

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

MMDS 2014

June, 2014

Xavier Amatriain
Director - Algorithms Engineering

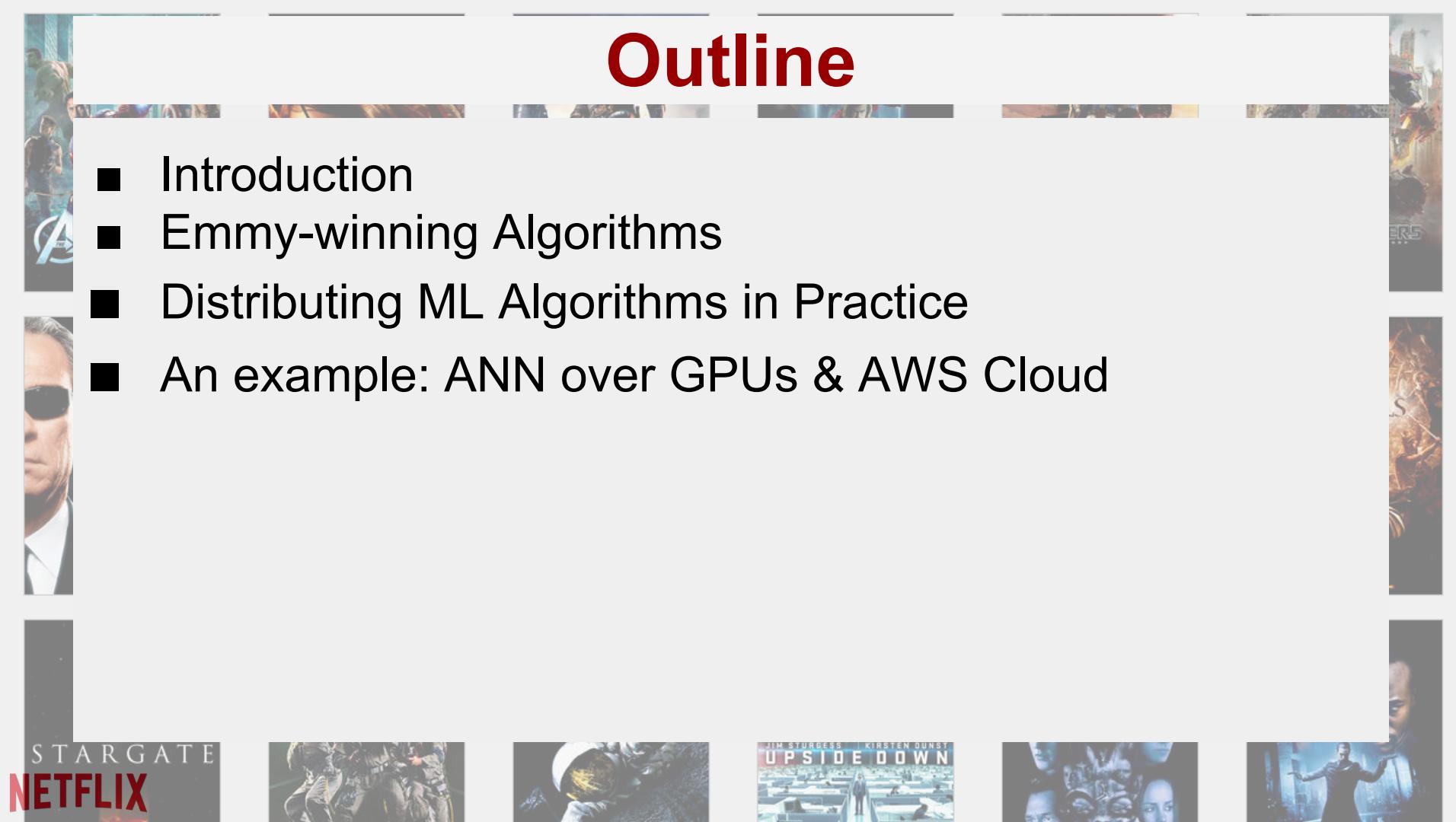


@xamat



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

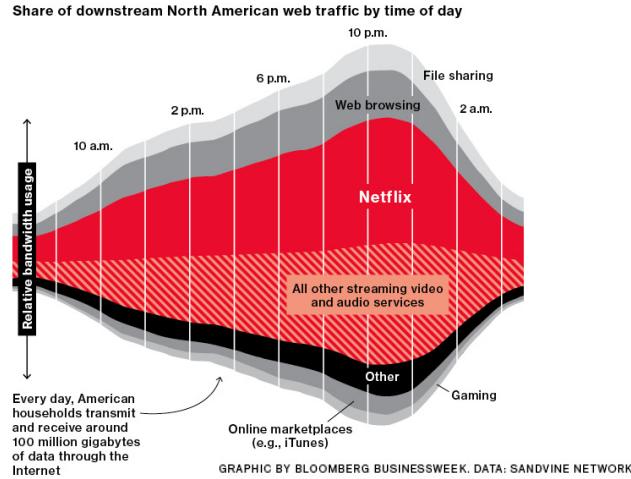


2006

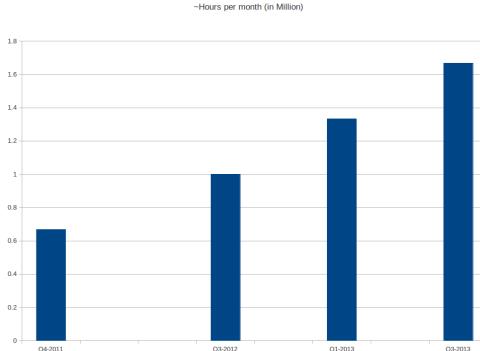


2014

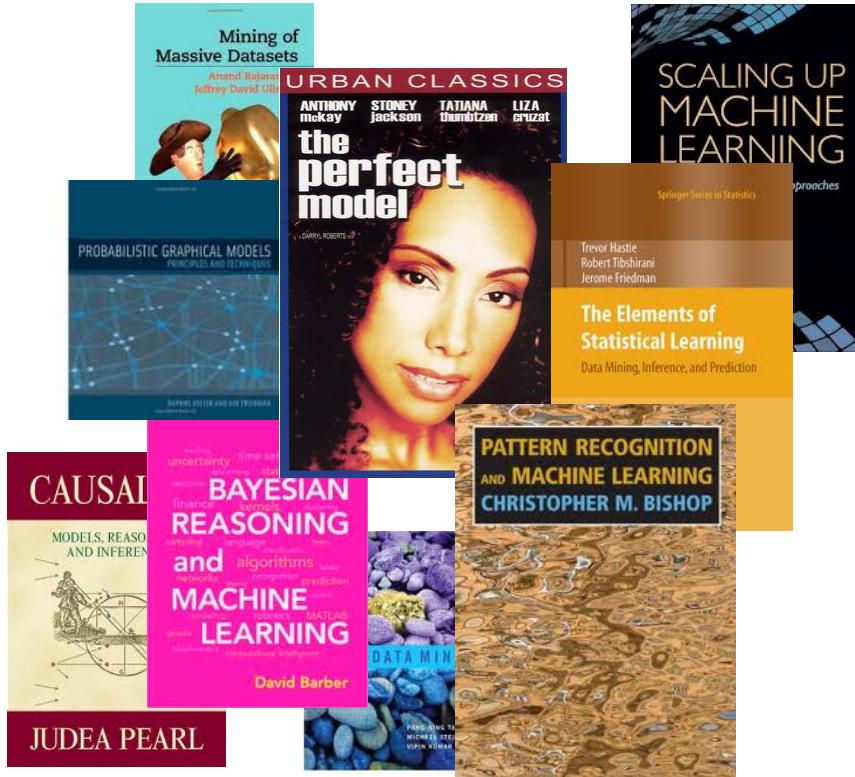
Netflix Scale



- > 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
- ...

A NETFLIX ORIGINAL

RICKY GERVAIS
IS

Derek



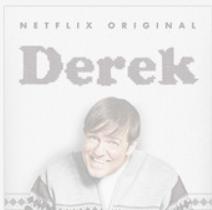
Derek

★★★★★ 2012 TV-14 1 Season

Ricky Gervais created and stars in this heartwarming comedy-drama series as a loyal English nursing home caretaker who sees only the good in everyone. [More info](#)

Recently Watched

My List [See All](#)



“Emmy Winning”

Netflix Algorithms





Could Iron Man's Lab Soon Be A Reality?

Photos Photos of You Your Photos Albums Add Contributors Add Photos Tag

Our Trip to Yellowstone
Contributors: Tom Kammann, Jennifer, Leslie Morris and Mike Morris (posted about 2 weeks ago) Taken at Yellowstone National Park (9)

We decided to go to Yellowstone for the weekend to reconnect with nature. It was a memorable trip with great friends.

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



55 people like this. Be the first of your friends.



Getty

30

12

2

7

107

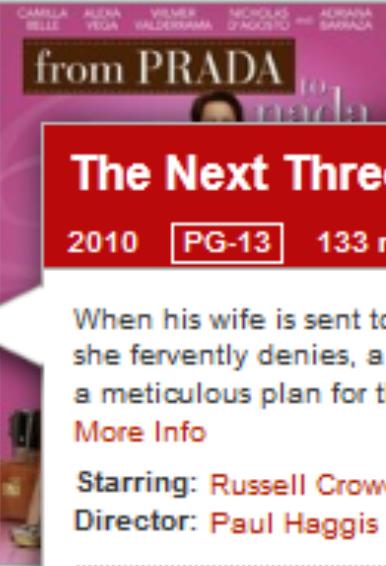


GET TECHNOLOGY NEWSLETTERS:

Enter email

SUBSCRIBE

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:



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

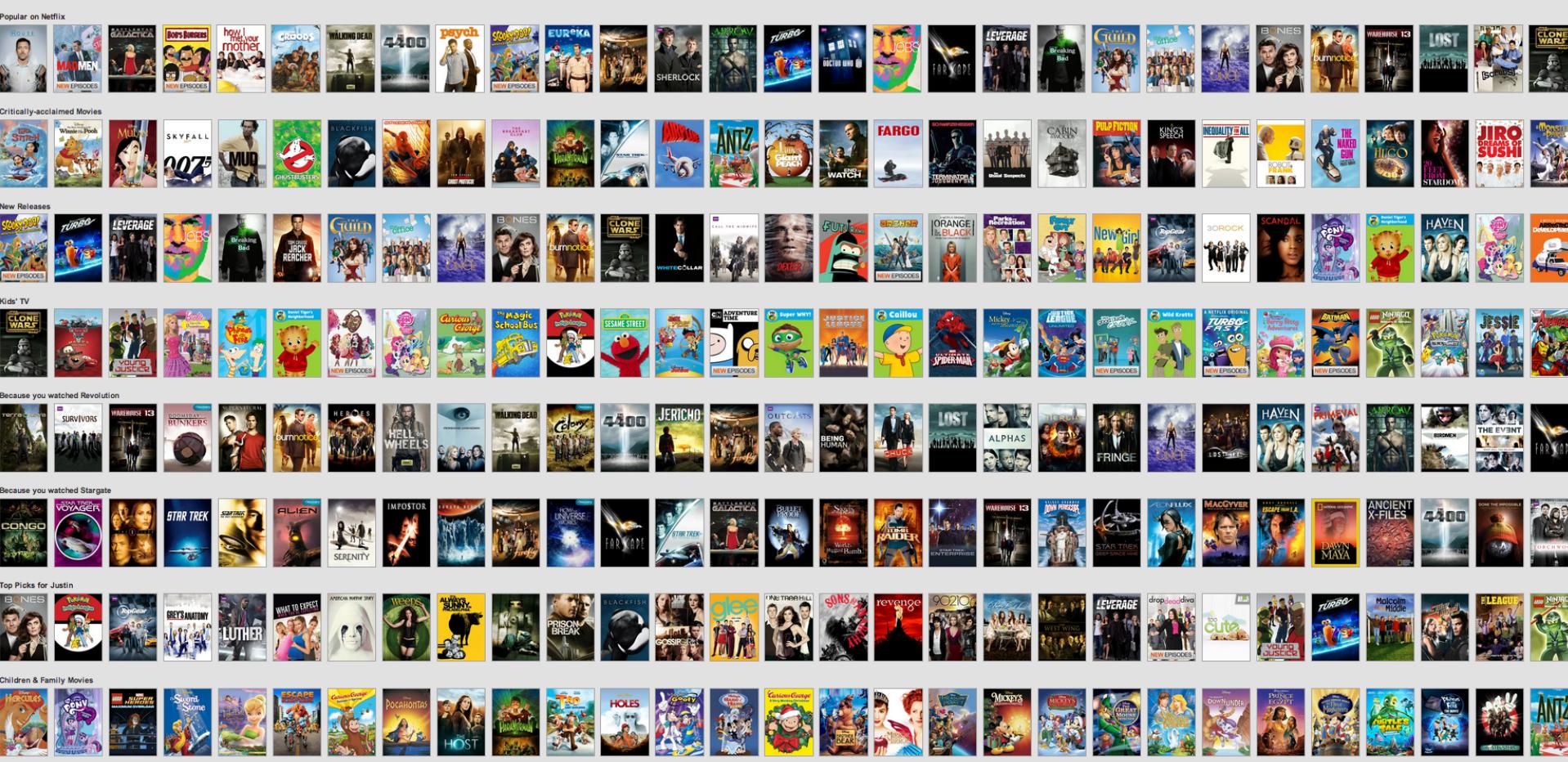
Ranking

The screenshot shows the Netflix homepage with several recommendation sections. At the top, there's a search bar with 'Netflix' and a '+' button. Below it, a large blue arrow points right over the 'Ranking' section. The 'Ranking' section itself has a blue header with the word 'Ranking'. Below this, there are four main categories of recommendations:

- Recently watched:** Shows like *Spaced*, *How I Met Your Mother*, *Frasier*, *Breaking Bad*, *Bones*, *MAD MEN*, *Young Frankenstein*, *Eddie Izzard*, *Tangled*, and *Manhattan Mystery*.
- Popular on Facebook:** Shows like *It's Not Funny*, *Katt Williams*, *Cloudy with a Chance of Meatballs*, *The Confession*, *The Little Engine That Could*, *Go! Go! Go!*, *Kipper*, *S.W.A.T.*, *Live*, and *The Suite Life*.
- Goofy TV Shows:** Shows like *30 Rock*, *Workaholics*, *Community*, *Parks and Recreation*, *South Park*, *Archer*, *The League*, *Todd Margaret*, and *Reno 911!*. A note says 'Your taste preferences created this row.'
- Visually-striking Exciting Foreign Movies:** Shows like *Iron Monkey*, *Shaolin Soccer*, *Fei Fei*, *Good Bad the Weird*, *Dark Lightning*, *Kung Fu Dunks*, *Goemon*, *Exiled*, *Arahan*, and *Execution*. A note says 'Your taste preferences created this row.'

At the bottom left, there's a 'Sci-Fi & Fantasy' section with movies like *Keaton*, *Sherlock Jr.*, *Toy Story 3*, *Serenity*, *Iron Man 2*, *Tron*, *HOT TUB TIME MACHINE*, *Beetlejuice*, *Terminator Genisys*, *Troll Hunter*, and *Labyrinth*. To the left of this section is a sidebar with a list of genres: Imaginative, Exciting, Fantasy, Supernatural, Sci-Fi Thrillers, and Suspenseful. The 'Top Rated' and 'Most Popular' filters are also visible for some of the movie sections.

Page composition



Similarity

NETFLIX Watch Instantly - Just for Kids - Taste Profile - DVDs - DVD Queue

Because you watched Family Guy

Because you watched The Following

Because you added The Way

One Week ONE WEEK 180° South 180° South

Play Play Play Play

Not Interested Not Interested Not Interested Not Interested

National Geographic: Appalachian Trail

I Am I AM APPALACHIAN TRAIL APPALACHIAN TRAIL

Hide Away HIDE AWAY

Seven Days in Utopia SEVEN DAYS IN UTOPIA

Play Play Play Play

Not Interested Not Interested Not Interested Not Interested

The Intouchables

Albert Nobbs ALBERT NOBBS

Play Play Play Play

Not Interested Not Interested Not Interested Not Interested

the office SPACED THE LEAGUE THE IT CROWD TODD MARG

Because you watched Derek

NETFLIX

Verizon 10:24 PM 93%

Search Megadeth: That On...

Filmed during one of the group's most creative periods, this 2005 show sees a return to form of founder Dave Mustaine, who'd been sidelined by injury.

Cast: Megadeth

Similar titles to watch instantly:

METALLICA: PHANTOM PUPPETS 2006 NR 1h 30min

★★★★★

GIGANTOUR 2005 R 1h 25min

★★★★★

Home Genres Search Instant Queue

Search Recommendations

MOVIES & TV SHOWS

Marriage Italiano

Smalltown, Italy

Perlasca

PEOPLE

Italia Almirante

Italo Renda

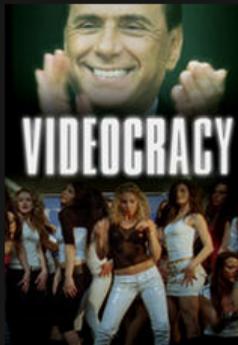
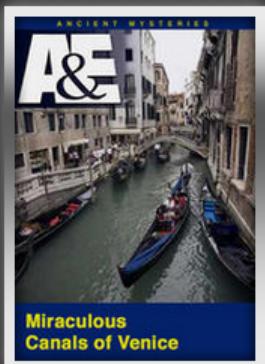
EXPLORE TITLES F

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

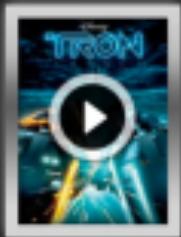


Back to browse Search

Rate The Fighter



Suggestions for you
to watch now...

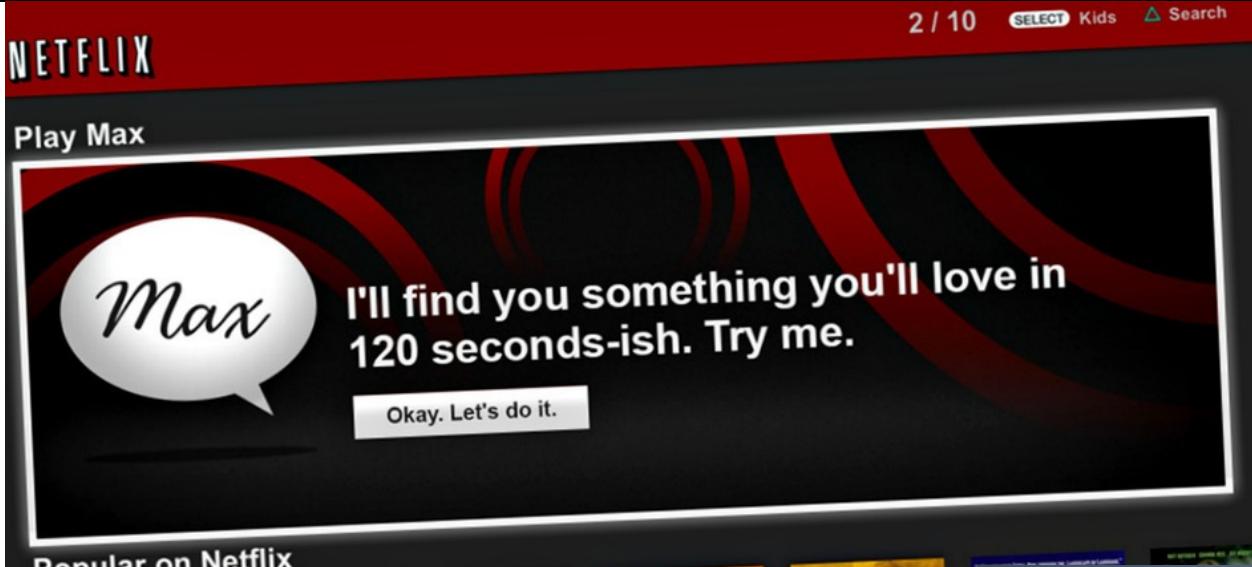


Tron: Legacy

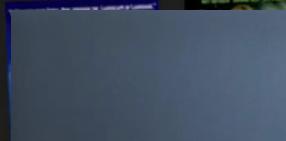
PG 2h 05m

While investigating the mysterious disappearance of his father, Kevin, techie Sam Flynn lands in a beguiling computerized world of enslaved...

Gamification



Popular on Netflix



NETFLIX



Dysfunctional Families or Space Travel



Max's Mystery Call

BROWSE CATALOG

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, using both published algorithms and novel ideas. Of those, the results from two of

A B O U T

[Anand Rajaraman](#)

[Datawocky](#)

Really?

R E C E N T P O S

[Goodbye, Kosmix. H
@WalmartLabs](#)

[Retail + Social + Mo
@WalmartLabs](#)

[Creating a Culture of
Innovation: Why 20%](#)
[not Enough](#)

[Reboot: How to Rein
Technology Startup](#)

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

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

Sometimes, it's not
about more data

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.

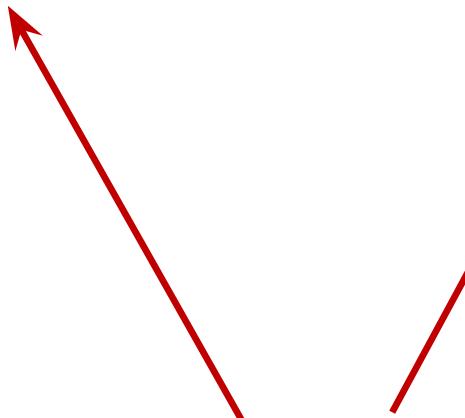
The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which



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]

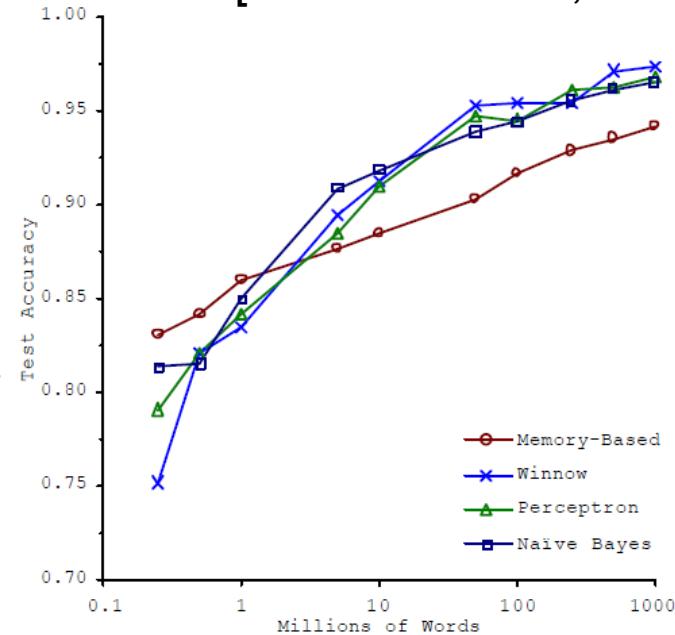
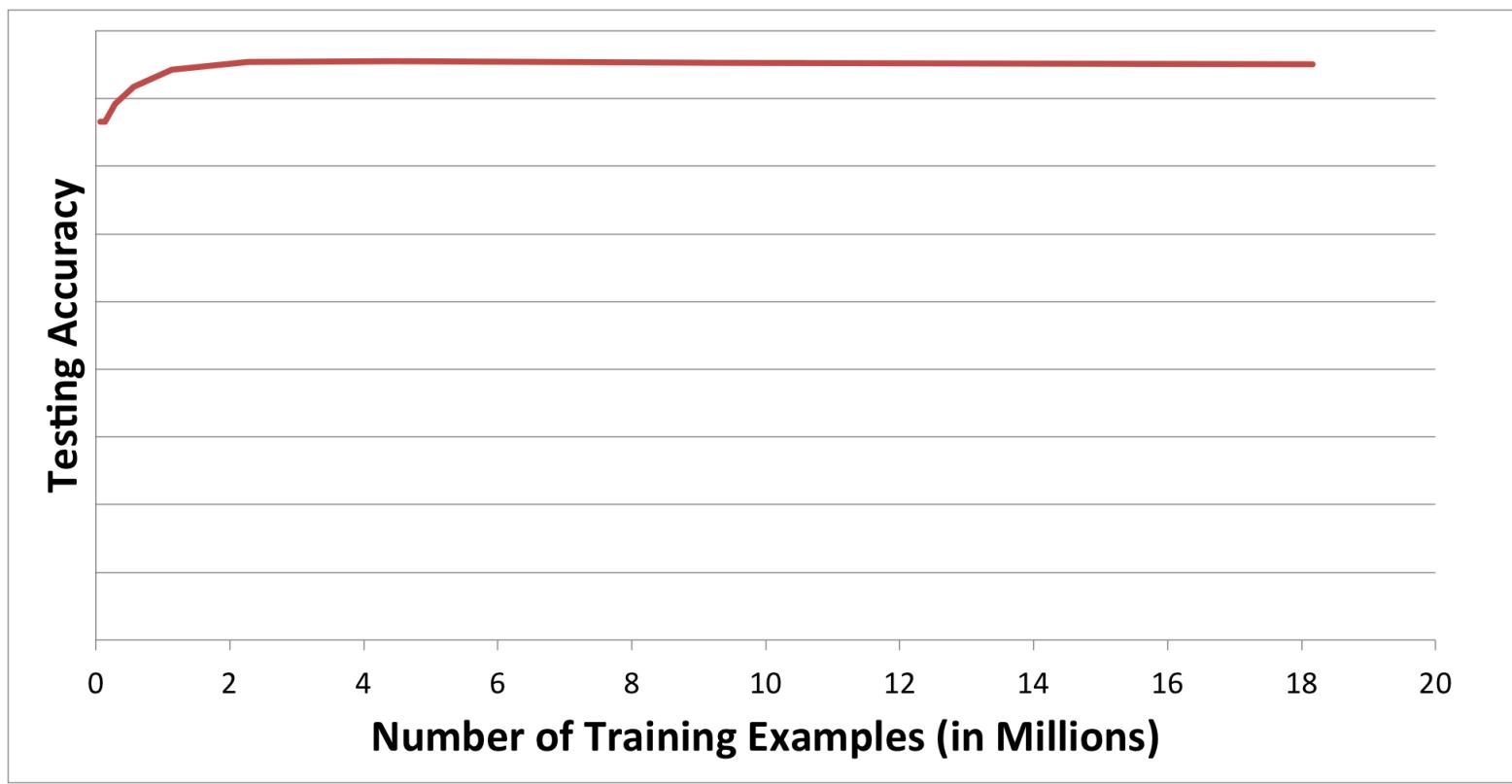


Figure 1. Learning Curves for Confusion Set Disambiguation

Sometimes, it's not
about more data





At what level should I parallelize?

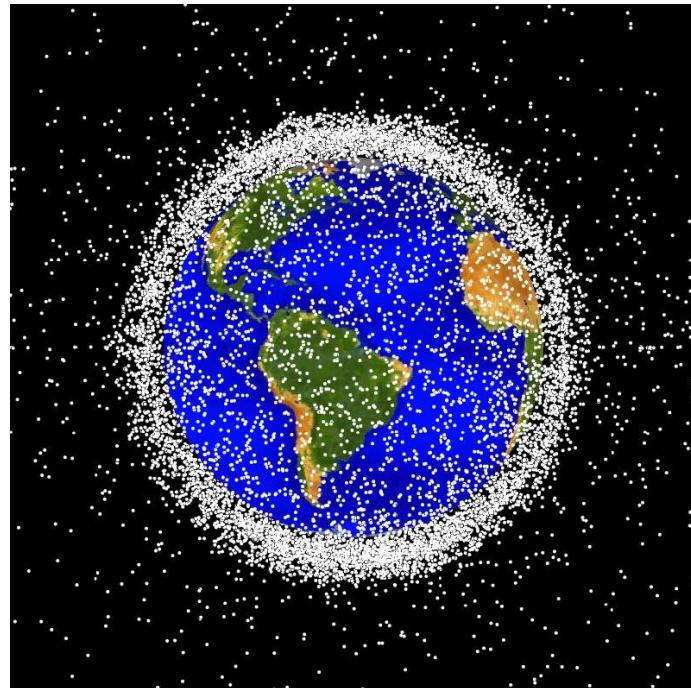
The three levels of Distribution/Parallelization

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

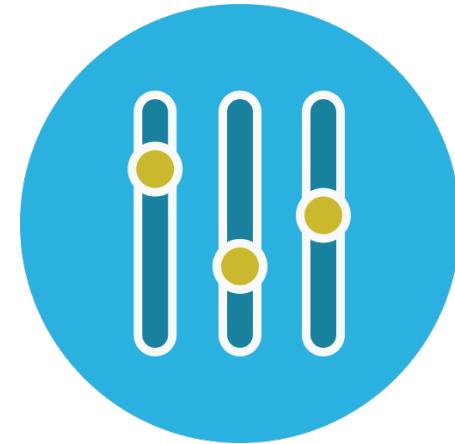
Level 1 Distribution

- 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



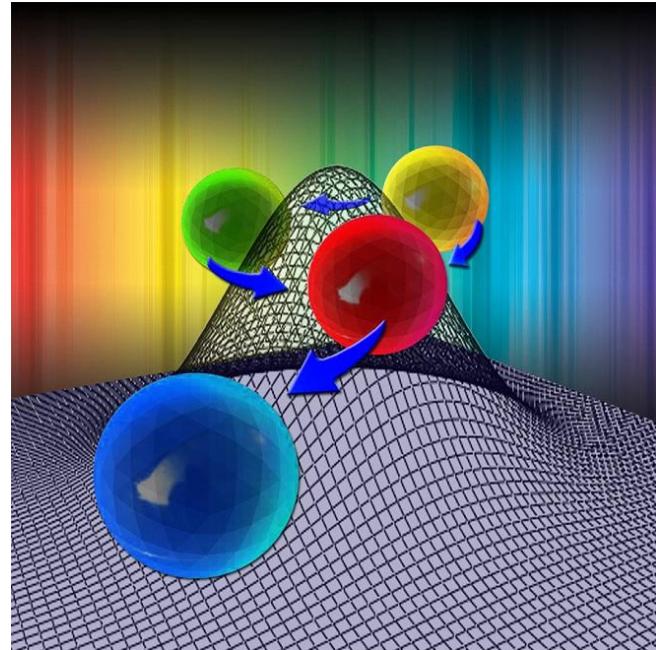
Level 2 Distribution

- 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

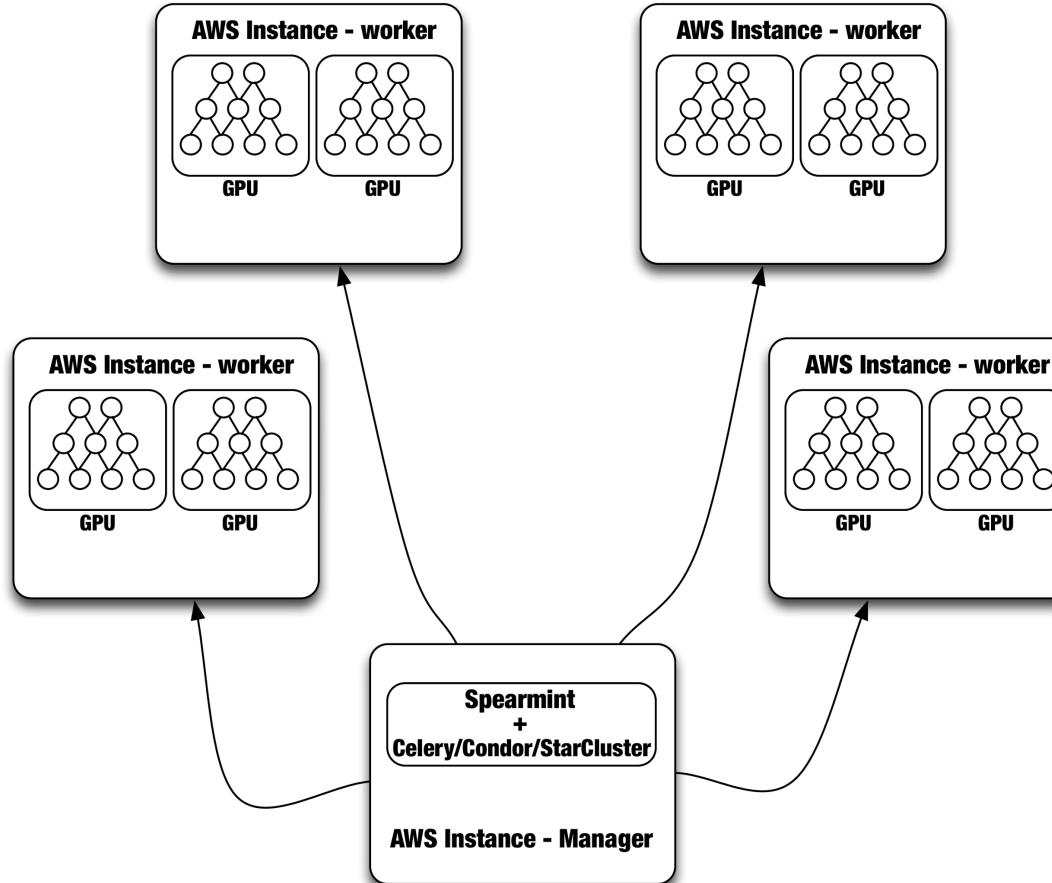


Level 3 Distribution

- 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



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

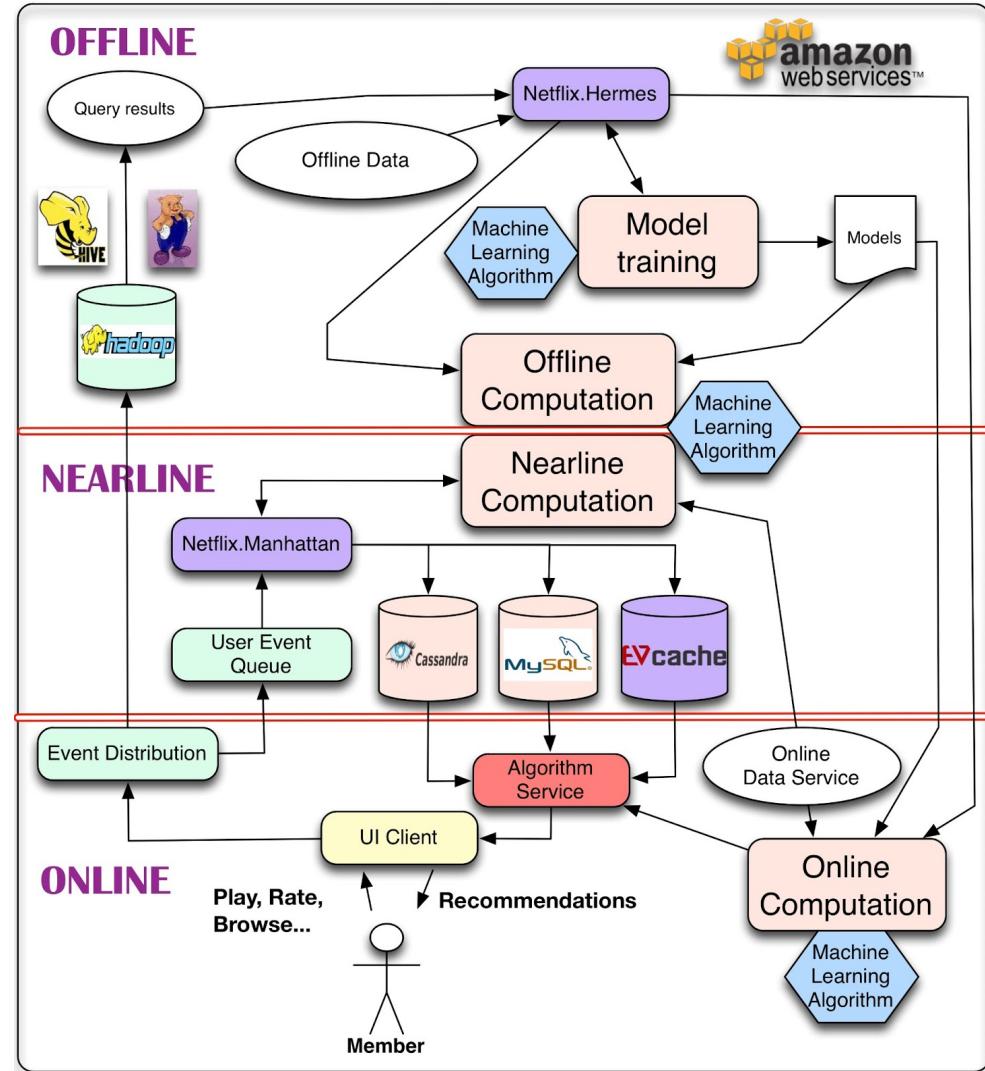
LILYHAMMER



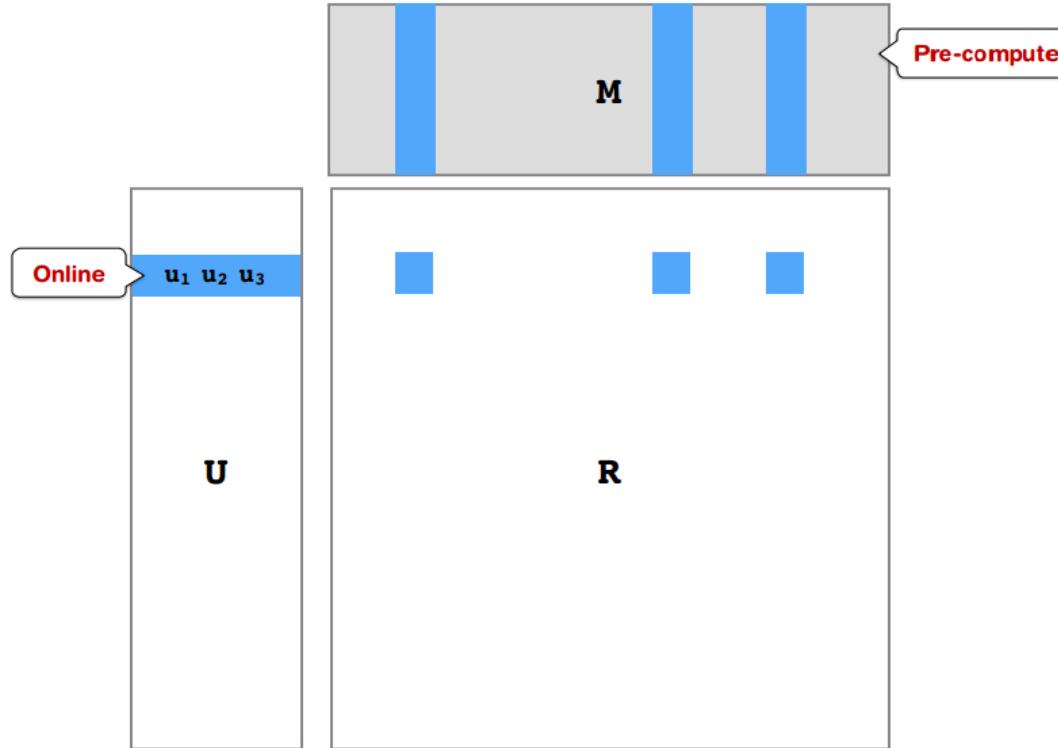
What latency can I afford?

3 shades of latency

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



Matrix Factorization Example



Xavier Amatriain (@xamat)
xavier@netflix.com



Thanks!
(and yes, we are hiring)