Data Mining & Machine Learning

CS37300
Purdue University

August 21, 2017

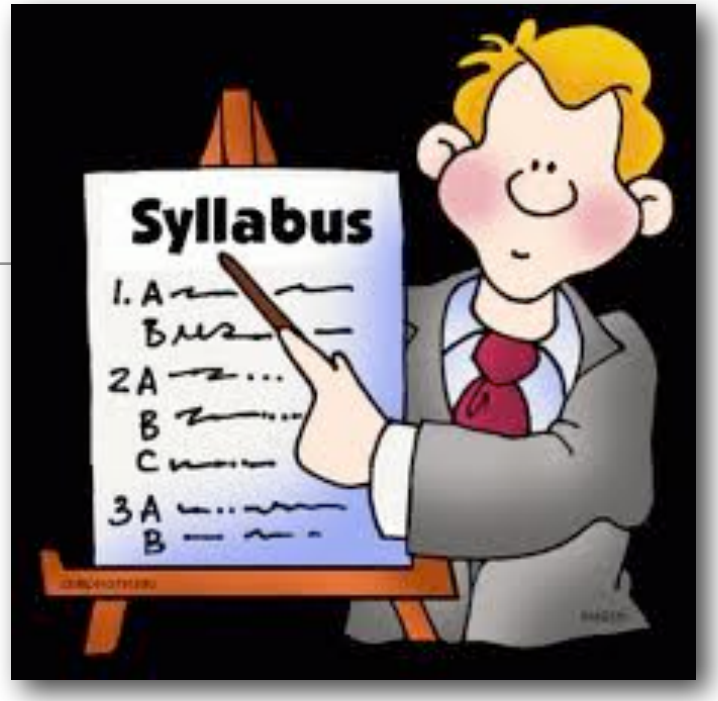
Course overview

Goals

- Identify key elements of data mining and machine learning algorithms
- Understand how algorithmic elements interact to impact performance
- Understand how to choose algorithms for different analysis tasks
- Analyze data in both an exploratory and targeted manner
- Implement and apply basic algorithms for supervised and unsupervised learning
- Accurately evaluate the performance of algorithms, as well as formulate and test hypotheses

Topics

- Elements of data science algorithms
    - Machine Learning
    - Data Mining
    - Statistics
- Statistical basics and background
- Data preparation and exploration
- Predictive modeling
- Methodology, evaluation
- Descriptive modeling



Logistics

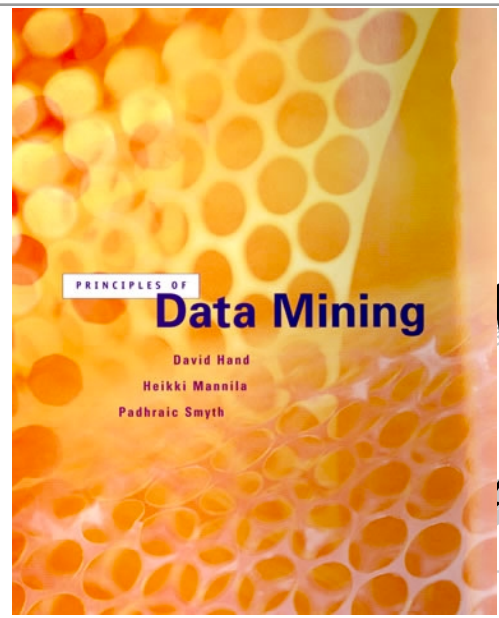
- Time and location: MWF 1:30-2:20pm, Grissom Hall 103
- Instructor: Bruno Ribeiro
  ribeiro@cs.purdue.edu, LWSN 2142C, office hours: M 12pm-1pm
- Teaching assistants: Israa Al-Qassem , Treavor Bonjour, Leonardo Teixeira, office hours:
- Mon 12:30pm - 1:30pm (Trevor)
- Thu 4:30pm - 5:30pm (Leo)
- Fri 3:00pm- 4:00pm (Israa)

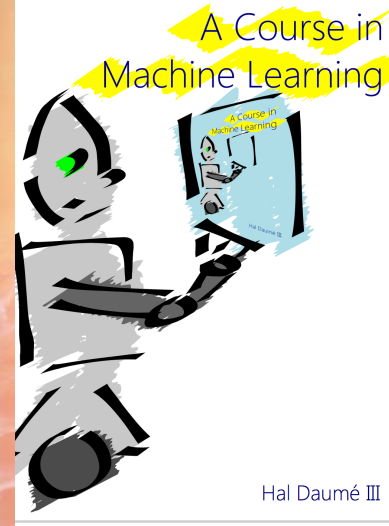
Lectures and homeworks
are password protected
Username: *cs373*
Password: *fall17p*

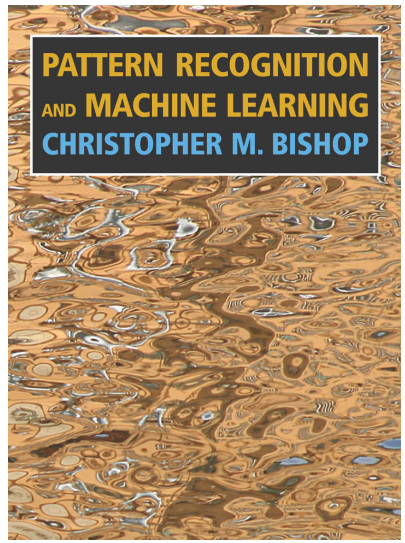
- Webpage: http://www.cs.purdue.edu/~ribeirob/courses/Fall2017
- Email list: fall-2017-cs-37300-le1@lists.purdue.edu
- Piazza signup: https://piazza.com/purdue/fall2017/cs37300
- Prerequisites: CS182, CS251
  Concurrent prerequisite: STAT350 or STAT511

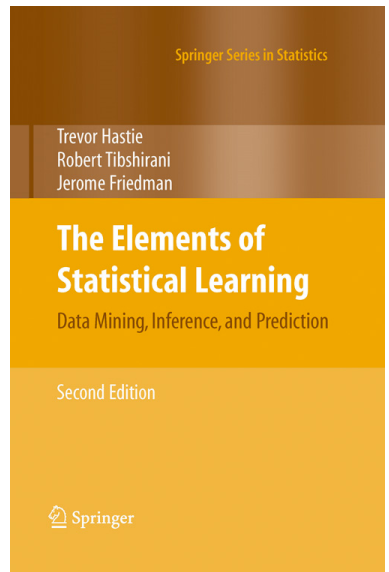
Readings

- No required text, readings will be announced/distributed on course webpage.
- Recommended texts
  - Principles of Data Mining (Free with PUID)
  - A Course in Machine Learning (Free Online)
  - Elements of Statistical Learning (Free Online)
  - Patterns Recognition and Machine Learning (Hard copy only)





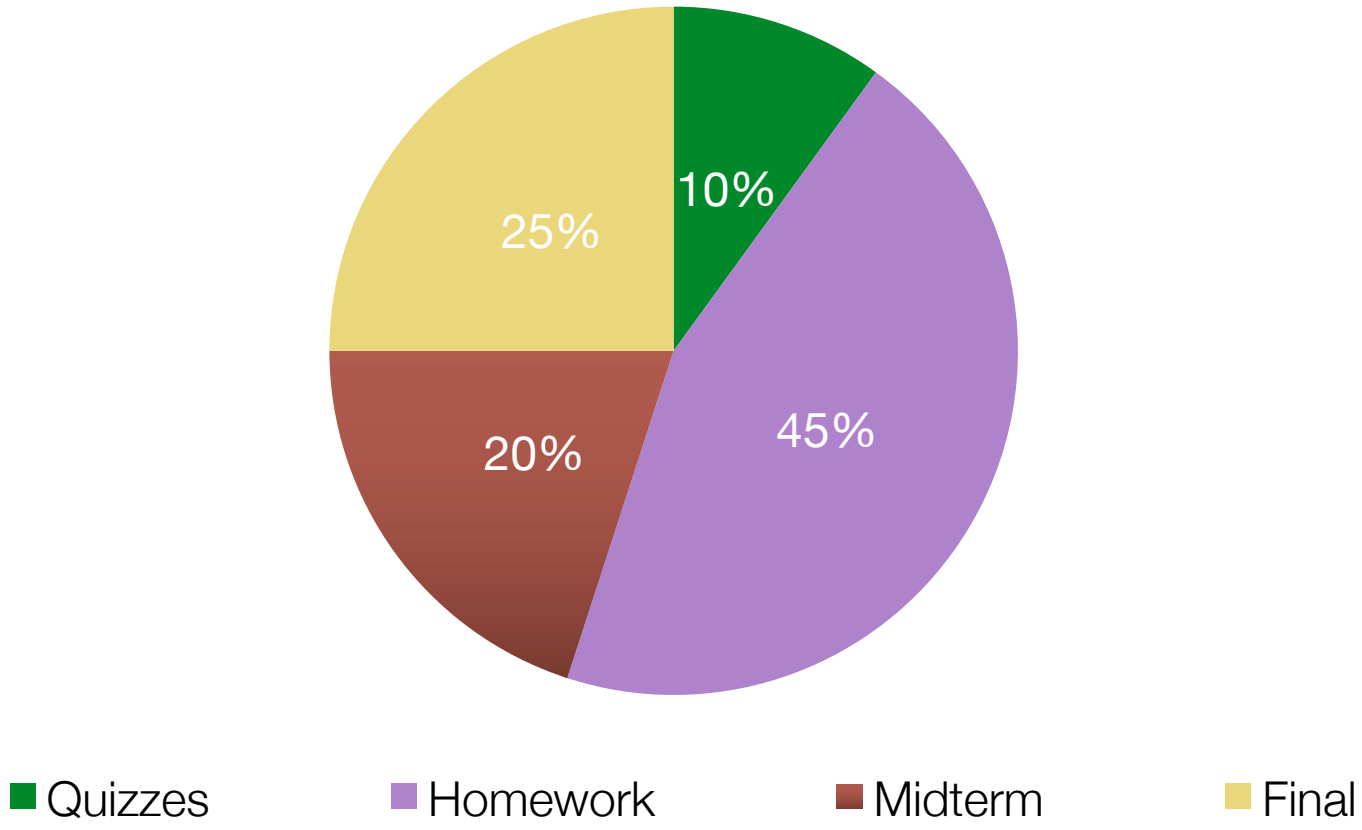




Workload

- Homeworks (7 assigments)
  - Seven assignments including written/math exercises, programming assignments in python
  - Late policy: No Late Homework (Grade = zero after deadline)
    - Submission on Blackboard
    - Firm deadline (Fridays 11:59pm)
    - On Blackboard we will give some time slack to account for acts of nature (e.g., WiFi not working) E.g.: Deadline is Friday 11:59pm but Blackboard says Saturday night (8pm).
      The deadlines are on FRIDAYS 11:59pm
      (the extra time is a buffer, don't count on it)
  - Only top 6 of 7 HW grades will count towards final HW grade average
- Exams
  - Midterm and final exam

Grading

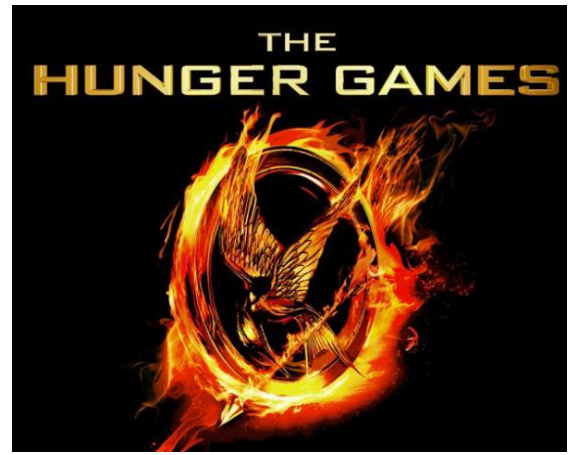


+5% extra credits given to the top 50% of our Kaggle competition

Kaggle Competition (up to +5% credit)

kaggle

CS373
Data Prediction Task



Task details soon

- +5% extra credit to the top 1%
- +4% extra credit to the 2%-10%
- +3% extra credit to the 11%-20%
- +2% extra credit to the 21%-30%
- +1% extra credit to the 41%-50%
- 0% extra to bottom <50%

Survey

1. Are you in the Machine Intelligence Track?

2. Background

    a) When did you take CS182 and CS251?

    b) What Stat course did you take and when?

    c) Are you familiar with Python?

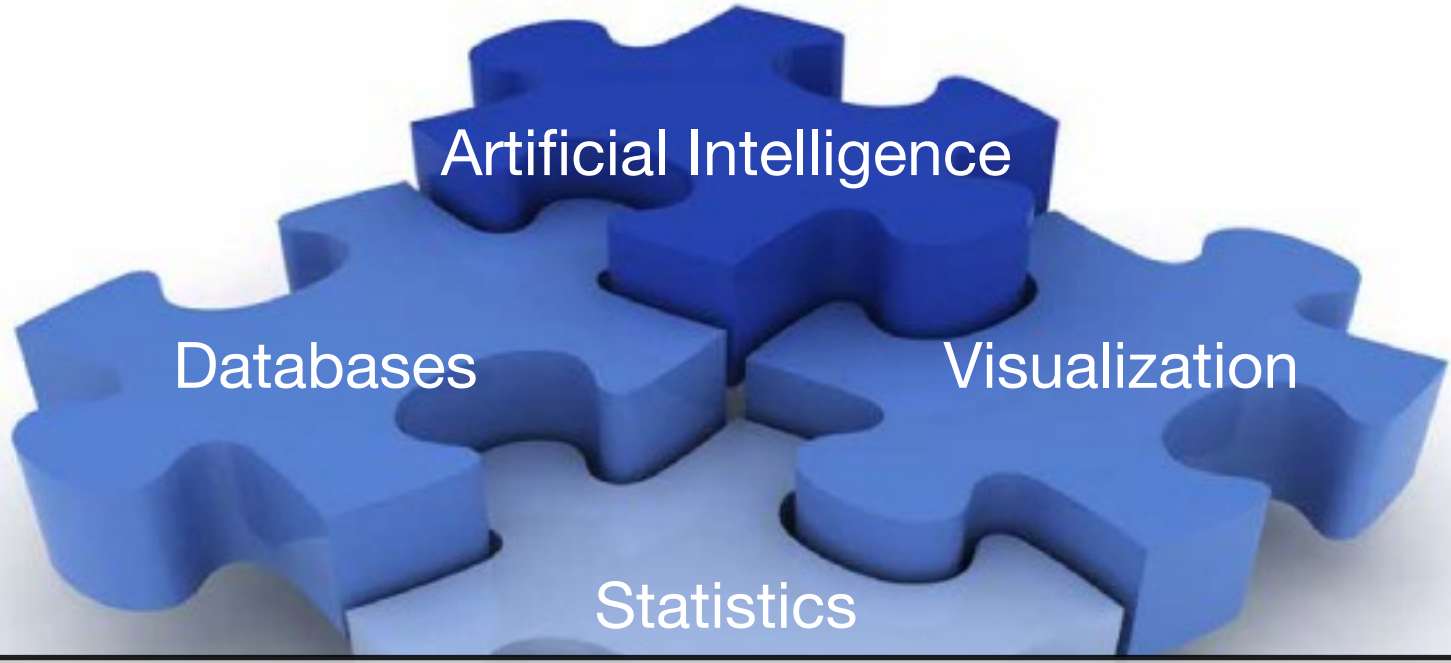
    Due by this Friday (8/25) 11:59pm

Course introduction

Data mining

The process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

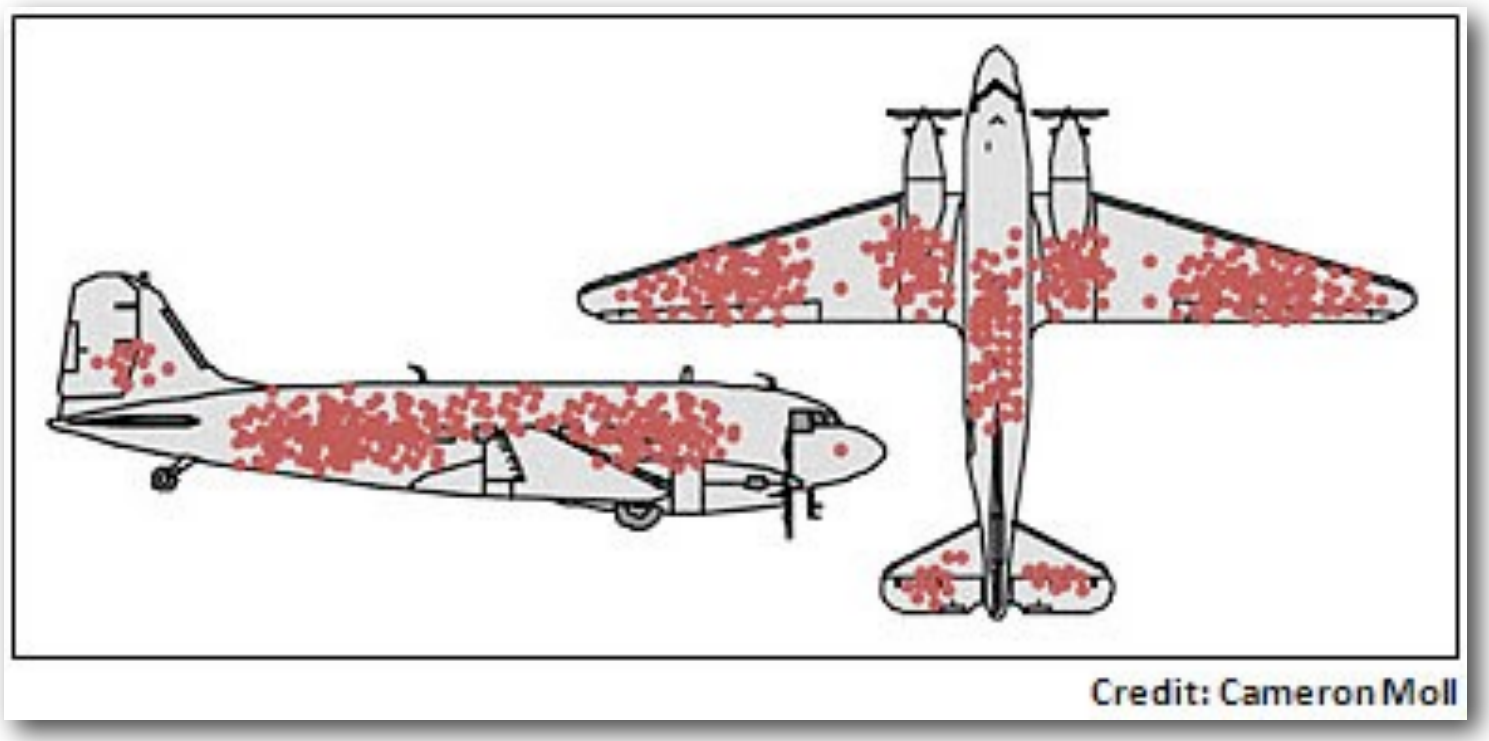
(Fayyad, Piatetsky-Shapiro & Smith 1996)



Machine learning: How can we build computer systems that automatically improve with experience? *(Mitchell 2006)*

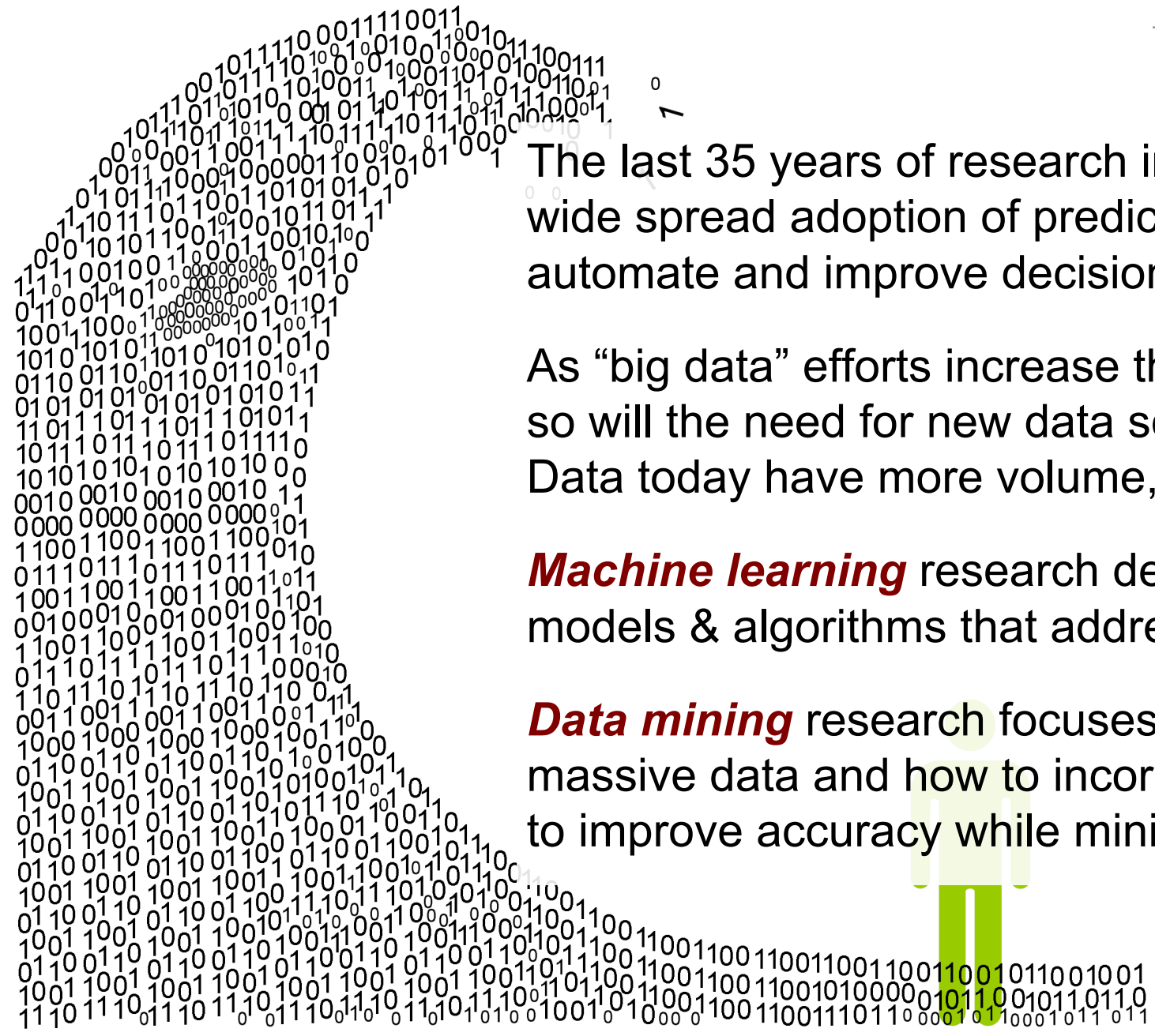
Example

Bullet holes of surviving airplanes



During WWII, statistician Abraham Wald was asked to help the British decide where to add armor to their planes

The data revolution



The last 35 years of research in ML/DM has resulted in wide spread adoption of predictive analytics to automate and improve decision making.

As “big data” efforts increase the collection of data… so will the need for new data science methodology. Data today have more volume, velocity, variety, etc.

Machine learning research develops statistical tools, models & algorithms that address these complexities.

Data mining research focuses on how to scale to massive data and how to incorporate feedback to improve accuracy while minimizing effort.

Bringing big data to the enterprise

#ibmbigdata

What is big data?

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data.**

The New York Times

Magazine

How Companies Learn Your Secrets

And among life events, none are more important than the arrival of a baby. At that moment, new parents' habits are more flexible than at almost any other time in their adult lives. If companies can identify pregnant shoppers, they can earn millions.

As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

Soon after the new ad campaign began, Target's Mom and Baby sales exploded. The company doesn't break out figures for specific divisions, but between 2002 — when Pole was hired — and 2010, Target's revenues grew from $44 billion to $67 billion. In 2005, the company's president, Gregg Steinhafel, boasted to a room of investors about the company's "heightened focus on items and categories that appeal to specific guest segments such as mom and baby."

Antonio Bolfo/Reportage for The New York Times

By CHARLES DUHIGG

Published: February 16, 2012 | 570 Comments

The data mining process

Data

Selection

Target data

Preprocessing

Processed data

Machine Learning

Knowledge

Interpretation evaluation

Patterns

Mining

Data mining process

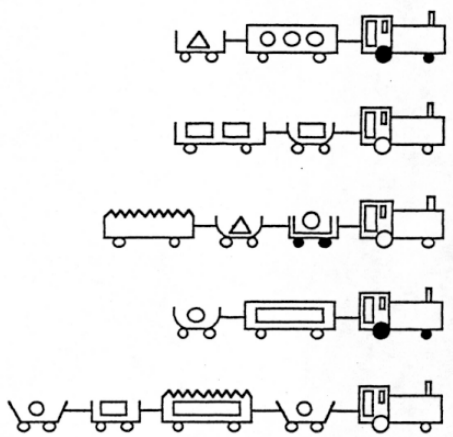
1. Application setup:
    - Acquire relevant domain knowledge
    - Assess user goals
2. Data selection
    - Choose data sources
    - Identify relevant attributes
    - Sample data
3. Data preprocessing
    - Remove noise or outliers
    - Handle missing values
    - Account for time or other changes
4. Data transformation
    - Find useful features
    - Reduce dimensionality

Data mining process

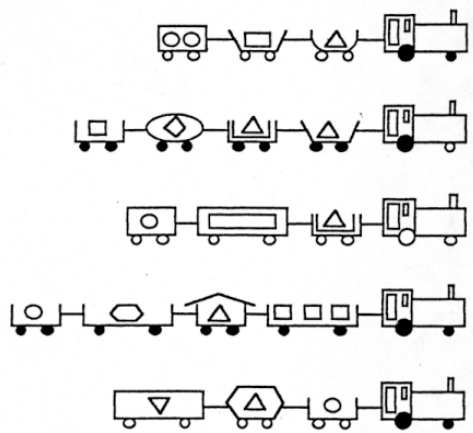
5. Data mining:
    - Choose task (e.g., classification, regression, clustering)
    - Choose algorithms for learning and inference
    - Set parameters
    - Apply algorithms to search for patterns of interest
6. Interpretation/evaluation
    - Assess accuracy of model/results
    - Interpret model for end-users
    - Consolidate knowledge
7. Repeat...

Example

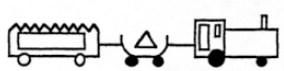
These trains carry toxic chemicals.



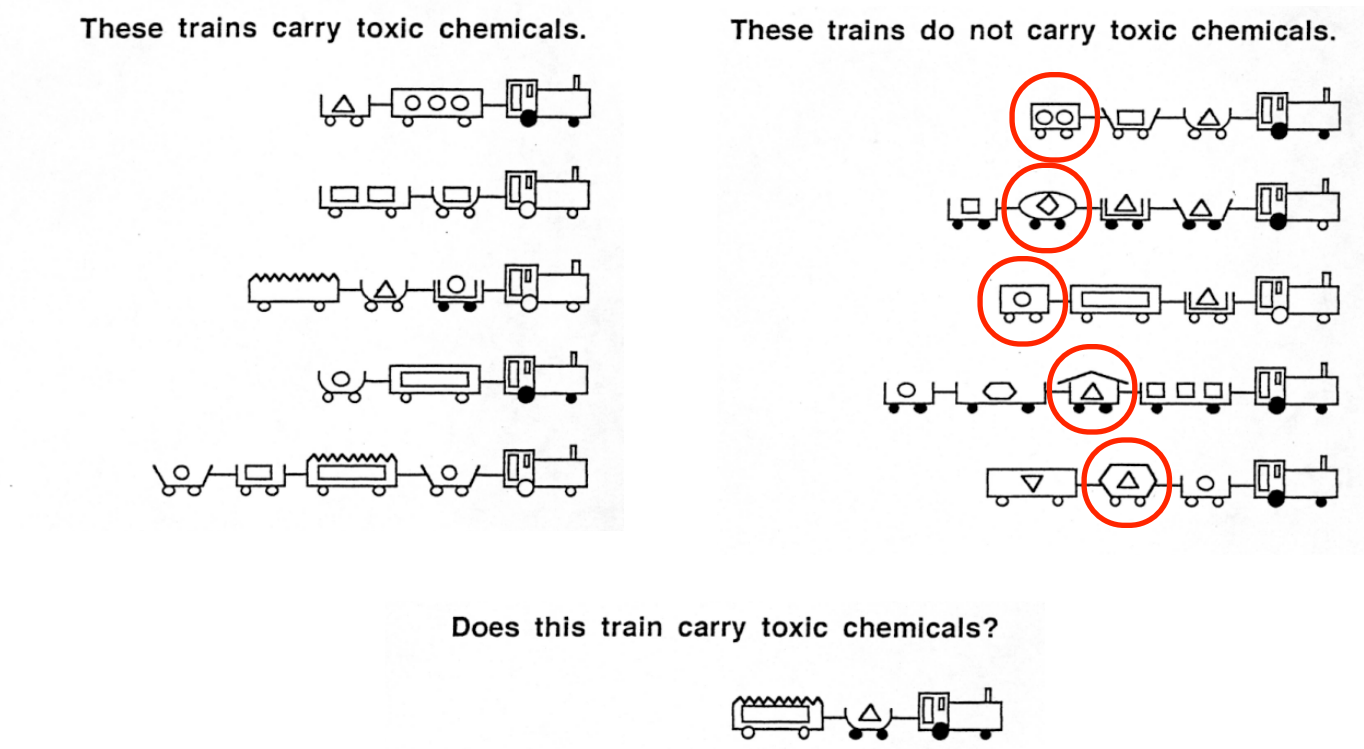
These trains do not carry toxic chemicals.



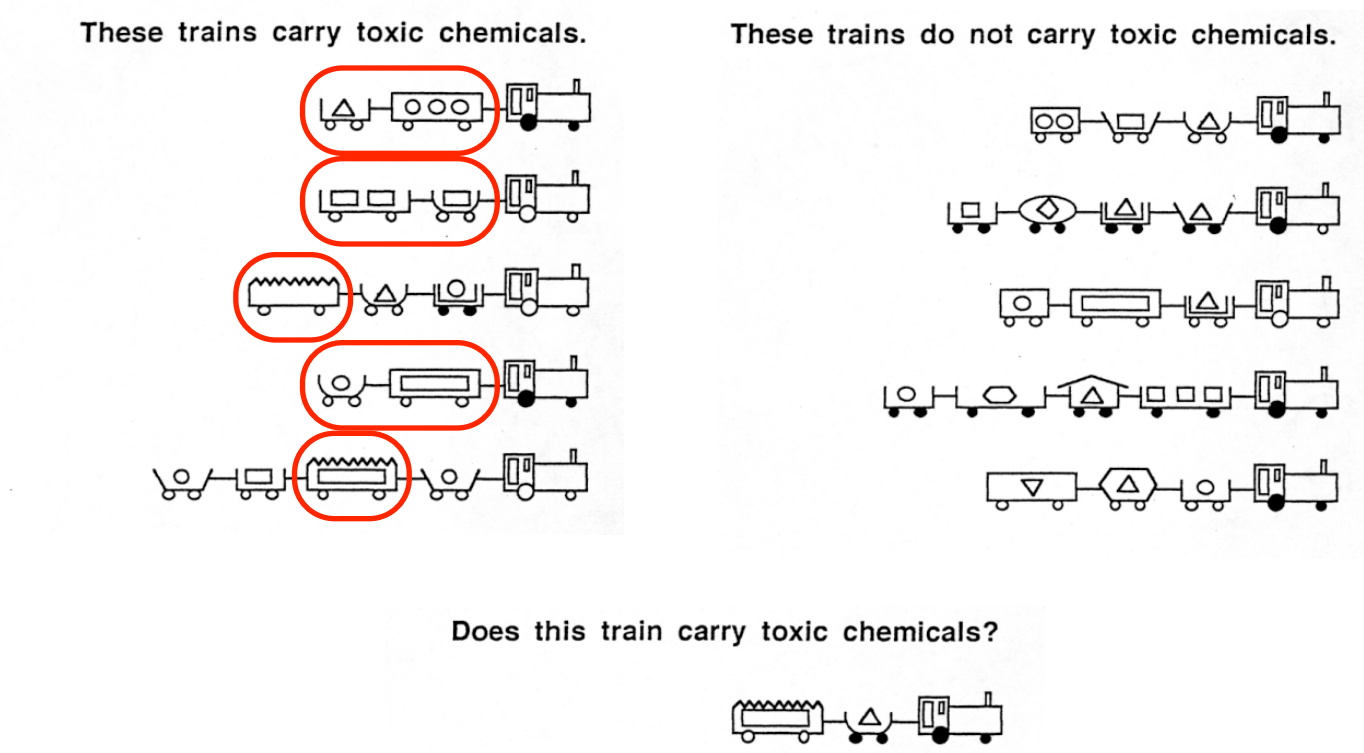
Does this train carry toxic chemicals?



Example rule (1)



Example rule (2)



How did you devise rules?

- Look for characteristics of one set but not the other?
- Reject potential rules that didn’t cover enough examples?
- Examine several potential rules?
- Consider simple rules first?

This is data mining...

- Data representation: Describe the data
- Task specification: Outline the goal(s)
- Knowledge representation: Describe the rules
- Learning technique:
  - Search: Identify a rule
  - Evaluation function: Estimate confidence
- Prediction technique: Apply the rule
- Data mining system: Do above in combination

Complexities

- Data size: vastly larger or changing rapidly
- Data representation: can affect ability to learn and interpret models
- Knowledge representation: needs to capture more subtle forms of probabilistic dependence
- Search space: vastly larger
- Evaluation functions: difficult to assess confidence in model utility

The data mining process

Data

Selection

Target data

Preprocessing

Processed data

Compute centralized view of the system from distributed logs

Define system states
Construct features
Identify training data

Knowledge

Interpretation evaluation

Patterns

Mining

Evaluate prediction lead time
Find new (predicted) failures in unlabeled data to explore

Formalize temporal prediction task
Avoid overfitting w/sparse models
Use ensemble over system states

Take-home quiz

TODO on Blacboard by Friday Aug 25th:
(1) Survey and (2) Take home quiz

Examples ***claims*** that are supported by data analysis from recent news articles:

- *The temperature of the planet is rising and the increase is due to human activities such as fossil fuel use and deforestation.*
- *Aspirin is effective in reducing cancer risk.*
- *Fathers who perform an equal share of household chores are more likely to have daughters who aspire to less traditionally feminine occupations.*

Due by this Friday (8/25) 11:59pm

Task: Identify three specific claims in news articles

1. Briefly state the claim
2. Describe the data that is (or could be) used to support the claim
3. Include a reference to the article

Length: One paragraph per claim.