Open Source ML Systems That Need To Be Built

Nikhil Garg

@nikhilgarg28

Quora

#MLSummit 6/5/17

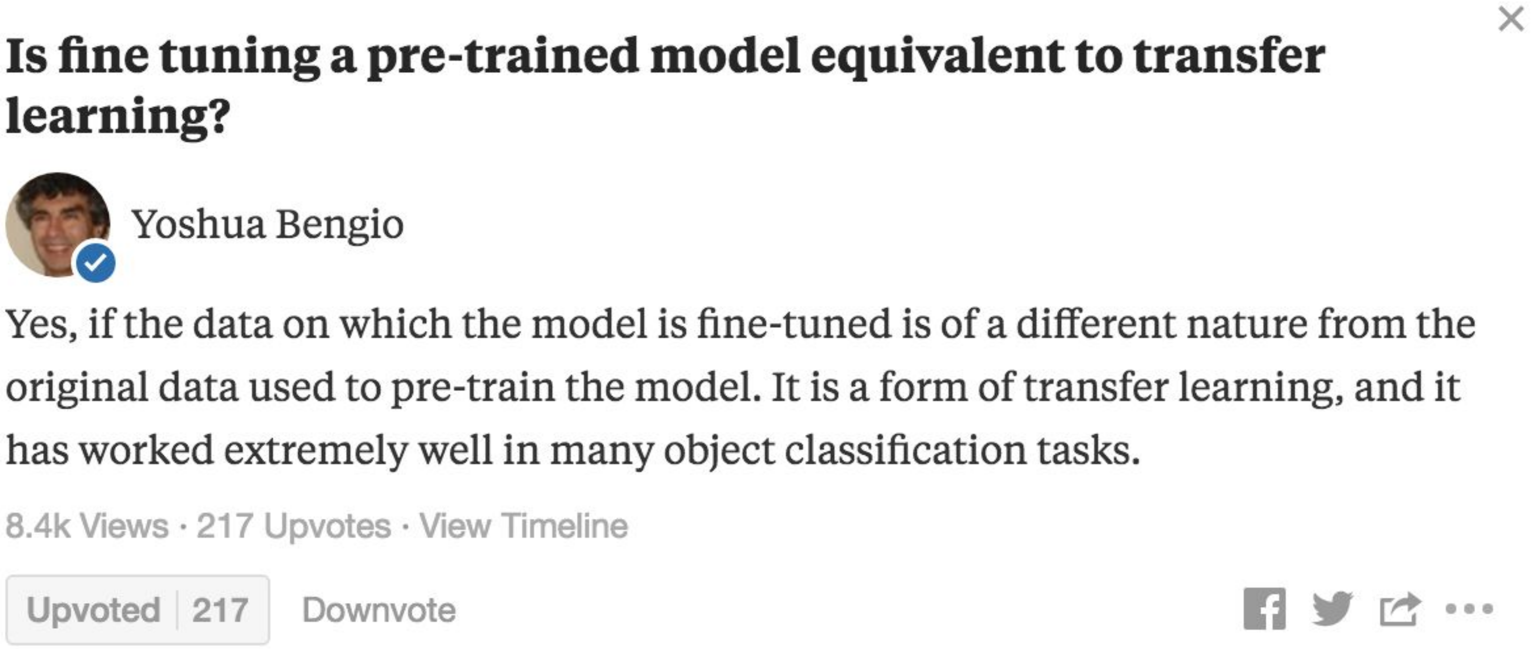
A bit about me...

- Currently leading two ML teams at Quora:
  - Ads
  - ML Platform
- Previously, led Content Quality and Core-product teams
- Interested in the intersection of distributed systems, machine learning and human psychology

@nikhilgarg28

Quora

The best answer to any question



Is fine tuning a pre-trained model equivalent to transfer learning?

Yoshua Bengio

Yes, if the data on which the model is fine-tuned is of a different nature from the original data used to pre-train the model. It is a form of transfer learning, and it has worked extremely well in many object classification tasks.

To Grow And Share World’s Knowledge

Over 200 million monthly uniques

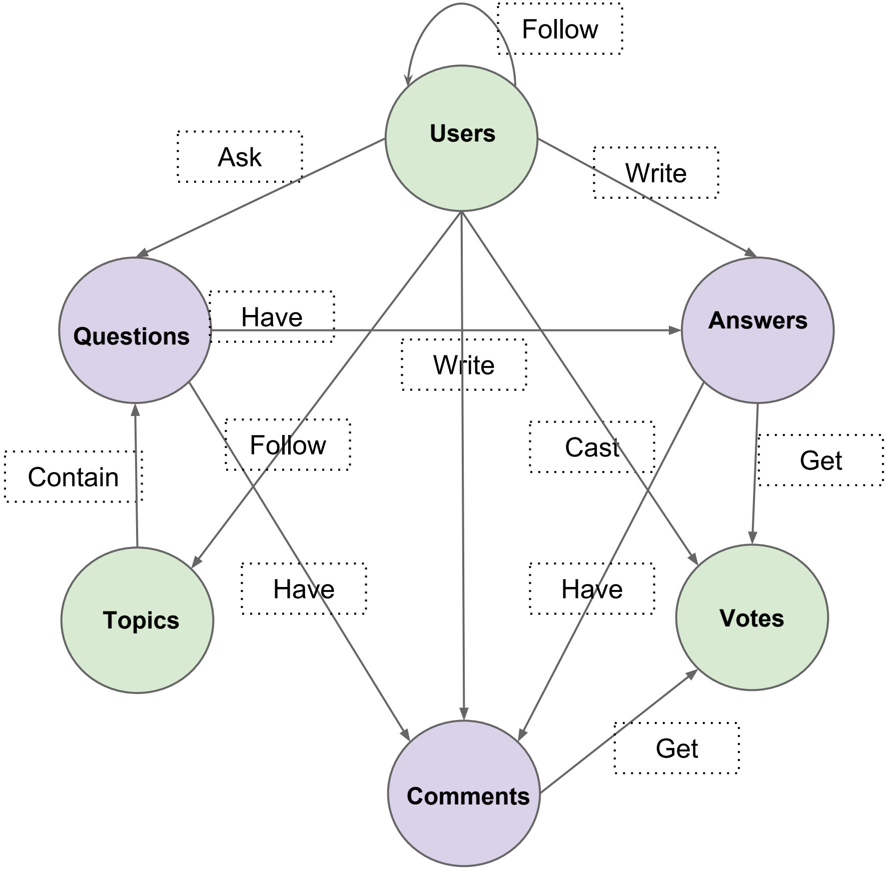
Millions of questions & answers

In hundreds of thousands of topics

Supported by < 100 engineers

ML @ Quora

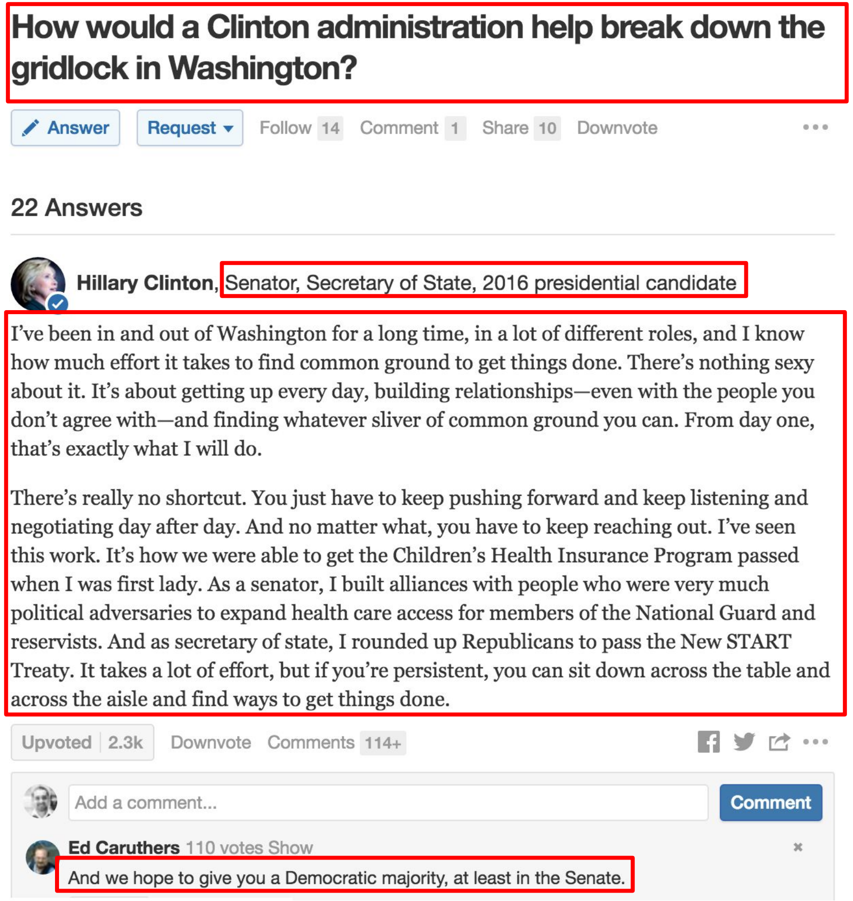
Data: Billions of relationships



Data: Billions of words in high quality corpus

Quora

- Questions
- Answers
- Comments
- Topic biographies
- ...



How would a Clinton administration help break down the gridlock in Washington?

22 Answers

Hillary Clinton, Senator, Secretary of State, 2016 presidential candidate

I've been in and out of Washington for a long time, in a lot of different roles, and I know how much effort it takes to find common ground to get things done. There's nothing sexy about it. It's about getting up every day, building relationships—even with the people you don't agree with—and finding whatever sliver of common ground you can. From day one, that's exactly what I will do.

There's really no shortcut. You just have to keep pushing forward and keep listening and negotiating day after day. And no matter what, you have to keep reaching out. I've seen this work. It's how we were able to get the Children's Health Insurance Program passed when I was first lady. As a senator, I built alliances with people who were very much political adversaries to expand health care access for members of the National Guard and reservists. And as secretary of state, I rounded up Republicans to pass the New START Treaty. It takes a lot of effort, but if you're persistent, you can sit down across the table and across the aisle and find ways to get things done.

Ed Caruthers 110 votes Show

And we hope to give you a Democratic majority, at least in the Senate.

Data: Interaction History

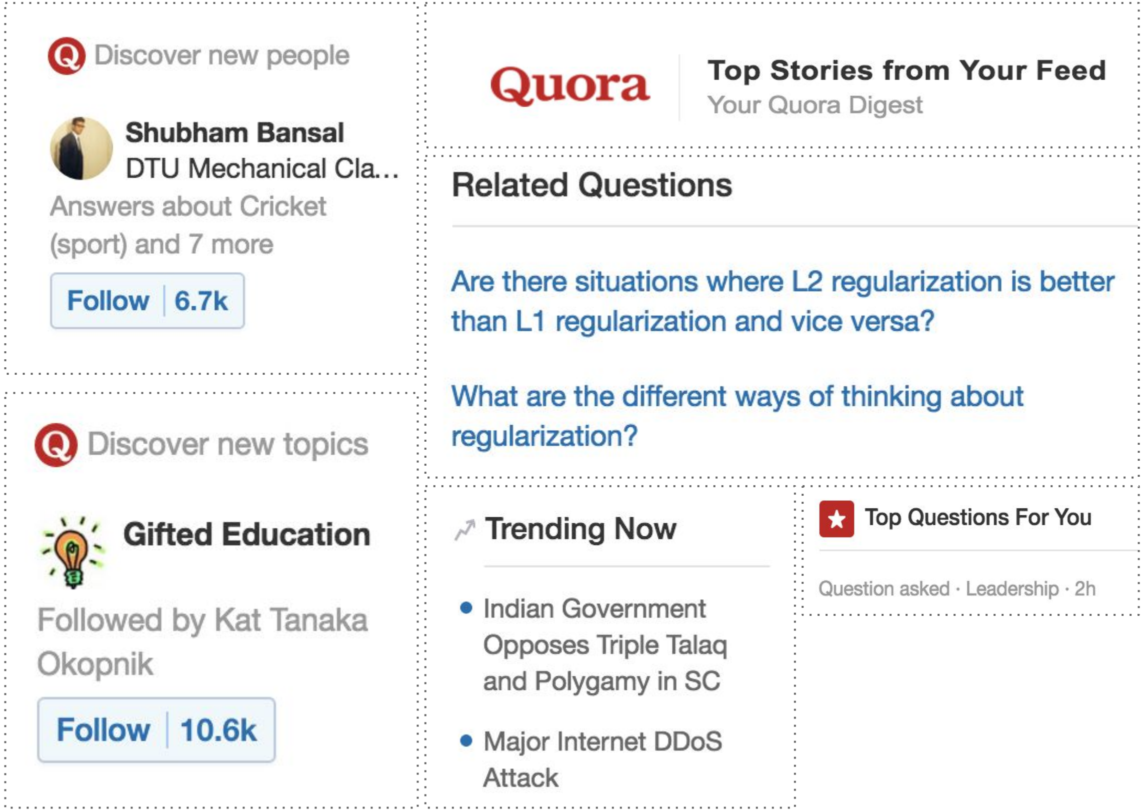
- Highly engaged users => long history of activity e.g search queries, upvotes etc.

- Ever-green content => long history of users engaging with the content in search, feed etc.

ML Applications At Quora

Quora

- Answer ranking
- Feed ranking
- Search ranking
- User recommendations
- Topic recommendations
- Duplicate questions
- Email Digest
- Request Answers
- Trending now
- Topic expertise prediction
- Spam, abuse detection
- ….



Discover new people

Shubham Bansal
DTU Mechanical Cla…
Answers about Cricket (sport) and 7 more

Follow | 6.7k

Quora | **Top Stories from Your Feed**
Your Quora Digest

Related Questions

Are there situations where L2 regularization is better than L1 regularization and vice versa?

What are the different ways of thinking about regularization?

Discover new topics

Gifted Education

Followed by Kat Tanaka Okopnik

Follow | 10.6k

Trending Now

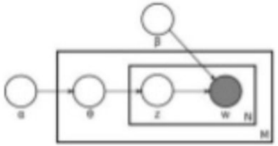
- Indian Government Opposes Triple Talaq and Polygamy in SC
- Major Internet DDoS Attack

★ Top Questions For You

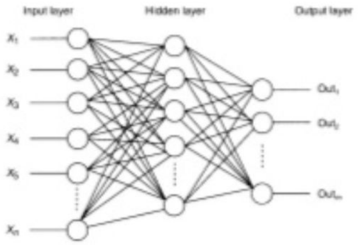
Question asked · Leadership · 2h

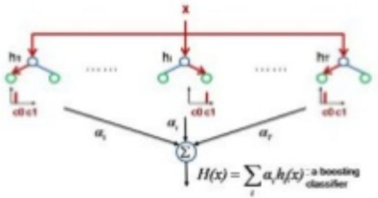
ML Algorithms At Quora

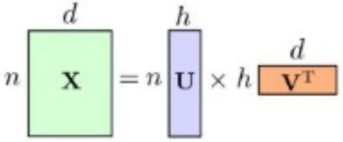
- Logistic Regression
- Elastic Nets
- Random Forests
- Gradient Boosted Decision Trees
- Matrix Factorization
- (Deep) Neural Networks
- LambdaMart
- Clustering
- Random walk based methods
- Word Embeddings
- LDA
- ...

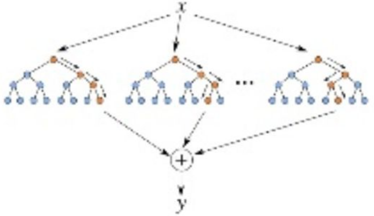


$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$



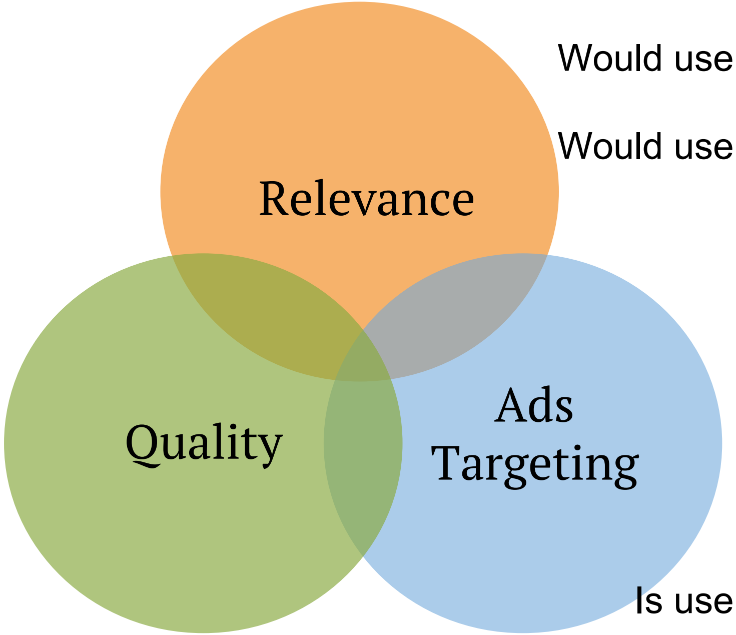






$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

What We Care About



Would user be interested in reading answer?

Would user be able to answer the question?

Is content high quality?

Is user an expert in the topic?

Is user deliberating a purchase decision?

Will user click on an ad?

ML As Quora’s Core Competency

- ML is the most promising tool for all our core problems
- ML can make our network effects even more powerful

Why ML Platform Team?

Why ML Platform Team?

1. Applied ML is bottlenecked on engineering

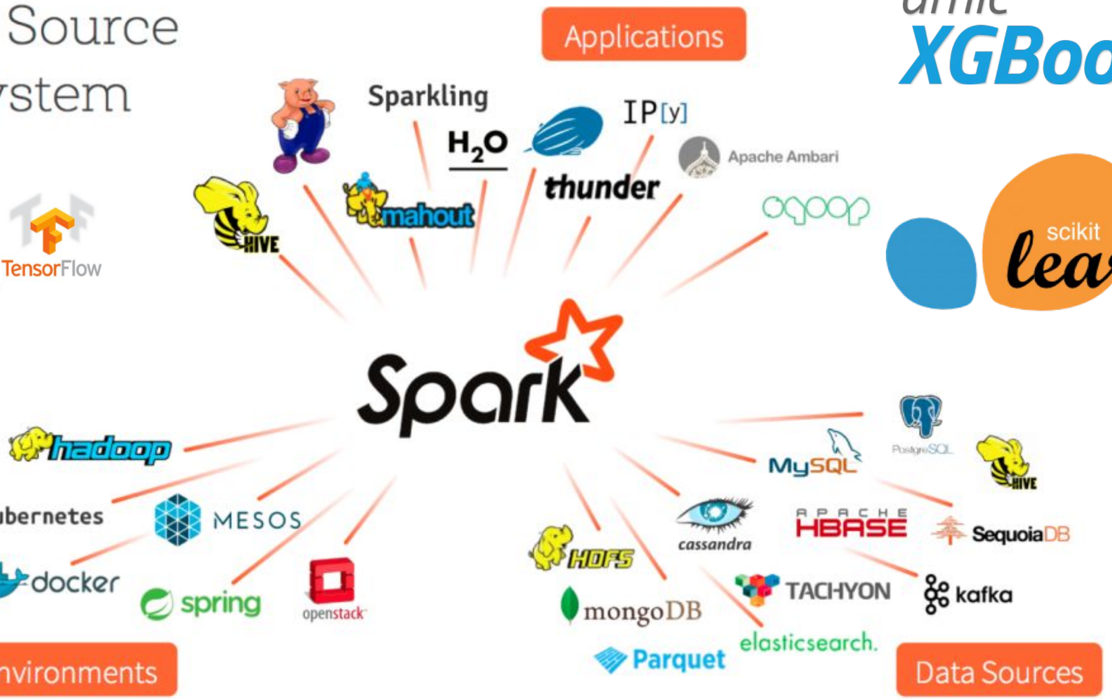
2. Most ML tasks require similar system primitives

Defining Times For ML Systems

Similar to Big Data 10-15 years ago

Open Source Ecosystem

dmlc
XGBoost



scikit
learn

GOAL

Mobilize Discussions In Open Source ML Systems Community

DISCLAIMER

All my ideas are probably wrong/unoriginal/incomplete

…and I’m shit scared right now!

1. Model Management

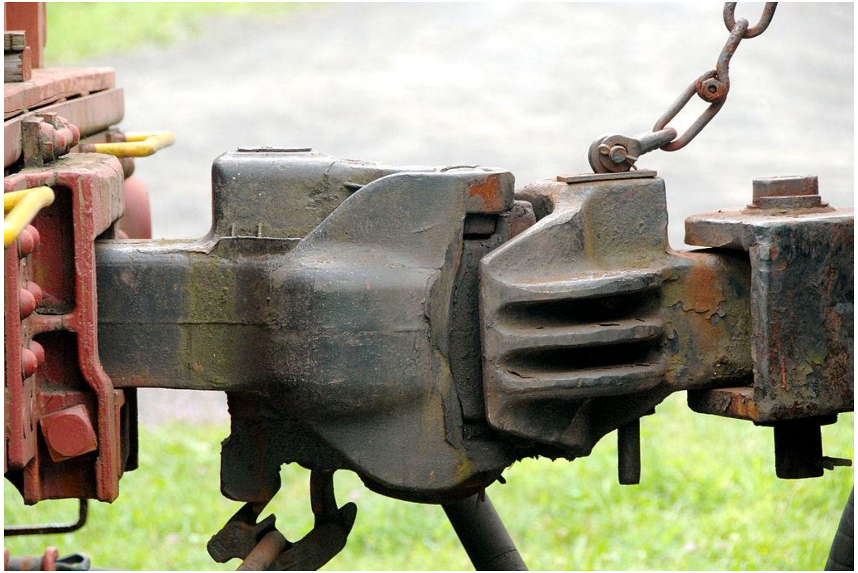
2. Feature Extraction Framework

1. **Model Management**

2. Feature Extraction Framework

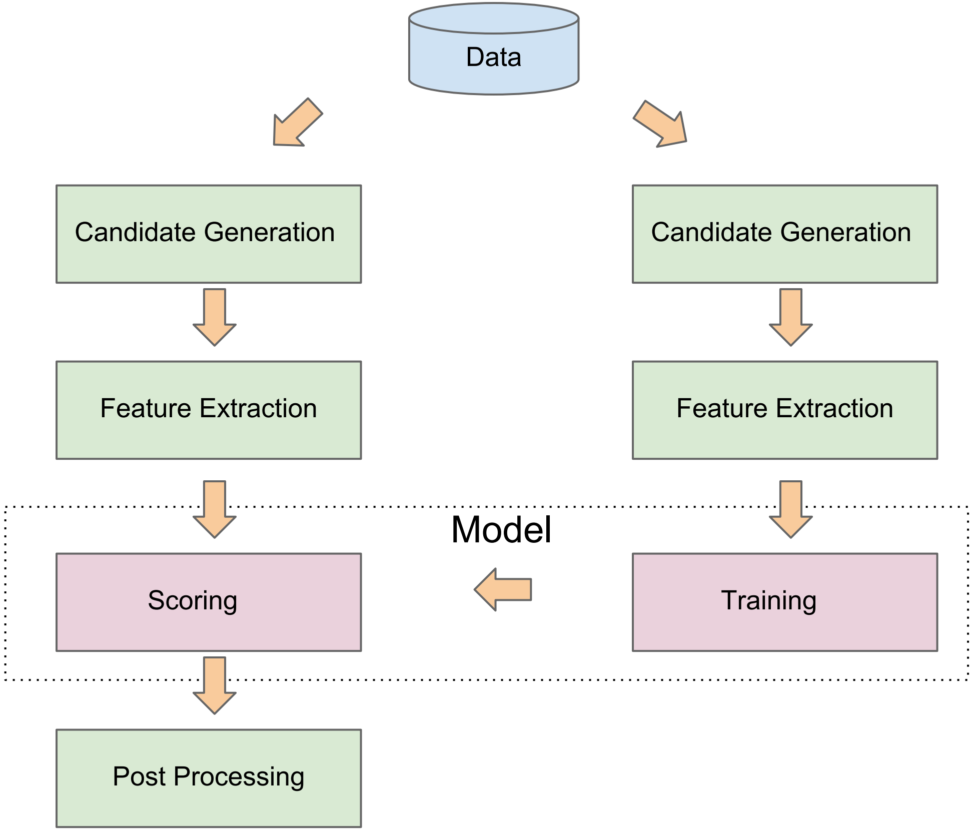
Sounds Familiar?

- Difficulty reproducing a model trained in R/Python in production on C++/Java
- Training using new library requires changing production too
- New library gives good metrics but is too slow in production
- Hard to manage too many versions of the same ML model in production

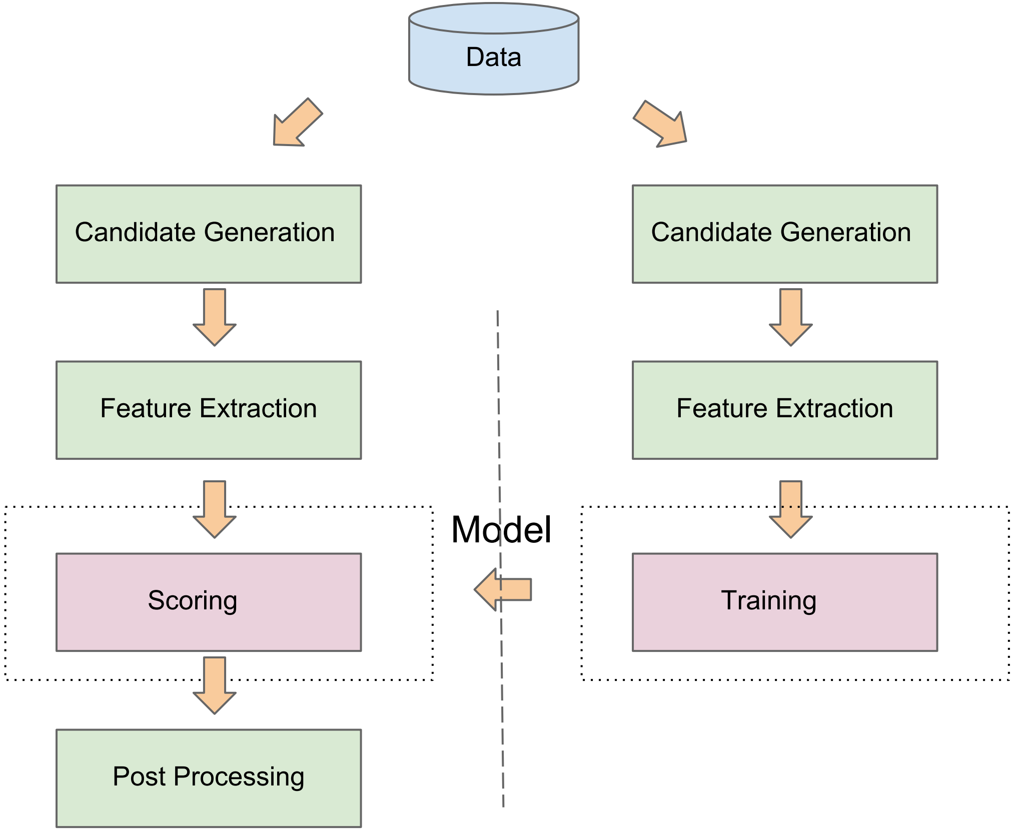


Coupling Between Model Training And Serving

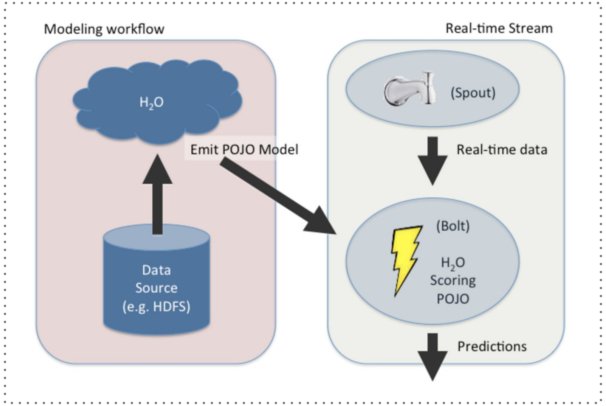
Quora

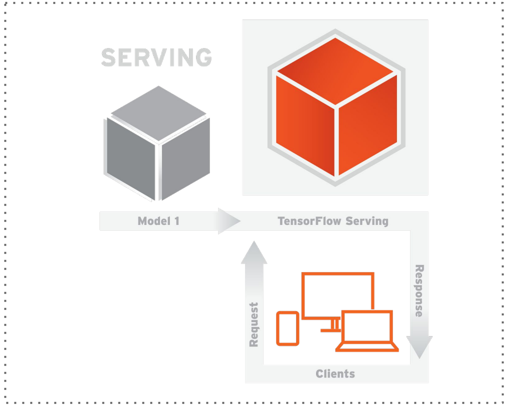


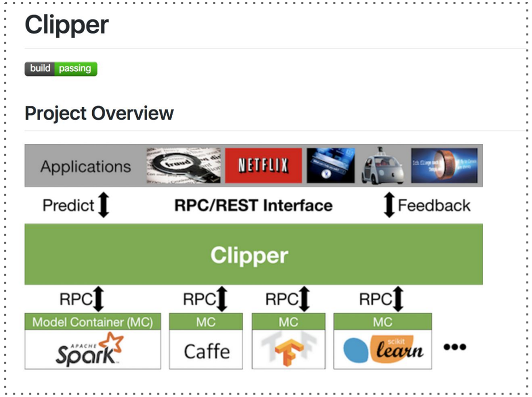
Coupling Between Model Training And Serving



Not a new idea...







MODEL

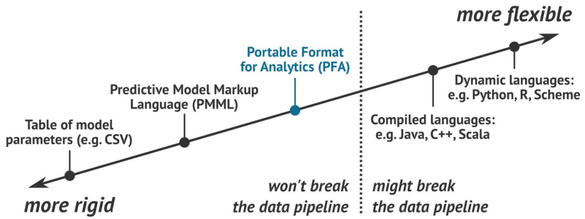
Collection (file) of learnt parameters

Universal model definition language

- Model files will be agnostic of training library/language

- Library plugins to convert existing models to a file in the universal model language

Language-agnostic production systems to serve models



Fast standardized serving

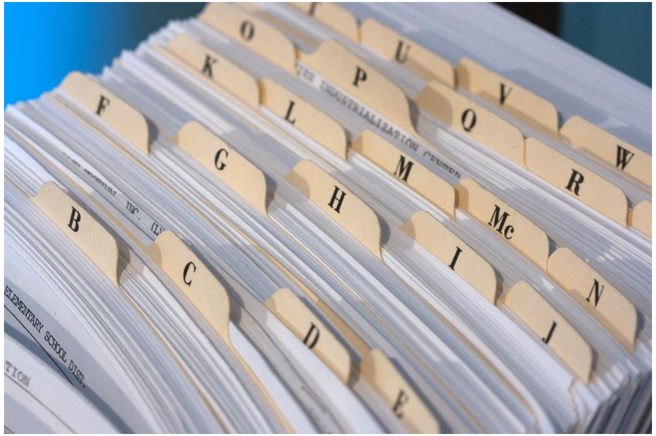
- A remote service usually works well and is sometimes necessary (e.g large memory footprint of a model)
- Local serving for cases where network round trip is too costly
- Fast standard model serving systems, supporting smart batching, GPU support etc.
- ‘Compiling’ the model for cases where interpreting it is too slow

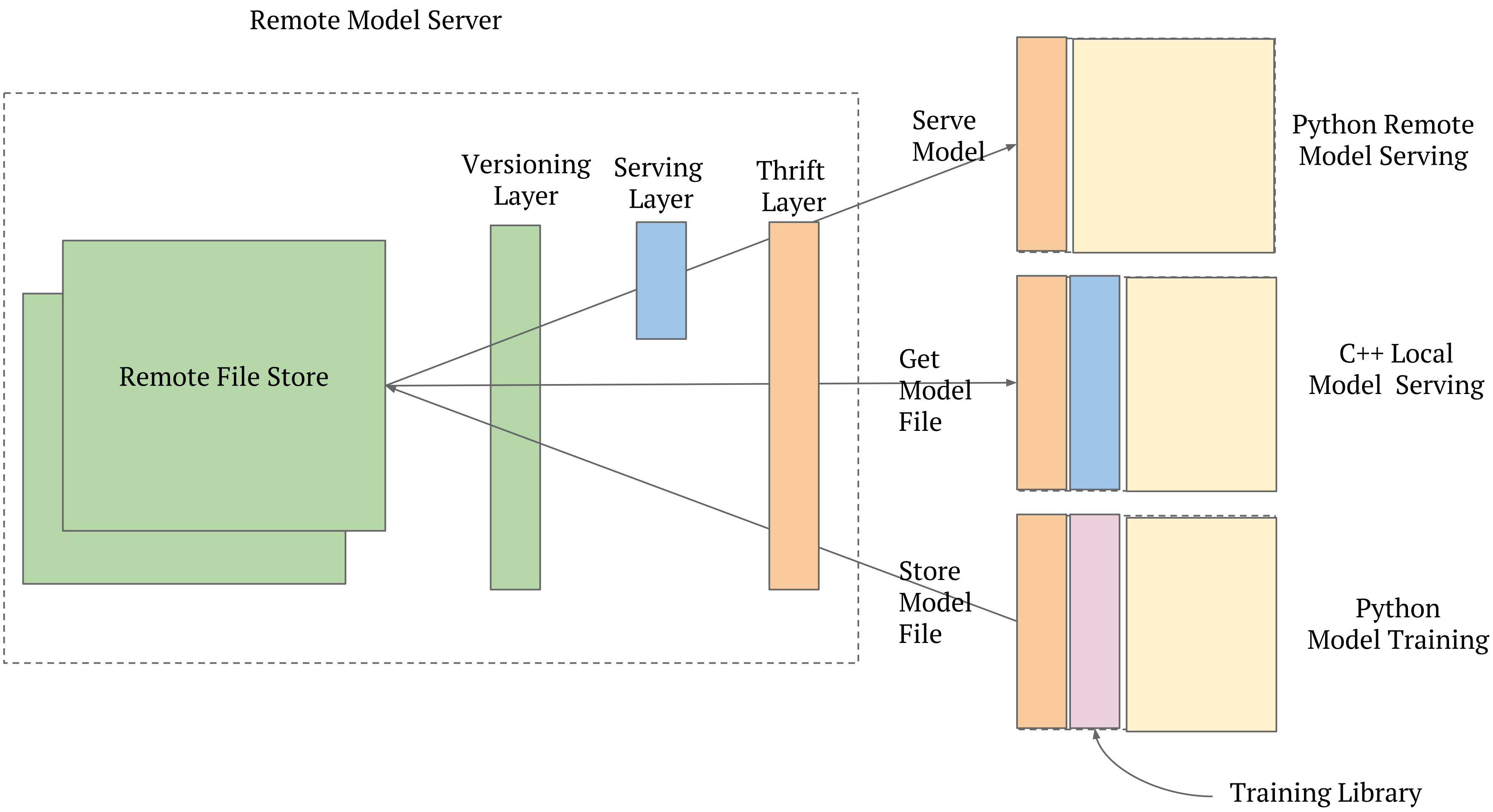
Versioning support

Quora

- Running multiple versions of a model - gradual roll outs, hot-swaps etc.

- Tensorflow serving does this very well, though need to add support for general model definition language.

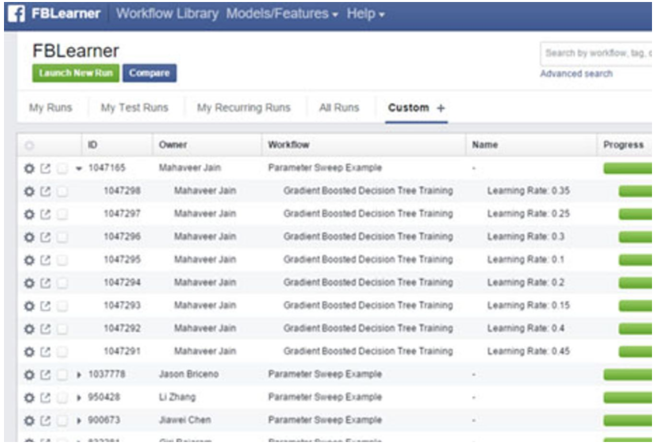



Remote Model Server
Versioning Layer
Serving Layer
Thrift Layer
Remote File Store
Serve Model
Get Model File
Store Model File
Python Remote Model Serving
C++ Local Model Serving
Python Model Training
Training Library

Model Repository

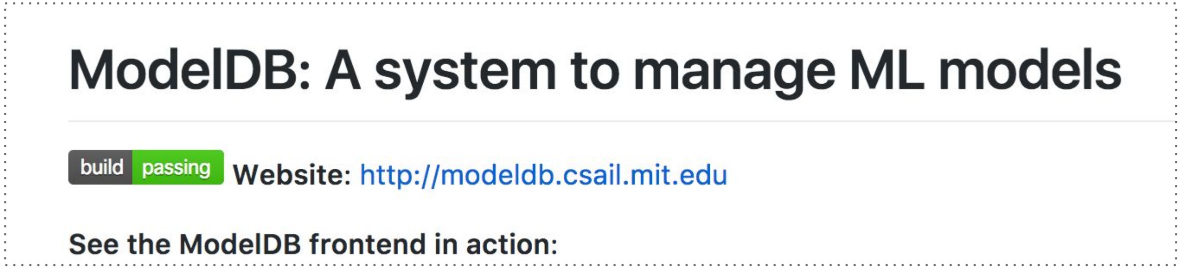
- Reproducibility -- could store features, hyper-parameters, algorithms, datasets and metrics used to train a model

- Repository of all previously trained models



FBLearner

ID	Owner	Workflow	Name	Progress
▾ 1047165	Mahaveer Jain	Parameter Sweep Example	-	🟩
1047298	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.35	🟩
1047297	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.25	🟩
1047296	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.3	🟩
1047295	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.1	🟩
1047294	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.2	🟩
1047293	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.15	🟩
1047292	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.4	🟩
1047291	Mahaveer Jain	Gradient Boosted Decision Tree Training	Learning Rate: 0.45	🟩
▸ 1037778	Jason Briceno	Parameter Sweep Example	-	🟩
▸ 950428	Li Zhang	Parameter Sweep Example	-	🟩
▸ 900673	Jiawei Chen	Parameter Sweep Example	-	🟩



ModelDB: A system to manage ML models

Website: http://modeldb.csail.mit.edu

See the ModelDB frontend in action:

Many Open Questions…

- Where does online-learning happen?
- Who takes care of the availability of the model service?
- Should versioning be a concern of the model service or the application?
- …

1. Model Management

2. **Feature Extraction Framework**

Sounds familiar?

- Diverging implementations of ‘BaseFeature’ classes
- Trouble discovering and reusing features across applications
- Problems integrating features across languages
- Hard to manage feature dependency graph, sometimes across applications and languages
- Ad-hoc testing/monitoring for feature values

```
class AnswerLength(BaseFeature):

    …

    def extract(self, aid):
           <some code>


    …
```

Feature extraction framework for standardization and reusability

Feature Extractors

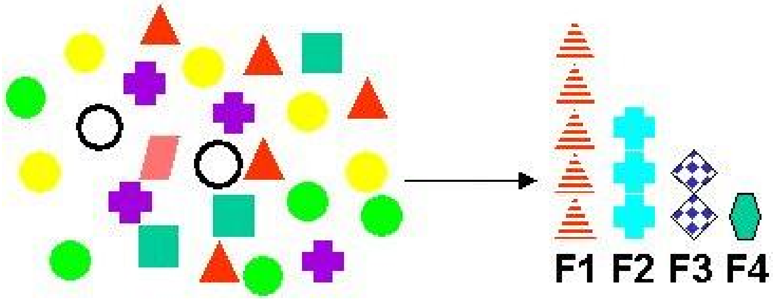
- Libraries/plugins for domain specific extractor building blocks e.g text, image, video

- Native support for distributed counting in a rolling window

- Feature transformers e.g log, bucketizer, centering, normalizing

- Encoders for categorical features e.g one-hot

- Combining multiple features e.g max, sum



Feature Storage And Serving

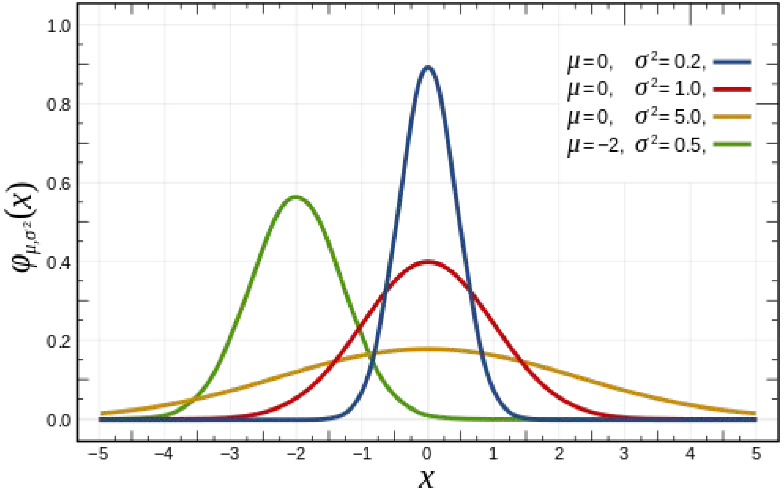
Quora

- Storage/caching/dirtying mechanisms
- Columnar storage for offline storage and training
- Central feature repository with discovery mechanism
- Central service serving all features behind language agnostic declarations
- Code can also be shipped to Spark workers

Feature Reliability

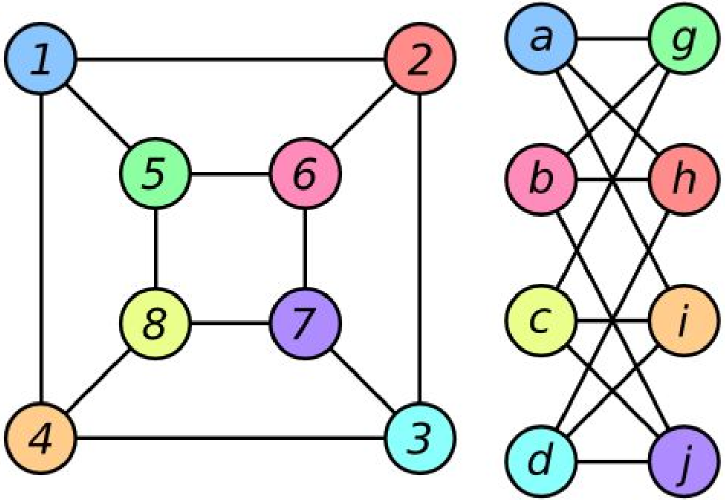
Quora

- Anomaly detection in feature value distributions
- Ground-truth feature tables
- Strong versioning support
- Feature debug/introspection UI



Models and features are functionally isomorphic

- Both models and features can depend on other features
- Features can work as a simple model
- Models can be a feature into another model
- Both need similar tooling support -- versioning, monitoring, debugging, repository etc.



https://proopnarine.wordpress.com/2010/02/20/random-graphs-and-food-webs/

Summary

- Defining times for ML Systems space
- Need powerful abstractions higher up in the ML stack
- Model management & feature extraction could use more open-source love
- Models & features are more similar than we might think

Thank You!

YES, WE ARE HIRING :)

Nikhil Garg
@nikhilgarg28