Machine Learning for Big Data
(CSE 547 / STAT 548)

(…what is “big data” anyways?)

Course Staff

Instructor:

- Sham Kakade

Two Great TAs: (interact with them. learn.)

- Aravind Rajeswaran
- Yali Wan

CONTENT

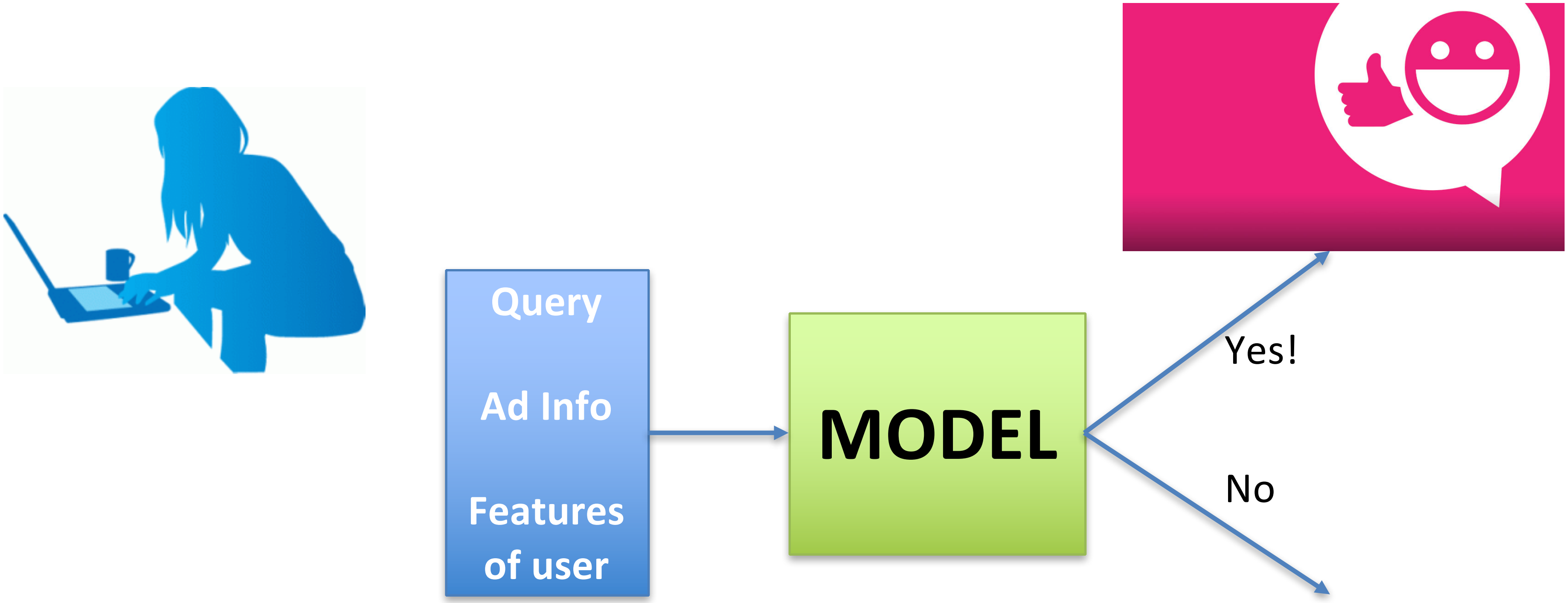
What is the course about?

Course Structure

- Some “case studies”
  - Estimating Click Probabilities
  - Document Retrieval
  - fMRI Prediction
  - Collaborative Filtering
  - ??
- Not comprehensive, but a sample of tasks and associated solution methods
- Methods broadly applicable beyond these case studies

1. Estimating Click Probabilities

- **Goal:** Predict whether a person clicks on an ad
- **Basic method:** logistic regression, online learning



Query

Ad Info

Features of user

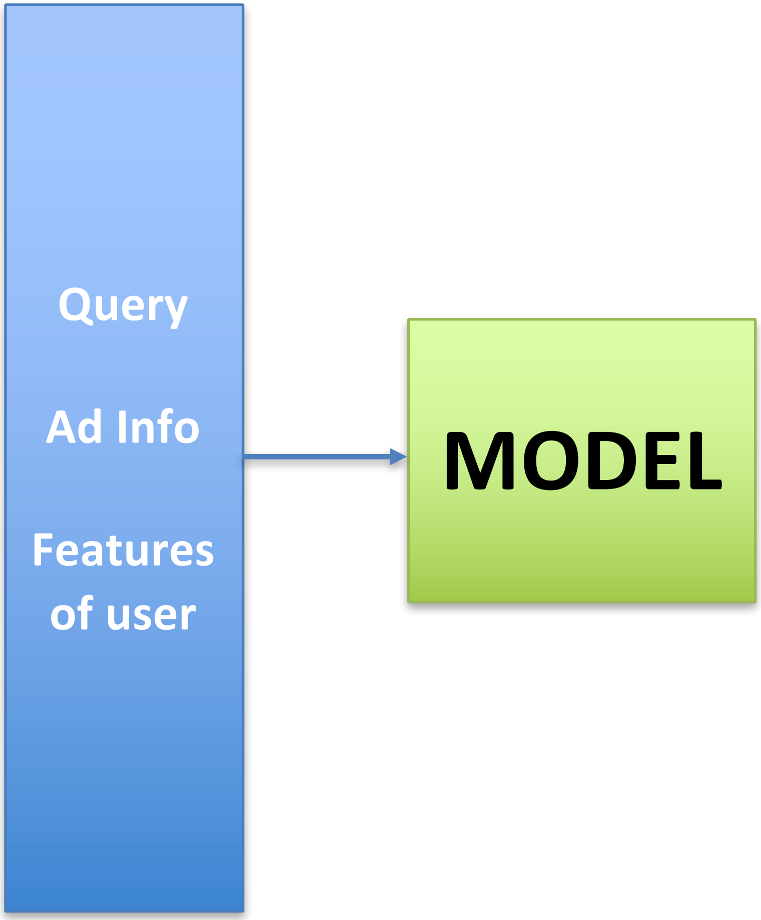
MODEL

Yes!

No

1. Estimating Click Probabilities

- **Challenge I:** Overfitting, high-dimensional feature space
- **Advanced method:** L2 regularization, hashing

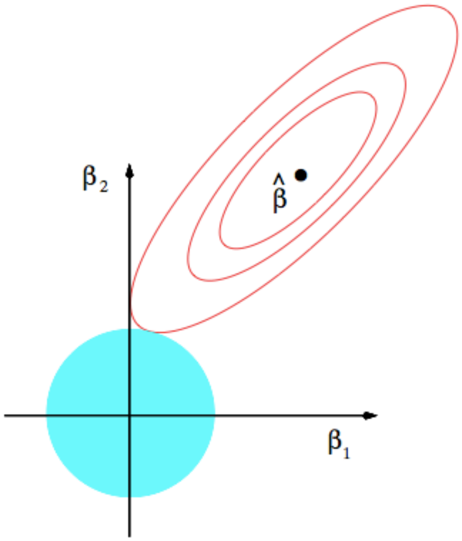


Query

Ad Info

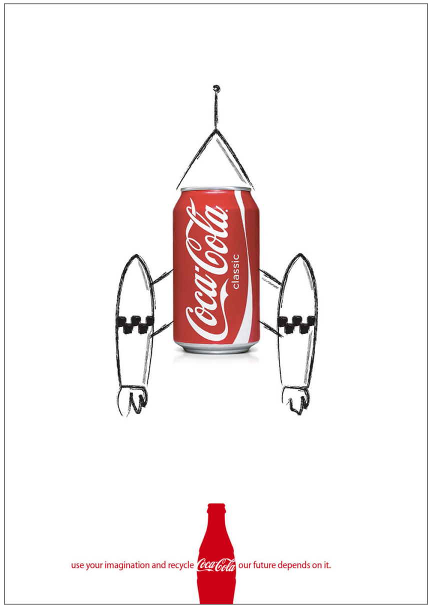
Features of user

MODEL



1. Estimating Click Probabilities

- **Challenge II:** Dimension of feature space changes
  - New word, new user attribute, etc.
- **Advanced method:** sketching, hashing



Coca-Cola
classic
use your imagination and recycle Coca-Cola our future depends on it.

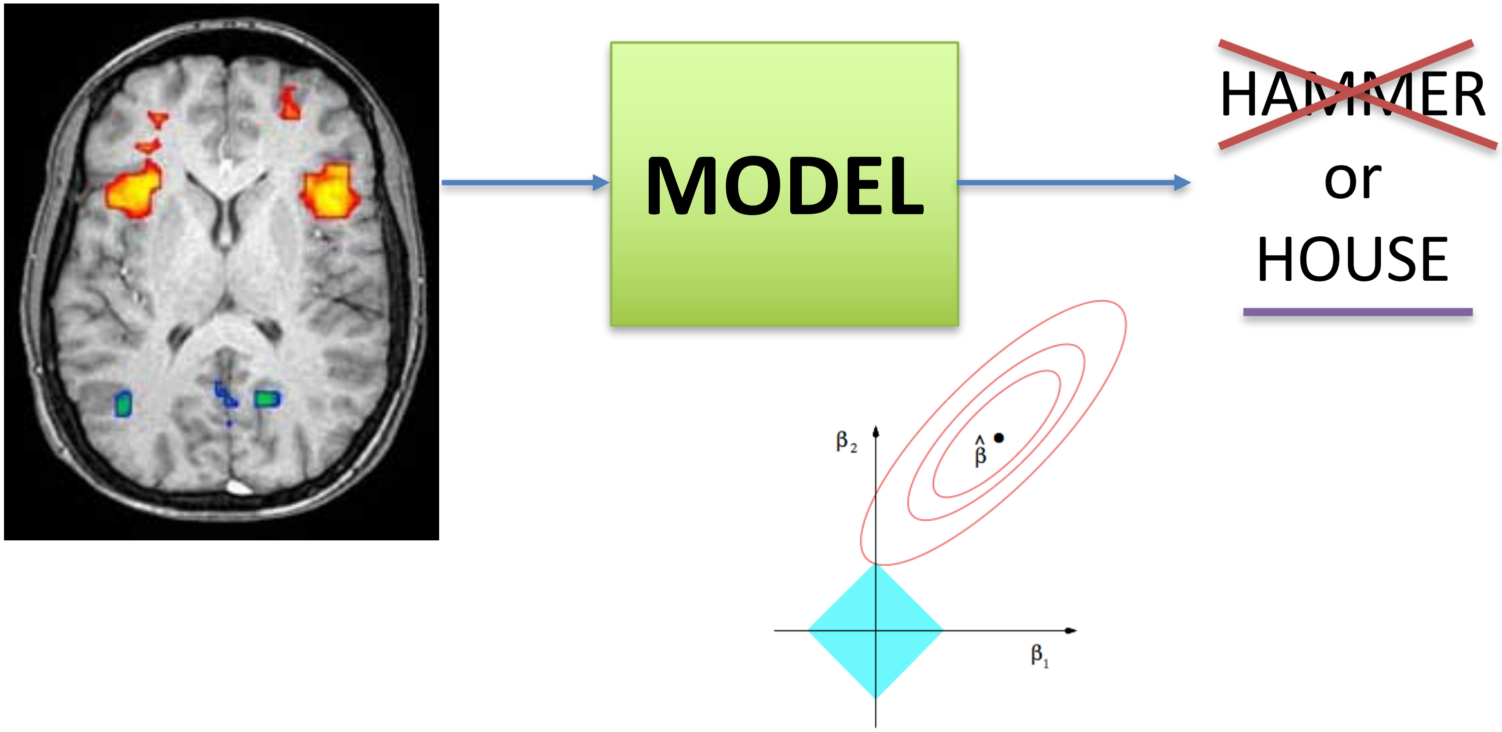
2. Document Retrieval

- **Goal:** Retrieve documents of interest
- **Methods:** fast K-NN, k-means, mixture models, Hadoop

Abbott Reveals Views on Trial
FINAL EXTRA
5 FLYERS BAIL OUT OVER PACIFIC---2 SAVED

3. fMRI Prediction

- **Goal:** Predict word probability from fMRI image
- **Challenge:** p >> n (feature dimension >> sample size)
- **Methods:** L1 regularization (LASSO), parallel learning



MODEL

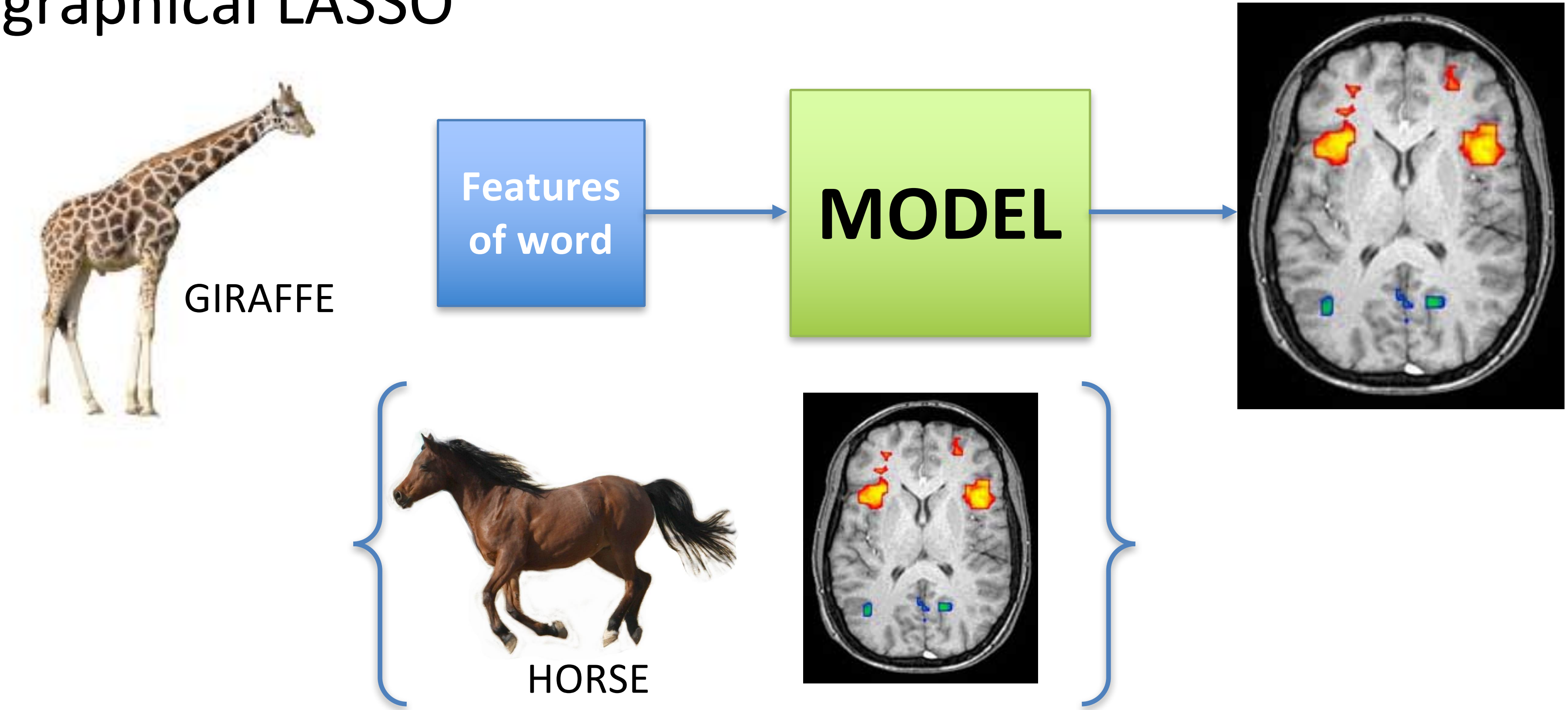
~~HAMMER~~

or

HOUSE

3. fMRI Prediction

- **Goal:** Predict fMRI image for given stimulus
- **Challenge:** zero shot learning (generalization)
- **Methods:** features of words, Mechanical Turk, graphical LASSO



GIRAFFE

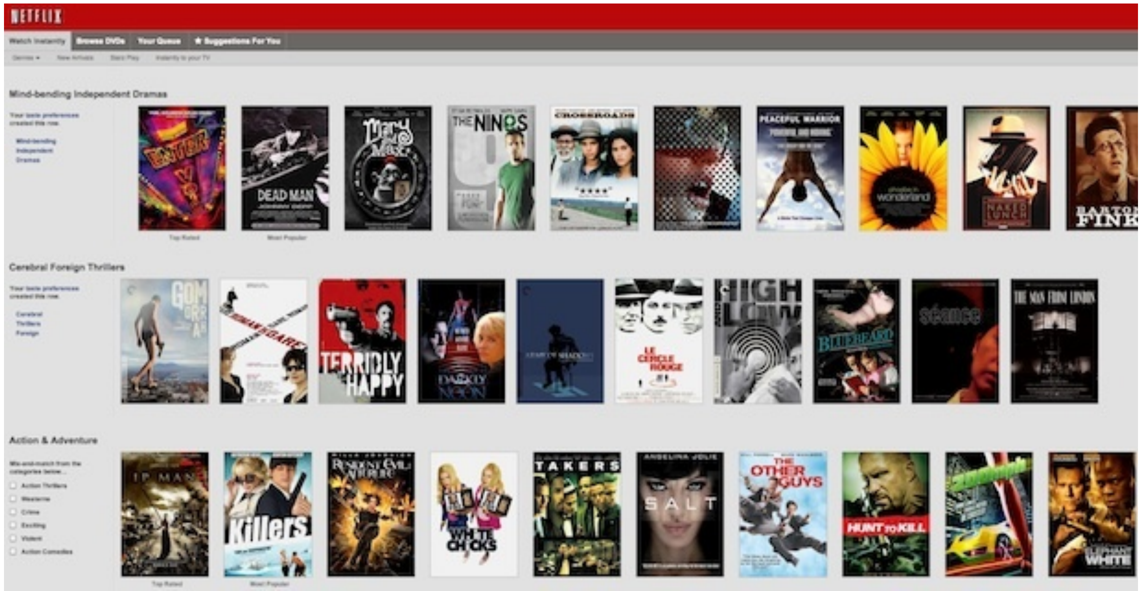
Features of word

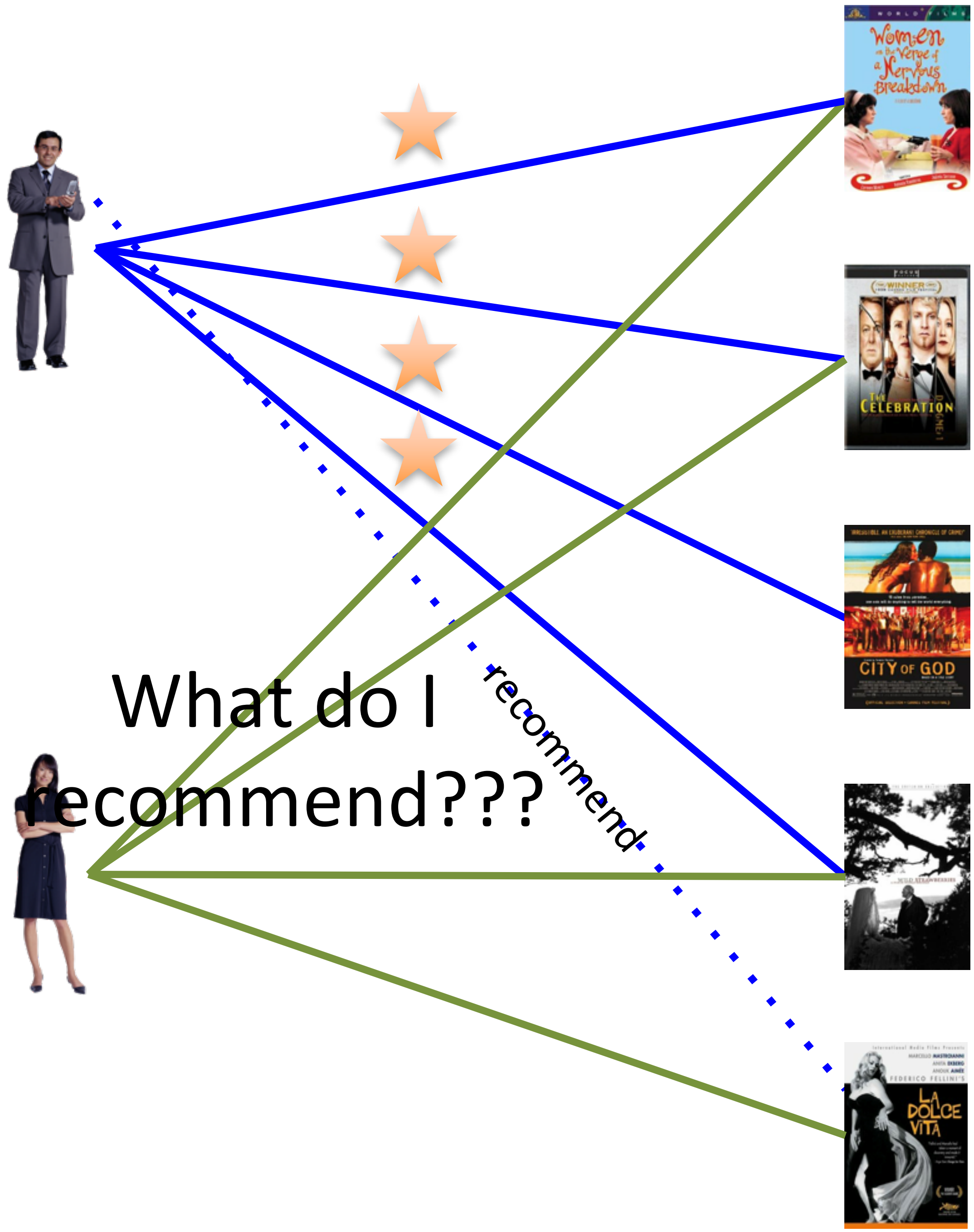
MODEL

HORSE

4. Collaborative Filtering

- **Goal:** Find movies of interest to a user based on movies watched by the user and others
- **Methods:** matrix factorization, latent factor models, GraphLab





What do I recommend???

Women on the Verge of a Nervous Breakdown

The Celebration

City of God

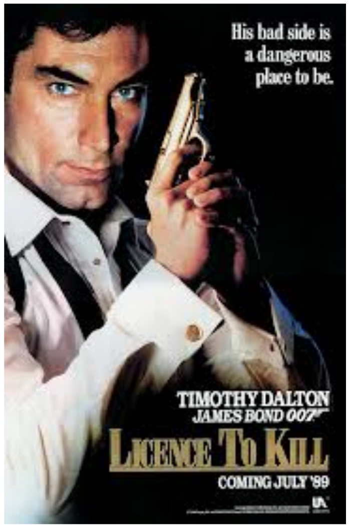
Wild Strawberries

La Dolce Vita

4. Collaborative Filtering

- **Challenge:** Cold-start problem (new movie or user)
- **Methods:** use features of movie/user

IN THEATERS


His bad side is a dangerous place to be.
TIMOTHY DALTON JAMES BOND 007
LICENCE TO KILL
COMING JULY '89

Scalability

- Throughout case studies, introduce notions of parallel learning and distributed computations

Assumed Background

Official Prereq (strict): CSE 546 or STAT 535

Know specific topics:

- Linear and logistic regression, ridge regression, LASSO
- Basic optimization (e.g., gradient descent, SGD)
- Perceptron algorithm
- K-NN, k-means, EM algorithm

Comfortable with:

- Java or Python
- Ability to learn programming languages (TensorFlow?)
- Probabilistic and statistical reasoning
- Linear Algebra

Computational and mathematical maturity

LOGISTICS

How is the course going to operate?

Diversity/Gender Issues

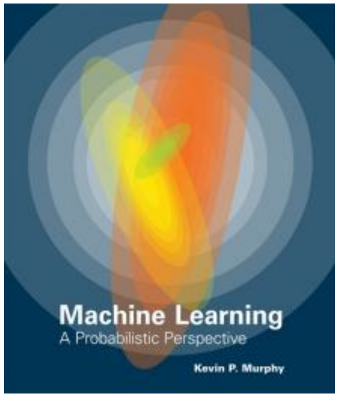
- An acknowledgement: there are diversity/gender issues to overcome.
  - Please be mindful of this.

Website and Catalyst

- Course website: courses.cs.washington.edu/courses/cse547/17sp/index.html

- Canvas:
  - Used for all discussions!!
  - Post all questions there (unless personal)
  - Homework collection
  - Personal: cse547-instructors@cs.washington.edu

Reading

- Required textbook:
- 

  "Machine Learning: A Probabilistic Perspective"

  Kevin P. Murphy
- Also, readings will be from papers linked to on course website
- Please do reading before lecture on topic

Homework

- 4 HWs, approx one for each case study
- Collaboration allowed, but write-ups and coding must be done individually
- You must submit your code.
- Due on posted date/time.
- Late: (up to) 1 day late 33%, (up to) 2 day late 66%, etc
- If you plan to be late, DO NOT TAKE THE COURSE.
- YOU MUST SUBMIT ALL HW TO PASS THE COURSE (EVEN IT IS FOR 0 CREDIT)

Project

- Individual, or teams of two
- New work, but can be connected to research
- Schedule: SEE WEBSITE FOR CHANGES TO DATES
  - Proposal (1 page) – April 7
  - Progress report /Milestone (3 pages) – May 5
  - Poster presentation –
    Thursday, June 1, 9:00-11:30am (YOU MUST MAKE THIS)
  - Final report (8 pages, NIPS format) – June 6

Grading

- HWs 1, 2, 3, 4 (15% each)
- Final project (40%)

- GRADING QUESTIONS: All regrading/policy change questions must be requested by email at cse547-instructors@cs.washington.edu. All in personal discussions (for TAs/instructors) are limited to knowledge based questions. Regrading may result in any part of the HW set going up or down.

Support/Resources

- Office Hours
  - TBD
- Discussion Board

Conclusion

- It will be hard work and fun…
- ML is having tremendous impact in technology/society.

  - What about social impact?
  - And social good?