NLP with MLlib: Global Empire-Building for Fun & Profit

Michelle Casbon
June 7, 2017
Spark Summit
San Francisco

Qordoba

`whoami` Michelle Casbon

- Where I work
    - Qordoba, Director of Data Science
- What I've done
- What I love
    - Natural language processing
    - Distributed systems
    - Emoji One

@texasmichelle

Data Science Engineer

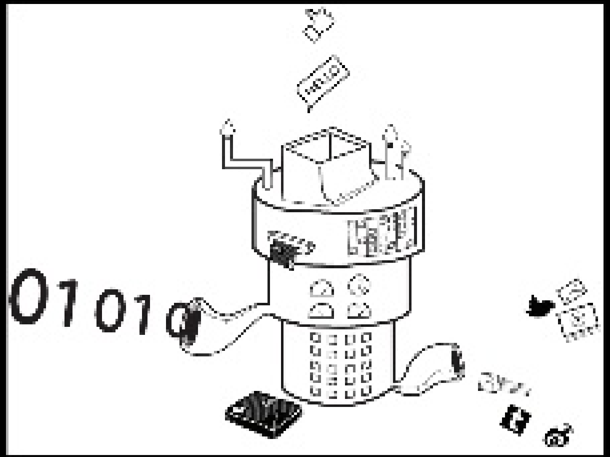
What my friends think I do

What my parents think I do

What society thinks I do

What my boss thinks I do

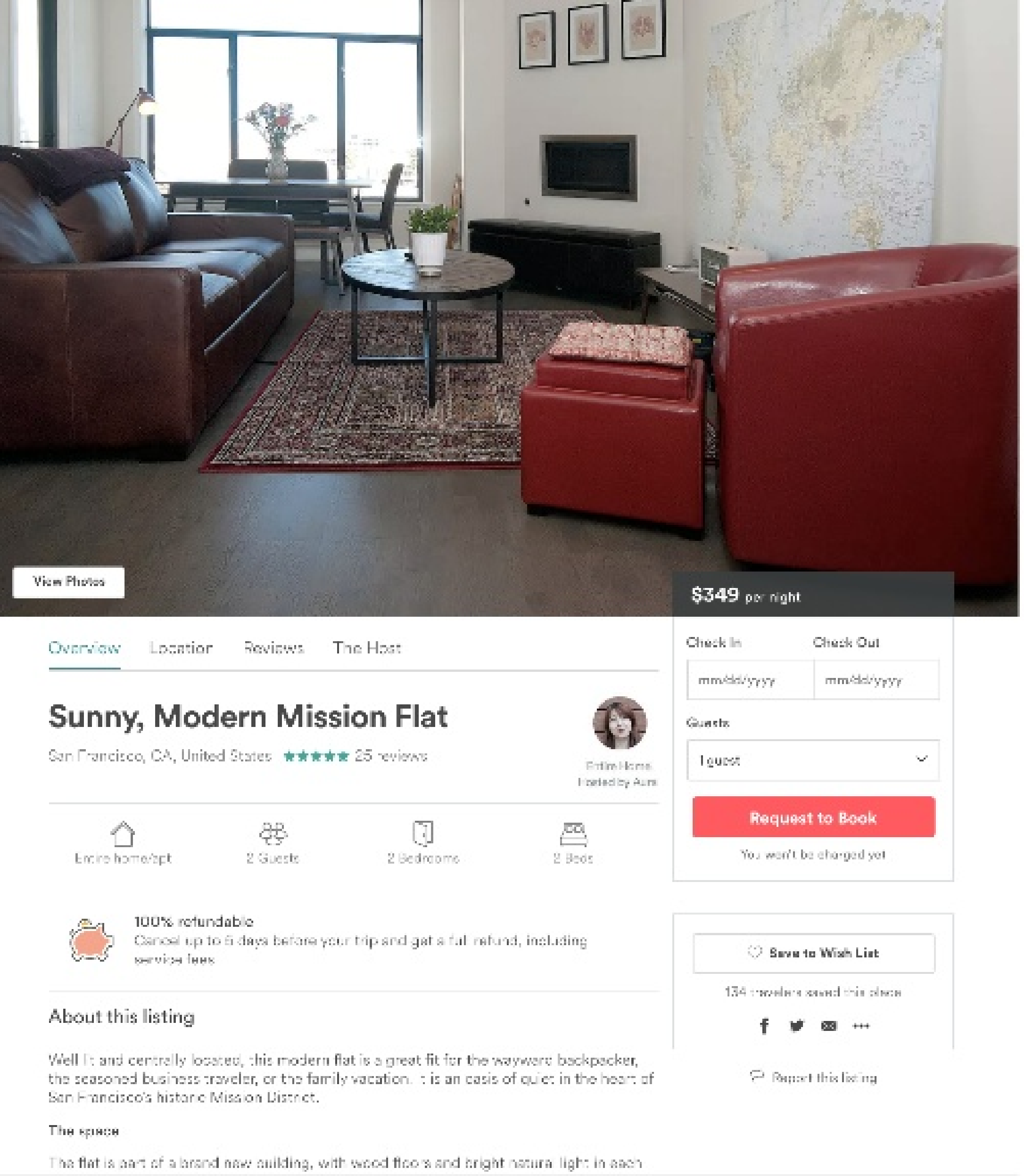
What I think I do



What I actually do

TL;DR

- Importance of localization
  - Why it’s hard
  - How it can be made easier
  - Examples
- The role of NLP in solving this problem
  - Brief introduction to NLP
- Spark’s role in NLP
- Infrastructure
- Deployment



$349 per night

Overview Location Reviews The Host

Check In Check Out

mm/dd/yyyy mm/dd/yyyy

Guests

1 guest

Request to Book

You won't be charged yet

Sunny, Modern Mission Flat

San Francisco, CA, United States ★★★★★ 25 reviews

Entire Home
Hosted by Aura

Entire home/apt 2 Guests 2 Bedrooms 2 Beds

100% refundable
Cancel up to 5 days before your trip and get a full refund, including service fees

Save to Wish List

134 travelers saved this place

Report this listing

About this listing

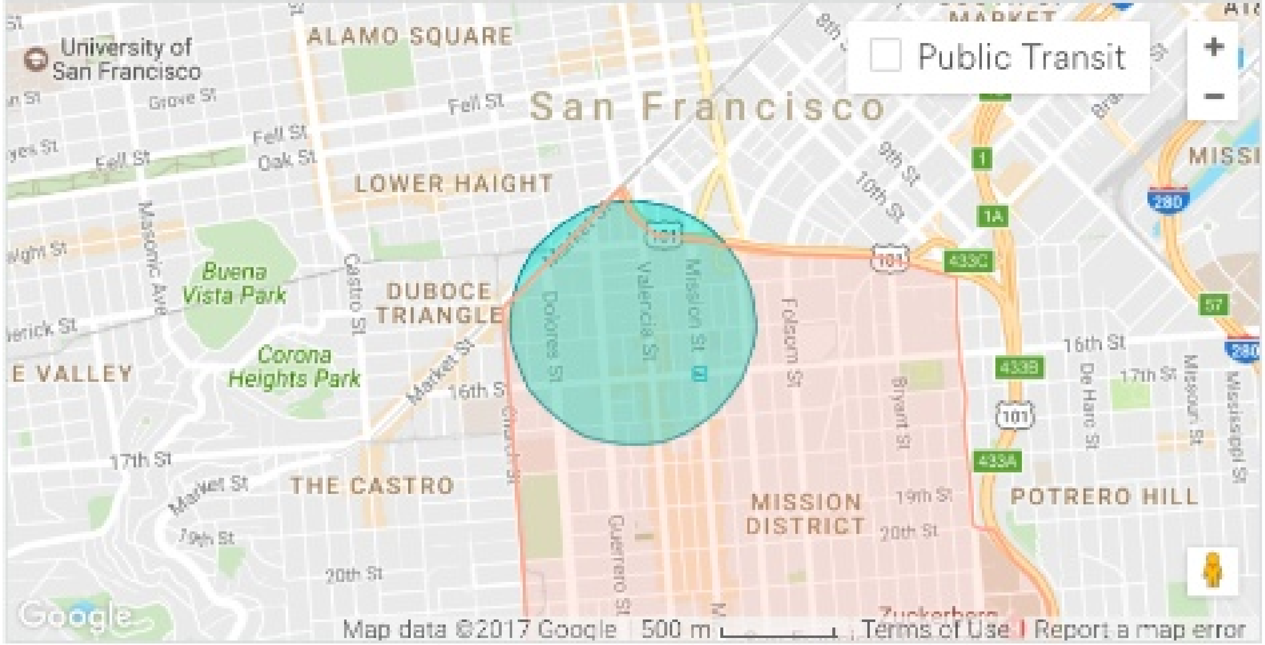
Well lit and centrally located, this modern flat is a great fit for the wayward backpacker, the seasoned business traveler, or the family vacation. It is an oasis of quiet in the heart of San Francisco's historic Mission District.

The space

The flat is part of a brand new building, with wood floors and bright natural light in each

This apartment is in a killer location

Diese Wohnung befindet sich an einem mörderischen Standort



Exact location information is provided after a booking is confirmed.

What is Qordoba?

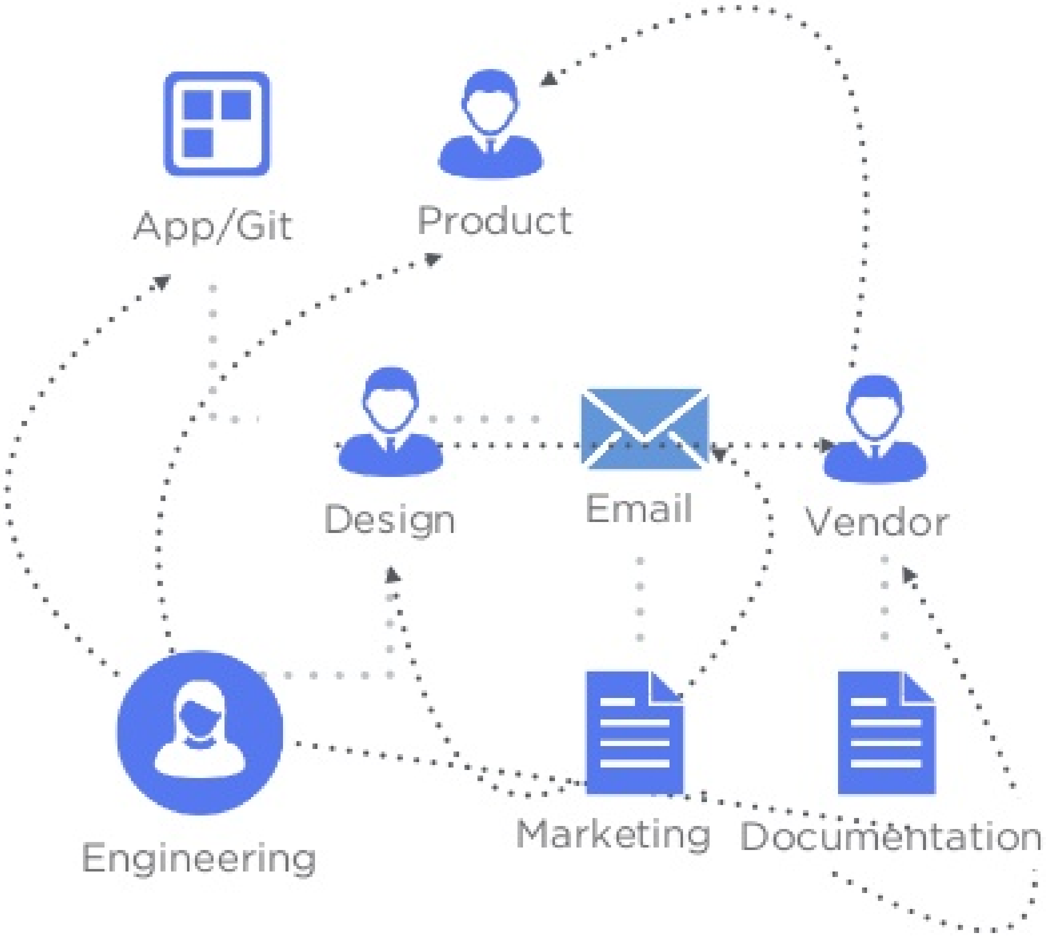
Every product in every language, by default

Fast, scalable way to go from one market to many

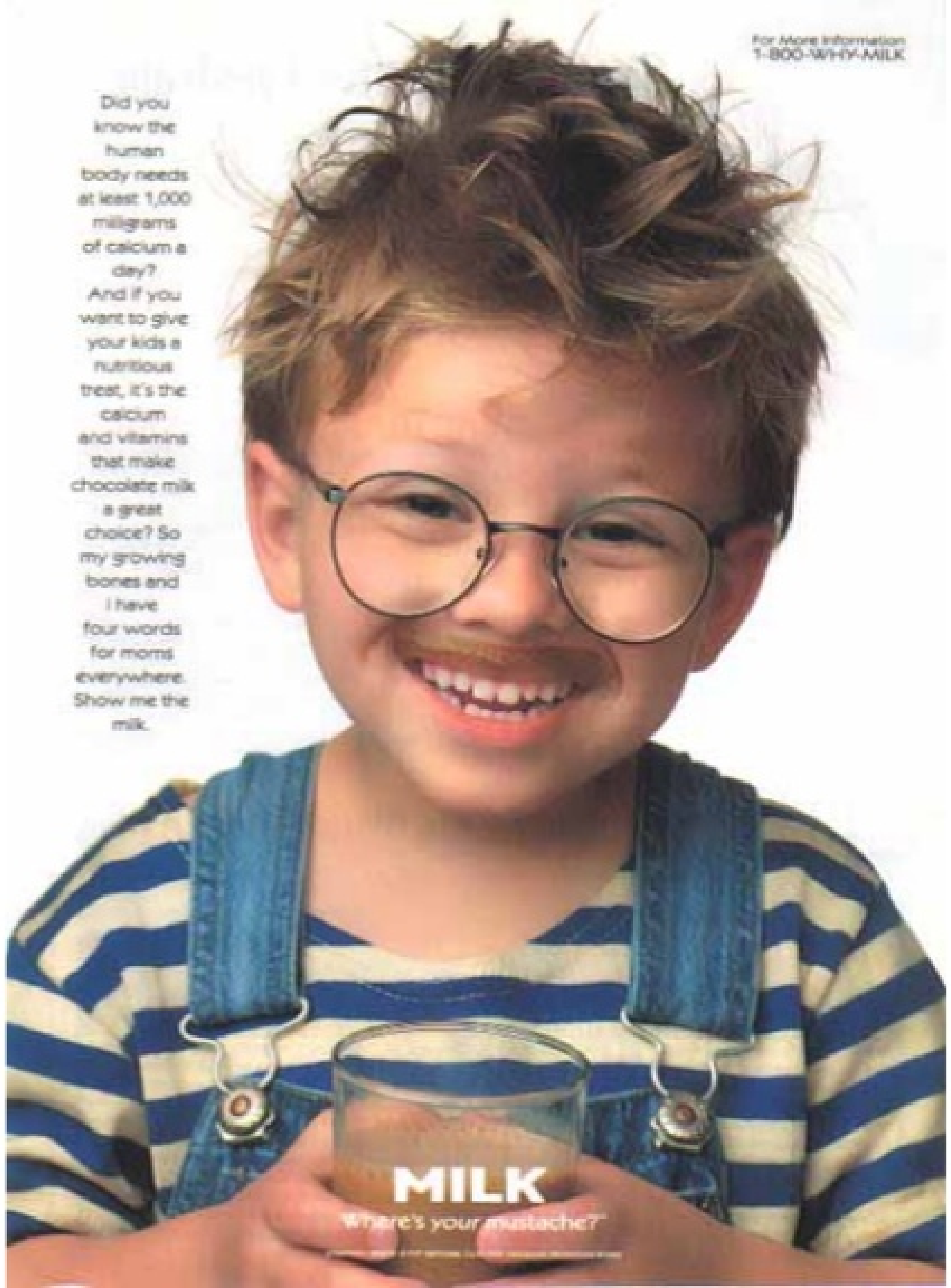
Together, our team has lived & worked in 34 countries

Localization is hard

- People
  - Product managers
  - Marketers
  - Designers
  - Linguists
  - Engineers
- Things
  - Copies of copy
  - String files
  - Emails
  - Pull requests
- Wash, rinse, repeat



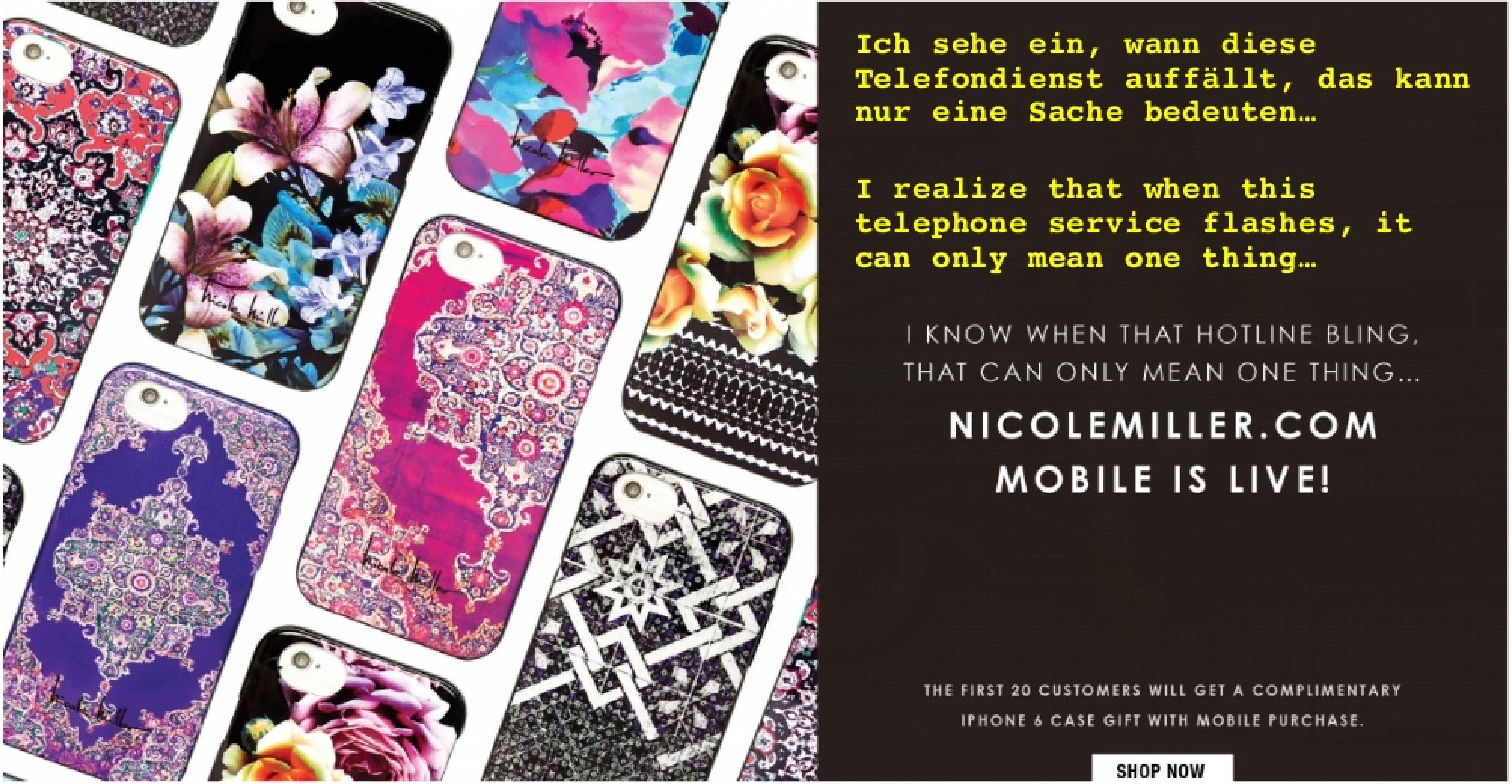
@texasmichelle



For More Information
1-800-WHY-MILK

Did you know the human body needs at least 1,000 milligrams of calcium a day? And if you want to give your kids a nutritious treat, it's the calcium and vitamins that make chocolate milk a great choice? So my growing bones and I have four words for moms everywhere. Show me the milk.

got milk?®



Ich sehe ein, wann diese Telefondienst auffällt, das kann nur eine Sache bedeuten…

I realize that when this telephone service flashes, it can only mean one thing…

I KNOW WHEN THAT HOTLINE BLING,
THAT CAN ONLY MEAN ONE THING...

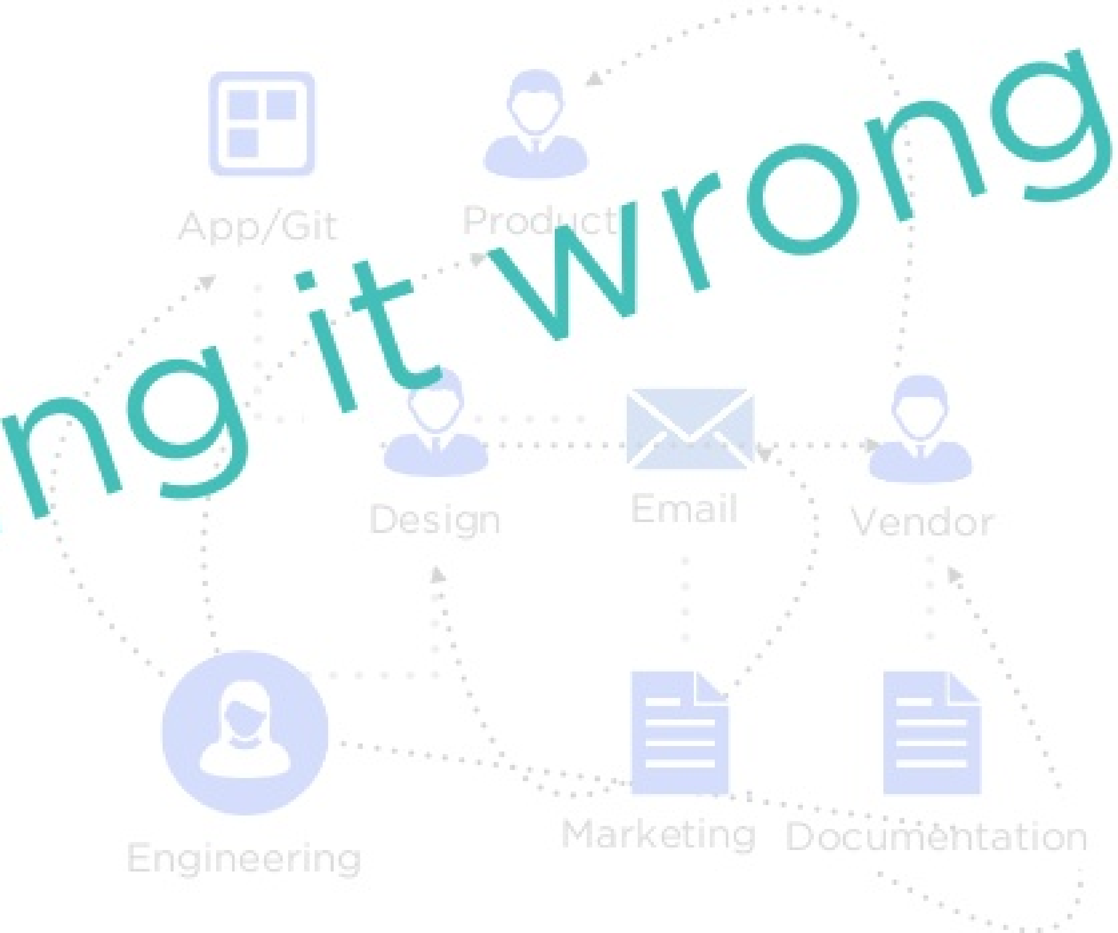
NICOLEMILLER.COM
MOBILE IS LIVE!

THE FIRST 20 CUSTOMERS WILL GET A COMPLIMENTARY
IPHONE 6 CASE GIFT WITH MOBILE PURCHASE.

SHOP NOW

Localization is hard

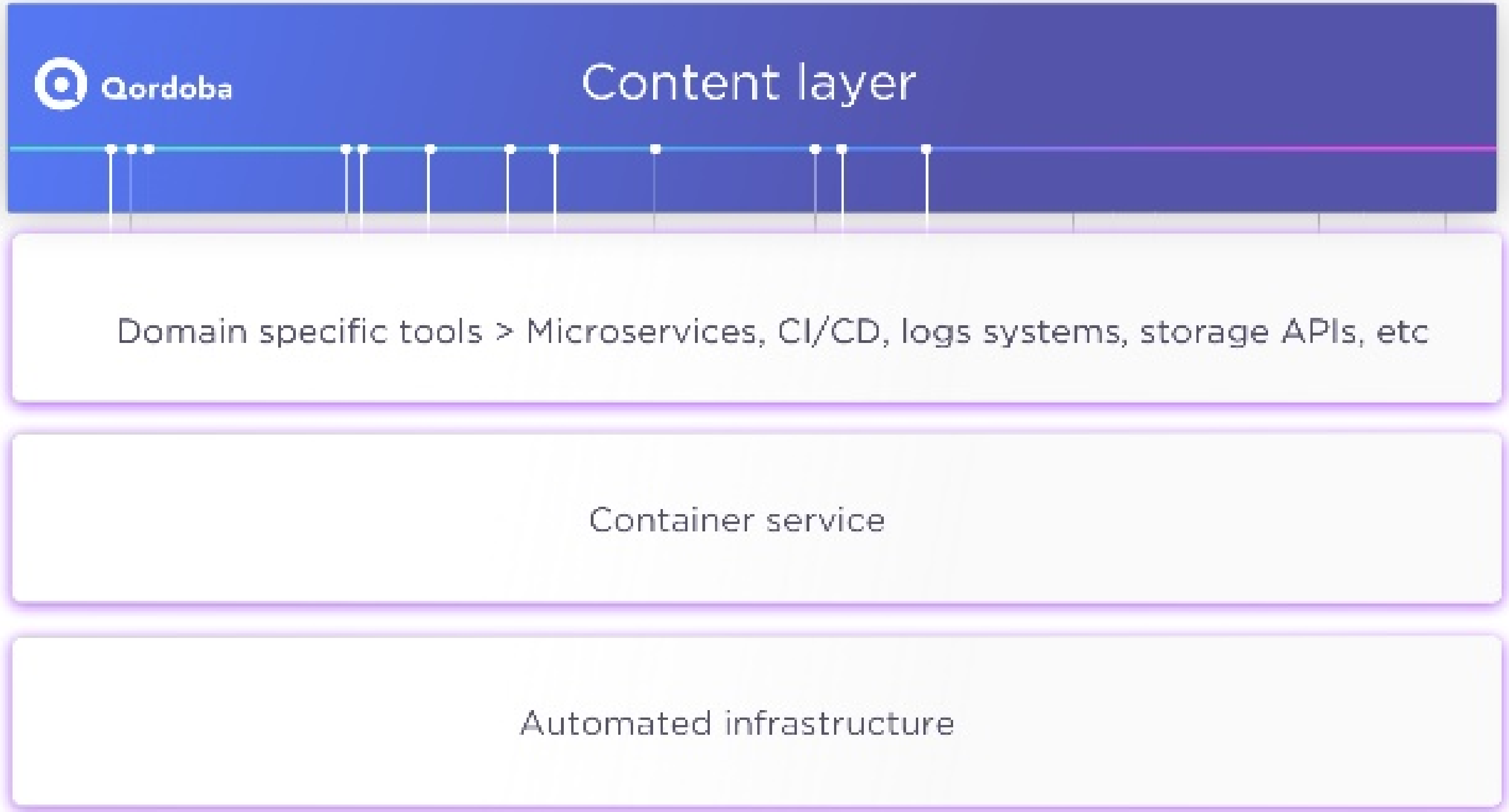
- People
  - Product managers
  - Marketers
  - Designers
  - Linguists
  - Engineers
- Things
  - Copies of copy
  - String files
  - Emails
  - Pull requests
- Results
  - Got milk?
- Wash, rinse, repeat



You're doing it wrong

@texasmichelle

The 🐳 of content



Qordoba

Content layer

Domain specific tools > Microservices, CI/CD, logs systems, storage APIs, etc

Container service

Automated infrastructure

@texasmichelle

Problem	Solution
Content is disorganized & deeply embedded	Content stored in one central location, separate from the application logic
Similar strings are translated over & over	Translation Memory & machine learning
A translation is wrong & needs to be updated everywhere	The text is corrected once & changes are reflected everywhere, without a deployment
Adding new content takes time	Real-time translation of new strings
Translations don't reflect the correct context	Live editor All checks have passed This branch has no conflicts with the base branch Merge pull request
Making a change to content takes forever	A github integration propagates changes automatically

Japantown

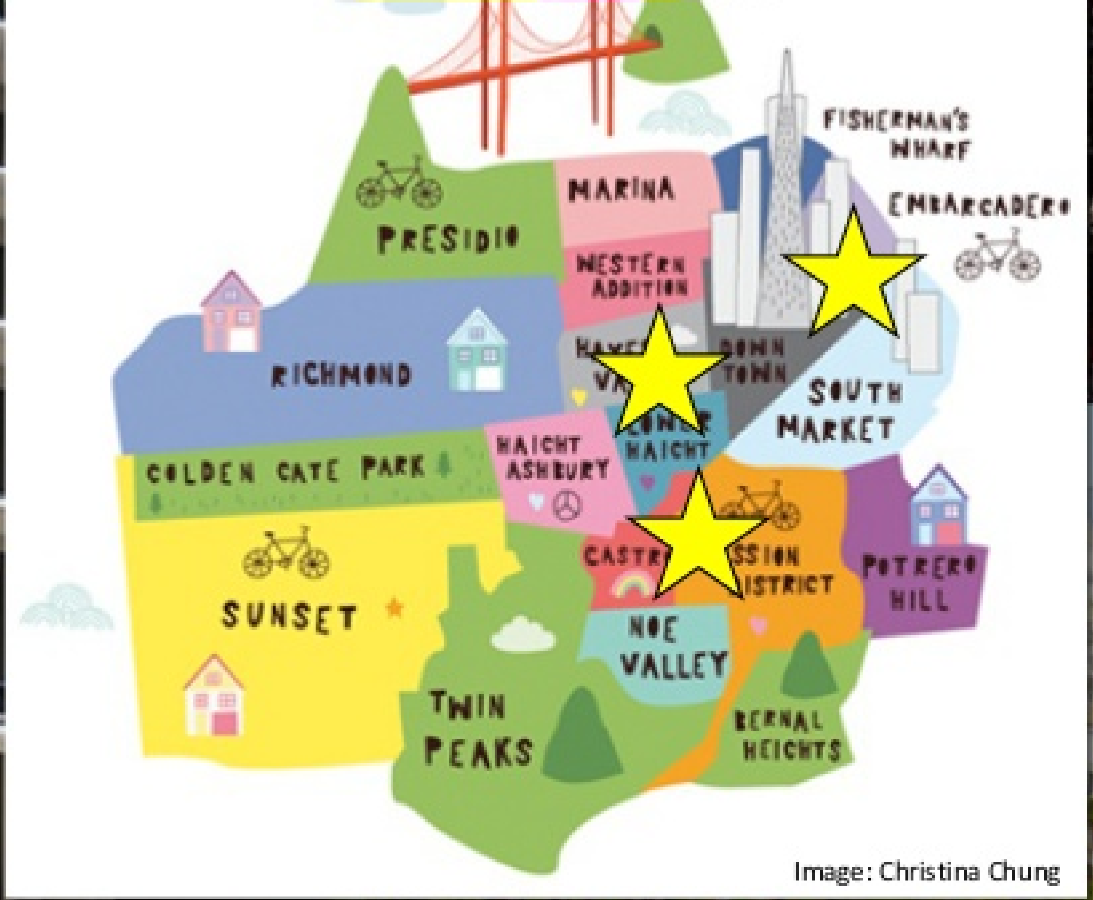
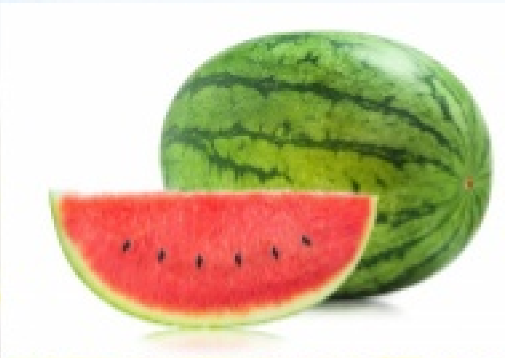


Image: Christina Chung

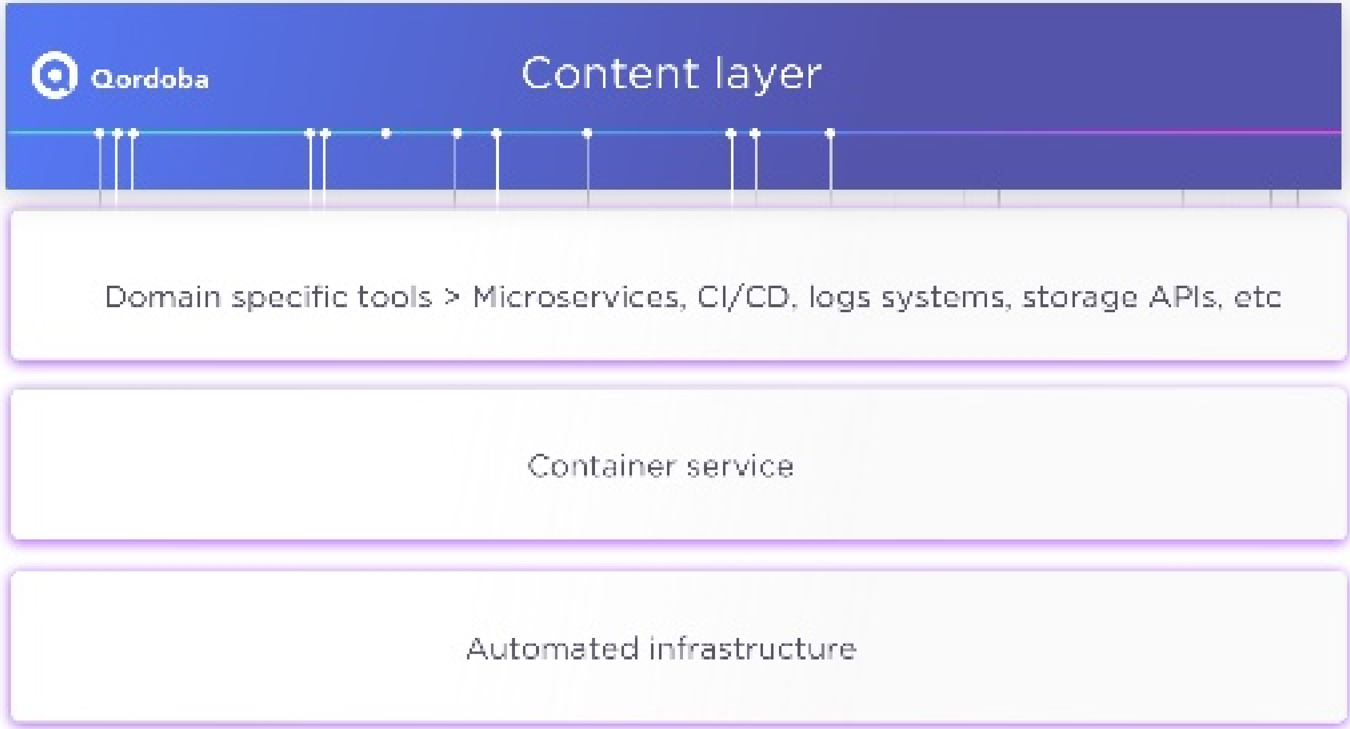
Chinatown



Mission Dolores

Containers

- Powered by
- All the way down 🐢
  - Legal 🇫🇷
  - Informal 🇫🇷
  - Somber 🇫🇷



Qordoba

Content layer

Domain specific tools > Microservices, CI/CD, logs systems, storage APIs, etc

Container service

Automated infrastructure

Affect Detection

- What is it?
- Why do we want it?
  - Hands-off translations 👐
  - Workflow transitions 🔁
- What does this have to do with ☆ ?

Affect Detection

• I had dinner with my wife 🍽️	fear
• I had dinner with my girlfriend 🍽️	joy
• Help Wanted 🙀	sadness
• Busca empleo 🙀	anger
• This apartment is in a killer location 🏙️	joy
• Diese Wohnung befindet sich an einem mörderischen Standort 🏙️	fear

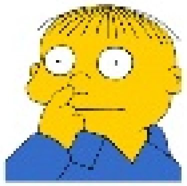
Requirements

- REST interface 💨
  - Response time
- Scalability 🇩🇪 🇩🇰 🇩🇴
  - Languages
  - Deployment
  - Models
- Accuracy 💯
- Open source

PredictionIO

Initial Model

Wow, your first day at the new school! Lisa, have fun. Bart, don't!

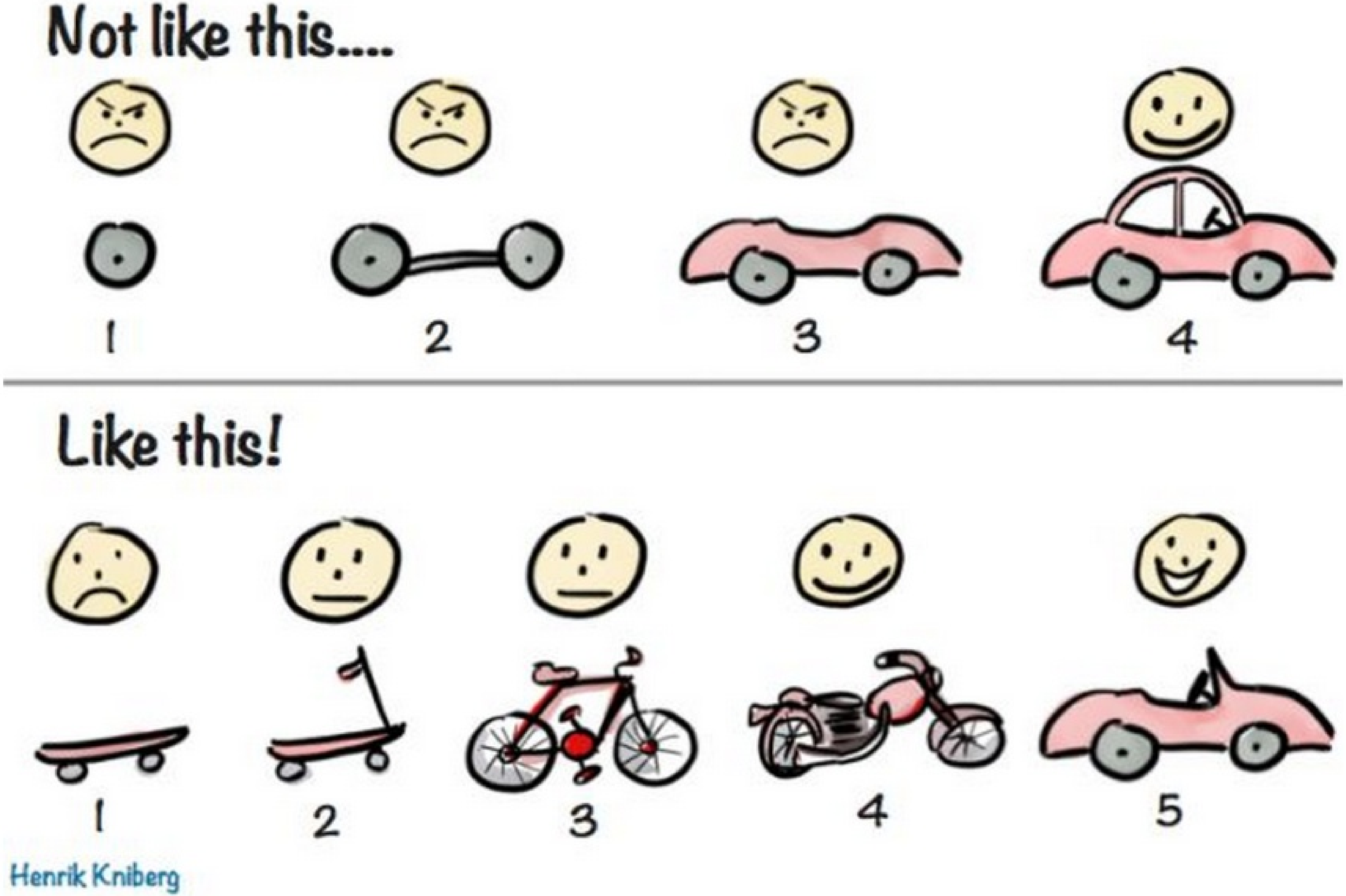


Oh boy, sleep! That's where I'm a viking.

When she sees you'll do anything she says, she's bound to respect you!

It takes two to lie. One to lie and one to listen.

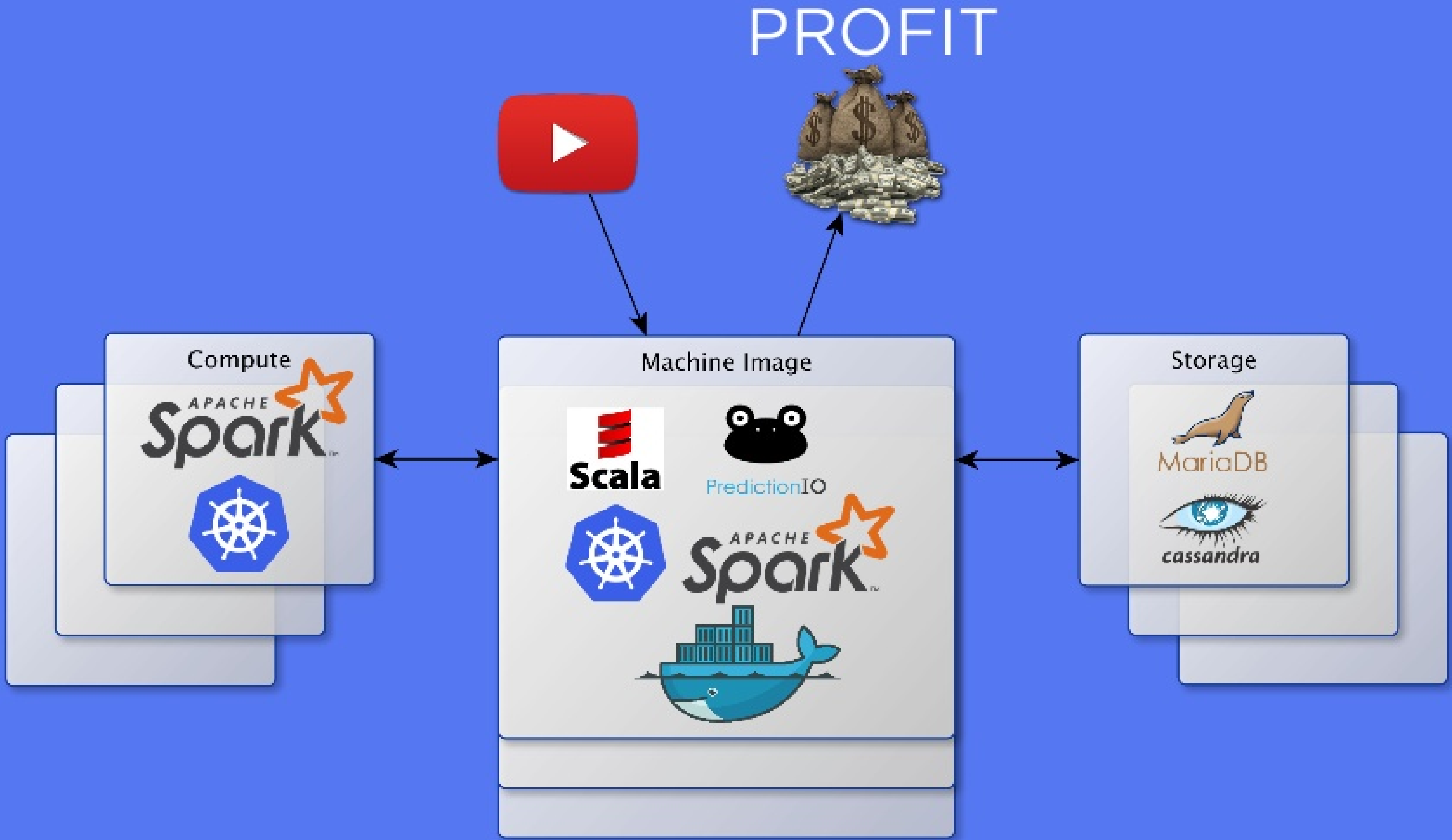
Credit: Matt Groening, Fox


Not like this....
1
2
3
4
Like this!
1
2
3
4
5
Henrik Kniberg

What does the 🚲 look like?

APOLLO 13

APOLLO 13



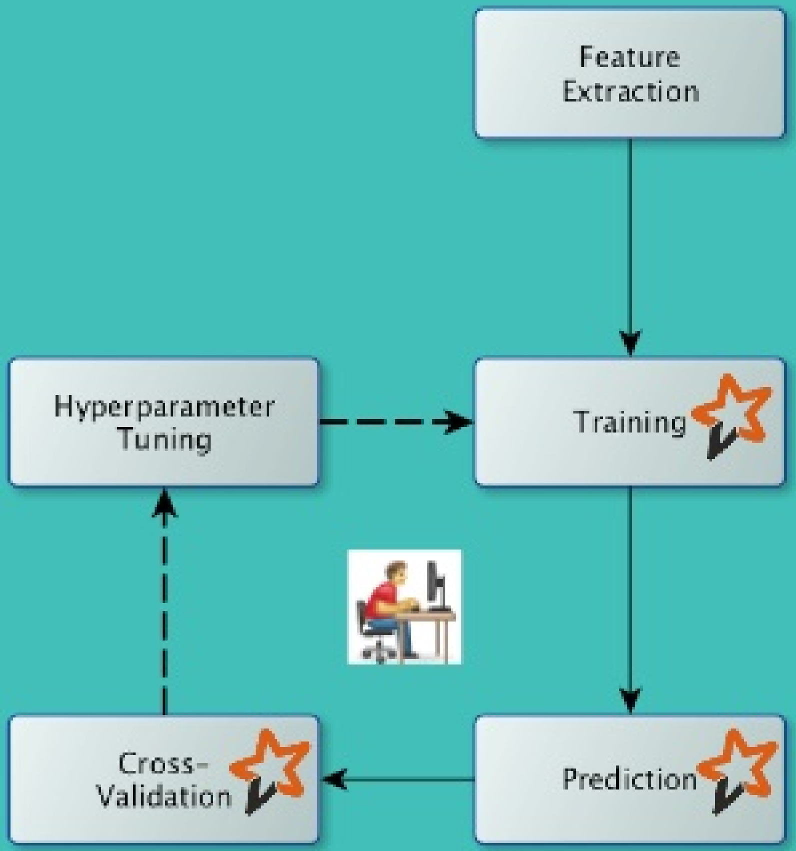
PROFIT

Compute

Machine Image

Storage

NLP



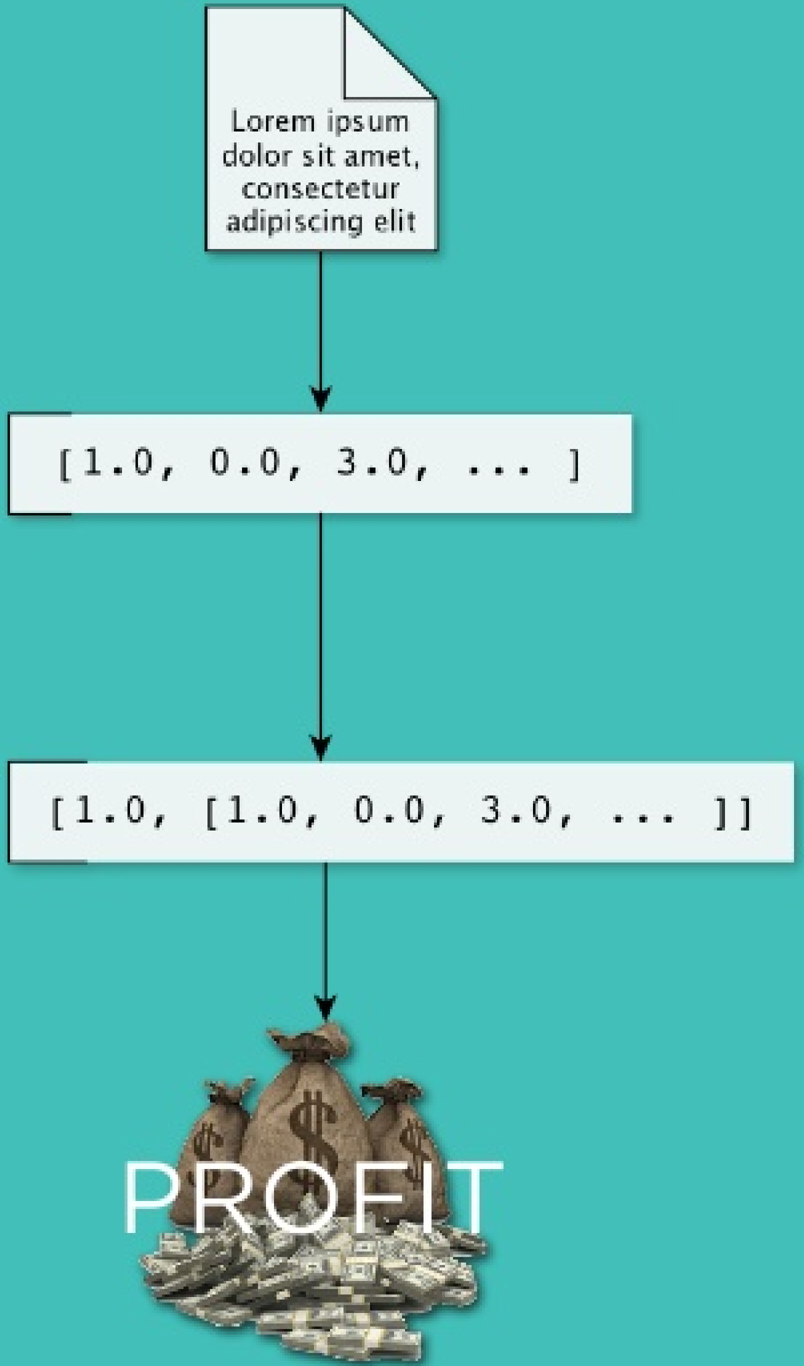
Feature Extraction

Hyperparameter Tuning

Training

Cross-Validation

Prediction

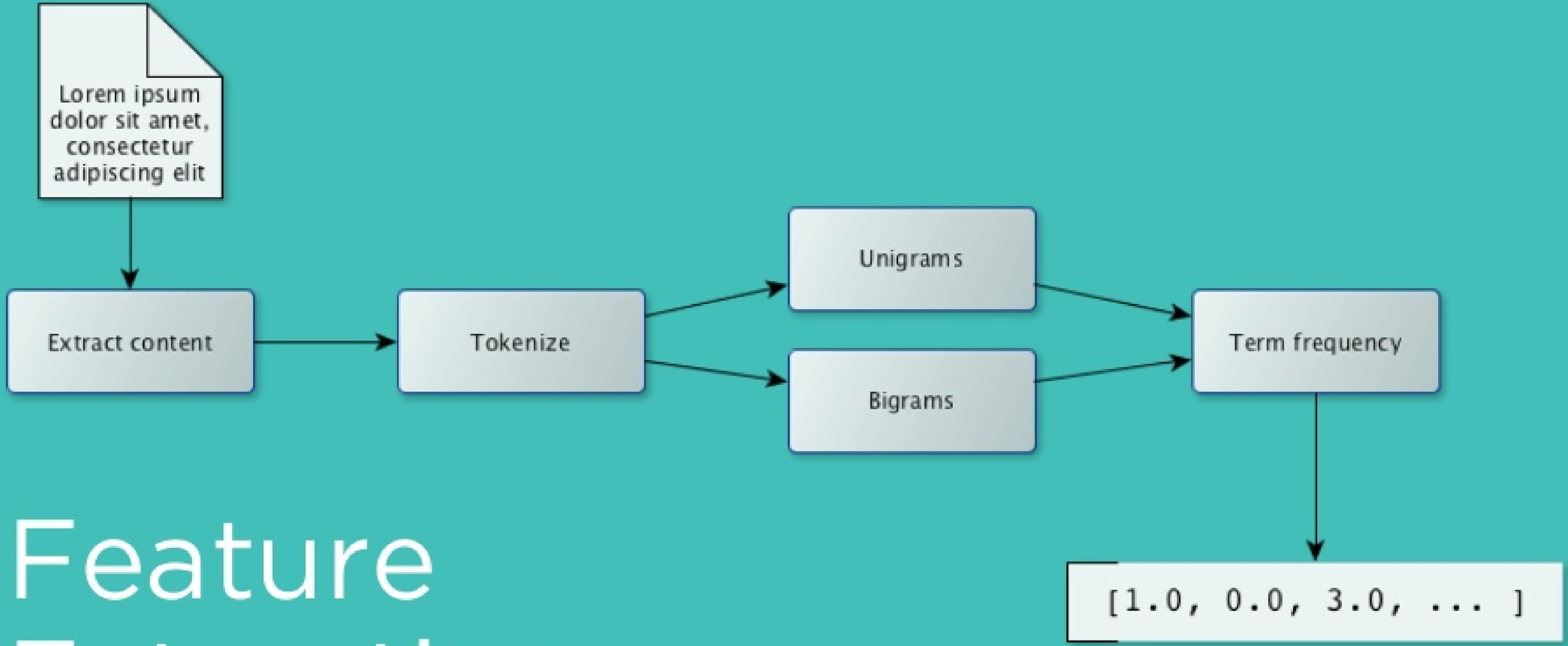


Lorem ipsum dolor sit amet, consectetur adipiscing elit

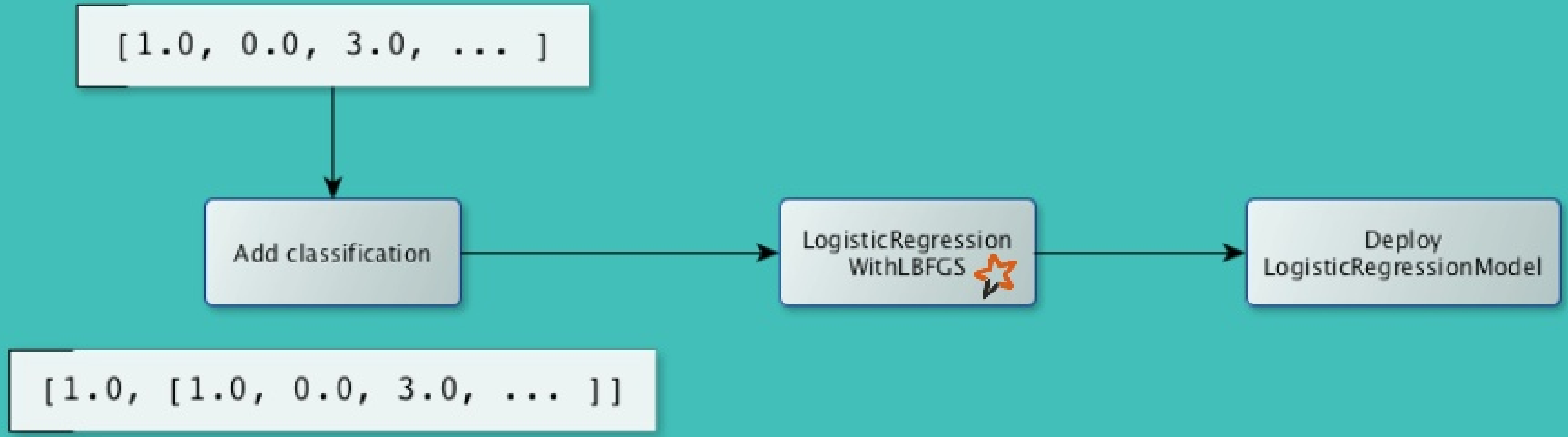
[1.0, 0.0, 3.0, ...]

[1.0, [1.0, 0.0, 3.0, ...]]

PROFIT



Feature Extraction



[1.0, 0.0, 3.0, ...]

Add classification

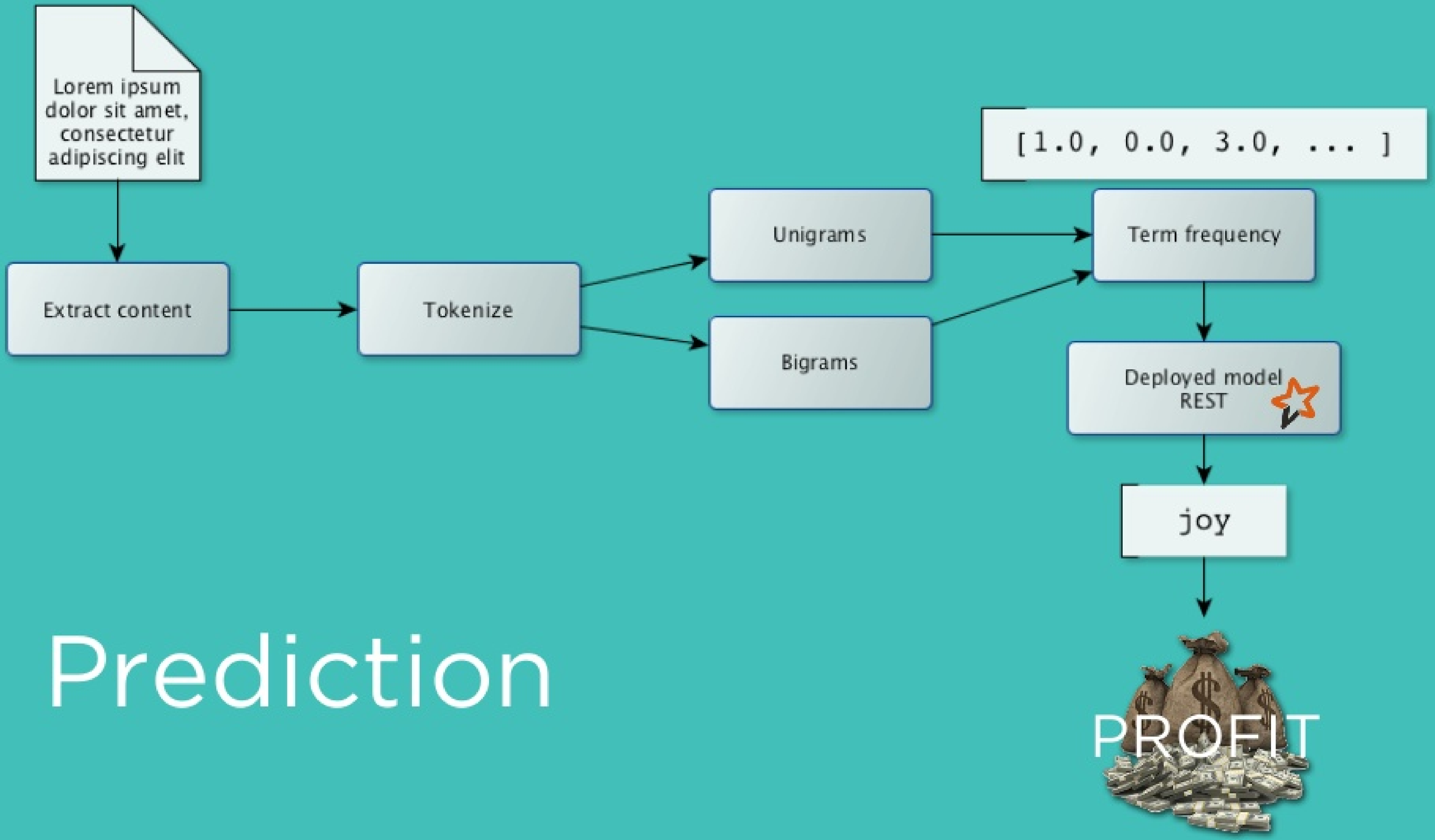
[1.0, [1.0, 0.0, 3.0, ...]]

LogisticRegression WithLBFGS

Deploy LogisticRegressionModel

Training

Prediction



Lorem ipsum dolor sit amet, consectetur adipiscing elit

Extract content

Tokenize

Unigrams

Bigrams

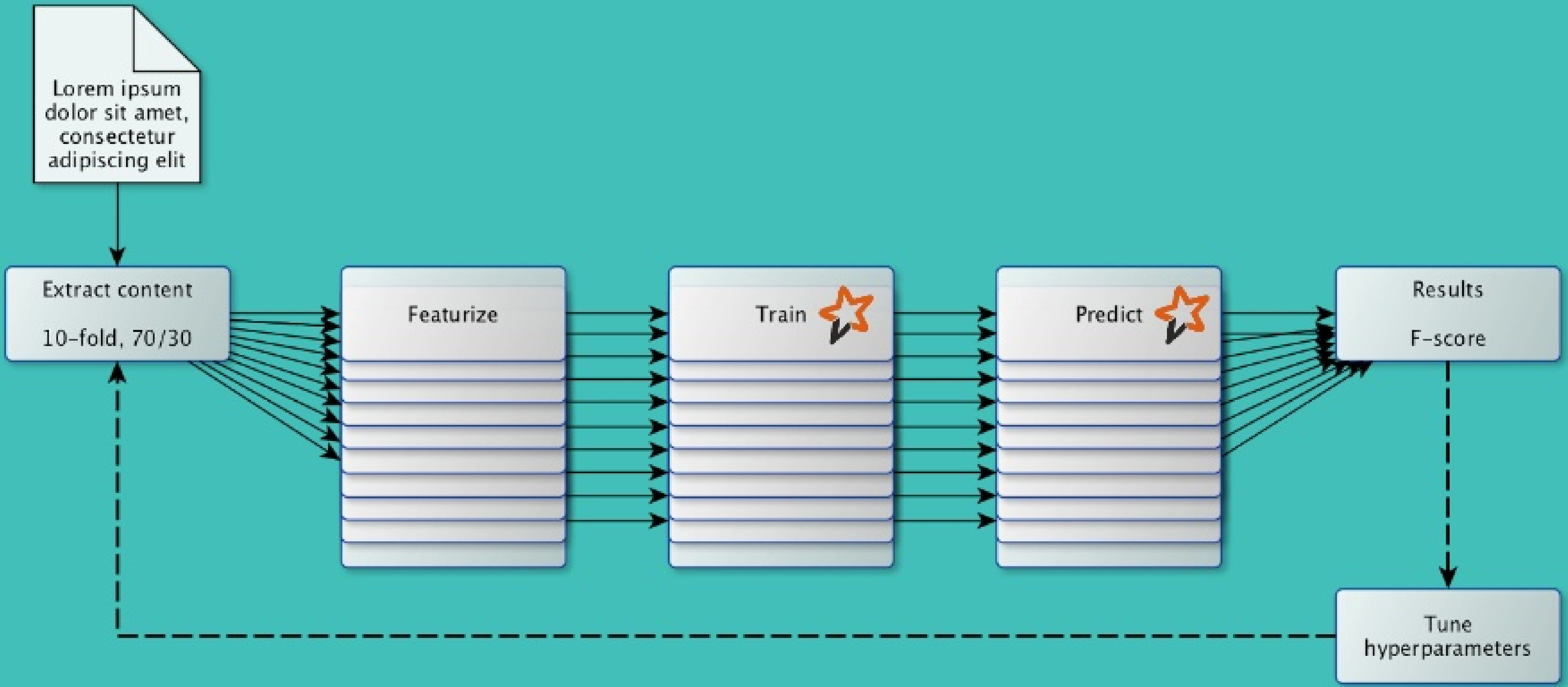
Term frequency

```
[1.0, 0.0, 3.0, ... ]
```

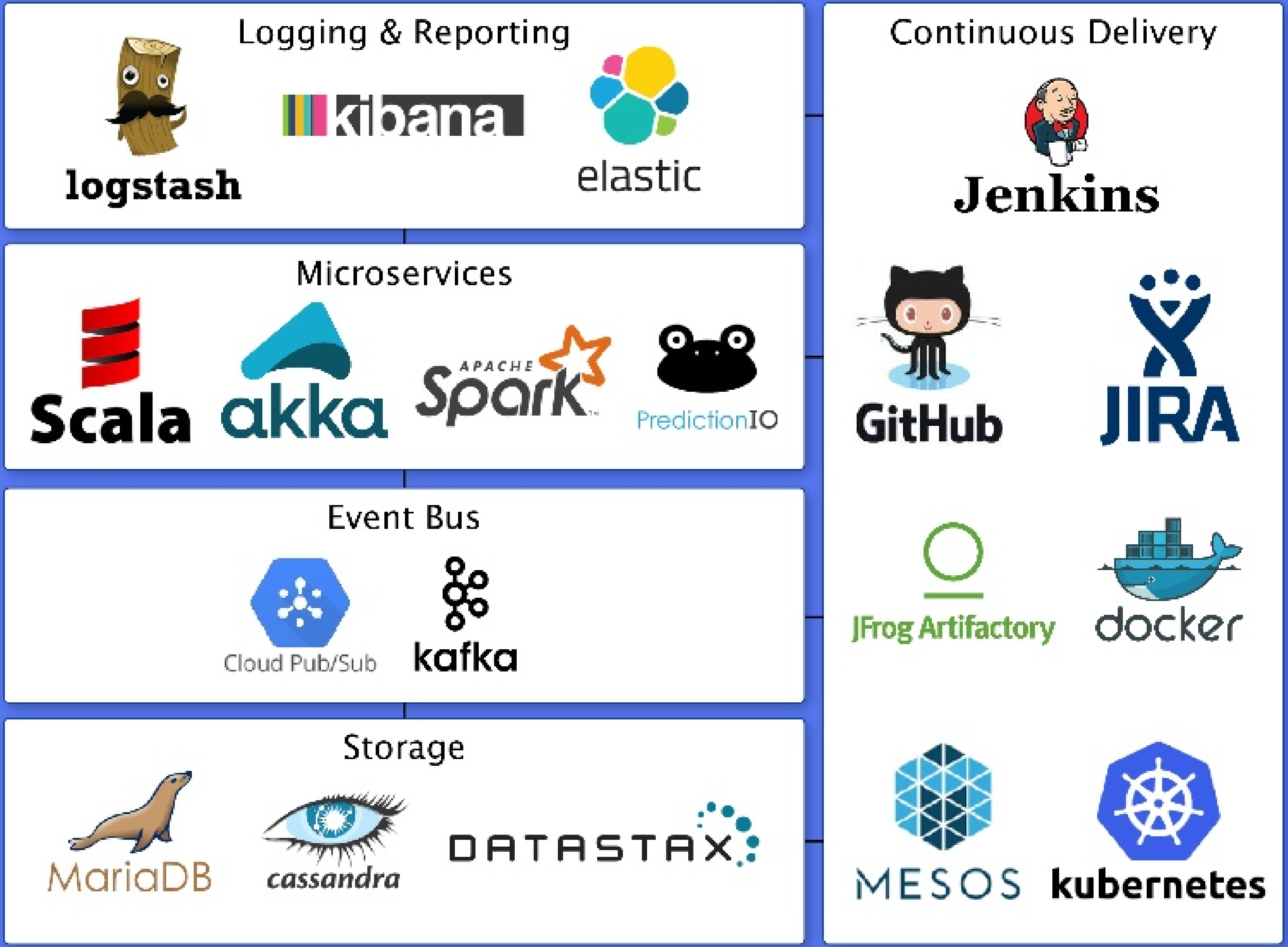
Deployed model REST

```
joy
```

PROFIT



Cross-validation



Logging & Reporting

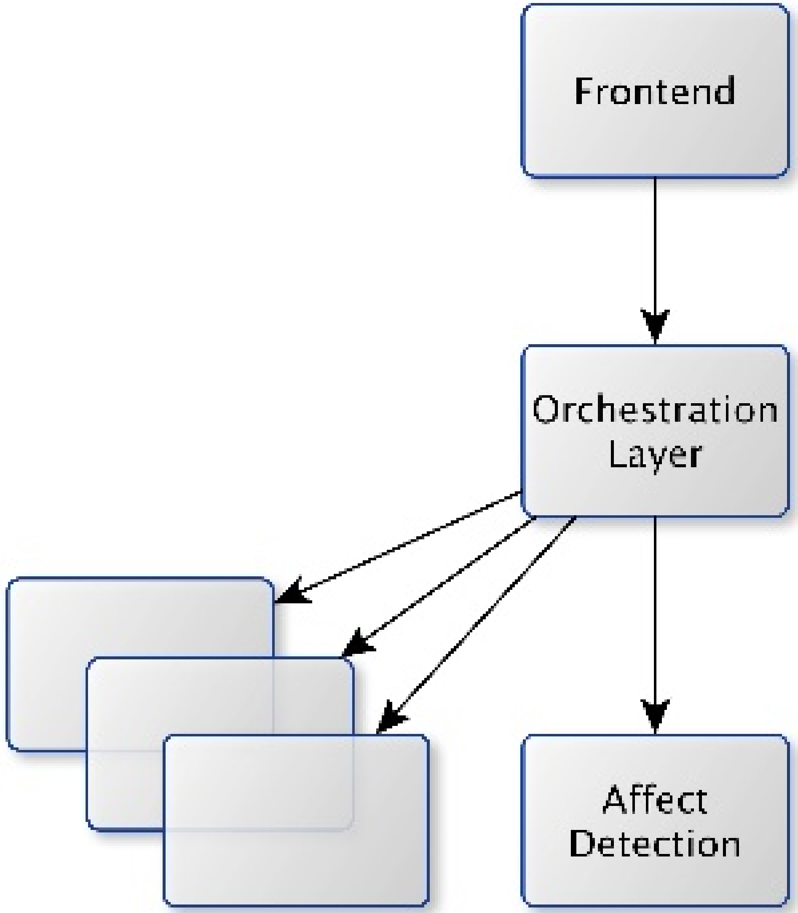
Microservices

Event Bus

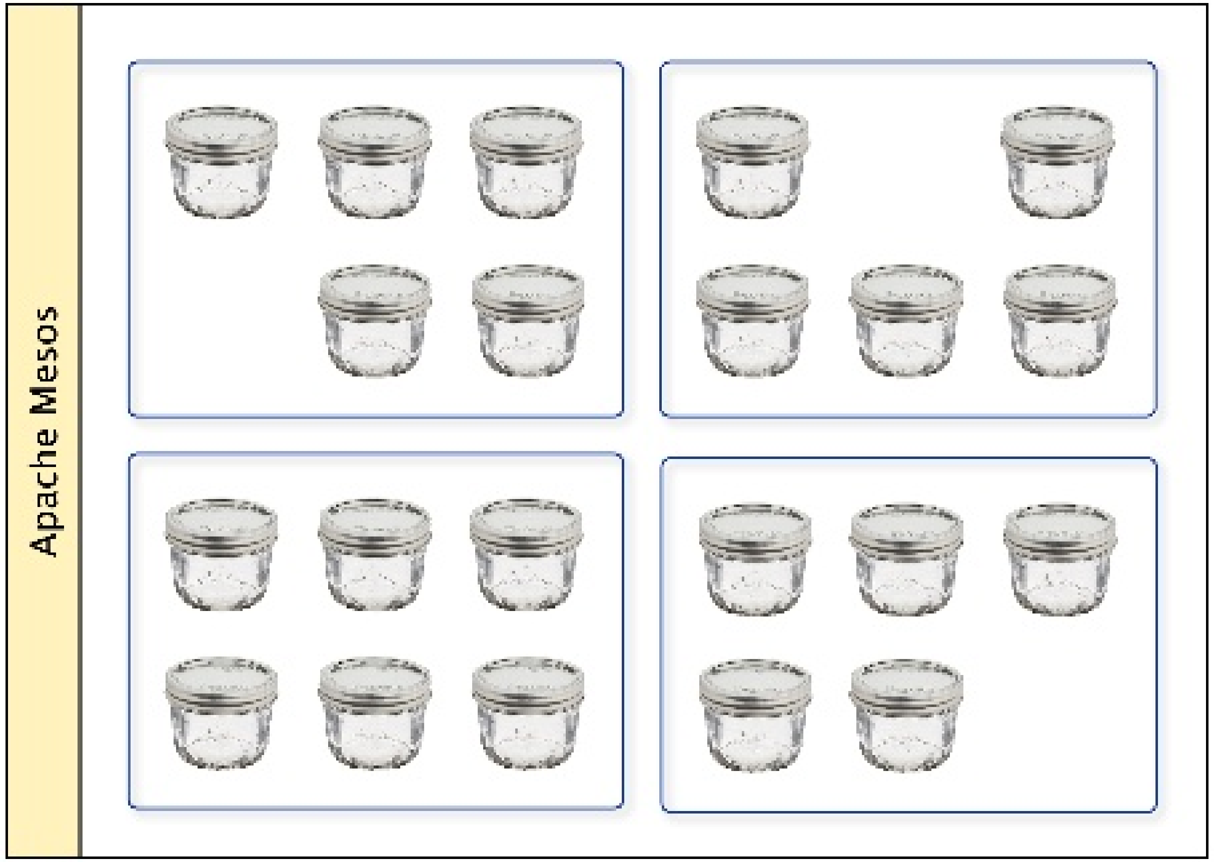
Storage

Continuous Delivery

System Components

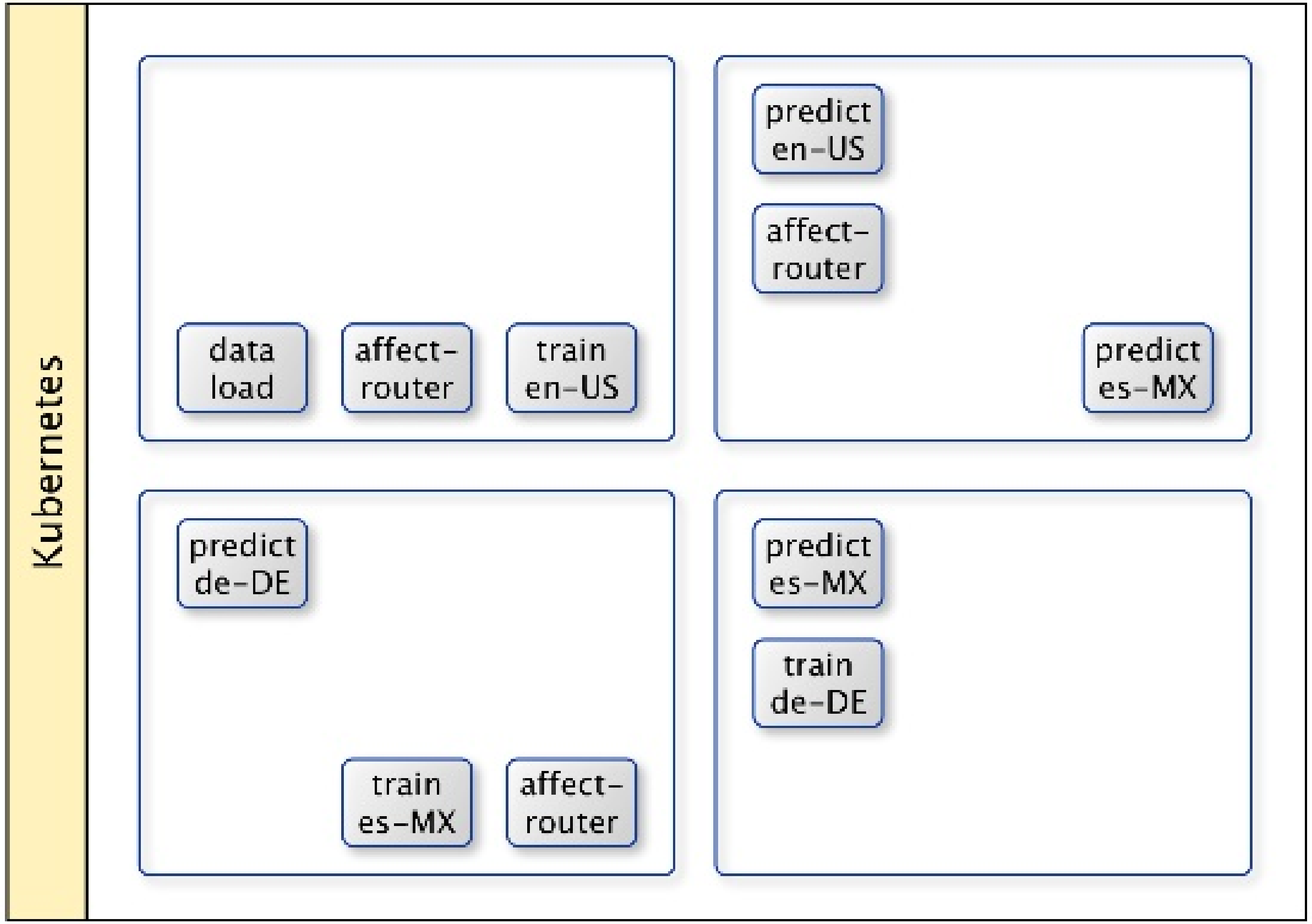


Before



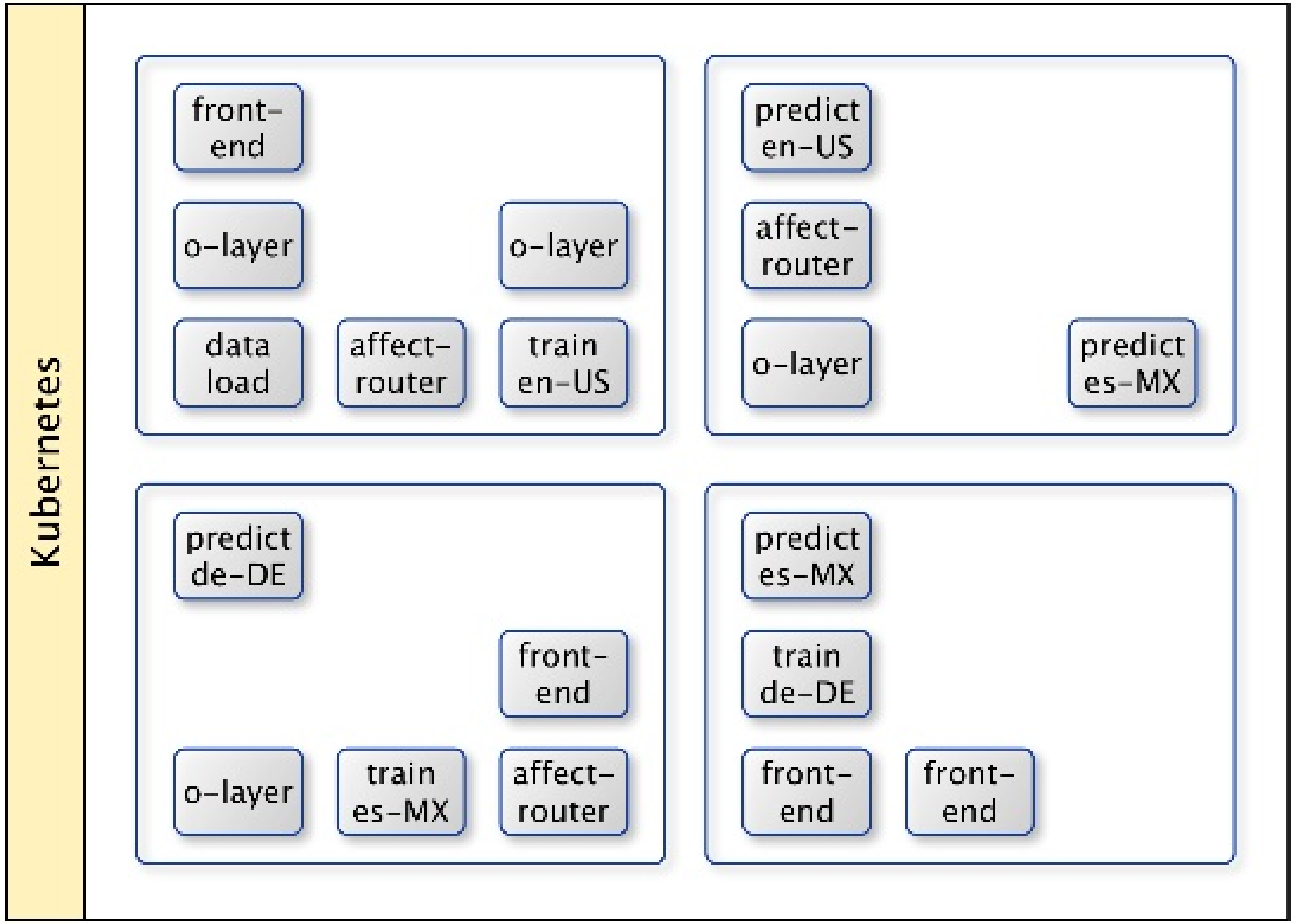
How to prevent conflict between data science & engineering

Affect detection deployment



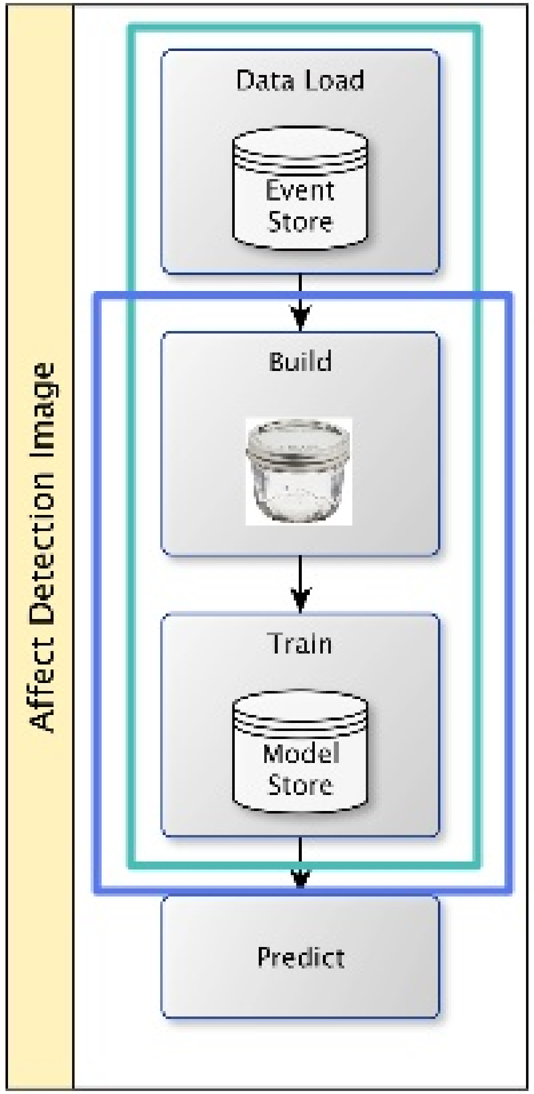
@texasmichelle

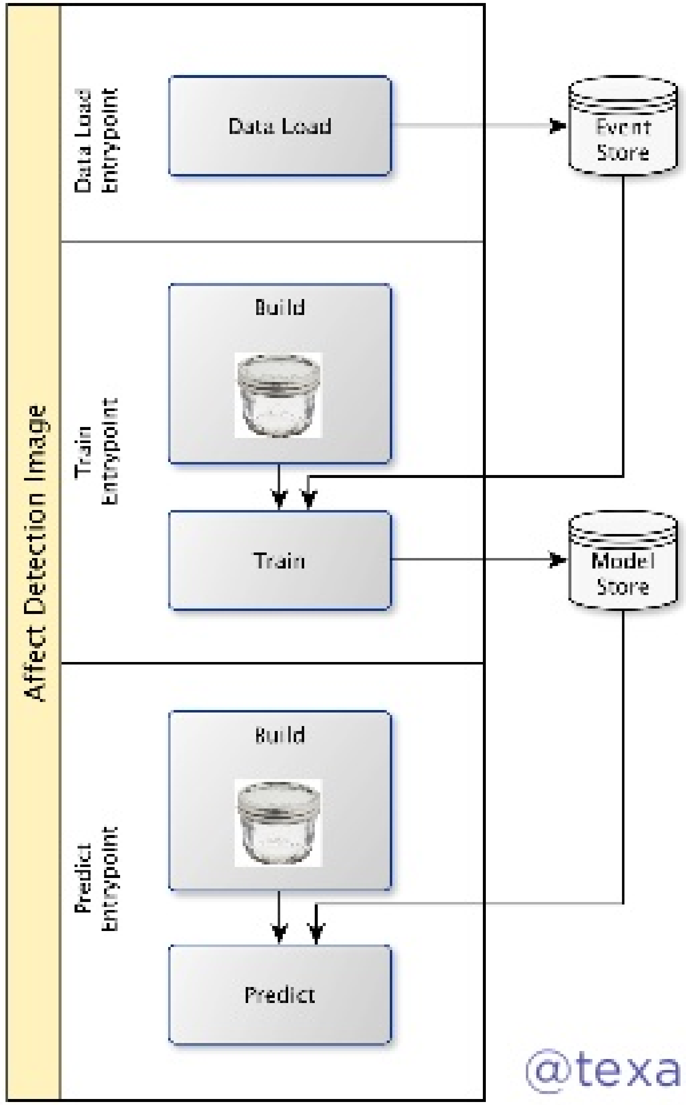
After



@texasmichelle

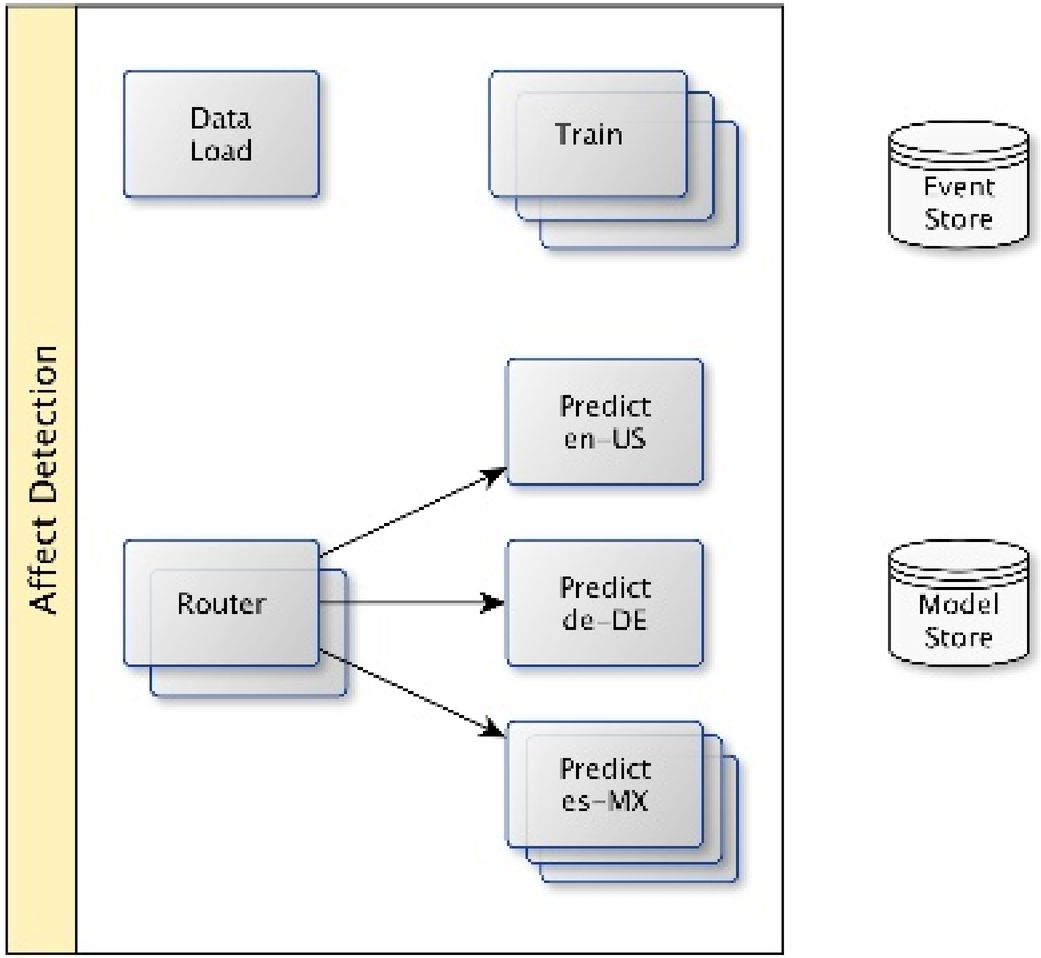
Single Model





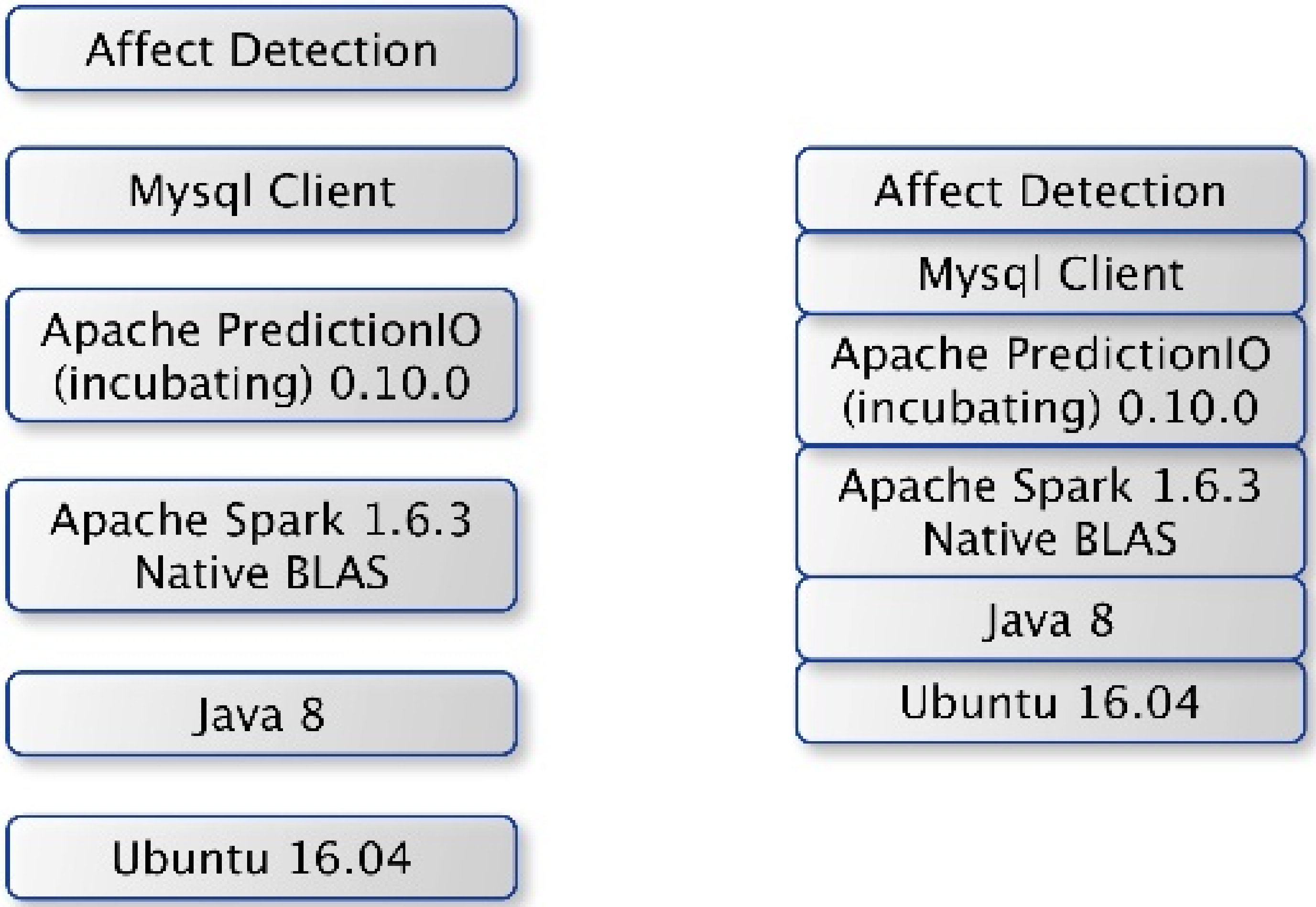
@texasmichelle

Many Models



@texasmichelle

Affect Detection Image



Affect Detection

Mysql Client

Apache PredictionIO (incubating) 0.10.0

Apache Spark 1.6.3 Native BLAS

Java 8

Ubuntu 16.04

Affect Detection

Mysql Client

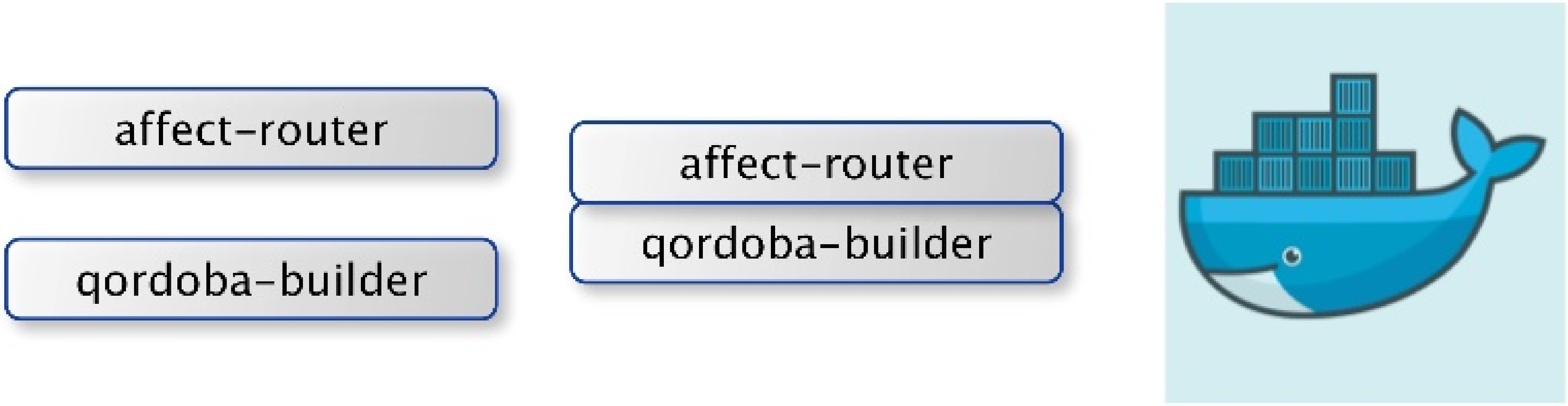
Apache PredictionIO (incubating) 0.10.0

Apache Spark 1.6.3 Native BLAS

Java 8

Ubuntu 16.04

Router Image



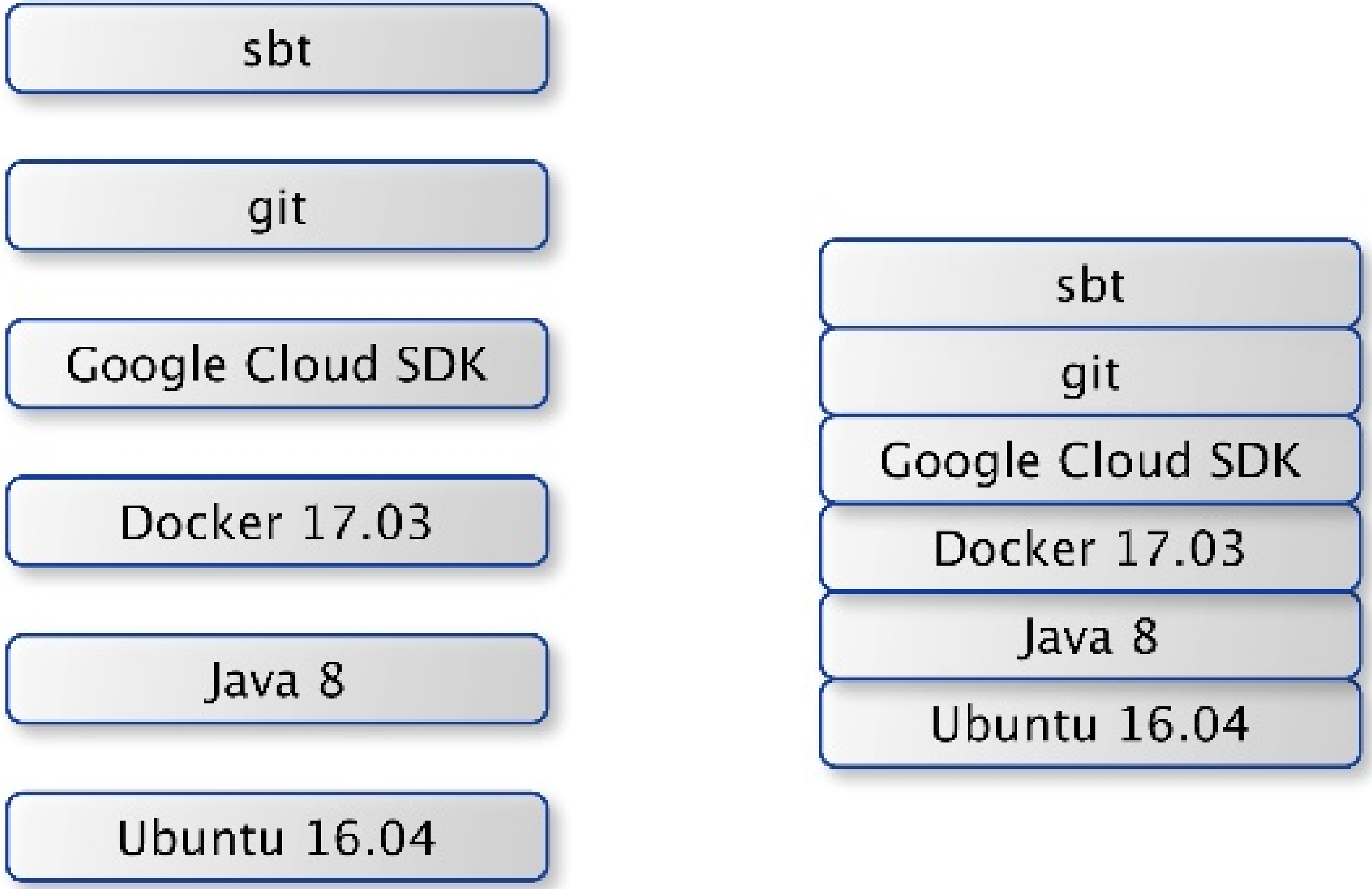
affect-router

qordoba-builder

affect-router

qordoba-builder

Qordoba Builder Image



sbt

git

Google Cloud SDK

Docker 17.03

Java 8

Ubuntu 16.04

sbt

git

Google Cloud SDK

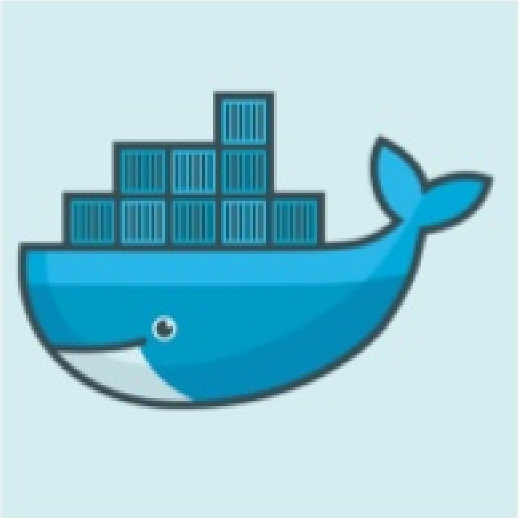
Docker 17.03

Java 8

Ubuntu 16.04

Qordoba Builder Image

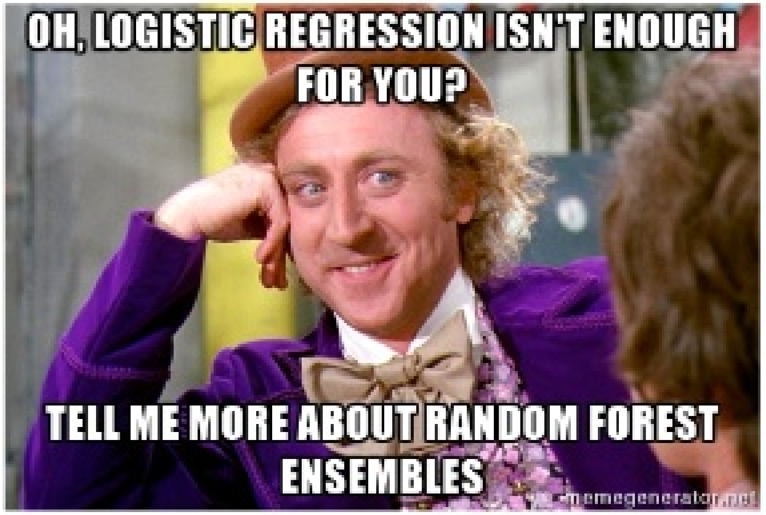
YO DAWG I HEARD YOU LIKE DOCKER
SO WE PUT A DOCKER IN YOUR DOCKER
memegenerator.net



Qordoba Builder Image

Affect Detection

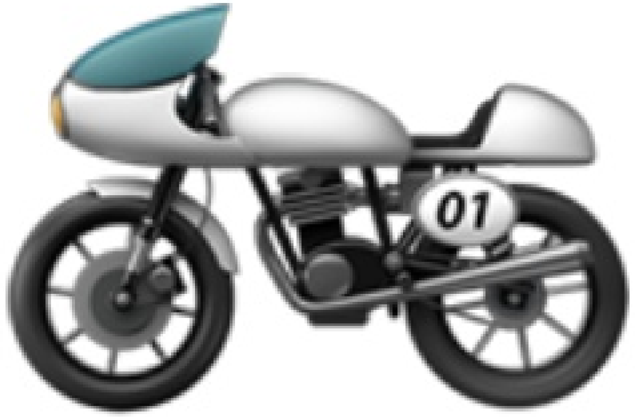
- Featurization improvements
- Multi-label instead of mutual exclusivity
- Confidence scores
- Additional algorithms
- Expanded training set
- More evaluation metrics
- ~~Classification~~ Graph?
- Unsupervised models


OH, LOGISTIC REGRESSION ISN'T ENOUGH FOR YOU?
TELL ME MORE ABOUT RANDOM FOREST ENSEMBLES
memegenerator.net

@texasmichelle

Affect Detection

- Lighter images
- Lighter deployment
- Weaveworks



TL;DR

- Importance of localization
    - Why it’s hard
    - How it can be made easier
    - Examples
- The role of NLP in solving this problem
    - Brief introduction to NLP
- Spark’s role in NLP
- Infrastructure
- Deployment

Challenge

Don’t just build products that scale. Build products that scale content.

Q&A

michelle@qordoba.com

@texasmichelle