

# Global Empire-Building for Fun & Profit

Michelle Casbon  
February 8, 2017  
Spark Summit East  
Boston



**Qordoba**

# whoami

- Where I work: Qordoba, Director of Data Science
- Where I used to work: **iDIBON**
- 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



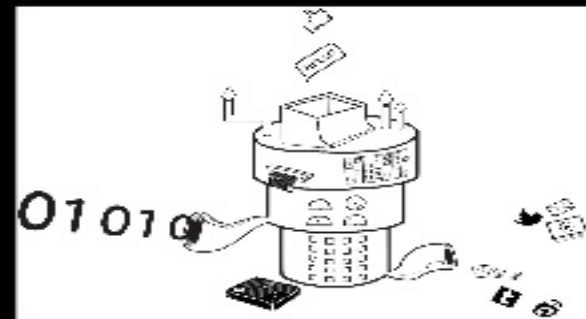
What society thinks I do



What my boss thinks I do



What I think I do



What I actually do

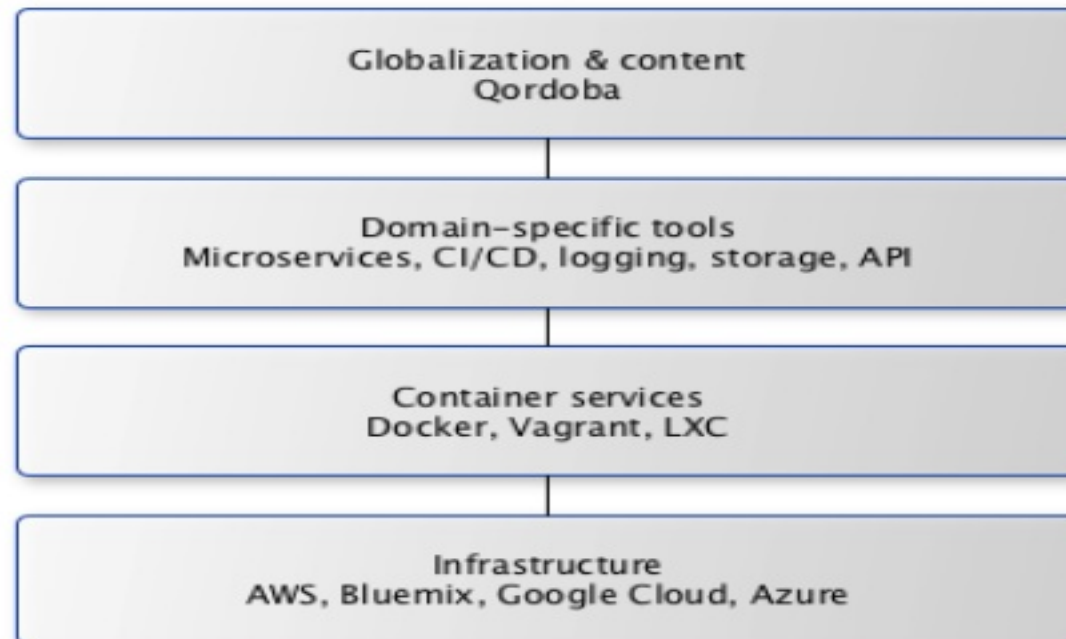
# TL;DR

- Mission
  - To enable products that feel native to every user
  - Default: a product is in 100+ markets
- There's a better way to do localization
  - Hint: it involves ✨ MLlib
- Affect detection is fun & useful

@texasmichelle



# The of content

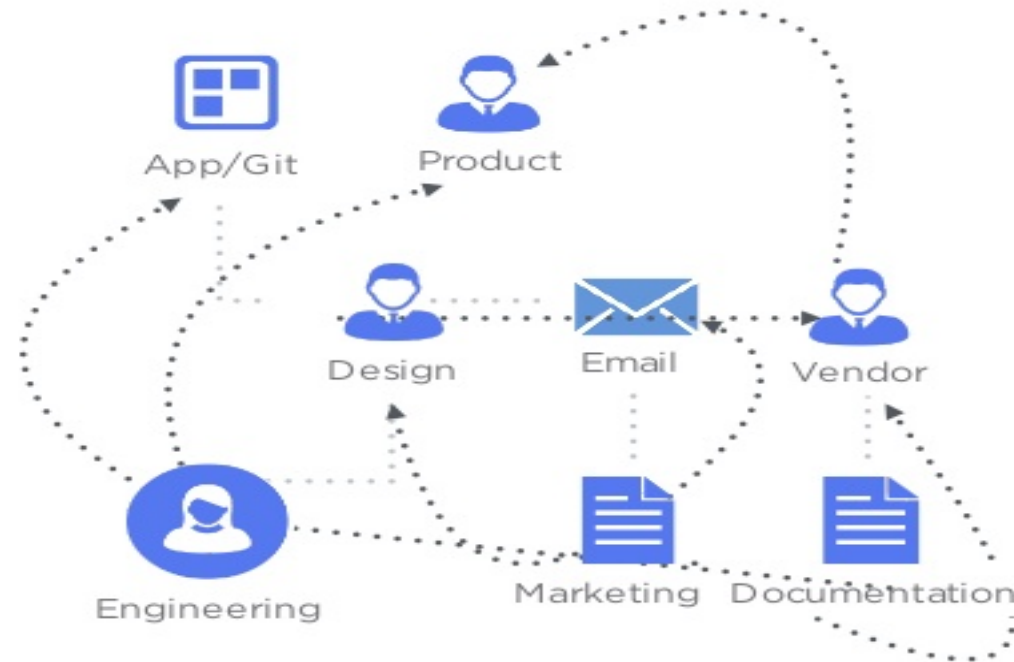


@texasmichelle



# Localization is hard

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



@texasmichelle







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



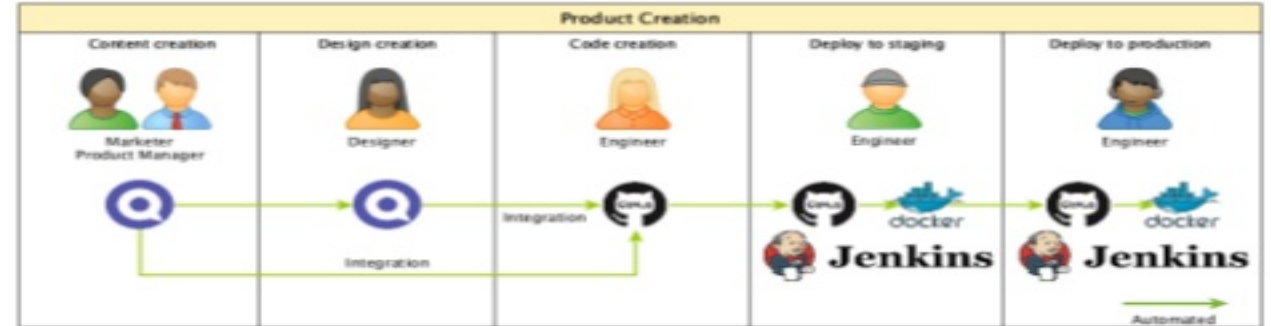
@texasmichelle





# Platform

- Code points to a dynamic link 
  - Other entities update the 
    - Marketers
    - Linguists
    - ML models
  - Live changes
  - Consistency among mediums
- Real-time translations



@texasmichelle



# Platform

- Context
  - Linguists
  - Designers
- Github repo receives PRs



@texasmichelle















## I Followed My Stolen iPhone Across The World, Became A Celebrity In China, And Found A Friend For Life?

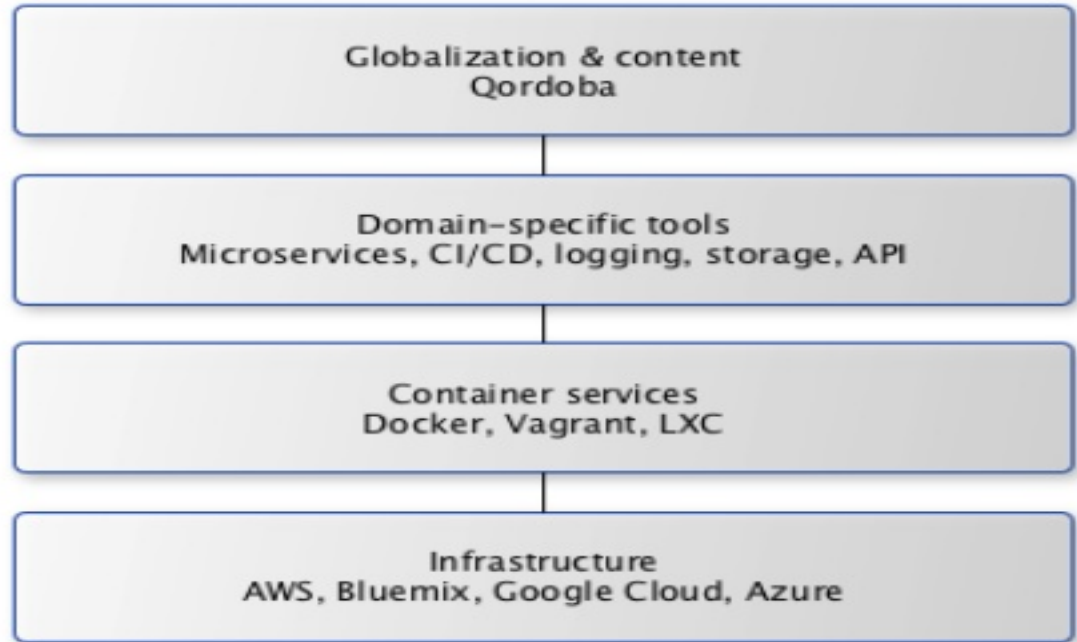
[https://www.buzzfeed.com/mjs538/i-followed-my-stolen-iphone-across-the-world-became-a-celebr?utm\\_term=.ut02j4dGN4#.txnrxyjGk](https://www.buzzfeed.com/mjs538/i-followed-my-stolen-iphone-across-the-world-became-a-celebr?utm_term=.ut02j4dGN4#.txnrxyjGk)





# Containers

- Powered by  & 
- All the way down 
  - Legal 
  - Informal 
  - Somber 



@texasmichelle



# Affect Detection

- What is it?
- Why do we want it?
  - Hands-off translations 🙌
  - Workflow transitions 🔁



@texasmichelle





# Affect Detection

- I had dinner with my wife 
- I had dinner with my girlfriend 
- Help Wanted 
- Busca empleo 
- This apartment is in a killer location 
- Diese Wohnung befindet sich an einem mörderischen Standort 

fear

joy

sadness

anger

joy

fear

@texasmichelle



# Requirements

- REST interface 
  - Response time
- Scalability   
  - Deployment
  - Models
  - Languages
- Accuracy **100**
- Open source 

@texasmichelle



01010

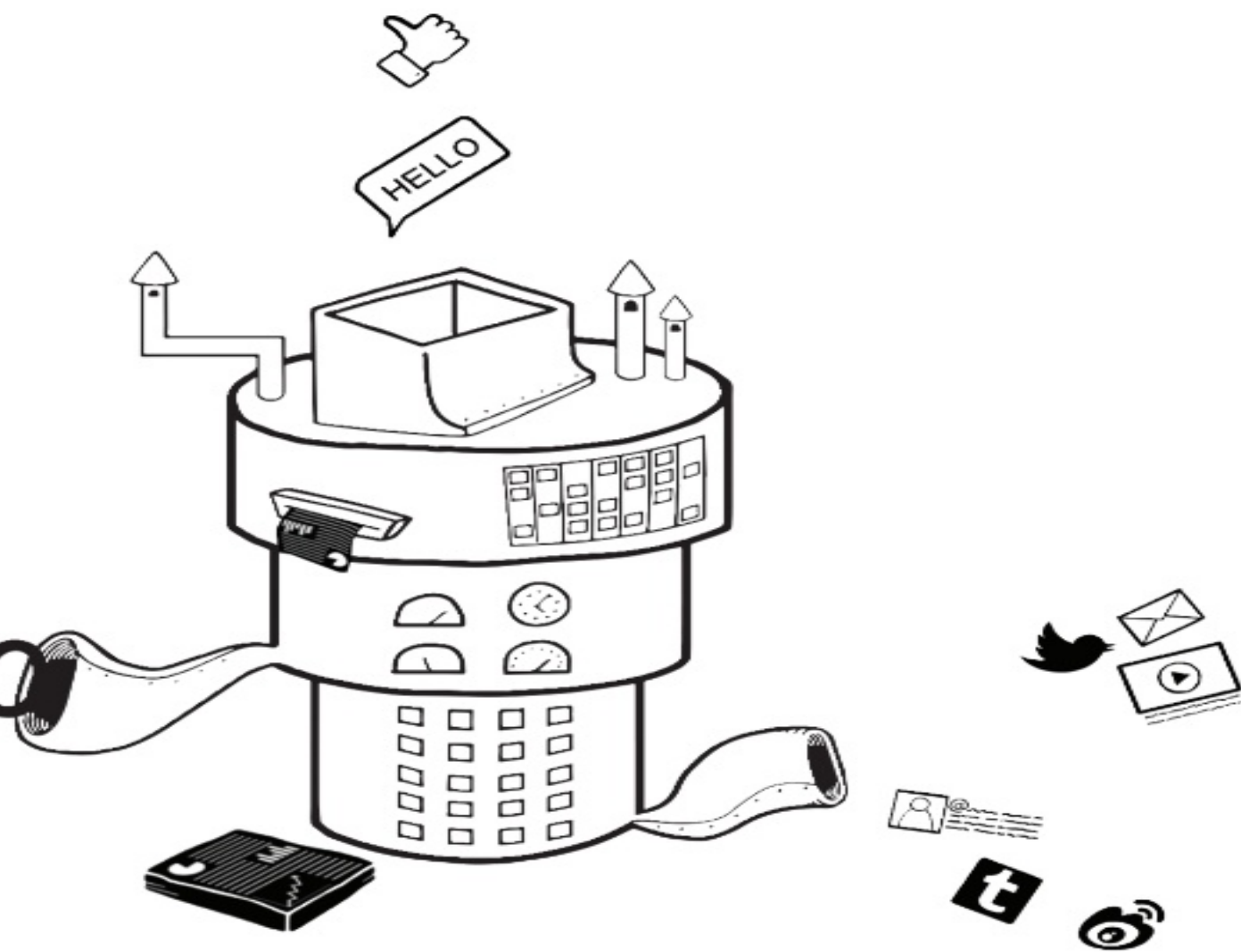





Photo credit: Evan Amos



# build vs. buy

Not from scratch. Because  MLlib!



Not like this....



1



2



3



4

---

Like this!



1



2



3








4



5

# Affect Detection

-  “What do you mean switching from fabric to leather seats involves a different drivetrain?”
-  came early
  - PredictionIO entered ASF Incubator 
    - Existing Apache community 
    - REST interface
    - Response time
    - Scalability
    - Clean objects 
      - Engine
      - Evaluator
      - Algorithm
      - ...



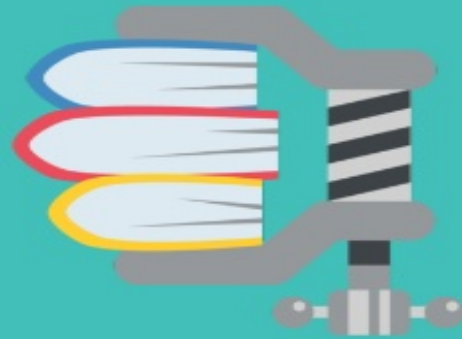
@texasmichelle



But!







Not like this....



1



2



3



4

---

Like this!



1



2



3








4



5

# Affect Detection

- Trade-in Rube Goldberg machine with  wheels for a 
- Focus on the hard parts 
  - Model building 
  - Featurization
  - Hyperparameters
  - Training data 

@texasmichelle





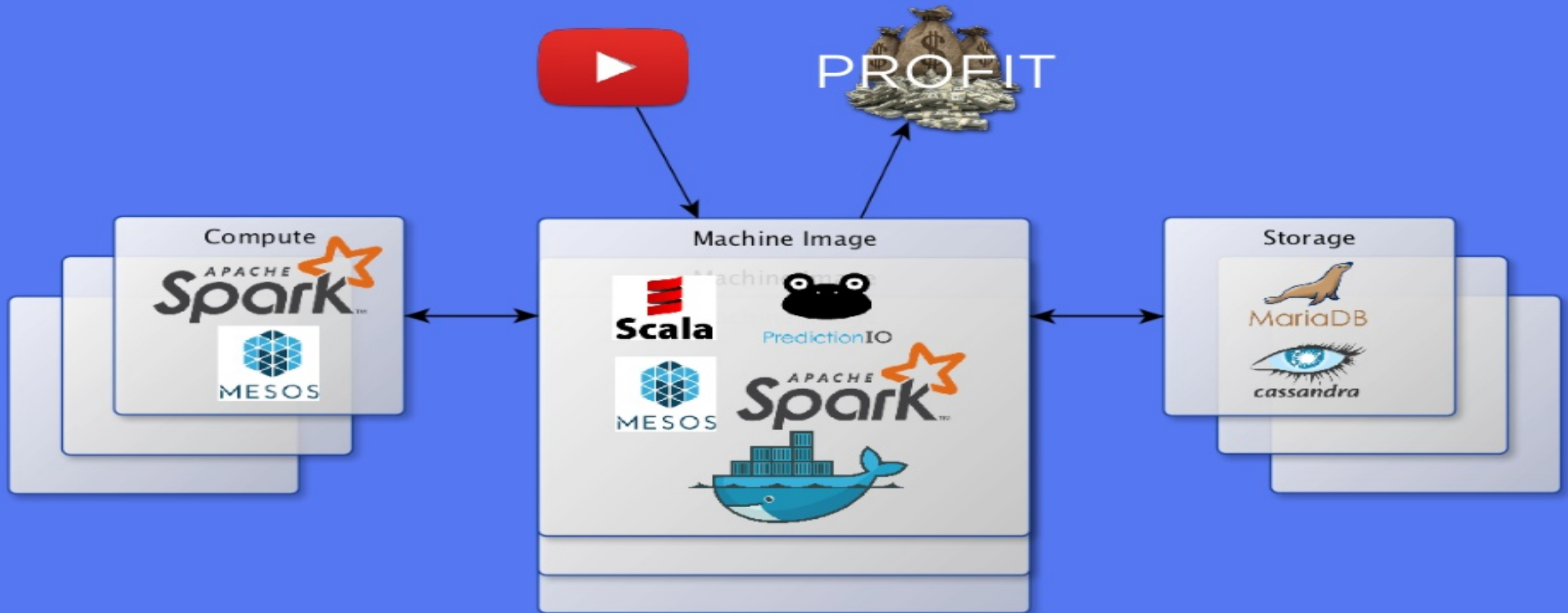
What does the  look like?



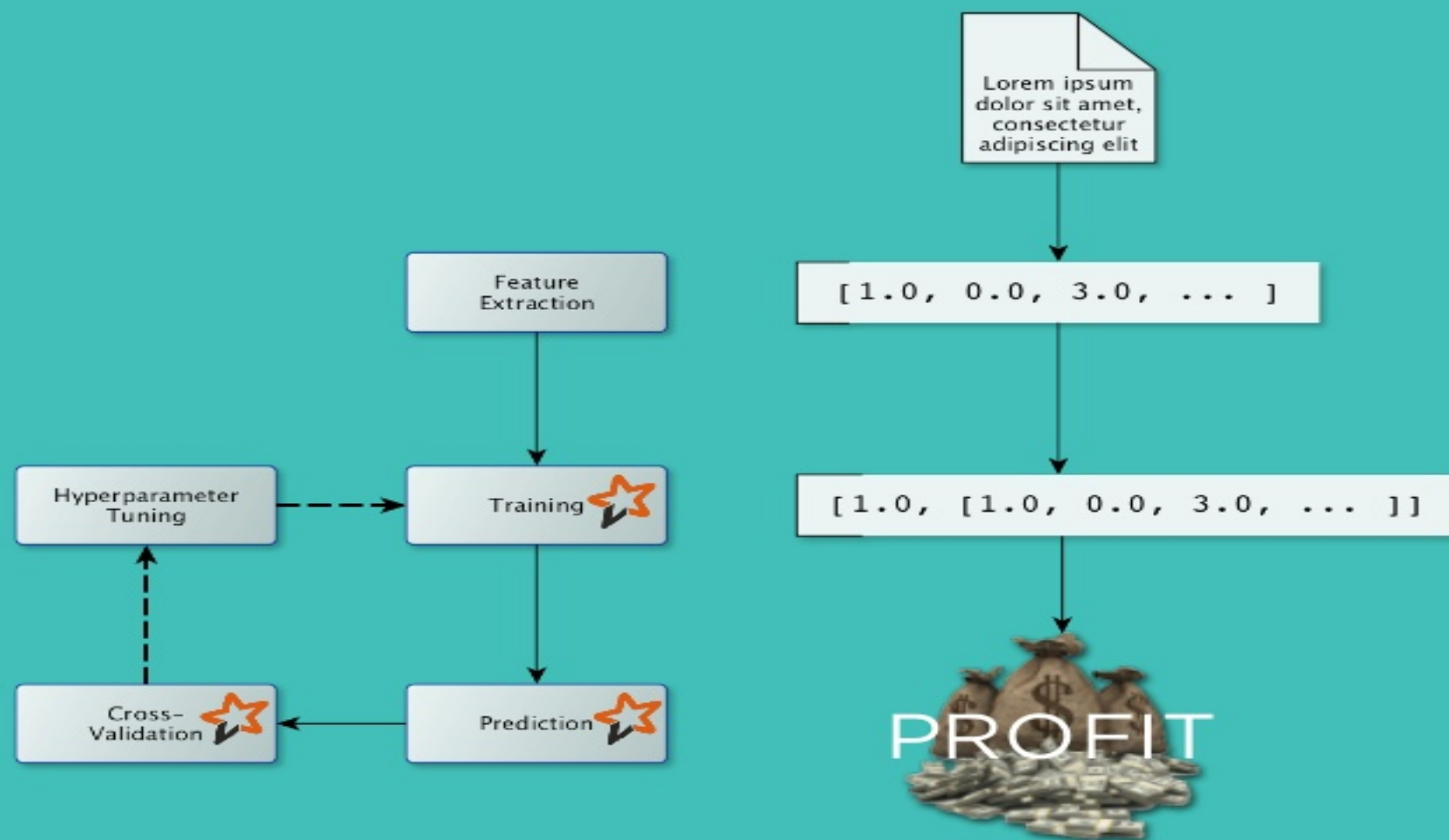


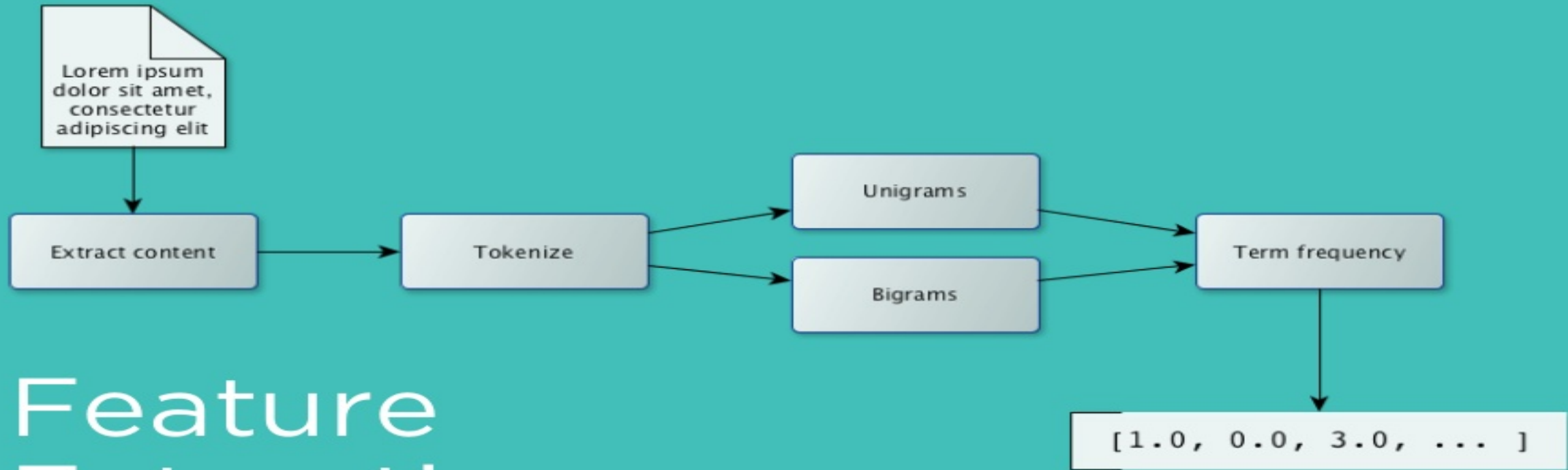


PROFIT



# NLP





# Feature Extraction



[ 1.0, 0.0, 3.0, ... ]

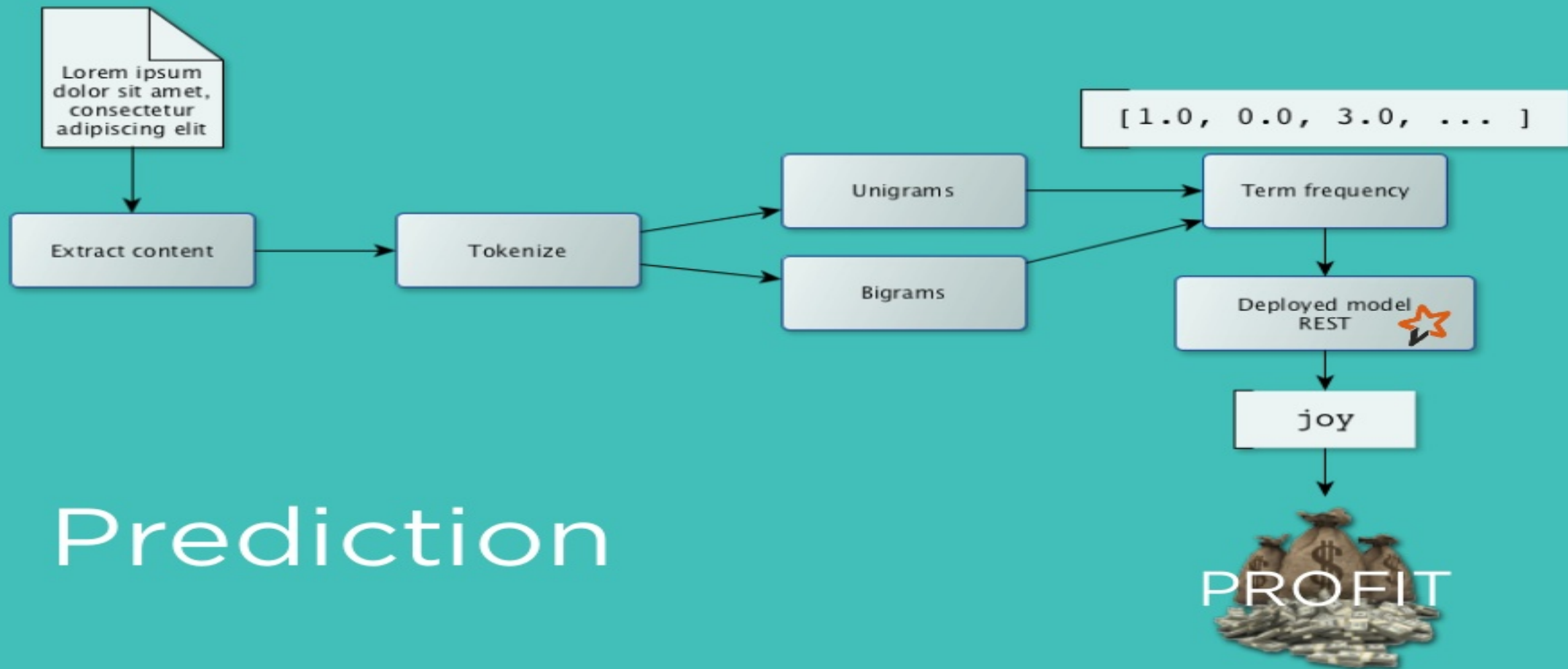
Add classification

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

LogisticRegression  
WithLBFGS

Deploy  
LogisticRegressionModel

# Training



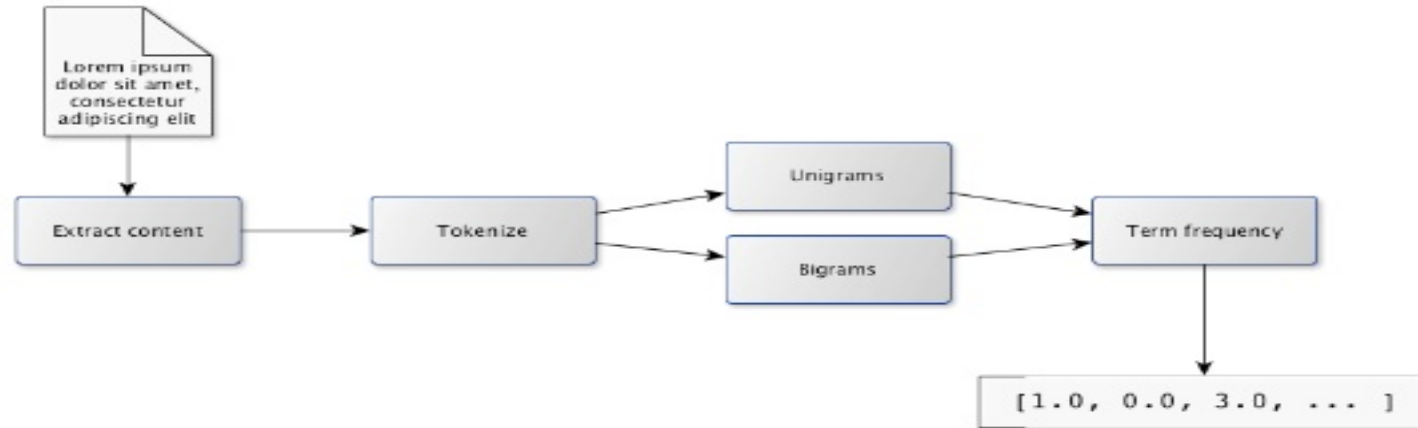
# Prediction



# Cross-validation

# Featurization

- UTF8 support
- Simple
  - Term frequency, descending
  - Ngrams



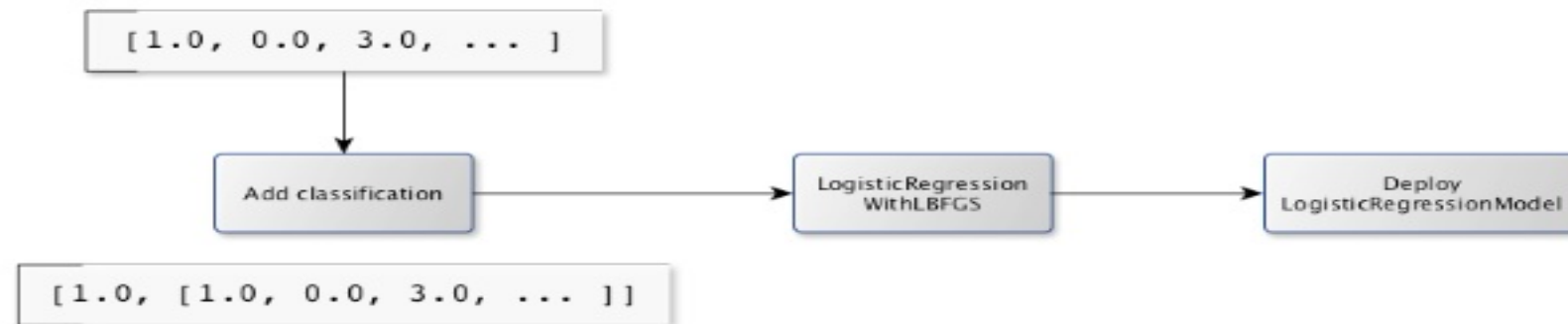
@texasmichelle





# Training

- Supervised models
- Multi-class classification
  - Logistic regression

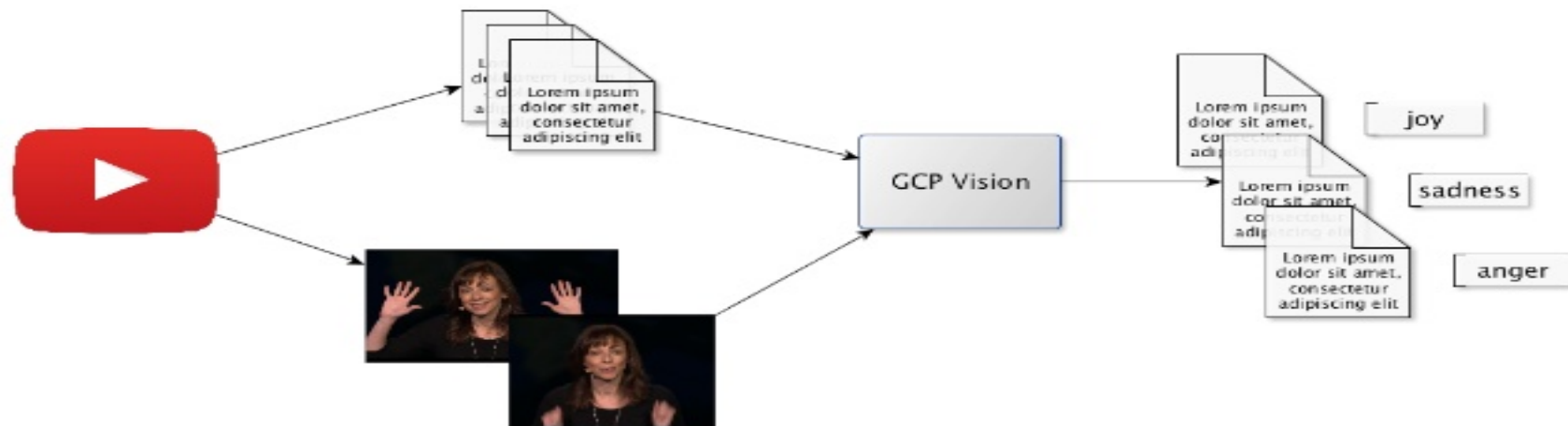


@texasmichelle



# Training

- Auto-generated

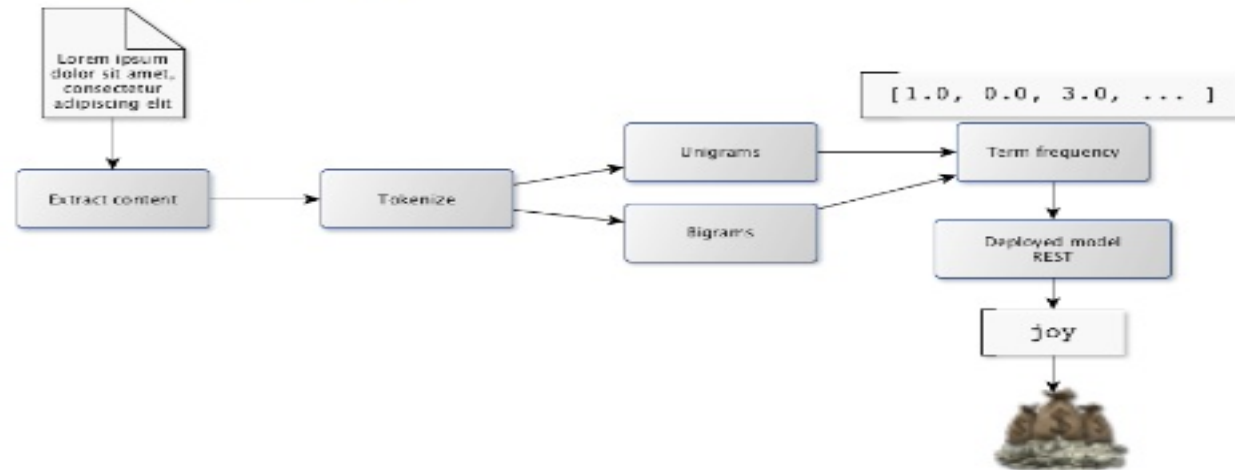


@texasmichelle



# Prediction

- Featurization
- Router
  - REST
  - Pub/Sub

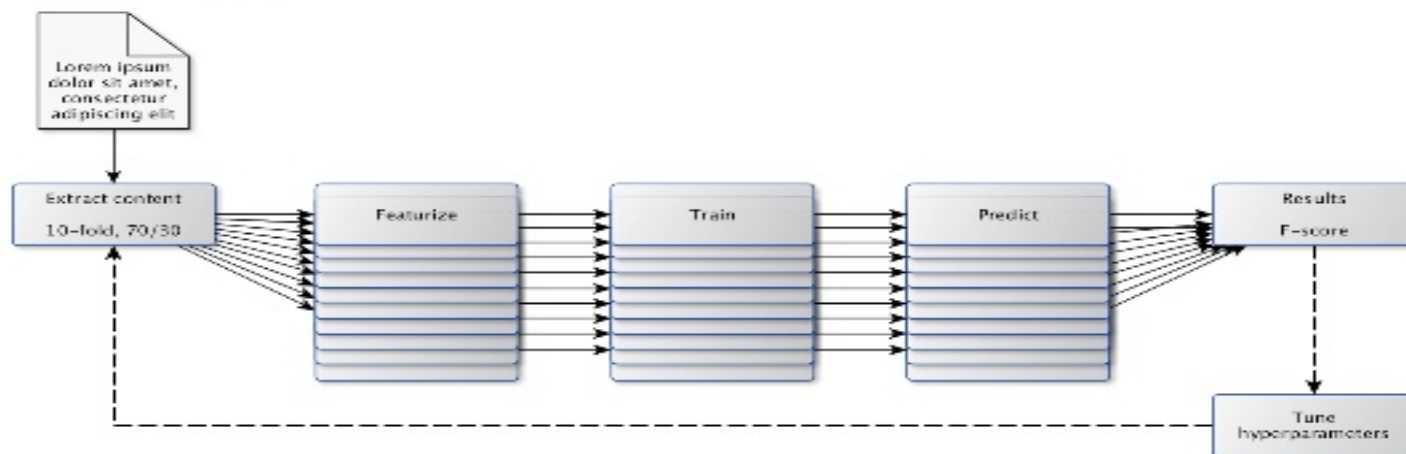


@texasmichelle



# Cross-validation

- F-score
  - Precision
  - Recall
  - Accuracy
- AUC



@texasmichelle





# Affect Detection

- More sophisticated featurization
- Multi-label instead of mutual exclusivity
- Confidence scores
- Additional algorithms 
- More evaluation metrics
- Automation 
- ~~Classification~~ Graph?



@texasmichelle



# Affect Detection

- Expanded training set
  - Language
  - Source
  - Filter
- Unsupervised models
- ~~All the domains~~ Generalization



@texasmichelle



# TL;DR

- There's a better way to ~~take over the world~~ do globalization
- What do you want to do tonight?
  - Affect detection!

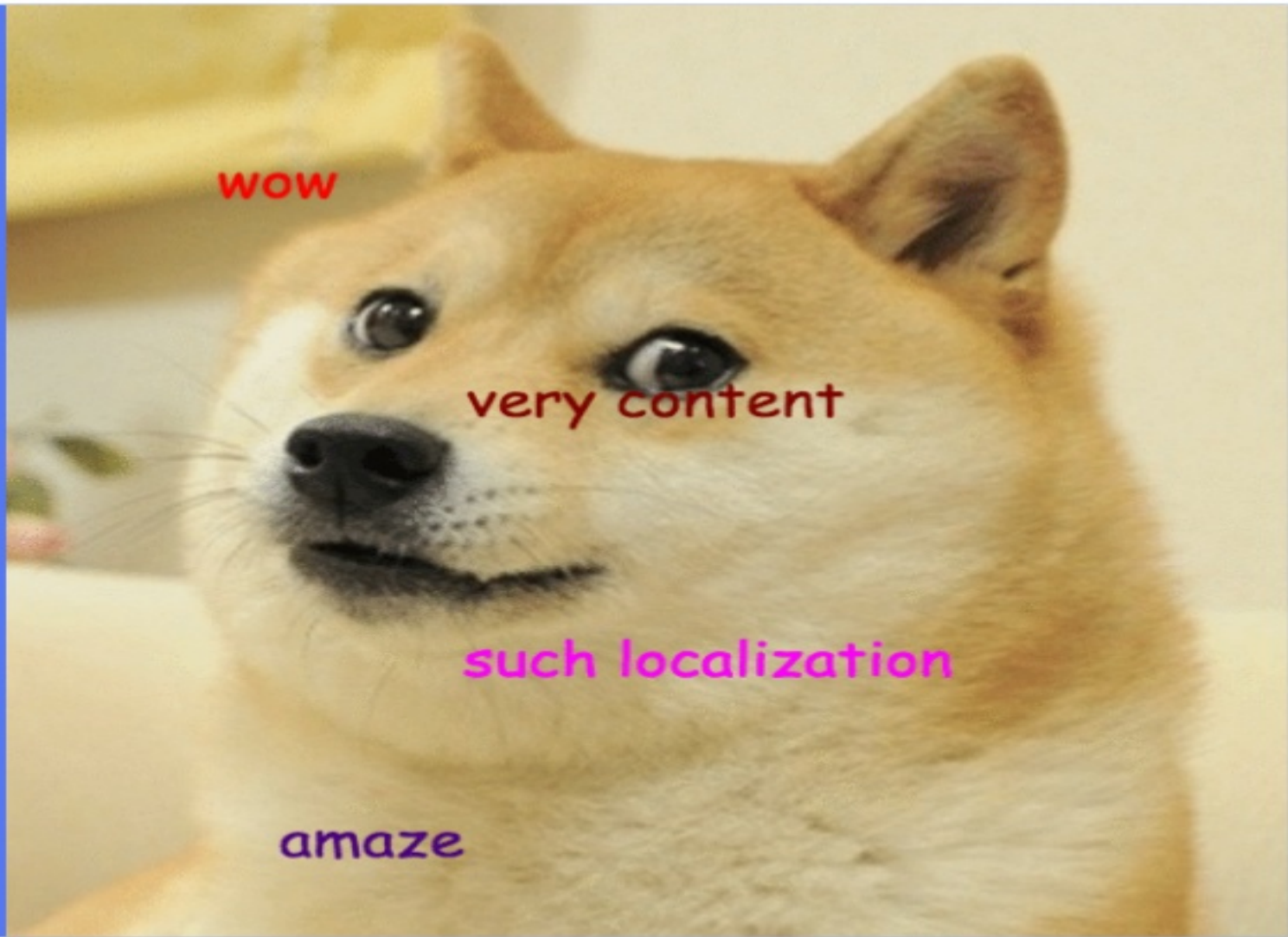


@texasmichelle



# Q&A

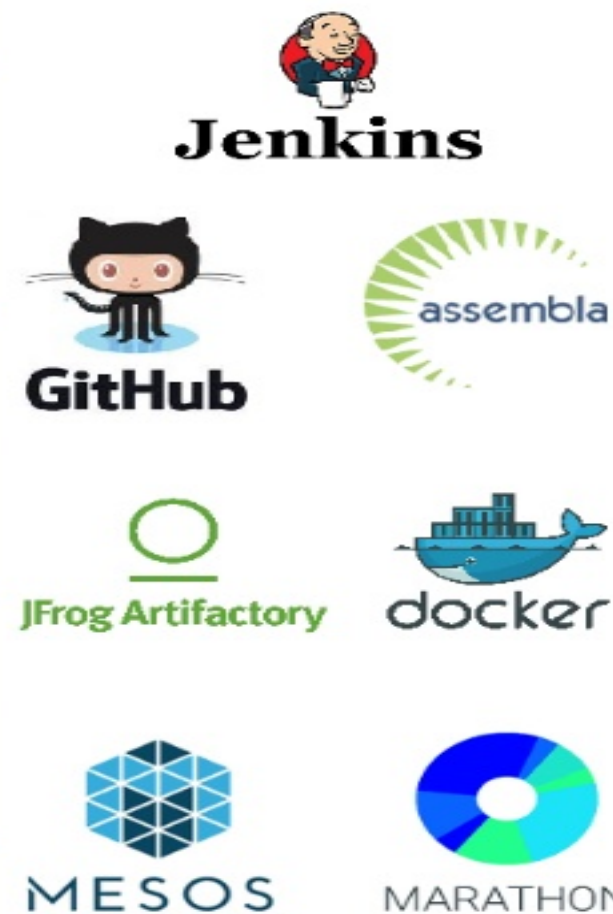
michelle@qordoba.com  
@texasmichelle

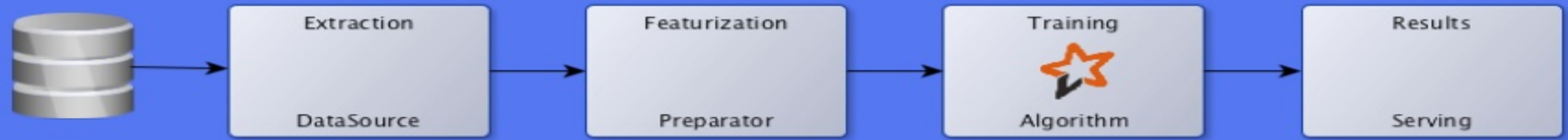




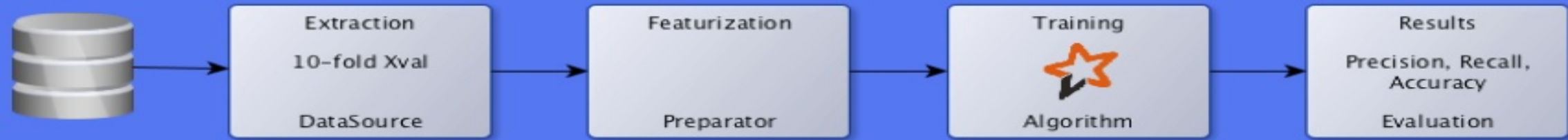


## Continuous Delivery





# Training with PredictionIO



# Cross-validation with PredictionIO