

# Scaling Quality On Quora Using Machine Learning

Nikhil Garg

@nikhilgarg28

@Quora

@QconSF 11/7/16

- Introducing specific product problems we need to solve to stay high-quality
- Describing our formulation and approach to these problems.
- Identifying common themes of ML problems in the quality domain.
- Sharing high level lessons that we have learnt over time.

# A bit about me...

- At Quora since 2012
- Currently leading two engineering teams:
  - ML Platform
  - Content Quality
- Interested in the intersection of distributed systems, machine learning and human behavior



@nikhilgarg28

# Quora

The best answer to any question



**To Grow And Share World's Knowledge**

# What good things does the TPP hold for underemployed American citizens?

Will this trade agreement raise the U.S. workforce participation rate?



Follow

6

Comment

Share

Downvote



## 9 Answers

---



**Barack Obama**, President of the United States

Written Thu · Featured in Inc · Upvoted by Marc Bodnick, [Harvard Gov major, Stanford PoliSci PhD student, Nadia Singer, and 27 others you follow](#)

The underemployed – folks who work part-time and want to work full-time – have a lot to gain from the Trans-Pacific Partnership, or TPP. I'm not the only one who believes this – a number of independent, credible analyses of TPP have shown it will grow exports, grow our economy, and raise incomes.

Underemployment and unemployment are driven by an economy that's not producing at its full potential – factories that aren't humming at capacity and workers who aren't being out to their highest and best use. It makes sense – fewer goods, services, and exports

# How did Ashton Kutcher prepare for his role as Steve Jobs in the new movie Jobs?

 Answer

Request ▾

Follow

538

Comments

13+

Share

54

Downvote



## 5 Answers




---



**Ashton Kutcher**, I live it



I spent about 3 months preparing the character.

I started by consuming content about Steve Jobs. The script was a fantastic resource but after reading it I was left with as many questions as answers as to why he was the way he was and why he made some of the decisions he made. I started by watching documentaries and interviews [Silicon Valley Historical Association](#)  about him and collecting youtube content [Inspired By Jobs: Technology](#)  and Soundcloud files on [jobs](#) .

This was to try to understand some of the broad themes of his persona. What I was looking for was natterns on consistent behavior and ideals. I heard him reneat his storv

# Does grunting help tennis players improve their game?

 Answer

Request ▼

Follow 15

Comment

Share

Downvote



## 11 Answers

---



**Pat Cash**, 1987 men's Wimbledon champion

Updated Sep 11, 2013 · Upvoted by Nadia Singer, Robert Cezar Matei, and 13 others you follow

I recently wrote an article about this titled "[What Everyone Ought to Know About Tennis Grunting](#)" which dealt specifically with this.

I also answered a similar question here: [Why do female professional tennis players appear to grunt when they hit the ball, while males do not?](#)

In short, exhaling while swinging towards the ball does help tennis players improve their game. Making noise while breathing out is not required but is often natural to players.



# What is the best way to educate kids not to ask for things while crying?



Follow

107

Comment

Share

25

Downvote



## 73 Answers



**Arie Anna Farnam**, former child, parent, ESL preschool teacher



Written May 18, 2015

Essentially you don't respond if they're screaming at you, unless it involves blood or fire. And you tell them that you won't respond until they stop screaming... when they take a breath so they can theoretically hear you. And you have consistent routines set up so that they know the consequences of various things, including when they are going to have to go someplace away from people to calm down.

And if you do that with 99 percent consistency and almost never make a mistake, you will be SLIGHTLY better off than the parents who go at it without thinking and just appease screaming kids to get peace in the moment. And you'll get criticized by know-it-alls with no

# What are the best Caltrain hacks?

 Answer

Request ▾

Follow 72

Comment 1

Share Downvote



## 5 Answers




**Matt Laroche**, Ex-Caltrain Regular.

Updated Feb 25 · Upvoted by Jeremy Lipps, I am the Social Media Officer for Caltrain., Nadia Singer, and 23 others you follow



### Tickets/Passes

- Load your Caltrain monthly (and other passes/credit) at Walgreen's (or another Clipper retail location: <https://clippercard.com/ClipperW...> ). It avoids the 3 day wait when you buy a pass online. This also means you can use your monthly without doing the zone based tagging in and out. (There's a few edge cases that fail with online purchase of monthly passes)
- Monthly passes are valid for all zones on weekends and holidays!
- 2+ zone monthly passes are treated, basically, as monthly passes for VTA and

**Over 100 million monthly uniques**

**Millions of questions & answers**

**In hundreds of thousands of topics**

**Supported by 80 engineers**

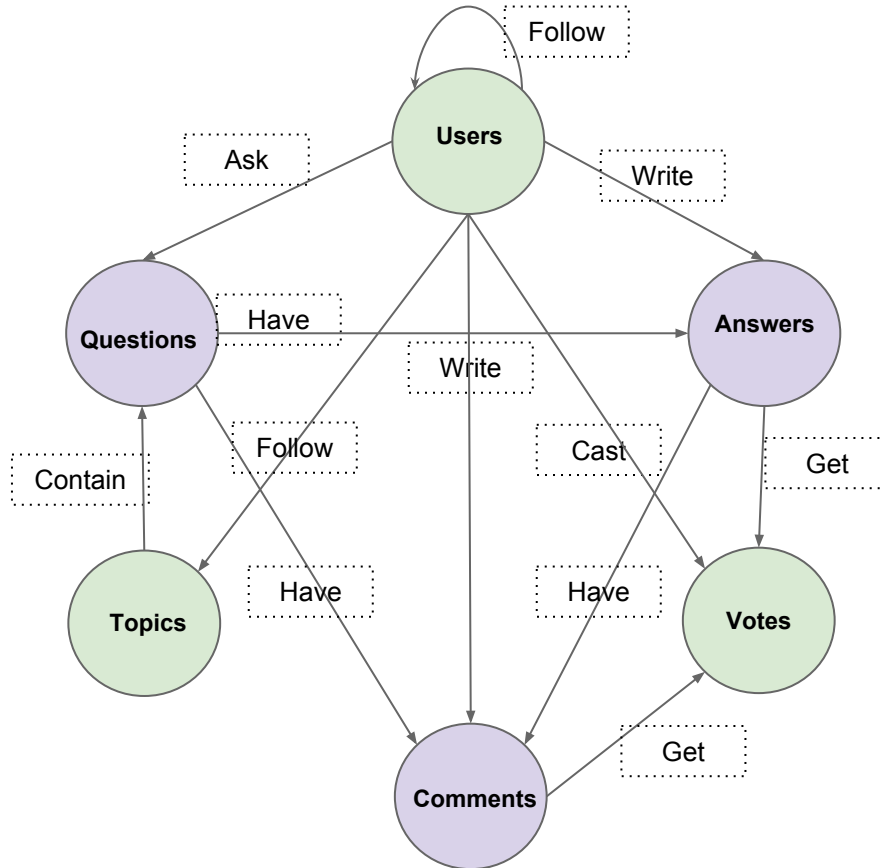
**ML @ Quora**

# ML's Importance For Quora

---

- ML is not just something we do on the side, it is mission critical for us.
- It's one of the most important core competencies for us.

# Data: Billions of relationships



# Data: Billions of words in high quality corpus



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

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

[Answer](#) [Request](#) Follow 14 Comment 1 Share 10 Downvote ...

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.

Upvoted 2.3k Downvote Comments 114+ ...



Add a comment...

Comment



**Ed Caruthers** 110 votes Show

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


- 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




- 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

 Discover new topics

 **Gifted Education**

Followed by Kat Tanaka Okopnik

[Follow](#) | 10.6k

 **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?](#)

 **Trending Now**

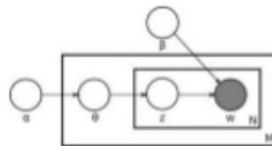
- Indian Government Opposes Triple Talaq and Polygamy in SC
- Major Internet DDoS Attack

 **Top Questions For You**

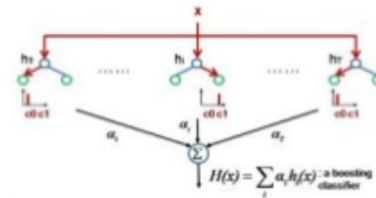
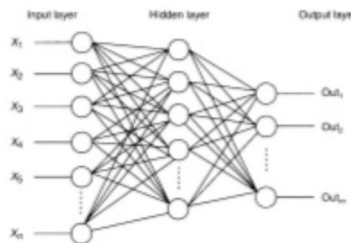
Question asked · Leadership · 2h

# ML Algorithms At Quora

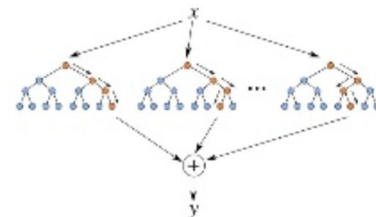
- Logistic Regression
- Elastic Nets
- Random Forests
- Gradient Boosted Decision Trees
- Matrix Factorization
- (Deep) Neural Networks
- LambdaMart
- Clustering
- Random walk based methods
- Word Embeddings
- LDA
- ...



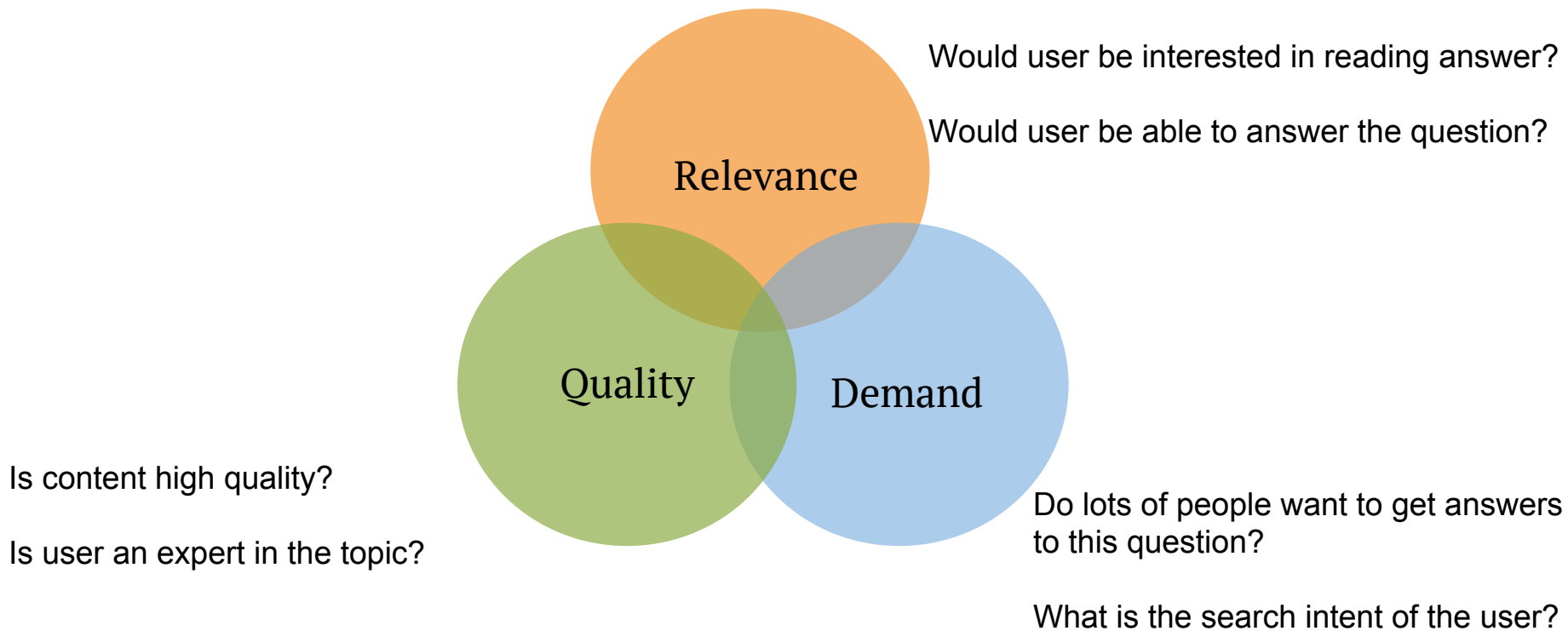
$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

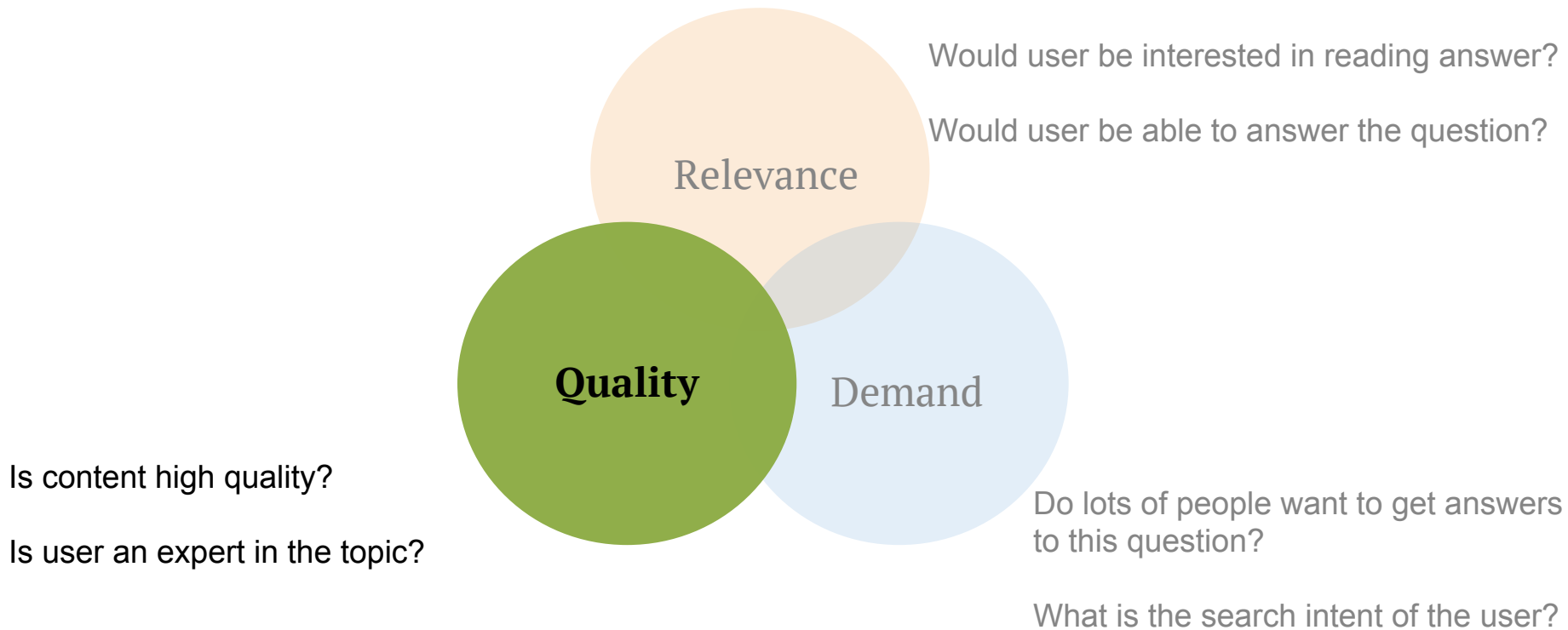


$$n \begin{matrix} d \\ \mathbf{X} \end{matrix} = n \begin{matrix} h \\ \mathbf{U} \end{matrix} \times h \begin{matrix} d \\ \mathbf{V}^T \end{matrix}$$



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





1. Duplicate Question Detection
2. Answer Ranking
3. Topic Expertise Detection
4. Moderation

**1. Duplicate Question Detection**

2. Answer Ranking

3. Topic Expertise Detection

4. Moderation

# Why Duplicate Questions Are Bad



- Energy of people who can answer the question gets divided
- No single question page becomes the best resource for that question
- People looking for answers have to search and read many question pages
- Bad experience if the “same” question shows up in feed again and again
- Search engines can not rank any one page very highly

# Duplicate Question Detection

**Check For Existing Questions** ✕

"what is the sun's temperature?"

---

**What is the temperature of the Sun's surface?**

**What is the sun's pressure, temperature and volume?**

How is the temperature of the Sun's core ascertained?

Is the temperature of the sun increasing?

**What is the sun's core density?**

---

Write Question Details

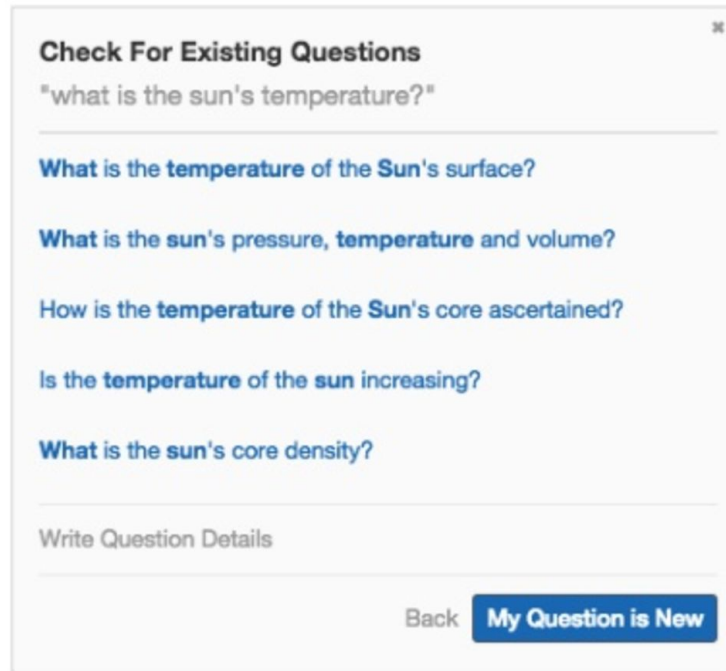
---

[Back](#) [My Question is New](#)



# Duplicate Question Detection

- Need to detect duplicates even before question reaches the system
- When user adds a question, we search ALL our questions to check for duplicates.
- Latency: tens of milliseconds.
- ML algorithm aside, this is also a crazy hard engineering problems.



## Problem Statement

**Detect if a new question is duplicate of an existing question.**

- Syntax
- Semantics
- Generality
- High precision
- High recall

What is the Sun's temperature?

How hot does the Sun get?

What is the average temperature of Sun?

What is the temperature of Sun's surface and that of Sun's core?

What is the hottest object in our solar system? How hot is it?

What is the temperature, pressure and density of Sun?

What is the temperature of a yellow star like our Sun?

## Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks

**Kai Sheng Tai, Richard Socher\*, Christopher D. Manning**

Computer Science Department, Stanford University, \*MetaMind Inc.

kst@cs.stanford.edu, richard@metamind.io, manning@stanford.edu

---

## Distributed Representations of Sentences and Documents

---

**Quoc Le**

**Tomas Mikolov**

Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043

QVL@GOOGLE.COM

TMIKOLOV@GOOGLE.COM

## Illinois-LH: A Denotational and Distributional Approach to Semantics

**Alice Lai and Julia Hockenmaier**

Department of Computer Science

University of Illinois at Urbana-Champaign

{aylai2, juliahmr}@illinois.edu

## Siamese Recurrent Architectures for Learning Sentence Similarity

**Jonas Mueller**

Computer Science

& Artificial Intelligence Laboratory

Massachusetts Institute of Technology

**Aditya Thyagarajan**

Department of Computer Science

and Engineering

M. S. Ramaiah Institute of Technology

## **Problem Formulation**

Binary classification on pairs of questions

## **Training Data Sources**

Hand labeled data, Semi-supervised approaches to bootstrap data, Random negative sampling, User browse/search behavior, Language model on standard datasets, ...

## **Models**

Logistic Regression, Random Forests, GBDT, Deep Neural Networks, Ensembles...

## **Features**

Word embeddings, conventional IR features, usage based features, ...

# Duplicate Questions: Problem Properties



- Judgements are hairy for even humans.
- Can't optimize some user action directly.
- Training data is scarce -- need to fuse multiple data sources together.

1. Duplicate Question Detection
2. **Answer Ranking**
3. Topic Expertise Detection
4. Moderation

## Is A.I. an existential threat to humanity?



**Oren Etzioni**, CEO of the Allen AI, Founder of Farecast, Professor at UW, CSE

2.7k Views

Originally Answered: Is AI an existential threat to humanity?

Andrew Ng famously said **"worrying about AI turning evil is like worrying about overpopulation on Mars"**.

Hollywood (and the press's) obsession with an "existential threat" ignores the following:

- The formidable technical challenges of producing "real" AI systems as opposed to narrowly focused "savants" like AlphaGo or Deep Blue.
- The huge potential benefits of AI to society in preventing car accidents, medical errors, and more.

In fact, the obsession with a hypothetical long-term scenario distracts from the real concerns about AI and its potentially disruptive impact on jobs, on weapons systems, etc.

I talk more about this in this brief article: [AI won't exterminate us. It will empower us. — Backchannel](#)

Also, remember that there will not be a single AI---multiple AI systems can offer "checks and balances" on each other. See: [To Keep AI Safe — Use AI](#)

Written Apr 8 · View Upvotes · Answer requested by Xavier Amatriain, Sizheng Chen (陈思政), and 22 more

Upvote 40

Downvote

Comments 2



## Is A.I. an existential threat to humanity?



**Sharan Gala**, I sincerely hope that we will be alright.

31 Views

Originally Answered: Do you think Machine Learning pose an existentialist threat to humanity?

No.

Humanity poses a threat to humanity.

We invented machine learning. It is our responsibility to use it wisely.

Machine learning itself is a small segment of what makes up Full A.I.

And we haven't even come close to making Full A.I. yet.

A suggestion is that if we stack several machine learning algorithms together, they will keep refining input at the level of a human, then we can begin to get a computer to understand awareness through density of knowledge.

But all of this is in a future at least twenty years from now. And when it will come true, it will surely have safeguards built into it.

And if not safeguards, I doubt scientists will create a shitty persona for the computer.

It might just be the nicest conscience to ever grace the planet. Perhaps it helps us achieve our goals instead of trying to kill us!

Written Apr 20

Upvote

Downvote

Comment





## Problem Statement

**Rank answers to a question by their “quality”**

## **Learning to Rank Answers on Large Online QA Collections**

**Mihai Surdeanu, Massimiliano Ciaramita, Hugo Zaragoza**

Barcelona Media Innovation Center, Yahoo! Research Barcelona

`mihai.surdeanu@barcelonamedia.org, {massi, hugo}@yahoo-inc.com`

## **Ranking Community Answers by Modeling Question-Answer Relationships via Analogical Reasoning**

Xin-Jing Wang  
Microsoft Research Asia  
4F Sigma, 49 Zhichun Road  
Beijing, P.R.China  
`xjwang@microsoft.com`

Xudong Tu<sup>\*</sup>, Dan Feng  
Huazhong Sci.&Tech. Univ.  
1037 Luoyu Road, Wu Han  
Hu Bei, P.R.China  
`{tuxudong, dfeng}@hust.edu.cn`

Lei Zhang  
Microsoft Research Asia  
4F Sigma, 49 Zhichun Road  
Beijing, P.R.China  
`leizhang@microsoft.com`

## Previous Approach

**A simple function of upvotes and downvotes,  
with some precomputed author priors.**

# Great Baseline, but...



- Popular answers != factually correct
- Joke answers get disproportionately many upvotes
- Expert answers ranked lower than answers by popular writers
- Rich get richer
- Poor ranking for new answers
- ...

# Why do all these problems exist?



- Upvote means different things to different people e.g funny, correct, useful.
- Doesn't always correspond to quality
- ...what is quality?

# Defining High Quality Answers



- Answers the question
- Is factually correct
- Is clear and easy to read
- Supported with rationale
- Demonstrates credibility
- ...

## Where is deep learning research going towards?



**Yoshua Bengio**, My lab has been one of the three that started the deep learning approach, bac...

18.8k Views · Featured in The Huffington Post · Upvoted by Xavier Amatriain, Research Scientist and manager for over 10 years both in academia and industry, Nikhil Garg, I lead a team of Quora engineers working on ML/NLP problems, Alberto Bietti, and 23 others you follow  
Most Viewed Writer in Machine Learning with 90+ answers

Originally Answered: Yoshua Bengio: Where is deep learning research headed?

Research is by definition exploratory, which means that (a) we do not know what will work and (b) we need to explore many paths, we need a lot of diversity of research directions in the scientific community. So I can only tell you about my current gut feelings and visions of where I see important challenges and opportunities that appeal to my personal aesthetics and instincts. Here are some elements of this:

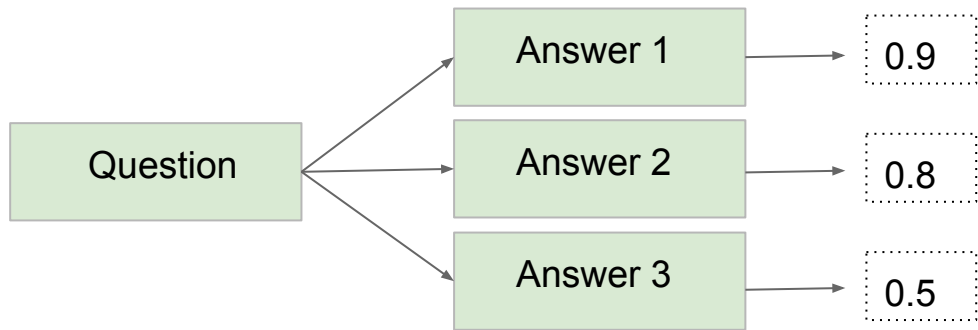
\* unsupervised learning is crucial and we do not do it right yet (there are many arguments I and others have written and talked about to justify this)

\* deep learning research is likely to continue its expansion from traditional pattern recognition jobs to full-scale AI tasks involving symbolic manipulation, memory, planning and reasoning. This will be important for reaching to full understanding of natural language and dialogue with humans (i.e., pass the Turing test). Similarly, we are seeing deep learning expanding into the territories of reinforcement learning, control and

Upvote 291 Downvote Comments 2+



- Item-wise regression on answers.
- Also tried item-wise multi-class classification on score buckets
- Item-wise enable comparing answers across different questions.
- Can also discover “Quora Gold” and “really bad” answers



- R2
- Weighted R2 with different weights for different parts of the quality spectrum
- NDCG
- Kendall's Tau
- ...

$$DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

$IDCG_k$  is the maximum possible (ideal)  $DCG$  for a given set of queries, documents, and relevances.

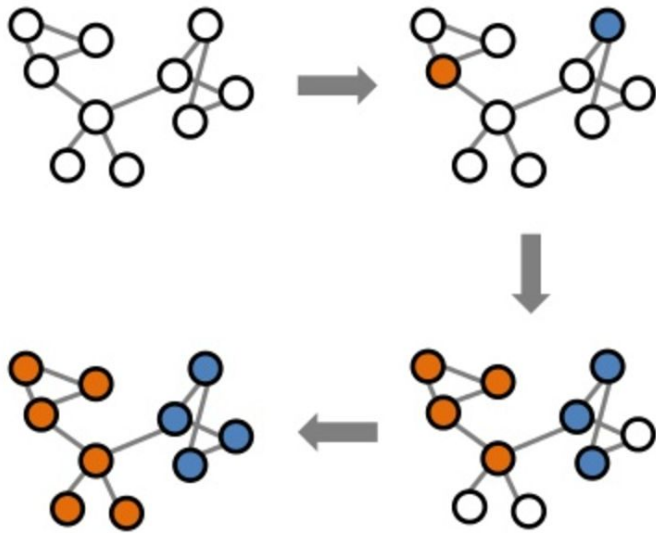
$$nDCG_k = \frac{DCG_k}{IDCG_k}$$



# Answer Ranking: Training Data

- Hand labeled data
- Language model on standard datasets
- Explicit quality survey shown to users
- Implicit data from usage
- Semi-supervised approaches for label propagation
- Surrogate learning (e.g predicting if “topic experts” will upvote the answer)
- ...

Did you find this answer helpful? Yes · No



We tested 100+ features, the final model uses ~50 features after feature pruning

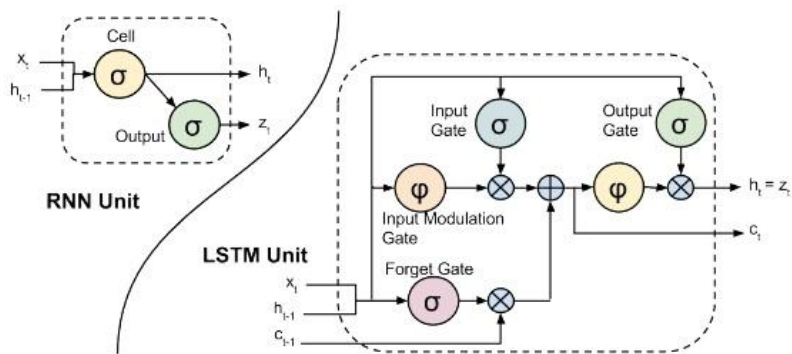
- **User features** -- e.g “Is the author an expert in the topic?”
- **Answer text features** -- e.g “What is the syntactic complexity of the text?”
- **Question/Answer features** -- e.g “Is the answer answering the question?”
- **Voter features** -- e.g “Is voter an expert in the topic?”
- **Metadata features** -- e.g “How many answers did the question have when the answer was written?”

## Models

- Logistic Regression
- Random Forests
- Gradient Boosted Decision Trees
- Recurrent Neural Networks
- Ensembles
- ...

GBDTs provide a good balance between accuracy, complexity, training time, prediction time and ease of deploying in production.

## Sequence Learning



Long Short-Term Memory (LSTM)  
Proposed by Hochreiter and Schmidhuber, 1997

# Answer Ranking: Interpretability

Author	Clint Lewis	Stephen Black	David Mays	Robert Mullen	Wynne Sharp
Answer Score	0.8953	0.8708	0.8552	0.8473	0.7991
Answer Length	0.8137	0.9073	0.9803	0.9475	0.8024
Answer Age	0.0099	0.0067	0.0049	0.0144	0.0017
Answer Upvotes	1.4908	2.2760	2.2760	2.2760	1.2828
Answer Downvotes	1286328285975790.0000	1322366874628287.0000	1374711950197888.0000	1319868647696886.0000	1372429721373320.0000
Downvote Count	0.0000	1.0000	1.0000	1.0000	0.0000

Training

Sample  
Answers

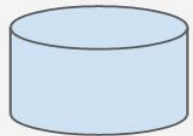
User Surveys  
Usage Metrics  
...

Ground Truth Data

Scores

Features

Feature Extraction



User Actions  
Upvotes  
Downvotes  
...

Predictive Model

Prediction

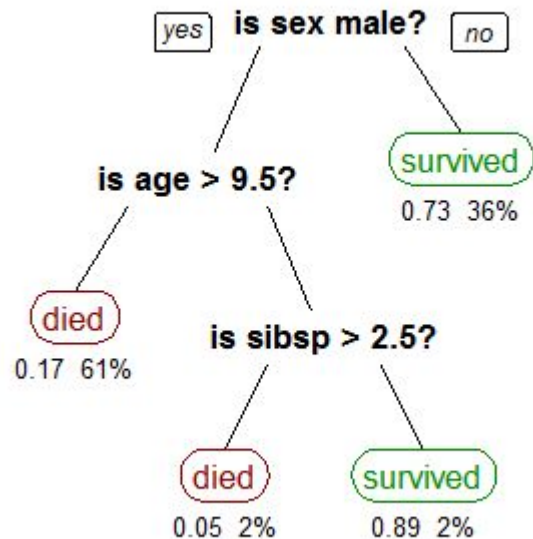
Answer

Features

Scores

- Latency: tens of milliseconds
- Computing 100 features each for 100 answers, even at 10us per feature computation, can take 100ms.
- Need to parallelize computation, and also cache feature values/scores.
- Caching → need to support real-time cache dirties/updates.

- Trick -- don't recompute scores if the feature doesn't flip any 'decision branch'.



- Need to start with defining what we want the model to learn.
- Feature engineering and interpretability are important.
- Class imbalance for classification problems.
- Training data is scarce -- need to fuse multiple data sources together.



1. Duplicate Question Detection
2. Answer Ranking
- 3. Topic Expertise Detection**
4. Moderation

## How long does it take to get an H-1B receipt from USCIS if the petition has been filed under standard processing?



Priya Alagiri, I am an immigration lawyer who specializes in business immigration law matters.

Written Jan 24, 2013 · Upvoted by Perna Lal, [Immigration Attorney](#)

## How is Rheumatoid Arthritis caused?



Peri Dwyer, Chiropractic Physician, BS in Biology, Anatomy and Physiology instructor

Written Oct 23, 2015 · Upvoted by Kate Simmons, [myofascial pain specialist](#)

- Important signal to all other quality systems.
- Can make content more trustworthy.
- Helps retaining and engaging experts

**How long does it take to get an H-1B receipt from USCIS if the petition has been filed under standard processing?**



Priya Alagiri, I am an immigration lawyer who specializes in business immigration law matters.

Written Jan 24, 2013 · Upvoted by Perna Lal, [Immigration Attorney](#)

**How is Rheumatoid Arthritis caused?**



Peri Dwyer, Chiropractic Physician, BS in Biology, Anatomy and Physiology instructor

Written Oct 23, 2015 · Upvoted by Kate Simmons, [myofascial pain specialist](#)

# Given the rise of Neural Networks, what is the future of probabilistic graphical models?

 Answer

Request ▾

Follow 98

Comment 1

Share Downvote

⋮

## 4 Answers

Topic Experts



**Ian Goodfellow**

Research scientist at OpenAI

Written Aug 11 · Upvoted by Waleed Kadous, PhD in ML and AI, Shehroz Khan, ML Researcher, Postdoc @U of Toronto, Yoshua Bengio and Yair Livne

Relevant Credentials

The two aren't mutually exclusive. Most applications of neural nets can be considered graphical models that use neural nets to provide some of the conditional probability distributions. You could argue that the graphical model perspective is growing less useful because so many recent neural models have such simple graph structure (GANs, VAEs and NICE are bipartite graphs with every latent variable connected to every observed variable, PixelRNNs/MADE/NADE are the complete graph with no latent variables). These graphs are not very structured compared to neural models that were popular a few years ago like DBMs. But there are some recent models that make a little bit of use of graph structure, like VAEs with auxiliary variables.

3.1k Views · View Upvotes · Answer requested by Eric Dortmans, Piyush Gupta, and 8 more

Upvote | 36

Downvote Comment

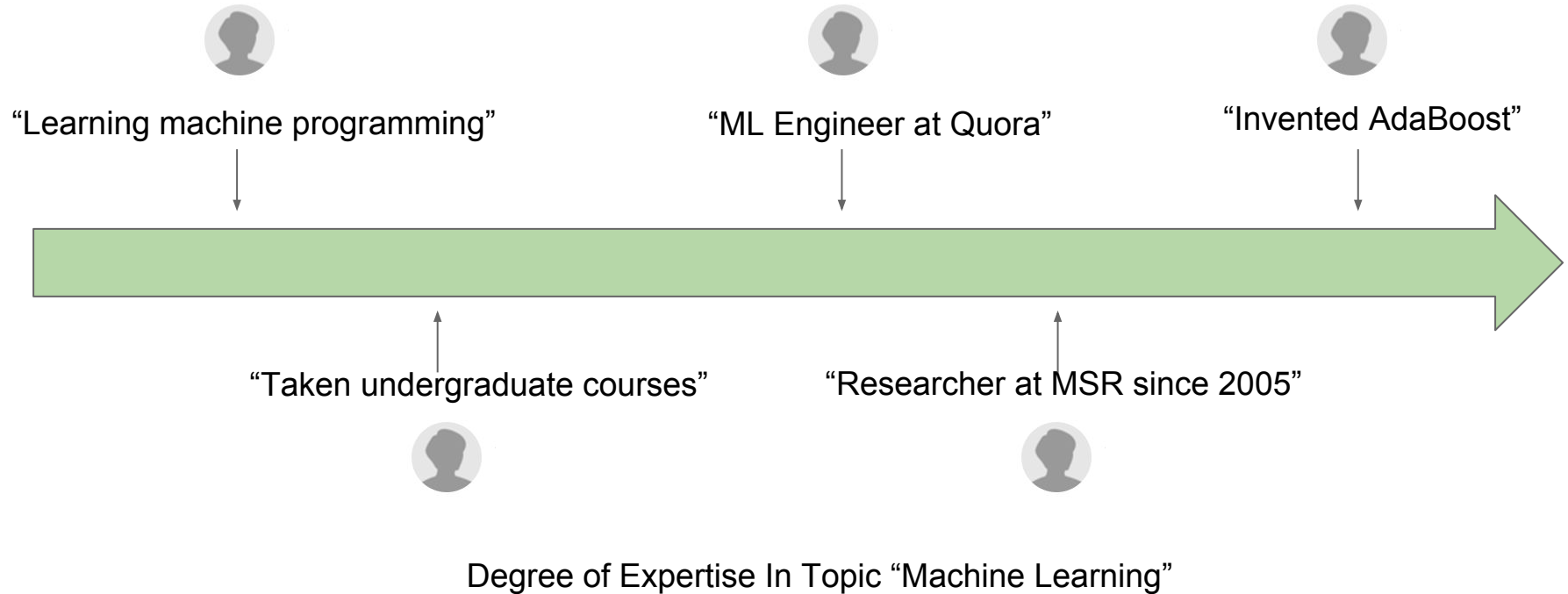


## Problem Statement

**Predict topic expertise level of users.**

# Deducing Expertise From Topic Biography

Quora



## **Problem Formulation**

Multi-class classification on text of biography, classes being discrete buckets on the expertise spectrum.

Experts are sparse → class imbalance.

## **Training Data Sources**

Hand labeled data, Data from other quality measures, Label propagation, Users can “report” bios...

## Models

Logistic Regression, Random Forests, Gradient Boosted Decision Trees, ...

Regularization is very important

## Features

Ngrams, Named Entities, Cosine similarity between topic name and biography text, ...



- Andrew Ng is an ML expert → his ML answers must be good.
- What about answers upvoted by him?
- What about answers written by someone whose own answers were upvoted by him?
- ...

## How should you start a career in Machine Learning?



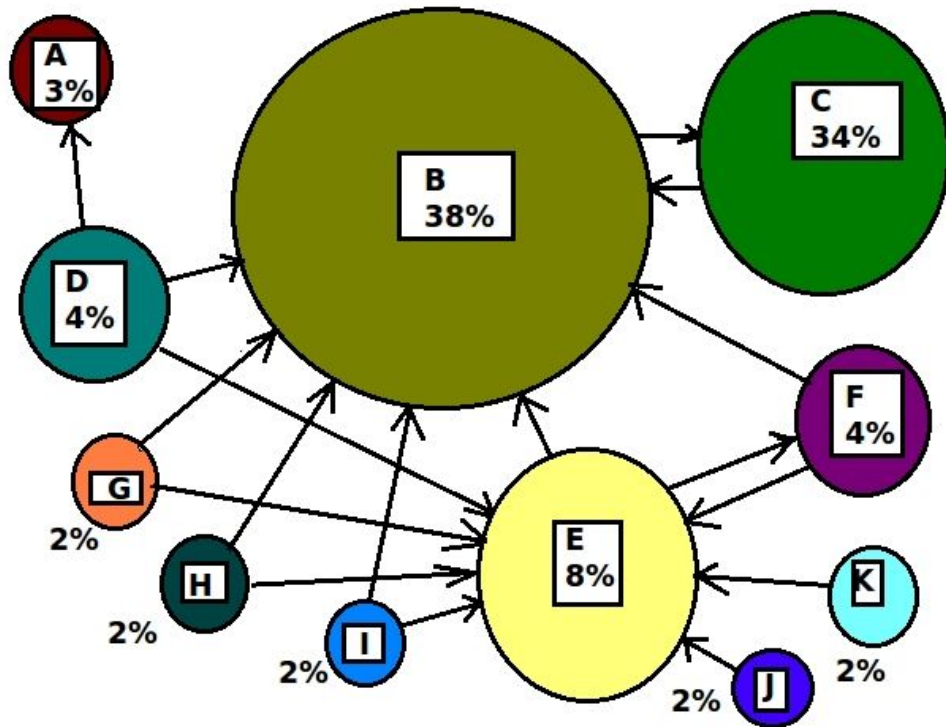
Andrew Ng, Chief Scientist at Baidu; Chairman/Co-Founder of Coursera; Stanford faculty

The Stanford machine learning on Coursera was designed to help a broad audience get started in machine learning. If you're familiar with basic programming (in any language), I'd recommend starting there.

Many people have gotten jobs in machine learning just by completing that MOOC. There're other similar online courses that help: for example the John Hopkins Data

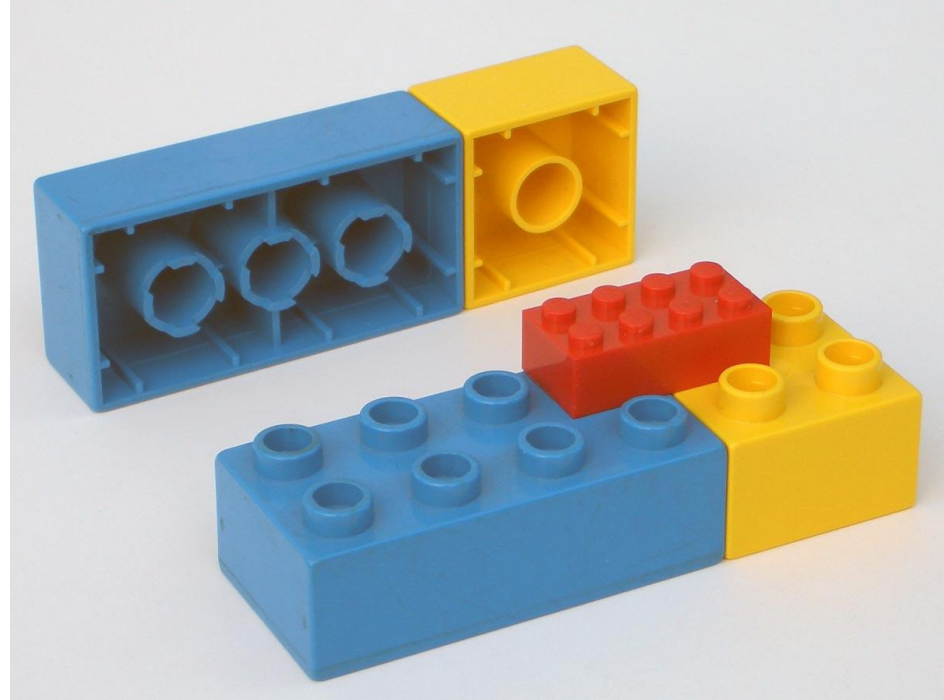
# Topic Expertise Propagation Graph

- Trust in expertise is transitive:  
 $A \rightarrow B$ , and  $B \rightarrow C \Rightarrow A \rightarrow C$
- So trust in expertise propagates through the network
- We can mine the graph to discover topic expertise using graph algorithms like PageRank
- Unsupervised learning!



# Ensembles To Combine Submodels

- Models trained on heterogeneous data
- Models trained using supervised, semi-supervised and unsupervised approaches
- Low correlation between different models
- Can combine them using ensembles



- Class imbalance for classification problems.
- Unsupervised learning is powerful.
- Ensembles can help combine learners trained on data from different sources.

1. Duplicate Question Detection
2. Answer Ranking
3. Topic Expertise Detection
4. **Moderation**

- Any user-generated-content product has lots of moderation challenges.
- Content -- spam, hate-speech, porn, plagiarism etc.
- Account -- fake names, impersonation, sockpuppets
- Quora specific policies -- e.g. answers making fun of questions, insincere questions.



# Moderation: ML Challenges



- Super nuanced judgements, too hard for even trained humans.
- Noisy labeled data
- Too hard a problem for a computer.

Shut up yourself.

Reply...

[Upvote](#) · [Downvote](#) · [Report](#) · Aug 26

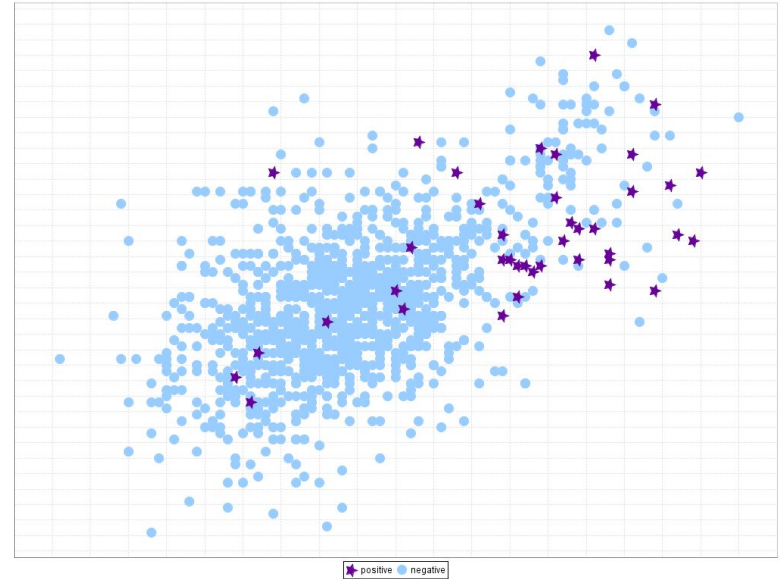
More like bollywood is made for idiots..

Reply...

[Upvote](#) · [Downvote](#) · [Report](#) · Aug 26

# Moderation: ML Challenges

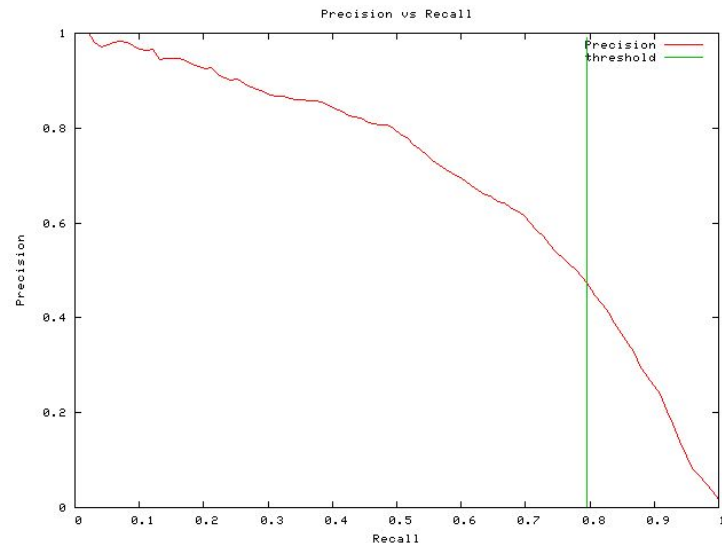
- Difficulty in learning due to severe class imbalance.
- Metrics can be deceiving -- useless models at 99 % accuracy
- Scarce labeled data, class imbalance makes it worse





# Moderation: ML Challenges

- High precision needed.
- Usually get extremely low recall at desired precision levels.
- Want to detect problems users see bad content, so can't rely on any user interactions.



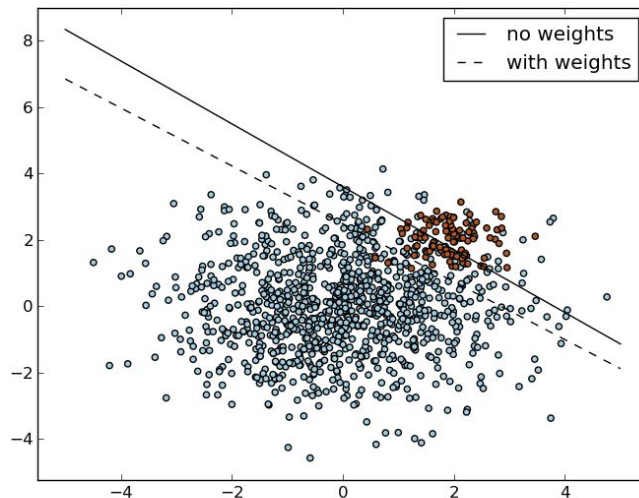
# Moderation: Tricks and Approaches

- Start with looking at the right metrics.
- Can use standard metrics like F1, AUC.
- Or fine tune metrics based on your application needs.

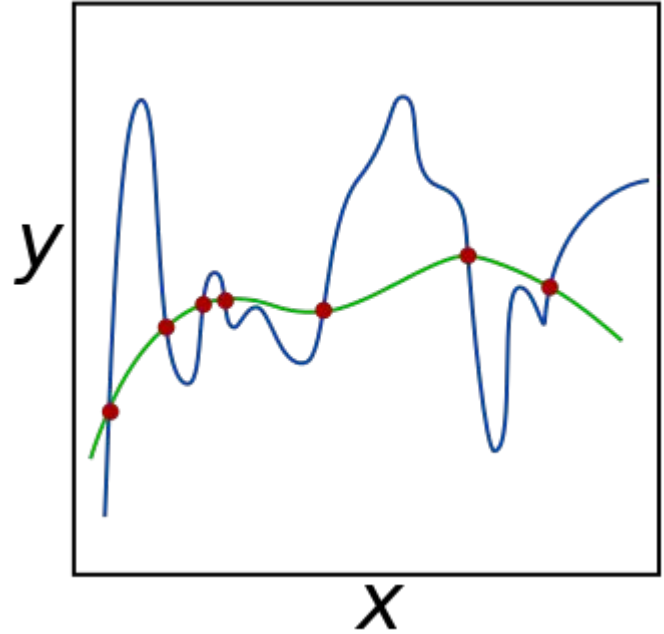
	Predicted: NO	Predicted: YES	
n=165 Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

# Moderation: Tricks and Approaches

- Oversampling minority class, undersampling majority class
- Random sampling to generate negative examples
- Higher cost of mistakes on the minority class.



- Successive iterations of: hand labeling → model to reduce the space → hand labeling ...
- Scarce data, high feature dimensionality → simple models with regularization often work very well
- Using low precision classifiers for making human review efficient



# Moderation: Problem Properties

Quora

- Severe class imbalance.
- Need to look at the “right” evaluation metrics.
- Sampling techniques can be very effective.

# Summary

What do all quality problems have in common?

# Machine Learning + Product Understanding

- Start with defining what we want the model to do.
- Product intuition is important for feature engineering.
- Interpretability of models is important for iteration.

# Training Data Scarcity

- Good and cheap training data is unavailable/costly.
- Often need to combine data from different sources.
- Unsupervised/semi-supervised learning are very useful.
- Ensembles are your friends.



# Dealing With Class Imbalance

- Look at the right metrics.
- Re-sampling techniques can be very effective.
- Can incorporate “data collection” cost into the algorithm.

# Topics For Some Other Day

- More specific semi-supervised learning approaches.
- Combining Quality, Relevance and Demand together.
- Avoiding (and sometimes creating) feedback loops.
- Engineering challenges behind these ML problems.

# Thank You! Questions?

Standard Disclaimer: Quora Is Hiring :)

Nikhil Garg

@nikhilgarg28