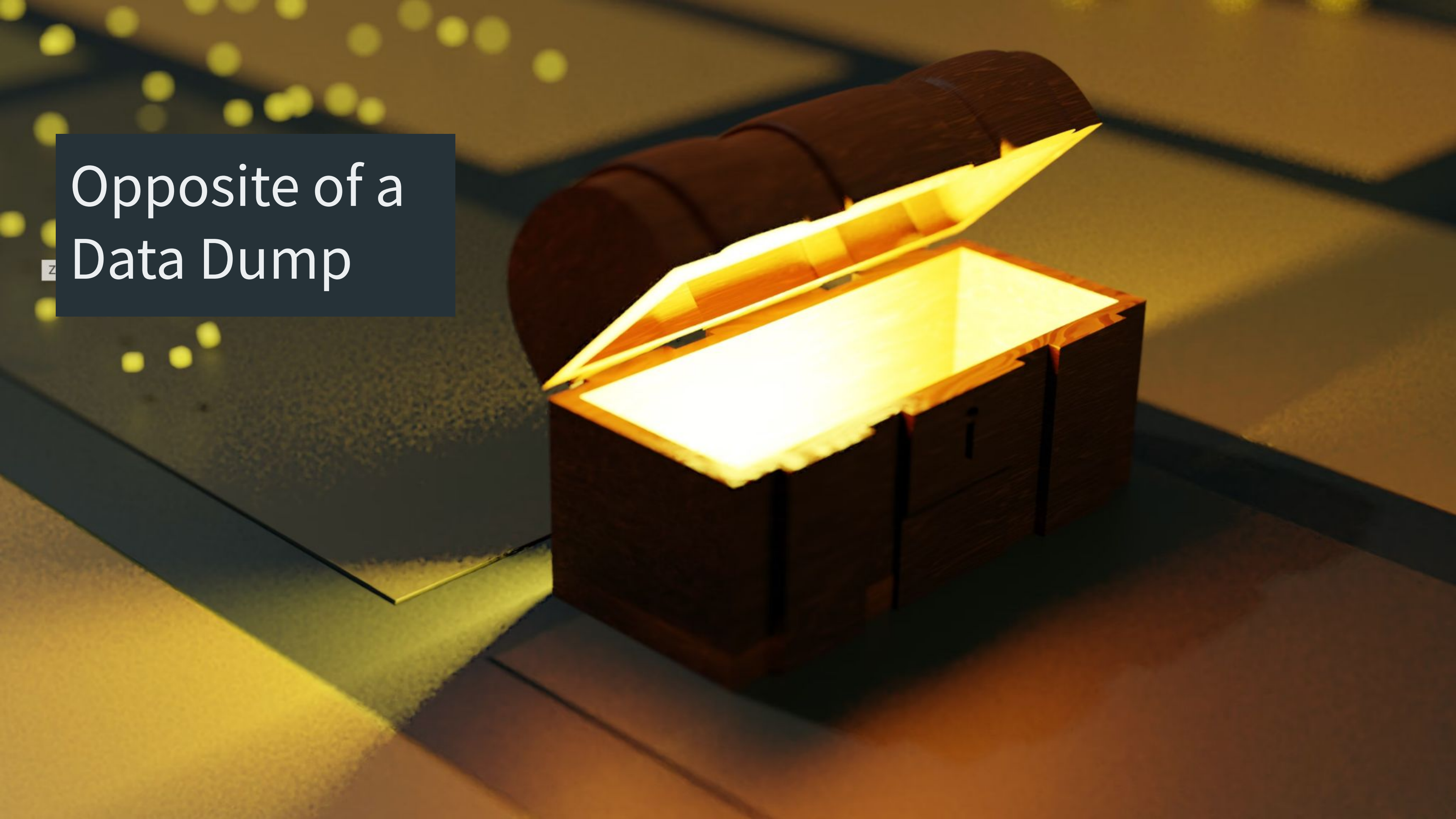
Data Products - Multi-faceted



Data Products - Multi-faceted



Opposite of a Data Dump



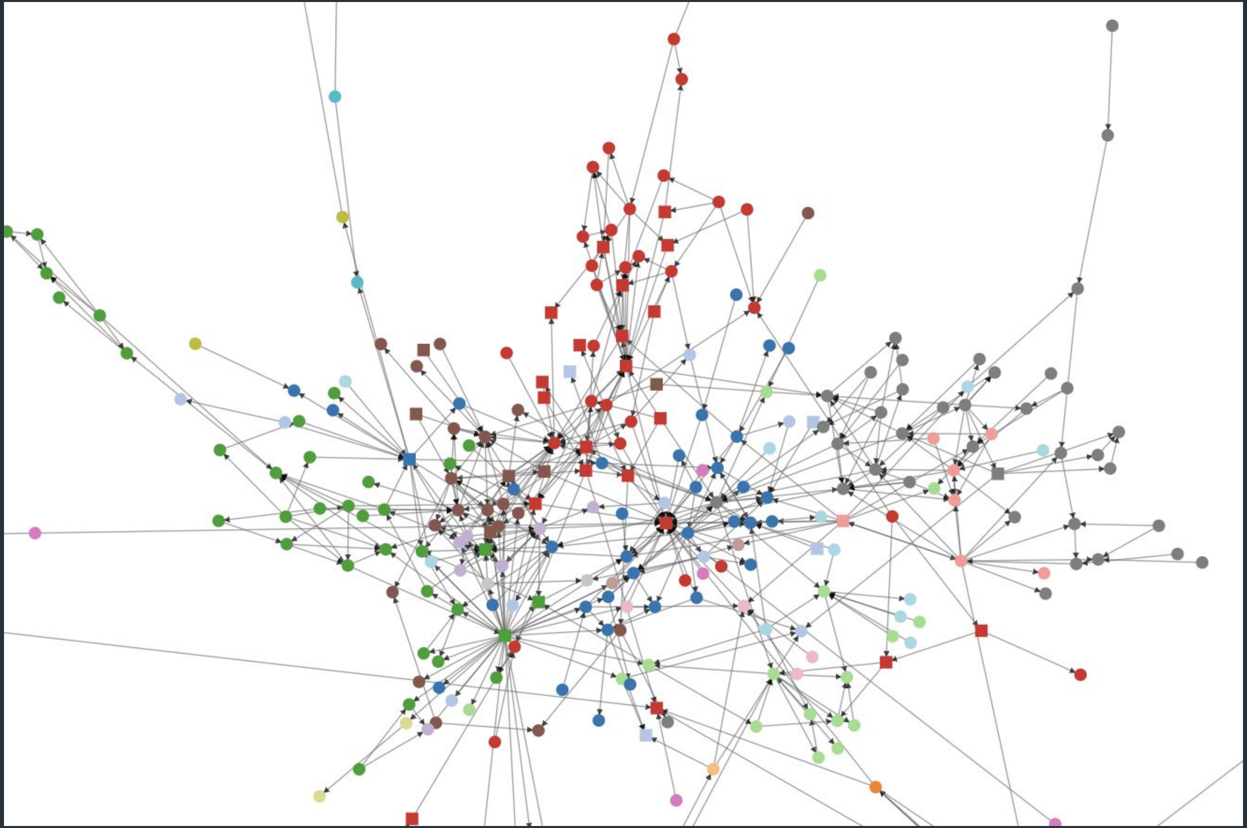
Incentives


YOU WANT ME
TO DO WHAT!?
memegenerator.net

Storytelling with data

- Data product development is abstract and not as visible as app development

- Need to craft our own narrative with real data to demonstrate

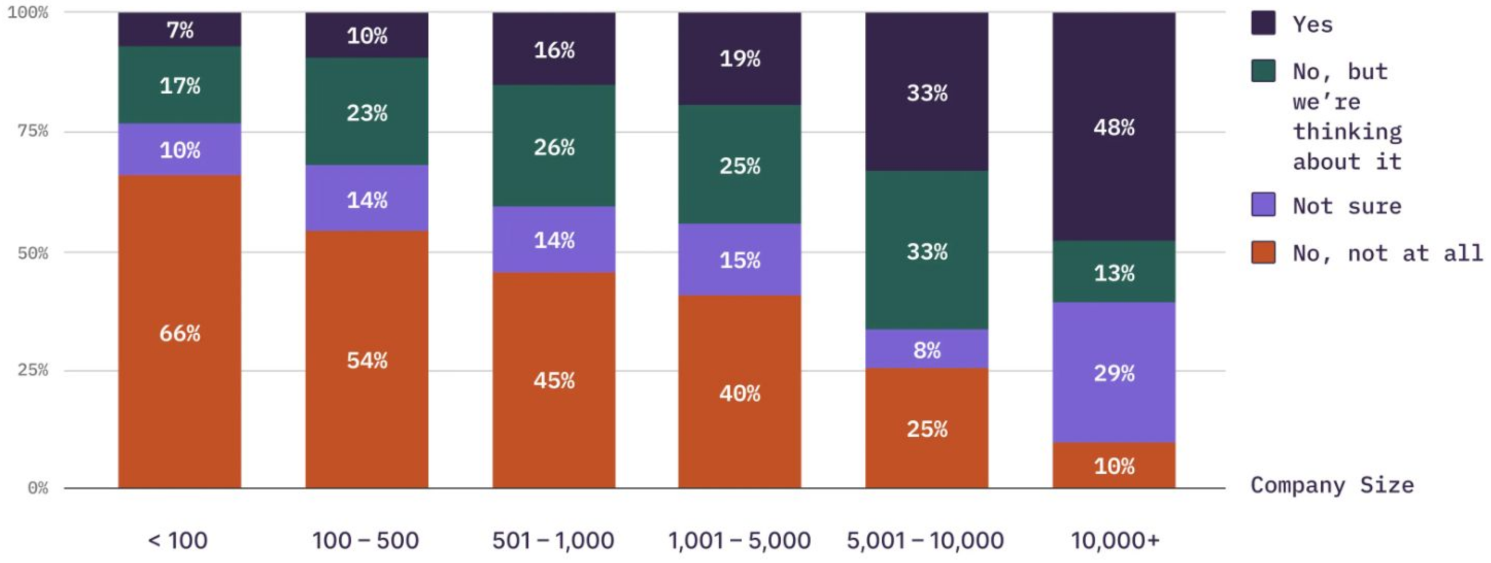


Teams

Data Team Topologies

- Product teams delivering DaaP
- DaaP teams
  - With Data scientists
- Domain data teams
  - Delivering DaaP on behalf of product teams

Who is moving to decentralized data architecture?



Data as a product - Summary

- Data as a product for reliable data connectivity
- Teams close to the source transform data
- Trust in data increases

4.

Delivery

Deliverables

- Notebooks
- Serialised ML models
- ML models as APIs

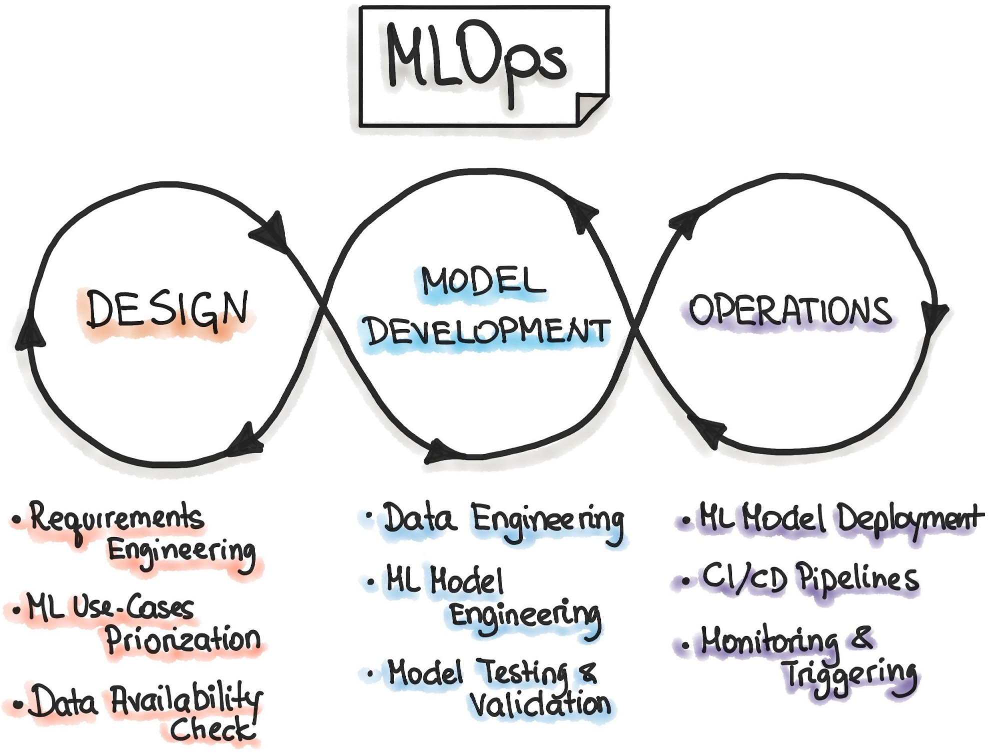
Deliverables

- Notebooks
- Serialised ML models
- ML models as APIs

Notebooks are for discovery and examples

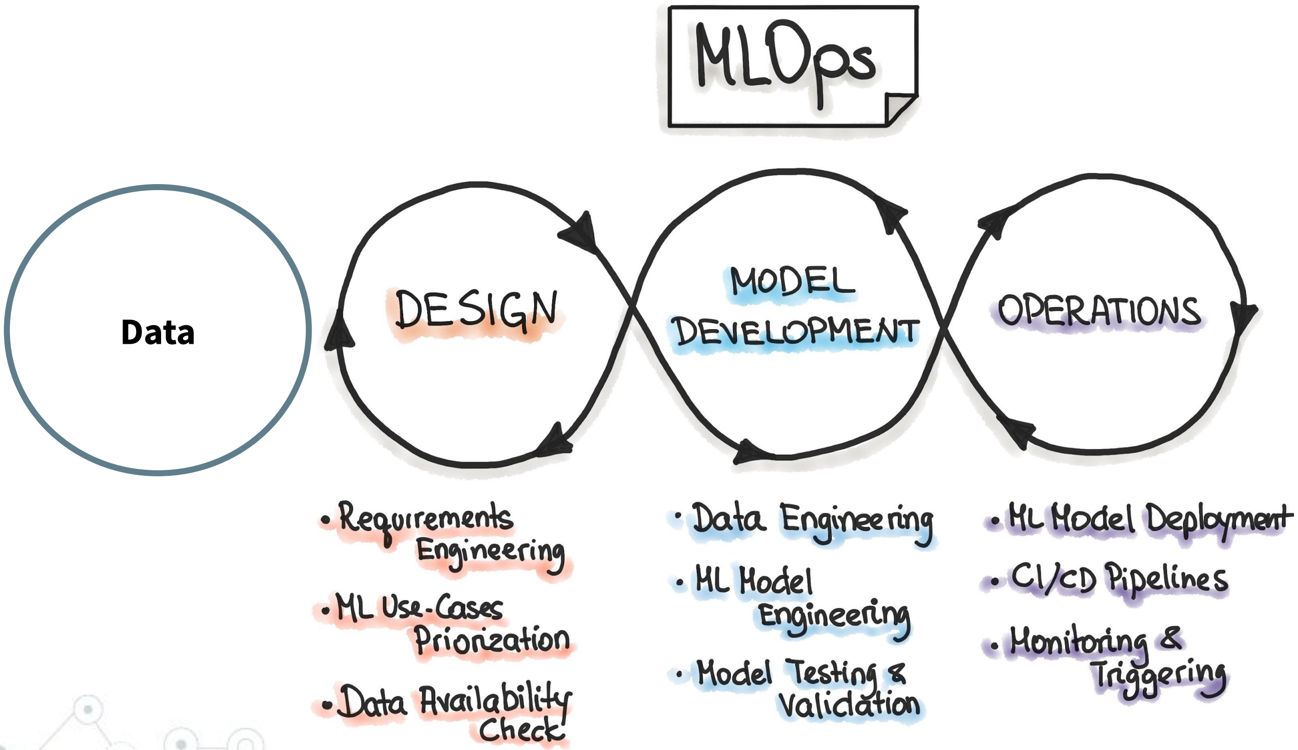
Model deployment needs to be part of an automatic, reproducible process

Deliverables

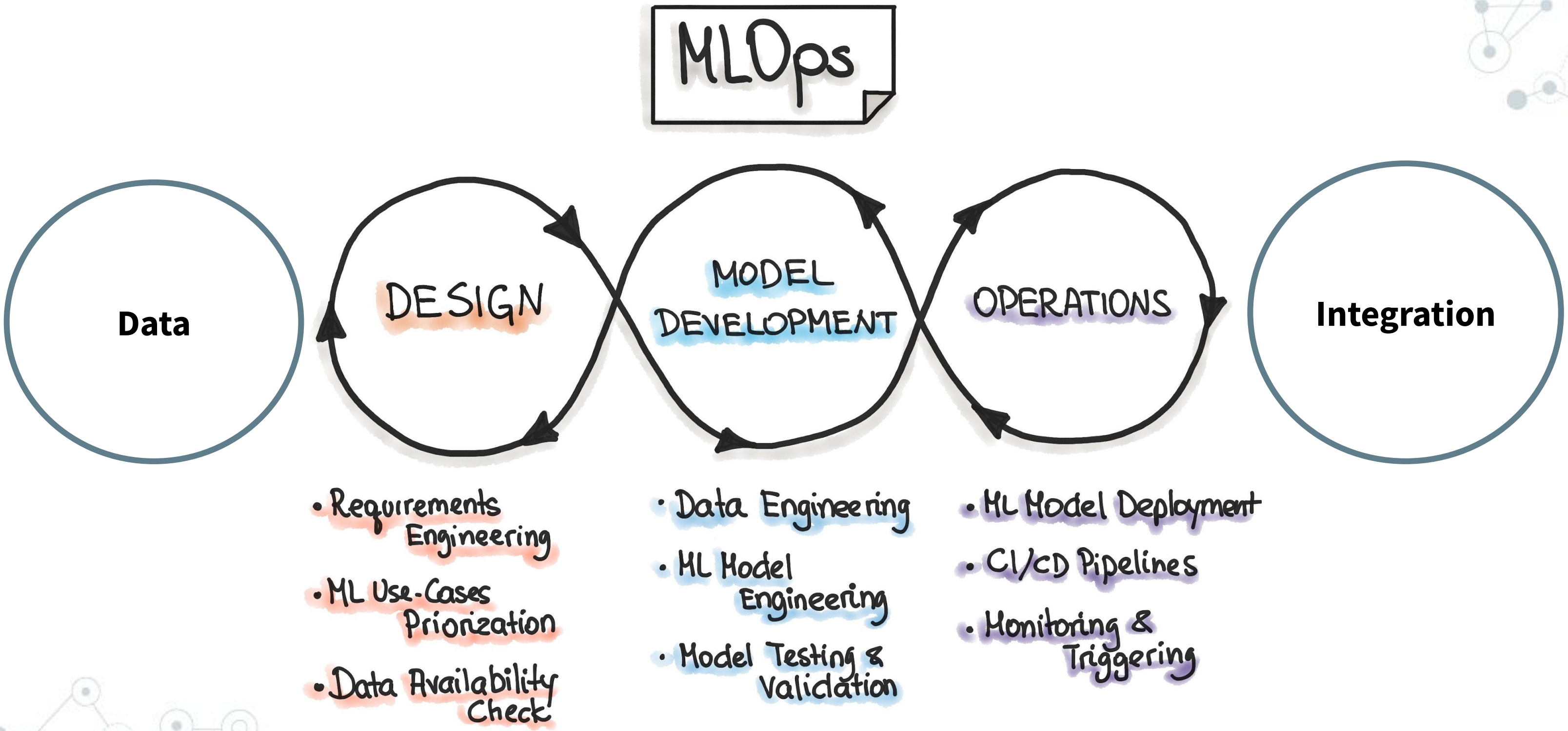


https://ml-ops.org/content/mlops-principles

Deliverables



Deliverables



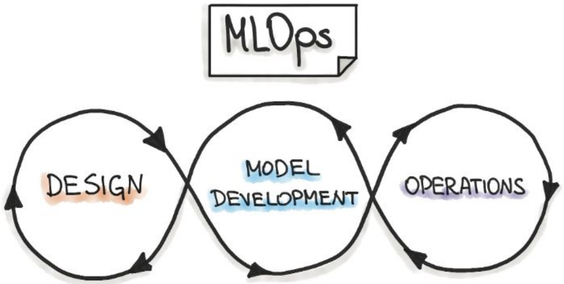
Deliverables

Challenging Hand-off

- Which operations environment?

- Unpredictable costs when scaling
  - PoC models may become unviable
  -

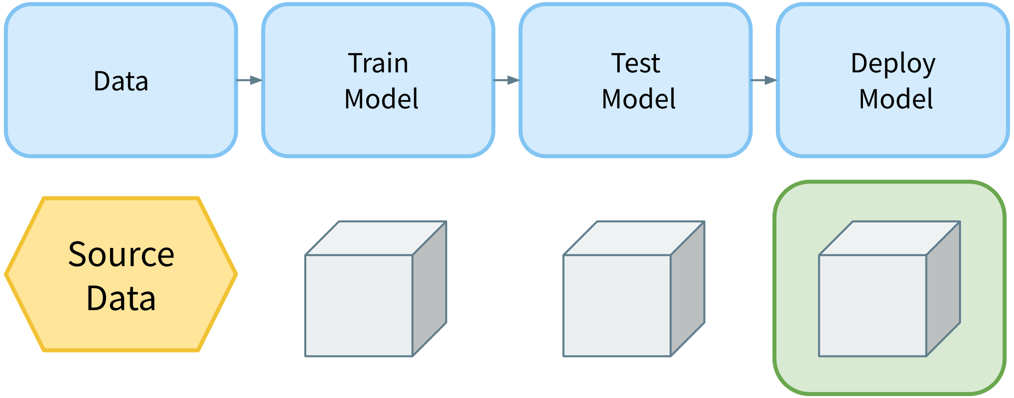
- Not the only option…



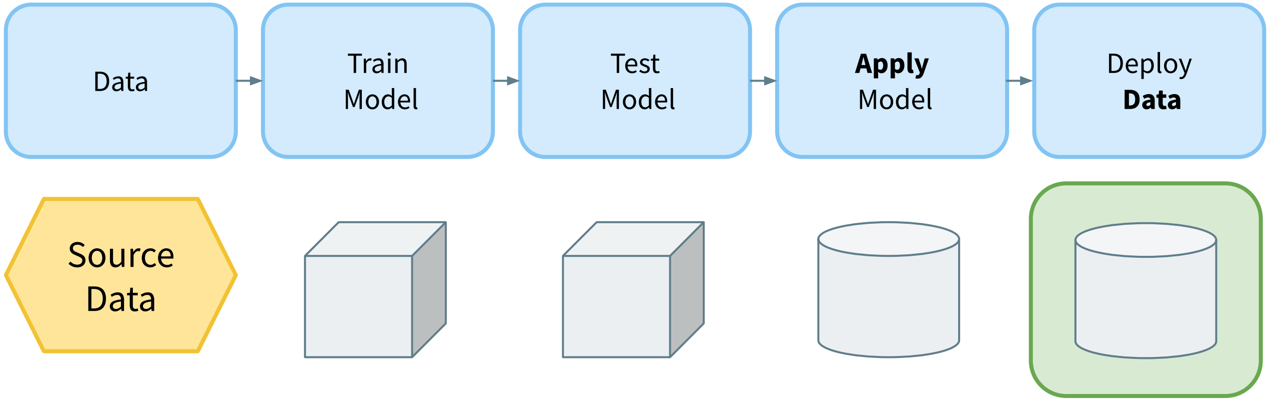
Data API principles

- Minimise work at request time; Maximise work at ingestion
  - Fast
  - Simple
  - Resilient
  - Reduced Compute
  - Failures handled in advance
- What can we pre-compute?

Model deployment



Model deployment



Delivery Summary

- Not every problem needs real-time ML
- Move compute to the left if possible

Takeaways

- Make each data hand-off part of the process and make each hand off fully visible to all teams involved
- Ensure key metrics are calculated once, as close to the data source as possible
- Treat data-as-a-product to shift data responsibilities to the most effective point in the data supply chain

Takeaways

- We can treat a delivery process like science experiments

Data Science and Agility

At Springer Nature

Charles Kubicek

Agile on the Beach 2024

@ckubicek
/in/ckubicek