MLconf

The Machine Learning Conference

10 Lessons Learned from building ML systems

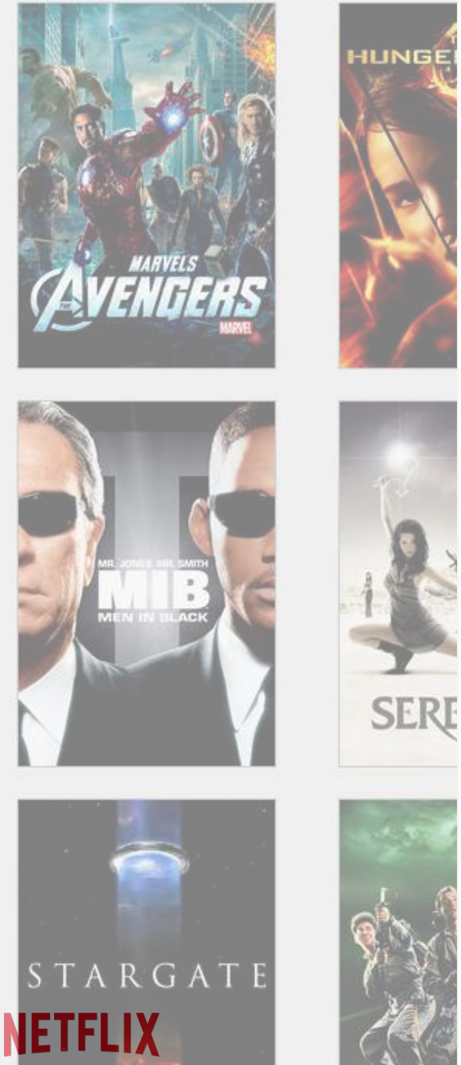
Xavier Amatriain - Director Algorithms Engineering

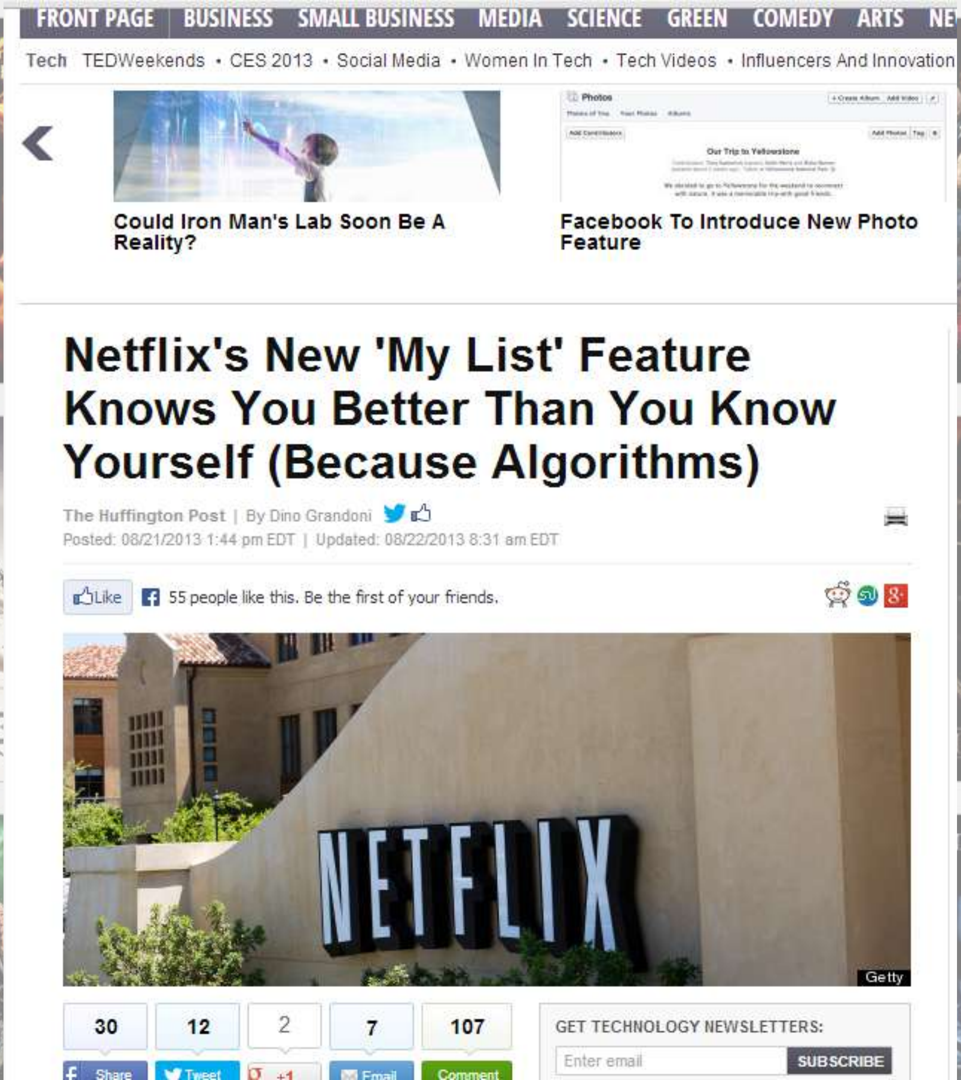
November 2014

NETFLIX

Machine Learning @ Netflix

NETFLIX





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

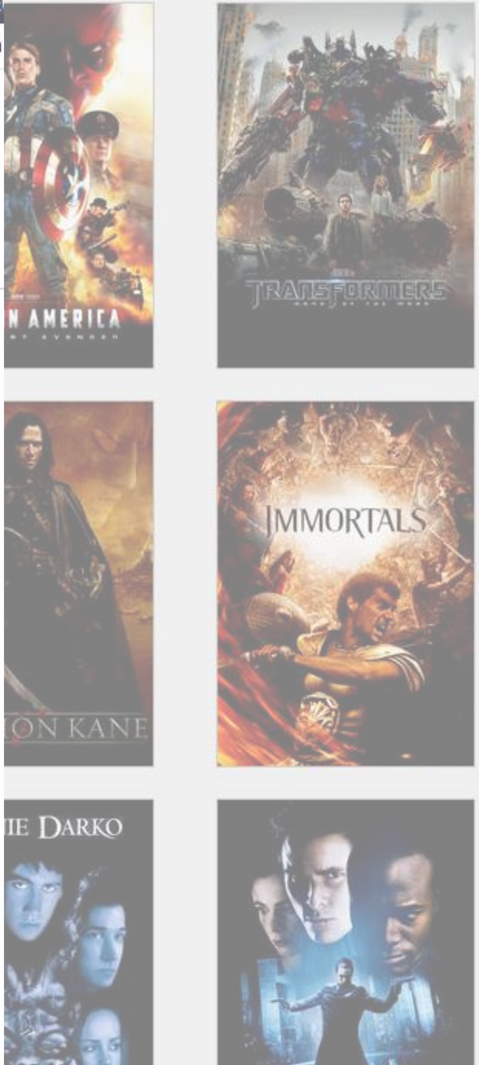
Like 55 people like this. Be the first of your friends.

Getty

30	12	2	7	107
Share	Tweet	+1	Email	Comment

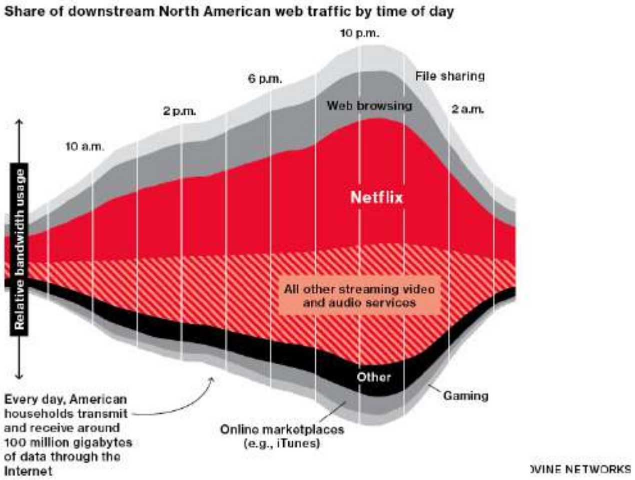
GET TECHNOLOGY NEWSLETTERS:

Enter email SUBSCRIBE



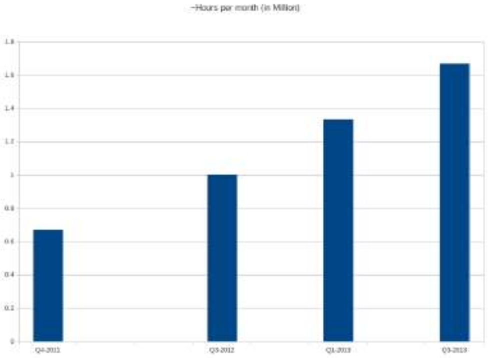
NETFLIX

Netflix Scale



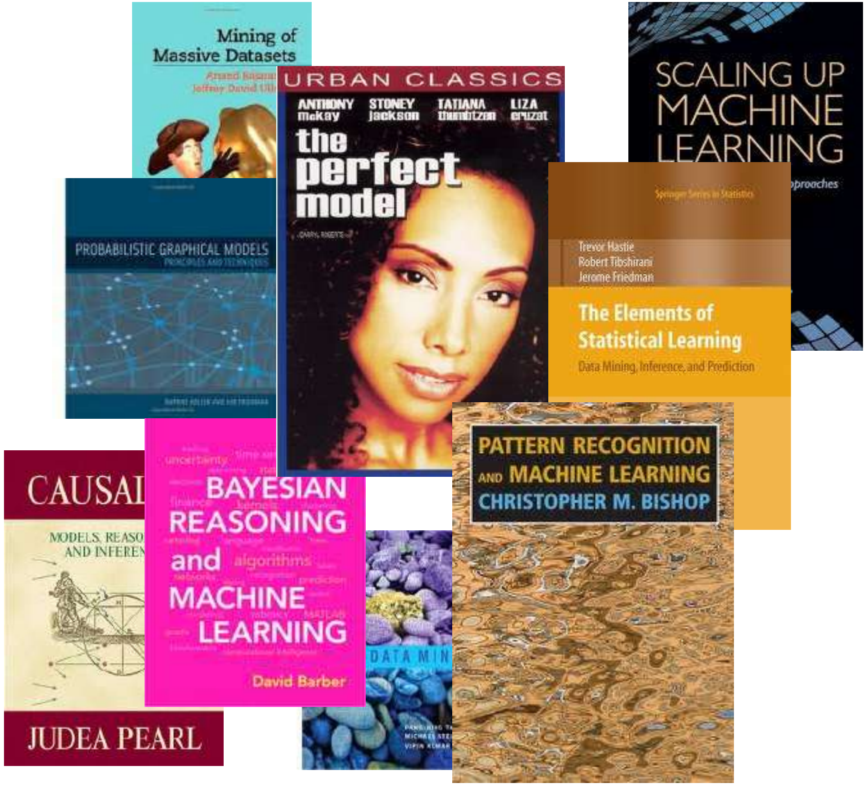
Every day, American households transmit and receive around 100 million gigabytes of data through the Internet

)VINE NETWORKS



- > 50M members
- > 40 countries
- > 1000 device types
- > 7B hours in Q2 2014
- Plays: > 70M/day
- Searches: > 4M/day
- Ratings: > 6M/day
- Log 100B events/day
- 31.62% of peak US downstream traffic

Smart Models


Mining of Massive Datasets
URBAN CLASSICS
the perfect model
SCALING UP MACHINE LEARNING
PROBABILISTIC GRAPHICAL MODELS
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
CAUSAL
MODELS, REASO
AND INFEREN
BAYESIAN REASONING and MACHINE LEARNING
David Barber
DATA MIN
PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
JUDEA PEARL

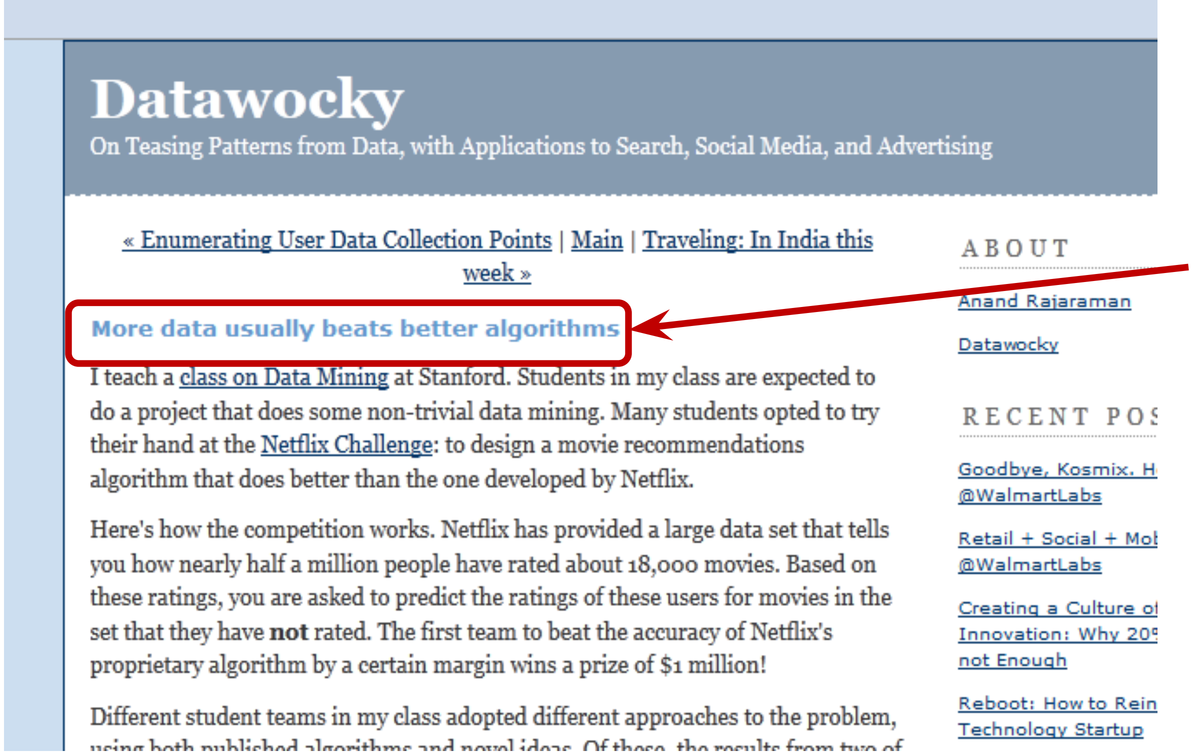
- Regression models (Logistic, Linear, Elastic nets)
- GBDT/RF
- SVD & other MF models
- Factorization Machines
- Restricted Boltzmann Machines
- Markov Chains & other graphical models
- Clustering (from k-means to HDP)
- Deep ANN
- LDA
- Association Rules
- …

10 Lessons

NETFLIX

1. More data ~~vs.~~ & Better Models

More data or better models?



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 these, the results from two of

ABOUT

Anand Rajaraman

Datawocky

RECENT POS

Goodbye, Kosmix. H @WalmartLabs

Retail + Social + Mol @WalmartLabs

Creating a Culture of Innovation: Why 20% not Enough

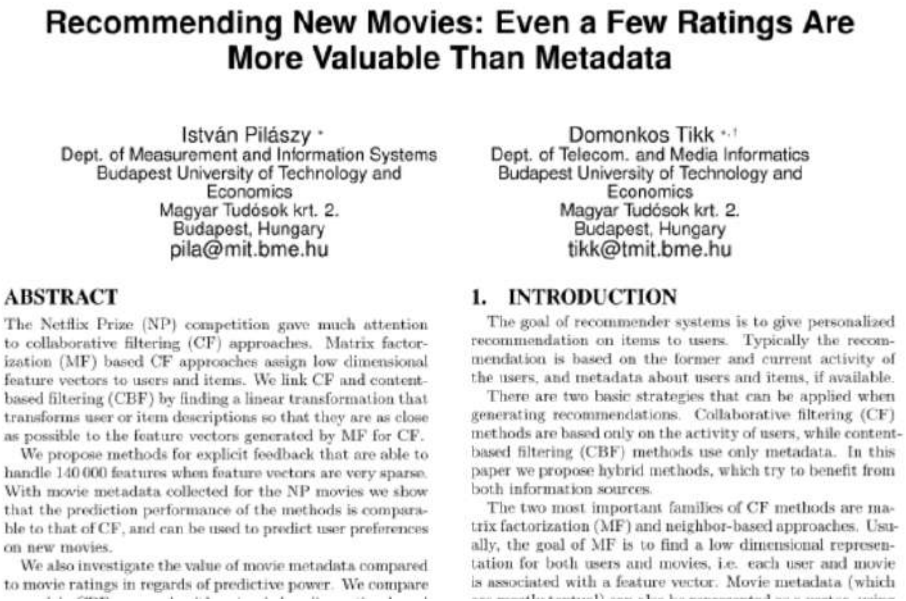
Reboot: How to Rein Technology Startup

Really?

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

More data or better models?

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

More data or better models?

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

The Unreasonable Effectiveness of Data

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

[Banko and Brill, 2001]

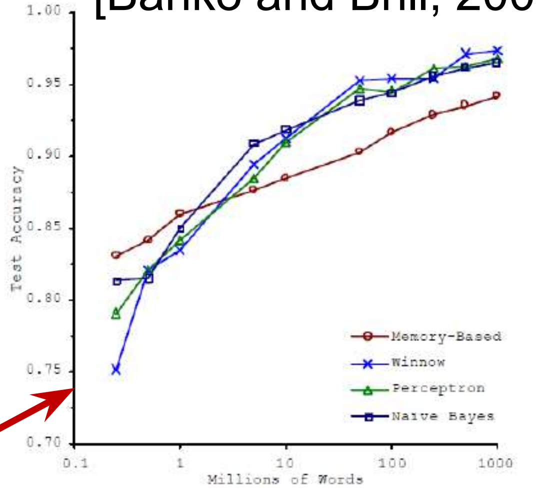
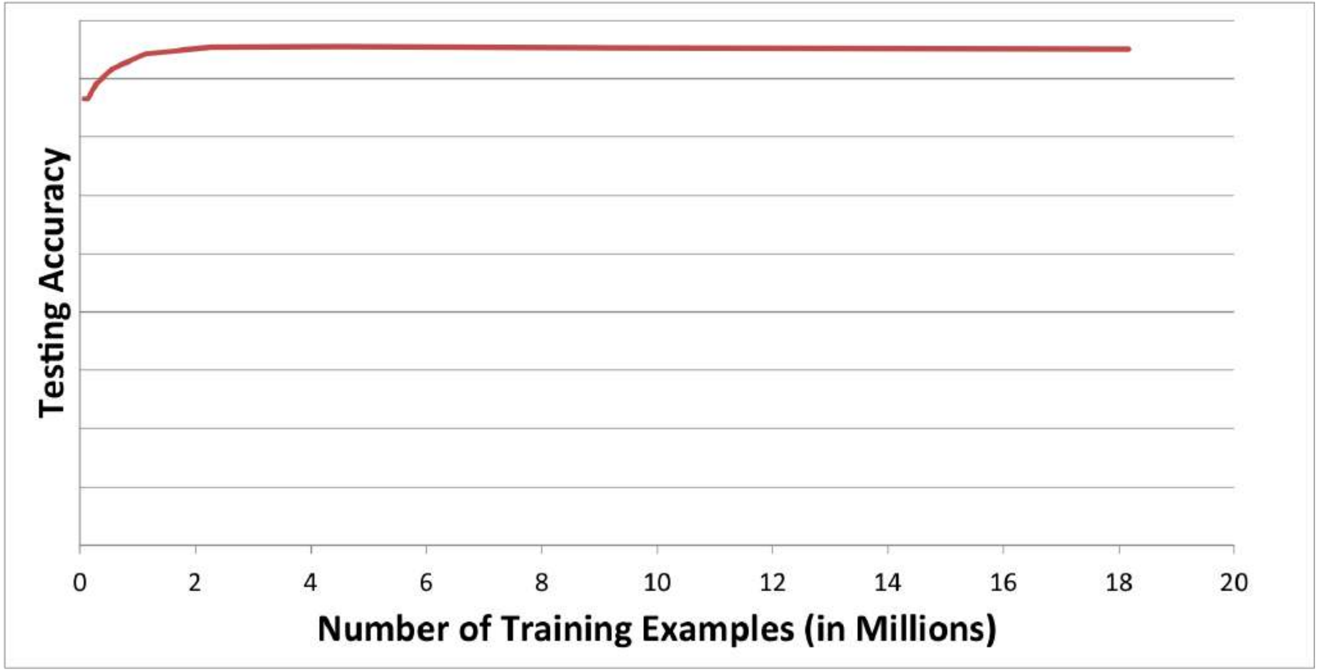


Figure 1. Learning Curves for Confusion Set Disambiguation

Many features/ low-bias models

More data or better models?



Sometimes, it's not about more data

2. You might not need all your Big Data

NETFLIX

How useful is Big Data

- “Everybody” has Big Data
  - But not everybody needs it
  - E.g. Do you need many millions of users if the goal is to compute a MF of, say, 100 factors?
- Many times, doing some kind of smart (e.g. stratified) sampling can produce as good or even better results as using it all

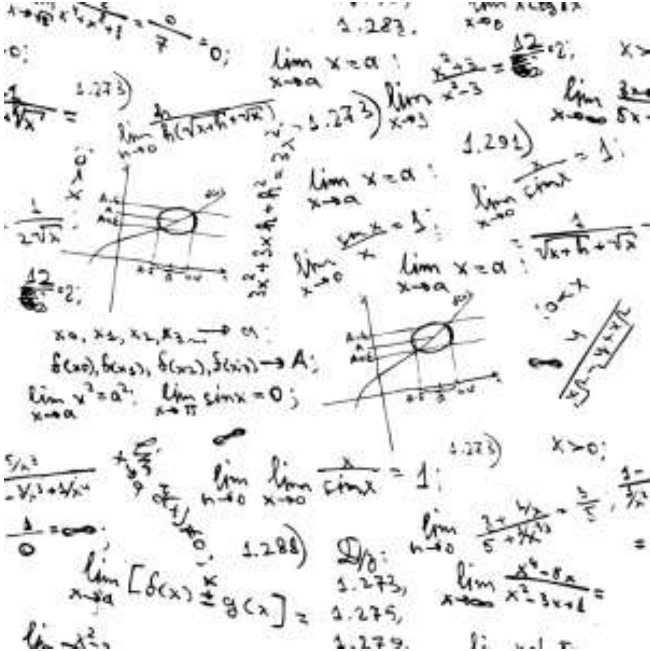
3. The fact that a more complex model does not improve things does not mean you don't need one

Better Models and features that “don’t work”

- Imagine the following scenario:
  - You have a linear model and for some time you have been selecting and optimizing features for that model
    - If you try a more complex (e.g. non-linear) model with the same features you are not likely to see any improvement
    - If you try to add more expressive features, the existing model is likely not to capture them and you are not likely to see any improvement

Better Models and features that “don’t work”

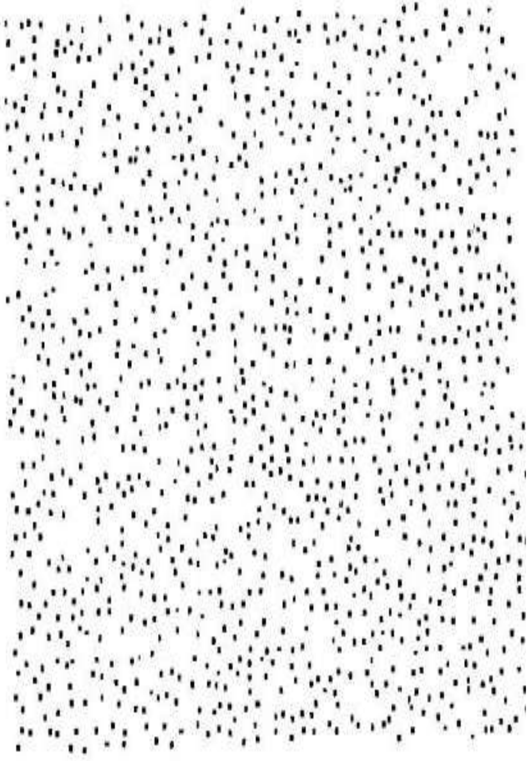
- More complex features may require a more complex model
- A more complex model may not show improvements with a feature set that is too simple



4. Be thoughtful about your training data

Defining training/testing data

- Imagine you are training a simple binary classifier
  - Defining positive and negative labels -> Non-trivial task
  - E.g. Is this a positive or a negative?
    - User watches a movie to completion and rates it 1 star
    - User watches the same movie again (maybe because she can’t find anything else)
    - User abandons movie after 5 minutes, or 15 minutes… or 1 hour
    - User abandons TV show after 2 episodes, or 10 episode… or 1 season
    - User adds something to her list but never watches it



Other training data issues: Time traveling

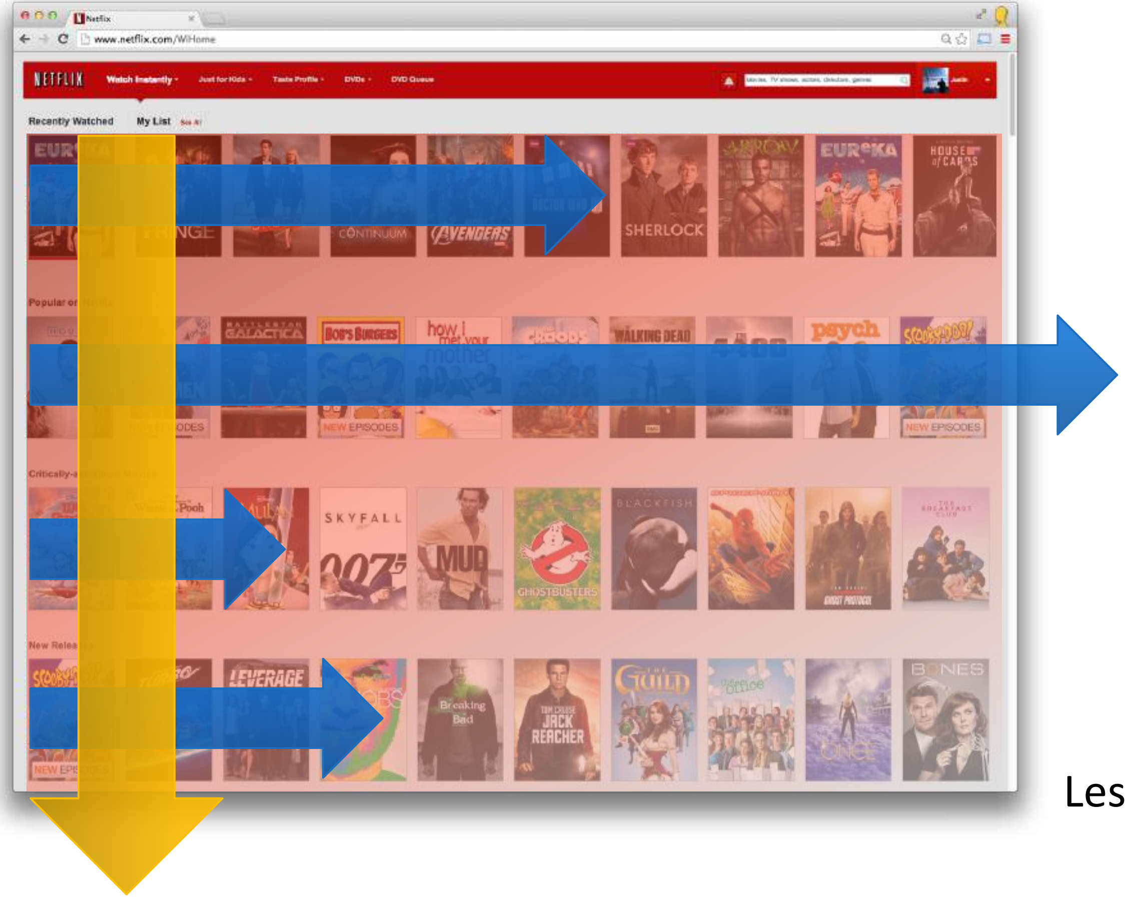
- Time traveling: usage of features that originated after the event you are trying to predict
  - E.g. *Your rating a movie is a pretty good prediction of you watching that movie, especially because most ratings happen AFTER you watch the movie*
  - It can get tricky when you have many features that relate to each other
  - Whenever we see an offline experiment with huge wins, the first question we ask ourselves is: “Is there time traveling?”

5. Learn to deal with (The curse of) Presentation Bias

NETFLIX

2D Navigational Modeling

More likely
to see



Less likely

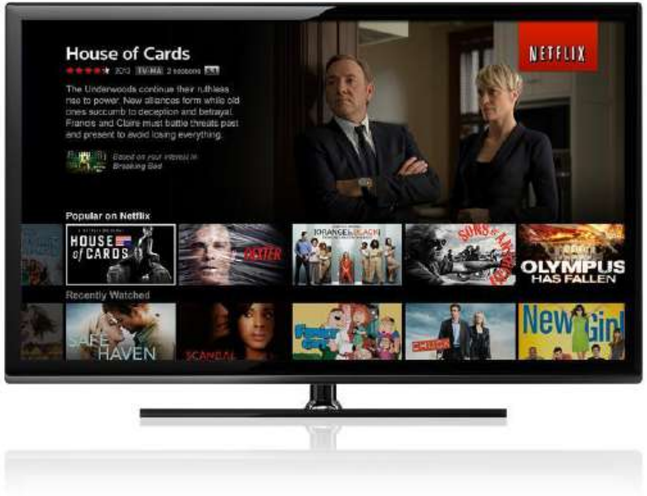
The curse of presentation bias

- User can only click on what you decide to show
  - But, what you decide to show is the result of what your model predicted is good
- Simply treating things you show as negatives is not likely to work
- Better options
  - Correcting for the probability a user will click on a position -> Attention models
  - Explore/exploit approaches such as MAB

6. The UI is the algorithm's only communication channel with that which matters most: the users

UI->Algorithm->UI

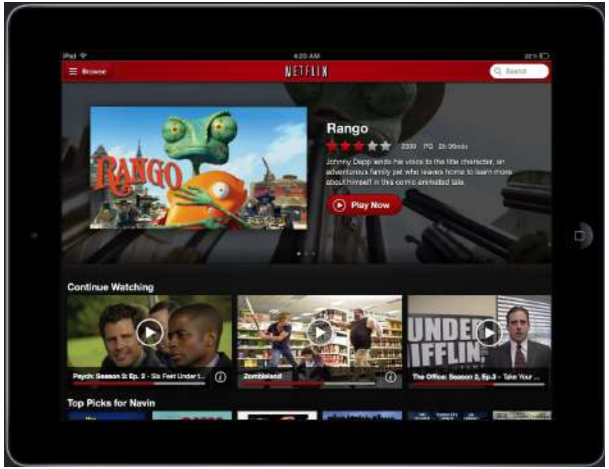
- The UI generates the user feedback that we will input into the algorithms
- The UI is also where the results of our algorithms will be shown
- A change in the UI might require a change in algorithms and viceversa



House of Cards

The Underwoods continue their ruthless rise to power. New alliances form while old ones succumb to deception and betrayal. Francis and Claire must battle threats past and present to avoid losing everything.

Based on your interest in Breaking Bad



Rango

2009 PG

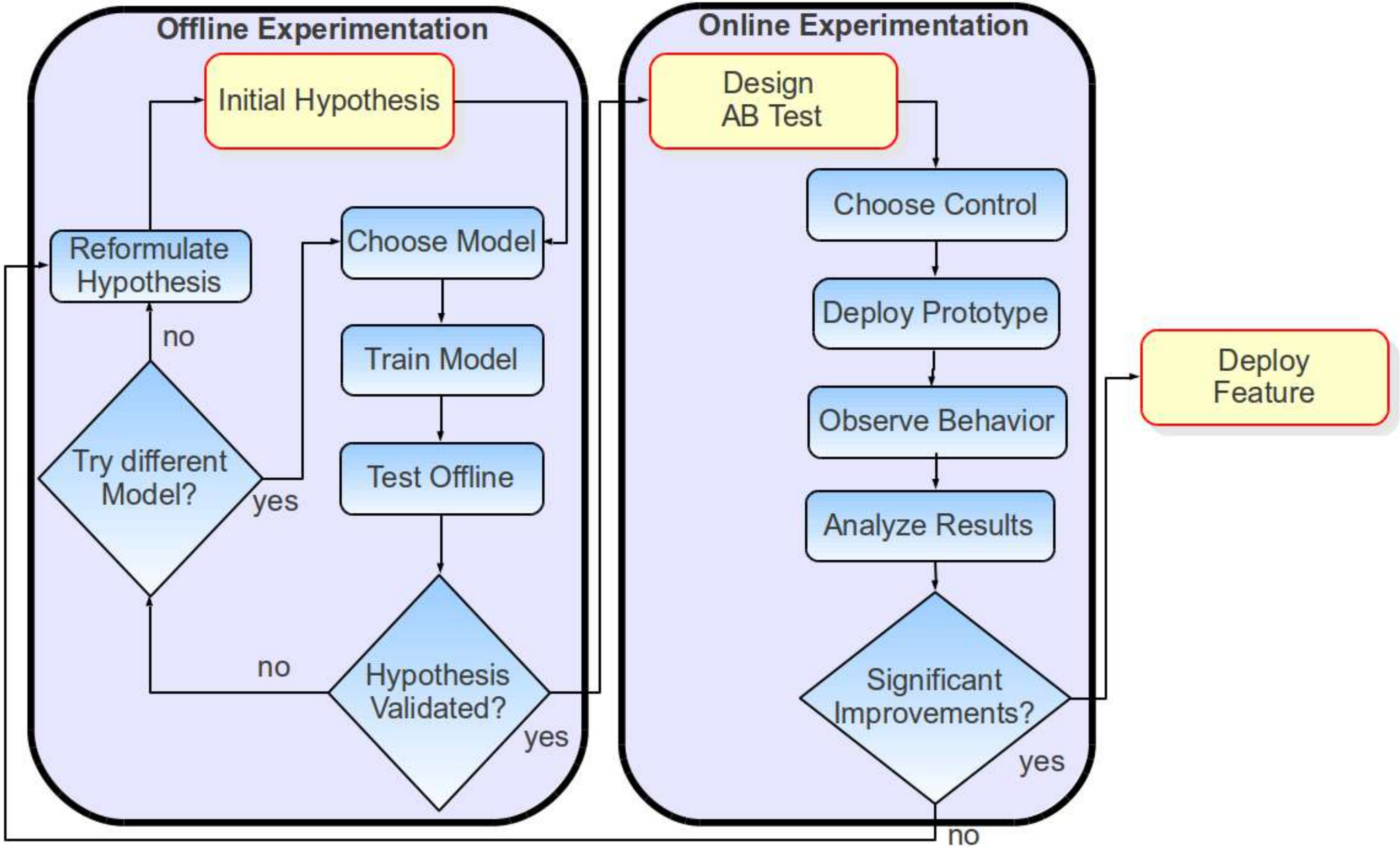
Johnny Depp lends his voice to the title character, an adventurous family pet who leaves home to learn more about himself in this comic animated tale.

iPad
Browse
NETFLIX
Search
RANGO
Play Now
Continue Watching
Psych: Season 5: Ep. 3 - Six Feet Under t...
Zombieland
The Office: Season 5, Ep. 3 - Take Your ...
Top Picks for Nevin

NETFLIX

7. Data and Models are great. You know what's even better? The right evaluation approach

Offline/Online testing process

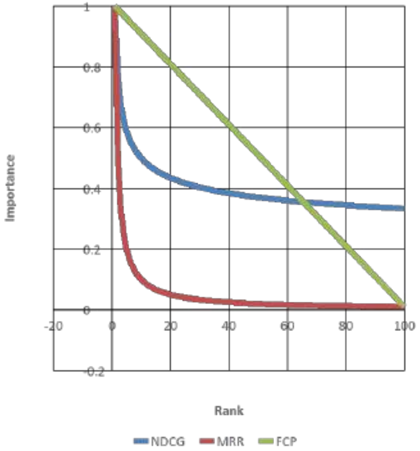


Executing A/B tests

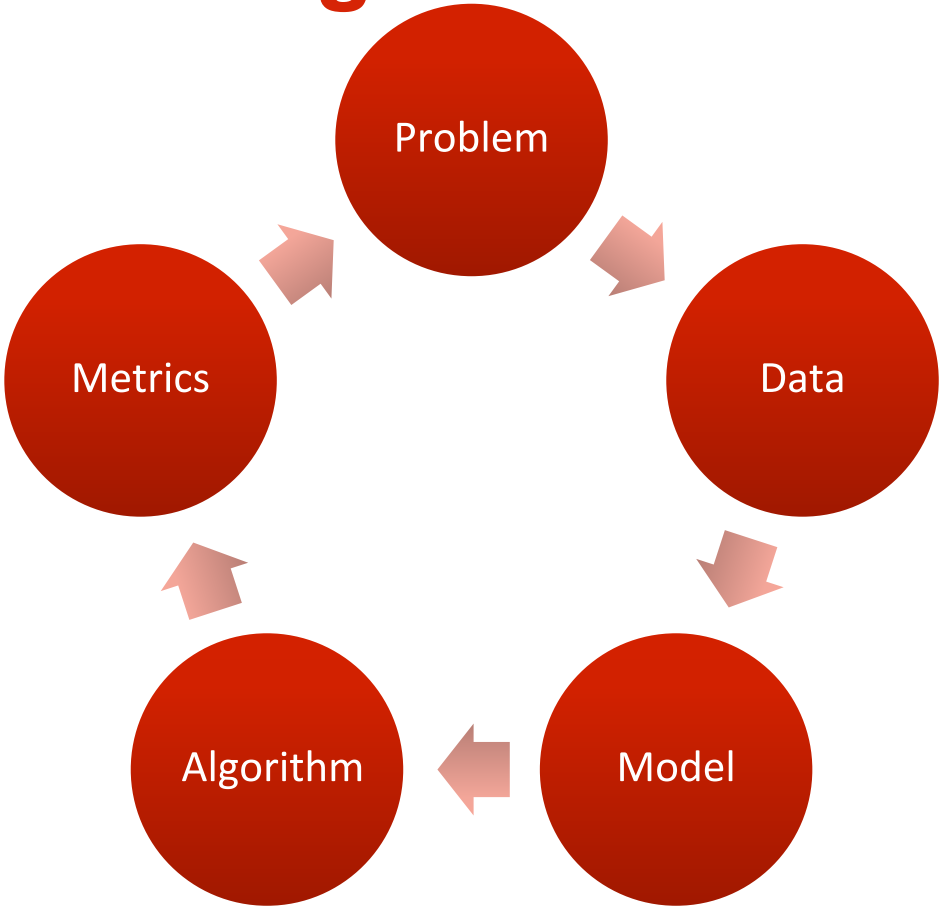
Measure *differences* in metrics across *statistically identical* populations that each experience a different algorithm.

- Decisions on the product always data-driven
- Overall Evaluation Criteria (OEC) = member retention
  - Use long-term metrics whenever possible
  - Short-term metrics can be informative and allow faster decisions
    - But, not always aligned with OEC

Offline testing



- Measure model performance, using (IR) metrics
- Offline performance used as an indication to make informed decisions on follow-up A/B tests
- A critical (and mostly unsolved) issue is how offline metrics can correlate with A/B test results.



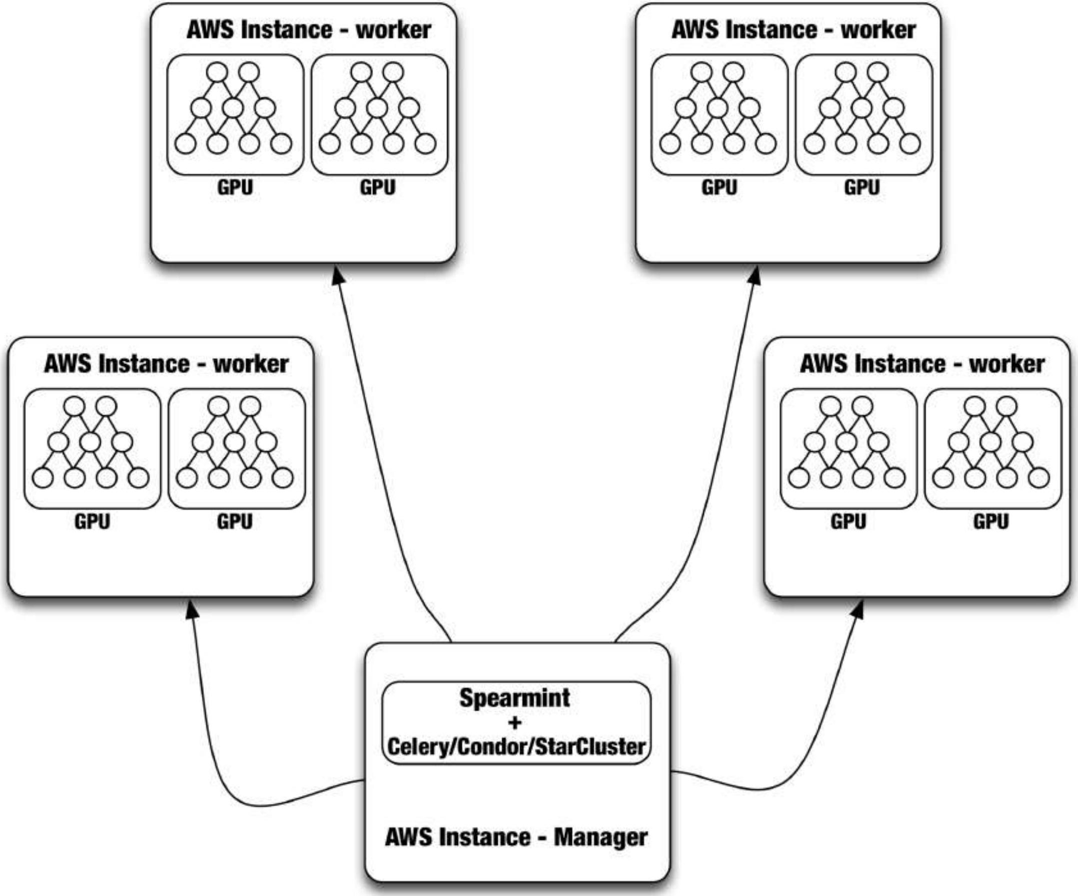
8. Distributing algorithms is important, but knowing at what level to do it is even more important

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

ANN Training over GPUS and AWS



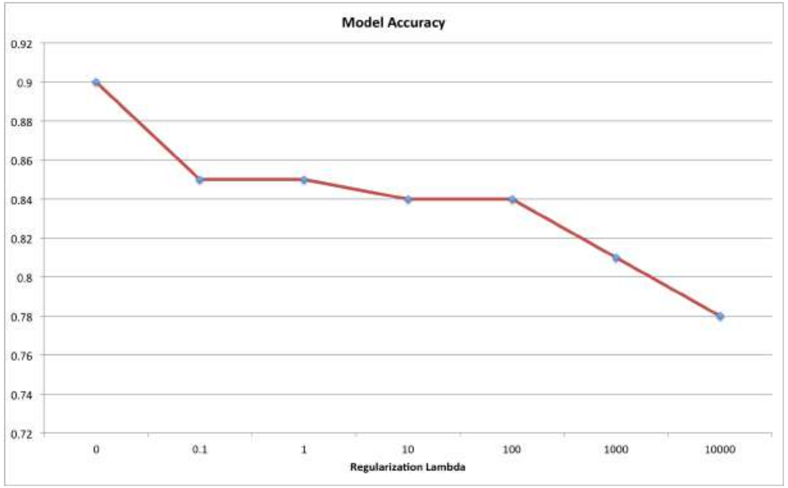
NETFLIX

9. It pays off to be smart about choosing your hyperparameters

NETFLIX

Hyperparameter optimization

- Automate hyperparameter optimization by choosing the right metric.
  - But, is it enough to choose the value that maximizes the metric?
  - E.g. Is a regularization lambda of 0 better than a lambda = 1000 that decreases your metric by only 1%?
- Also, think about using Bayesian Optimization (Gaussian Processes) instead of grid search

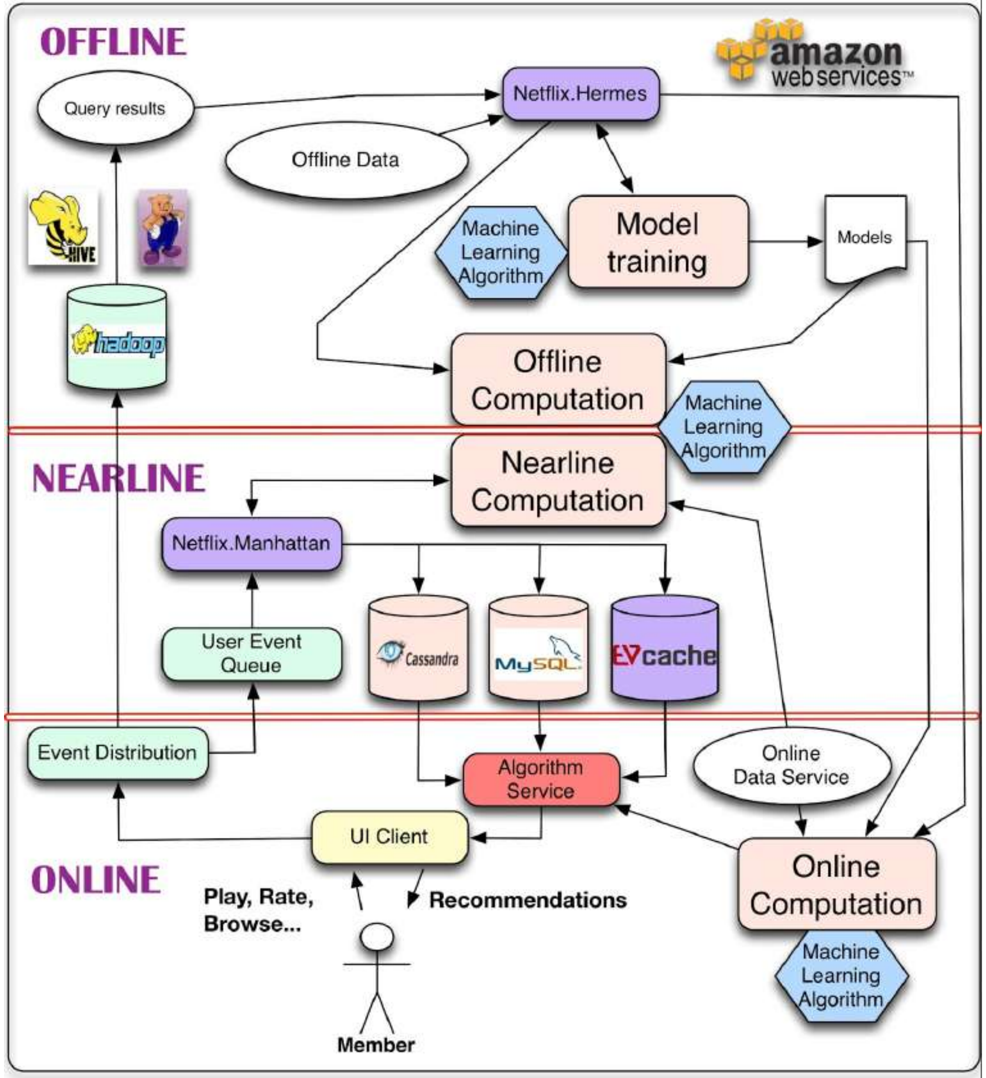


10. There are things you can do offline and there are things you can't… and there is nearline for everything in between

NETFLIX

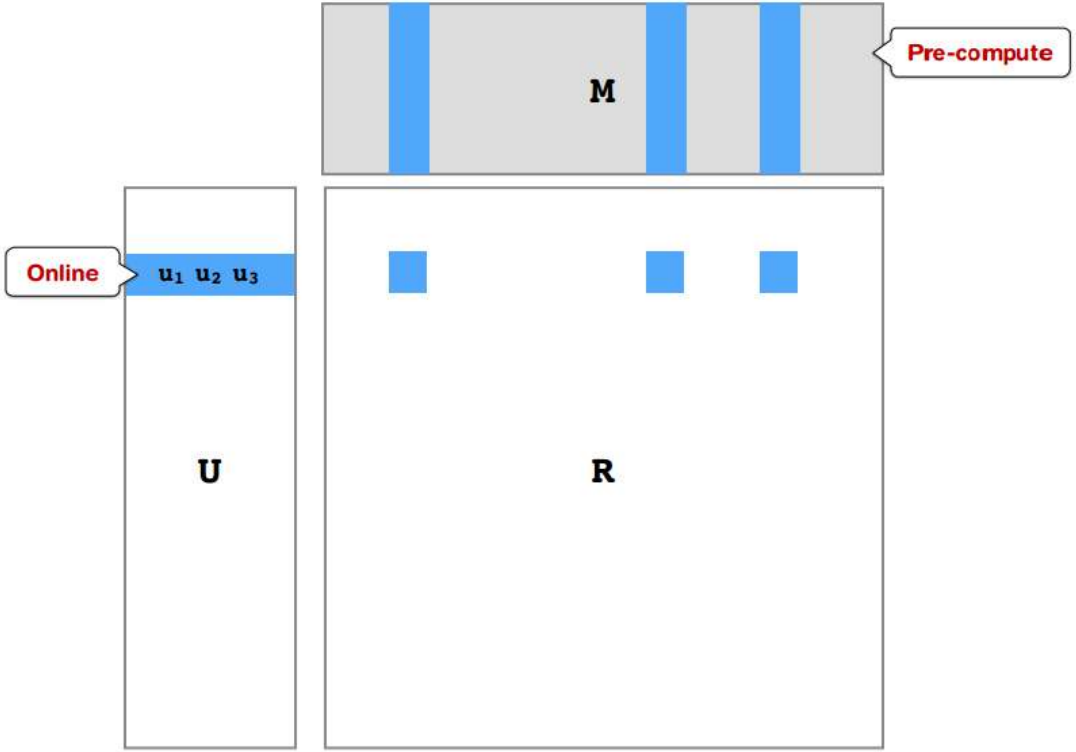
System Overview

- Blueprint for multiple personalization algorithm services
  - Ranking
  - Row selection
  - Ratings
  - …
- Recommendation involving multi-layered Machine Learning



NETFLIX

Matrix Factorization Example



NETFLIX

Conclusions

1. Choose the right metric
2. Be thoughtful about your data
3. Understand dependencies between data and models
4. Optimize only what matters

NETFLIX

Xavier Amatriain (@xamat)
xavier@netflix.com

NETFLIX

Thanks!

(and yes, we are hiring)