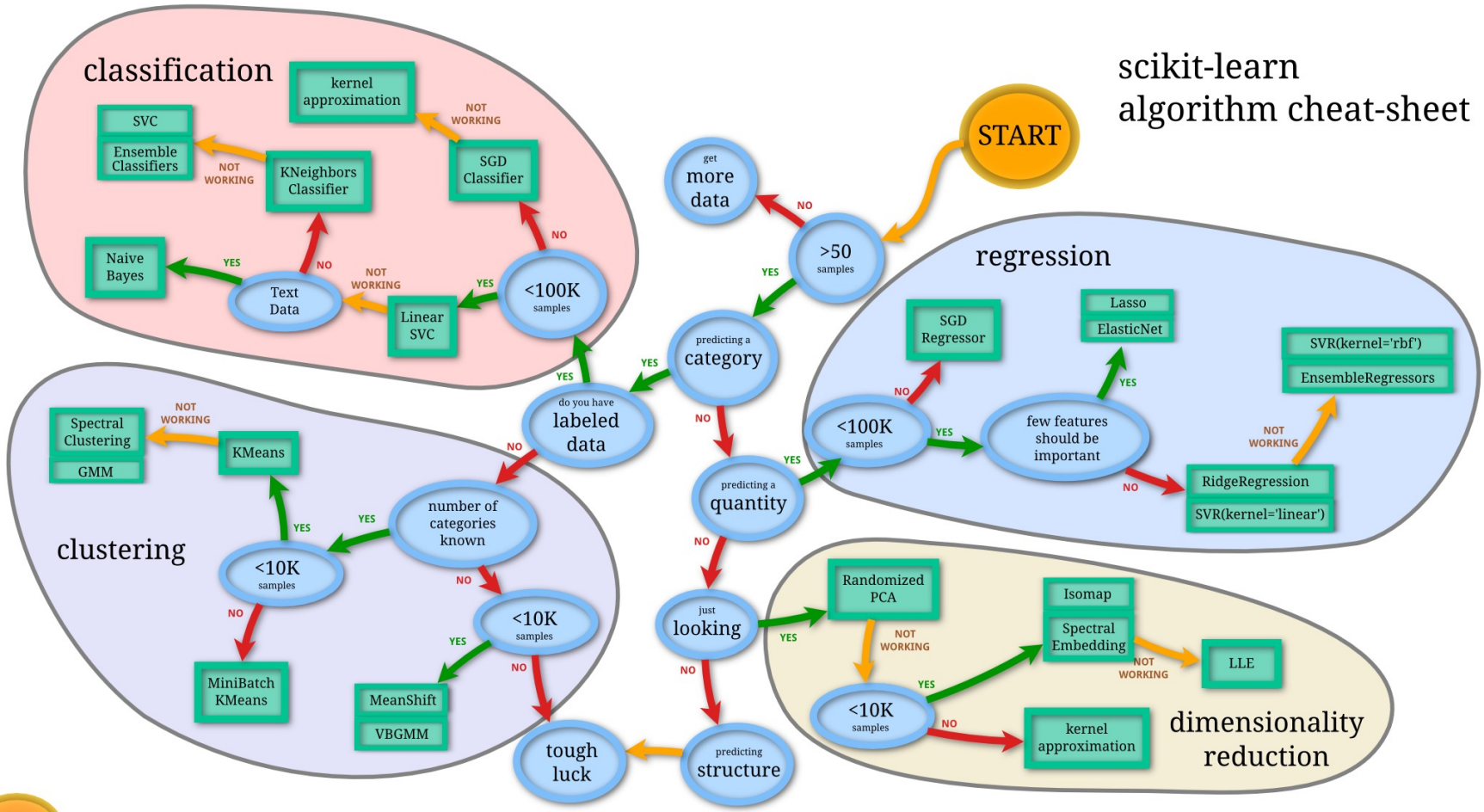
# Model selection and experimentation

SOEN 691: Engineering AI-based Software Systems

Emad Shihab, Diego Elias Costa
Concordia University

# How to select the best model?



# Certain problems can be best solved by certain models


Convolutional Neural Networks


Naïve Bayes
Embeddings + Traditional Classifiers
Transformers

# Sometimes the model is selected based on constraints


Smaller models with fast inference time

The sky is the limit

- Ensemble models
- Classical models
- Deep Learning

# Models' trade-offs

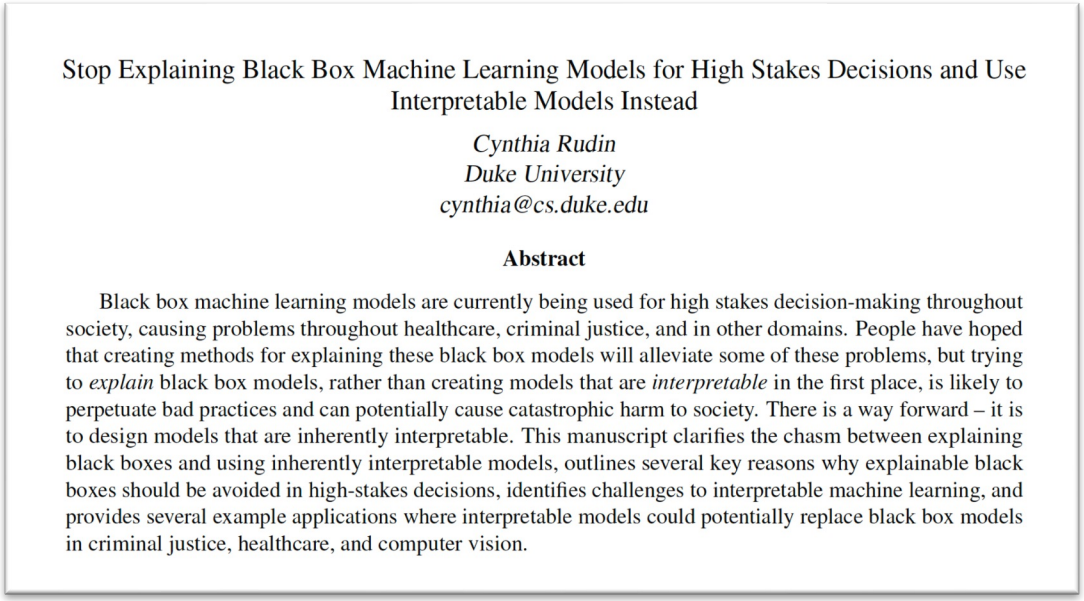
- Models may have different trade-offs
  - Interpretability
  - Robustness
  - Model size
  - Training time
  - Inference time

- Our goal is not to give an overview of all models
- But to showcase a debate in research and industry of solutions and challenges of model selection

# Use Interpretable Models in High Stakes Decisions

Paper arguments that:

- Explainable models should not be used in high stake decisions
- Instead, we should use Interpretable models



Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

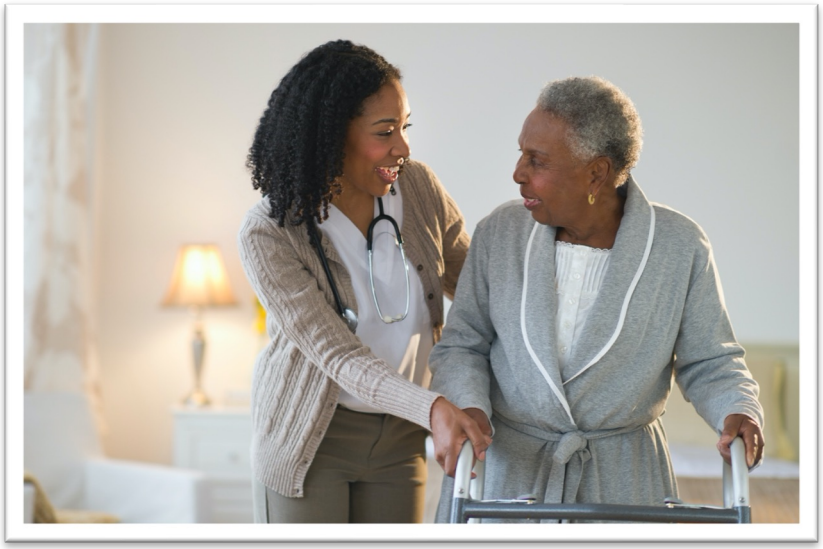
*Cynthia Rudin*
*Duke University*
*cynthia@cs.duke.edu*

**Abstract**

Black box machine learning models are currently being used for high stakes decision-making throughout society, causing problems throughout healthcare, criminal justice, and in other domains. People have hoped that creating methods for explaining these black box models will alleviate some of these problems, but trying to *explain* black box models, rather than creating models that are *interpretable* in the first place, is likely to perpetuate bad practices and can potentially cause catastrophic harm to society. There is a way forward – it is to design models that are inherently interpretable. This manuscript clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare, and computer vision.

# What are High Stake Decisions?

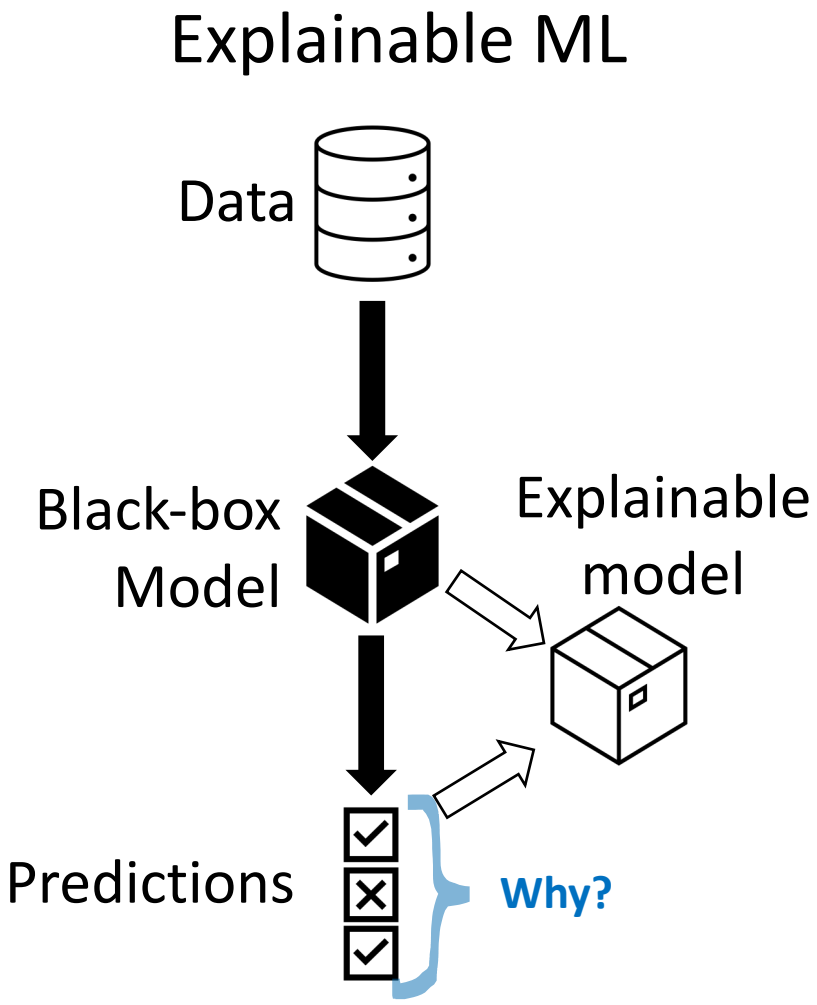
- Applications that deeply impact human-life

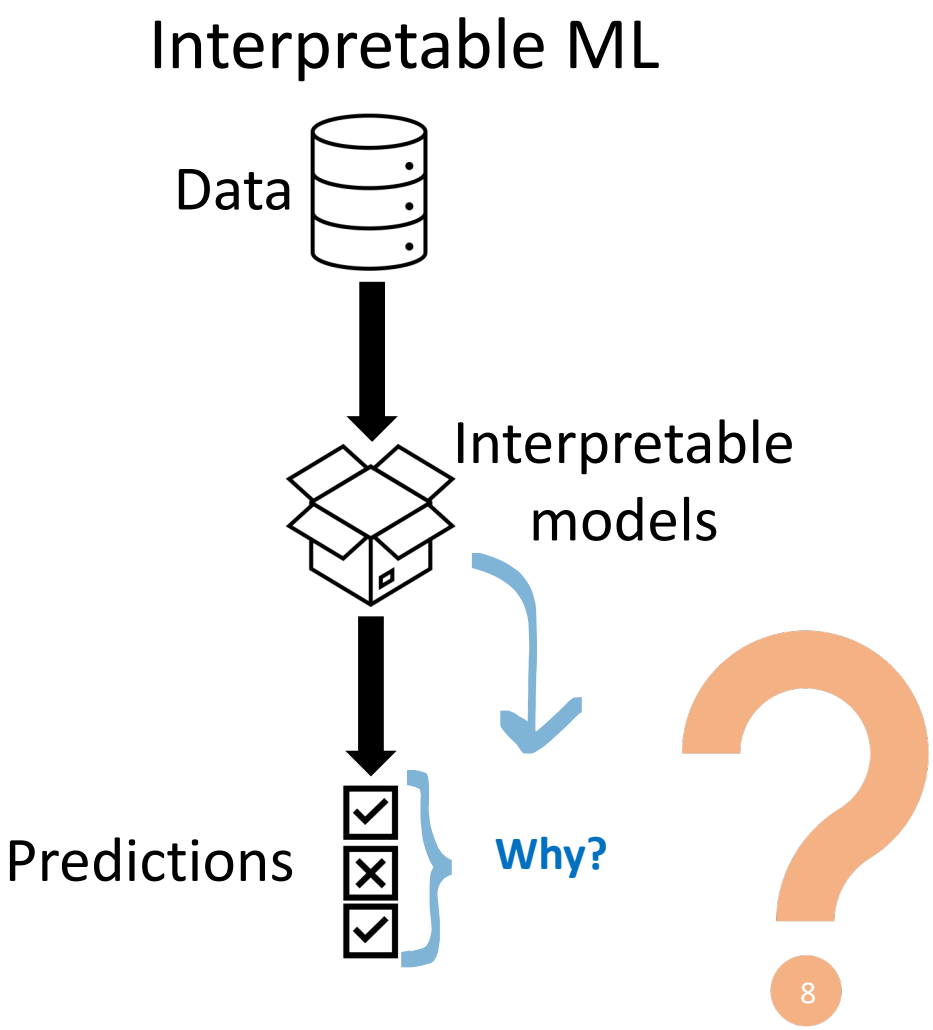


Health care

Criminal Justice

# Interpretable vs Explainable ML



# Interpretable vs Explainable ML
## Cont'd

- Interpretable Models
  - Tend to be simpler models (e.g., linear, rule-based)
  - Explanations come from the model itself
  - Decision process is transparent

- Explainable Models
  - Second (posthoc) model is created to explain the first more complex black-box model
  - Explanations are indirect
  - Decision process is opaque

# Study Design?

- Argumentative (persuasive) paper
  - Review of Related Work
  - Position of an expert in the field

- Key issues with Explainable ML
- Key issues with Interpretable ML
- Encouraging responsible ML governance

# Problem

- The lack of transparency and accountability of predictive models have severe consequences

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

230

Inmate denied parole despite having a nearly perfect record of rehabilitation

- Due to an automated system COMPAS
- The system was faulty
- The inmate had to appeal and to prove the system was wrong

https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

# Explainable ML is not good enough for High Stake decisions

- Using Explainable ML is not enough
  - Explanations are often not reliable and misleading

- We should use Interpretable ML
  - Case-based reasoning (rule-based)
  - Sparse models (easier to identify interaction)
  - Obeys structural knowledge of the domain
    - Theory of causality

# Key issues with Explainable ML

- Explainable ML explains a **black-box** model
  - A model that is too complicated for humans
  - A model that is proprietary

- Key issues
  1. Myth of accuracy and interpretability trade-off
  2. Explanations are not faithful to the original model
  3. Explanations often do not make sense
  4. Not compatible with external risk assessment
  5. Leads to overcomplicated decision path

# Myth of accuracy and interpretability trade-off

- In problems with **meaningful features** and **adequate data preprocessing** there is no significant difference between…
  - **Complex classifiers** (DNN, Random Forest)
  - **Much simpler classifiers** (Logistic regression, Decision lists)

- In a realistic scenario (iterative), the small differences in accuracy are often overwhelmed by the ability to interpret results

# The case of the New York Grid

- Goal: predict electrical grid failures across New York
  - Messy data, free-form text, legacy data (1890)
  - Initially, more complex models yield better results
    - Differences in the type of models at most 1%
- The ability to interpret the results and reinterpret data lead to
  - Revealing false assumptions, data problems
- Most accurate prediction were made by simple and sparse models

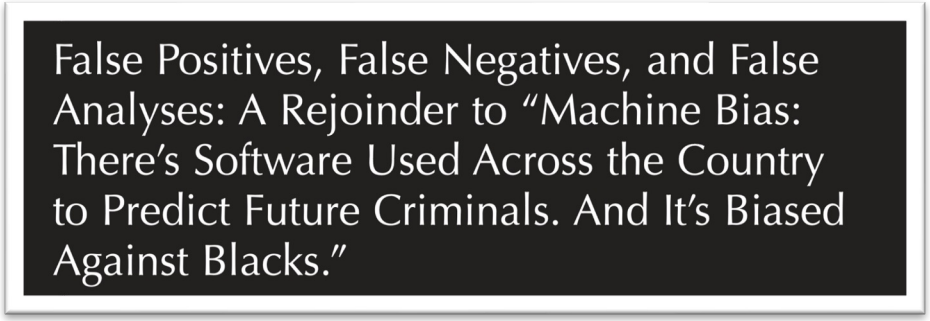
# Myth of accuracy and interpretability trade-off? (caveat)

- “Recent review and commentary articles on this topic imply (implicitly or explicitly) that the trade-off between interpretability and accuracy generally occurs.”
  - Some domains may require black-box solutions

- The author’s experience in has not corroborated with this statement
  - Healthcare and criminal justice
  - Energy reliability
  - Financial risk assessment

# Explanations are not faithful to the model

- Explanations **must** be wrong
  - Otherwise, do we need the original black box model?
- Explanations do not (always) use the same features of the models


Used explanation model and accused COMPAS to be racially biased.

Explanation model was wrong.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder

# Explanations are not informative enough

- Explanations often leave do much information out that it is not possible to “understand” the model

| | Test Image | Evidence for Animal Being a Siberian Husky | Evidence for Animal Being a Transverse Flute |
|---|---|---|---|
| Explanations Using Attention Maps |  |  |  |

Figure 2: Saliency does not explain anything except where the network is looking. We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Figure credit: Chaofan Chen and [28].

# Black-box models are not compatible with risk assessment

- In high stake decisions there are considerations (not embedded in data) that need to be combined with a risk calculation.

- Leading to a COMPAS example:
  - What if the circumstances of the crime changes someone's risk?
  - Current COMPAS does not take the type of crime into account.
    - Many judges may not know this fact.
    - A transparent model will explicitly present this limitation

# Leads to overcomplicated decision path

- For example: COMPAS has 137 factors
  - Even if a typographical error occurs 1% of the time, one in every 2 surveys might have typos
  - Typographical errors sometimes determine the decision outcomes


https://hdsr.mitpress.mit.edu/pub/7z10o269/release/4

# Key Issues with Interpretable ML

- Interpretable ML models are simpler, sparse and easier to understand

- Key issues with Interpretable models:
  1. Not easily profitable (hard to protect intellectual property)
  2. Require more effort to construct (particularly in terms of domain expertise)
  3. Black-box models uncover “hidden” patterns

# Corporations make profit of black-box models

- Interpretable models are harder to protect the intellectual property
- Another example of the COMPAS case:

Much simpler + Similar performance

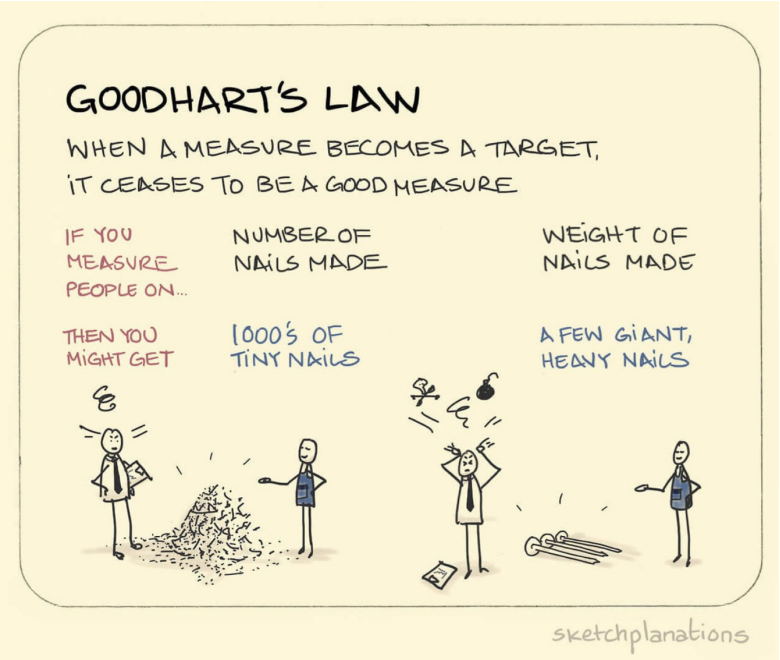
| COMPAS | CORELS |
|---|---|
| black box<br>130+ factors<br>might include socio-economic info<br>expensive (software license),<br>within software used in U.S. Justice System | full model is in Figure 3<br>only age, priors, (optional) gender<br>no other information<br>free, transparent |

Table 1: Comparison of COMPAS and CORELS models. Both models have similar true and false positive rates and true and false negative rates on data from Broward County, Florida.

# Black-box models are harder to gamify

- Keeping models hidden prevents them from being gamed or reversed-engineered.
  - A **badly defined** system may suffer from this

A **transparent** system will help users to genuinely **align** with the overall goal of improvement



# Interpretable models are harder to construct

- Domain expertise is needed to construct the definitions of **interpretability** of the domain
  - It tends to be easier to use black-box models to solve computationally hard problems

- For High Stakes decisions
  - A flawed prediction **costs more** than the analyst time

# Black-models uncover “hidden patterns’

- Models may uncover patterns not previously known by experts

- The author argue that a transparent model might be able to discover the same patterns
  - Caveat: Selected models need to be interpretable and flexible enough to fit the data

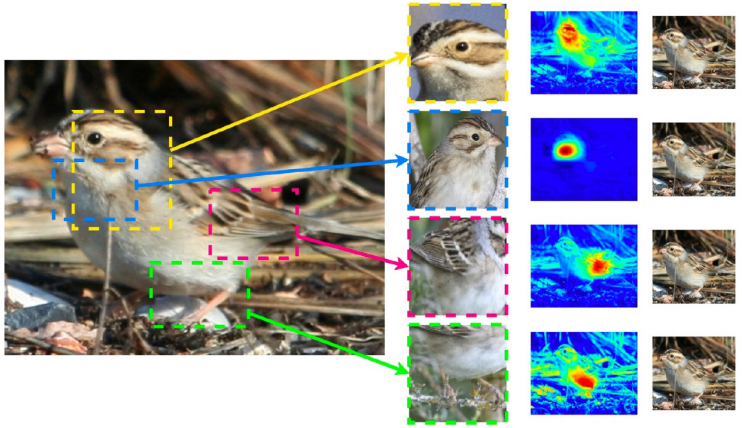
# Encouraging ML Governance

- **G**eneral **D**ata **P**rotection **R**egulation (GDPR)
  - User has the right of an explanation
  - Explanation may come from inaccurate models

- Proposal I: No **black-box** should be deployed when there exists an **interpretable** model with **similar performance**
  - Use of black-box models could be considered false-advertisement
- Proposal II: Black-box models should report accuracy

# Algorithmic Challenges

1. Constructing optimal **logic models**
   - Logical models are easier to interpret but harder to reach comparable performance
2. Construct optimal sparse **scoring algorithms**
   - Using ML algorithms (Logistic Regression) often does not reach a good performance
3. Define interpretability for specific domains
   - E.g., Computer Vision



# Conclusion

- Challenge the basic assumption of Explainable ML: **Black box is necessary for accurate predictions**
- Encourage policy makers to not accept black box solutions without attempting **interpretable variants**
- Bring awareness to the **problems** of interpretable models

# Open Discussion

# Netflix Recommendations

- Focus on describing the early days of Netflix model selection
  - Netflix Prize
  - ML Experimentation

# The Netflix Prize

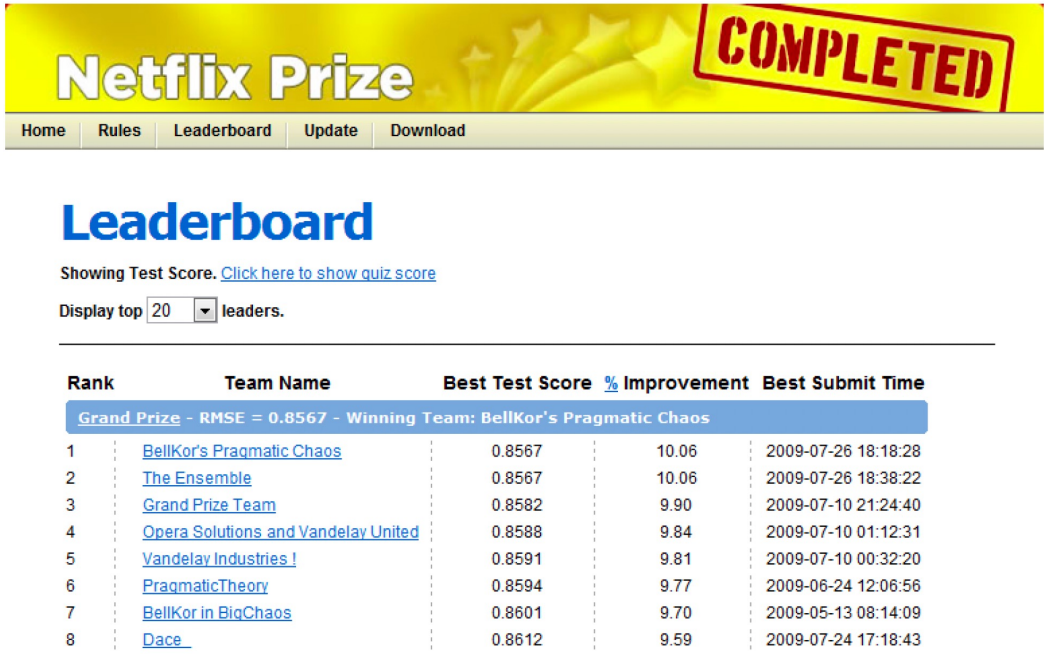
- Machine learning and data mining competition
  - $1 million to whoever improved the accuracy of their system (Cinematch) by 10%
  - Root Mean Squared Error (RMSE): 0.9525 (< 0.8572)

# The first winning team

- Korbell team won the Progress Prize
  - Improvement of 8.43%
  - More than 2000 hours of work
  - Solution: Combination of 107 algorithms
- The two best algorithms (RMSE 0.88)
  - Singular Value Decomposition (SVD)
  - Restricted Boltzmann Machines (RBM)
- Large engineering effort to deploy those models at Netflix scale
  - From 100 thousand to 5 billion recommendations

# The final winning team

- Solution blended hundreds of predictive models
  - Engineering effort to put the solution in production was not worth it



# Changes in the Landscape

- Netflix released their streaming service
  - The original Netflix prize was planned to help their DVD recommendation system
  - Hundreds of different types of devices
- 75% of what people watch is based on recommendation

# Everything is personalized

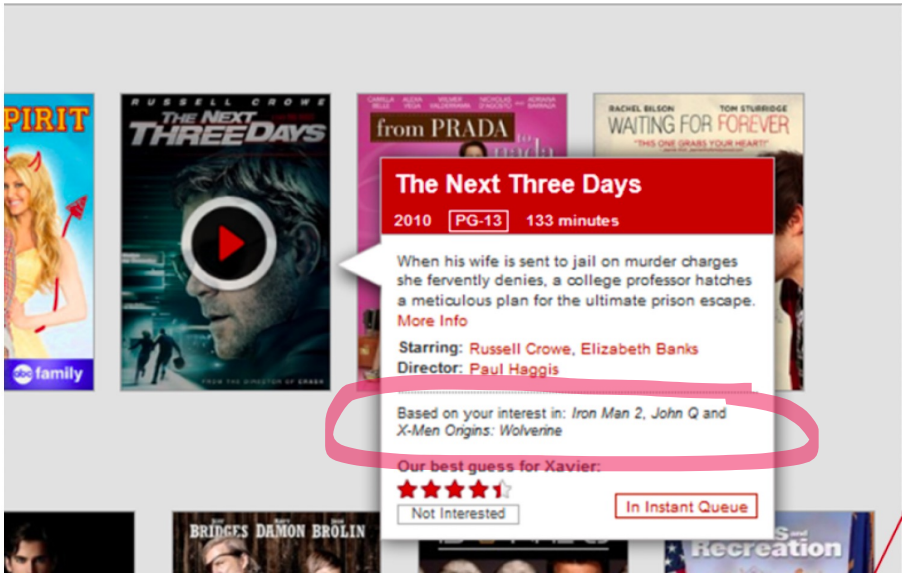
# Everything is a recommendation

- Recommendation starts with the rows
  - Groups of videos with meaningful connection

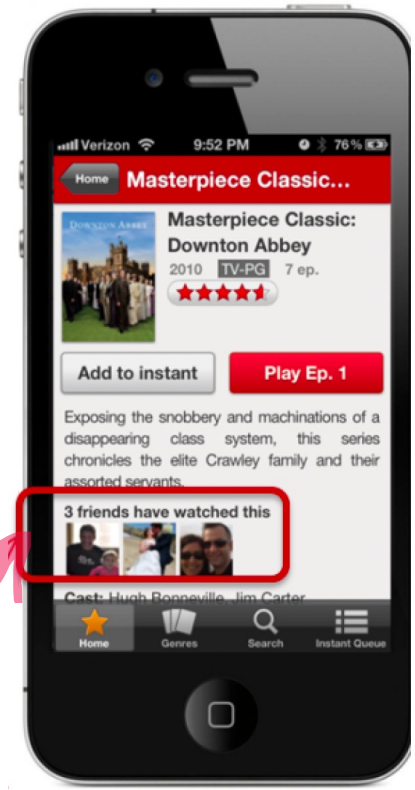
- E.g., The 10 videos you are most likely to watch

# Awareness is key

- Recommendations need to provide an **explanation**



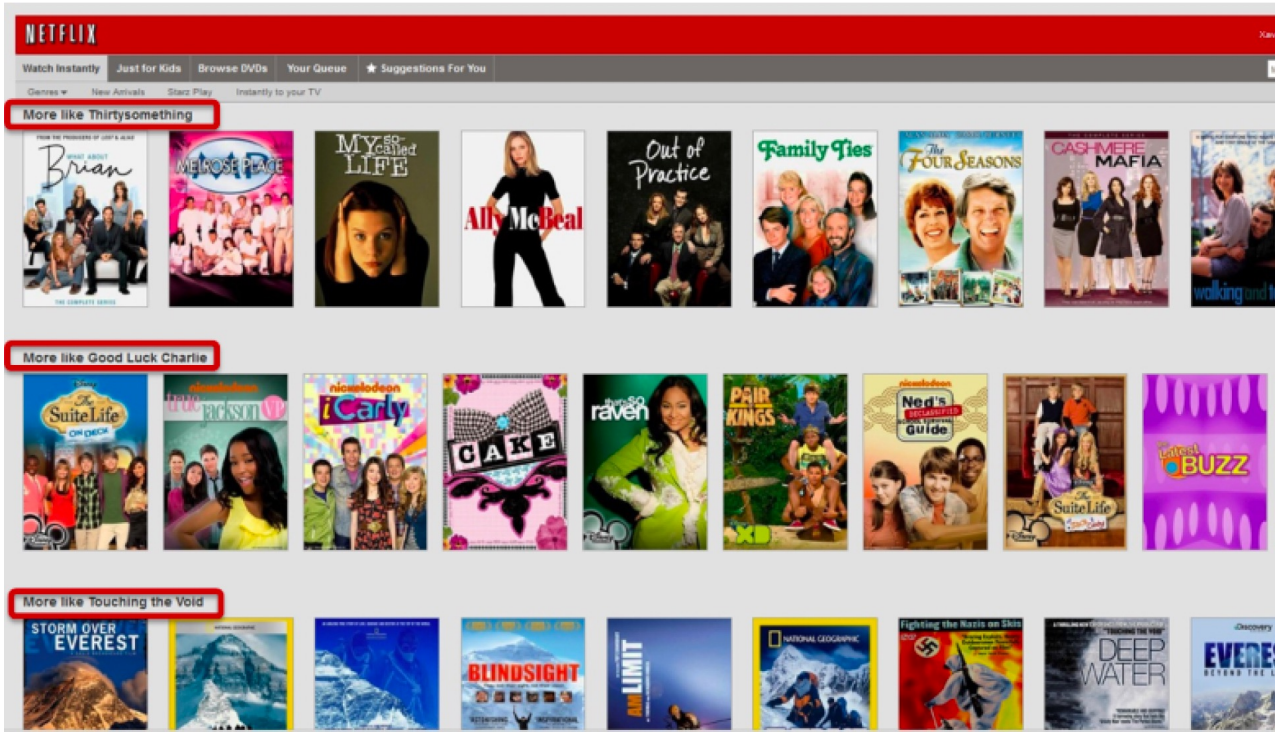

Explanation




Social support

# Similarity is an important source of personalization

- Similarity between movies, members
  - Multiple dimensions: metadata, ratings, viewing data
  - Similarity is blended and **used as features** for models



# Ranking is one aspect of effective recommendation

- The original Netflix Prize was aimed to predict the movie's ratings

- Effective recommendation needs to account:
  - Context
  - Title's popularity
  - Interest
  - Diversity
  - Freshness
  - …

# Popularity is not Everything

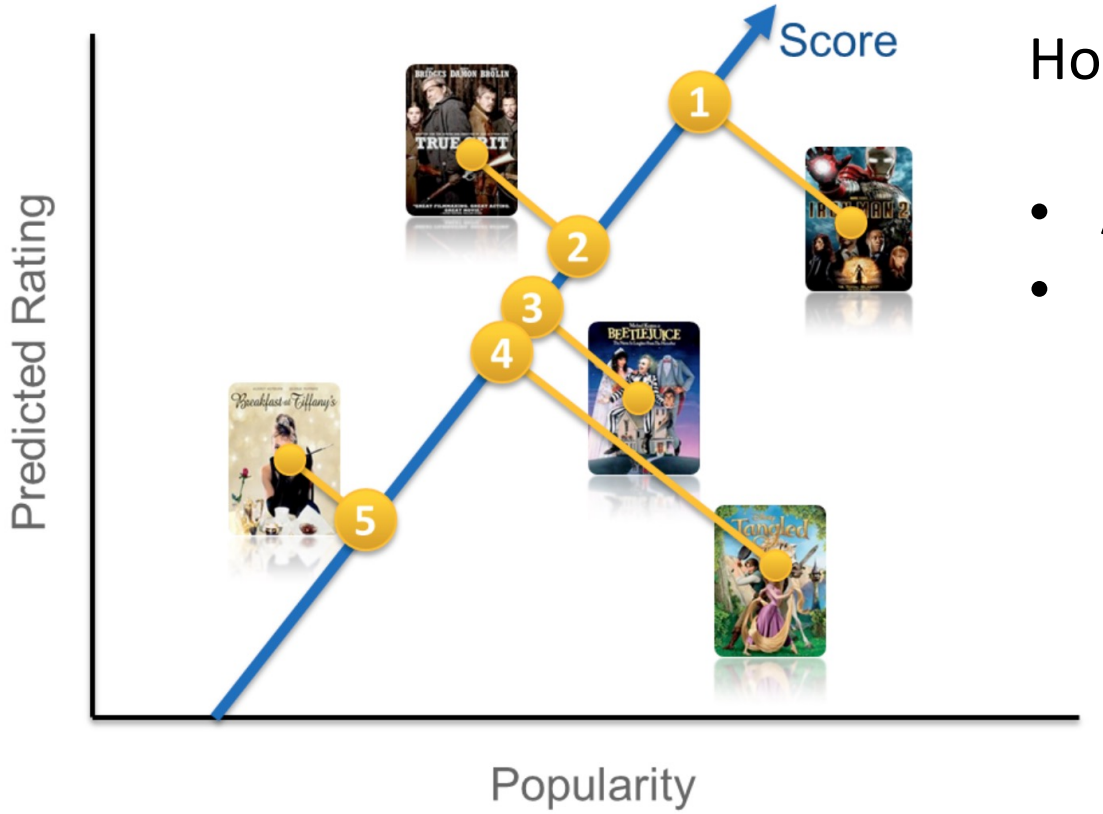
- Goal is to recommend the titles that each member is most likely to play and enjoy
  - Maximize consumption!
  - Ranking = Scoring + Sorting + Filtering

- Obvious baseline: Popularity
  - This leads to niche and less popular movies that will never be watched (even if a user may like them)

# Balancing Popularity and Personalization

Popularity Personalization

$$\text{frank}(\text{user}, \text{video}) = \mathbf{w1}\ \text{popularity}(\text{video}) + \mathbf{w2}\ \text{rating}(\text{user}, \text{video}) + b$$




How to find the weights?

- A/B testing?
- Machine Learning!

# Data Sources

- Many relevant data sources can be used to optimize recommendations
  - Billions of item ratings from members
  - Item popularity (different time frames)
  - Millions of plays every day
  - Millions of items in the queue
  - Title's metadata: actors, directors, genre, …
  - Presentations: how many recommendations were successful with that title?
  - Social data, search terms, …
  - External data: box-office, critic reviews, …

# Models used at Netflix

- Given the plethora of data, Netflix uses multiple types of models
  - Linear regression
  - Logistic regression
  - Elastic nets
  - Singular Value Decomposition
  - Restricted Boltzmann Machines
  - Association Rules
  - Random Forests
  - Clustering algorithms (k-means)

Use anything that solves their problem

# Singular Value Decomposition

- Identify correlation patterns
- Simple and interpretable algebra

$$\mathbf{X}_{[n \times m]} = \mathbf{U}_{[n \times r]}\, \mathbf{S}_{[r \times r]}\, (\mathbf{V}_{[m \times r]})^{\mathsf{T}}$$

$$\underset{m \times n}{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix} = \underset{m \times r}{U} \begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix} \underset{r \times r}{S} \begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix} \underset{r \times n}{V^{\mathsf{T}}} \begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}$$

- **X**: *m x n* matrix (e.g., m users, n videos)
- **U**: *m x r* matrix (m users, r concepts)
- **S**: *r x r* diagonal matrix (strength of each 'concept') (r: rank of the matrix)
- **V**: r x n matrix (n videos, r concepts)

# Restricted Boltzmann Machines

- Simplified Neural Nets
  - One hidden layer
  - Hidden neurons are not connected
- Scalable

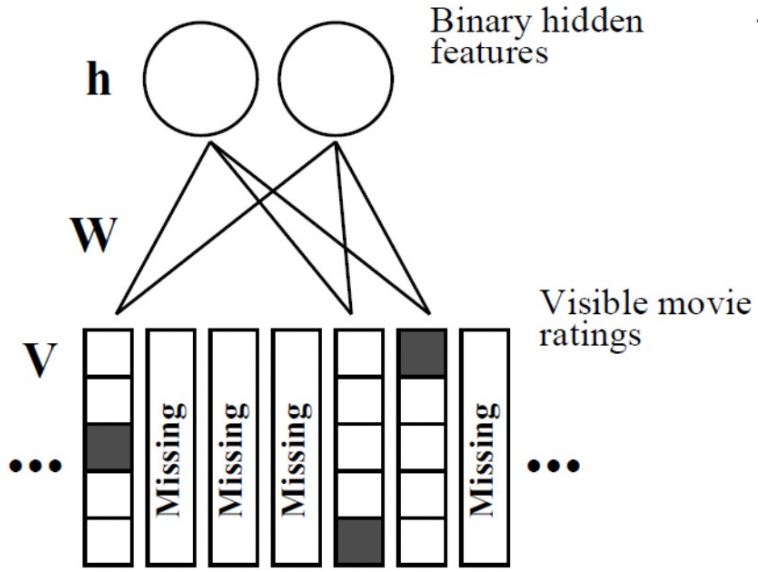
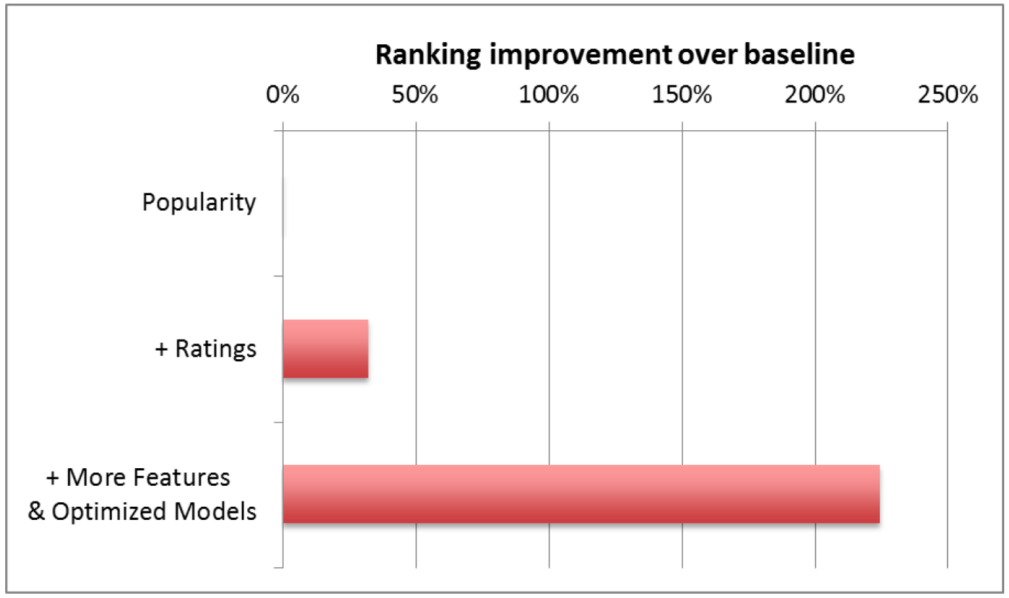



*Figure 1.* A restricted Boltzmann machine with binary hidden units and softmax visible units. For each user, the RBM only includes softmax units for the movies that user has rated. In addition to the symmetric weights between each hidden unit and each of the $K = 5$ values of a softmax unit, there are 5 biases for each softmax unit and one for each hidden unit. When modeling user ratings with an RBM that has Gaussian hidden units, the top layer is composed of linear units with Gaussian noise.

# Improving the ranking

- Baseline based on popularity
  - Using ratings already improved the baseline in ~40%

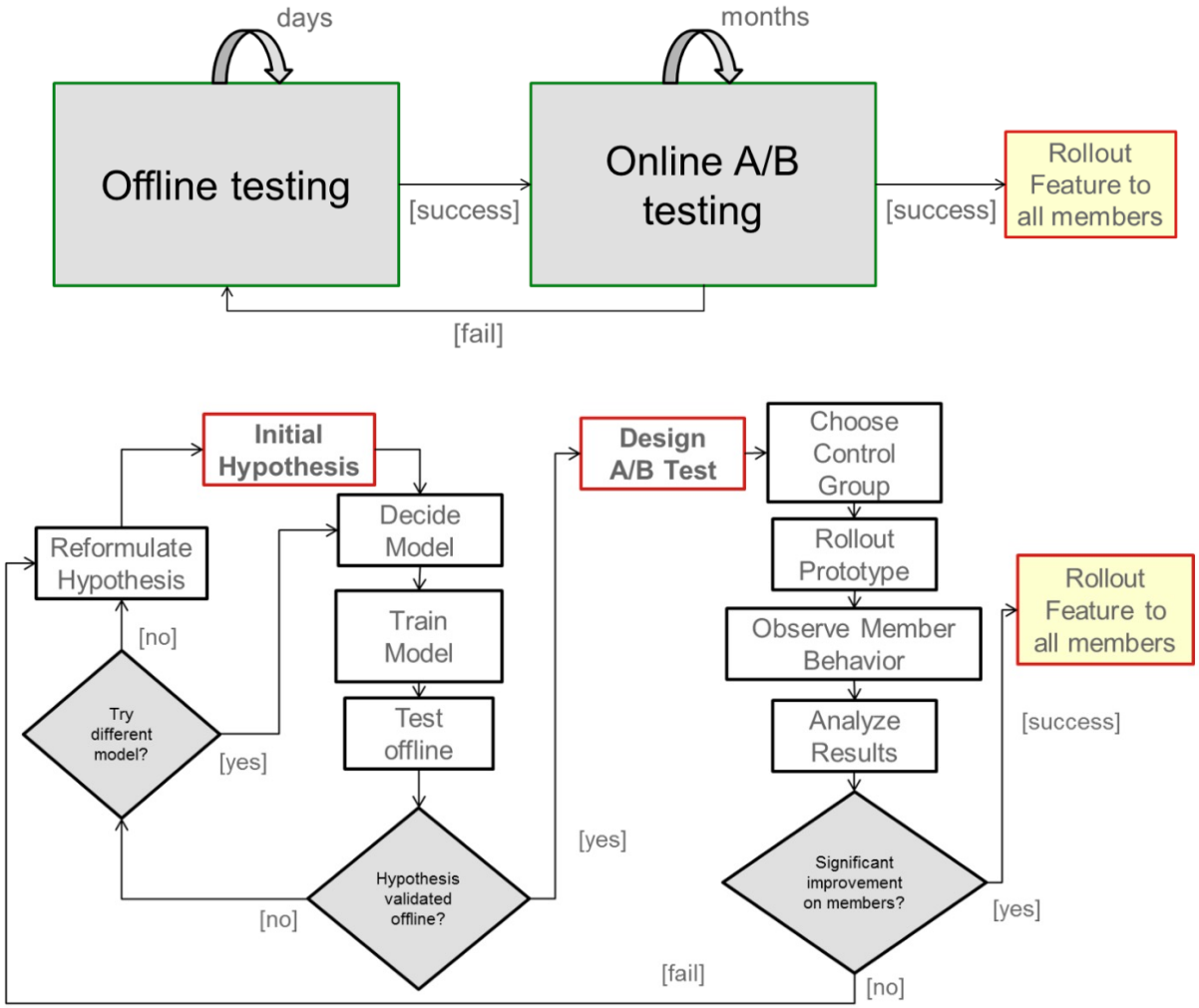
More features + Optimized Models improved the performance >2x



# Consumer Data Science Approach

1. Start with a hypothesis
   - Algorithm/feature/design X improves user engagement by X%
2. Design a test
   - Prototyping
   - Design the experiment: control, dependent and independent features, significance, etc…
3. Execute the test
4. Let the data speak for itself

# Offline testing to A/B testing



# Conclusion

- The Netflix Prize abstracted the recommendation problem to a proxy question: predicting ratings

- Recommendation is far beyond just the five stars ratings and became essential to Netflix's success

- Data mining and other experimental approaches to:
  - incrementally inform intuition and
  - prioritize effort investment

# Open Discussion

# Homework

- Two papers **will be** posted on Moodle
- For one of the papers, write a summary (aprox. 1/3 of a page)
- For the other paper, write a critique, which includes a summary, at least 3 strong points and at least 3 weaknesses (aprox. 1 page).
- Submit your summary and critique on Moodle by Friday, March. 11th at 5pm

# References

- Beyond the 5 Stars – Slide presentation