

Data Analysis with Hadoop & Spark

Part One



Agenda

- Introduction to Distributed Computing
 - The Age of Data Products
 - Building Data Products at Scale
 - Data Product Architectures
- Hadoop: An Operating System for Big Data
 - Hadoop Architecture
 - What's In A Cluster?
 - HDFS Caveats
- Hands-on Lab

Introduction to Distributed Computing

The Age of Data Products

What is a data product?

“A data product is a product that is based on the combination of data and algorithms.”

-- Hilary Mason



Google

[Google Search](#) [I'm Feeling Lucky](#)

Carrier 3:19 PM

Search Results

San Francisco → Hong Kong, 3/11/13

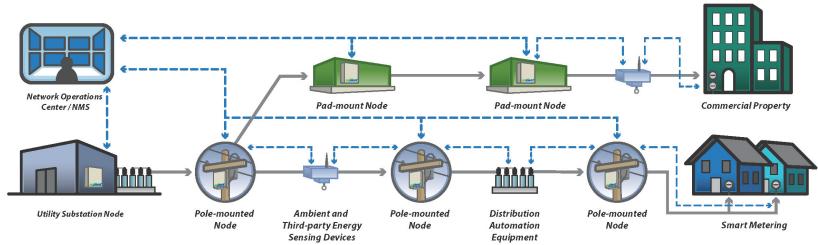
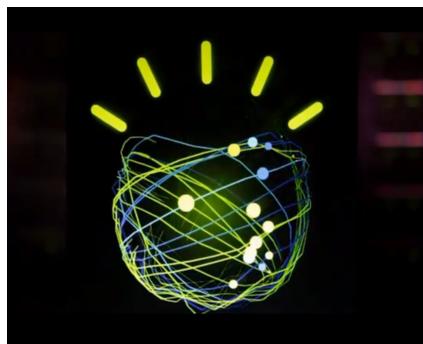
flights	Depart:	12PM	12AM	12PM	12AM
	Arrive:	3AM	3PM	3AM	3PM
\$1,017	Singapore				
\$1,048	United				
\$1,030	Air Canada				
\$1,067	China Air				
\$1,148	Cathay*				
\$1,055	Korean				
\$1,068	ANA*				
\$1,068	United				
\$1,068	ANA*				
\$1,079	Air China				

Agony Price Depart Length

“A data application acquires its value from the data itself, and creates more data as a result. It’s not just an application with data; it’s a data product.

*Data science enables
the creation of data products.”*

-- Mike Loukides

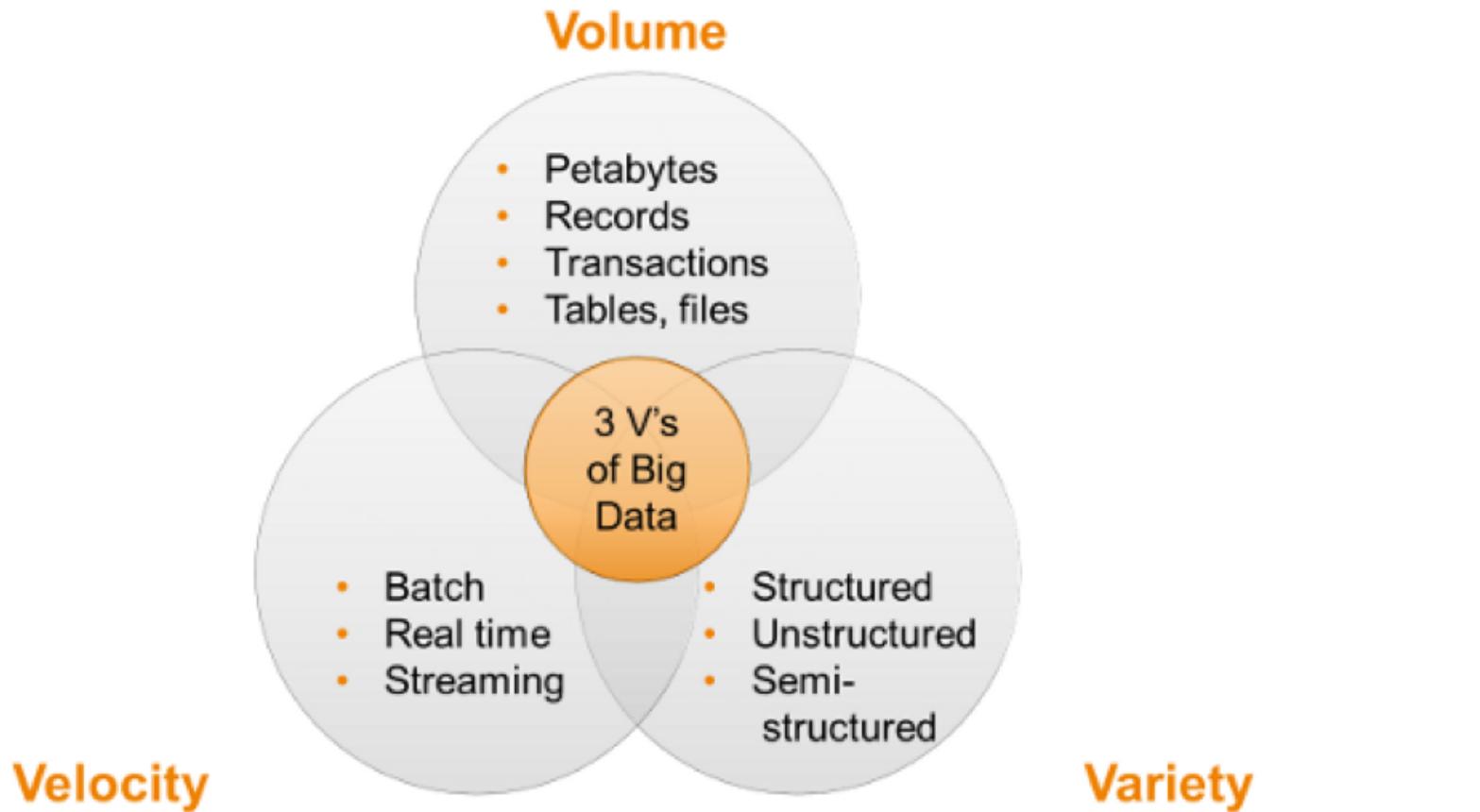


Data Products are self-adapting, broadly applicable software-based engines that derive their value from data and generate more data by influencing human behavior or by making inferences or predictions upon new data.

- District Data Labs

Building Data Products at Scale

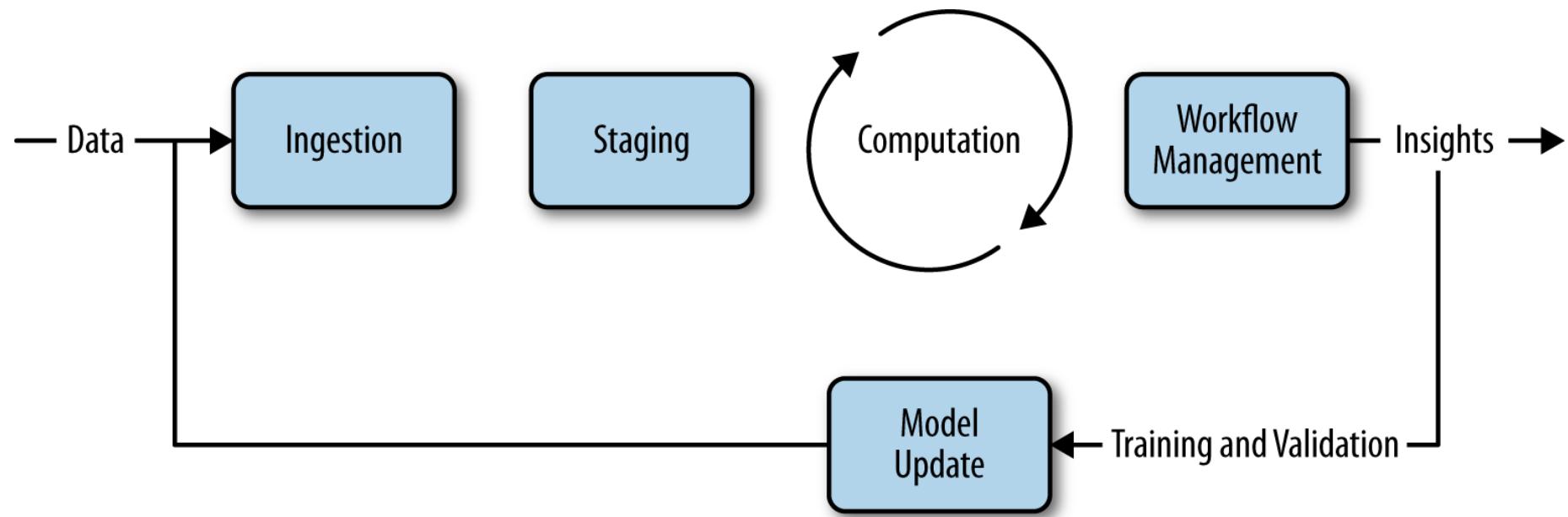
What Makes Data Big?



Possibilities for Big Data

- Higher dimensionality → increased complexity
- Solution: Machine learning (ML)
- ML learns by example → need lots of examples!

Big Data Pipeline



Let's Build a Data Machine

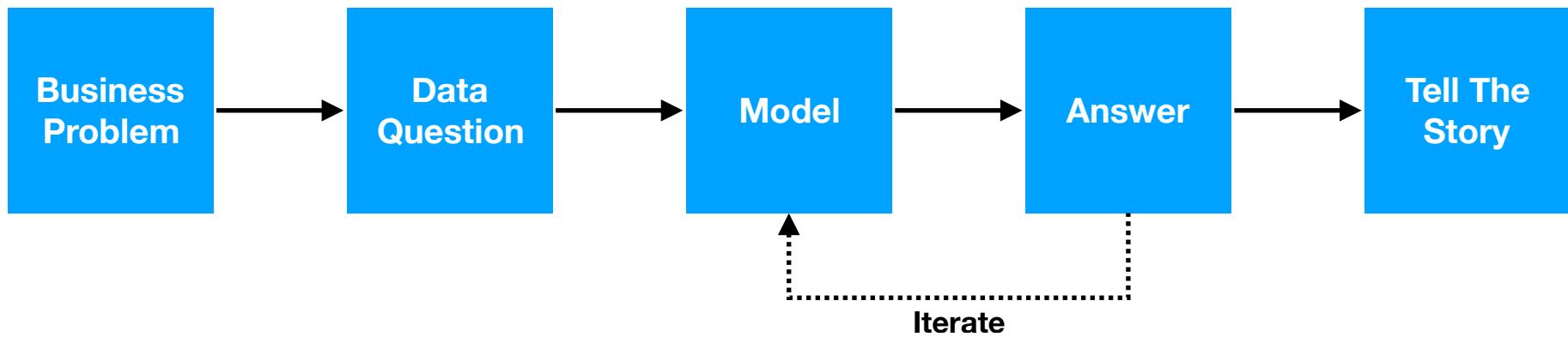
Input [] Output

Data Product Architectures

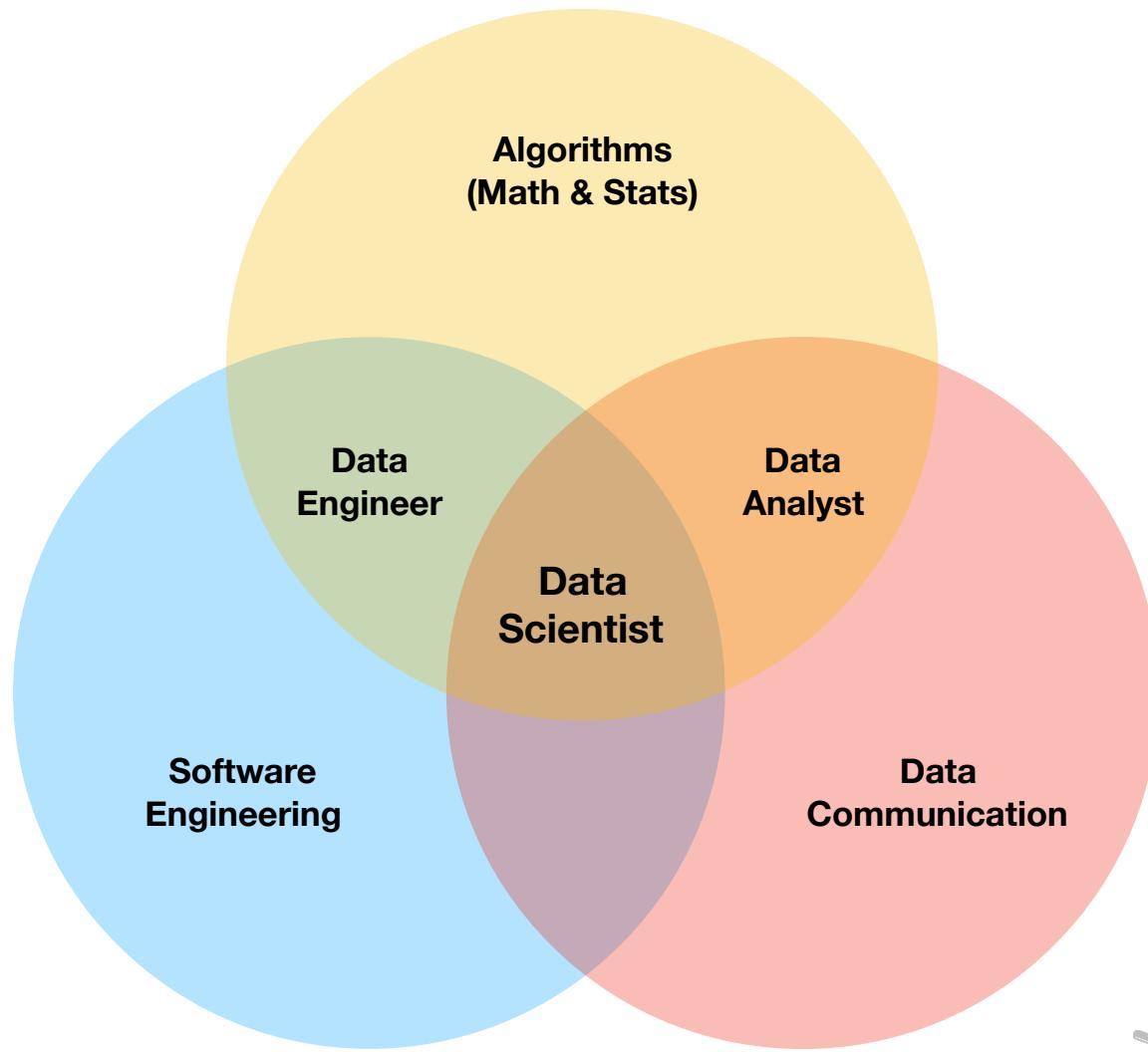
What is data science?

What is the goal of data science?

The Data Science Process



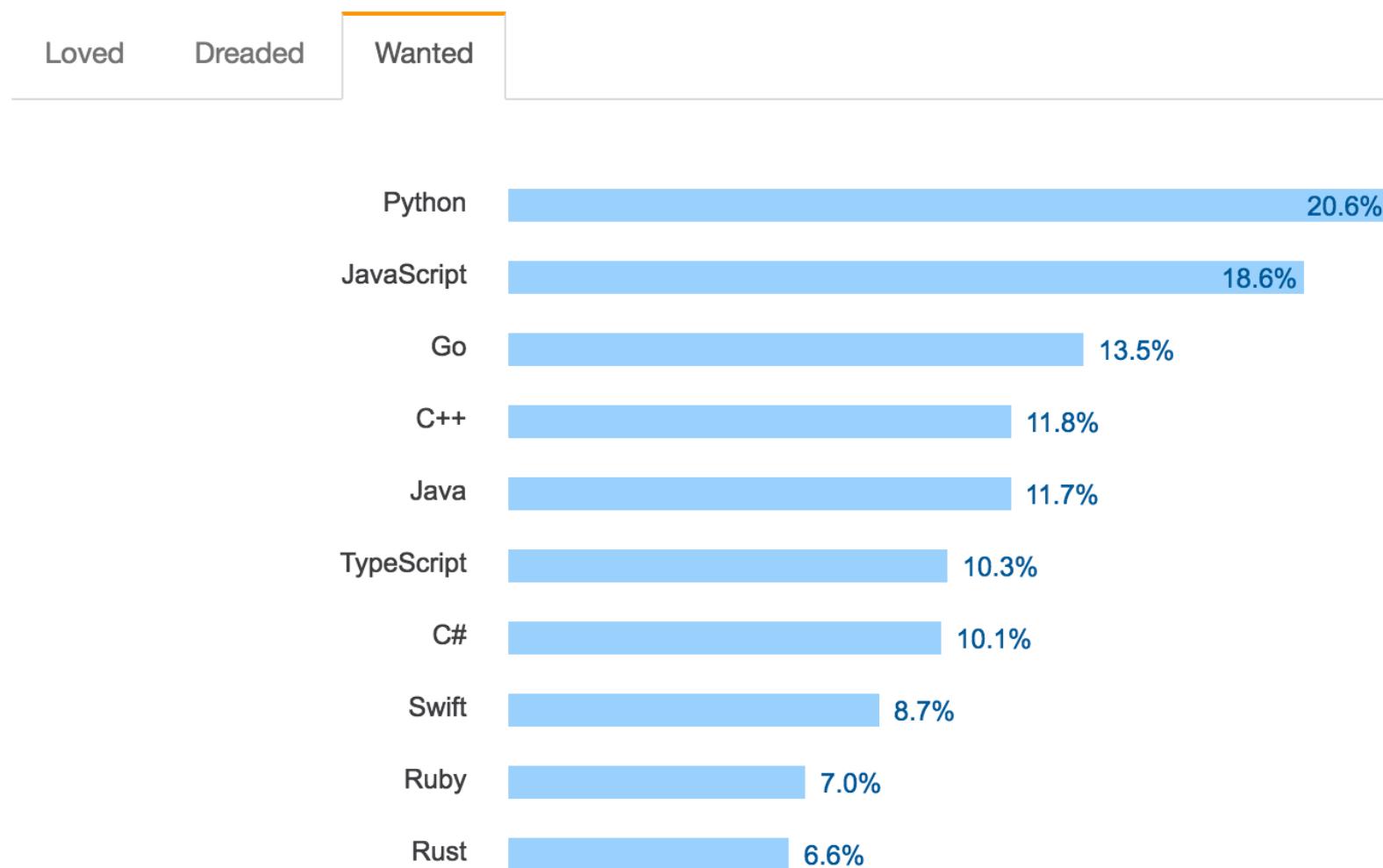
Data Scientist Skillset



Software Engineering

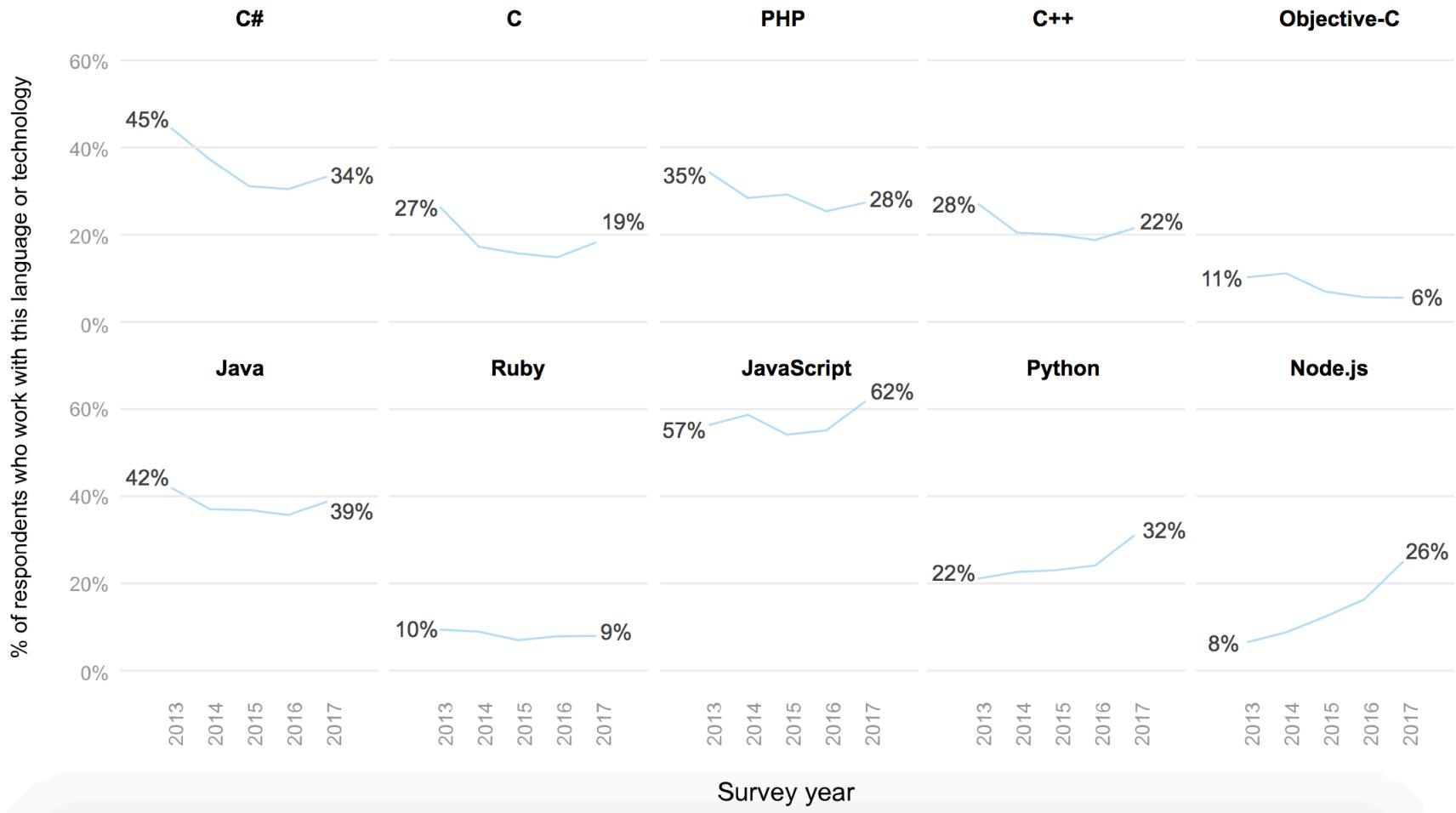
- Build systems that can gather, process and store data
- Know 1+ programming languages
- Can apply your knowledge and skills to large datasets
- Familiar with distributed computing & the technologies

Most Loved, Dreaded, and Wanted Languages



Change in technology popularity over time

Arranged in ascending order of total 2013 to 2017 change.



Distributed Computing

- Linux
- Amazon Web Services (EC2, S3, RDS, etc.)
- Docker, Amazon ECS, Zookeeper, Kubernetes
- Apache Kafka
- Apache Spark, Apache Storm
- Redis, PostgreSQL, Cassandra
- HDFS

Suggested Engineering Path

- Python
- Amazon EC2 & S3
- PostgreSQL
- Apache Spark

Algorithms (Math & Stats)

- Understand stats, math and modeling rules
- Understand existing algorithms and create new ones
- Can design and measure experiments

Is ML Stats?

- Goal of both: learn from data.
- They are nearly the same and are converging.
- ML produces statistical models
- A model is a mathematical function defining the relationship between a set of inputs and a set of outputs.

Base Concepts to Learn

- Classification
- Regression
- Clustering
- Dimensionality reduction
- Model selection
- Preprocessing

Suggested Algorithms Path

- Jupyter Notebook
- pandas
- scikit-learn: classification, regression, clustering
- Spark MLlib: data science at scale

Data Communication

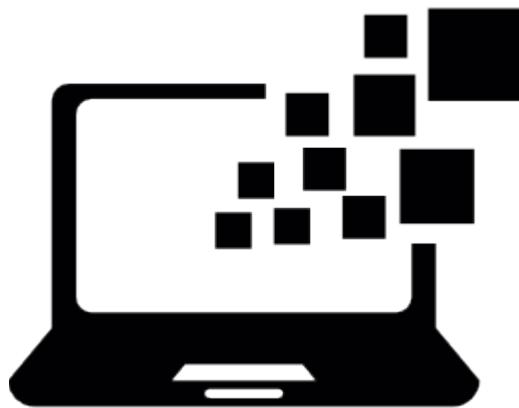
- Communicate the solution verbally.
- Communicate the solution with charty goodness.
- Relate the solution to business problems.
- Recommend actions to be taken using the solution.

Suggested Data Communication Path

- Jupyter Notebook
- Pandas
- Matplotlib
- Meetups & brown bag lunches

Ultimately it's all about
the users...

Two Objectives Orient Data Science to Users



Data Products

Improve product performance via automatic decision making.

Decision Science

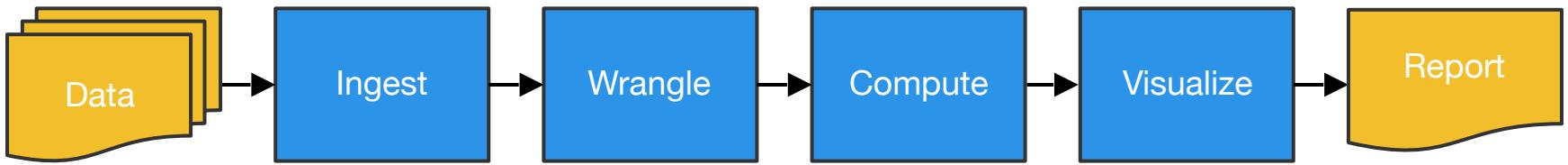
Inform strategy and key decisions by analysis of business metrics.

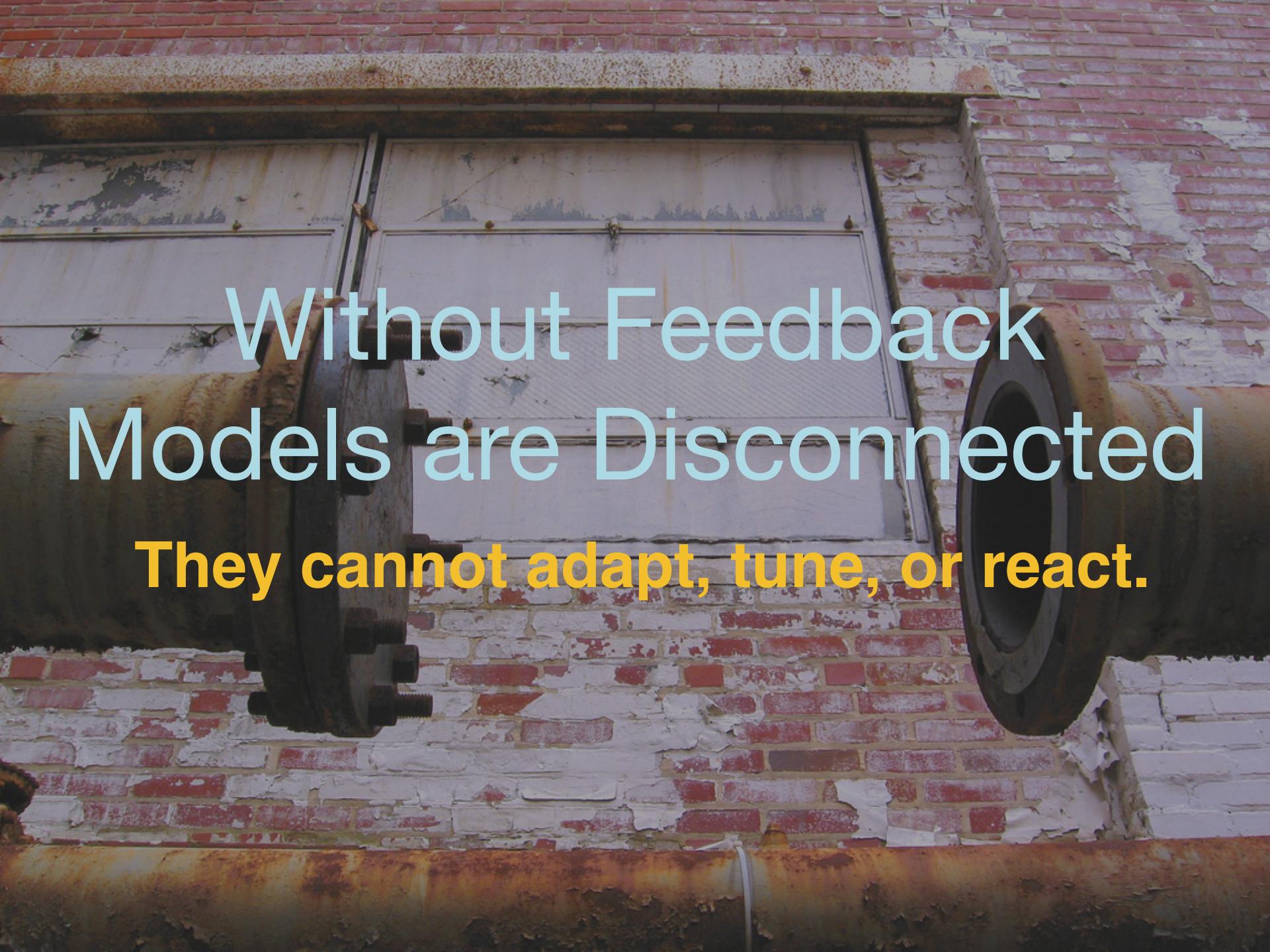
Data Products are Applications that



Employ Many Machine Learning Models

How Many Are Taught Data Science

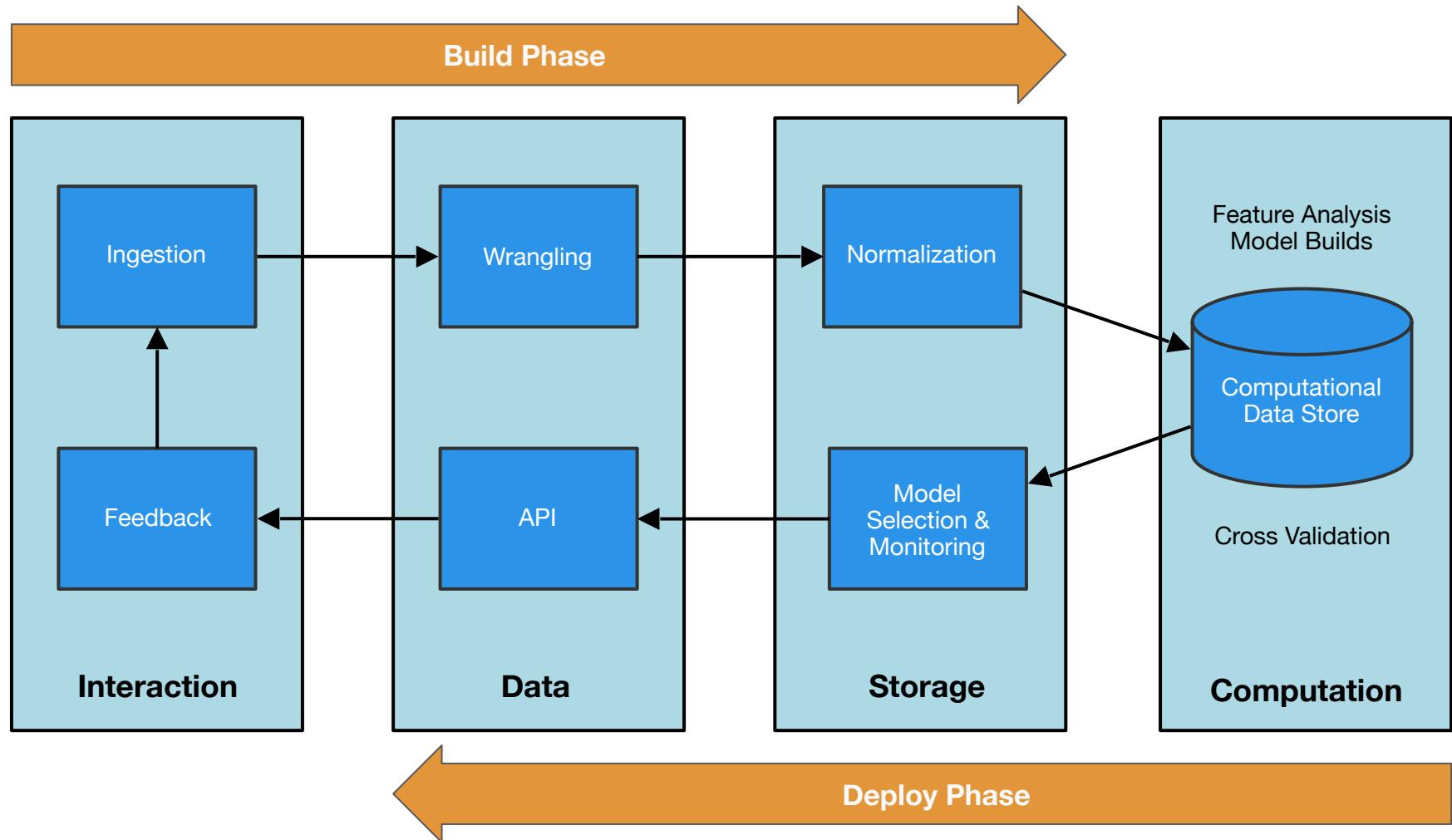


A photograph of a large, weathered metal door set into a brick wall. The door is light-colored with visible rust and damage. Two large, dark circular plates are mounted on the wall on either side of the door frame. The brick wall is made of red and white bricks, some of which are missing or broken.

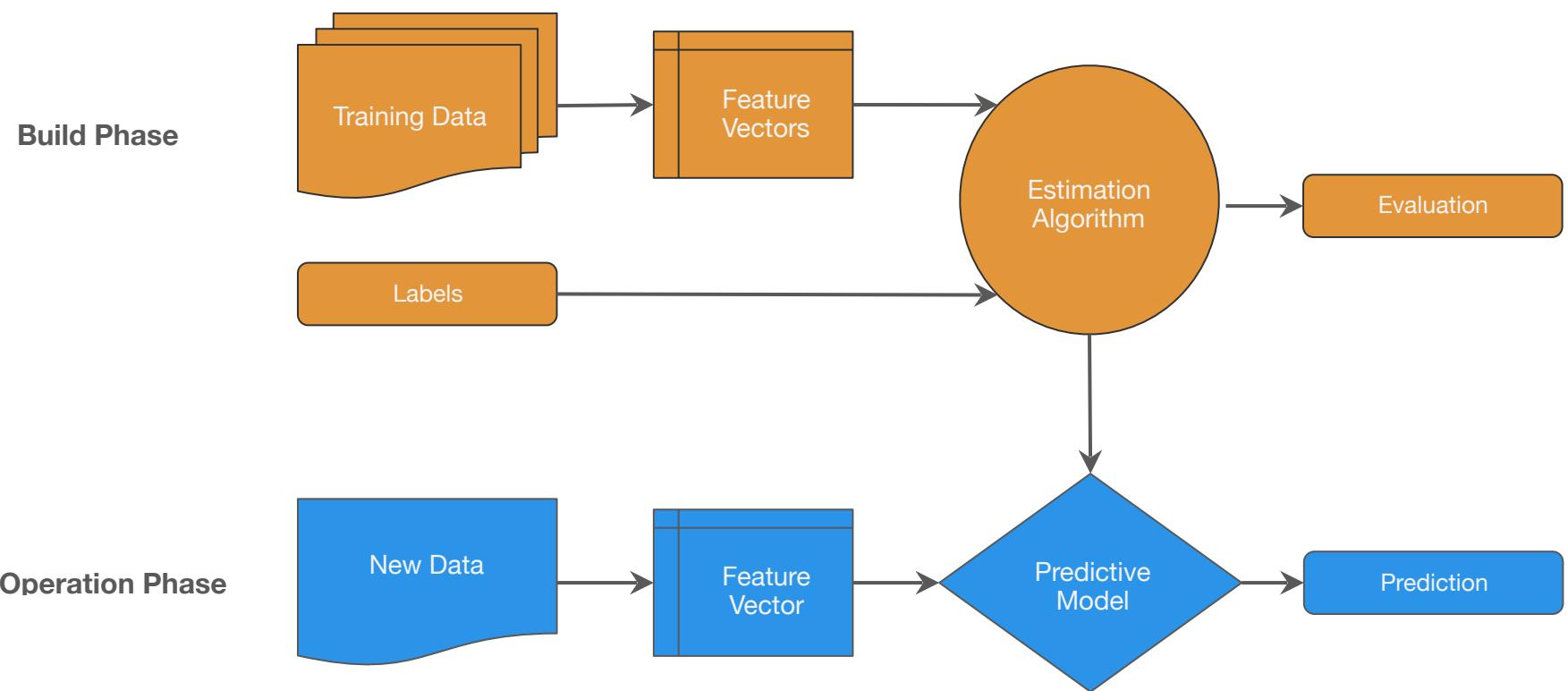
Without Feedback
Models are Disconnected

They cannot adapt, tune, or react.

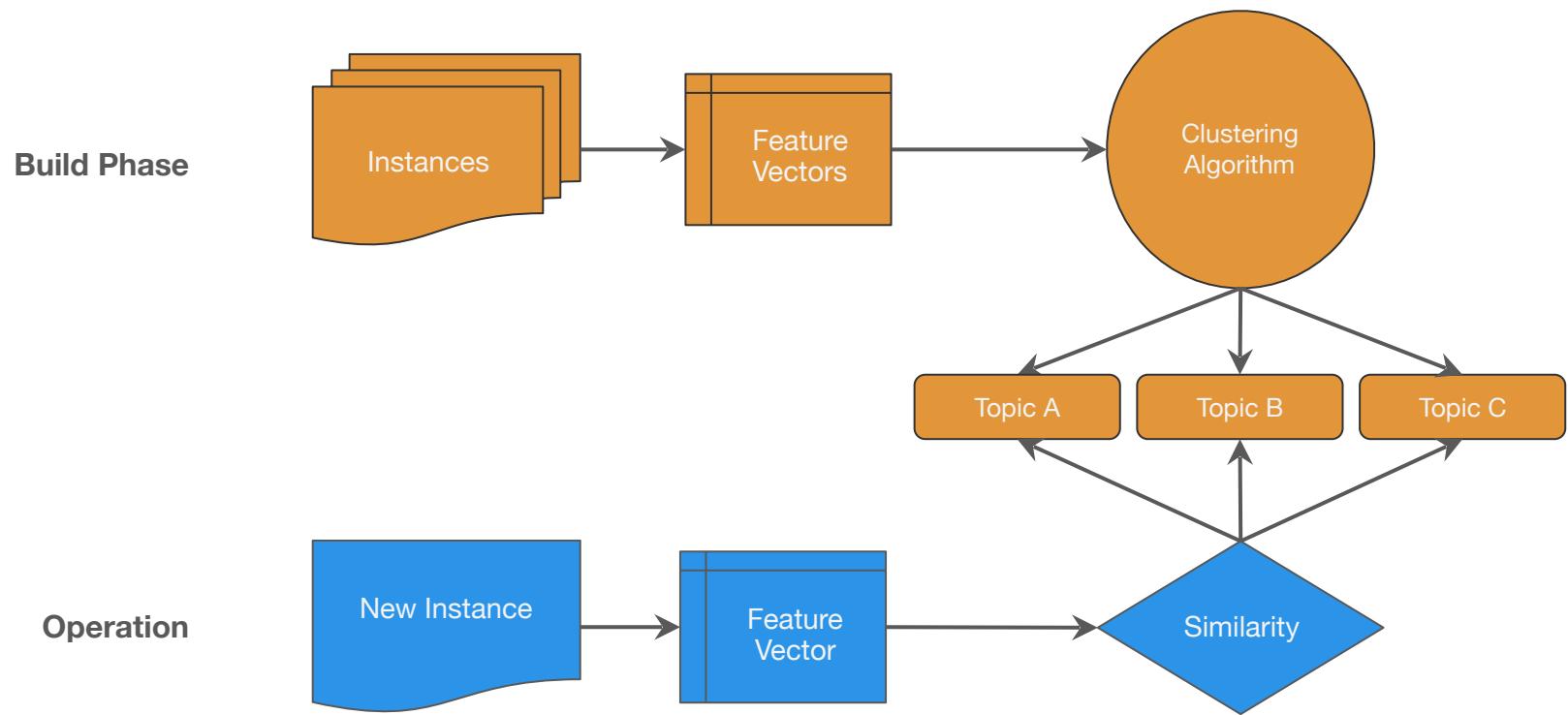
The Data Product Pipeline



Supervised Learning



Unsupervised Learning

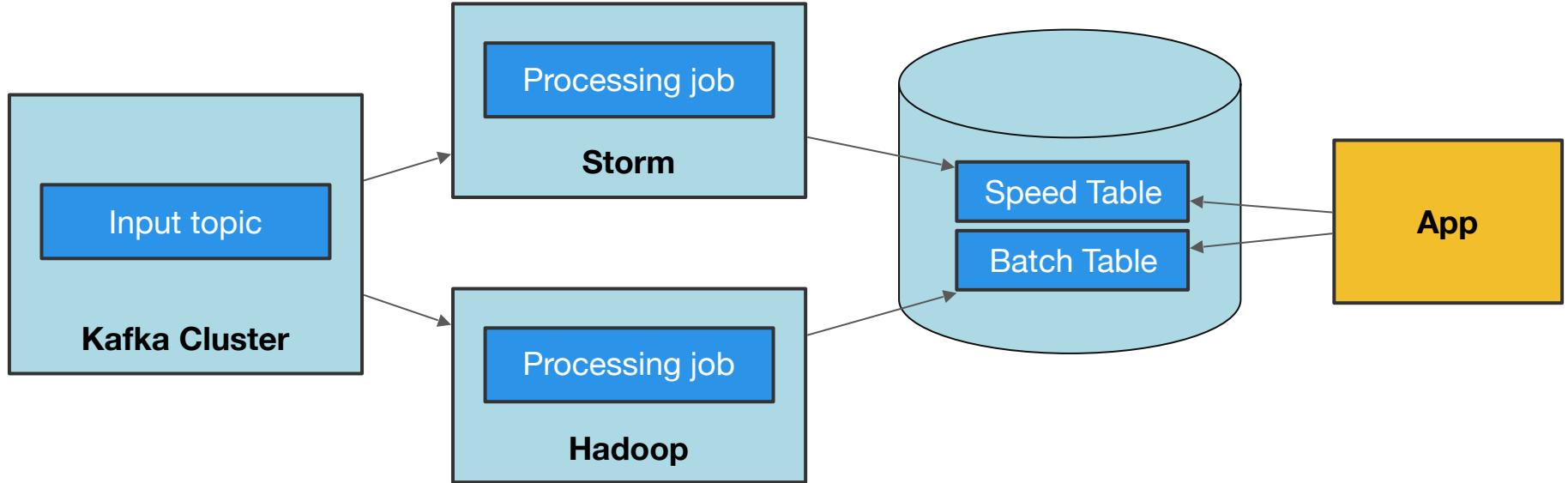




Data products aren't single models

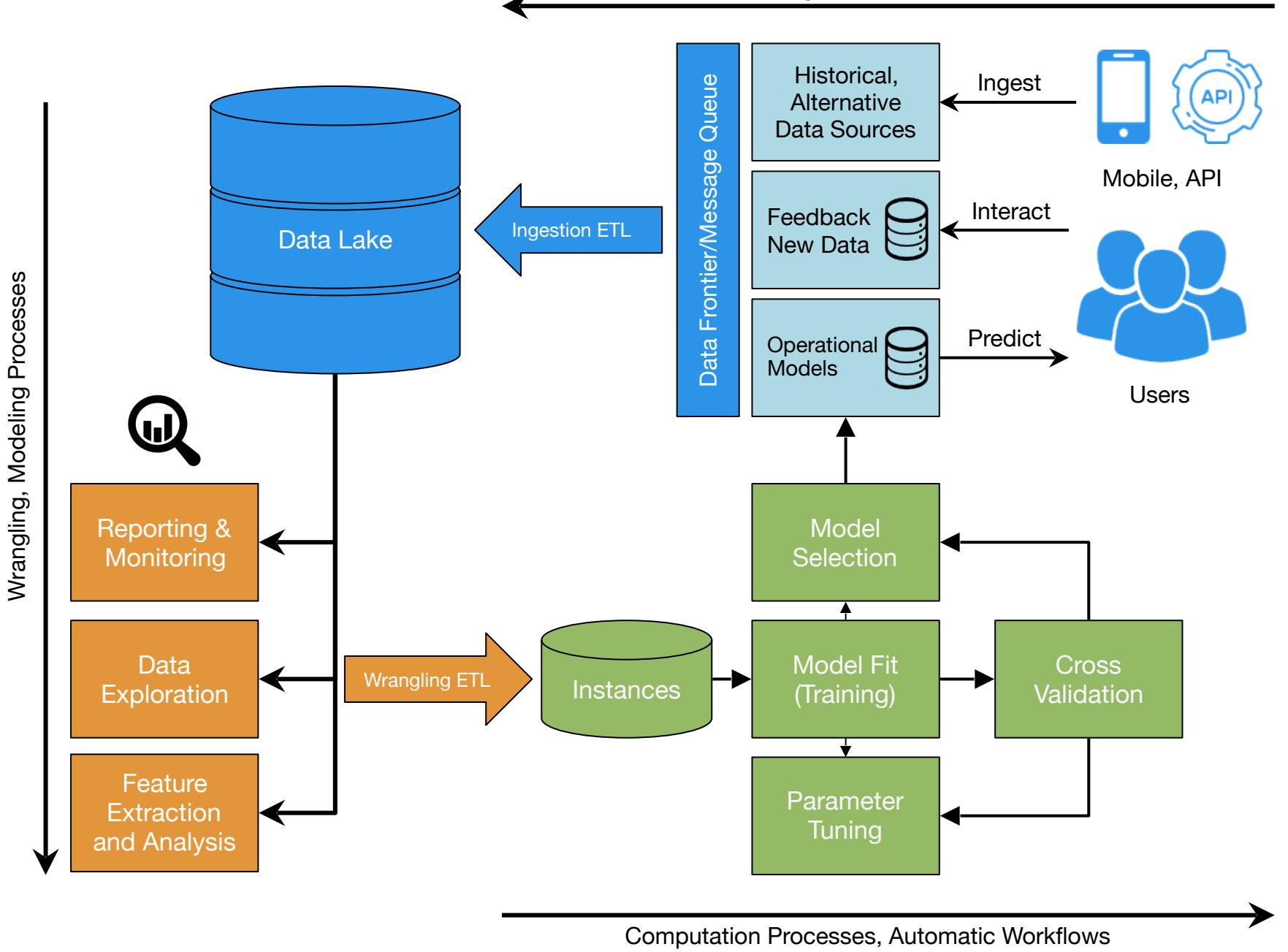
So how do we *architect* data
products?

The Lambda Architecture

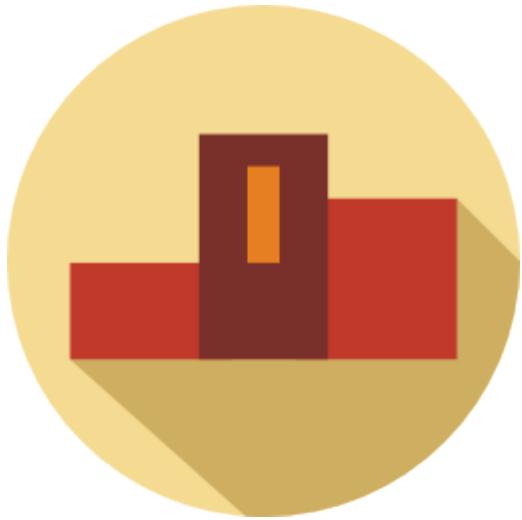


Questioning the Lambda Architecture -- Jay Kreps
<http://oreil.ly/29RK7dQ>

Wrangling, Modeling Processes



Computation Processes, Automatic Workflows



Semantic Search

Enabling and improving search results for a domain specific application.



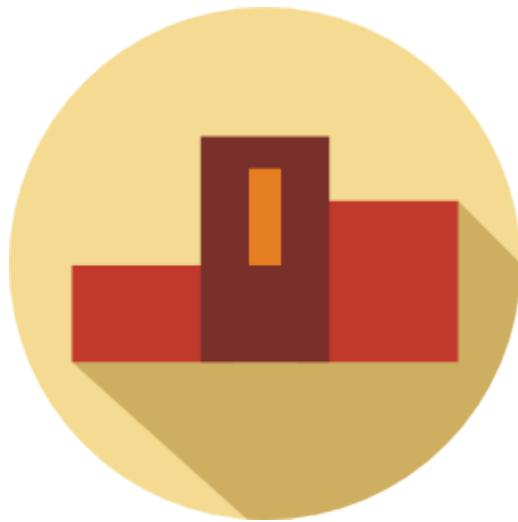
Recommendation

Creating collective models using collaborative filtering and graphs.



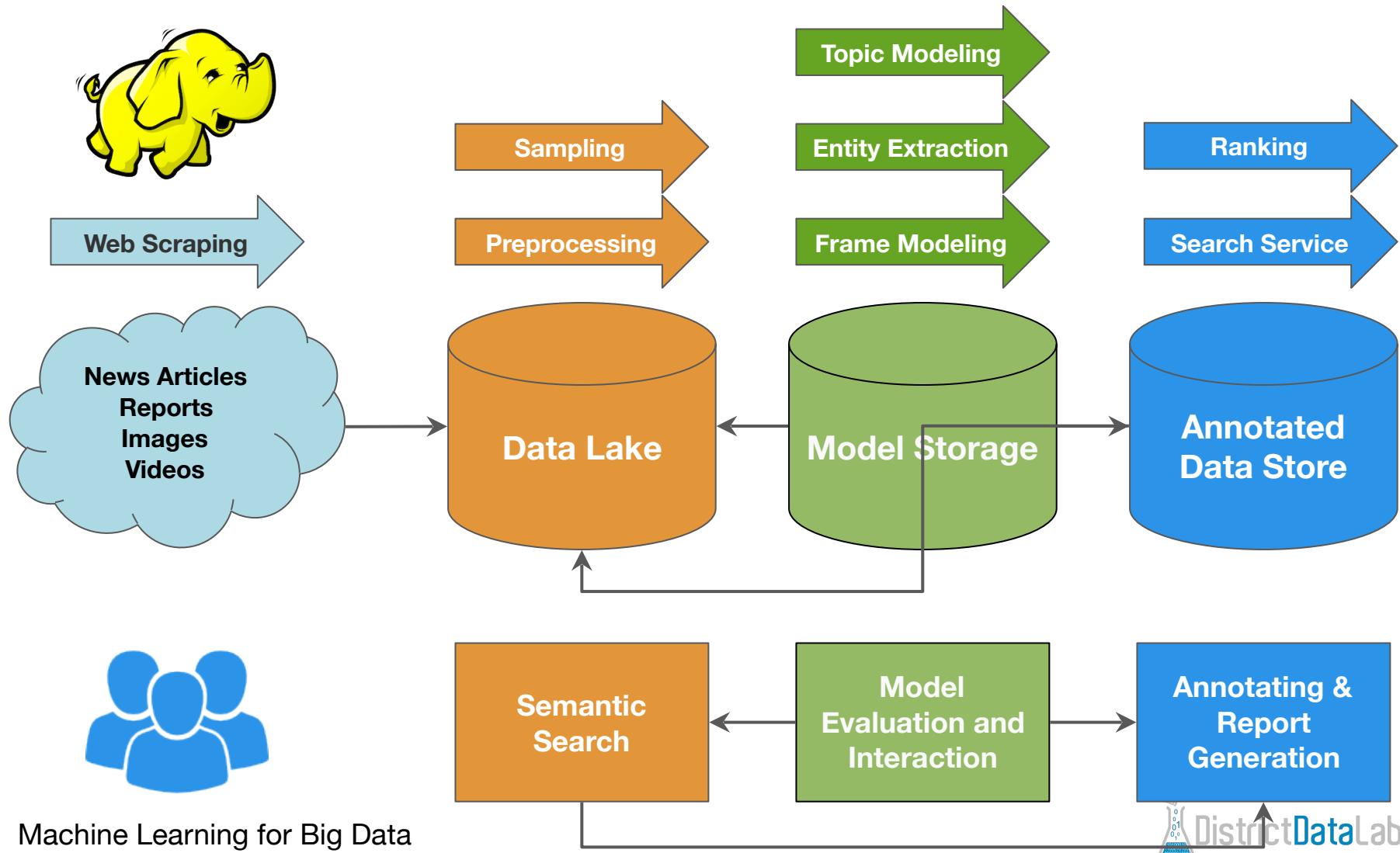
Political Classifier

Create per-user models as well as global models for dynamic interaction.



Semantic Search

Analyst Architecture





Analyst Architecture: Search

President Go!

Results

<http://reliefweb.int/disaster/tc-2013-000139-phl>
artifact: af2597893ef67406066d1c06633e5cd8744fe75cc11fcbb61f98a960082c8b
60 Relevant Sentences · 245 Extracted Frames **PERSON: Seleka** more ...

In Bangui, disarmament operations conducted by Seleka elements against reportedly former supporters of President Bozizé had resulted in widespread lootings of houses and indiscriminate attacks on civilians, according to international observers. ⓘ

In early August, President Djotodia had barred fighters from the Seleka coalition from participating in policing operations in Bangui and had declared that the task be left to the MISCA. ⓘ

<http://reliefweb.int/disaster/tc-2013-000139-phl>
artifact: aa4867e598de70ef5d9c71fbcb6fc9e4f1bb2acdfea37dbfb9011b39ba1c59529
70 Relevant Sentences · 242 Extracted Frames more ...

The political scene is also likely to be dominated by the planned trial of the president and vice-president at the International Criminal Court (ICC) in the coming year. ⓘ

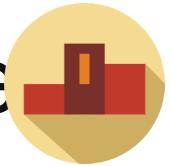
According to international observers, there are currently no favourites to succeed President Karzai, as the country is growing increasingly turbulent. ⓘ

<http://reliefweb.int/disaster/tc-2013-000139-phl>
artifact: 3367f49e97b426c7826981dbfa9ee617a2943a66639f0d7d28b0d739cfcb79c
37 Relevant Sentences · 65 Extracted Frames more ...

ANDRY NIRINA RAJOELINA, President of Madagascar, noting that new conflicts arose in the world each year demanding attention, urged the international community to analyse their root causes and take

Found 255 Relevant Documents

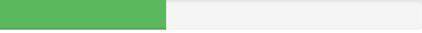
Analyst Architecture: Document Review



Reuters Corpus

The Reuters Corpus contains 10,788 news documents totaling 1.3 million words. The documents have been classified into 90 topics, and grouped into two sets, called "training" and "test"; thus, the text with fileid 'test/14826' is a document drawn from the test set. This split is for training and testing algorithms that automatically detect the topic of a document.

Documents	10,788
Annotations	8,320

A horizontal progress bar consisting of a green segment followed by a grey segment.

[Edit](#) [Upload](#) [New](#)

Available Tags

-  PERSON
-  ORGANIZATION
-  LOCATION

Frames Entities

BOWATER INDUSTRIES PROFIT EXCEED EXPECTATIONS

Bowater Industries Plc <BWTR.L> 1986 pretax profits of 48.0 mln stg exceeded market expectations of around 40 mln and pushed the company's shares up sharply to a high of 491p from 468p last night, dealers said. The shares later eased back to 481p. **Bowater** reported a 32.4 mln stg profit in 1985. The company said in a statement accompanying the results that the underlying trend showed improvement and it intended to expand further by developing existing businesses and seeking new opportunities. **ORGANIZATION** It added that it had appointed **David Lyon**, currently managing director of **Redland Plc** <RDLD.L> as its new chief executive. Analysts noted that **Bowater's** profits of 18.9 mln stg from 13.2 mln previously had been given a boost by pension benefits of 4.5 mln stg. Profit from **Australia** and the **Far East** showed the greatest percentage rise, jumping 55.0 pct to 15.5 mln from 10.0 mln, while the profit from **U.K.** Operations rose 30.7 pct to 24.7 mln, and **Europe**, 42.9 pct to 11.0 mln.

Reuters Corpus Document #14872 9

[← Prev](#) [Next →](#)

Analyst Architecture: Triggers



Trigger Events

Last Trigger:
16 Dec 2013 13:23 UTC

Data Trigger	1284 a
Time Trigger	8h 30m
User Defined Trigger A	90%
User Defined Trigger B	18%

Immediate Recalibrate

Statistical Models

Coreference Resolution	0 ⚙️	Last Updated: 24 Jan 2014 16:15 UTC
Frame Extraction	0 ⚙️	Last Updated: 24 Jan 2014 16:15 UTC
Entity Disambiguation	0 ⚙️	Last Updated: 24 Jan 2014 16:14 UTC
Named Entity Classifier	0 ⚙️	Last Updated: 24 Jan 2014 16:14 UTC
Test Model	3 ⚙️	Last Updated: 24 Jan 2014 16:13 UTC

Trigger Management

41: Benjamin Bengfort's Trigger (Jan 24 2014)	
40: Benjamin Bengfort's Trigger (Jan 24 2014)	
37: Benjamin Bengfort's Trigger (Jan 23 2014)	

Add Trigger

Job 00004 at 24 Jan 2014 16:28 UTC

Model: Automatic Frame Extraction

Less Specific More Specific

Result Info 6m 28s

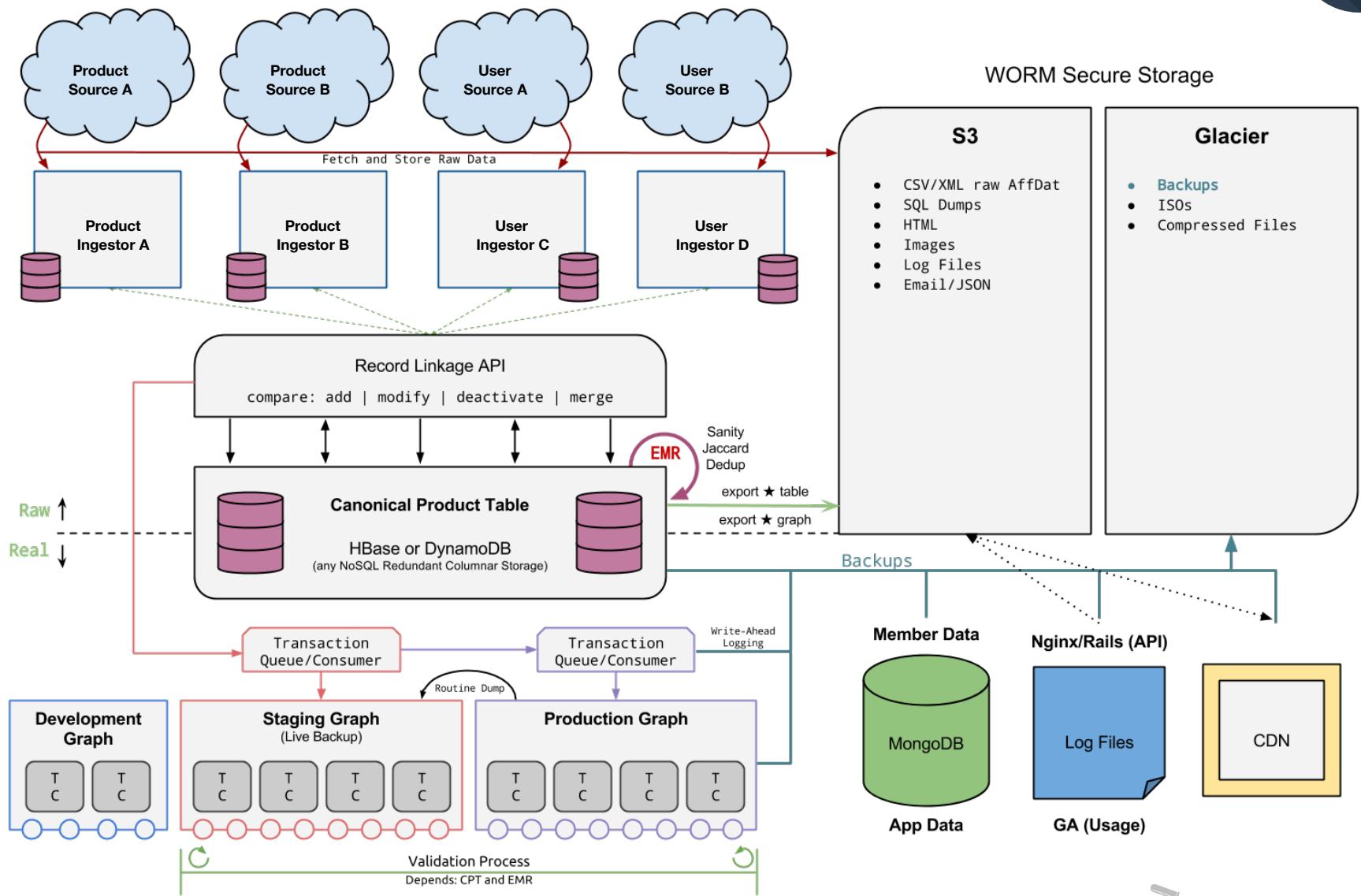
Recent Hadoop Jobs

Job 00004 at 24 Jan 2014 16:18 UTC	complete
Job 00003 at 24 Jan 2014 16:15 UTC	pending
Job 00002 at 24 Jan 2014 16:14 UTC	complete
Job 00001 at 24 Jan 2014 16:13 UTC	failed



Recommendation

Recommender Architecture



Recommender: Annotation Service



GUESS Self Made Printed T-Shirt



Get recognized with this GUESS t-shirt, printed with the phrase "Self Made" at the front.

Current Meta

Male Adult tops

Oberon Classification

Tops 97.858%

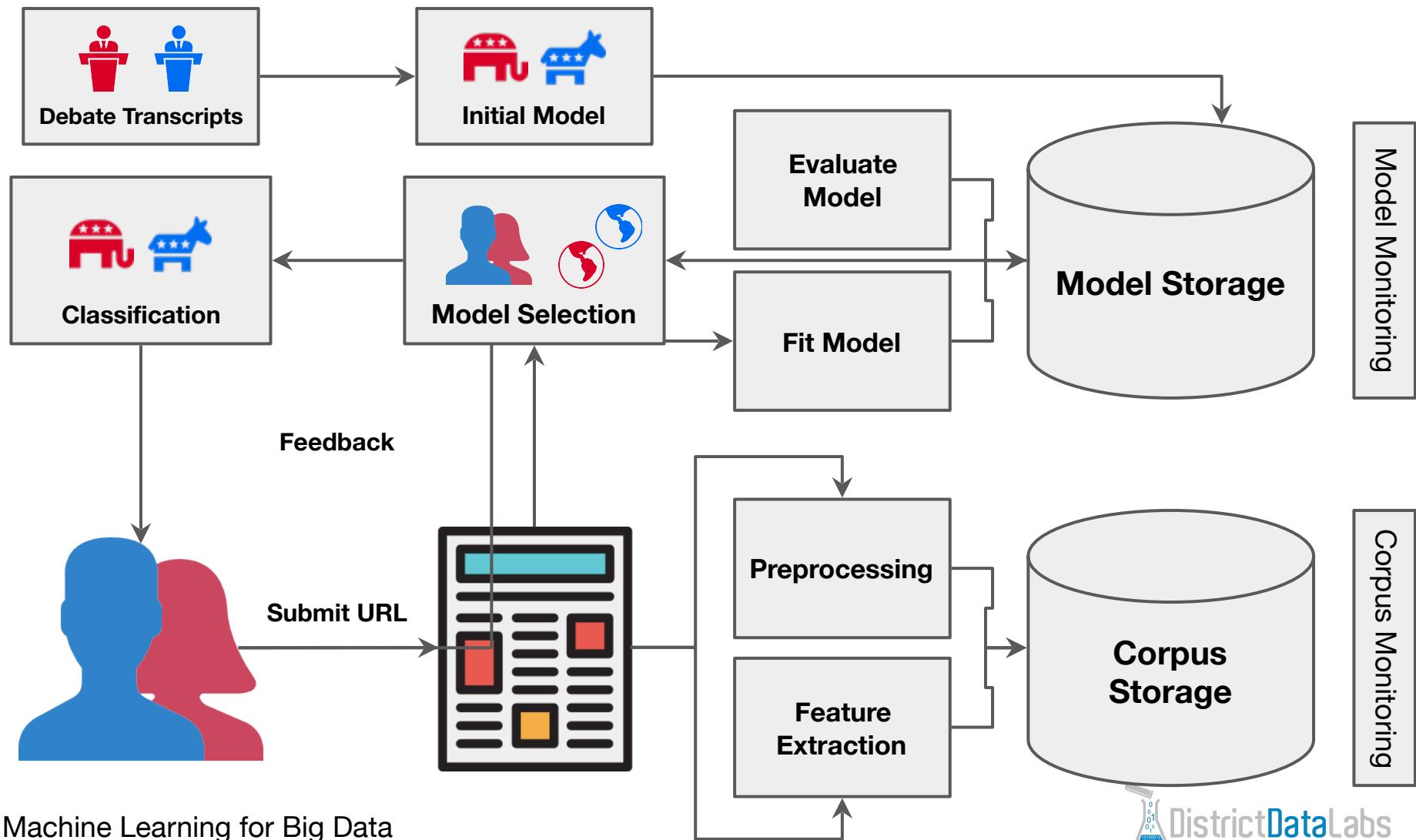
← Prev Right Wrong Next →

Benjamin up voted GUESS Self Made Printed T-Shirt



Political Classifier

Partisan Discourse Architecture



Partisan Discourse: Adding Document



Partisan Discourse beta

Benjamin Bengfort

Enter URL of webpage to classify ... Go!

Recent Activity

- [GOP brass rally party behind Trump after convention nomination | Fox News](#)
bbengfort annotated this document "Republican" on July 20, 2016
- [Republican National Convention: Scrutiny of Melania Trump's speech follows plagiarism allegations - The Washington Post](#)
bbengfort annotated this document "Democratic" on July 19, 2016
- [Trump panic drives progressives toward Clinton - POLITICO](#)
bbengfort annotated this document "Democratic" on July 19, 2016
- [Republican Convention: Newt Gingrich Pushed for Mike Pence](#)
bbengfort annotated this document "Republican" on July 19, 2016
- [Newt Gingrich: Donald Trump's Hot Wife Is Proof He Doesn't Hate Immigrants](#)
bbengfort annotated this document "Republican" on July 19, 2016
- [Jeb Bush: It's Obama's fault the GOP has become anti-gay - Metro Weekly](#)
bbengfort annotated this document "Republican" on July 19, 2016
- [A Conversation With Bernie Sanders Supporters : NPR](#)
rbilbro added this document on July 19, 2016
- [Bernie Sanders' Delegates: Photos & Stories of DNC Delegates | Heavy.com](#)
rbilbro added this document on July 19, 2016
- [Whither Sanders' 'revolution?' - Baltimore Sun](#)
rbilbro added this document on July 19, 2016



Partisan Discourse: Documents

Partisan Discourse beta

Benjamin Bengfort

Enter URL of webpage to classify ... Go!

GOP brass rally party behind Trump after convention nomination | Fox News

<http://fxn.ws/2a8l3MO> Word Count 879 Vocabulary 391

Republican Democratic

CLEVELAND – Republican congressional leaders , joined by vanquished primary candidates , immediately worked to rally the party behind Donald Trump Tuesday night after their national convention formally nominated him for president – with House Speaker Paul Ryan calling on voters to hit the polls like never before and “ see this thing through .”

“ Our candidates will be giving their all , they ’ ll be giving their utmost , and every one of us has got to go and do the same , ” Ryan said from the convention podium in Cleveland .

Night Two of the Republican convention contained plenty of rhetorical body slams against presumptive Democratic nominee Hillary Clinton . Chants from the crowd of “ lock her up ” were frequent . But the night was also an opportunity for the so - called GOP ‘ establishment ’ to make a very public show of unity and close the book on the raucous primary season .

Perhaps more than any other GOP leader on Capitol Hill , Ryan has had his share of scrapes with Trump over the nominee ’ s controversial remarks and tactics – but he closed his address Tuesday night with a call to action , saying , “ Only with Donald Trump and Mike Pence do we have a chance at a better way .”

“ Fellow Republicans , what we have begun here , let ’ s see this thing through , let ’ s win this thing , let ’ s show America our best and nothing less , ” Ryan said .

Retired neurosurgeon Ben Carson , a former primary rival turned supporter , later warned that the country “ may never recover ” from another Clinton presidency -- saying , “ I ’ m proud to support Donald Trump .”

“ Now is the time for us to rise up and take America back , ” Carson said .

The speeches came on the heels of Republicans formally nominating Trump for president , with Indiana Gov . Mike Pence named to the ticket as his running mate .

Trump afterward addressed the convention hall via video message , saying : “ This is a movement ... but we have to go all the way .”

Partisan Discourse: User Specific Model



Partisan Discourse beta

Benjamin Bengfort



Details

No biography added quite yet.

>Password Edit Profile

Your Recent Activity

 GOP brass rally party behind Trump after convention nomination | Fox News
bbengfort annotated this document "Republican" on July 20, 2016

 Republican National Convention: Scrutiny of Melania Trump's speech follows plagiarism allegations - The Washington Post
bbengfort annotated this document "Democratic" on July 19, 2016

 Trump panic drives progressives toward Clinton - POLITICO
bbengfort annotated this document "Democratic" on July 19, 2016

 Republican Convention: Newt Gingrich Pushed for Mike Pence
bbengfort annotated this document "Republican" on July 19, 2016

 Newt Gingrich: Donald Trump's Hot Wife Is Proof He Doesn't Hate Immigrants
bbengfort annotated this document "Republican" on July 19, 2016

 Jeb Bush: It's Obama's fault the GOP has become anti-gay - Metro Weekly
bbengfort annotated this document "Republican" on July 19, 2016

 On convention's opening night, Republicans appeal to their party's base - The Washington Post
bbengfort added this document on July 19, 2016

13 Documents 7 Republican 5 Democratic

Key Questions for Data Product Architectures



How do you detect deterioration?



When do you rebuild (triggers)?

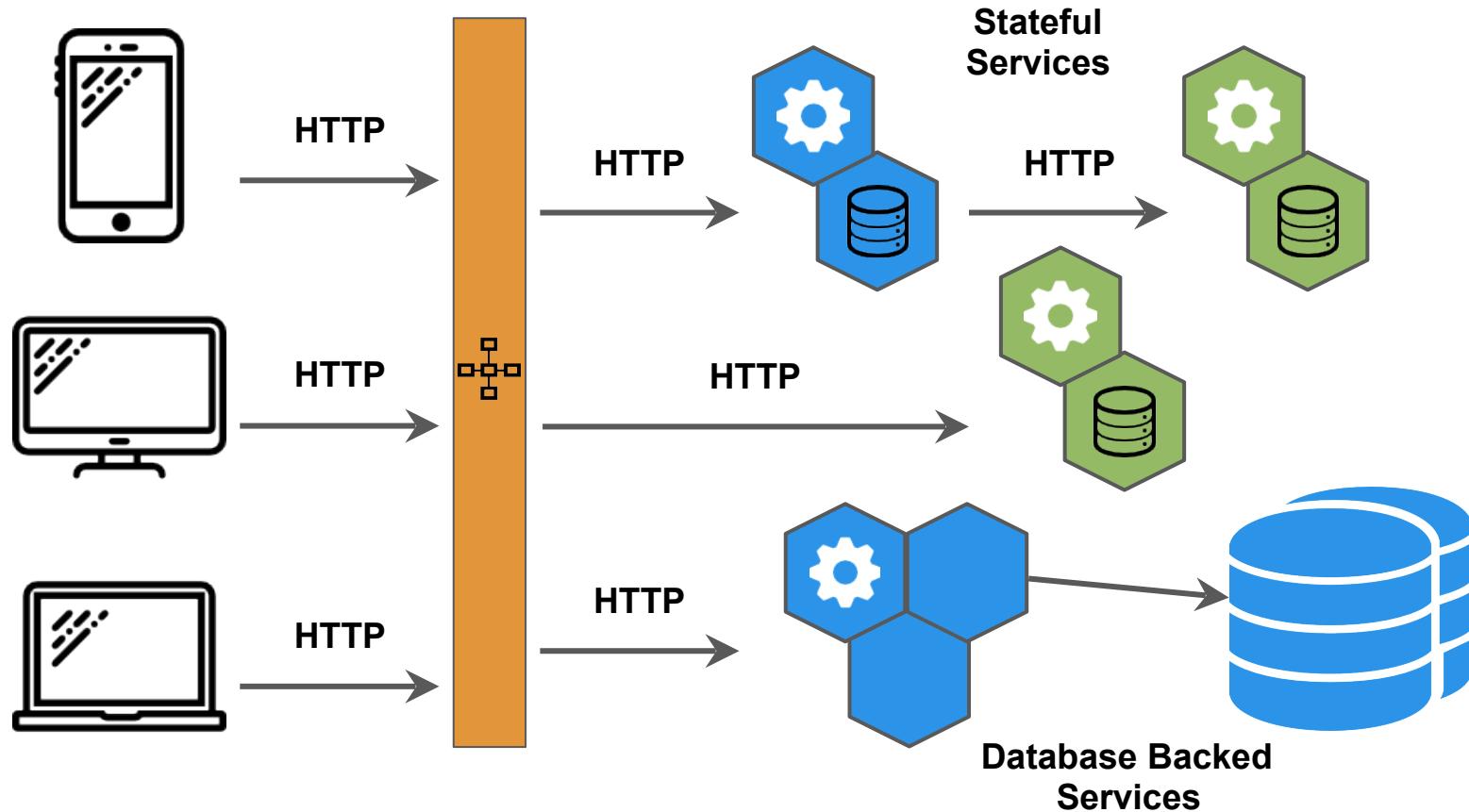


How to monitor change and decay old data?



How do you ensemble models to improve?

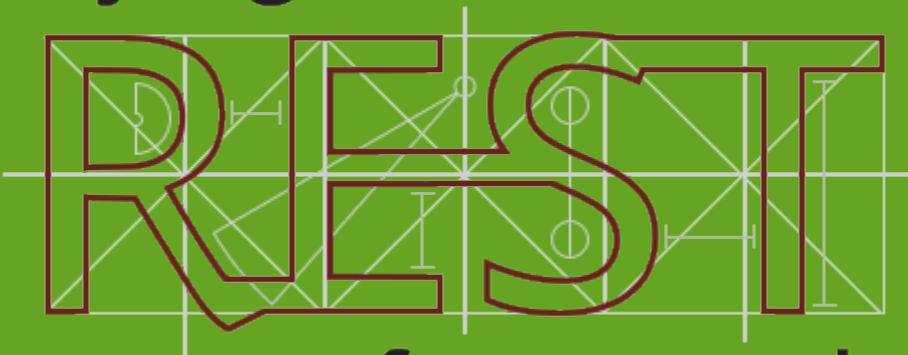
Microservices Architecture



Martin Fowler <http://bit.ly/2a2Lc17>

django

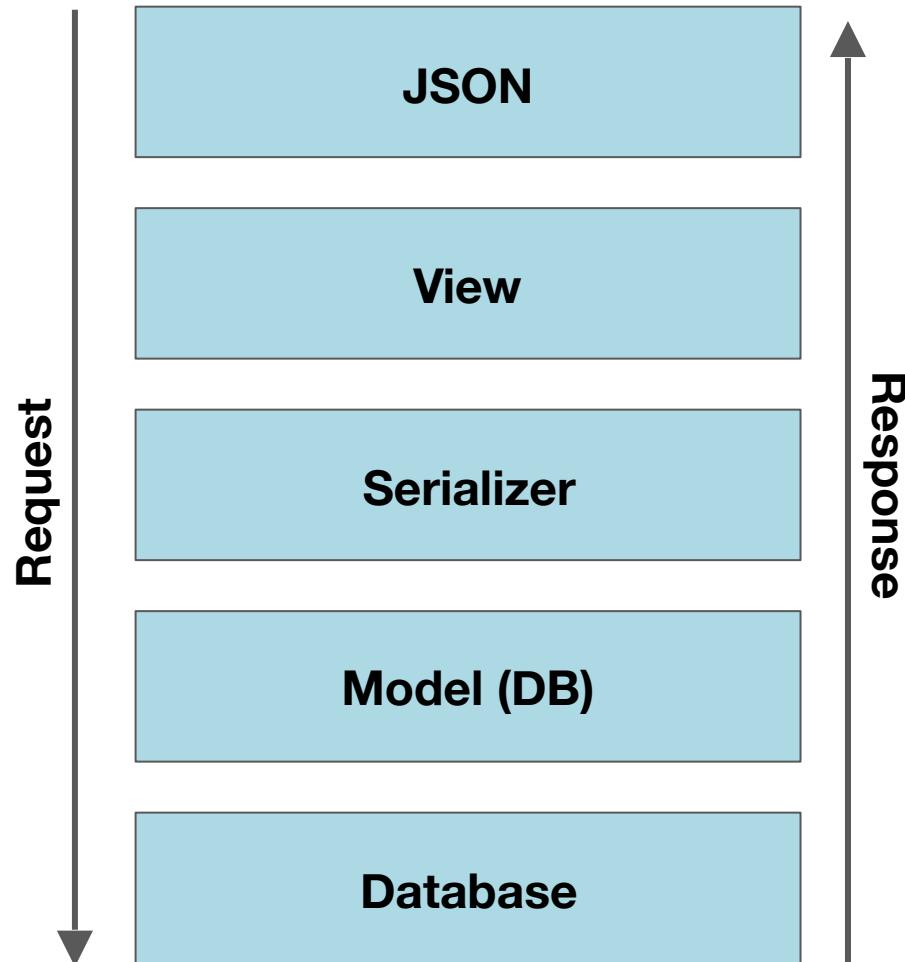
django



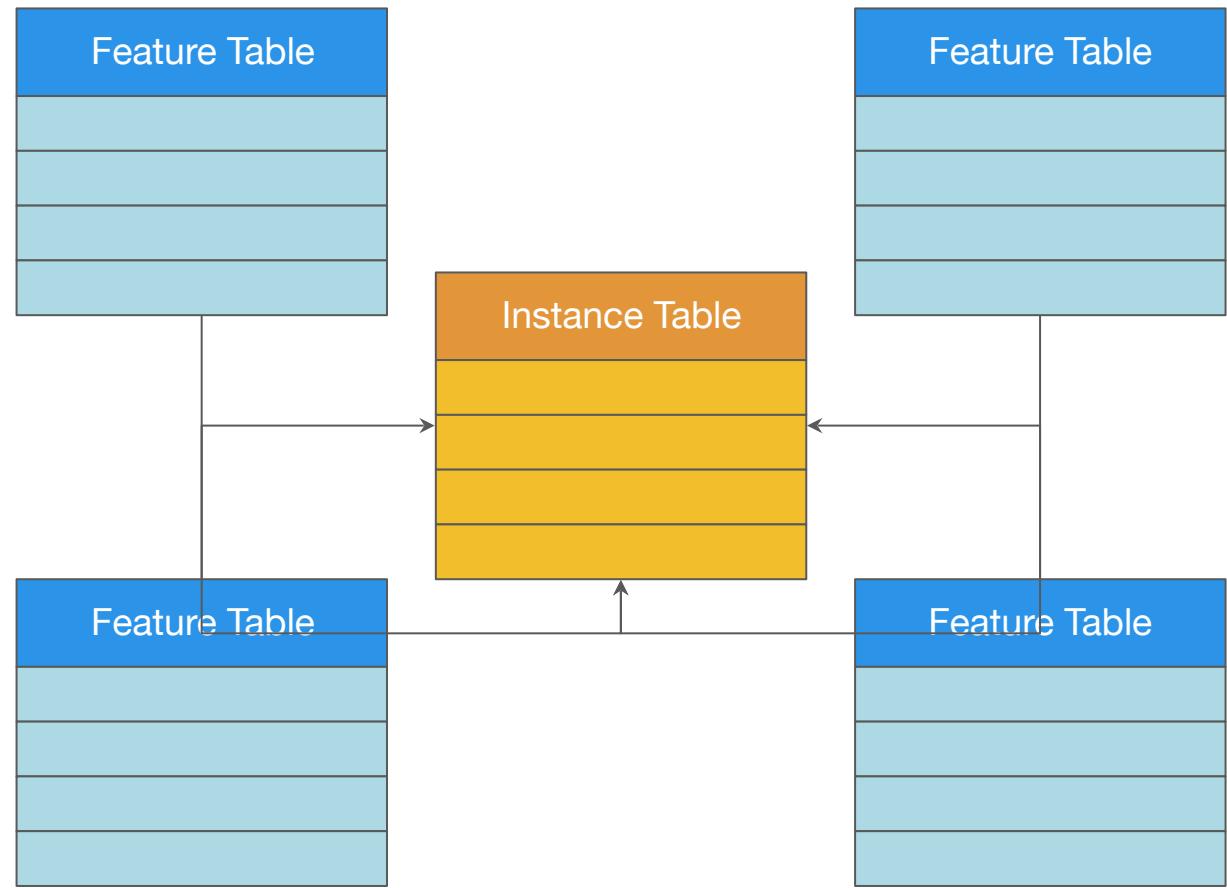
framework

Django Application Model

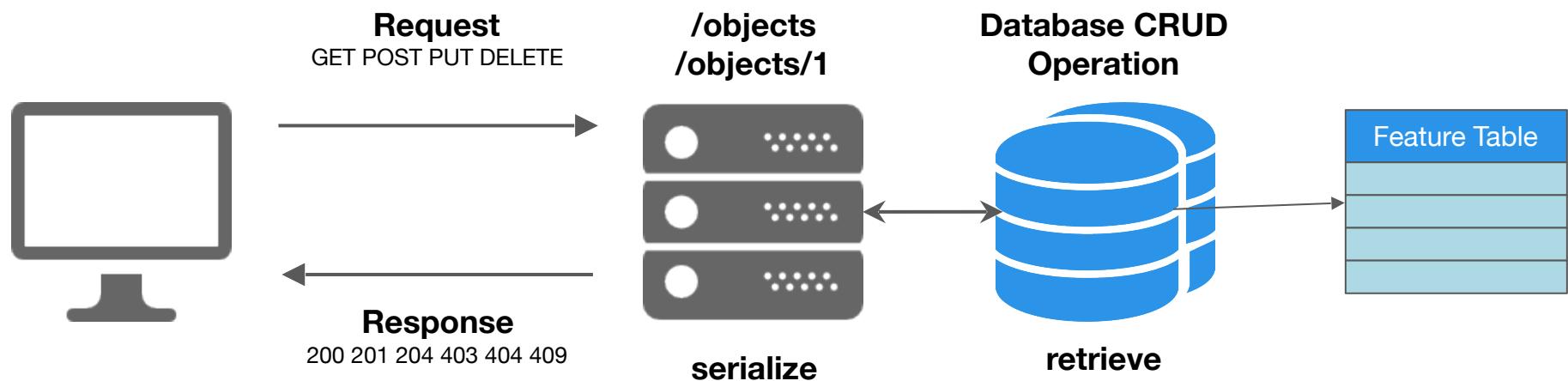
- Input and output of the application is JSON.
- Views are associated with endpoints and handle requests and responses.
- Serializers deal with the JSON database
- Models interact with the database.
- Almost any database backend is supported.



Features and Instances as Star Schema



REST API Feature Interaction



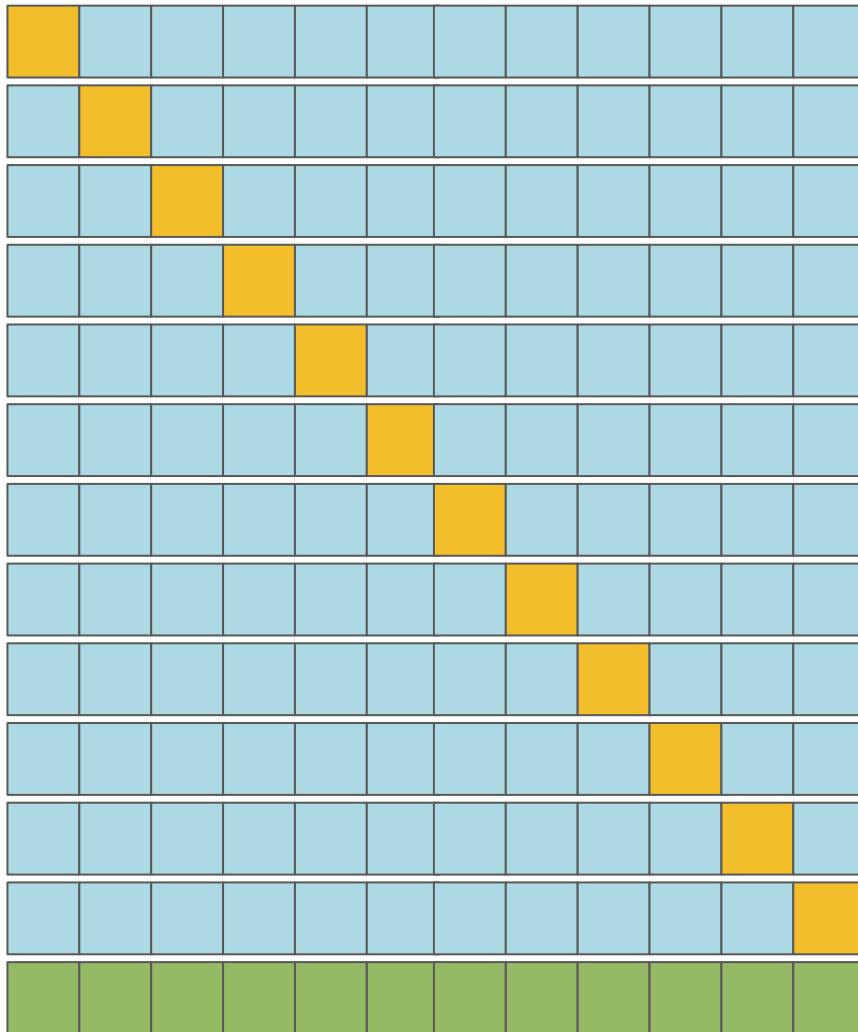
Model (ML) Build Process: Export Instance Table



```
COPY (
    SELECT instances.* FROM instances
        JOIN feature on feature.id =
instances.id
        ...
        ORDER BY instance.created DESC
        LIMIT 10000
    ) as instances
TO '/tmp/instances.csv' DELIMITER ',' CSV
HEADER;
```



Model (ML) Build Process: Build Model



```
import pandas as pd
from sklearn.svm import SVC
from sklearn.cross_validation import KFold

# Load Data
data = pd.read_csv('/tmp/
instances.csv')
scores = []

# Evaluation
folds = KFold(n=len(data), n_folds=12)
for train, test in folds:
    model = SVC()
    model.fit(data[train])
    score = model.score(data[test])
    scores.append(score)

# Build the actual model
model = SVC()
model.fit(data)
```

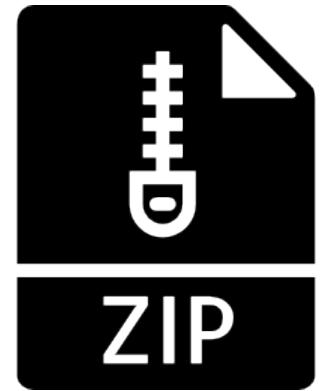
Model (ML) Build Process: Store Model



```
import json
import pickle
import base64
import datetime

data = pickle.dump(model)
data = base64.b64encode(data)

return {
    "model": data,
    "created": datetime.datetime.now(),
    "form": repr(model),
    "name": model.__class__.__name__,
    "scores": scores,
}
```



Model Data Storage

```
from django.db import models

class PredictiveModel(models.Model):

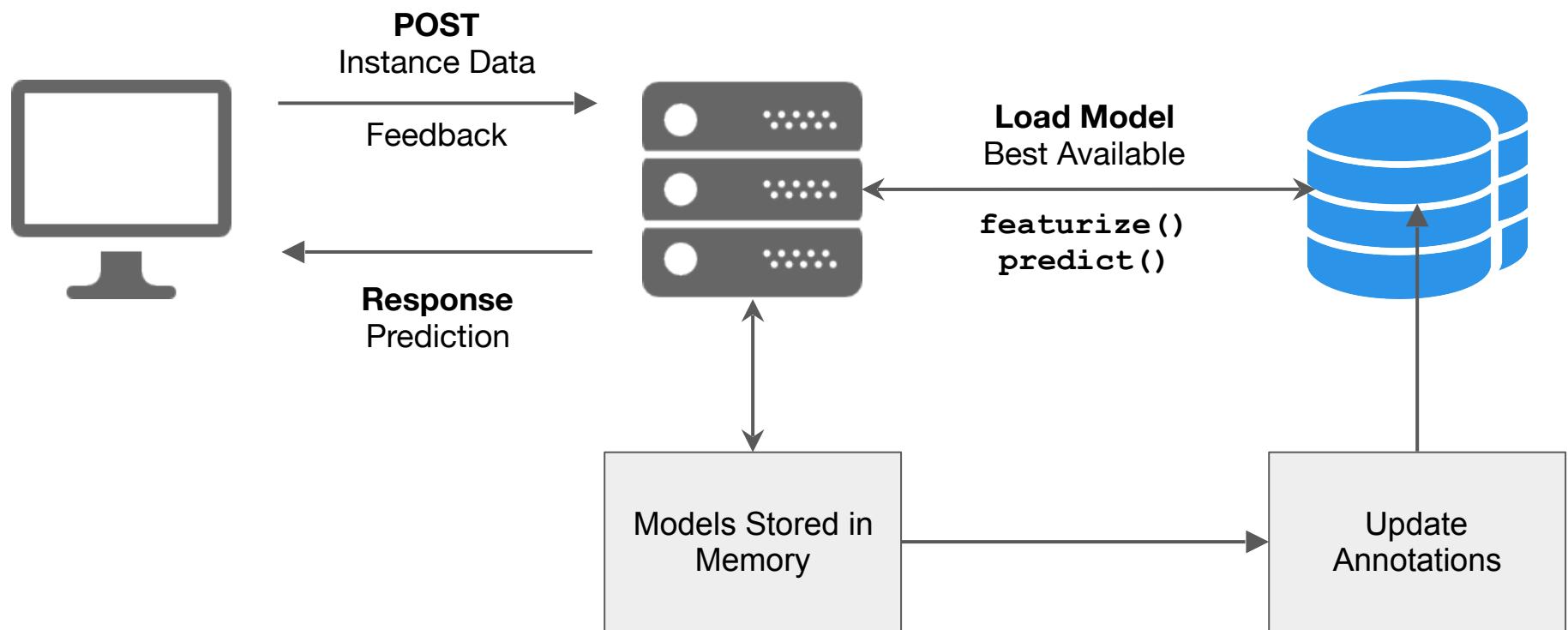
    name      = models.CharField()
    params    = models.JSONField()
    build     = models.FloatField()
    f1_score  = models.FloatField()
    created   = models.DateTimeField()
    data      = models.BinaryField()
```

Considerations

- FileField vs. BinaryField
- How much meta data can you store (the more the better)
- What fields can be monitored in an model selection management system? (E.g. CMS for ML)

ID	Model	Hyperparameters	Build Time	F1	Pickle
1	Naive Bayes	{"alpha": 1.0}	235.32	.832	BLOB
2	SVC	{"C": 1.0, "kernel": "linear"}	20.312	.861	BLOB
3	KNN	{"k": 5, "weights": "distance"}	482.129	.821	BLOB

REST API Model Interaction



Hands-On Lab

Tasks

Log in to the server.

Create a folder with your personal ID

I'll copy the code into the folders for you.

Let's view and run the notebooks:

1. Easy as Pi
2. Load a CSV File
3. Word Count

Hadoop: An Operating System for Big Data

Hadoop Architecture

Distributed System Requirements

Fault Tolerance - If a component fails, it should not result in the failure of the entire system.

Recoverability - In the event of a failure, no data should be lost.

Consistency - The failure of one job or task should not affect the final result.

Scalability - Adding load leads to a decline in performance, not failure

Hadoop Architecture - Basic Concepts

Data distributed immediately and stored on multiple nodes.

Stored in blocks of fixed size (128MB) and each block duplicated multiple times.

Computation referred to as a job, and jobs broken up into tasks.

Jobs are written at high level without concern for network programming, time, or low-level infrastructure.

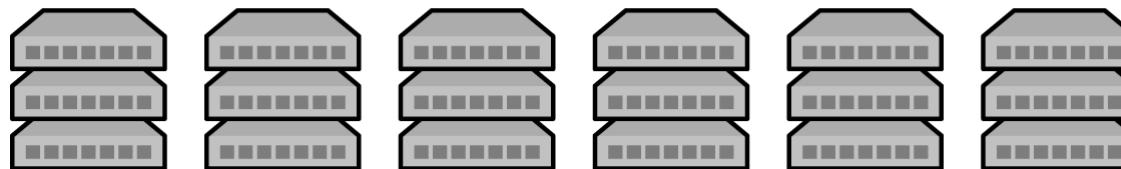
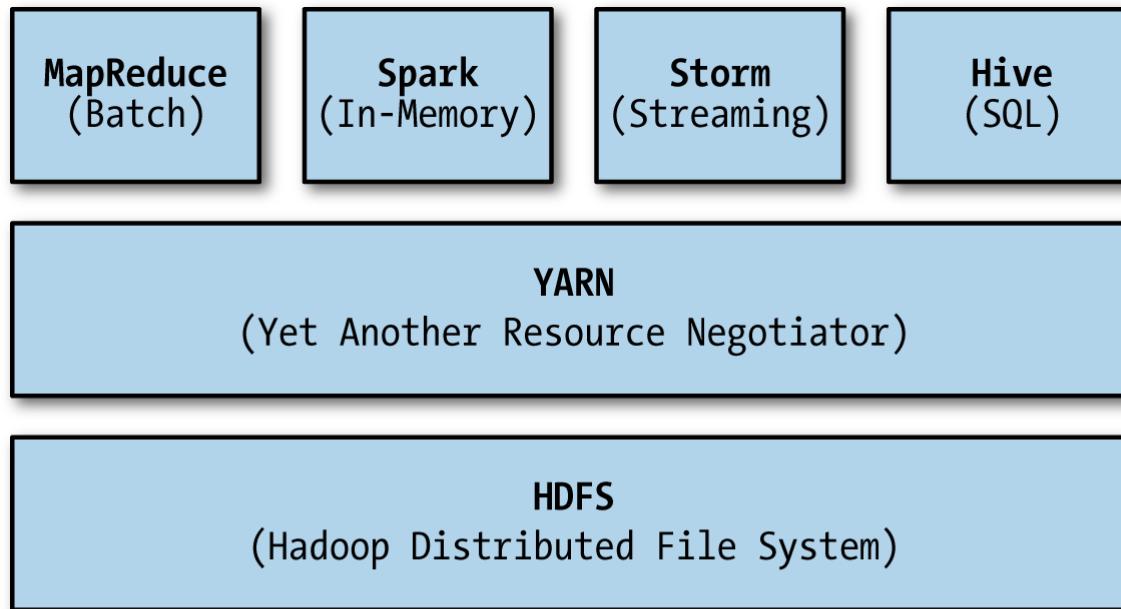
Hadoop Architecture - Basic Concepts

Amount of network traffic should be minimized transparently by the system.

Jobs are fault tolerant, usually through task redundancy.

Master programs allocate work to worker nodes such that many worker nodes can operate in parallel.

Hadoop Architecture



Cluster of economy disks and processors

What's In a Cluster?

What is a Cluster?

Set of machines running HDFS and YARN.

Individual horizontally scalable machines called **nodes**.

HDFS and YARN implemented by **daemon** processes.

Daemons run inside Java Virtual Machine (JVM).

Each mode identified by type of process it runs.

Types of Nodes

Master Nodes

Coordinate services for Hadoop workers.
Are usually the entry point for user access.

Worker Nodes

Most computers in the cluster.
Run services that accept tasks from Master Nodes.
Implement both HDFS and YARN worker services.

HDFS Services

NameNode (Master) - Stores directory tree, file metadata, and file locations.

Secondary NameNode (Master) - Performs housekeeping and checkpointing on behalf of NameNode.

DataNode (Worker) - Stores and manages HDFS blocks on the local disk and reports health and status to NameNode.

YARN Services

Resource Manager (Master) - Allocates and monitors available cluster resources and schedules jobs.

Application Manager (Master) - Coordinates applications being run on the cluster.

NodeManager (Worker) - Runs and manages processing tasks and reports health and status of tasks.

MapReduce vs. YARN

MapReduce

Very efficient for large-scale batch workloads, but also I/O intensive.

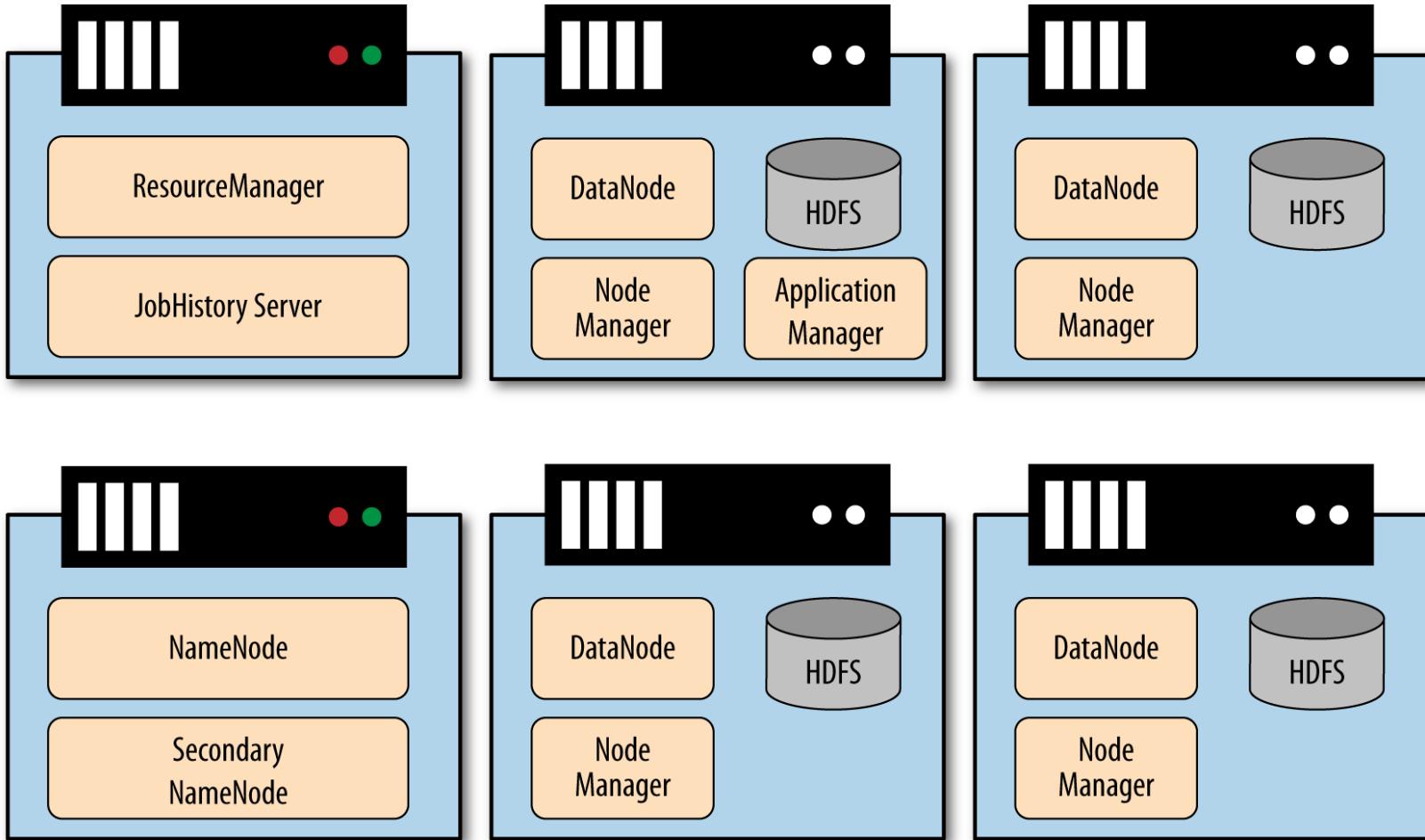
Limitations for interactive analysis, graph processing, machine learning, and other memory-intensive algorithms.

YARN

Decouples workload management from resource management.

Multiple applications can share centralized, common resource management service.

Cluster Example



HDFS Caveats

HDFS Caveats

Performs with modest number of very large files.

Implements the WORM (write once, read many) pattern.

Optimized for large, streaming file reads.

Best suited for storing raw input, intermediary results, and final results.

Not a good fit as a data backend for applications.

Blocks

HDFS files are split into blocks.

Typically 64MB, 128MB, or 256MB.

Allow large files to be distributed across many machines.

Replicated across DataNodes (3-fold by default)

Data Management

Master NameNode tracks
What blocks make up a file.
Where blocks are located.

Secondary NameNode merges snapshot of current data space with edit log.

When client application wants to read a file:
Requests metadata from NameNode to locate blocks.
Requests locations of DataNodes that store blocks.