Data-Intensive Distributed Computing
CS 431 (Fall 2026)



UNIVERSITY OF WATERLOO

The Data Flywheel
(v1.00)

Week 1: September 10, 2026

Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo

These slides are available at `https://lintool.github.io/cs431-2026f/`



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International
See https://creativecommons.org/licenses/by-nc-sa/4.0/ for details

Who am I?





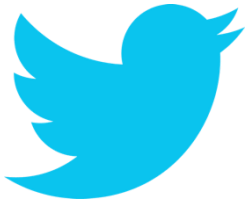


Massachusetts
Institute of
Technology



yupp

primal

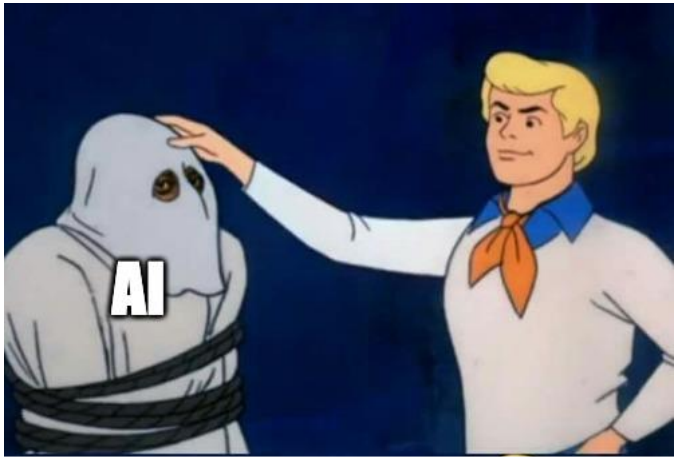


What's this course about?



What does it mean to an AI-first company?

What does it mean to a data-driven company?


AI



When people talk about AI these days,
they really mean ML…

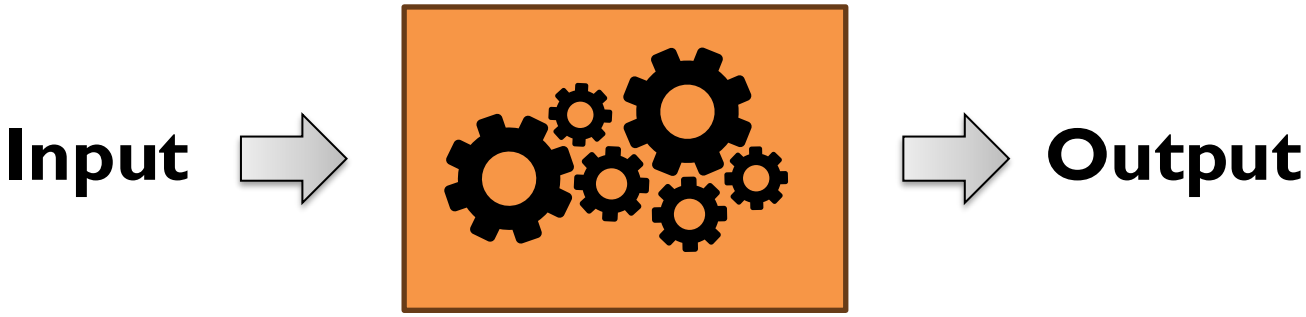

ML



When people talk about ML these days, they really mean supervised ML…



What's supervised machine learning?

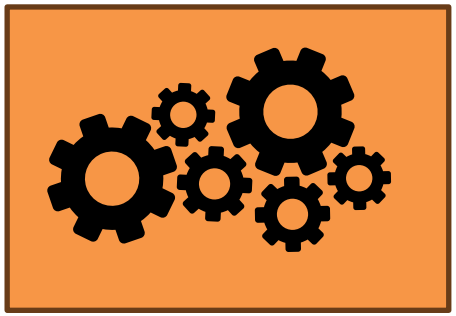


(accomplishes *some* task)
(hopefully *well*)



Input





Model

Output

owl

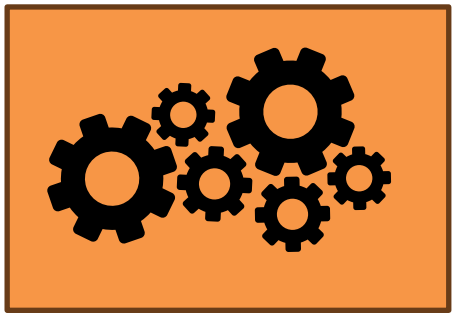
cat

(accomplishes *some* task)
(hopefully *well*)

amazing spot for good food & a fun time 🍕🍊 they offer a super unique dine-in experience with their interactive tables! also love that they have innovative weekly feature dishes 😋

Worst service I have ever experienced! Waited 25 mins for our waitress to come to our table once seated, no apology! Waited another 30 mins for our drinks, which we had to ask about.

Input



Model

Output





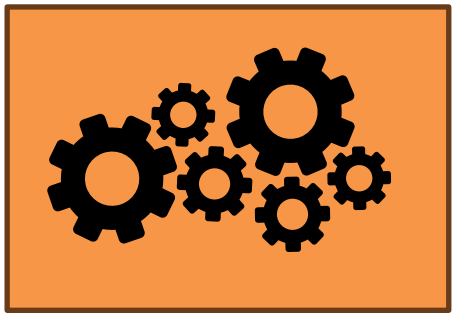
(accomplishes *some* task)

(hopefully *well*)

amazing spot for good food & a fun time 🍕🍊 they offer a super unique dine-in experience with their interactive tables! also love that they have innovative weekly feature dishes 😋

Input

Worst service I have ever experienced! Waited 25 mins for our waitress to come to our table once seated, no apology! Waited another 30 mins for our drinks, which we had to ask about.



Model



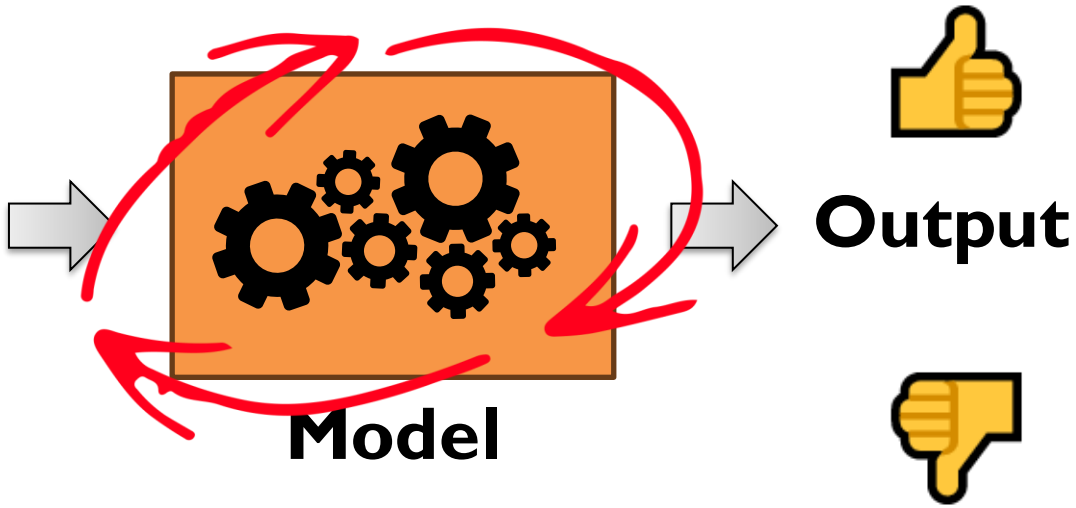
Output



How do you build such a model?

Model *learns* from the data

(review, 👍)
(review, 👎)
(review, 👍)
(review, 👍)
(review, 👎)
(review, 👎)



Machine learning algorithm
adjusts the model *parameters*

How do you build such a model?

More details later in the course…



What's supervised machine learning?



What does it mean to an AI-first company?

What does it mean to a data-driven company?

AI ~means ML

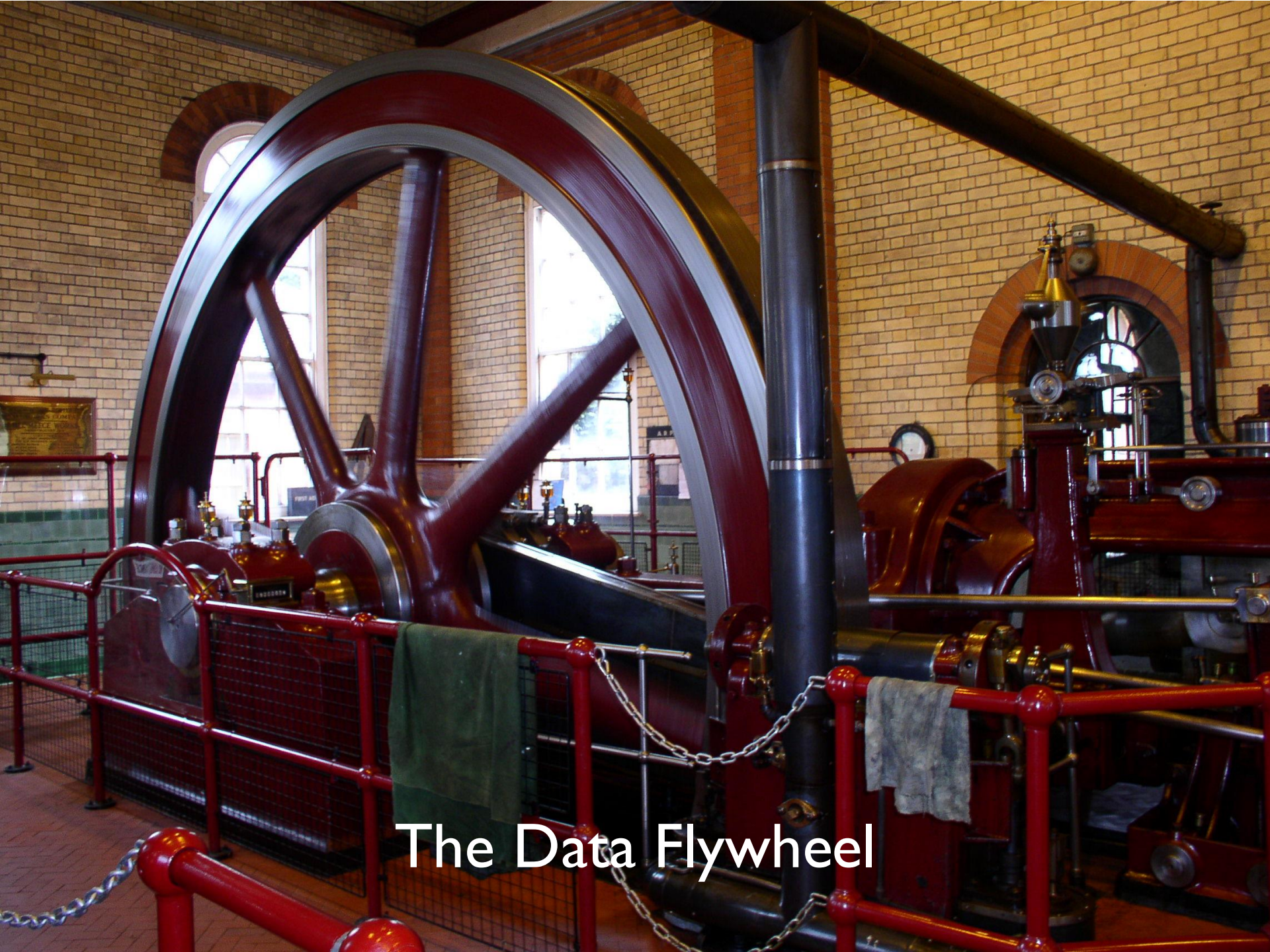
ML ~means supervised ML

supervised ML requires lots of data
("natural" and synthetic)

data is what powers AI

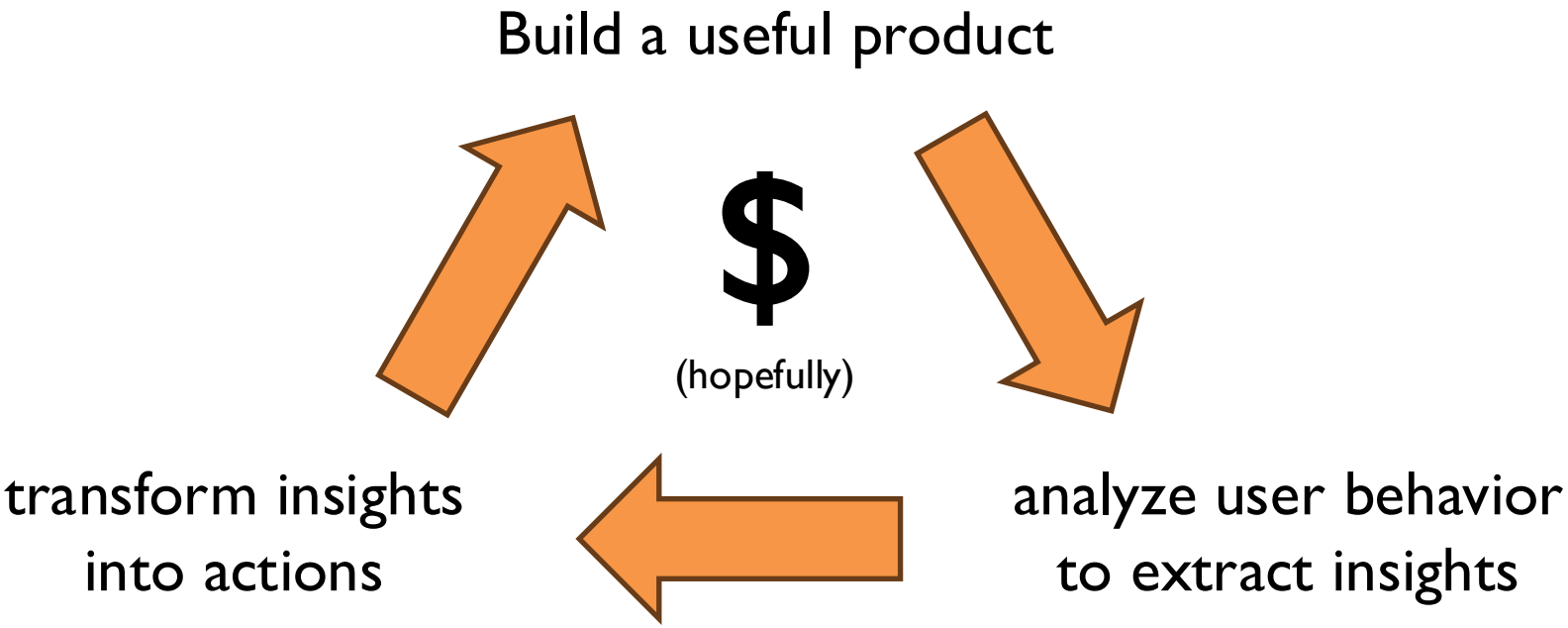
AI is what you "do" with data

What's this course about?


The Data Flywheel

The Data Flywheel

(a virtuous cycle)



Google. Facebook. Twitter. Amazon. Uber.

Business Intelligence

An organization should retain data that result from carrying out its mission and exploit those data to generate insights that benefit the organization, for example, market analysis, strategic planning, decision making, etc.

Duh!?

This is not a new idea…

Case Study

In the 1990s, Wal-Mart found that customers tended to buy diapers and beer together. So they put them next to each other and increased sales of both.*

Why didn't they do it earlier?

Relational databases weren't invented until 1970

Computers were slow and expensive

* BTW, this is completely apocryphal. (But it makes a nice story.)



5 MB hard drive in 1956

Case Study

In the 1990s, Wal-Mart found that customers tended to buy diapers and beer together. So they put them next to each other and increased sales of both.*

So what's changed?

More compute and storage

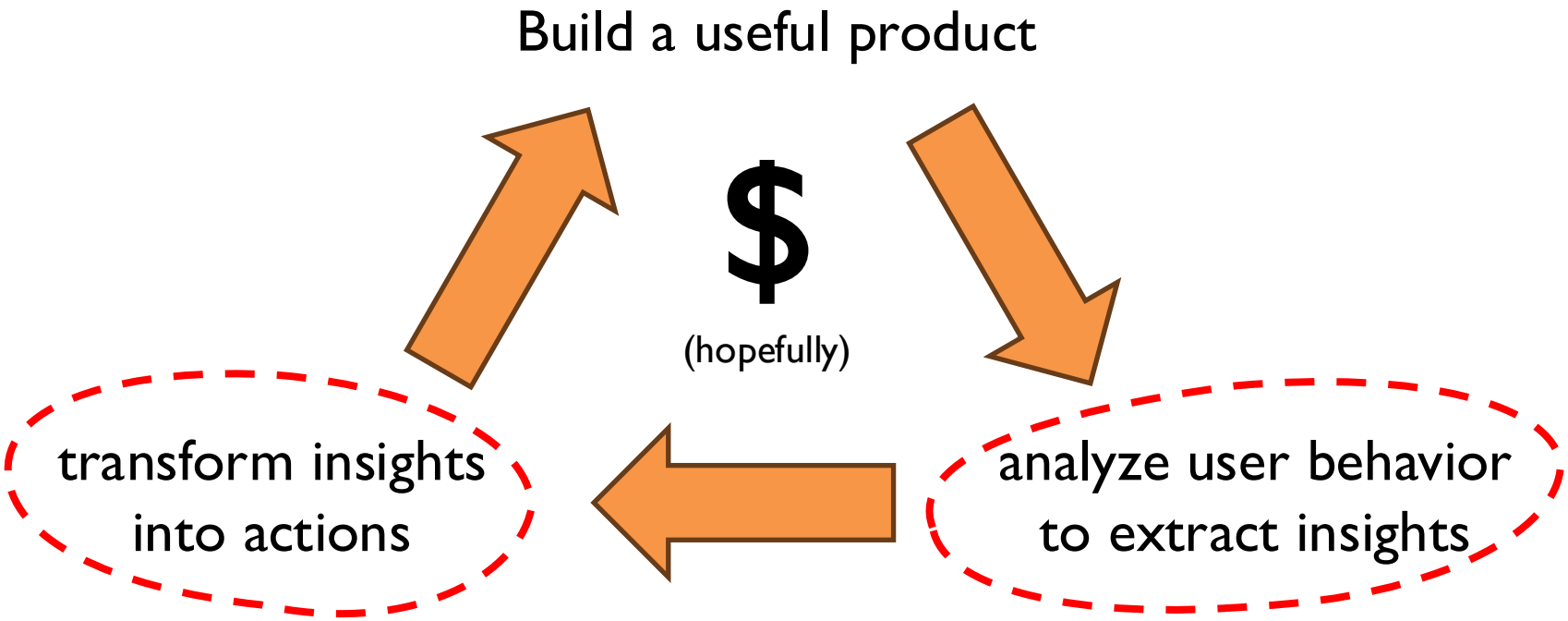
Ability to gather behavioral data

Advent of machine learning

* BTW, this is completely apocryphal. (But it makes a nice story.)

The Data Flywheel

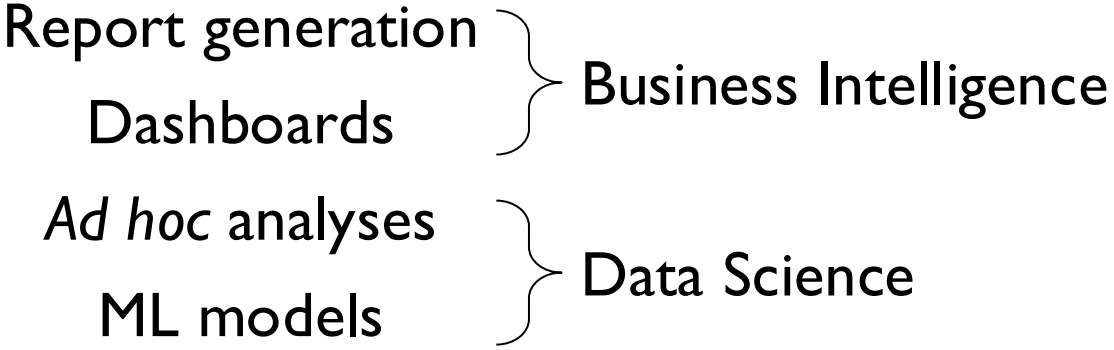
(a virtuous cycle)



Google. Facebook. Twitter. Amazon. Uber.

Transform Insights into Actions

What does that really mean?



What's this course about?

The *infrastructure* that supports the data flywheel.

data platforms

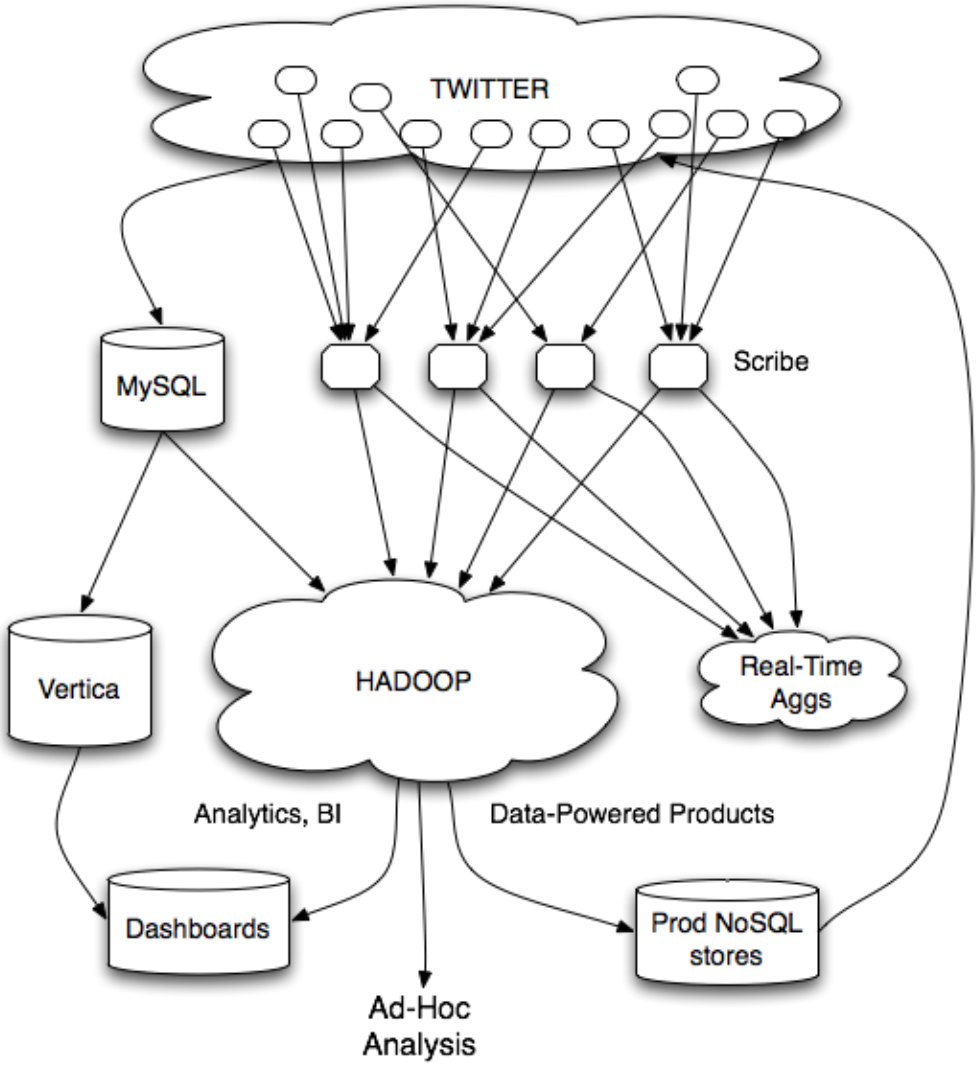
lakehouse = data warehouse + data lake

data engineering

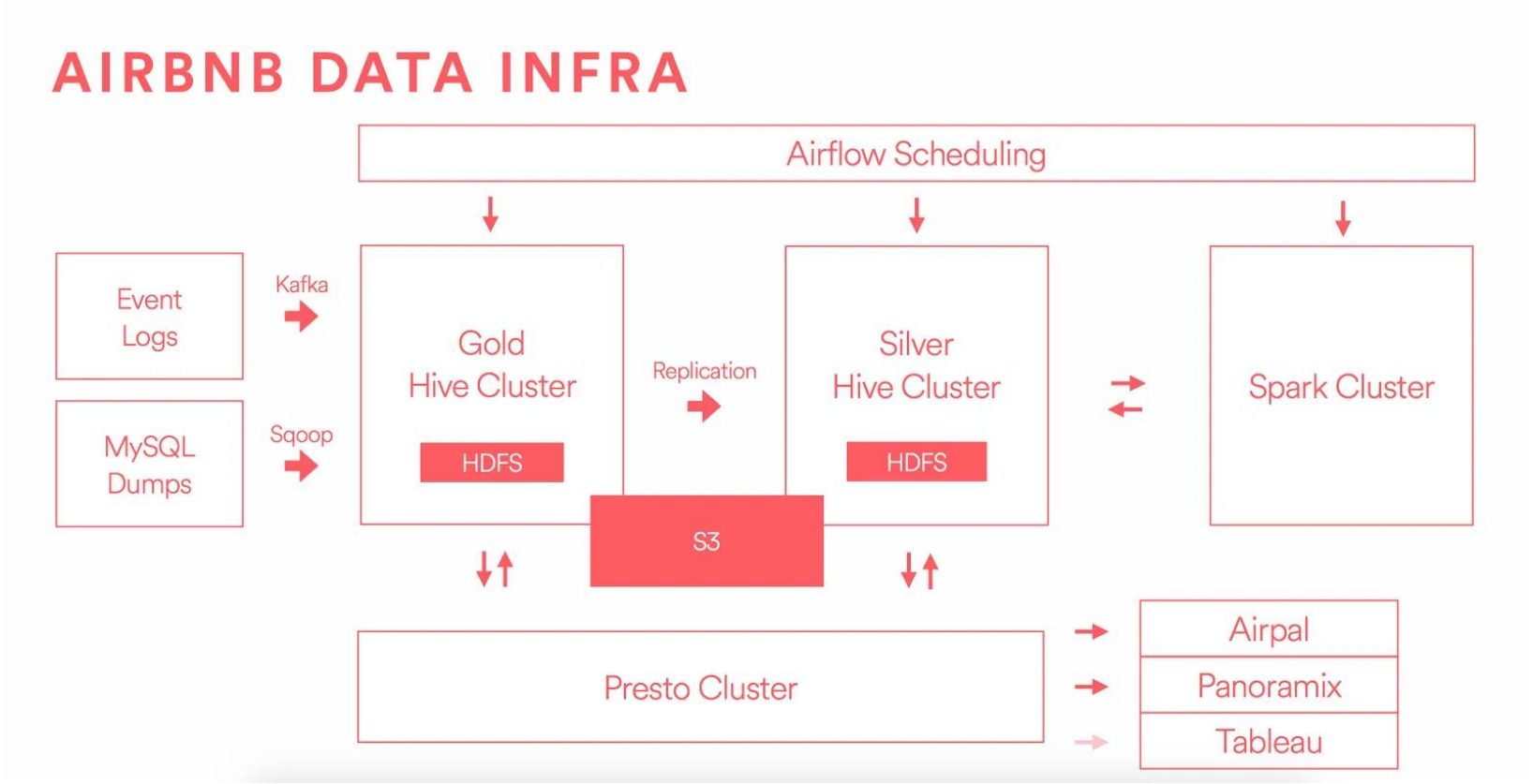
What problems do data platforms solve?

Ingesting, storing, manipulating, maintaining, serving…
the data that supports the data flywheel.

Some examples…

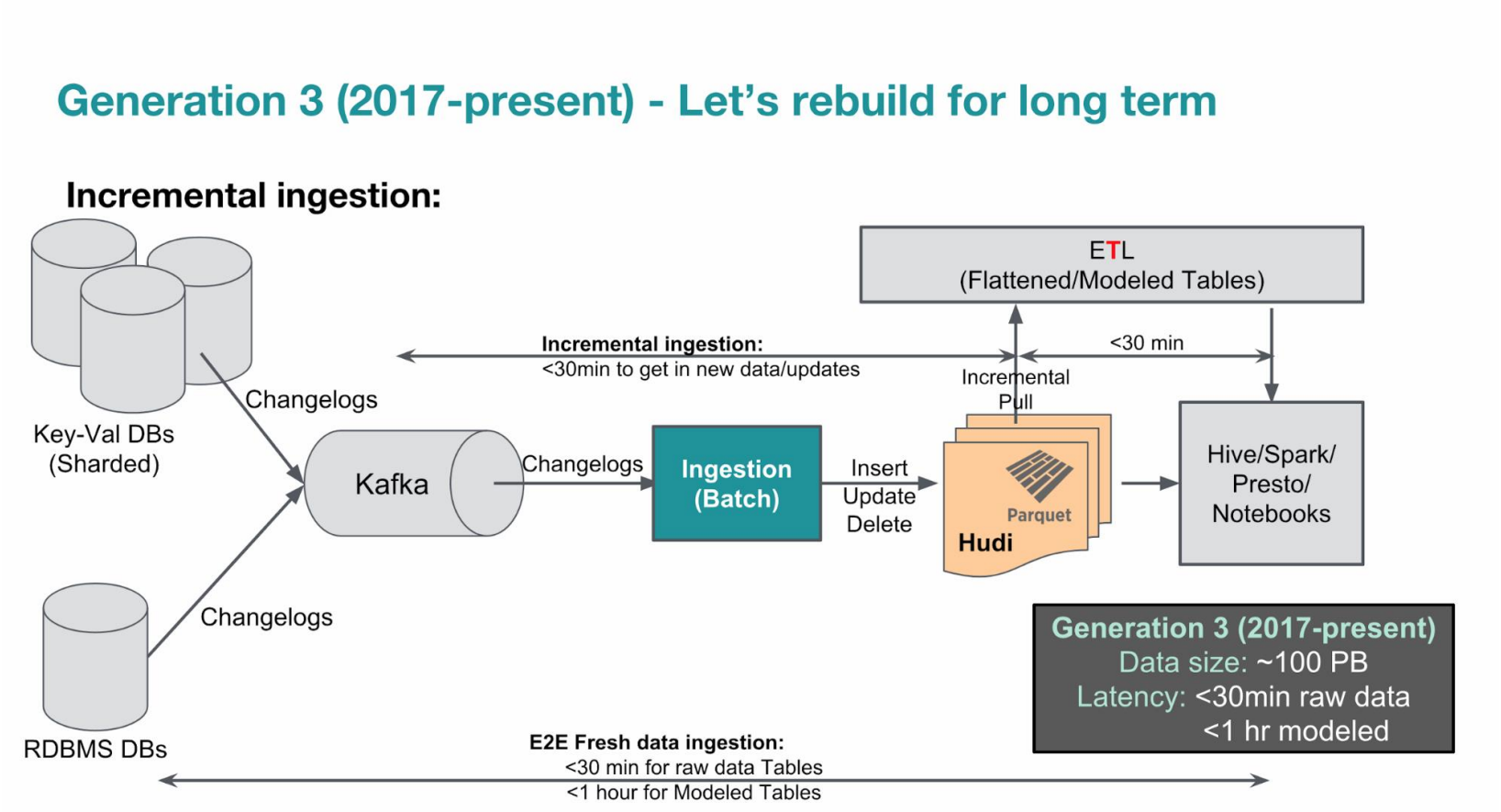


Twitter's data platform (circa 2012)



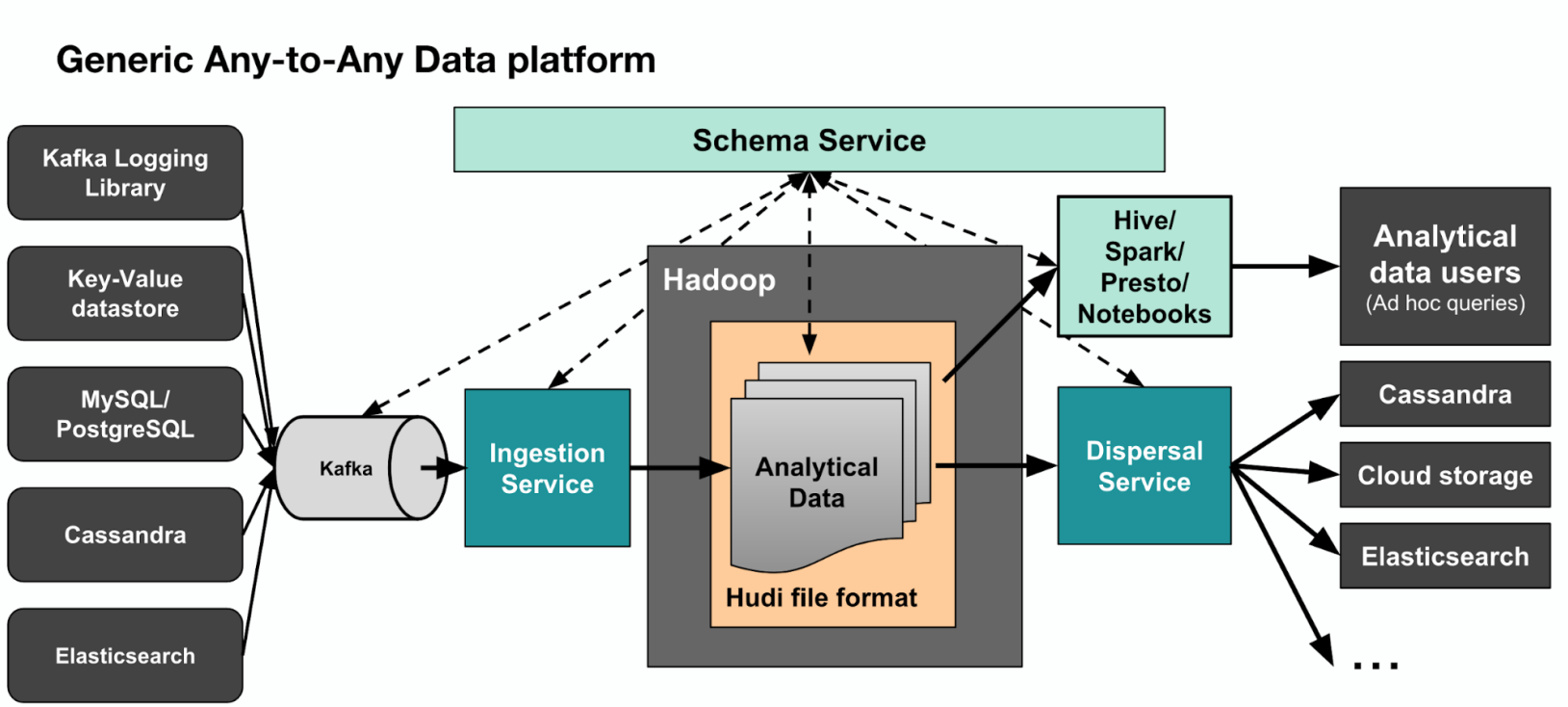
AirBnB's data platform (circa 2016)

https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c



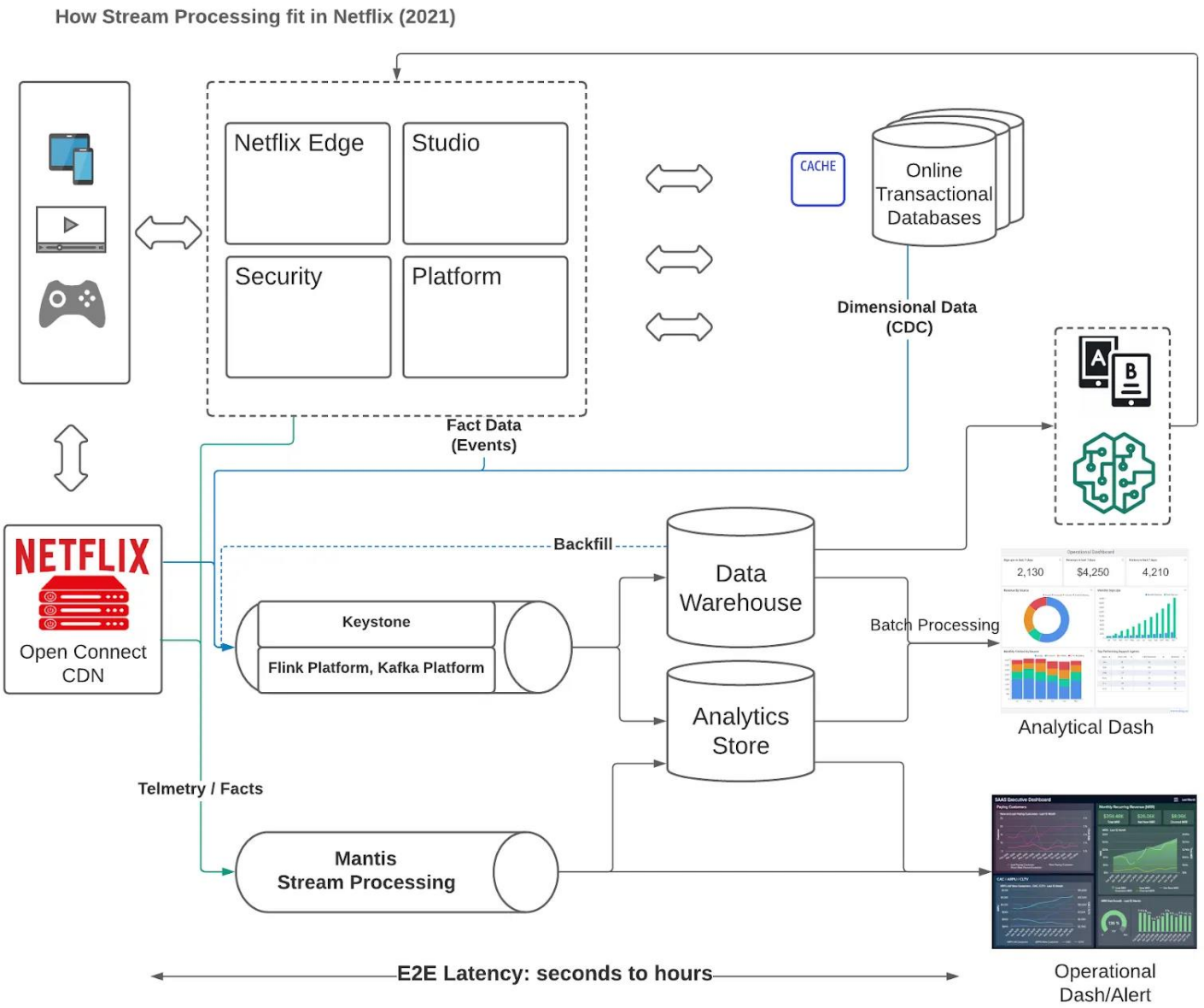
Uber's data platform (circa 2018)

https://www.uber.com/en-CA/blog/uber-big-data-platform/



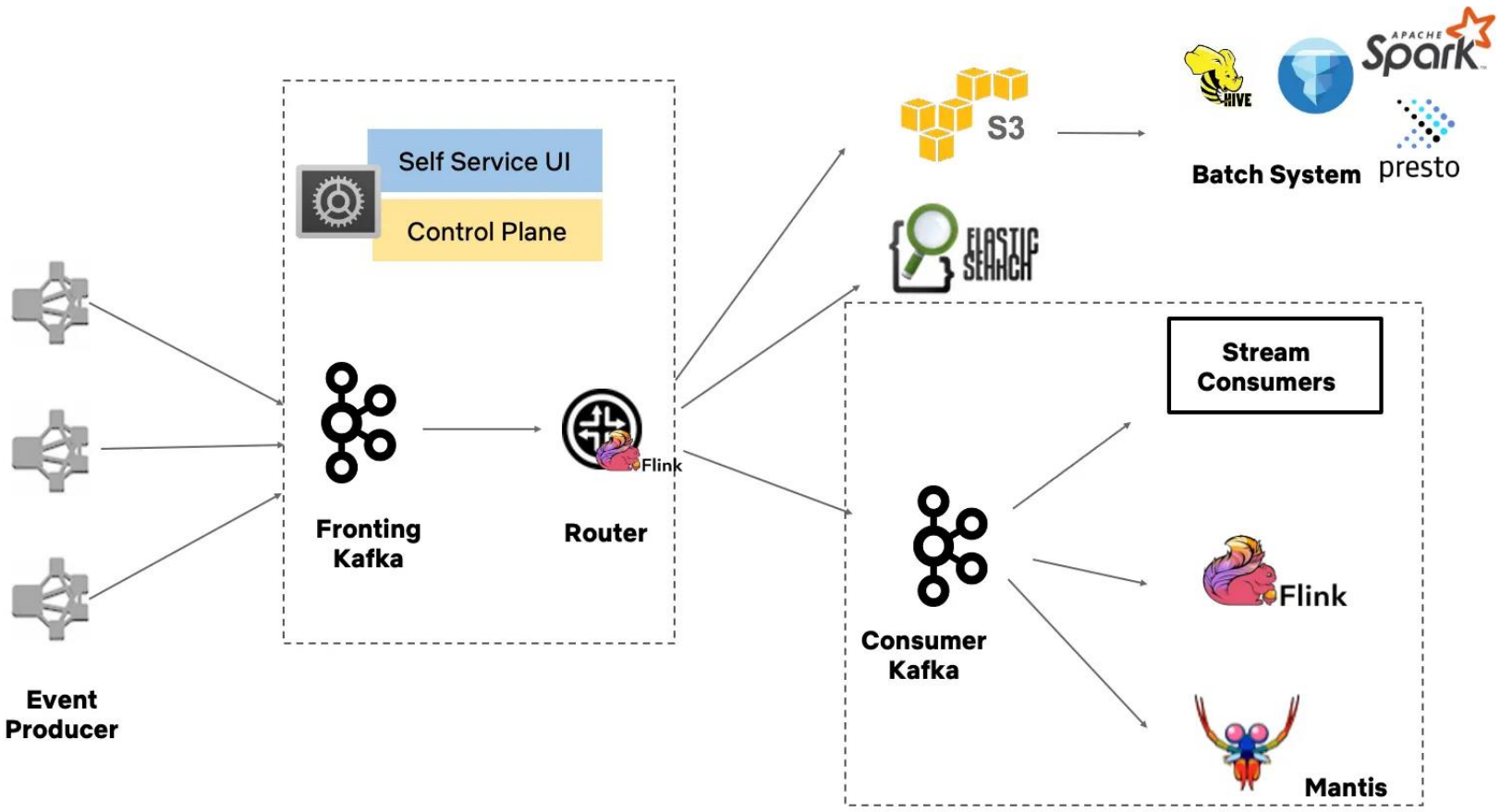
Uber's data platform (circa 2018)

https://www.uber.com/en-CA/blog/uber-big-data-platform/



Netflix's data platform (circa 2021)

https://zhenzhongxu.com/the-four-innovation-phases-of-netflixs-trillions-scale-real-time-data-infrastructure-2370938d7f01



Publish, Collect, Move & Compute event data

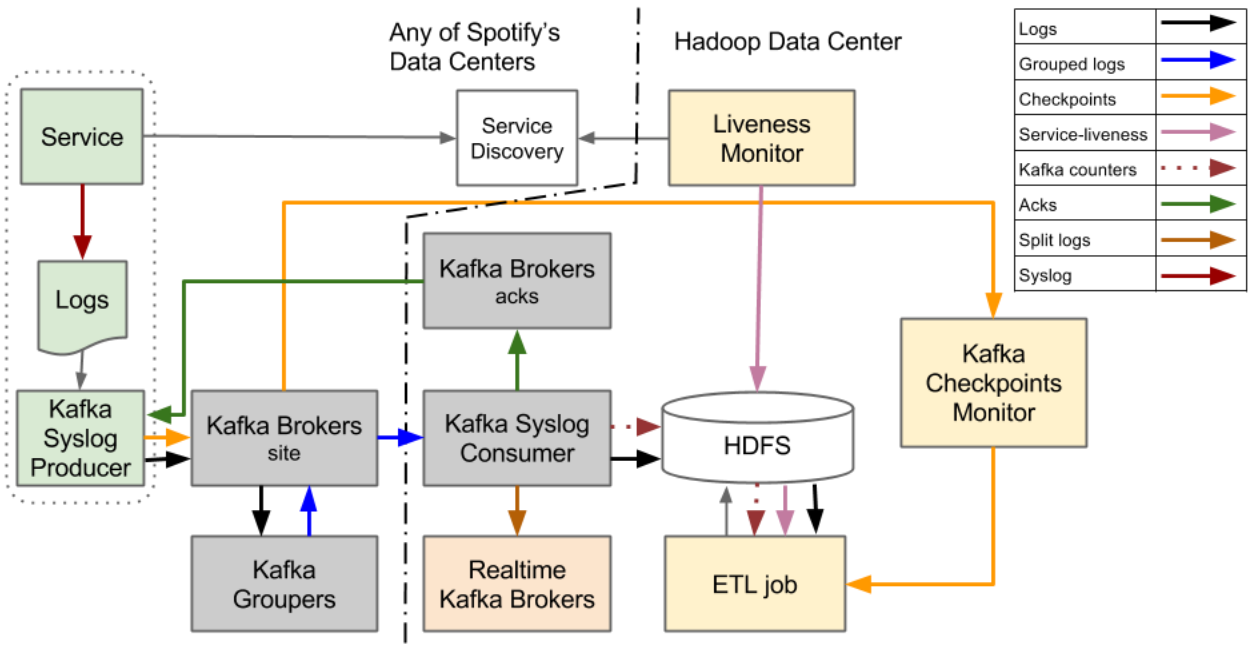
Netflix's data platform (circa 2021)

https://zhenzhongxu.com/the-four-innovation-phases-of-netflixs-trillions-scale-real-time-data-infrastructure-2370938d7f01



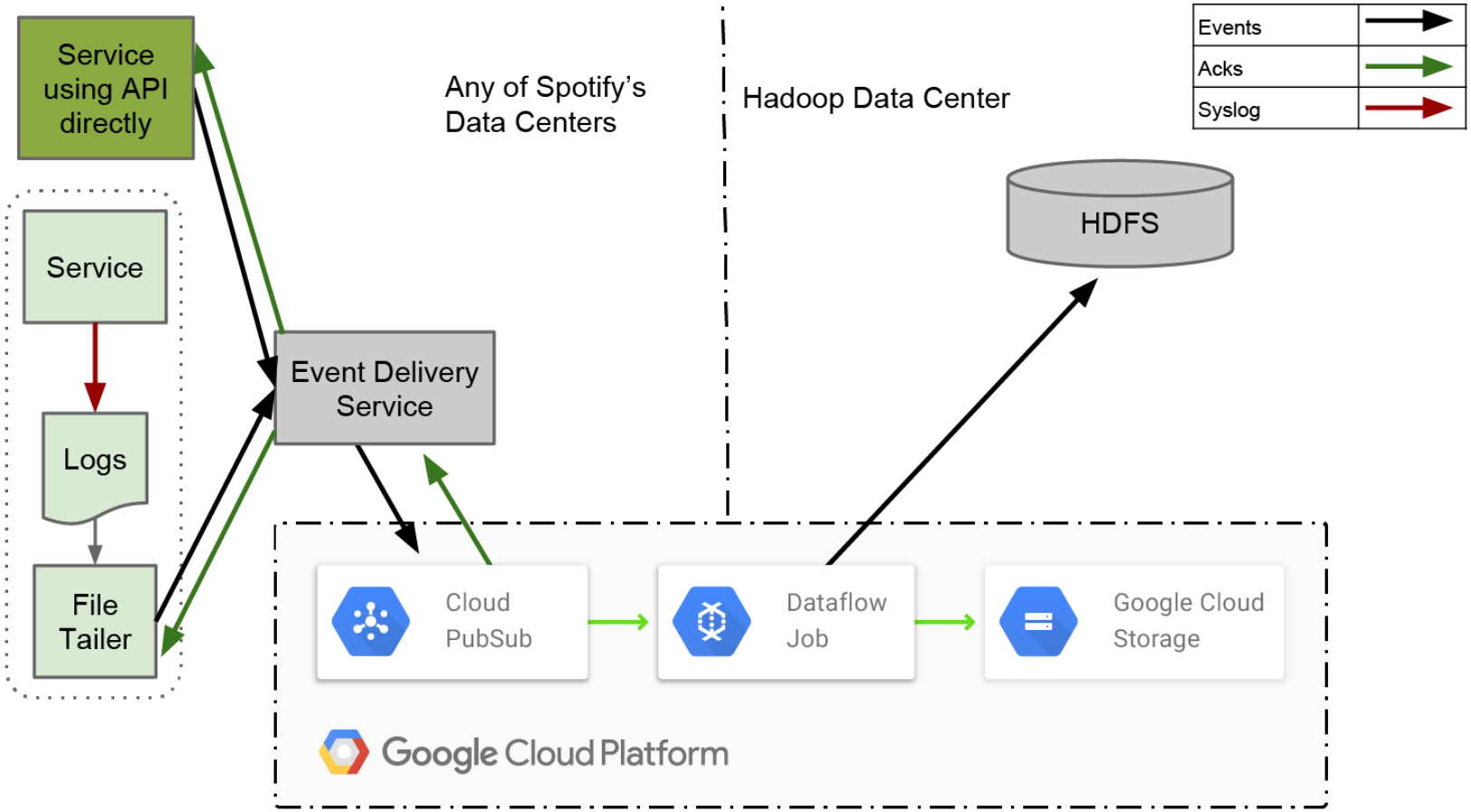
Spotify's data platform (circa 2024)

https://engineering.atspotify.com/2024/5/data-platform-explained-part-ii



Spotify's data delivery architecture (circa 2016, pre-cloud)

https://engineering.atspotify.com/2016/02/spotifys-event-delivery-the-road-to-the-cloud-part-i



Spotify's data delivery architecture (circa 2016, post-cloud)

https://engineering.atspotify.com/2016/3/spotifys-event-delivery-the-road-to-the-cloud-part-ii

What problem do data platforms solve?

Ingesting, storing, manipulating, maintaining, serving…
the data that supports the data flywheel.

By the end of the course you should be able
to "make sense" of any data platform.

Why is this a difficult problem?

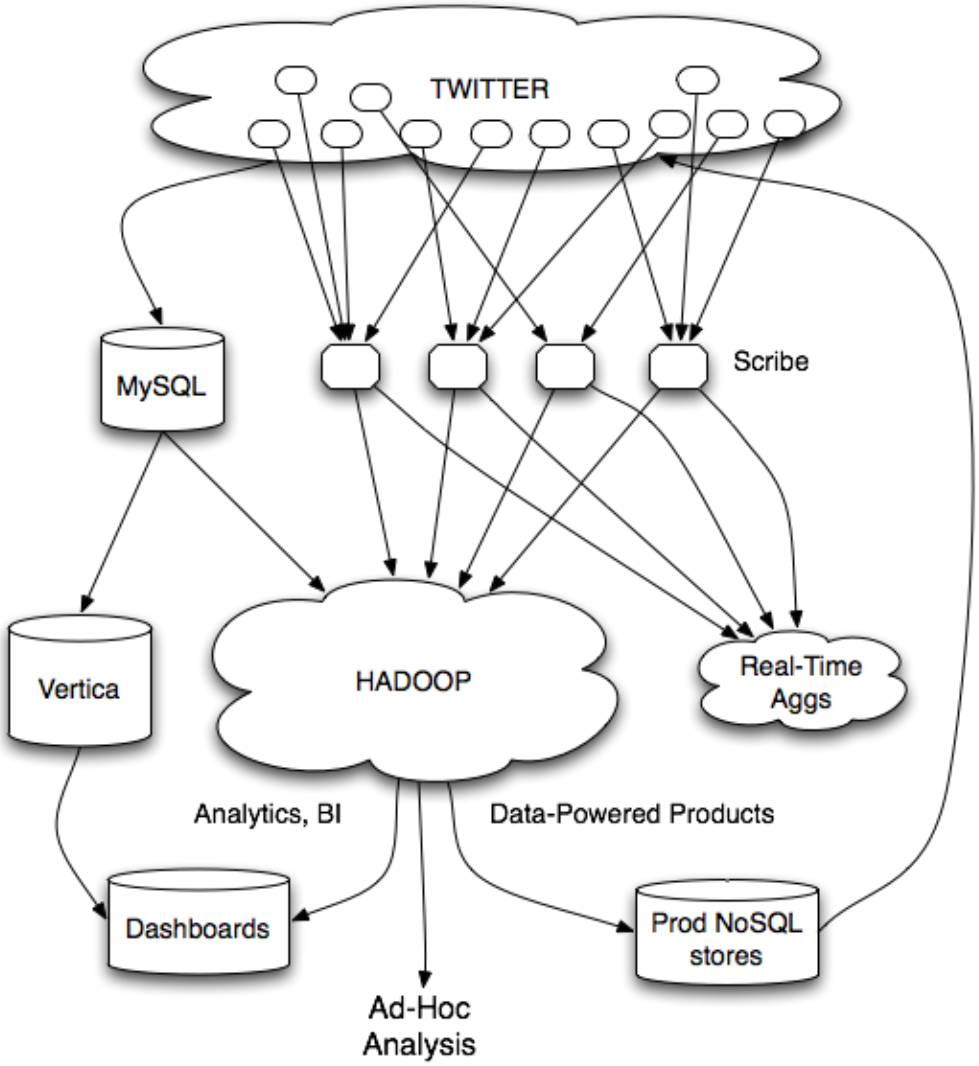
Characteristics of the data: Volume, Velocity, Variety* + Veracity

* Coined by Gartner analyst Doug Laney in 2001.

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

* Coined by Gartner analyst Doug Laney in 2001.



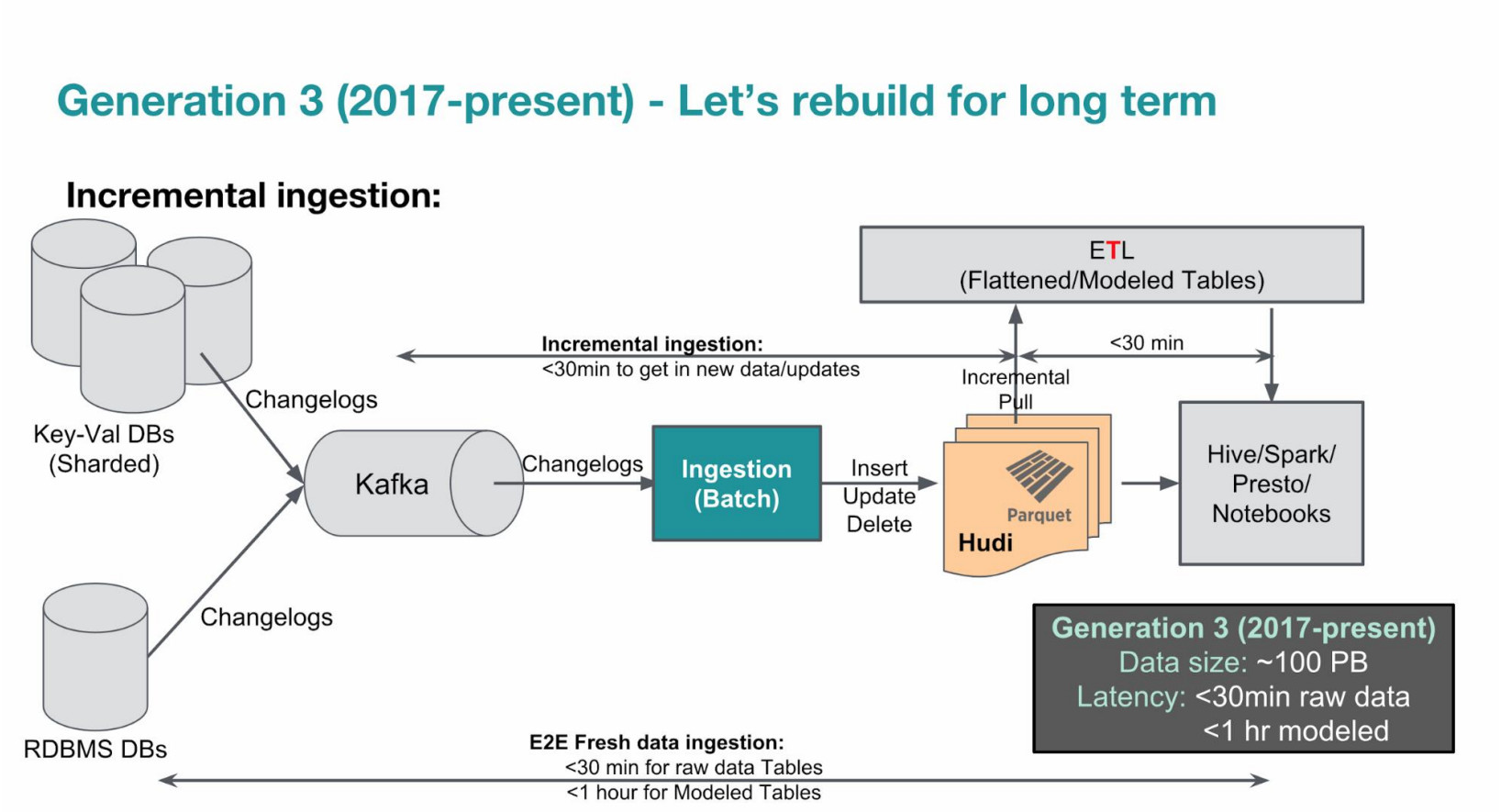
Twitter's data platform (circa 2012)

circa ~2010

~150 people total
~60 Hadoop nodes
~6 people use analytics stack daily

circa ~2012

~1400 people total
10s of Ks of Hadoop nodes, multiple DCs
10s of PBs total Hadoop DW capacity
~100 TB ingest daily
dozens of teams use Hadoop daily
10s of Ks of Hadoop jobs daily



Uber's data platform (circa 2018)

https://www.uber.com/en-CA/blog/uber-big-data-platform/

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

Want more?

`https://lintool.github.io/my-data-is-bigger-than-your-data/`

* Coined by Gartner analyst Doug Laney in 2001.

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

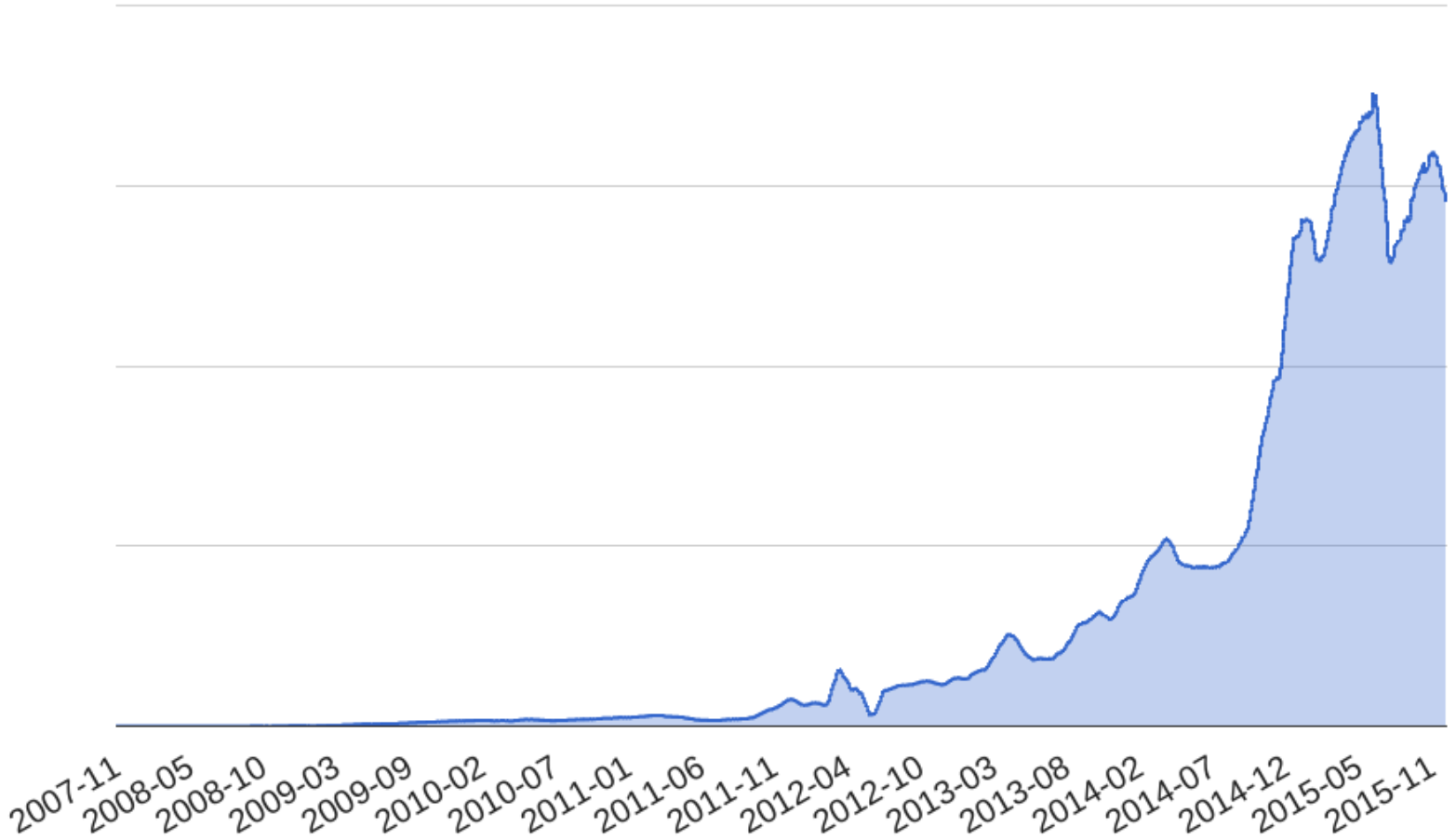
* Coined by Gartner analyst Doug Laney in 2001.



Spotify's data platform (circa 2024)

https://engineering.atspotify.com/2024/5/data-platform-explained-part-ii

When it comes to scalability, Spotify's Data Collection platform collects more than 1 trillion events per day. (circa 2024)



Amount of Spotify's Delivered Events over time (circa 2016)

https://engineering.atspotify.com/2016/02/spotifys-event-delivery-the-road-to-the-cloud-part-i

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

Want more?

`https://lintool.github.io/my-data-is-bigger-than-your-data/`

* Coined by Gartner analyst Doug Laney in 2001.

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

Previously – relational

Today – semi-structured, unstructured, graph, multimodal, etc.

* Coined by Gartner analyst Doug Laney in 2001.

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

GI / GO

* Coined by Gartner analyst Doug Laney in 2001.

Why is this a difficult problem?

Characteristics of the data: Volume, Velocity, Variety* + Veracity

By the end of the course you should be able understand and appreciate the challenges here.

* Coined by Gartner analyst Doug Laney in 2001.

Data Engineering "Truths"

(That are rarely acknowledged explicitly)

Data engineering is practical

No solutions, only tradeoffs (but best practices exist)

Data platforms…

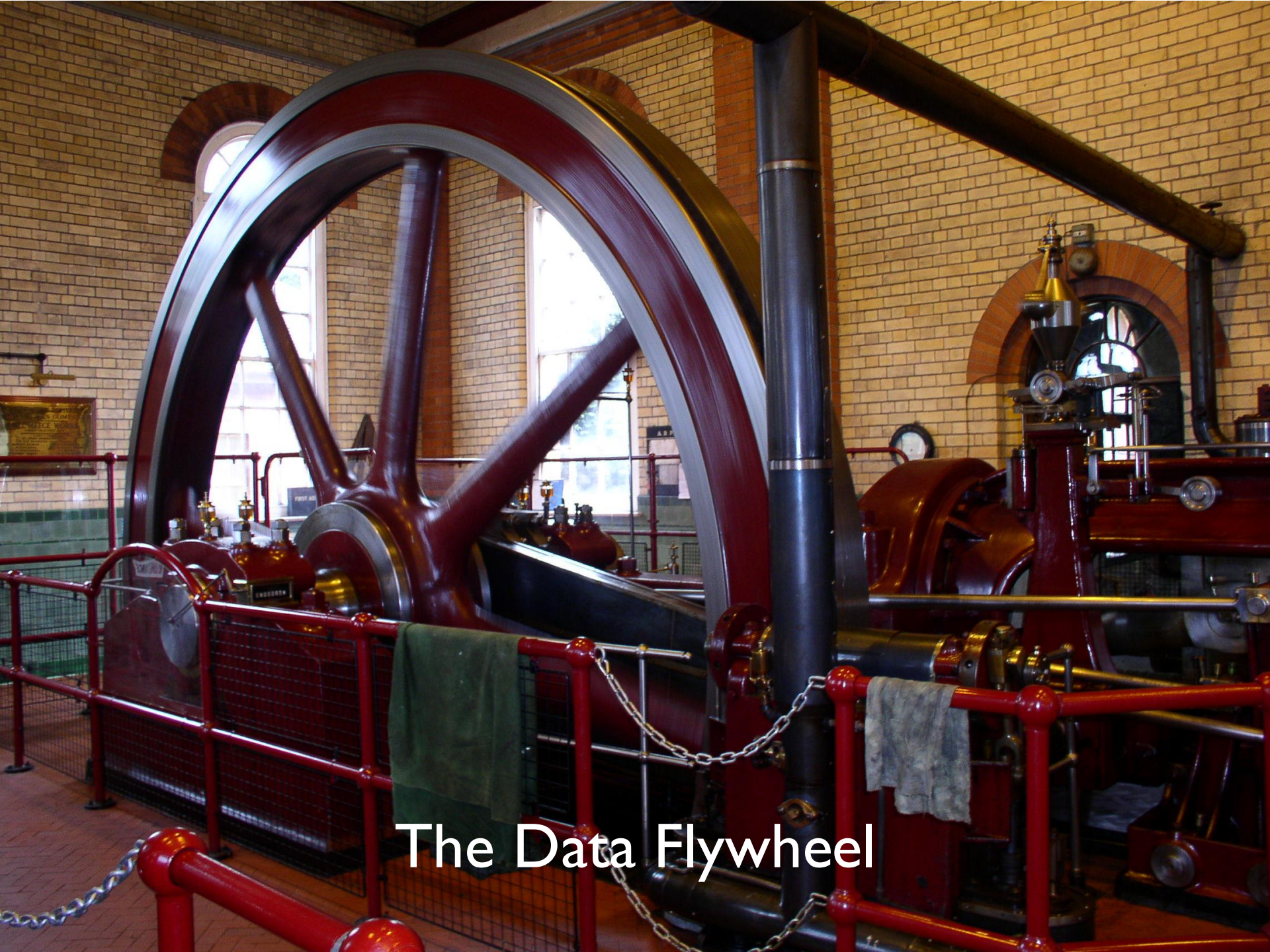
already exist in some form

are complex, messy, and evolving

combination of technologies and processes

subjected to technical, organizational, and business constraints

What's this course about?


The Data Flywheel

What's this course *not* about?

machine learning
data science
distributing processing
building database applications

Course Mechanics

Back to this…

Data Engineering "Truths"

(That are rarely acknowledged explicitly)

Data engineering is practical

No solutions, only tradeoffs (but best practices exist)

Data platforms…

already exist in some form

are complex, messy, and evolving

combination of technologies and processes

subjected to technical, organizational, and business constraints

Hence…

My Teaching Philosophy

Focus on concepts and intuition

Focus on reasoning about tradeoffs

Less emphasis on detailed algorithms

(don't confuse with non-technical)

Less emphasis on specific pieces of software

(how concepts are operationalized – in passing)

(how to actually use *x* – on your own)

Course Materials

All announcements will be posted on Piazza
All materials will be posted on the course homepage
You are responsible for keeping up to date.
"I didn't know" or "I didn't see it" won't be accepted as an excuse.

Readings will be assigned from textbooks and papers
You are responsible for the readings.
Assigned by week… when to do it?

Course Grade

six assignments

GitHub to manage logistics of submissions

midterm/final exam

Coding Agents

tl;dr – required!

Use coding agents effectively and responsibly!
What does that mean?

You can outsource *doing*….
… but you can't outsource *understanding*
… but you can't outsource *accountability*

(direct applications to assignments + exams)

Coding Agents: Starting Points

Just try it!

Talk to your data, code base, SQL server…

Learn the “surfaces” and personalize your own workflow…

Back to the beginning…



What does it mean to an AI-first company?

What does it mean to a data-driven company?

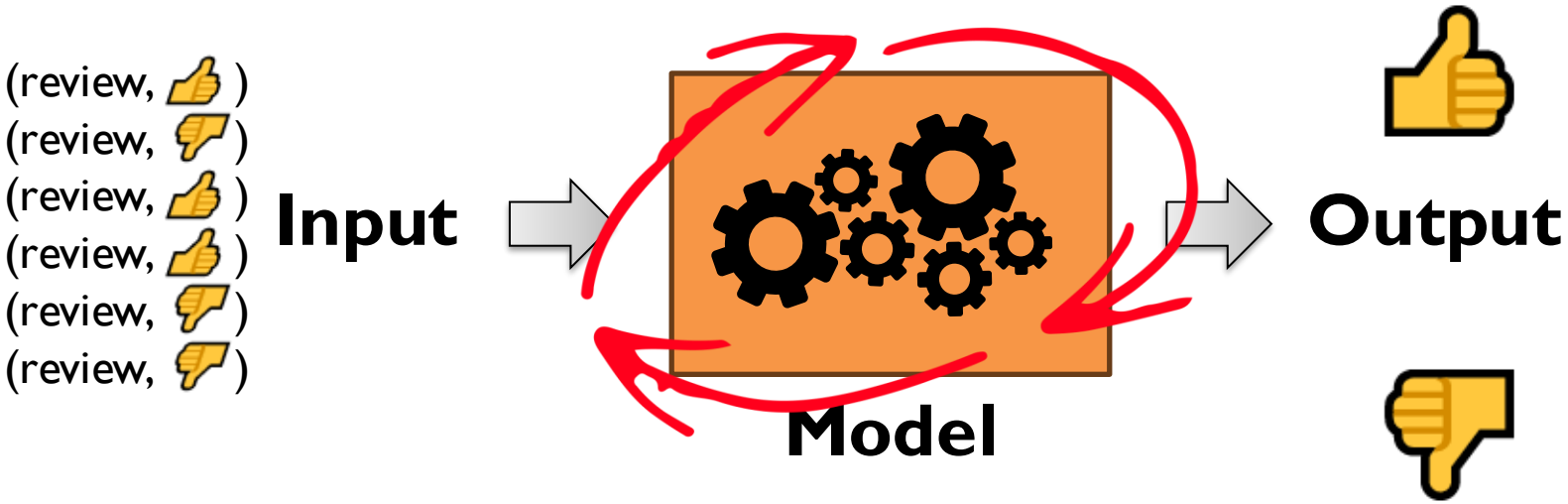
AI ~means ML

ML ~means supervised ML

supervised ML requires lots of data

("natural" and synthetic)

Model *learns* from the data



Machine learning algorithm
adjusts the model *parameters*

Where does the data *actually* come from?


富嶽三十六景 神奈川沖浪裏
Stay tuned…