Machine Learning @Quora: Beyond Deep Learning

STANFORD UNIVERSITY

08/02/2016

Xavier Amatriain (@xamat)

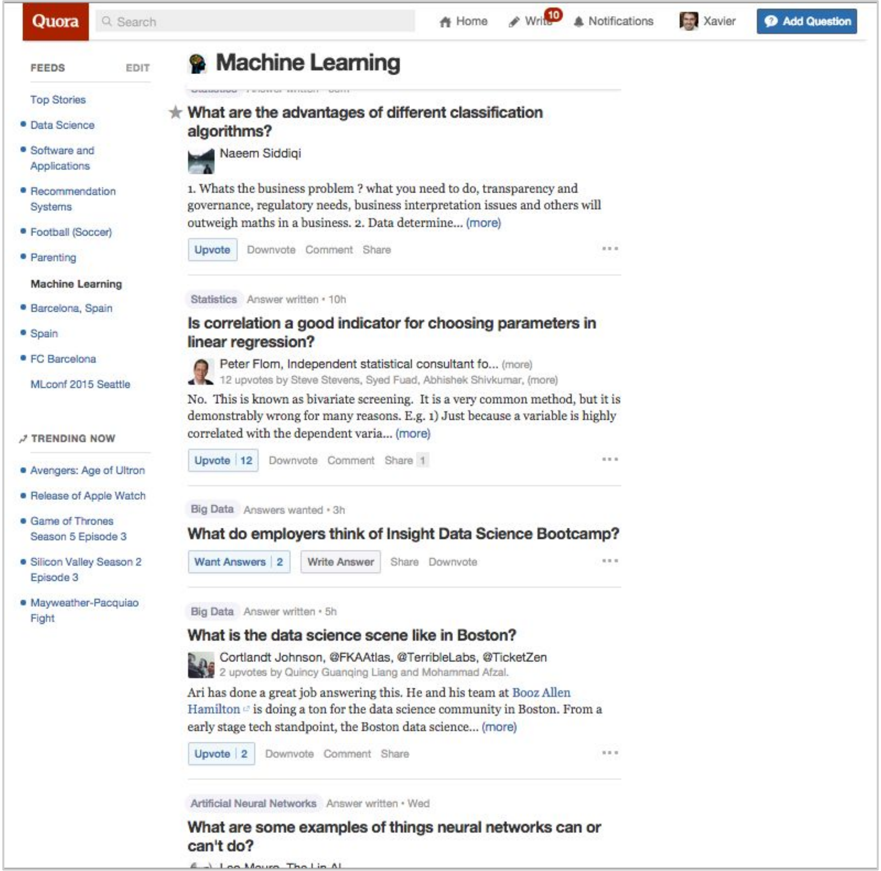
Quora

Our Mission

Quora

“To share and grow the world’s knowledge”

- Millions of questions
- Millions of answers
- Millions of users
- Thousands of topics
- …



Machine Learning

What are the advantages of different classification algorithms?

Naeem Siddiqi

1. Whats the business problem ? what you need to do, transparency and governance, regulatory needs, business interpretation issues and others will outweigh maths in a business. 2. Data determine... (more)

Statistics Answer written • 10h

Is correlation a good indicator for choosing parameters in linear regression?

Peter Flom, Independent statistical consultant fo... (more)
12 upvotes by Steve Stevens, Syed Fuad, Abhishek Shivkumar, (more)

No. This is known as bivariate screening. It is a very common method, but it is demonstrably wrong for many reasons. E.g. 1) Just because a variable is highly correlated with the dependent varia... (more)

Big Data Answers wanted • 3h

What do employers think of Insight Data Science Bootcamp?

Big Data Answer written • 5h

What is the data science scene like in Boston?

Cortlandt Johnson, @FKAAtlas, @TerribleLabs, @TicketZen
2 upvotes by Quincy Guanqing Liang and Mohammad Afzal.

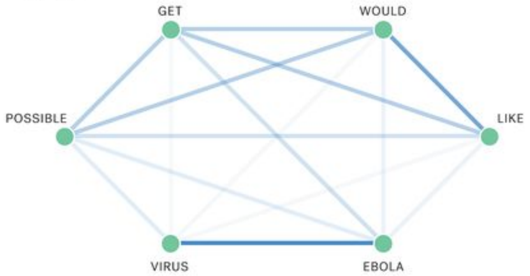
Ari has done a great job answering this. He and his team at Booz Allen Hamilton is doing a ton for the data science community in Boston. From a early stage tech standpoint, the Boston data science... (more)

Artificial Neural Networks Answer written • Wed

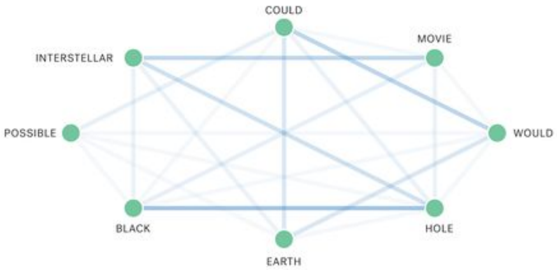
What are some examples of things neural networks can or can't do?

Lots of high-quality textual information

2014, Q4



2014, Q4



DATA @ QUORA

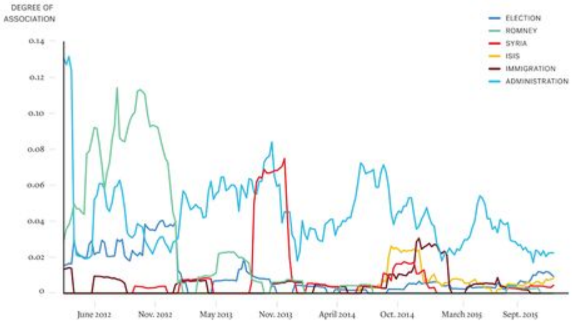
Mapping the Discussion on Quora Over Time through Question Text

WENWEN TAO (陶雯雯)

United States, 2014 Quarter 4

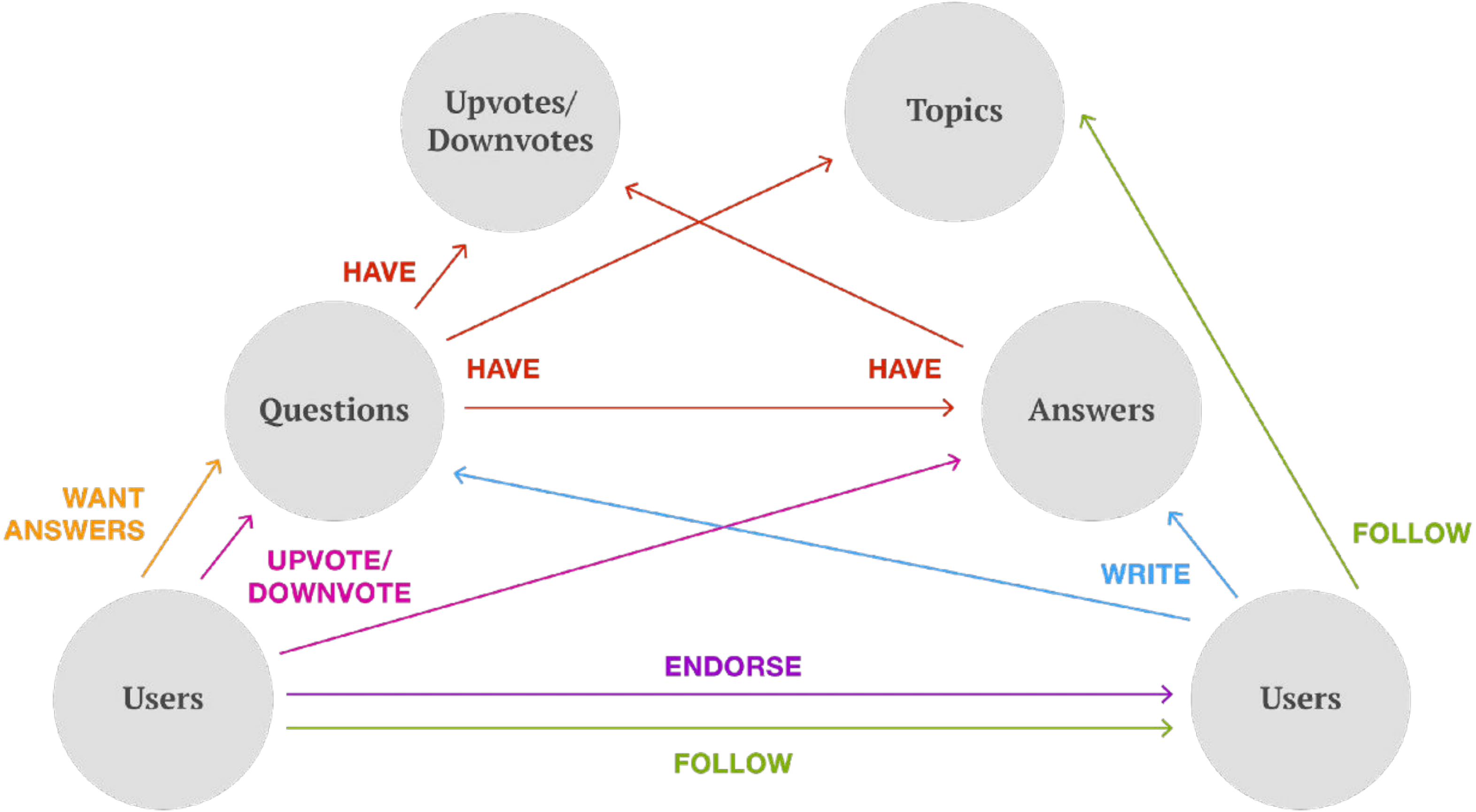
Uber
SRI
SKI
TCS
Reviews
Warming
Foreigner
Regression
Yishan
Lanka
Investor
Orbitals
JEE
Researchers
Ferguson
Ebola
Hindu
Elasticsearch
Torture
Indonesia
Spoiler
Einstein
Relativity
Alibaba
Nexus
Nasa
Thanksgiving
Shanghai
Mcdonald
Tinder
Interstellar
Sony
Wong
Reactor
Borderline
Tamil
Turkey
Rationale
Pakistan
Nobel

Focus Term: Obama

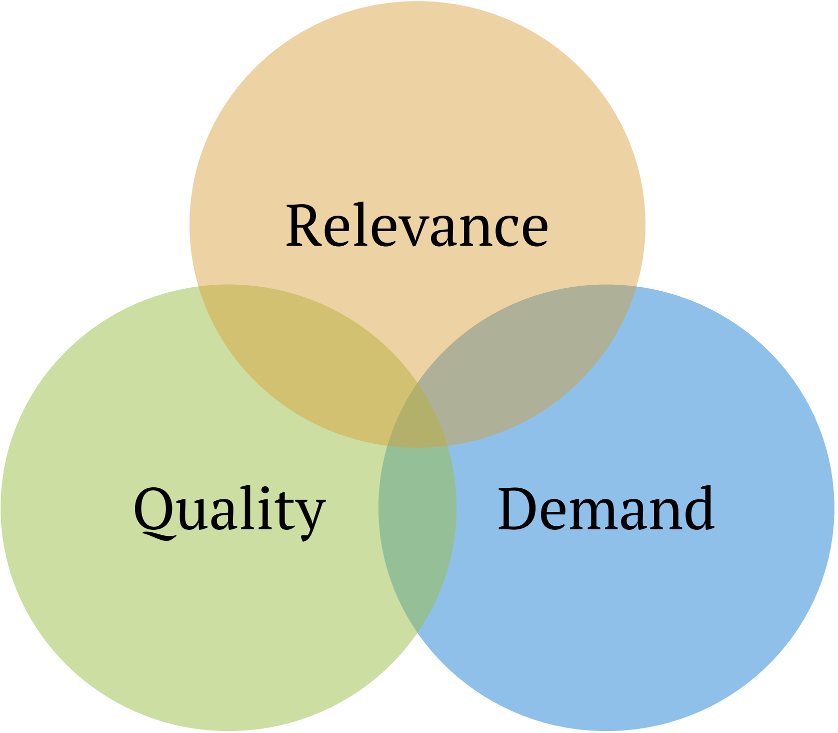


Text + all those other things

Quora

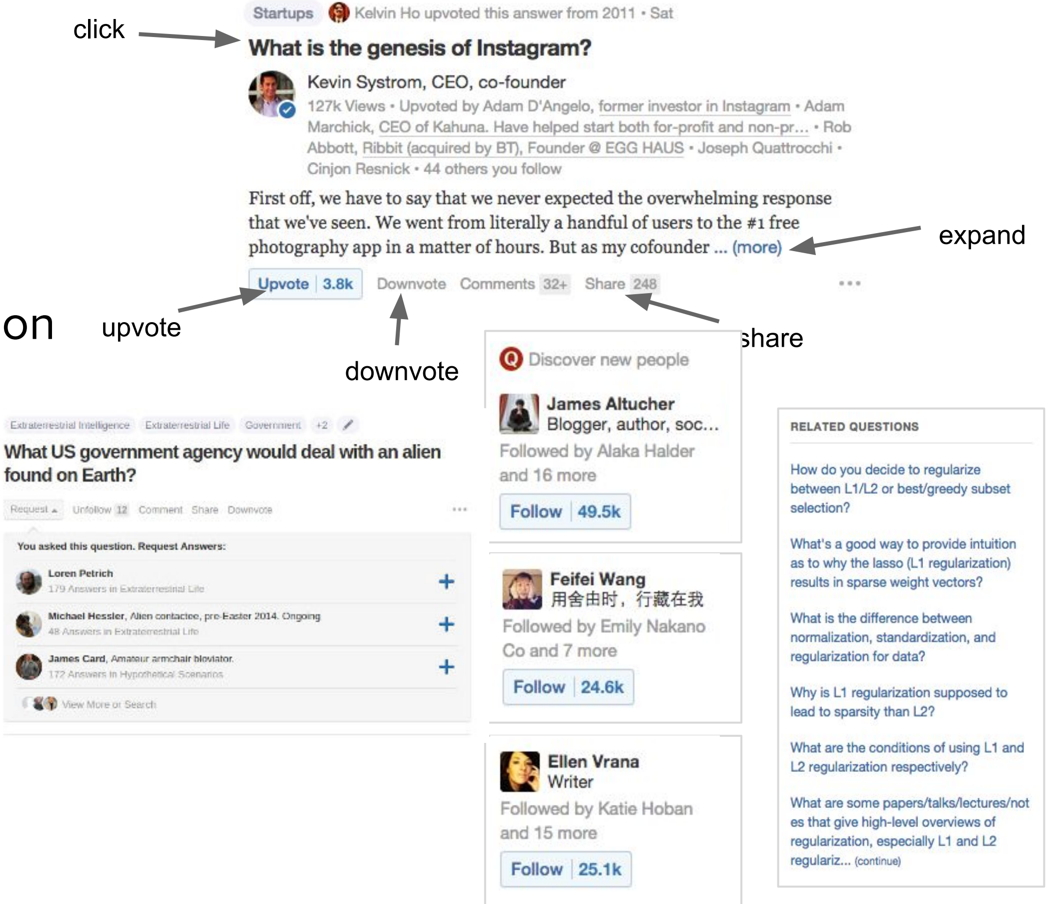


What we care about



ML Applications

- Homepage feed ranking
- Email digest
- Answer quality & ranking
- Spam & harassment classification
- Topic/User recommendation
- Trending Topics
- Automated Topic Labelling
- Related & Duplicate Question
- User trustworthiness
- ...



click

Startups · Kelvin Ho upvoted this answer from 2011 · Sat

What is the genesis of Instagram?

Kevin Systrom, CEO, co-founder

127k Views · Upvoted by Adam D'Angelo, former investor in Instagram · Adam Marchick, CEO of Kahuna. Have helped start both for-profit and non-pr... · Rob Abbott, Ribbit (acquired by BT), Founder @ EGG HAUS · Joseph Quattrocchi · Cinjon Resnick · 44 others you follow

First off, we have to say that we never expected the overwhelming response that we've seen. We went from literally a handful of users to the #1 free photography app in a matter of hours. But as my cofounder ... (more)

expand

Upvote 3.8k · Downvote · Comments 32+ · Share 248 · ...

upvote

downvote

share

Extraterrestrial Intelligence · Extraterrestrial Life · Government · +2

What US government agency would deal with an alien found on Earth?

Request · Unfollow 12 · Comment · Share · Downvote · ...

You asked this question. Request Answers:

- Loren Petrich — 179 Answers in Extraterrestrial Life +
- Michael Hessler, Alien contactee, pre-Easter 2014. Ongoing — 48 Answers in Extraterrestrial Life +
- James Card, Amateur armchair bloviator. — 172 Answers in Hypothetical Scenarios +

View More or Search

Discover new people

James Altucher
Blogger, author, soc...
Followed by Alaka Halder and 16 more
Follow 49.5k

Feifei Wang
用舍由时，行藏在我
Followed by Emily Nakano Co and 7 more
Follow 24.6k

Ellen Vrana
Writer
Followed by Katie Hoban and 15 more
Follow 25.1k

RELATED QUESTIONS

How do you decide to regularize between L1/L2 or best/greedy subset selection?

What's a good way to provide intuition as to why the lasso (L1 regularization) results in sparse weight vectors?

What is the difference between normalization, standardization, and regularization for data?

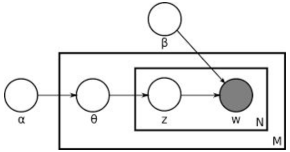
Why is L1 regularization supposed to lead to sparsity than L2?

What are the conditions of using L1 and L2 regularization respectively?

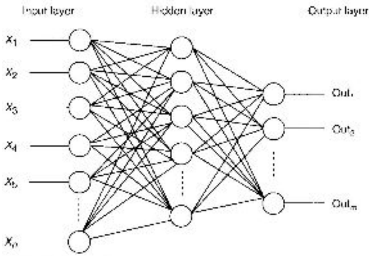
What are some papers/talks/lectures/notes that give high-level overviews of regularization, especially L1 and L2 regulariz... (continue)

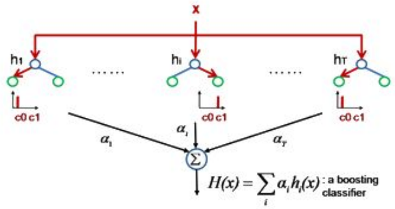
Models

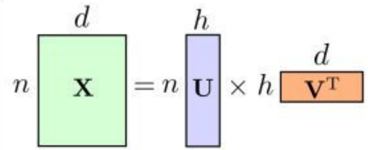
- Deep Neural Networks
- Logistic Regression
- Elastic Nets
- Gradient Boosted Decision Trees
- Random Forests
- LambdaMART
- Matrix Factorization
- LDA
- ...
-

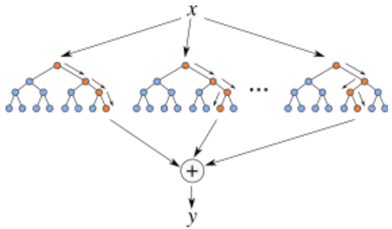


$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$





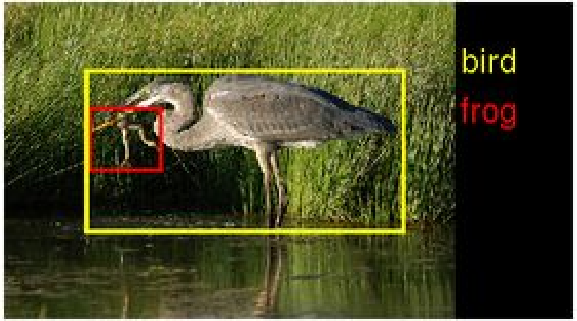


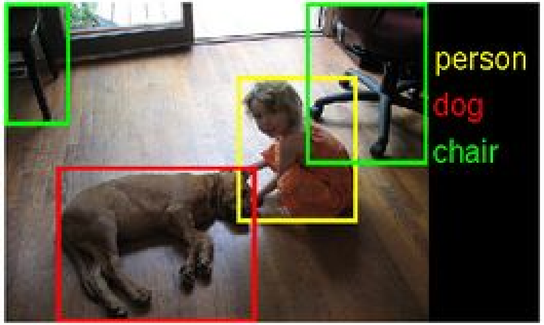


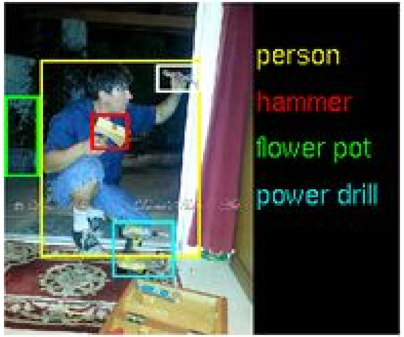
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

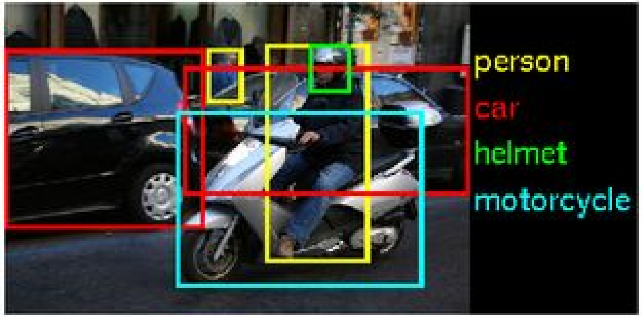
Deep Learning Works

Image Recognition

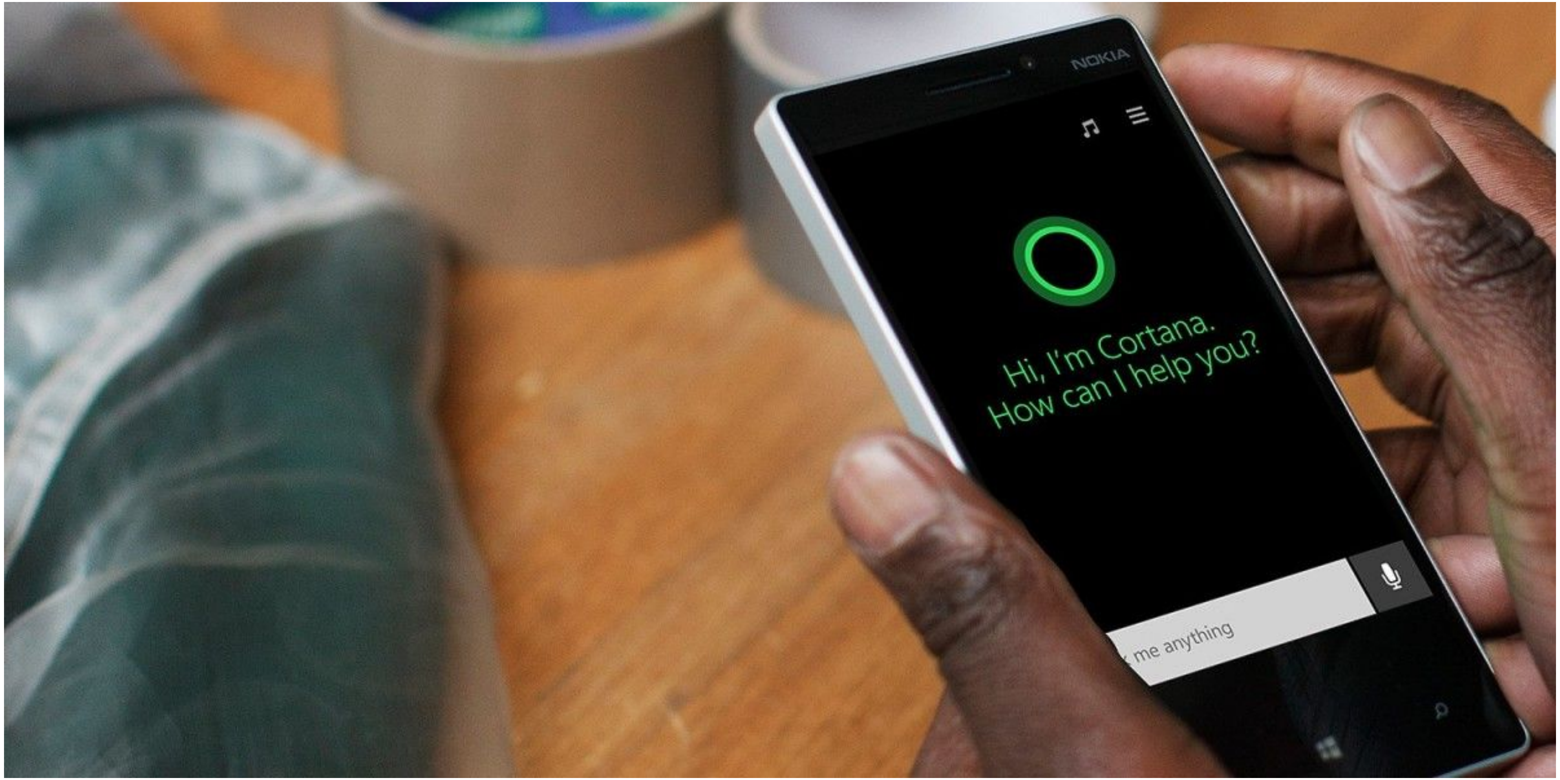




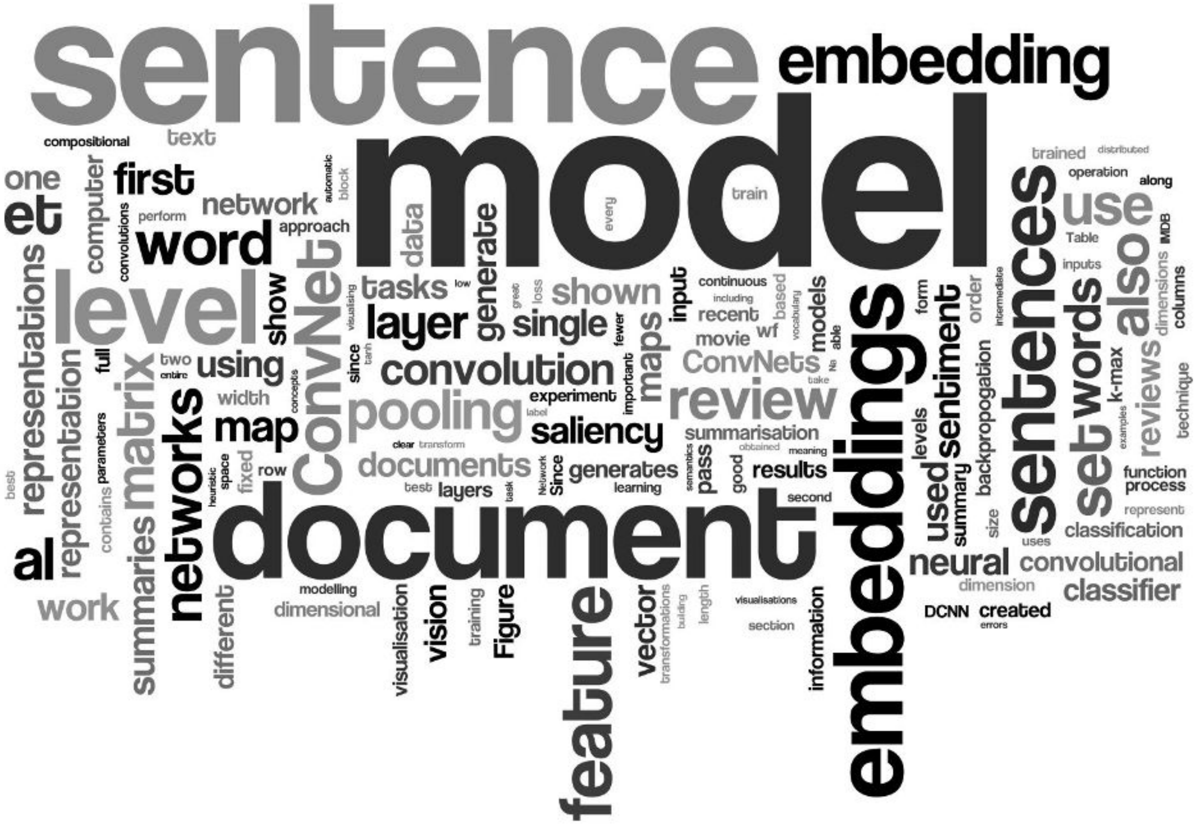




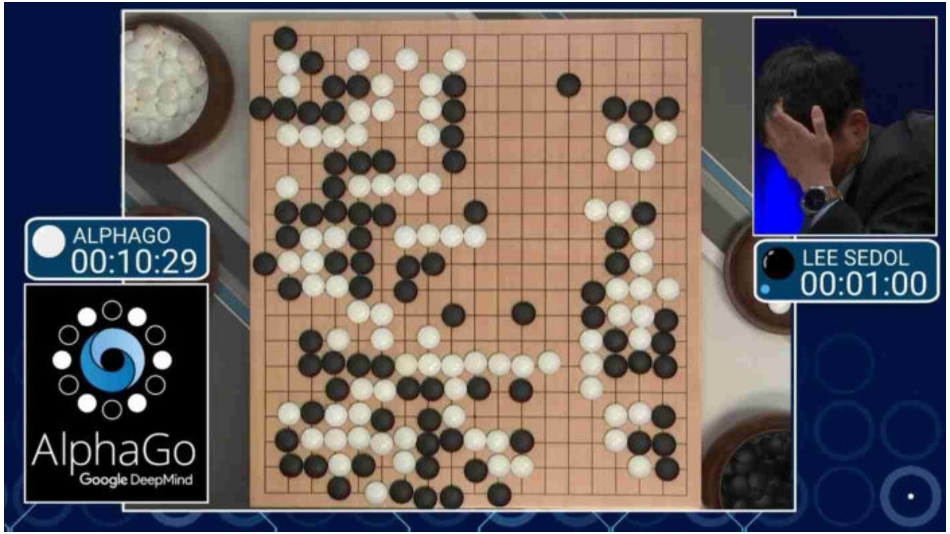
Speech Recognition

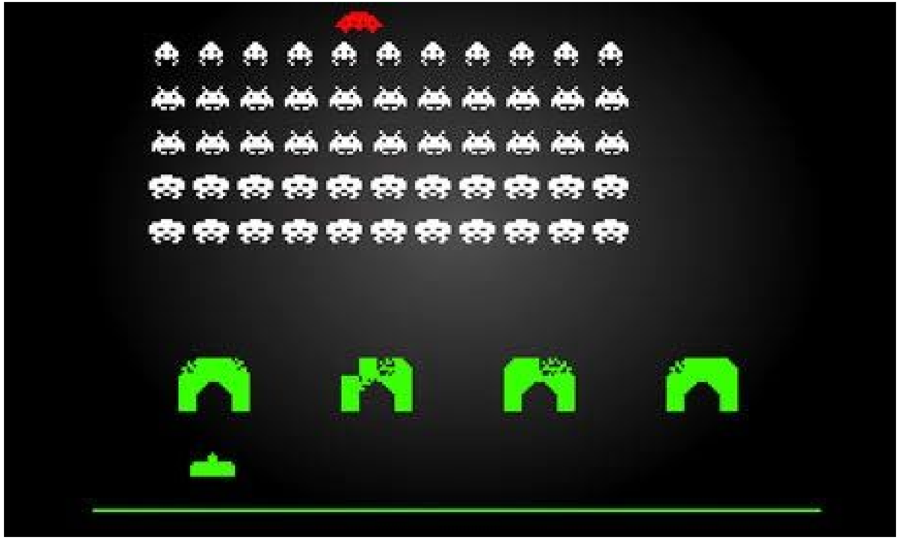


Natural Language Processing

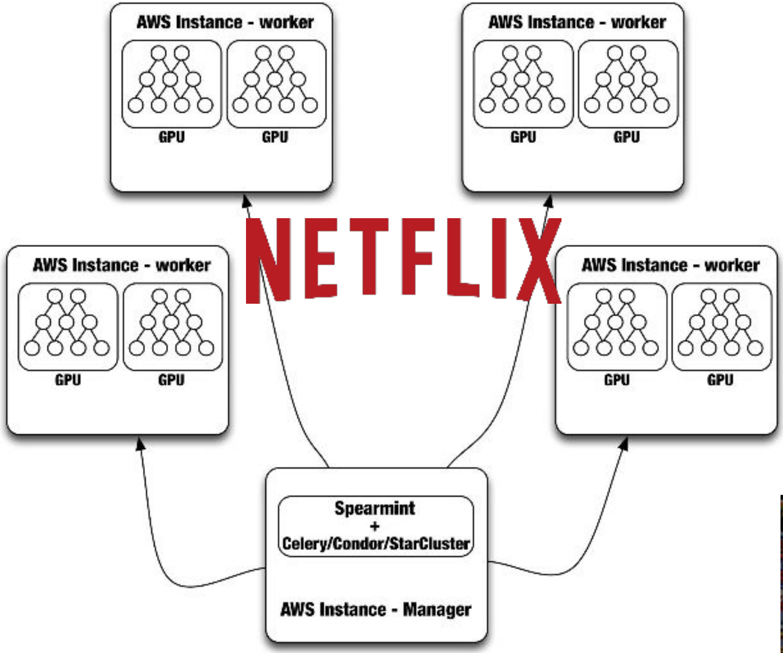


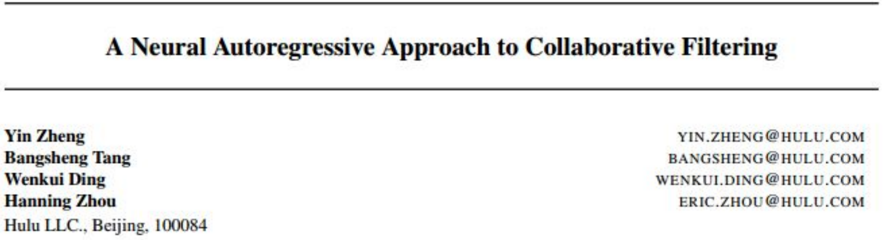
Game Playing





Recommender Systems





A Neural Autoregressive Approach to Collaborative Filtering

Yin Zheng YIN.ZHENG@HULU.COM
Bangsheng Tang BANGSHENG@HULU.COM
Wenkui Ding WENKUI.DING@HULU.COM
Hanning Zhou ERIC.ZHOU@HULU.COM
Hulu LLC., Beijing, 100084

Recommending music on Spotify with deep learning

AUGUST 05, 2014

But…

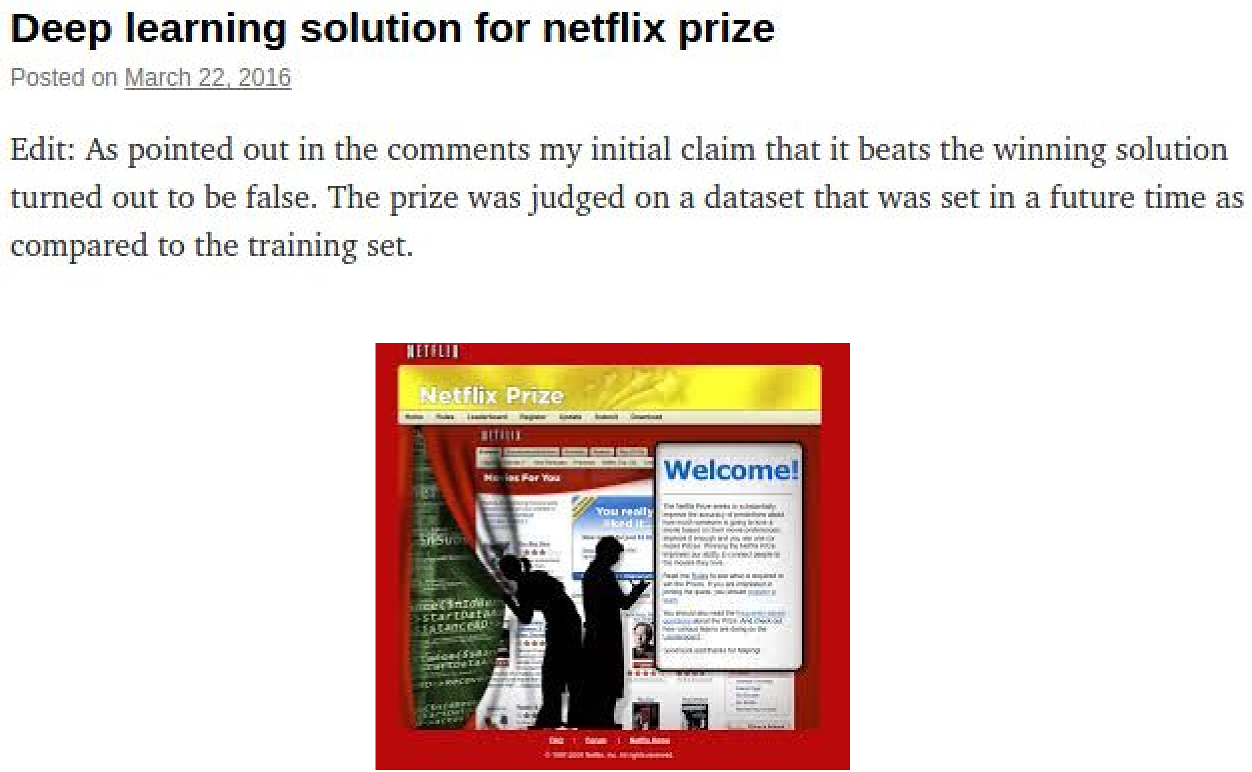
Deep Learning is not Magic

OREN ETZIONI BUSINESS 06.15.16 7:00 AM

DEEP LEARNING ISN'T A DANGEROUS MAGIC GENIE. IT'S JUST MATH

WIRED

Deep Learning is not always that “accurate”

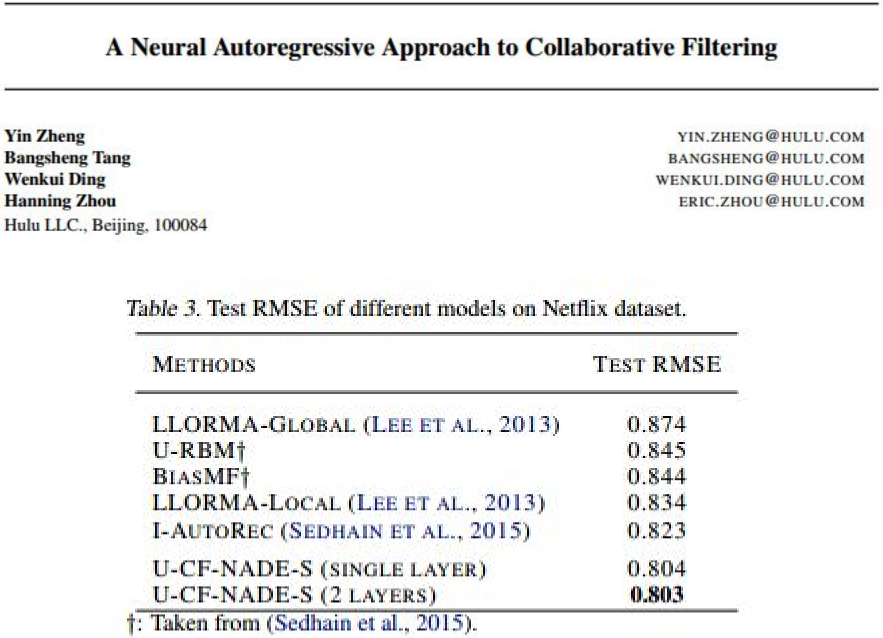


Deep learning solution for netflix prize

Posted on March 22, 2016

Edit: As pointed out in the comments my initial claim that it beats the winning solution turned out to be false. The prize was judged on a dataset that was set in a future time as compared to the training set.

… or that “deep”



A Neural Autoregressive Approach to Collaborative Filtering

Yin Zheng YIN.ZHENG@HULU.COM
Bangsheng Tang BANGSHENG@HULU.COM
Wenkui Ding WENKUI.DING@HULU.COM
Hanning Zhou ERIC.ZHOU@HULU.COM
Hulu LLC., Beijing, 100084

Table 3. Test RMSE of different models on Netflix dataset.

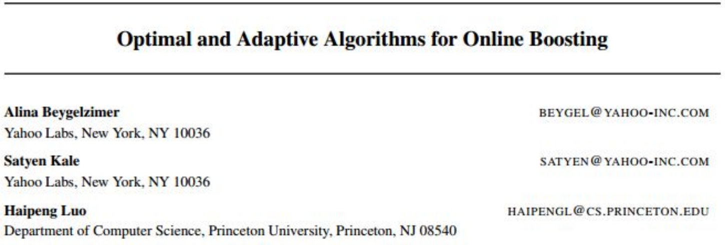
Methods	Test RMSE
LLORMA-Global (Lee et al., 2013)	0.874
U-RBM†	0.845
BiasMF†	0.844
LLORMA-Local (Lee et al., 2013)	0.834
I-AutoRec (Sedhain et al., 2015)	0.823
U-CF-NADE-S (single layer)	0.804
U-CF-NADE-S (2 layers)	**0.803**

†: Taken from (Sedhain et al., 2015).

Other ML Advances

Quora

- Factorization Machines
- Tensor Methods
- Non-parametric Bayesian models
- XGBoost
- Online Learning
- Reinforcement Learning
- Learning to rank
- ...

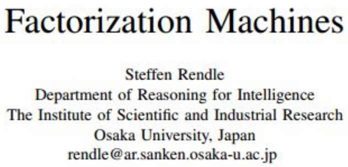


Optimal and Adaptive Algorithms for Online Boosting

Alina Beygelzimer BEYGEL@YAHOO-INC.COM
Yahoo Labs, New York, NY 10036

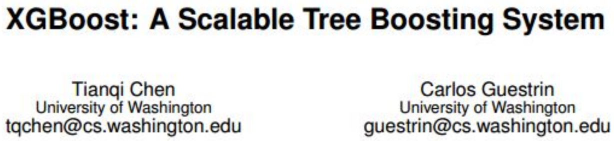
Satyen Kale SATYEN@YAHOO-INC.COM
Yahoo Labs, New York, NY 10036

Haipeng Luo HAIPENGL@CS.PRINCETON.EDU
Department of Computer Science, Princeton University, Princeton, NJ 08540



Factorization Machines

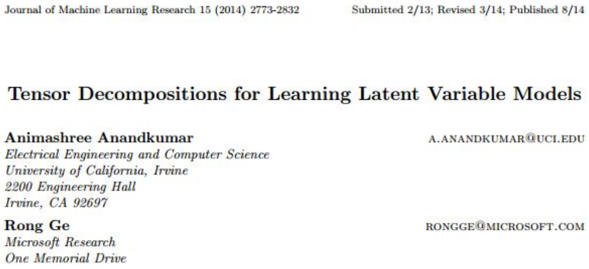
Steffen Rendle
Department of Reasoning for Intelligence
The Institute of Scientific and Industrial Research
Osaka University, Japan
rendle@ar.sanken.osaka-u.ac.jp



XGBoost: A Scalable Tree Boosting System

Tianqi Chen
University of Washington
tqchen@cs.washington.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

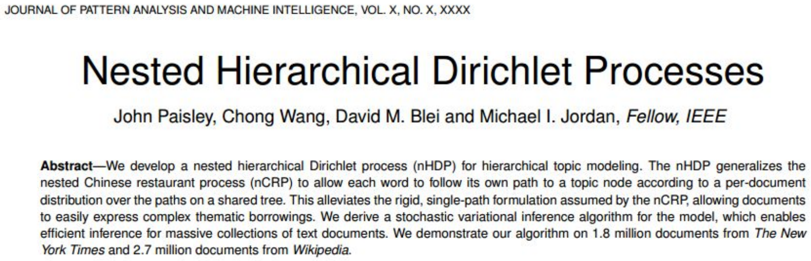


Journal of Machine Learning Research 15 (2014) 2773-2832 Submitted 2/13; Revised 3/14; Published 8/14

Tensor Decompositions for Learning Latent Variable Models

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Electrical Engineering and Computer Science
University of California, Irvine
2200 Engineering Hall
Irvine, CA 92697

Rong Ge RONGGE@MICROSOFT.COM
Microsoft Research
One Memorial Drive



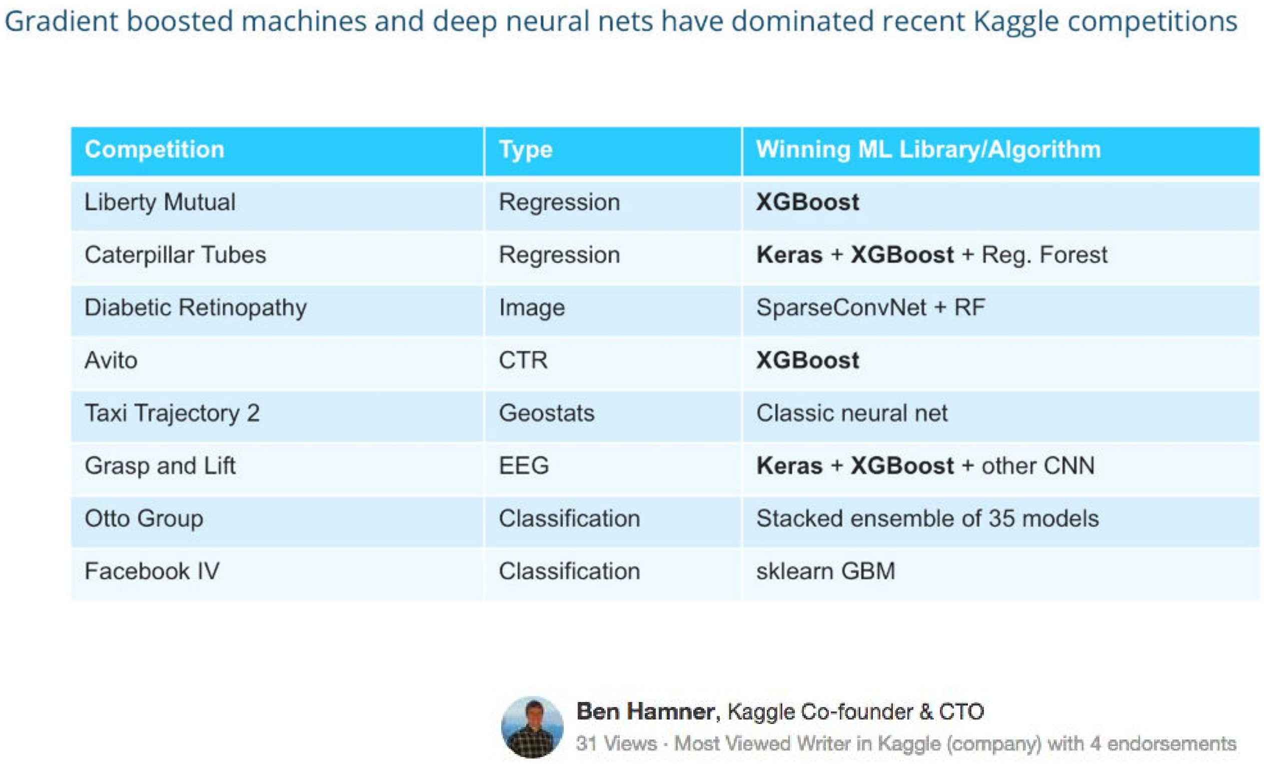
JOURNAL OF PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. X, NO. X, XXXX

Nested Hierarchical Dirichlet Processes

John Paisley, Chong Wang, David M. Blei and Michael I. Jordan, *Fellow, IEEE*

Abstract—We develop a nested hierarchical Dirichlet process (nHDP) for hierarchical topic modeling. The nHDP generalizes the nested Chinese restaurant process (nCRP) to allow each word to follow its own path to a topic node according to a per-document distribution over the paths on a shared tree. This alleviates the rigid, single-path formulation assumed by the nCRP, allowing documents to easily express complex thematic borrowings. We derive a stochastic variational inference algorithm for the model, which enables efficient inference for massive collections of text documents. We demonstrate our algorithm on 1.8 million documents from *The New York Times* and 2.7 million documents from *Wikipedia*.

Other very successful approaches



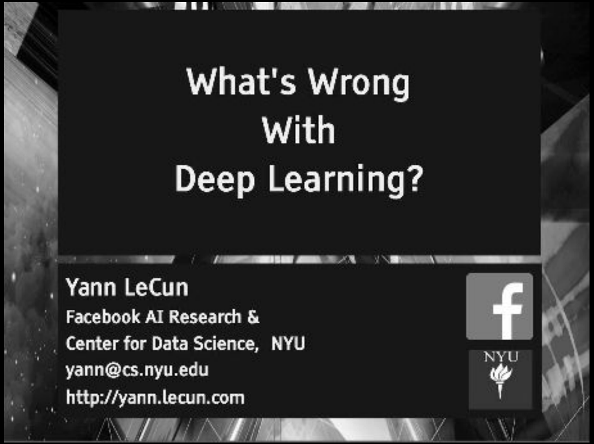
Gradient boosted machines and deep neural nets have dominated recent Kaggle competitions

Competition	Type	Winning ML Library/Algorithm
Liberty Mutual	Regression	**XGBoost**
Caterpillar Tubes	Regression	**Keras** + **XGBoost** + Reg. Forest
Diabetic Retinopathy	Image	SparseConvNet + RF
Avito	CTR	**XGBoost**
Taxi Trajectory 2	Geostats	Classic neural net
Grasp and Lift	EEG	**Keras** + **XGBoost** + other CNN
Otto Group	Classification	Stacked ensemble of 35 models
Facebook IV	Classification	sklearn GBM

Ben Hamner, Kaggle Co-founder & CTO

31 Views · Most Viewed Writer in Kaggle (company) with 4 endorsements

Is it bad to obsess over Deep Learning?

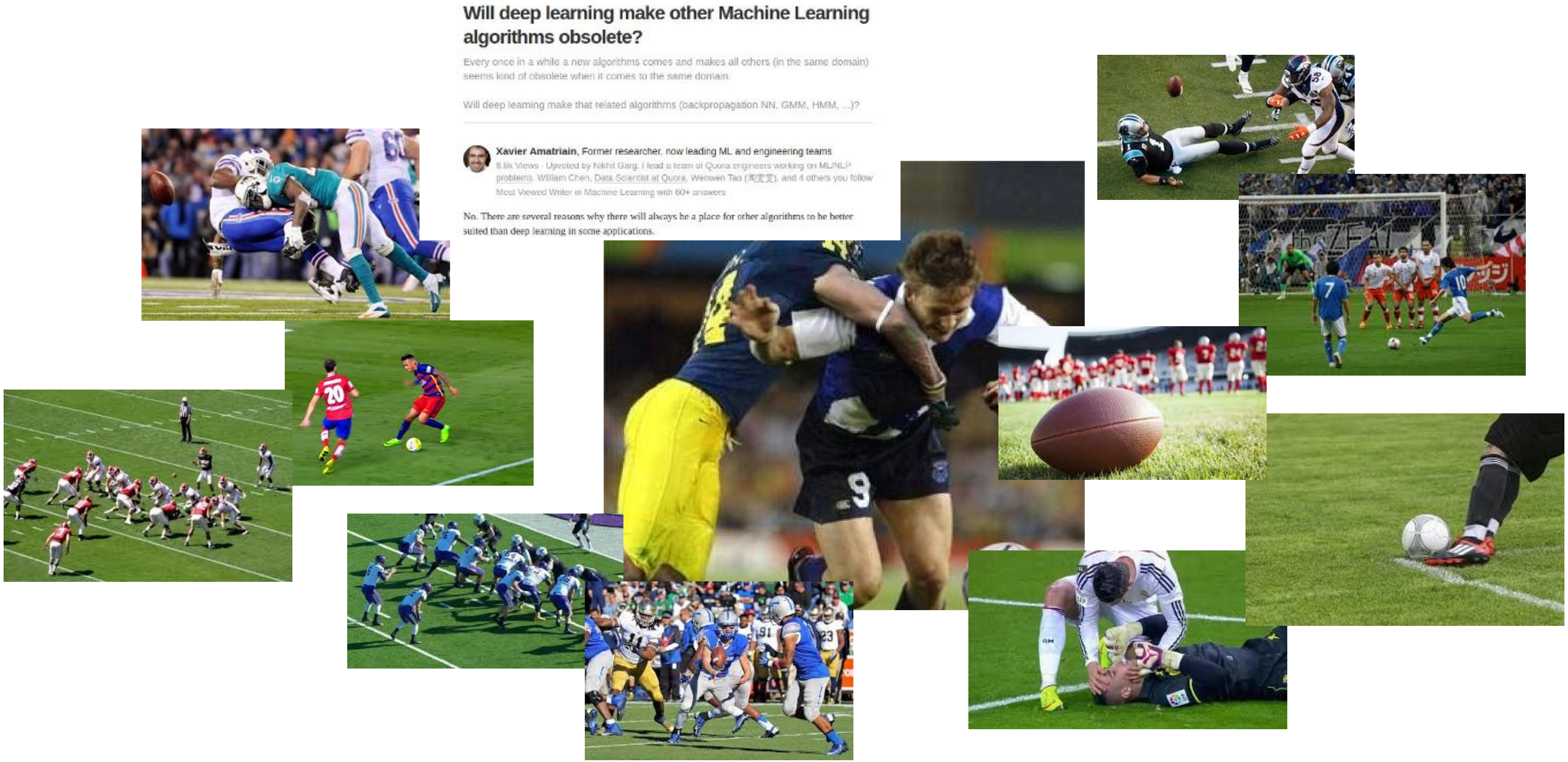


What's Wrong
With
Deep Learning?

Yann LeCun
Facebook AI Research &
Center for Data Science, NYU
yann@cs.nyu.edu
http://yann.lecun.com

Some examples

Football or Futbol?



Will deep learning make other Machine Learning algorithms obsolete?

Every once in a while a new algorithms comes and makes all others (in the same domain) seems kind of obsolete when it comes to the same domain.

Will deep learning make that related algorithms (backpropagation NN, GMM, HMM, ...)?

Xavier Amatriain, Former researcher, now leading ML and engineering teams

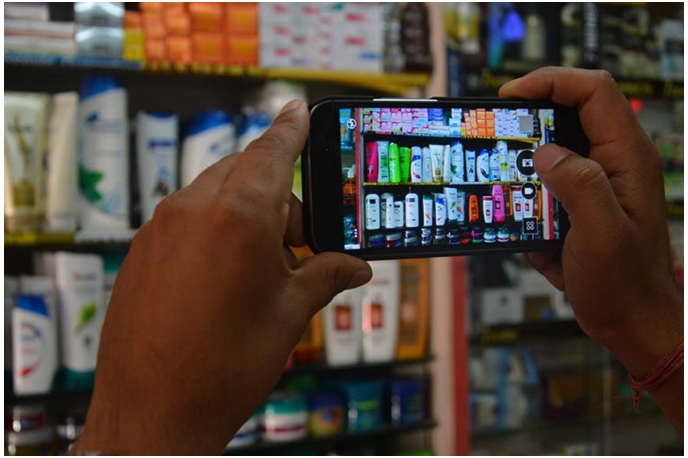
8.8k Views · Upvoted by Nikhil Garg, I lead a team of Quora engineers working on ML/NLP problems, William Chen, Data Scientist at Quora, Wenwen Tao (陶雯雯), and 4 others you follow

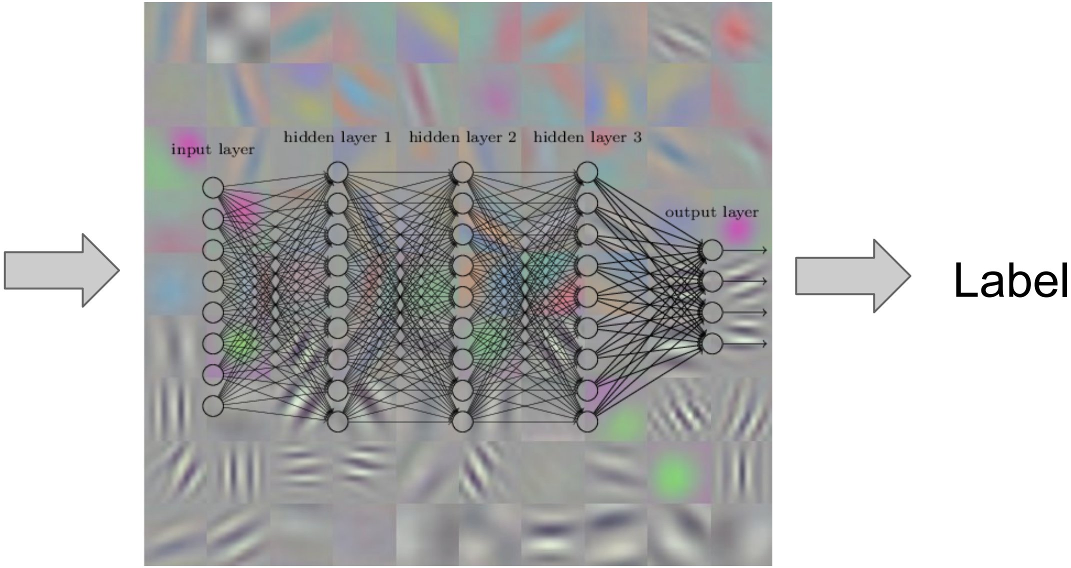
Most Viewed Writer in Machine Learning with 60+ answers

No. There are several reasons why there will always be a place for other algorithms to be better suited than deep learning in some applications.

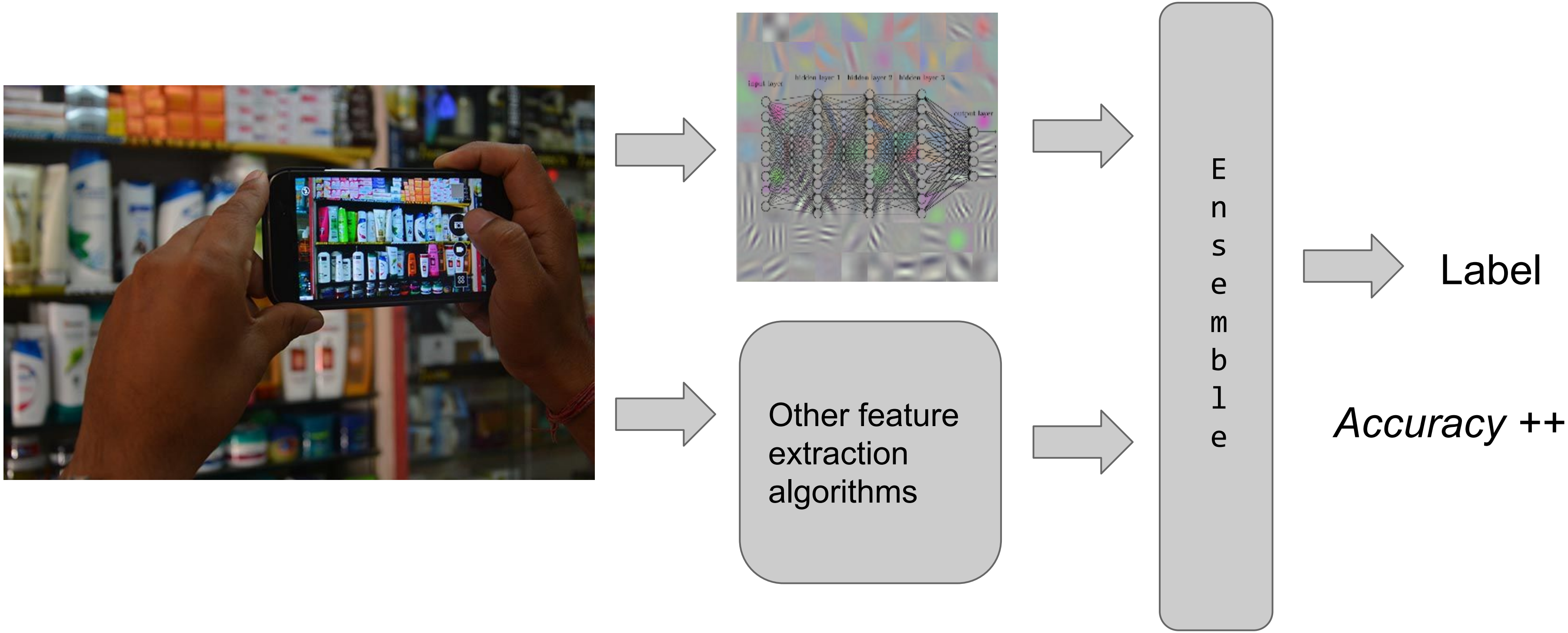
A real-life example

Quora





A real-life example: improved solution



Other feature extraction algorithms

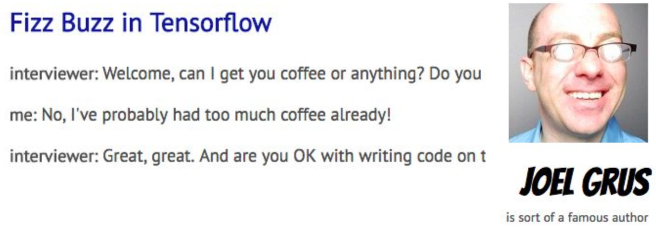
Ensemble

Label

Accuracy ++

Another real example

- Goal: Supervised Classification
  - 40 features
  - 10k examples
- What did the ML Engineer choose?
  - Multi-layer ANN trained with Tensor Flow
- What was his proposed next step?
  - Try ConvNets
- Where is the problem?
  - Hours to train, already looking into distributing
  - There are much simpler approaches



Fizz Buzz in Tensorflow

interviewer: Welcome, can I get you coffee or anything? Do you

me: No, I've probably had too much coffee already!

interviewer: Great, great. And are you OK with writing code on t

JOEL GRUS

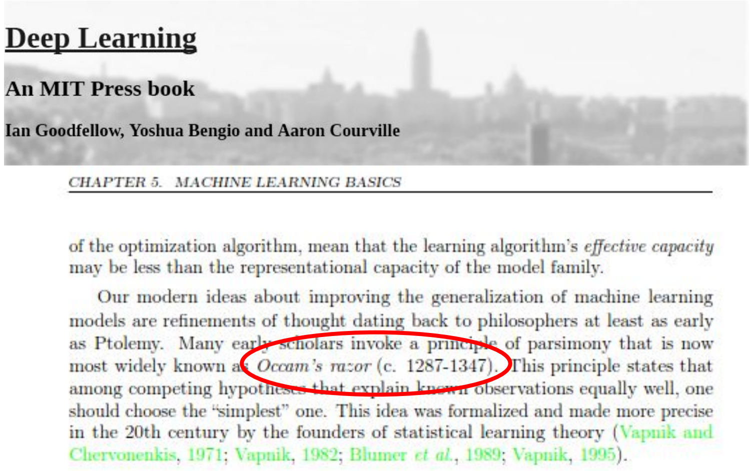
is sort of a famous author

Why DL is not the only/main solution

Occam’s Razor

Occam’s razor

- Given two models that perform more or less equally, you should always prefer the less complex
- Deep Learning might not be preferred, even if it squeezes a +1% in accuracy



Deep Learning

An MIT Press book

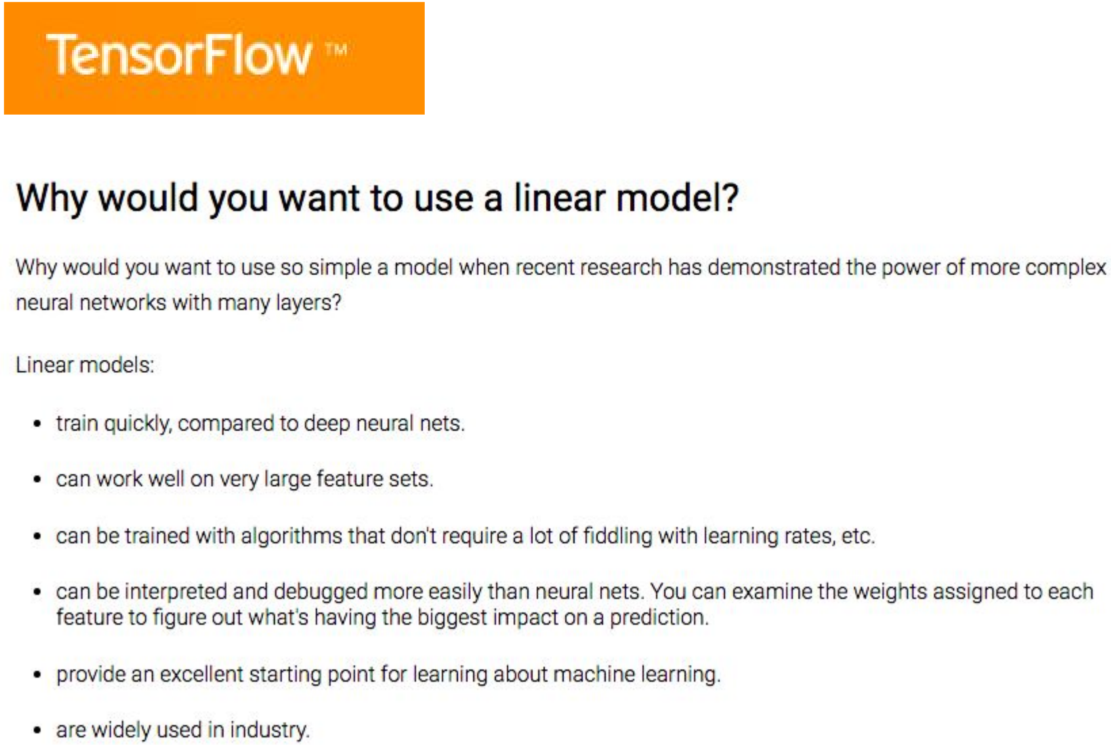
Ian Goodfellow, Yoshua Bengio and Aaron Courville

CHAPTER 5. MACHINE LEARNING BASICS

of the optimization algorithm, mean that the learning algorithm’s *effective capacity* may be less than the representational capacity of the model family.

Our modern ideas about improving the generalization of machine learning models are refinements of thought dating back to philosophers at least as early as Ptolemy. Many early scholars invoke a principle of parsimony that is now most widely known as *Occam’s razor* (c. 1287-1347). This principle states that among competing hypotheses that explain known observations equally well, one should choose the “simplest” one. This idea was formalized and made more precise in the 20th century by the founders of statistical learning theory (Vapnik and Chervonenkis, 1971; Vapnik, 1982; Blumer *et al.*, 1989; Vapnik, 1995).

Occam’s razor: reasons to prefer a simpler model



TensorFlow ™

Why would you want to use a linear model?

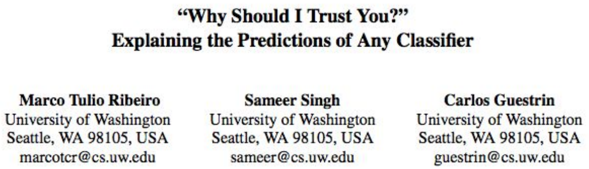
Why would you want to use so simple a model when recent research has demonstrated the power of more complex neural networks with many layers?

Linear models:

- train quickly, compared to deep neural nets.
- can work well on very large feature sets.
- can be trained with algorithms that don't require a lot of fiddling with learning rates, etc.
- can be interpreted and debugged more easily than neural nets. You can examine the weights assigned to each feature to figure out what's having the biggest impact on a prediction.
- provide an excellent starting point for learning about machine learning.
- are widely used in industry.

Occam’s razor: reasons to prefer a simpler model

- There are many others
  - ○ System complexity
  - ○ Maintenance
  - ○ Explainability
  - ○ ….

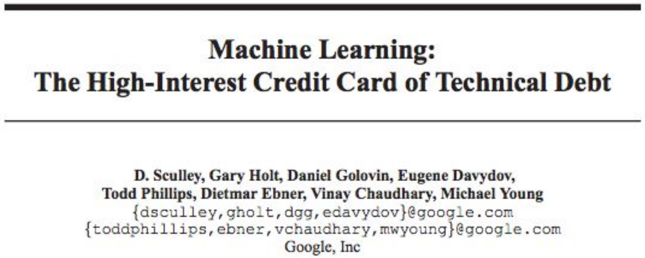


“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

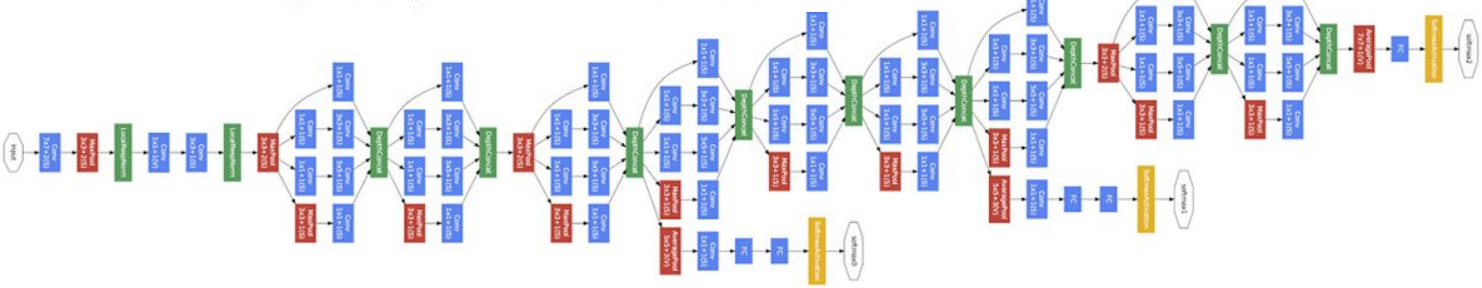
Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu



Machine Learning:
The High-Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov,
Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
{dsculley,gholt,dgg,edavydov}@google.com
{toddphillips,ebner,vchaudhary,mwyoung}@google.com
Google, Inc

Figure 3: GoogLeNet network with all the bells and whistles.



No Free Lunch

No Free Lunch Theorem

“ (…) any two optimization algorithms are equivalent when their performance is averaged across all possible problems".

“if an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.”

Feature Engineering

Feature Engineering

Need for feature engineering

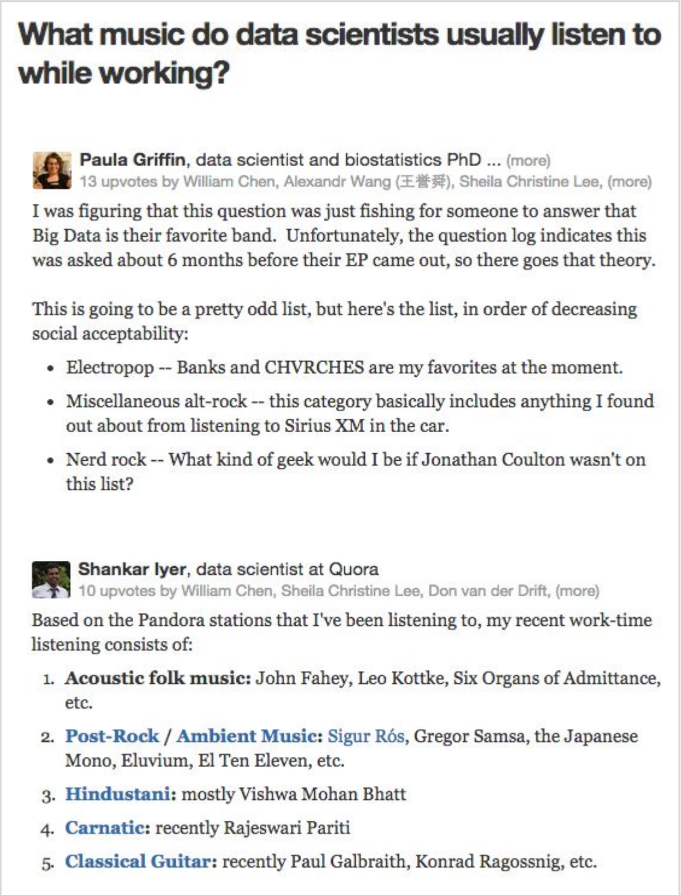
In many cases an understanding of the domain will lead to optimal results.

Feature Engineering Example - Quora Answer Ranking

Quora

What is a good Quora answer?

- truthful
- reusable
- provides explanation
- well formatted
- ...



What music do data scientists usually listen to while working?

Paula Griffin, data scientist and biostatistics PhD ... (more)
13 upvotes by William Chen, Alexandr Wang (王誉舜), Sheila Christine Lee, (more)

I was figuring that this question was just fishing for someone to answer that Big Data is their favorite band. Unfortunately, the question log indicates this was asked about 6 months before their EP came out, so there goes that theory.

This is going to be a pretty odd list, but here's the list, in order of decreasing social acceptability:

- Electropop -- Banks and CHVRCHES are my favorites at the moment.
- Miscellaneous alt-rock -- this category basically includes anything I found out about from listening to Sirius XM in the car.
- Nerd rock -- What kind of geek would I be if Jonathan Coulton wasn't on this list?

Shankar Iyer, data scientist at Quora
10 upvotes by William Chen, Sheila Christine Lee, Don van der Drift, (more)

Based on the Pandora stations that I've been listening to, my recent work-time listening consists of:

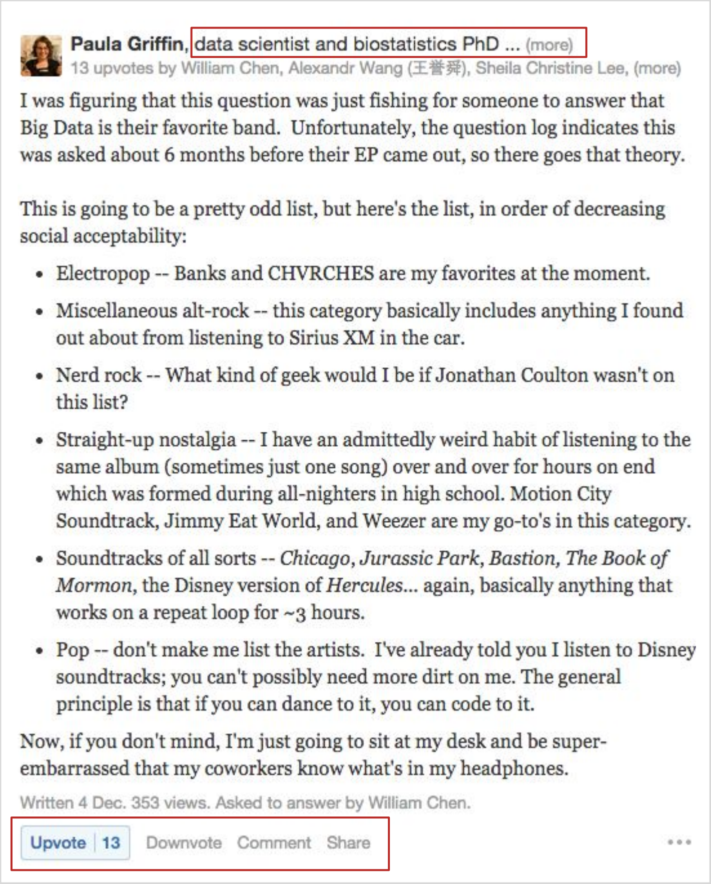
1. **Acoustic folk music:** John Fahey, Leo Kottke, Six Organs of Admittance, etc.
2. **Post-Rock / Ambient Music:** Sigur Rós, Gregor Samsa, the Japanese Mono, Eluvium, El Ten Eleven, etc.
3. **Hindustani:** mostly Vishwa Mohan Bhatt
4. **Carnatic:** recently Rajeswari Pariti
5. **Classical Guitar:** recently Paul Galbraith, Konrad Ragossnig, etc.

Feature Engineering Example - Quora Answer Ranking

Quora

How are those dimensions translated into features?

- Features that relate to the answer quality itself
- Interaction features (upvotes/downvotes, clicks, comments…)
- User features (e.g. expertise in topic)



Paula Griffin, data scientist and biostatistics PhD ... (more)
13 upvotes by William Chen, Alexandr Wang (王誉舜), Sheila Christine Lee, (more)

I was figuring that this question was just fishing for someone to answer that Big Data is their favorite band. Unfortunately, the question log indicates this was asked about 6 months before their EP came out, so there goes that theory.

This is going to be a pretty odd list, but here's the list, in order of decreasing social acceptability:

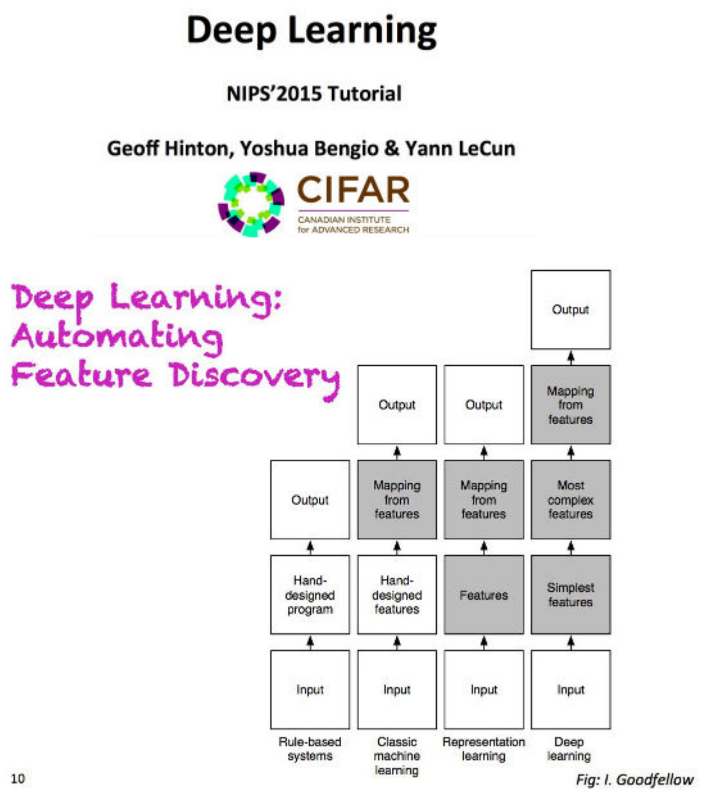
- Electropop -- Banks and CHVRCHES are my favorites at the moment.
- Miscellaneous alt-rock -- this category basically includes anything I found out about from listening to Sirius XM in the car.
- Nerd rock -- What kind of geek would I be if Jonathan Coulton wasn't on this list?
- Straight-up nostalgia -- I have an admittedly weird habit of listening to the same album (sometimes just one song) over and over for hours on end which was formed during all-nighters in high school. Motion City Soundtrack, Jimmy Eat World, and Weezer are my go-to's in this category.
- Soundtracks of all sorts -- *Chicago*, *Jurassic Park*, *Bastion*, *The Book of Mormon*, the Disney version of *Hercules*... again, basically anything that works on a repeat loop for ~3 hours.
- Pop -- don't make me list the artists. I've already told you I listen to Disney soundtracks; you can't possibly need more dirt on me. The general principle is that if you can dance to it, you can code to it.

Now, if you don't mind, I'm just going to sit at my desk and be super-embarrassed that my coworkers know what's in my headphones.

Written 4 Dec. 353 views. Asked to answer by William Chen.

Feature Engineering

- Properties of a well-behaved ML feature:
  - Reusable
  - Transformable
  - Interpretable
  - Reliable



Deep Learning

NIPS'2015 Tutorial

Geoff Hinton, Yoshua Bengio & Yann LeCun

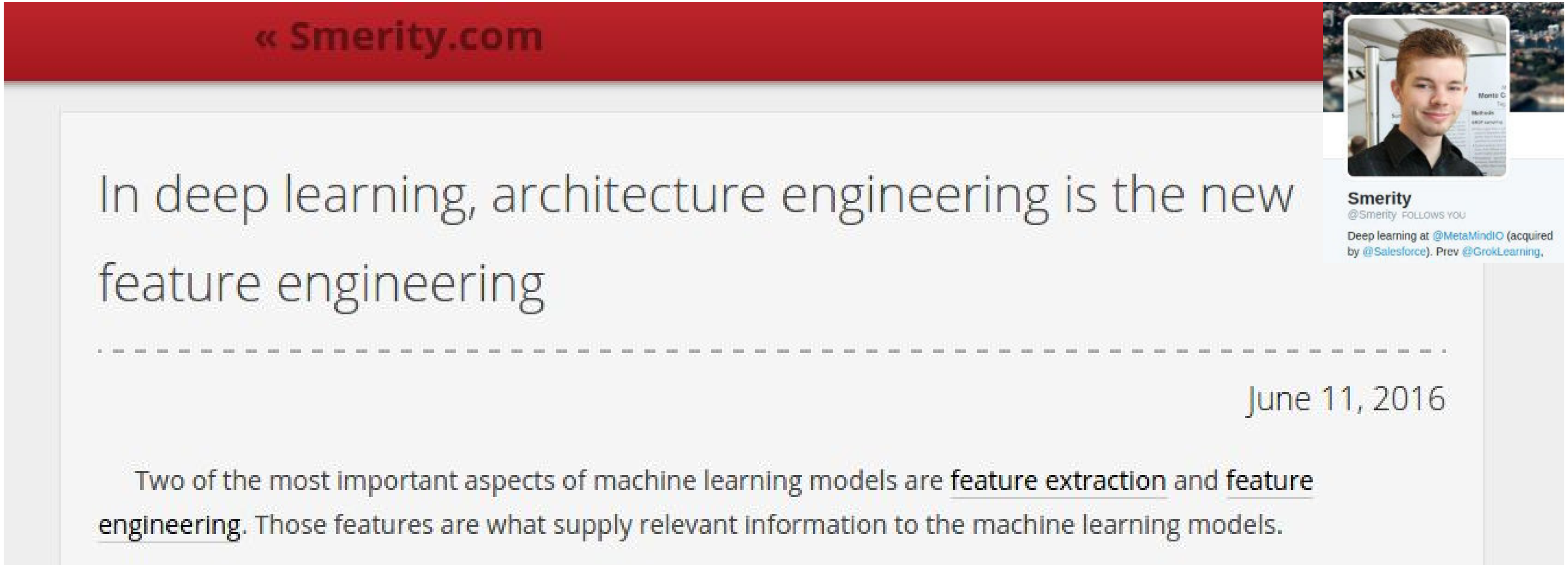
CIFAR CANADIAN INSTITUTE for ADVANCED RESEARCH

Deep Learning: Automating Feature Discovery

10

Fig: I. Goodfellow

Deep Learning and Feature Engineering



« Smerity.com

In deep learning, architecture engineering is the new feature engineering

June 11, 2016

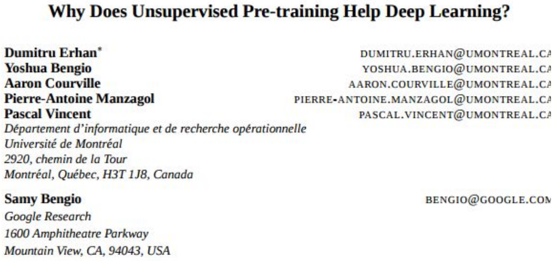
Two of the most important aspects of machine learning models are feature extraction and feature engineering. Those features are what supply relevant information to the machine learning models.

Smerity
@Smerity FOLLOWS YOU
Deep learning at @MetaMindIO (acquired by @Salesforce). Prev @GrokLearning,

Unsupervised Learning

Unsupervised Learning

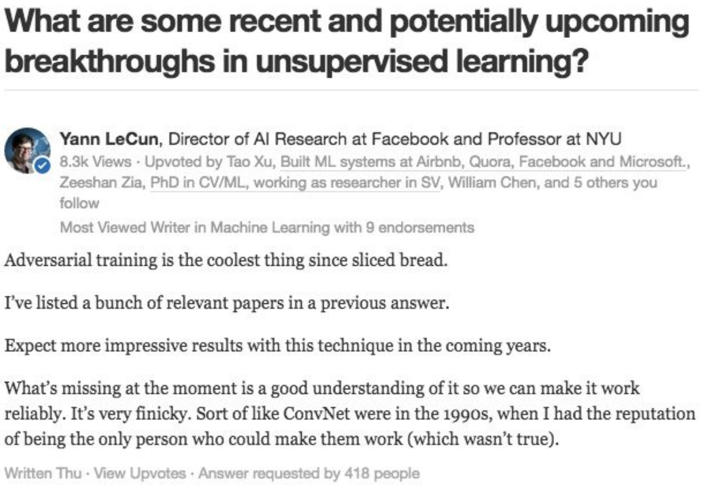
- Unsupervised learning is a very important paradigm in theory and in practice
- So far, unsupervised learning has helped deep learning, but the inverse is not true… yet



Why Does Unsupervised Pre-training Help Deep Learning?

Dumitru Erhan* DUMITRU.ERHAN@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Aaron Courville AARON.COURVILLE@UMONTREAL.CA
Pierre-Antoine Manzagol PIERRE-ANTOINE.MANZAGOL@UMONTREAL.CA
Pascal Vincent PASCAL.VINCENT@UMONTREAL.CA
Département d'informatique et de recherche opérationnelle
Université de Montréal
2920, chemin de la Tour
Montréal, Québec, H3T 1J8, Canada

Samy Bengio BENGIO@GOOGLE.COM
Google Research
1600 Amphitheatre Parkway
Mountain View, CA, 94043, USA



What are some recent and potentially upcoming breakthroughs in unsupervised learning?

Yann LeCun, Director of AI Research at Facebook and Professor at NYU
8.3k Views · Upvoted by Tao Xu, Built ML systems at Airbnb, Quora, Facebook and Microsoft., Zeeshan Zia, PhD in CV/ML, working as researcher in SV, William Chen, and 5 others you follow
Most Viewed Writer in Machine Learning with 9 endorsements

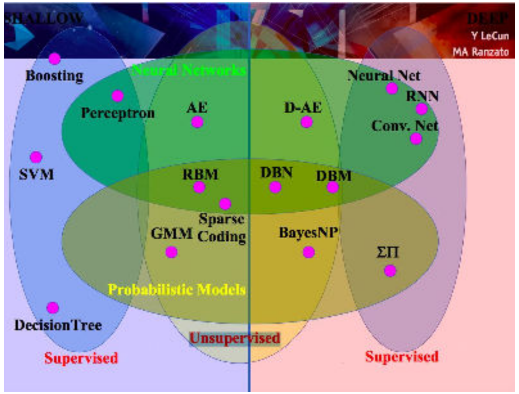
Adversarial training is the coolest thing since sliced bread.

I've listed a bunch of relevant papers in a previous answer.

Expect more impressive results with this technique in the coming years.

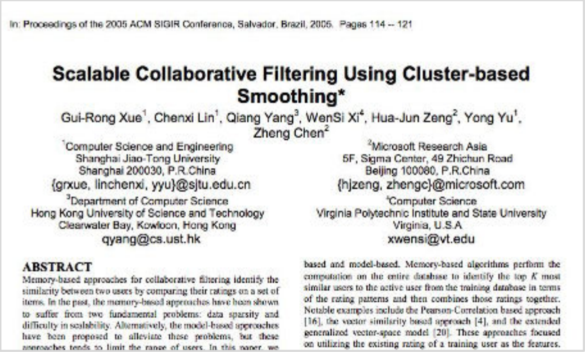
What's missing at the moment is a good understanding of it so we can make it work reliably. It's very finicky. Sort of like ConvNet were in the 1990s, when I had the reputation of being the only person who could make them work (which wasn't true).

Written Thu · View Upvotes · Answer requested by 418 people



Supervised/Unsupervised Learning

- Unsupervised learning as dimensionality reduction
- Unsupervised learning as feature engineering
- The “magic” behind combining unsupervised/supervised learning
    - E.g.1 clustering + knn
    - E.g.2 Matrix Factorization
        - MF can be interpreted as
            - Unsupervised:
                - Dimensionality Reduction a la PCA
                - Clustering (e.g. NMF)
            - Supervised
                - Labeled targets ~ regression



In: Proceedings of the 2005 ACM SIGIR Conference, Salvador, Brazil, 2005. Pages 114 -- 121

Scalable Collaborative Filtering Using Cluster-based Smoothing*

Gui-Rong Xue[1], Chenxi Lin[1], Qiang Yang[3], WenSi Xi[4], Hua-Jun Zeng[2], Yong Yu[1], Zheng Chen[2]

[1]Computer Science and Engineering
Shanghai Jiao-Tong University
Shanghai 200030, P.R.China
{grxue, linchenxi, yyu}@sjtu.edu.cn

[2]Microsoft Research Asia
5F, Sigma Center, 49 Zhichun Road
Beijing 100080, P.R.China
{hjzeng, zhengc}@microsoft.com

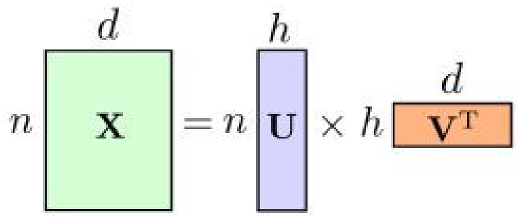
[3]Department of Computer Science
Hong Kong University of Science and Technology
Clearwater Bay, Kowloon, Hong Kong
qyang@cs.ust.hk

[4]Computer Science
Virginia Polytechnic Institute and State University
Virginia, U.S.A
xwensi@vt.edu

ABSTRACT

Memory-based approaches for collaborative filtering identify the similarity between two users by comparing their ratings on a set of items. In the past, the memory-based approaches have been shown to suffer from two fundamental problems: data sparsity and difficulty in scalability. Alternatively, the model-based approaches have been proposed to alleviate these problems, but these approaches tends to limit the range of users. In this paper, we

based and model-based. Memory-based algorithms perform the computation on the entire database to identify the top K most similar users to the active user from the training database in terms of the rating patterns and then combines those ratings together. Notable examples include the Pearson-Correlation based approach [16], the vector similarity based approach [4], and the extended generalized vector-space model [20]. These approaches focused on utilizing the existing rating of a training user as the features.



Ensembles

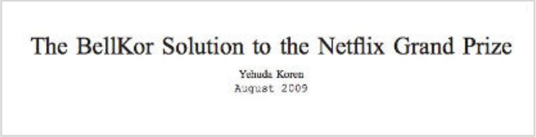
Ensembles

Even if all problems end up being suited for Deep Learning, there will always be a place for ensembles.

- Given the output of a Deep Learning prediction, you will be able to combine it with some other model or feature to improve the results.

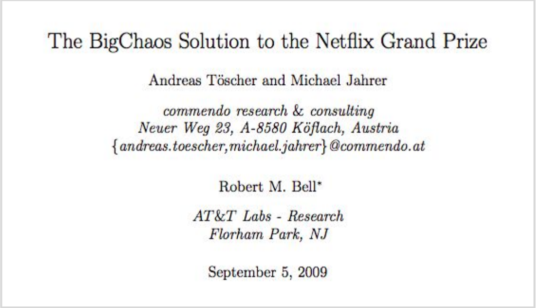
Ensembles

- Netflix Prize was won by an ensemble
  - Initially Bellkor was using GDBTs
  - BigChaos introduced ANN-based ensemble
- Most practical applications of ML run an ensemble
  - Why wouldn't you?
  - At least as good as the best of your methods
  - Can add completely different approaches



The BellKor Solution to the Netflix Grand Prize

Yehuda Koren
August 2009



The BigChaos Solution to the Netflix Grand Prize

Andreas Töscher and Michael Jahrer

commendo research & consulting
Neuer Weg 23, A-8580 Köflach, Austria
{andreas.toescher,michael.jahrer}@commendo.at

Robert M. Bell*

AT&T Labs - Research
Florham Park, NJ

September 5, 2009

Ensembles & Feature Engineering

Quora

- Ensembles are the way to turn any model into a feature!
- E.g. Don’t know if the way to go is to use Factorization Machines, Tensor Factorization, or RNNs?
    - Treat each model as a “feature”
    - Feed them into an ensemble

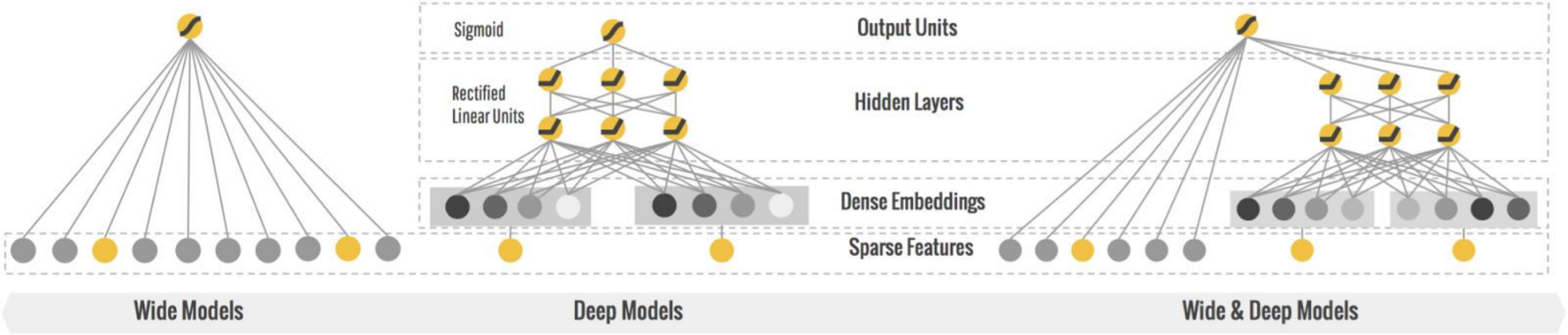
Google Research Blog

The latest news from Research at Google

Wide & Deep Learning: Better Together with TensorFlow

Wednesday, June 29, 2016

Posted by Heng-Tze Cheng, Senior Software Engineer, Google Research

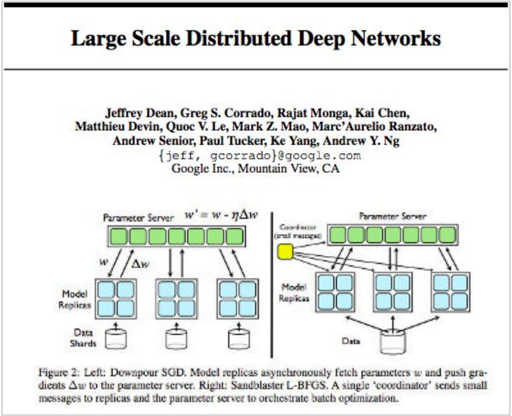


Distributing Algorithms

Distributing ML

- Most of what people do in practice can fit into a multi-core machine
  - Smart data sampling
  - Offline schemes
  - Efficient parallel code
- … but not Deep ANNs

- Do you care about costs? How about latencies or system complexity/debuggability?

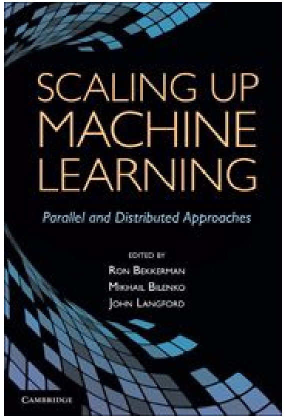


Large Scale Distributed Deep Networks

Jeffrey Dean, Greg S. Corrado, Rajat Monga, Kai Chen,
Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc'Aurelio Ranzato,
Andrew Senior, Paul Tucker, Ke Yang, Andrew Y. Ng
{jeff, gcorrado}@google.com
Google Inc., Mountain View, CA

Parameter Server
$w' = w - \eta\Delta w$
w
Δw
Model Replicas
Data Shards
Coordinator (small messages)
Parameter Server
Model Replicas
Data

Figure 2: Left: Downpour SGD. Model replicas asynchronously fetch parameters w and push gradients Δw to the parameter server. Right: Sandblaster L-BFGS. A single 'coordinator' sends small messages to replicas and the parameter server to orchestrate batch optimization.



SCALING UP MACHINE LEARNING

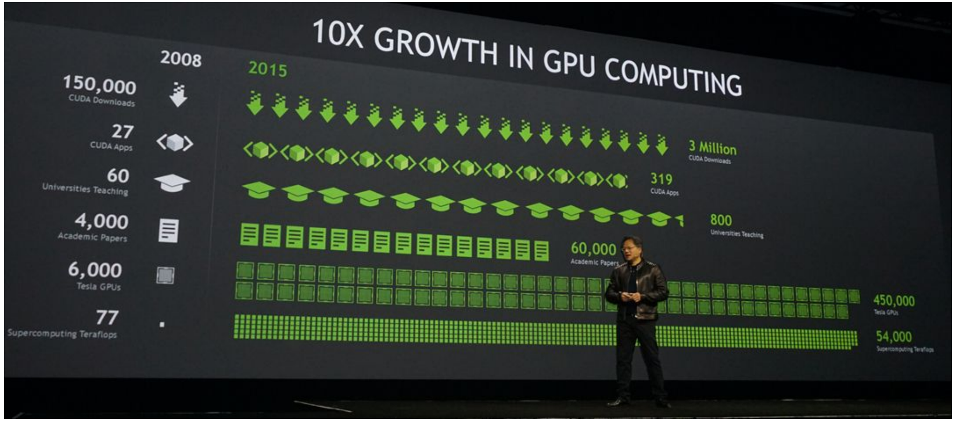
Parallel and Distributed Approaches

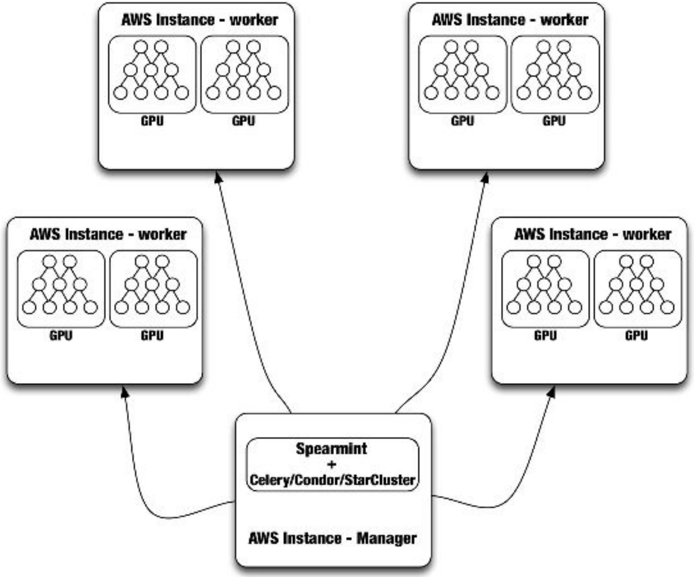
EDITED BY
RON BEKKERMAN
MIKHAIL BILENKO
JOHN LANGFORD

CAMBRIDGE

Distributing ML

- That said…
- Deep Learning has managed to get away by promoting a “new paradigm” of parallel computing: GPU’s



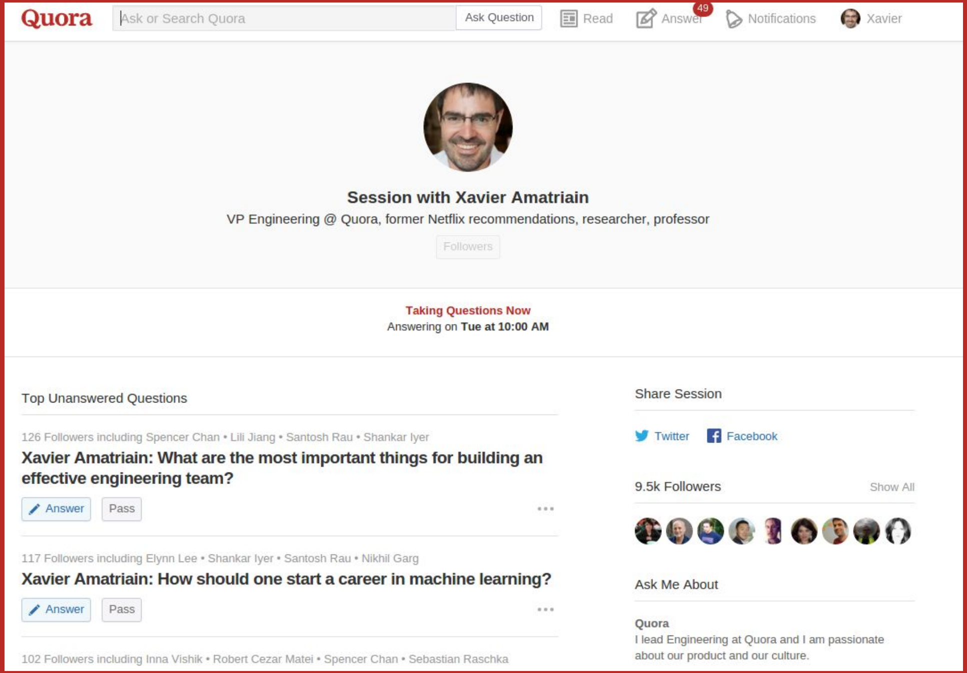


Conclusions

Conclusions

- Deep Learning has had some impressive results lately
- However, Deep Learning is not the only solution
  - It is dangerous to oversell Deep Learning
- Important to take other things into account
  - Other approaches/models
  - Feature Engineering
  - Unsupervised Learning
  - Ensembles
  - Need to distribute, costs, system complexity...

Questions?



Session with Xavier Amatriain

VP Engineering @ Quora, former Netflix recommendations, researcher, professor

Taking Questions Now

Answering on **Tue at 10:00 AM**

Top Unanswered Questions

126 Followers including Spencer Chan • Lili Jiang • Santosh Rau • Shankar Iyer

Xavier Amatriain: What are the most important things for building an effective engineering team?

117 Followers including Elynn Lee • Shankar Iyer • Santosh Rau • Nikhil Garg

Xavier Amatriain: How should one start a career in machine learning?

102 Followers including Inna Vishik • Robert Cezar Matei • Spencer Chan • Sebastian Raschka

Share Session

9.5k Followers

Ask Me About

Quora

I lead Engineering at Quora and I am passionate about our product and our culture.

Quora

We’re Hiring…
Deep & Shallow Learners