UW MLEARN 410 B

- Instructors:
  - Justin Donaldson
  - Zach Alexander
  - Sid Rajaram

Data Science: Service, Communities, and Search

Salesforce Einstein

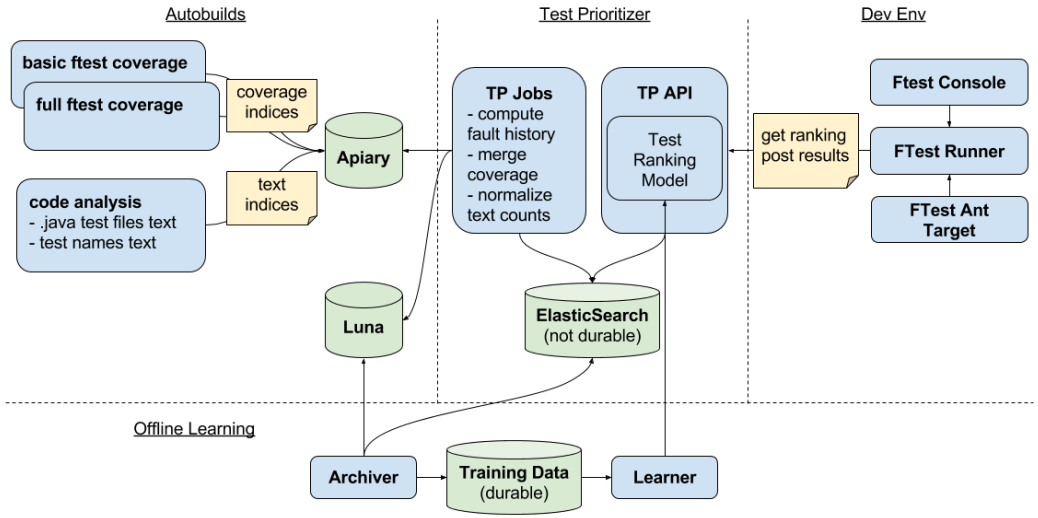
- TA: Jose Villalta

Service Cloud Einstein

- Is this e-mail Spam?
- What kind of help does the customer need?
- Is it a standard question? Can I just auto-reply?
- Is it a high priority case? Should it be bumped up the queue?
- Which of my agents has the skillset to handle this case? And what is their availability like? Who should get this case?
- Is there some bug/problem that is causing a bunch of new cases to occur right now (like trending hashtags)?

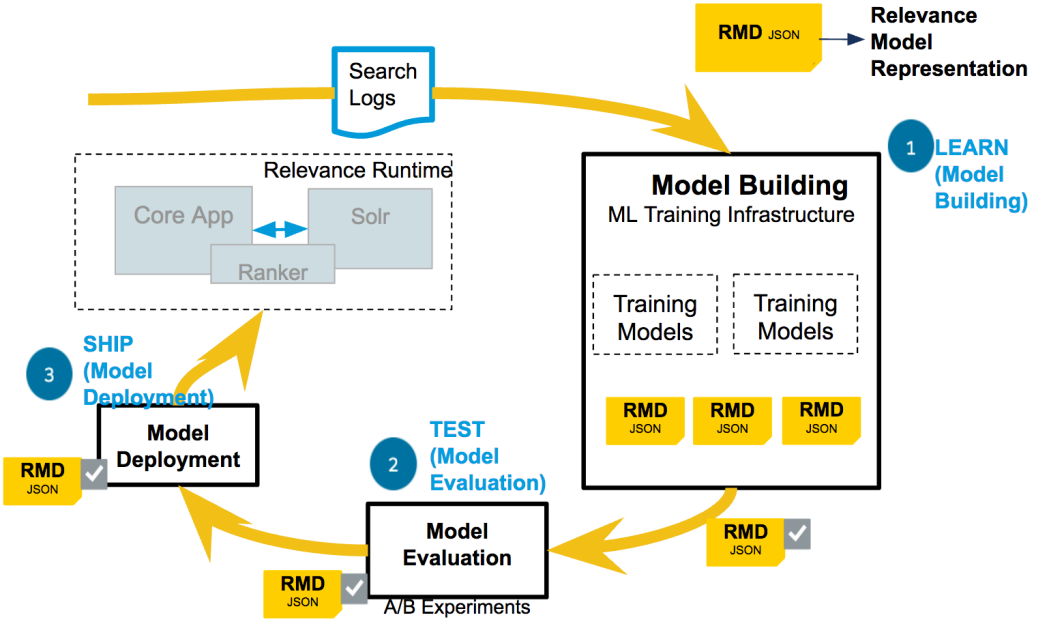
Productivity Intelligence

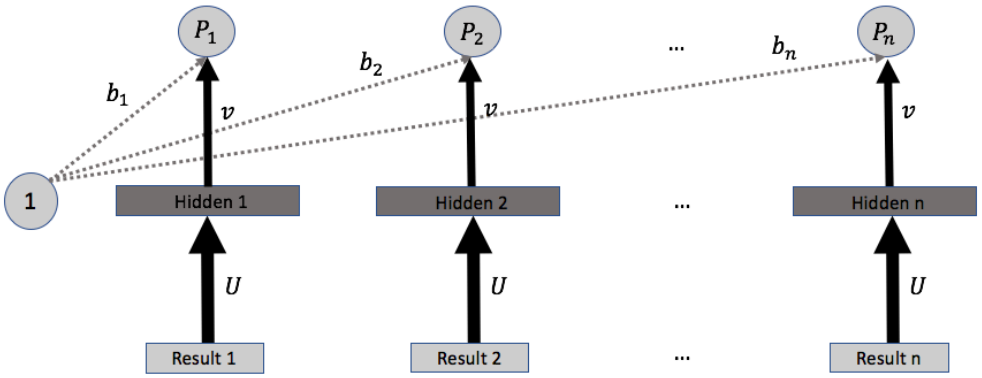
- Will this code cause a bug?
- Who should review this changelist?
- Who can fix this error?
- Are there some associated tests for the code I am trying to change?



Search Relevance

- Rank documents returned by a search engine.
- Need to understand the query…
- AND the documents…
- AND the user's intent…





Course Topics

- Supervised Learning (Classification and Regression)
  - Generalized Linear Models (1 class – Sid)
  - Trees and Forests (1 class – Justin)

- Unsupervised Learning
  - Clustering (1 class – Justin)
  - Dimensionality Reduction (1 class – Justin)
  - Topic Models (1 class – Zach)

- Domain-specific ML
  - Recommender Systems (1 class – Justin)
  - Advanced Topics : Neural Networks, Out-of-Core (1 class – Sid)
  - Anomaly Detection (1 class – Sid)
  - Learning to Rank (1 class – Zach)

- Project Presentations (1 class - Sid, Zach, Justin)

In-class Logistics

- Some lecturing and slides…but mostly
- "Flipped" classroom:
  - Will have R Markdown notebooks we're going to walk through where you fill in the steps
  - Try to work in small groups (2-3) to discuss, but then implement your own solutions

- **YOU WILL NEED R STUDIO TO WORK ON THESE NOTEBOOKS AND TO DO HOMEWORKS**

Homeworks

- 2 homeworks:
  - Supervised Learning
    - Covering the first two lectures
    - Will be put up on the Canvas site very soon
    - Due 4/20
  - Unsupervised Learning
    - Covering lecture 3-5
    - Will be released soon after we start the module
    - Due 5/18
- **R Markdown notebooks again**

Homeworks

- 2 homeworks:
  - Supervised Learning
    - Covering the first two lectures
    - Will be put up on the Canvas site very soon
    - Due 4/20
  - Unsupervised Learning
    - Covering lecture 3-5
    - Will be released soon after we start the module
    - Due 5/18
- **R Markdown notebooks again**

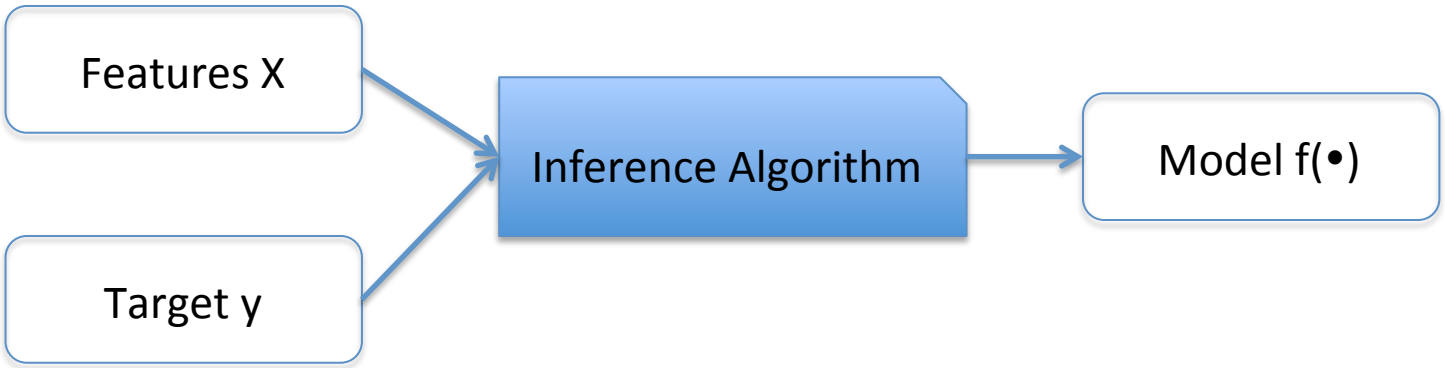
Project

- Work individually
- Use any language/tools you want – if it's not R, we may not be able to help with any issues that arise!
- Ideally, something that is relevant to you
- In-class presentation and written report
- Not pre-cleaned/processed data
- CHECK WITH US FIRST!
- Proposals due as part of Homework 1.

What the course is about

- Primarily about building the *end-to-end model building process*.
- We'll mostly explain topics via example
- We'll cover *some* theory and algorithms, but it won't be the focus of the course..

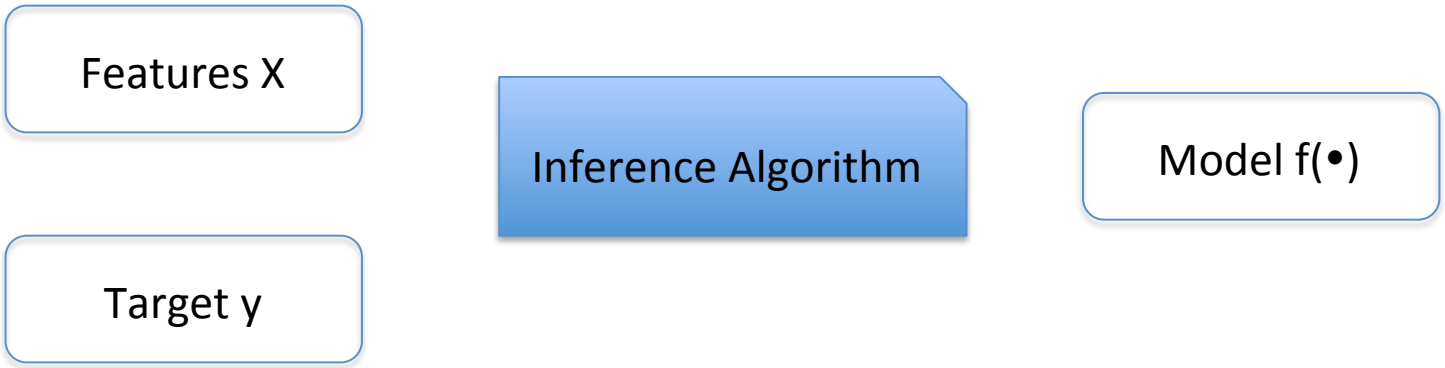
Textbook ML caricature



Supervised Learning

Given matrix X, with targets (labels) y,
learn f s.t. f(X) = y

Let’s decompose this



Real data never looks like this

- X is numeric – your raw data probably isn't
  - Raw text
  - Categorical features (Strings? Numeric?)
  - Date-Time features
  - Corrupt/missing data
  - Privacy concerns
- You will spend most of your time getting to a usable X

Featurizing Text Data

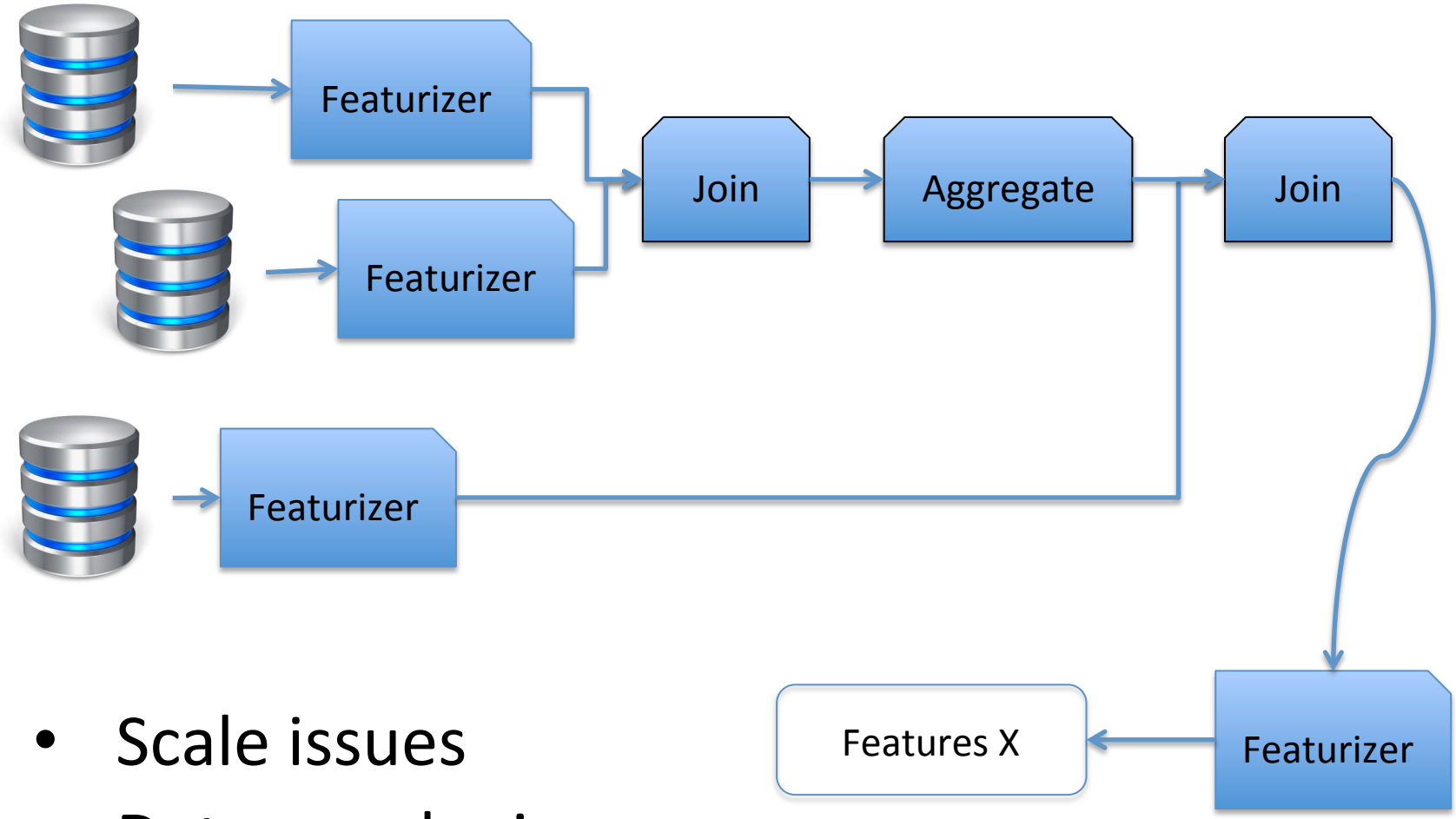
- Normalize text
  - foreign languages(?), remove punctuation, case
  - tokenize into words/phrases, drop stop words
- Vectorize documents
  - Counting/TF-IDF vectorizer
  - Word2vec (use external data?)
  - Topic models

Categorical Data

- Cardinal (cat/dog) or ordinal (high/medium/low)?
  - May already be numeric 
  - If ordinal, can your model work with that?
  - Does your model work with categorical features? Or One-hot representations?

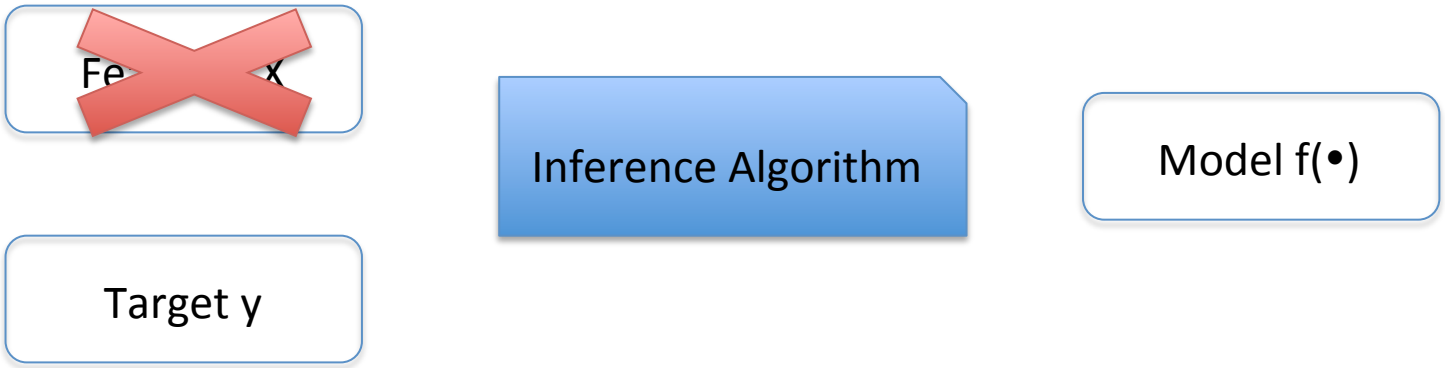
Other Data Types

- DateTimes
  - Datetime object, UTC timestamp, Formatted string (e.g. “YYYY-MM-DDT00:00:00”)
  - Need local time? Timezones?
  - Convert to number of seconds/hours/days from some fixed date
- Missing data/Nulls
  - Impute with mean? with a regression model? +/- Infinity?



- Scale issues
- Data may be in different places

Back to the big picture



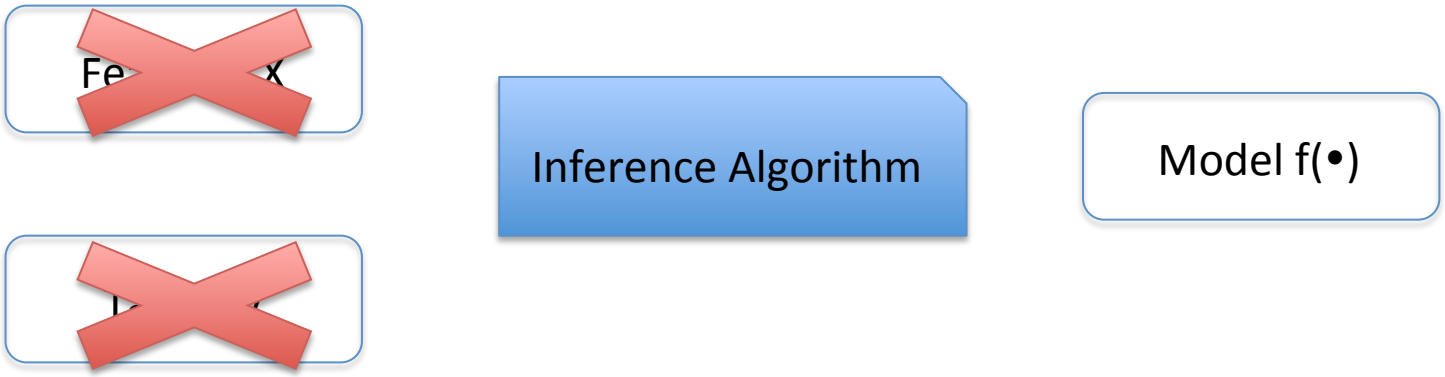
Target

Target y

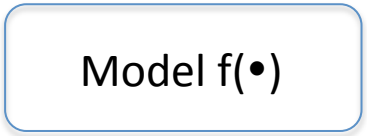
- Needs to be aligned to users needs
  - Can often be hard to define
  - E.g. - predicting whether a sales opportunity is likely to convert…in what timeframe

- Need to do everything we did to define X

Back to the big picture



Considerations for models



- Easy to debug?
  - Are the parameters directly interpretable to you?
  - Is it easy to understand prediction for a given sample?
  - Is the bug due to a poorly formulated problem? a bug in the code? bad data (includes bugs in the data pipeline)?
    - Need to track useful metrics

Can you explain *what* to a customer?

Frequently bought together

Total price: $204.64

Add all three to Cart

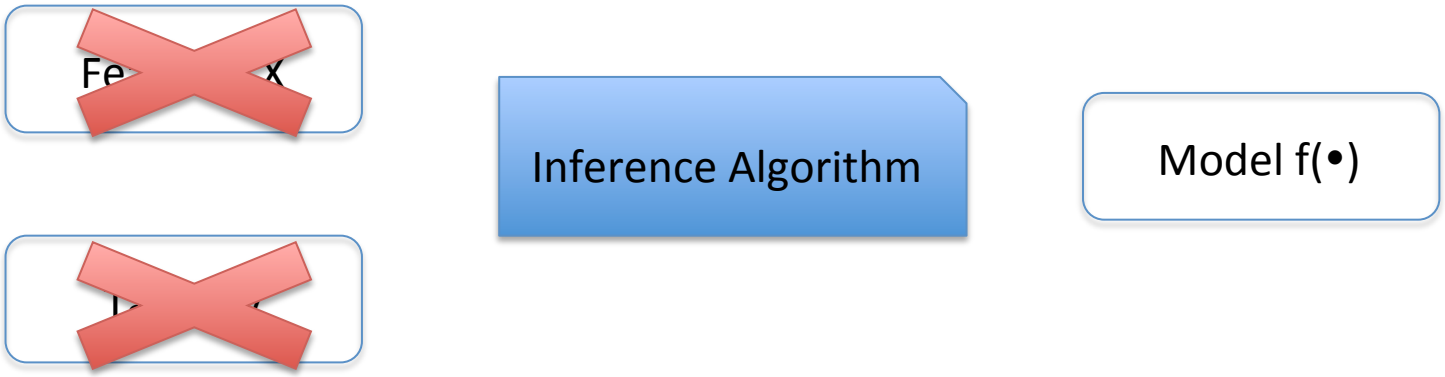
Add all three to List

- **This item:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition… by Trevor Hastie Hardcover $72.06
- An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) by Gareth James Hardcover $60.58
- Deep Learning (Adaptive Computation and Machine Learning series) by Ian Goodfellow Hardcover $72.00

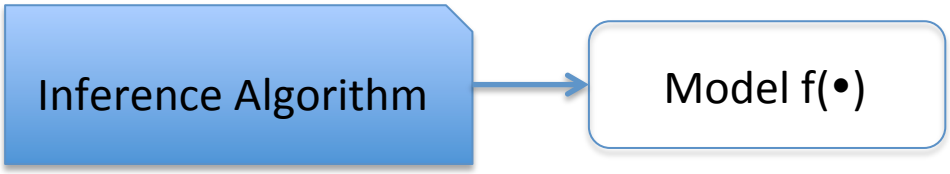
Can you explain *why* to a customer?

- Understand important factors / causal assessments
- Can be used for decision making/policy changes
  - E.g. – HR models to predict employee churn…use to change compensation/benefits structure

Back to the big picture



(Model, Training algorithm)



- Need to be selected jointly
  - Not all inference algorithms work with all models

- Need to consider scale/performance issues
  - Deep learning models need lots of data
  - Big Data -> online learning algorithm
  - Big Data -> distributed learning algorithm

The Netflix Prize (2006-2009)

- Prize = $1 million

- Want to predict movie rating, given <user, movie, date>

- Measurement metric – RMSE (*controversial*)

- Early leaders – Geoff Hinton's team using Restricted Boltzmann Machines (RBMs)

The Netflix Prize (2006-2009)

- RBMs are a type of neural network that is *very* slow to train

- Final winner – used an ensemble of models, including 100s of RBMs (> 10% improvement)

- Netflix *did not use* the winning solution techniques – too slow and expensive!

Course Topics

- Supervised Learning Methods (Classification/ Regression)
  - Basic Linear Models (today)
  - Trees and Forests
- Unsupervised Learning
  - Clustering
  - Dimensionality Reduction
  - Topic Models
- Domain-specific Machine Learning
  - Recommender Systems
  - Advanced Topics
  - Anomaly Detection
  - Learning to Rank
- Project Presentations

Course Topics

- What are you interested in learning?
  - Make suggestions in Discussion on Canvas site
  - ~~Deep Learning~~
- Any other suggestions or questions for logistics?