Distributing Large-scale ML Algorithms: from GPUs to the Cloud

MMDS 2014

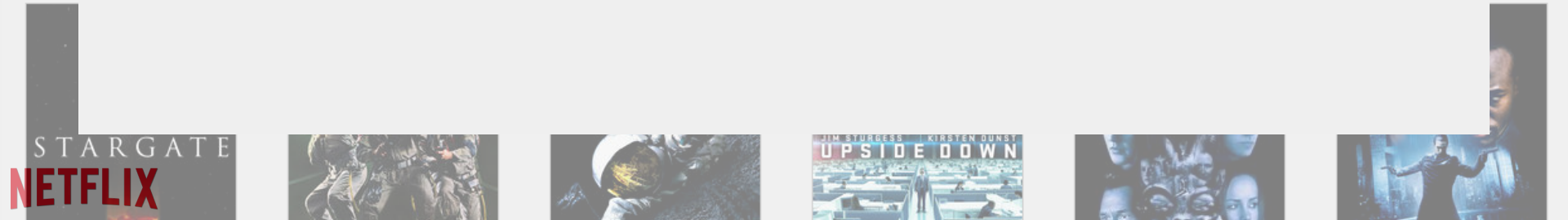
June, 2014

Xavier Amatriain
Director - Algorithms Engineering
@xamat

NETFLIX

Outline

- Introduction
- Emmy-winning Algorithms
- Distributing ML Algorithms in Practice
- An example: ANN over GPUs & AWS Cloud



Netflix Prize

COMPLETED

What we were interested in:

- High quality *recommendations*

Proxy question:

- Accuracy in predicted rating
- Improve by 10% = $1million!

Data size:

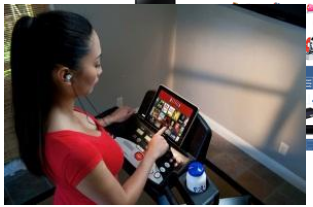
- 100M ratings (back then “almost massive”)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

NETFLIX

NETFLIX

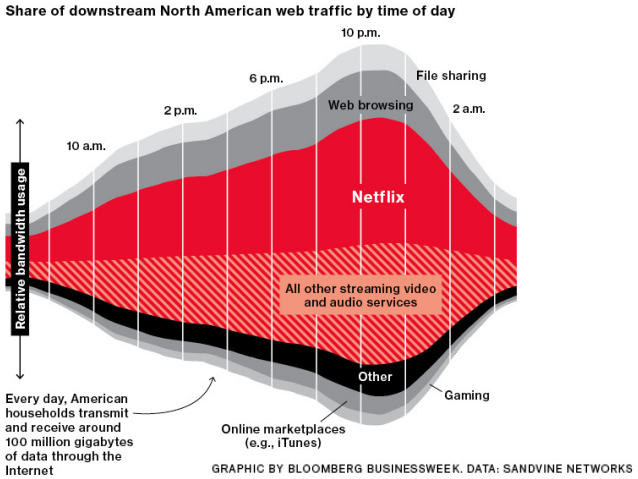
2006

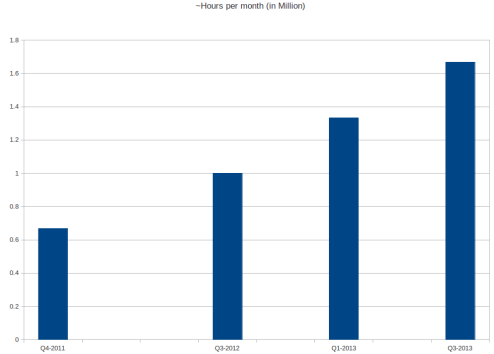


amazon web services™

2014

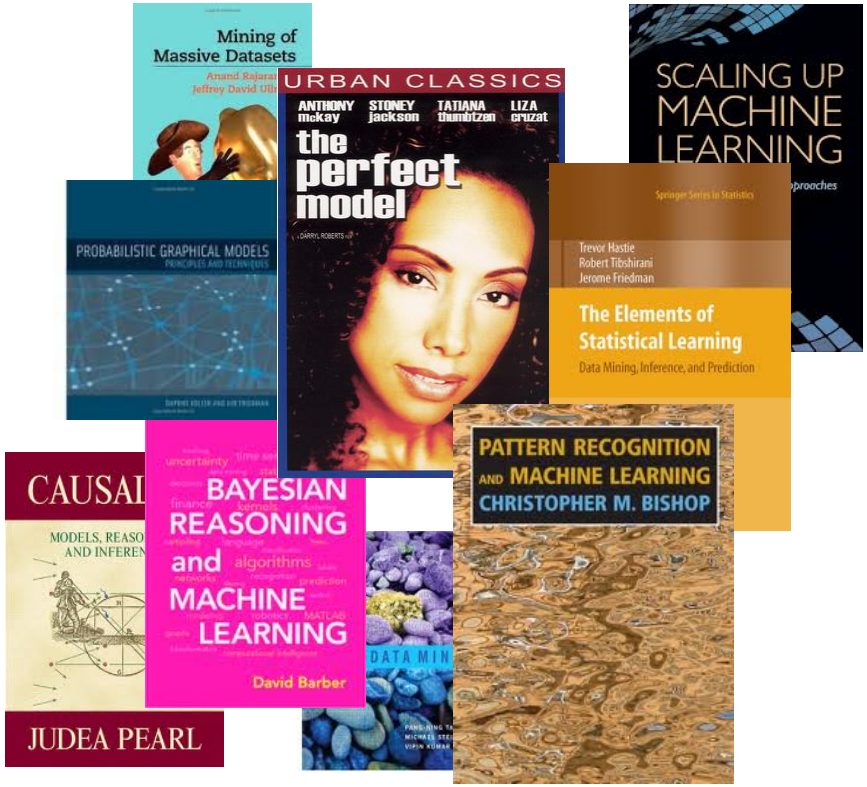
Netflix Scale





- > 44M members
- > 40 countries
- > 1000 device types
- > 5B hours in Q3 2013
- Plays: > 50M/day
- Searches: > 3M/day
- Ratings: > 5M/day
- Log 100B events/day
- 31.62% of peak US downstream traffic

Smart Models



- Regression models (Logistic, Linear, Elastic nets)
- GBDT/RF
- SVD & other MF models
- Factorization Machines
- Restricted Boltzmann Machines
- Markov Chains & other graphical models
- Clustering (from k-means to modern non-parametric models)
- Deep ANN
- LDA
- Association Rules
- …

NETFLIX

“Emmy Winning”

Netflix Algorithms

Could Iron Man's Lab Soon Be A Reality?

Facebook To Introduce New Photo Feature

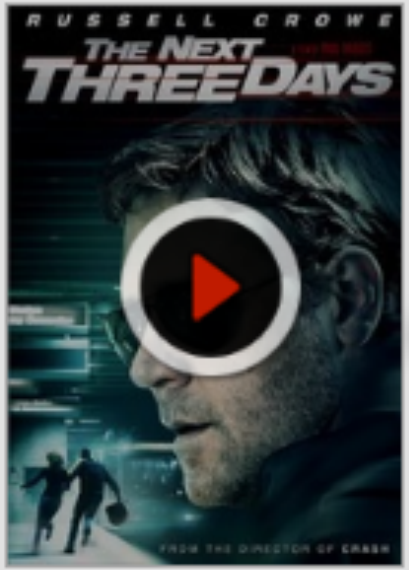
Netflix's New 'My List' Feature Knows You Better Than You Know Yourself (Because Algorithms)

The Huffington Post | By Dino Grandoni

Posted: 08/21/2013 1:44 pm EDT | Updated: 08/22/2013 8:31 am EDT

NETFLIX

Rating Prediction



The Next Three Days

2010 PG-13 133 minutes

When his wife is sent to jail on murder charges she fervently denies, a college professor hatches a meticulous plan for the ultimate prison escape. More Info

Starring: Russell Crowe, Elizabeth Banks
Director: Paul Haggis

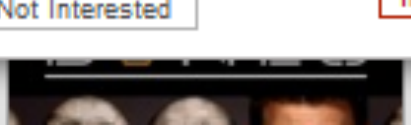
Based on your interest in: *Iron Man 2*, *John Q* and *X-Men Origins: Wolverine*

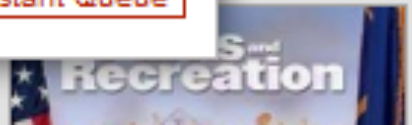
Our best guess for Xavier:

★★★★½

Not Interested

In Instant Queue

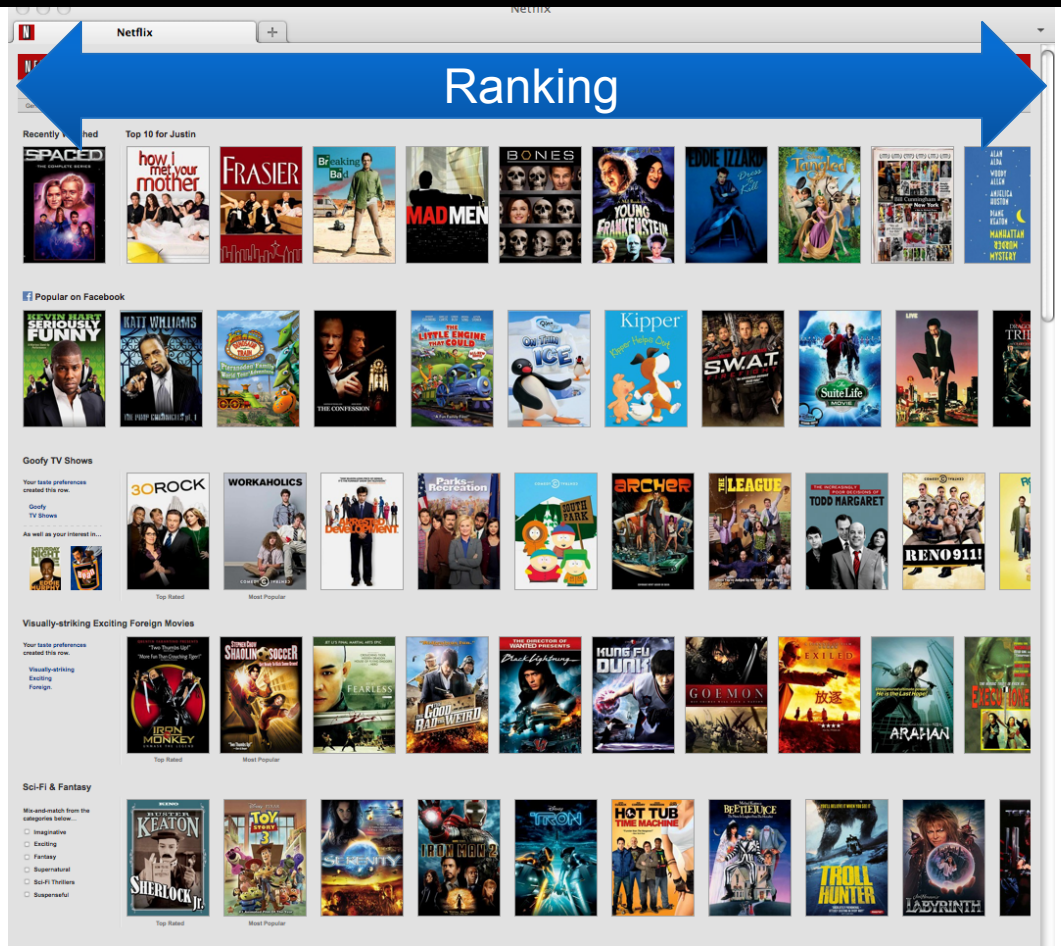




2007 Progress Prize

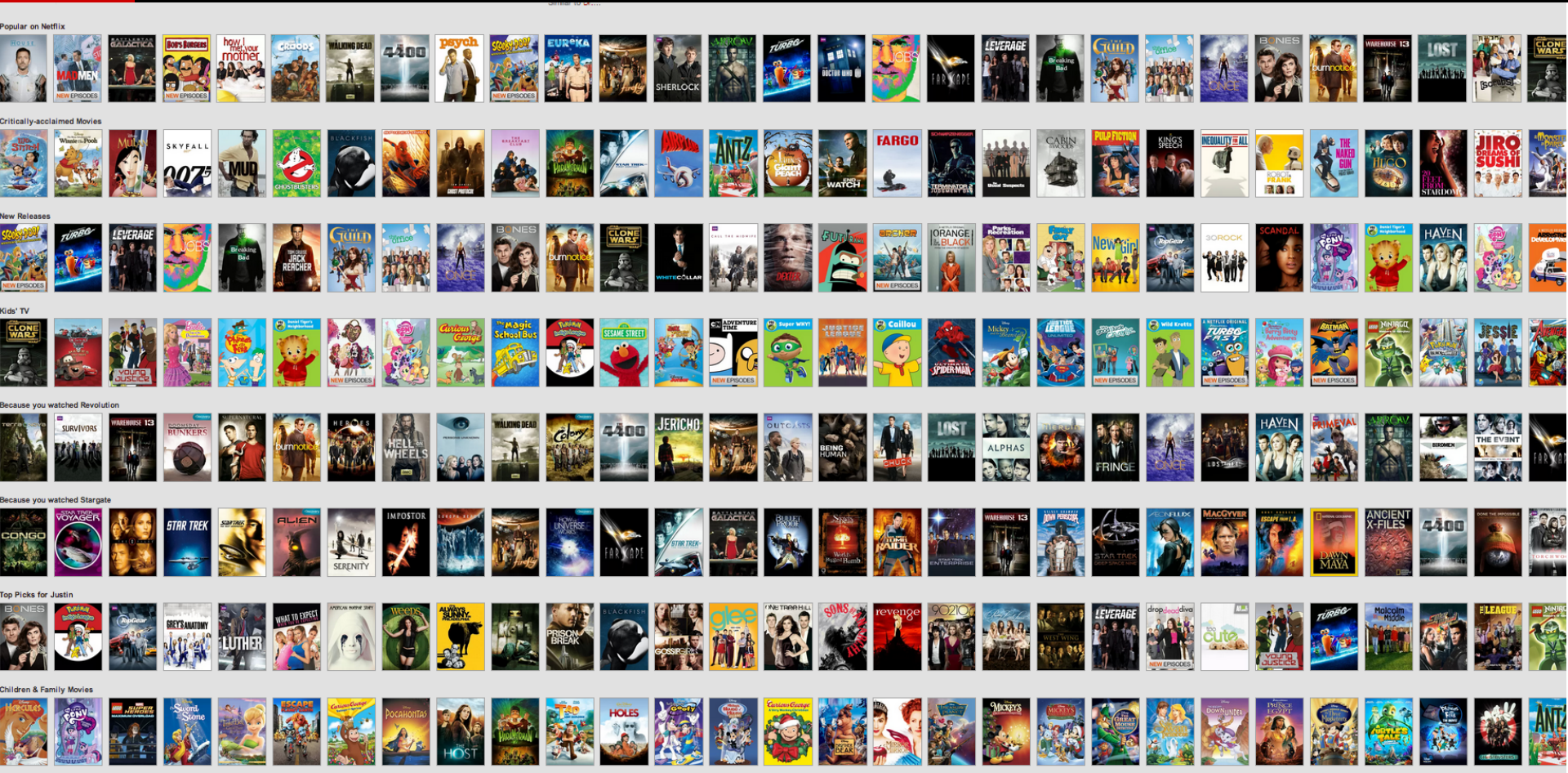
- Top 2 algorithms
  - MF/SVD - Prize RMSE: 0.8914
  - RBM - Prize RMSE: 0.8990
- Linear blend Prize RMSE: 0.88
- Currently in use as part of Netflix' rating prediction component
- Limitations
  - Designed for 100M ratings, we have 5B ratings
  - Not adaptable as users add ratings
  - Performance issues

Ranking

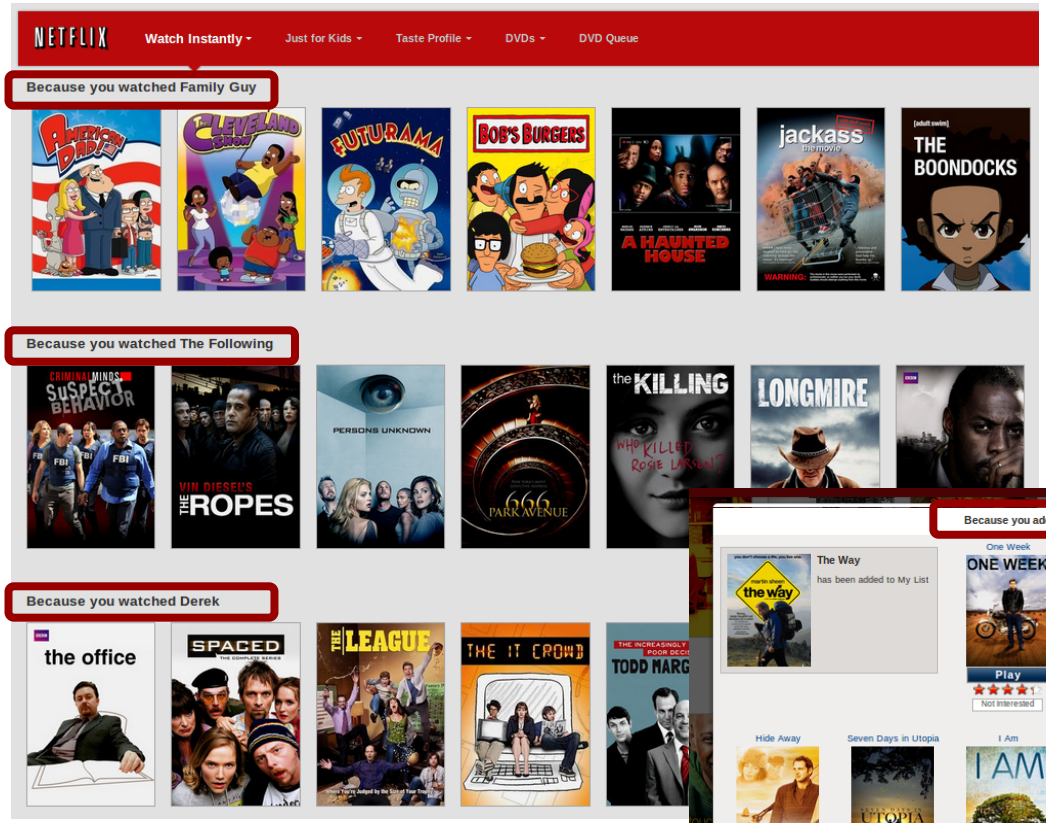


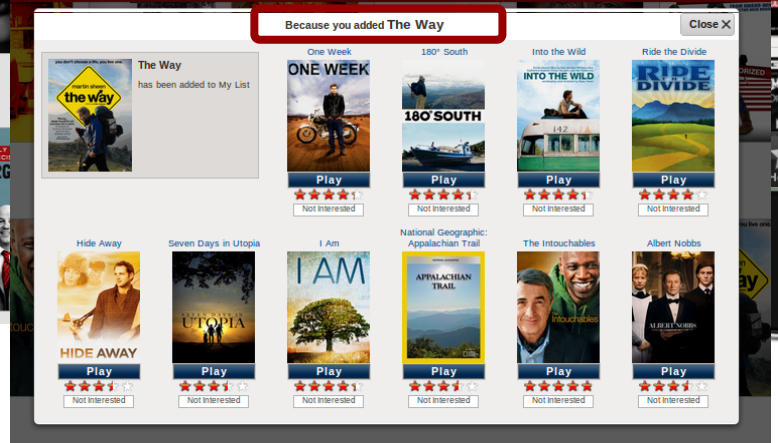
NETFLIX

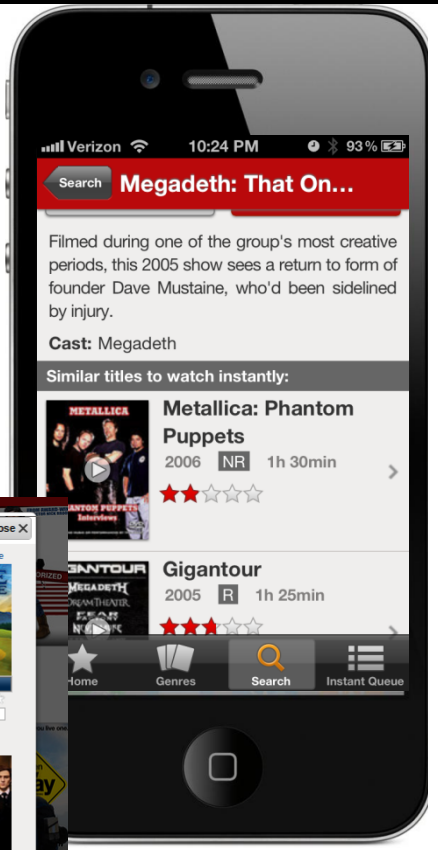
Page composition



Similarity







NETFLIX

Search Recommendations

MOVIES & TV SHOV
Marriage Italian
Smalltown, Italy
Perlasca
PEOPLE
Italia Almirante-M
Italo Renda
EXPLORE TITLES F
The Italian Job
The Italian
italy

Related to italy

Ancient Mysteries: Canals of Venice

2005 TV-G 46m

Known for its distinctive man-made canals and unparalleled aura of romance, the Italian city of Venice is like no other place on Earth.

TV Shows, Documentaries

Postplay

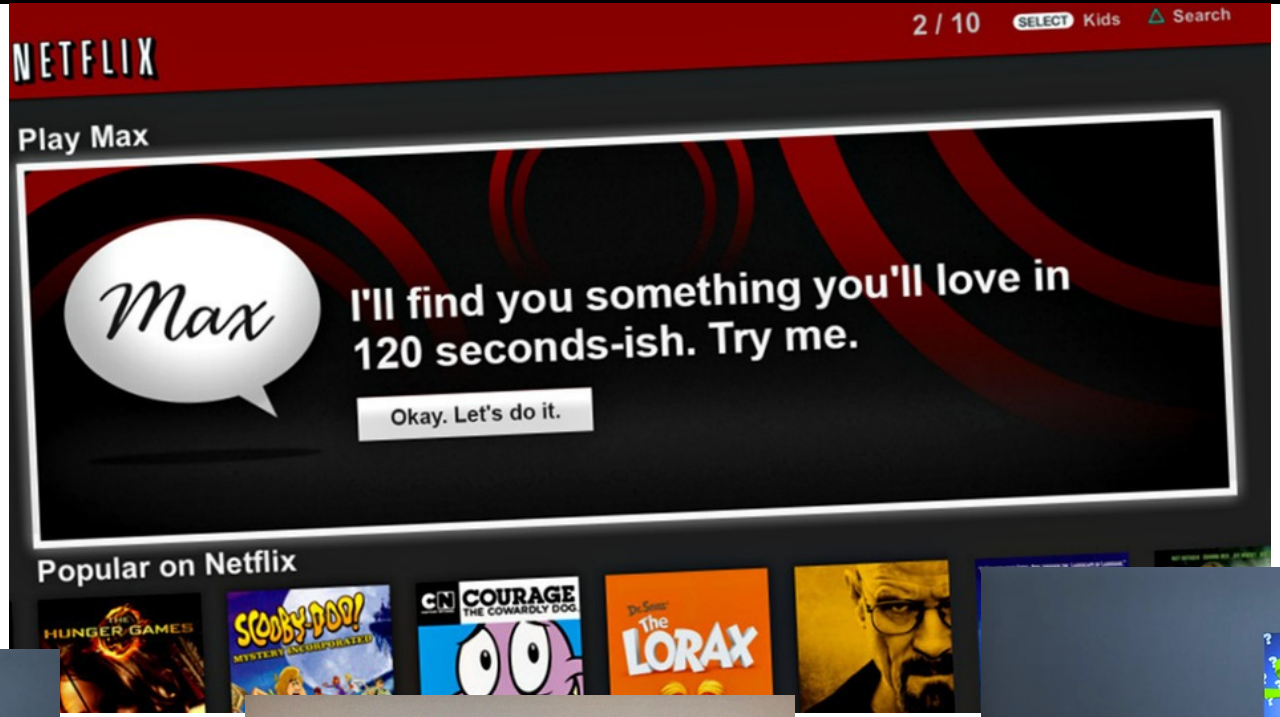
Back to browse Search

Rate The Fighter

★★★★★

Suggestions for you
to watch now...

Gamification



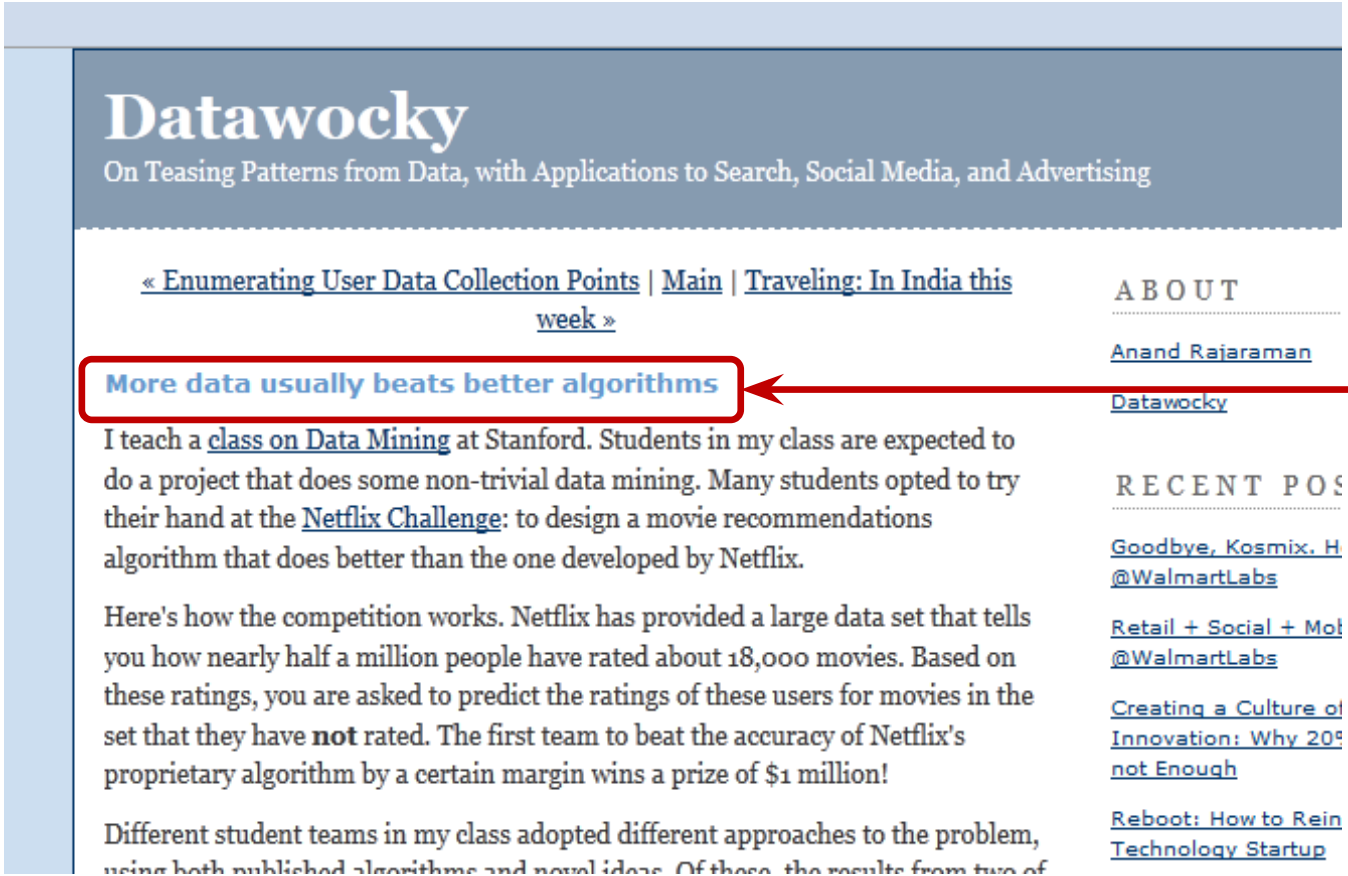
Distributing ML algorithms in practice

1. Do I need all that data?

2. At what level should I distribute/parallelize?

3. What latency can I afford?

Do I need all that data?



Datawocky

On Teasing Patterns from Data, with Applications to Search, Social Media, and Advertising

« Enumerating User Data Collection Points | Main | Traveling: In India this week »

More data usually beats better algorithms

I teach a class on Data Mining at Stanford. Students in my class are expected to do a project that does some non-trivial data mining. Many students opted to try their hand at the Netflix Challenge: to design a movie recommendations algorithm that does better than the one developed by Netflix.

Here's how the competition works. Netflix has provided a large data set that tells you how nearly half a million people have rated about 18,000 movies. Based on these ratings, you are asked to predict the ratings of these users for movies in the set that they have **not** rated. The first team to beat the accuracy of Netflix's proprietary algorithm by a certain margin wins a prize of $1 million!

Different student teams in my class adopted different approaches to the problem,

ABOUT

Anand Rajaraman

Datawocky

RECENT POS

Goodbye, Kosmix. H
@WalmartLabs

Retail + Social + Mob
@WalmartLabs

Creating a Culture of
Innovation: Why 20%
not Enough

Reboot: How to Rein
Technology Startup

Really?

Anand Rajaraman: Former Stanford Prof. & Senior VP at Walmart

NETFLIX

Sometimes, it's not about more data

Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy *
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk *,†
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that transforms user or item descriptions so that they are as close as possible to the feature vectors generated by MF for CF.

We propose methods for explicit feedback that are able to handle 140 000 features when feature vectors are very sparse. With movie metadata collected for the NP movies we show that the prediction performance of the methods is comparable to that of CF, and can be used to predict user preferences on new movies.

We also investigate the value of movie metadata compared to movie ratings in regards of predictive power. We compare

1. INTRODUCTION

The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available.

There are two basic strategies that can be applied when generating recommendations. Collaborative filtering (CF) methods are based only on the activity of users, while content-based filtering (CBF) methods use only metadata. In this paper we propose hybrid methods, which try to benefit from both information sources.

The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which

NETFLIX

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Norvig: “Google does not have better Algorithms, only more Data”

Many features/ low-bias models

[Banko and Brill, 2001]

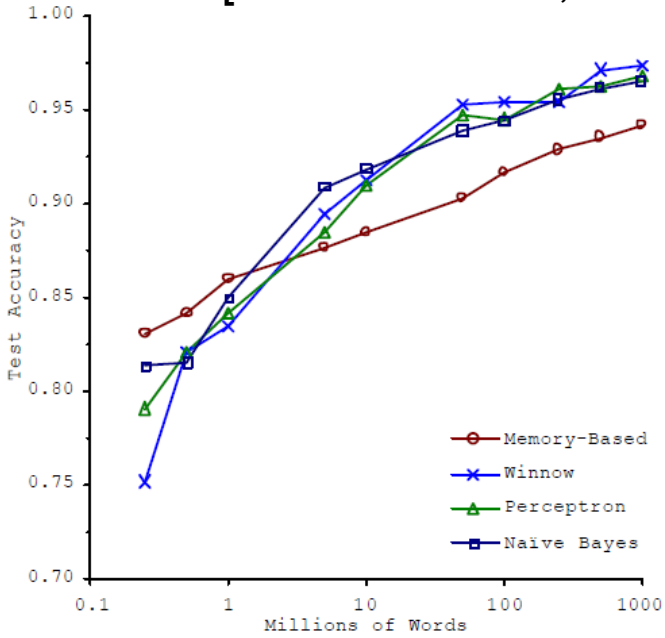
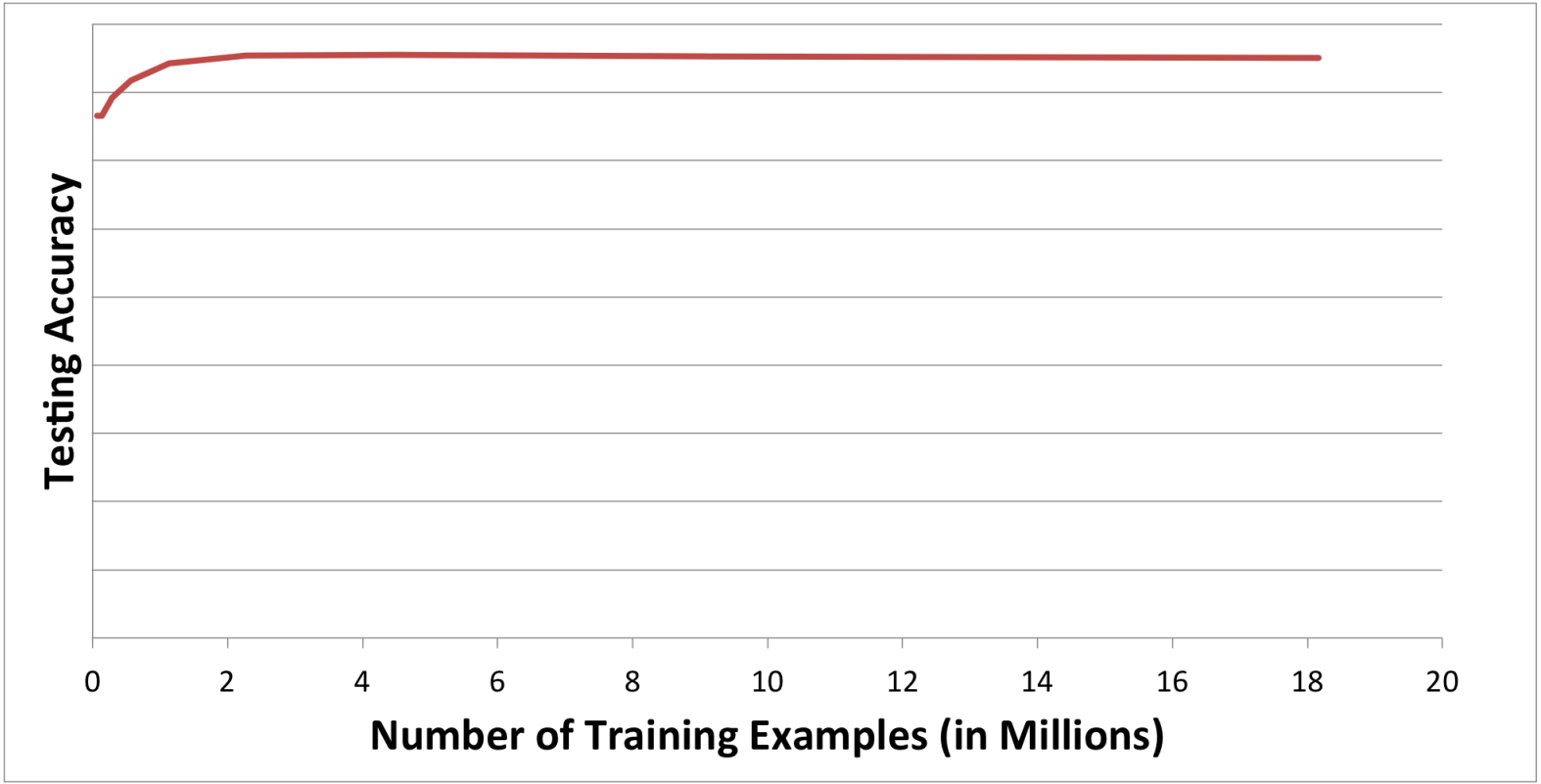


Figure 1. Learning Curves for Confusion Set Disambiguation

NETFLIX

Sometimes, it's not about more data



NETFLIX

At what level should I parallelize?

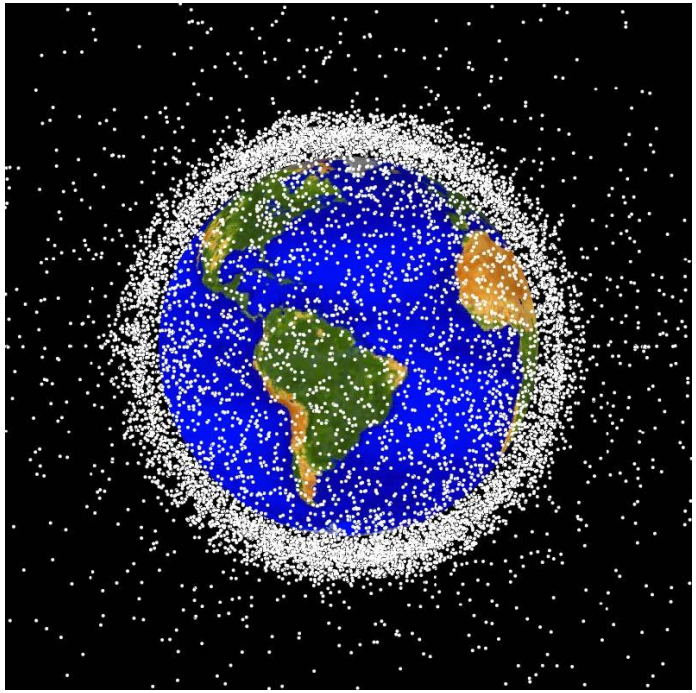
The three levels of Distribution/Parallelization

1. For each subset of the population (e.g. region)
2. For each combination of the hyperparameters
3. For each subset of the training data

Each level has different requirements

Level 1 Distribution

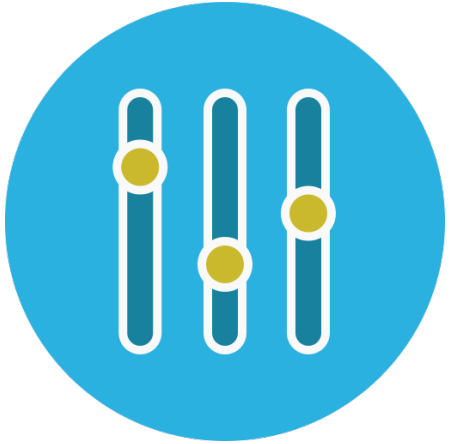
- We may have subsets of the population for which we need to train an independently optimized model.
  - Training can be fully distributed requiring no coordination or data communication



NETFLIX

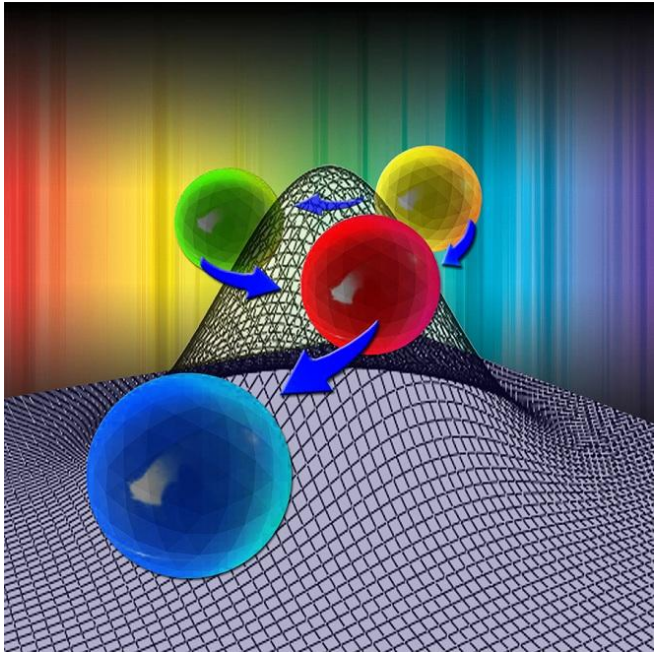
Level 2 Distribution

- For a given subset of the population we need to find the “optimal” model
- Train several models with different hyperparameter values
- Worst-case: grid search
  - Can do much better than this (E.g. Bayesian Optimization with Gaussian Process Priors)
- This process *does* require coordination
  - Need to decide on next “step”
  - Need to gather final optimal result
- Requires data distribution, not sharing

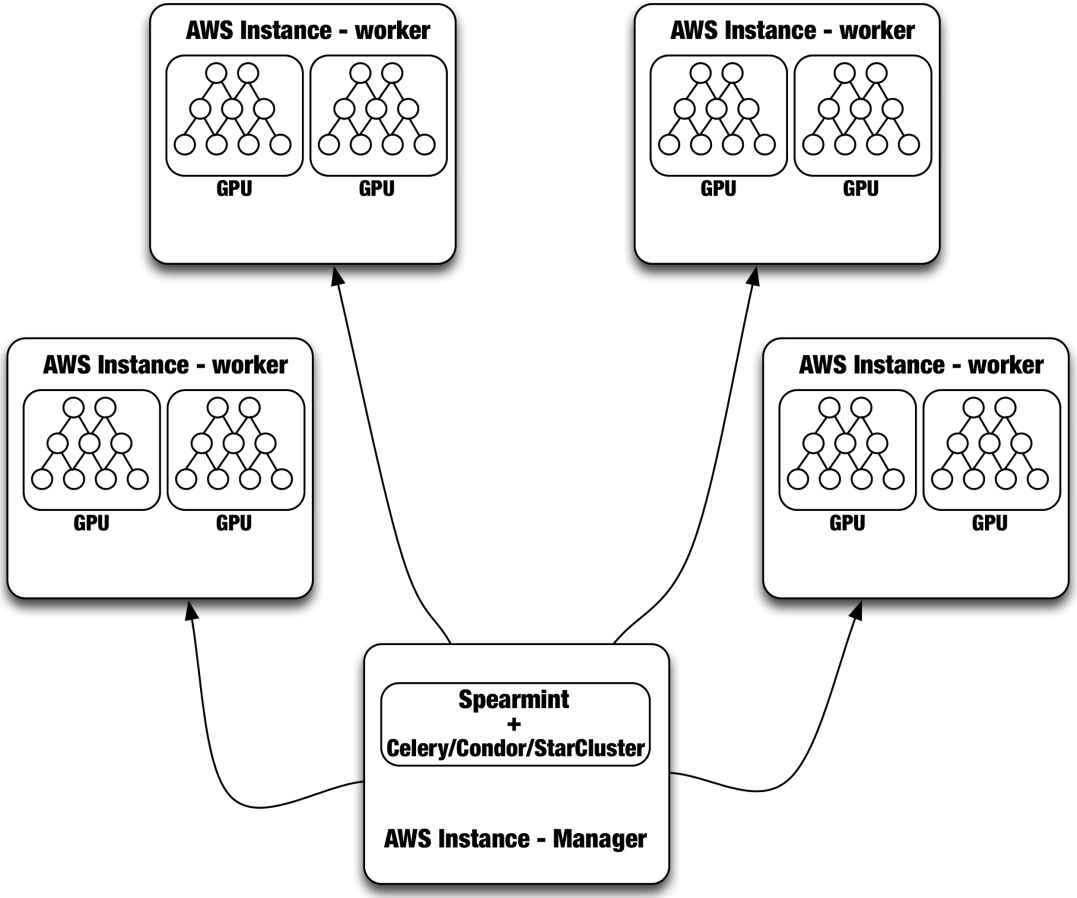


Level 3 Distribution

- For each combination of hyperparameters, model training may still be expensive
- Process requires coordination and data sharing/communication
- Can distribute computation over machines splitting examples or parameters (e.g. ADMM)
- Or parallelize on a single multicore machine (e.g. Hogwild)
- Or… use GPUs



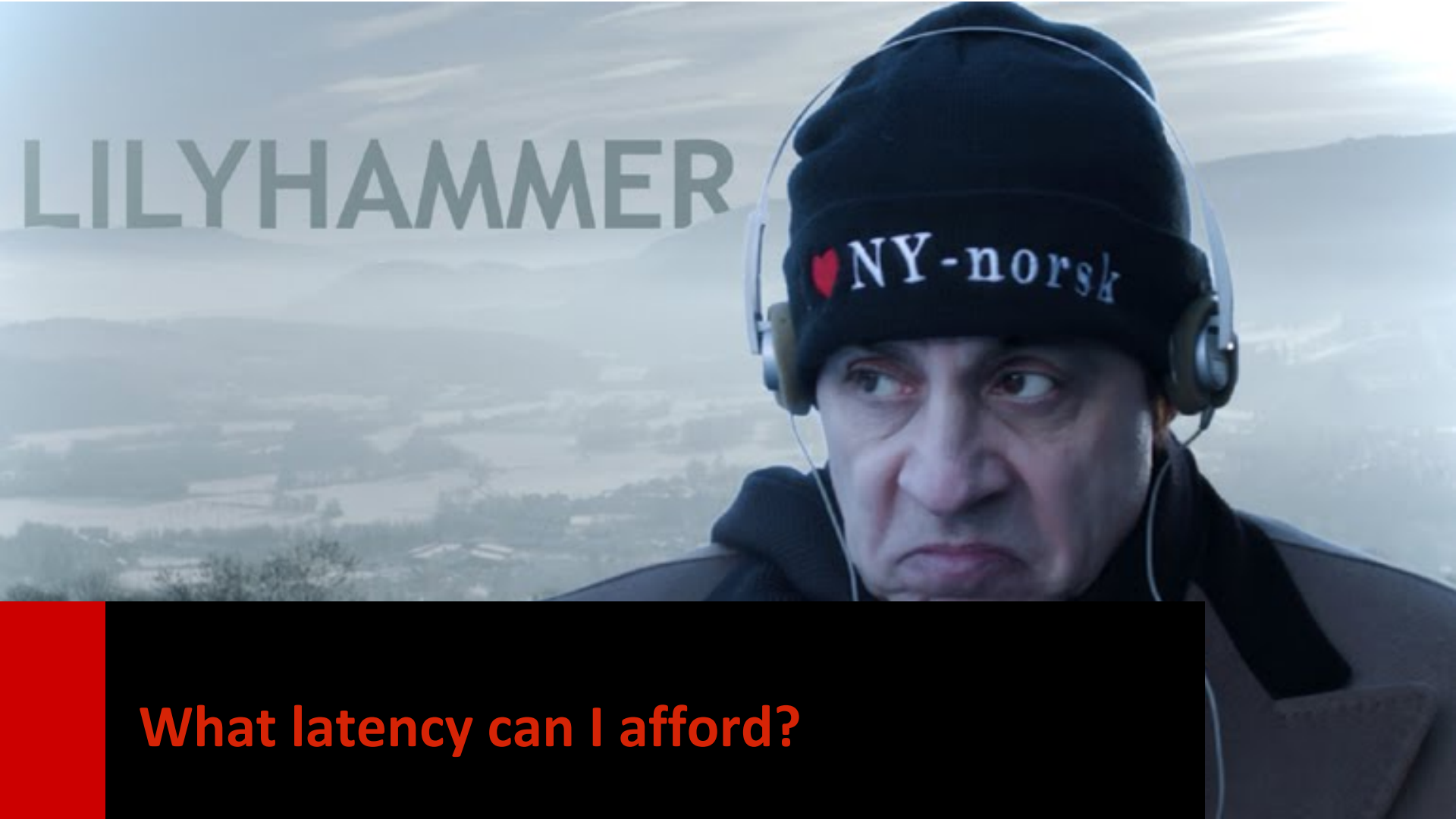
ANN Training over GPUS and AWS



NETFLIX

ANN Training over GPUS and AWS

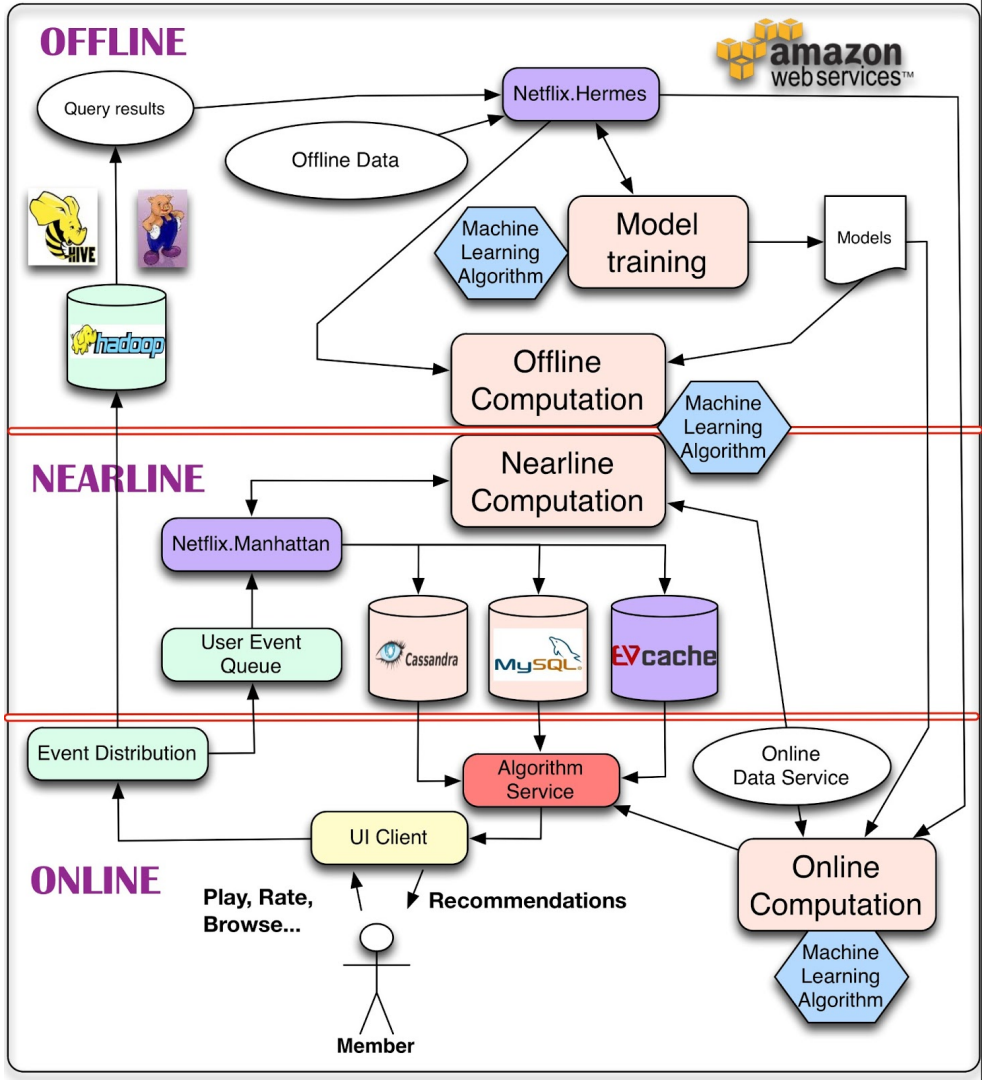
- Level 1 distribution: machines over different AWS regions
- Level 2 distribution: machines in AWS and same AWS region
  - Use coordination tools
    - Spearmint or similar for parameter optimization
    - Condor, StarCluster, Mesos… for distributed cluster coordination
- Level 3 parallelization: highly optimized parallel CUDA code on GPUs



What latency can I afford?

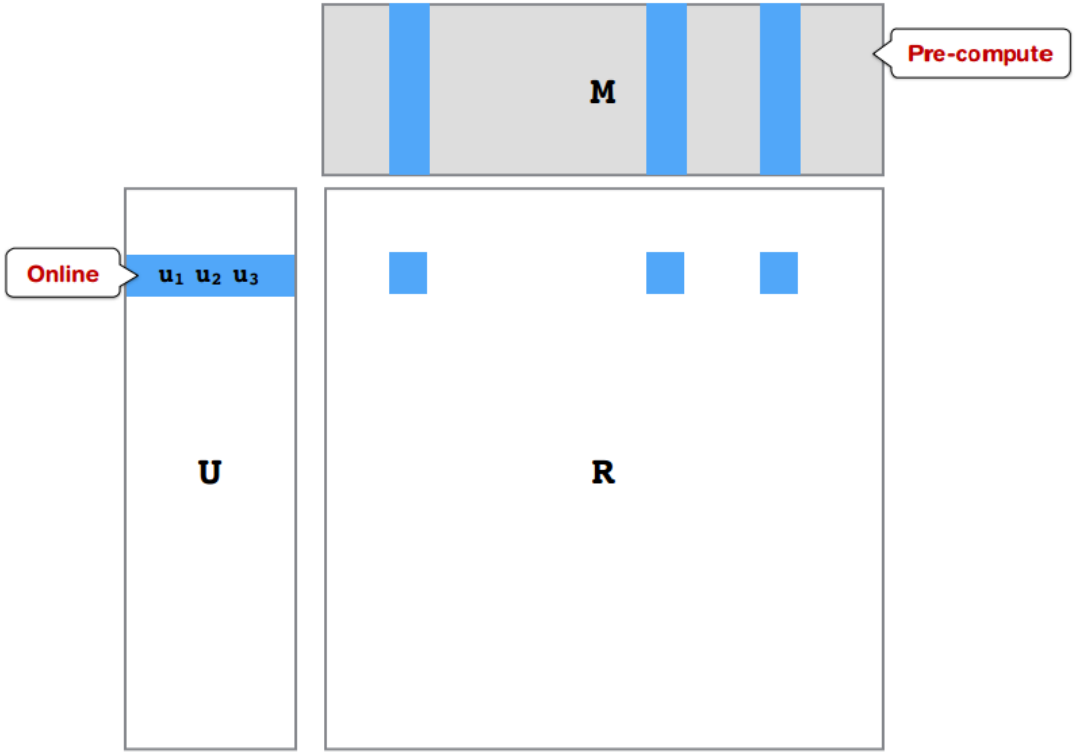
3 shades of latency

- Blueprint for multiple algorithm services
  - Ranking
  - Row selection
  - Ratings
  - Search
  - …
- Multi-layered Machine Learning



NETFLIX

Matrix Factorization Example



NETFLIX

Xavier Amatriain (@xamat)
xavier@netflix.com

Thanks!

(and yes, we are hiring)