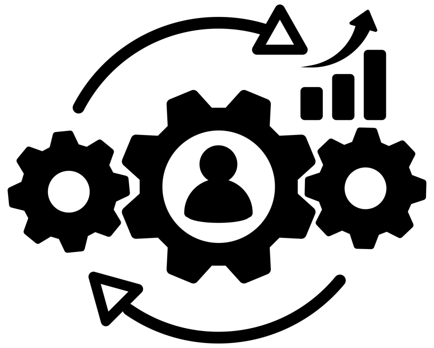
Data Intensive Applications

An Introduction

Tonight's Quest

What is a Data Intensive Application - Our definition?

Who Builds DIA?



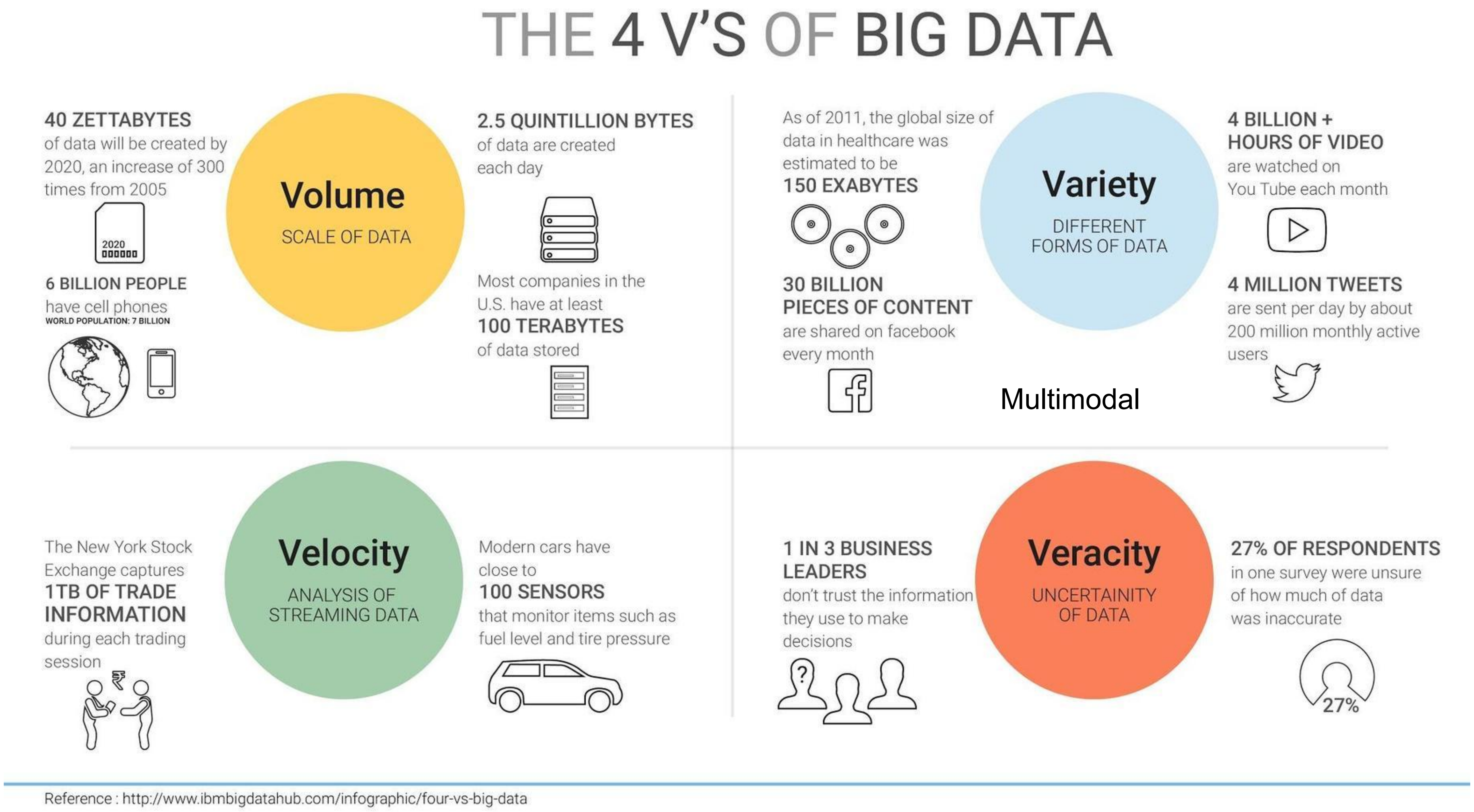
How are DIA Built?

What are some of the ***challenges*** in building DIA?

How do we ***classify*** and ***assess*** the value of a DIA?

Data intensive applications combine a large complement of data and compute to augment or replace human decision making for the benefit of optimizing important processes.

Data intensive applications combine a ***large*** complement of ***data*** and ***compute*** to ***augment*** or ***replace*** human **decision** making for the benefit of ***optimizing*** important ***processes***.

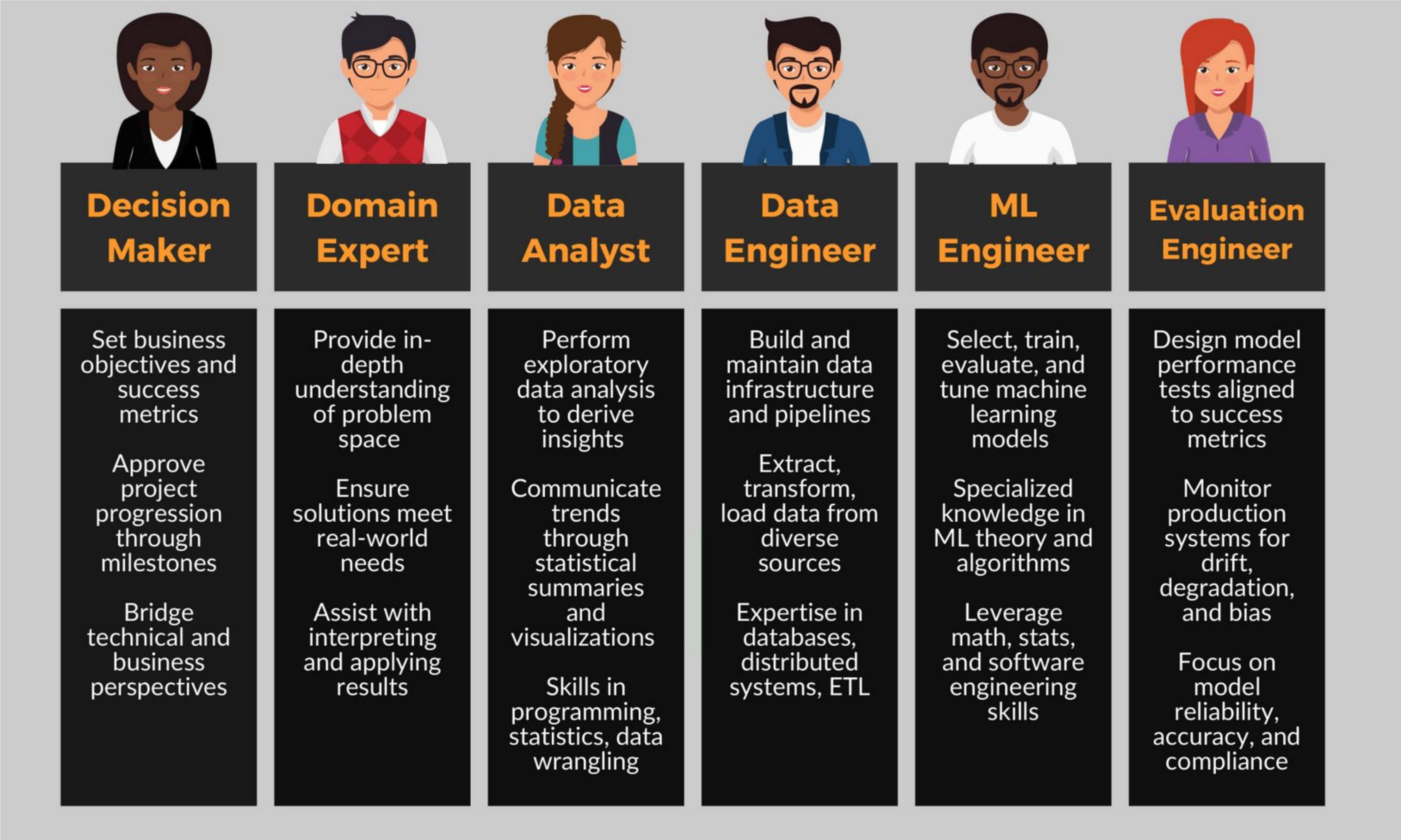


What do we mean by "large" complement of data?

Our work in this class is focused on "handling" these data challenges.

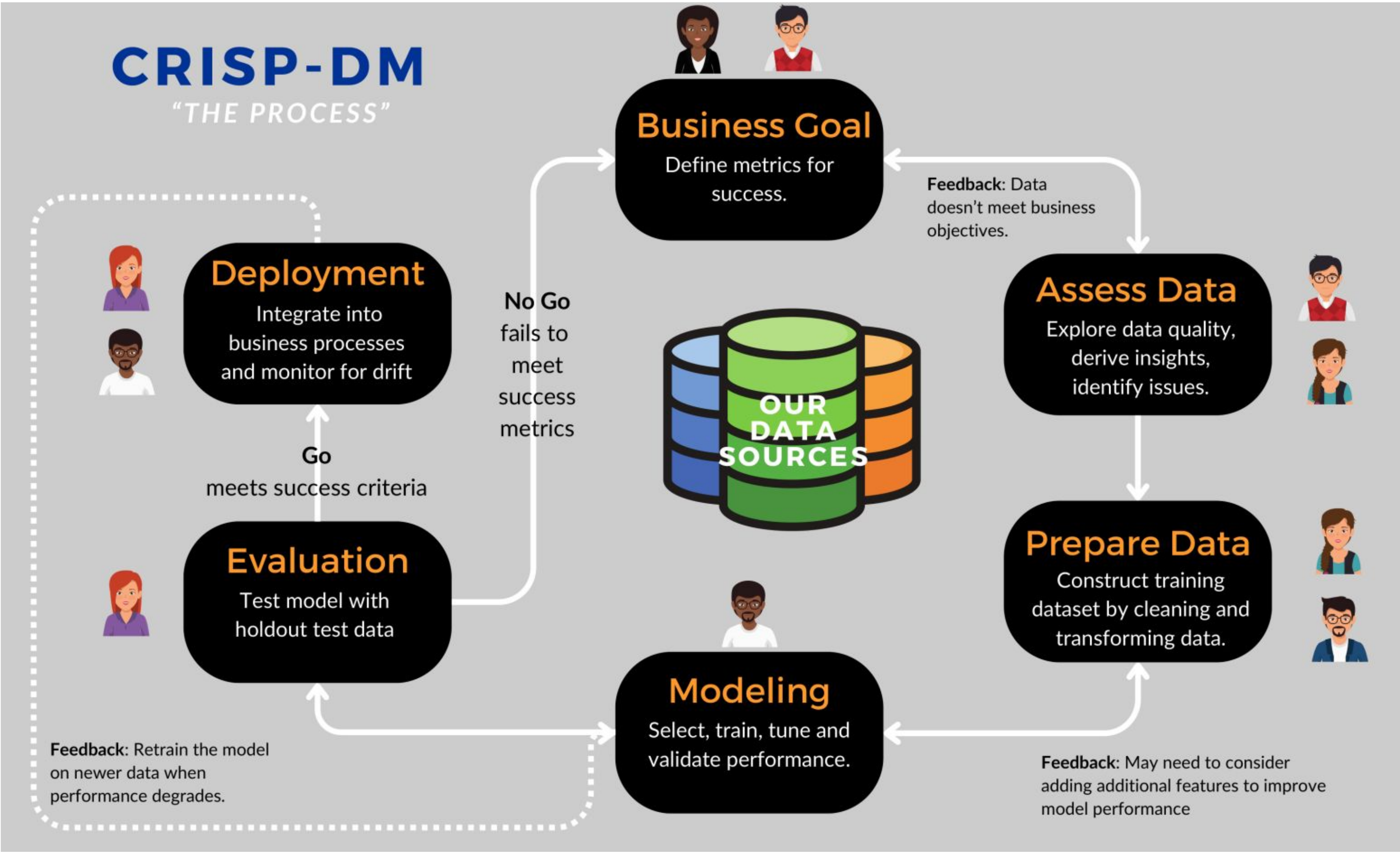
What do we mean by “large” complement of compute?

Building Data Intensive Applications is a team sport.



What we need in a process?

Structured Approach	Colaboration	Iteration	Predictability	Flexibiliy	Focus
Defines a clear, methodical sequence of steps to guide the end-to-end data value creation.	Clearly defines roles & responsibility to coordinate work across teams with diverse skills.	Allows folding back to previous steps when needed to refine the solution.	Following consistent standard methodology improves predictability.	The methodology can be tailored to fit different project needs and constraints.	Business goals anchor the process to ensure work stays aligned with delivering value.


CRISP-DM
"THE PROCESS"
Business Goal
Define metrics for success.
Feedback: Data doesn't meet business objectives.
Assess Data
Explore data quality, derive insights, identify issues.
Prepare Data
Construct training dataset by cleaning and transforming data.
Feedback: May need to consider adding additional features to improve model performance
Modeling
Select, train, tune and validate performance.
Evaluation
Test model with holdout test data
Go
meets success criteria
Deployment
Integrate into business processes and monitor for drift
No Go
fails to meet success metrics
OUR DATA SOURCES
Feedback: Retrain the model on newer data when performance degrades.

10 Ways Your Data Project Might Fail

- Lack of Clear **Objectives**
- Insufficient **Data Quality**
- Poor **Data Management**
- Inadequate **Skills** & **Expertise**
- Ignoring **Business Context**
- **Overcomplicating** Solutions
- Poor **Communication** & **Collaboration**
- Ignoring **Ethical Considerations**
- Lack of **Scalability**
- Insufficient **Monitoring** & **Maintenance**

Source: Adapted from Martin Goodson, "Ten Ways Your Data Project Is Going To Fail"

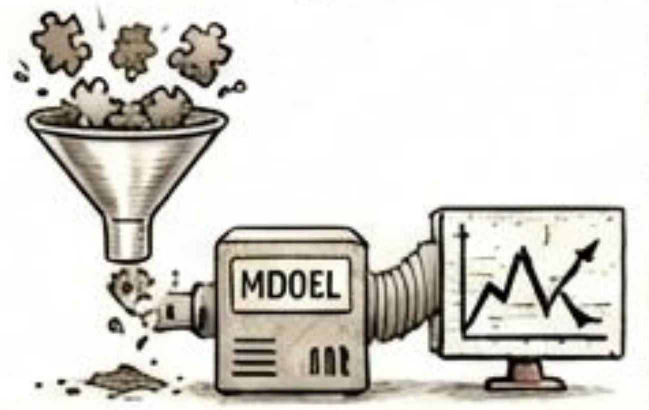
TEN WAYS YOUR DATA PROJECT IS GOING TO FAIL

Source: Martin Goodson's Analysis. *Pitfalls for Data Science Practitioners*.

1 LACK OF CLEAR OBJECTIVES

Example: Nonprofit aims to 'increase donor engagement" without defining metrics (frequency vs. size vs. retenlion).
Result: Generic reports, dissatisfied stakeholders.

2 INSUFFICIENT DATA QUALITY



Example; Healthcare predictive model for at-risk patients relies on inconsistent EHRx (misxing fields, errors).
Result: Unreliable predictions, ineffective project.

3 POOR DATA MANAGEMENT

Example: Media company has terabytes of unstructured, siloed customer data without governance.
Result: Impossible to extract insights for recommendation system.

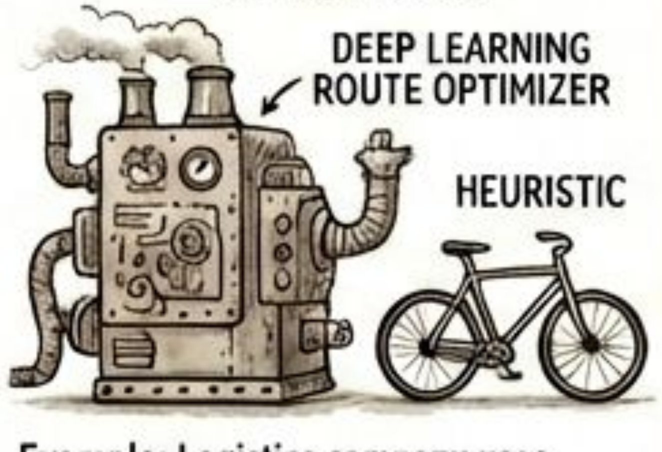
4 INADEQUATE SKILLS & EXPERTISE

Example: Retail chain hires data scientist without domain espertise.
Result: Complex, impractical modal incompatible with inventory workflows.

5 IGNORING BUSINESS CONTEXT

Example: Bank's accurate fraud model ignores manual approval process.
Result: Overwhelming faise positives, delayed transections, frustrated staff.

6 OVERCOMPLICATING SOLUTIONS

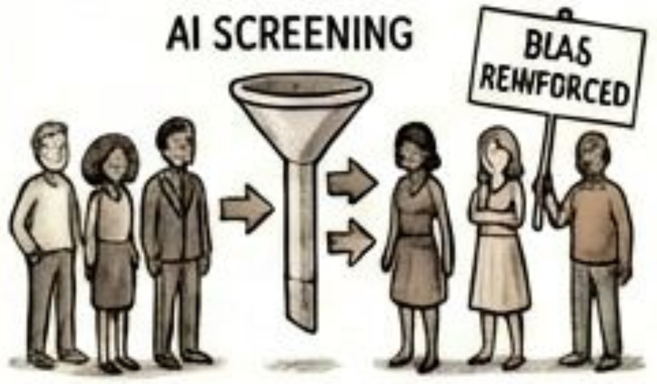


Example: Logistics company uses complex deep learning when simpie heuristic could achieve 9Q% benefit.
Result: unsustainable resource & expertise requirements.

7 POOR COMMUNICATION & COLLABORATION

Example: E-commerce customer segmentation model built without marketing input.
Result: Model too technical, misaligned with campaigns, underutilized.

8 IGNORING ETHICAL CONSIDERATIONS

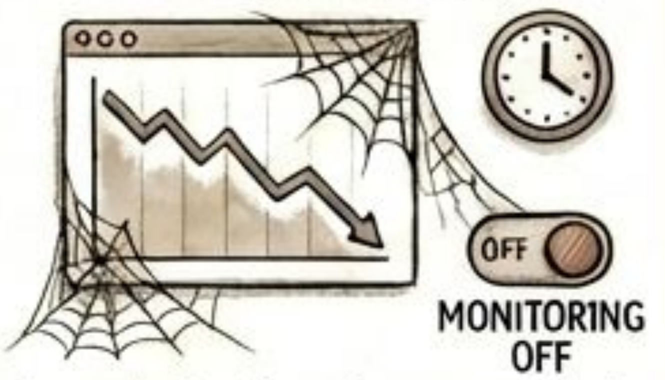


Example: Recruitment AI uses historical hiring data, reinforcing bias against underrepresented grouos.
Result: Public backlash, reputational damage.

9 LACK OF SCALABILITY

Esample: Startup's personalized email system on a simple database fails as customer base grows.
Result: Delays, failures, frustrated customers.

10 INSUFFICIENT MONITORING & MAINTENANCE



Example: Social media recommendation engine deployed without performance monitoring.
Result: Relevance drift due to behavior changes, decreased engagement.

Course Note: Success requires balancing technical rigor with clear goals, domain context, and ongoing stewardship. Avoid these traps!

CRISP-DM *covers a lot of ground, but must be complemented by strong communication, ethical considerations, and scalability planning.*

CRISP-DM Stage	**What Pitfall does it help mitigate?**
Business Goal	1: Clear Objectives & ROI - Defines success metrics upfront.
Assess Data	2: Data Quality - Identifies and addresses data issues early.
Prepare Data	2: Data Quality - Cleans & transforms data for reliability. 3: Data Management - Supports documentation and governance.
Modeling	4: Expertise - Emphasizes model validation & performance. 6: Overcomplicating - Promotes pragmatic solutions.
Evaluation	10: Monitoring - Tests model with holdout data prior to deployment.
Deployment	3: Data Management - Supports integration. 5: Context - Integrates into business processes. 10: Monitoring - Provides framework for monitoring after integration.

DATA + COMPUTE

necessary / not sufficient

MODELING + OPTIMIZATION

deliver the value

<table>
<tr><td colspan="2" rowspan="2"></td><th colspan="2">Human Decision Making</th></tr>
<tr><td>Augment - Narrow</td><td>Replace - General</td></tr>
<tr><th rowspan="2">Optimizing For …</th><td>More</td><td>Skin lesion Classification > Correct Diagnosis</td><td>Trading Bots > profits</td></tr>
<tr><td>Less</td><td>Delivery Route Planning < Fuel use</td><td>Self driving car < Accidents</td></tr>
</table>

Key Reason DIA are important: Data 2 Optimal Decisions

		Decision Frequency	
		High	Low
Decision Impact	High	**Prime Value Target** E.g. medical diagnosis and autonomous vehicles	**Mixed Bag**
	Low	**May be worth the investment** E.g. route planning and recommendations	**Forget It!** Too Costly

Key Reason DIA are important: Data 2 Optimal Decisions

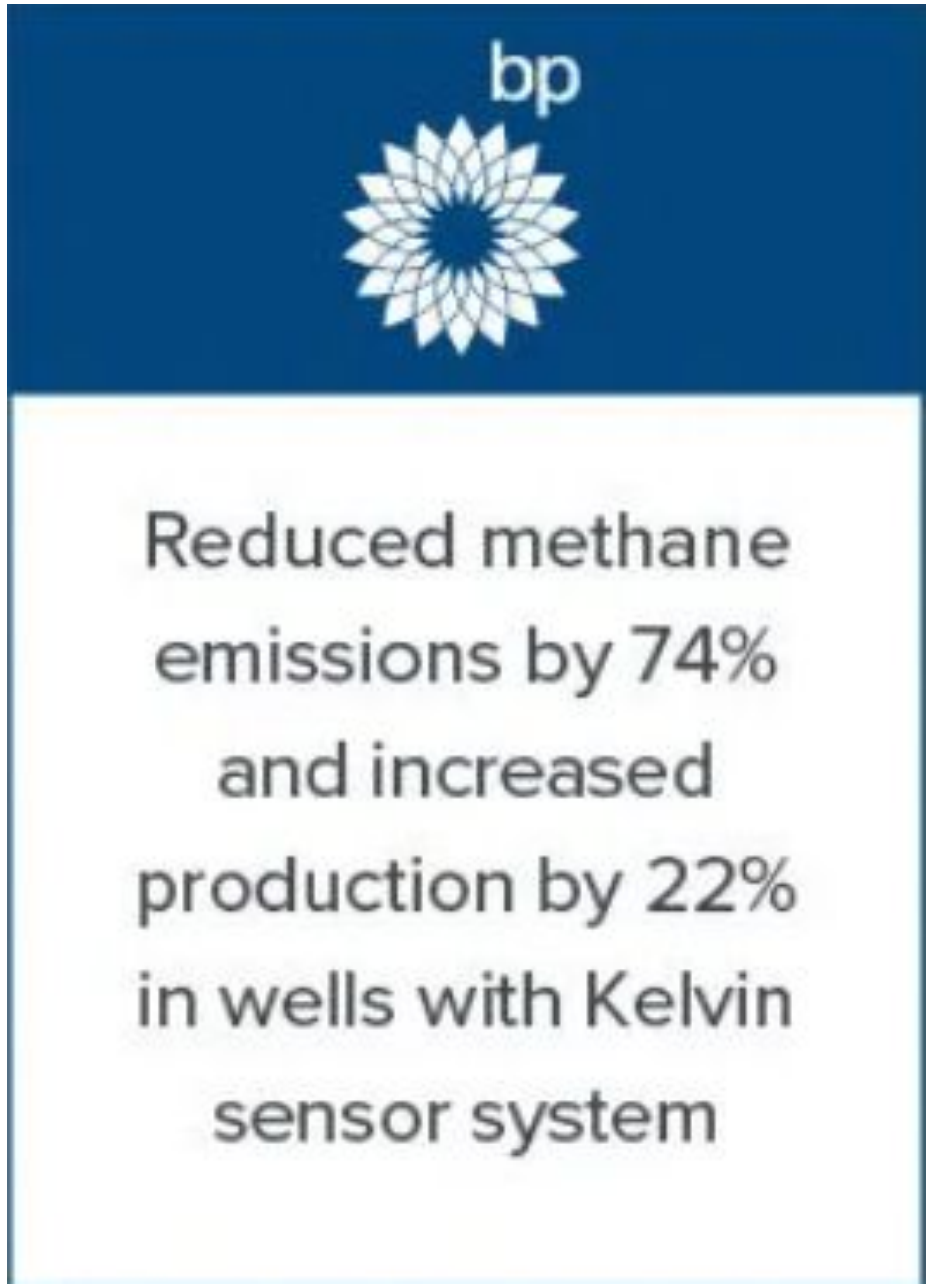
		Decision Making	
		Augment	Replace
Optimizing For …	More		
	Less	**X** Routes - Fuel	

		Decision Frequency	
		High	Low
Decision Impact	High		
	Low	**X** aggregation	

Augmenting
Planners and Drivers frequent decision process (every day). Low impact on a given delivery although aggregation across the high number of deliveries.

Outcome
Use Less fuel (KPI) by suggesting routes for each days deliveries.

Example: [UPS Route Optimization]



		Decision Making	
		Augment	Replace
Optimizing For …	More	X Oil Produced	
	Less	X GHG	

		Decision Frequency	
		High	Low
Decision Impact	High	X	
	Low		

Augmenting

Oil field operators with "digital twin" simulations leading to optimizations of operational parameters (flow and pressure) across the large scale operations.

Outcome

Less GHG and higher production. (KPI)

Example: BP's New Oilfield Roughneck Is An Algorithm airbnb

Increased booking conversion rate by ~4% by modeling likelihood of host acceptance

		Decision Making	
		Augment	Replace
Optimizing For …	More		X (rec)
	Less		

		Decision Frequency	
		High	Low
Decision Impact	High		
	Low	X (agg)	

Outcome
More booking revenue (KPI) by recommending listings that are likely to accept the renter. Beware of ethical considerations of algorithmic bias!

Replace
Replacing travel agent. Low impact on a given listing although aggregation across the high number of listings.

Example: AirBnB Recommendation - increase booking revenue

		Decision Making	
		Augment	Replace
Optimizing For …	More		**X** (rec)
	Less		

		Decision Frequency	
		High	Low
Decision Impact	High		
	Low	**X** (agg)	

Replacing

Search results based on user entered criteria. Low impact on a given search although aggregation across the high number of searches.

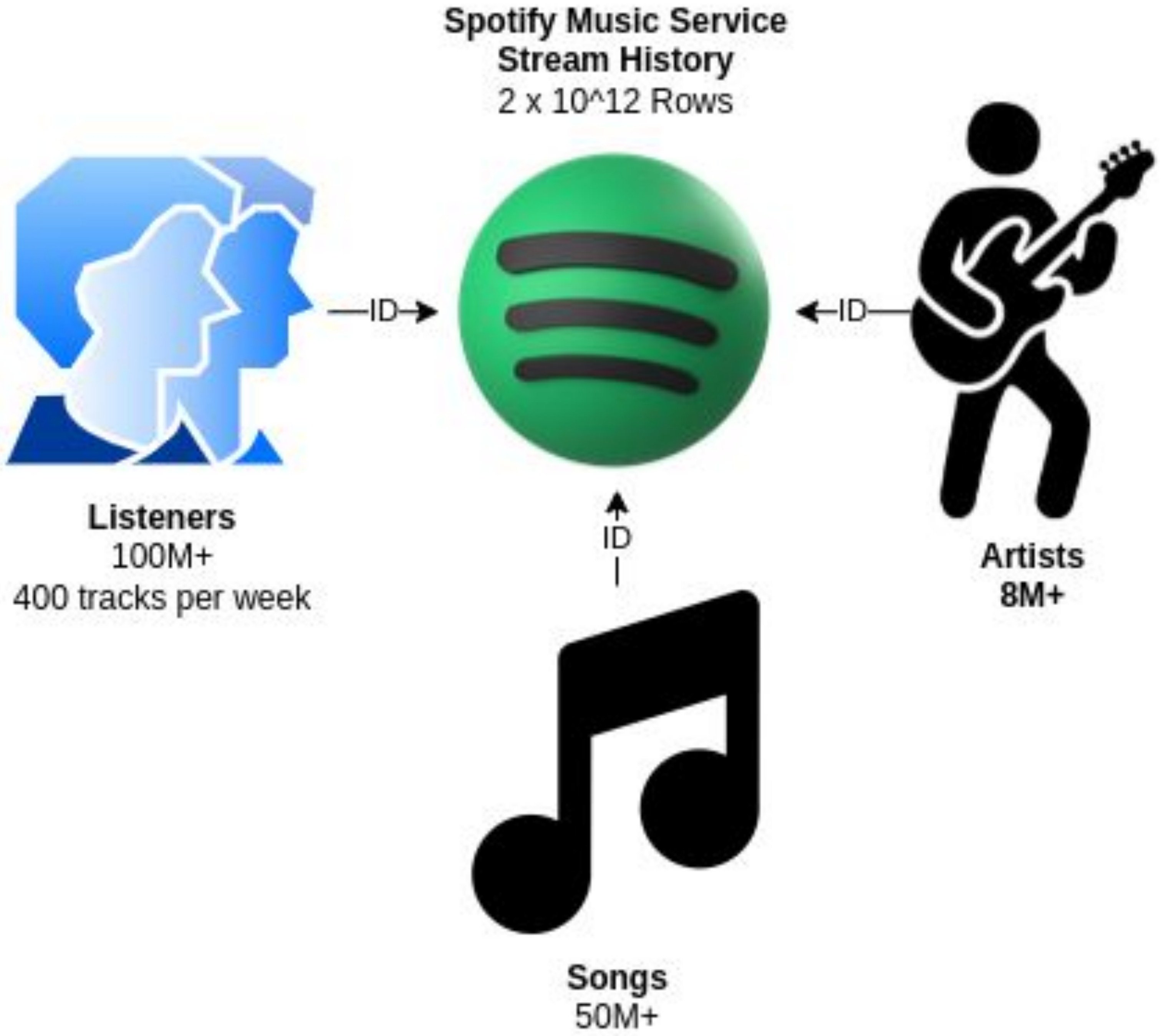
Outcome

Increase engagement - watch time (KPI) by recommending movies.

Example: [Netflix Recommendation - 80% of views]

A DATA & COMPUTE STORY

How Spotify ran the largest Google Dataflow job ever for Wrapped 2019




```
SELECT
    L.UserName,
    A.ArtistName,
    S.Track,
    COUNT(1) as ListenCount
FROM Stream_History SH
JOIN Artists A on A.ID = SH.AID
JOIN Songs S on S.ID = SH.SID
JOIN Listeners L on L.ID = SH.LID
GROUP BY L.UserName, A.ArtistName, S.Track
ORDER BY ListenCount DESC
```

Spotify Wrapped is a data-intensive application designed to:

		Decision Making	
		Augment	Replace
Optimizing For …	More	X (engagement)	
	Less		

		Decision Frequency	
		High	Low
Decision Impact	High		X (agg)
	Low		

<u>Outcome</u>

Strengthen emotional connections to the platform.

Drive business value through engagement, retention, and brand promotion.

<u>Augement</u>

Increases the users self-awareness of their listening habits Analytics for the people.

That's a Wrap!

Get going on your Labs!