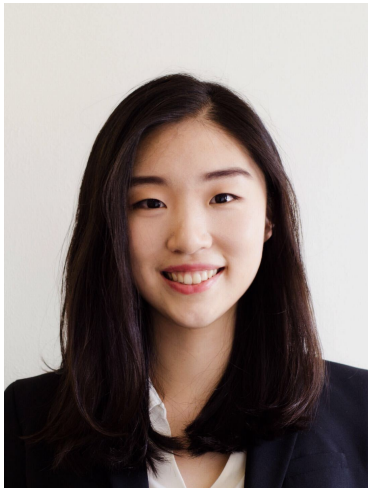
ORIE 4740: Statistical Data Mining

Damek Davis
School of ORIE, Cornell University
Lec 1 (Jan 25)

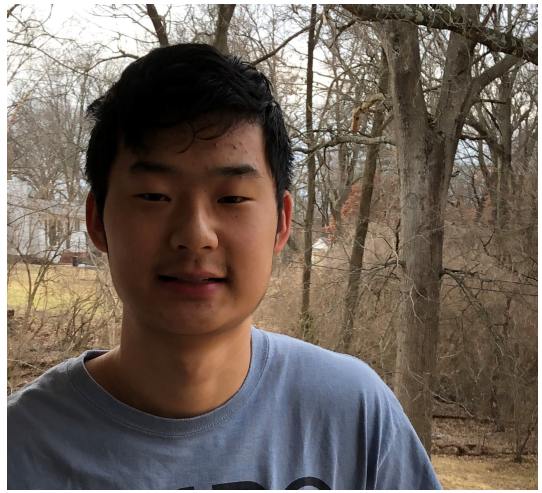
Our TAs



Duanduan Zhu

Tao Jiang

Tonghua Tian



Kevin Jiang

Some good friends

Doug

Piggy

Use iClicker

Check your email for signup instructions.

Use iClicker

Who is the dog in the picture:

A Doug

B Piggy

Data Mining

"[Data mining is] the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data... It employs pattern recognition technologies, as well as statistical and mathematical techniques."

– The Gartner Group

- Finding hidden and meaningful information in data.
- Often involves large data sets with many records/samples (e.g. customers) and many variables/attributes/features.

Data Mining

"[Data mining is] the process of discovering meaningful correlations, patterns, and trends by sifting through large amounts of data... It employs pattern recognition technologies, as well as statistical and mathematical techniques."

– The Gartner Group

- Finding hidden and meaningful information in data.
- Often involves large data sets with many records/samples (e.g. customers) and many variables/attributes/features.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Course goals

Course Goals:
To be able to take a large data set, decide on a set of data mining techniques depending on the question of interest, apply and compare those techniques, and draw conclusions.

- Handle large data sets
- Choose techniques based on your goals and the properties of the data.
- Understand the success/failure of an algorithm.
- Modify or extend an existing implementation of a data mining technique.
- Think like a statistician.

Review Syllabus

Available on

`canvas.cornell.edu`

Subject to changes (e.g., dates & times)
Email notification will be sent for important changes.

What is your major:

A : ORIE, System Engineering

B : Computer/Information Science

C : Statistics, Math, Physics

D : Business School

E : Other

You are a

A : Junior student

B : Senior student

C : Master student

D : Doctoral student

E : Other

Have you had a course that included multiple linear regression and logistic regression?

A. Yes

B. No

Have you used R, S-PLUS, Matlab?

A. Yes

B. No

Have you used C, Java, Python, or Julia?

A. Yes

B. No

Software

- R statistical software (open-source).
  - Implements many data mining and statistical algorithms. Has a large and growing set of additional packages.
  - Frequently used for statistical analysis and graphics in universities, industry and government
- RStudio: an integrated development environment (IDE).

Software

- R statistical software (open-source).
  - Implements many data mining and statistical algorithms. Has a large and growing set of additional packages.
  - Frequently used for statistical analysis and graphics in universities, industry and government
- RStudio: an integrated development environment (IDE).

Software

- R statistical software (open-source).
  - Implements many data mining and statistical algorithms. Has a large and growing set of additional packages.
  - Frequently used for statistical analysis and graphics in universities, industry and government
- RStudio: an integrated development environment (IDE).

Software

- R statistical software (open-source).
  - Implements many data mining and statistical algorithms. Has a large and growing set of additional packages.
  - Frequently used for statistical analysis and graphics in universities, industry and government
- RStudio: an integrated development environment (IDE).

Data Mining

Finding hidden and meaningful information in data.

Business Applications

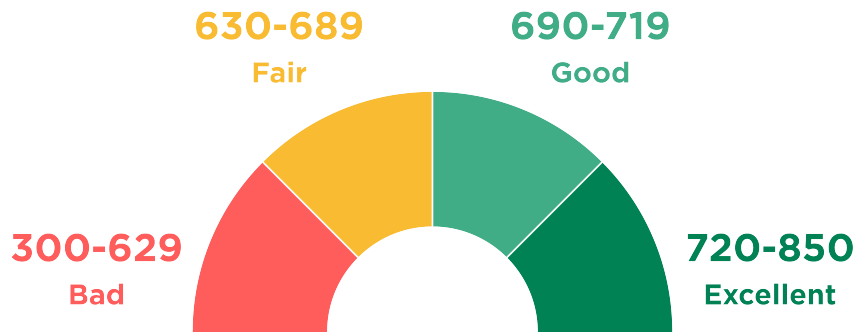
- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff
Tech
Welcome to The Not-So Private Parts where technology & privacy collide

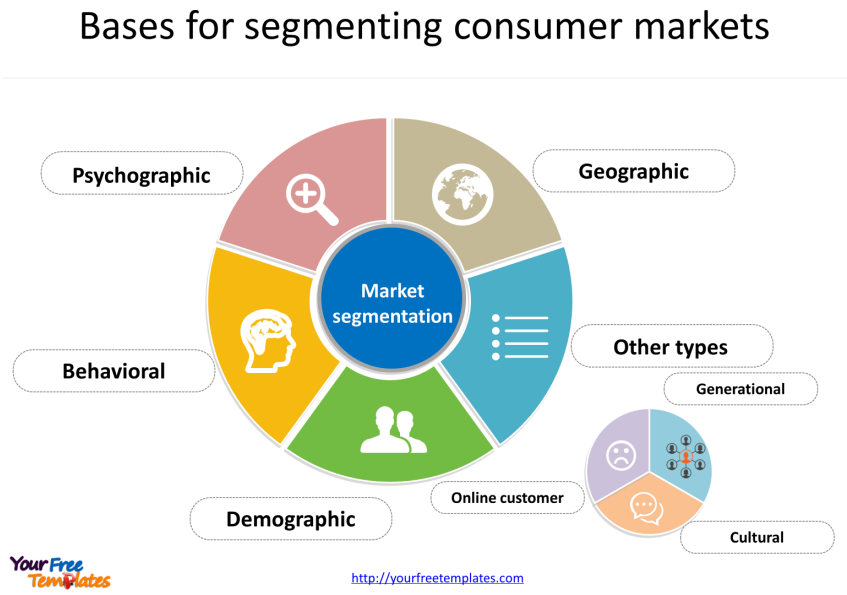
Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation



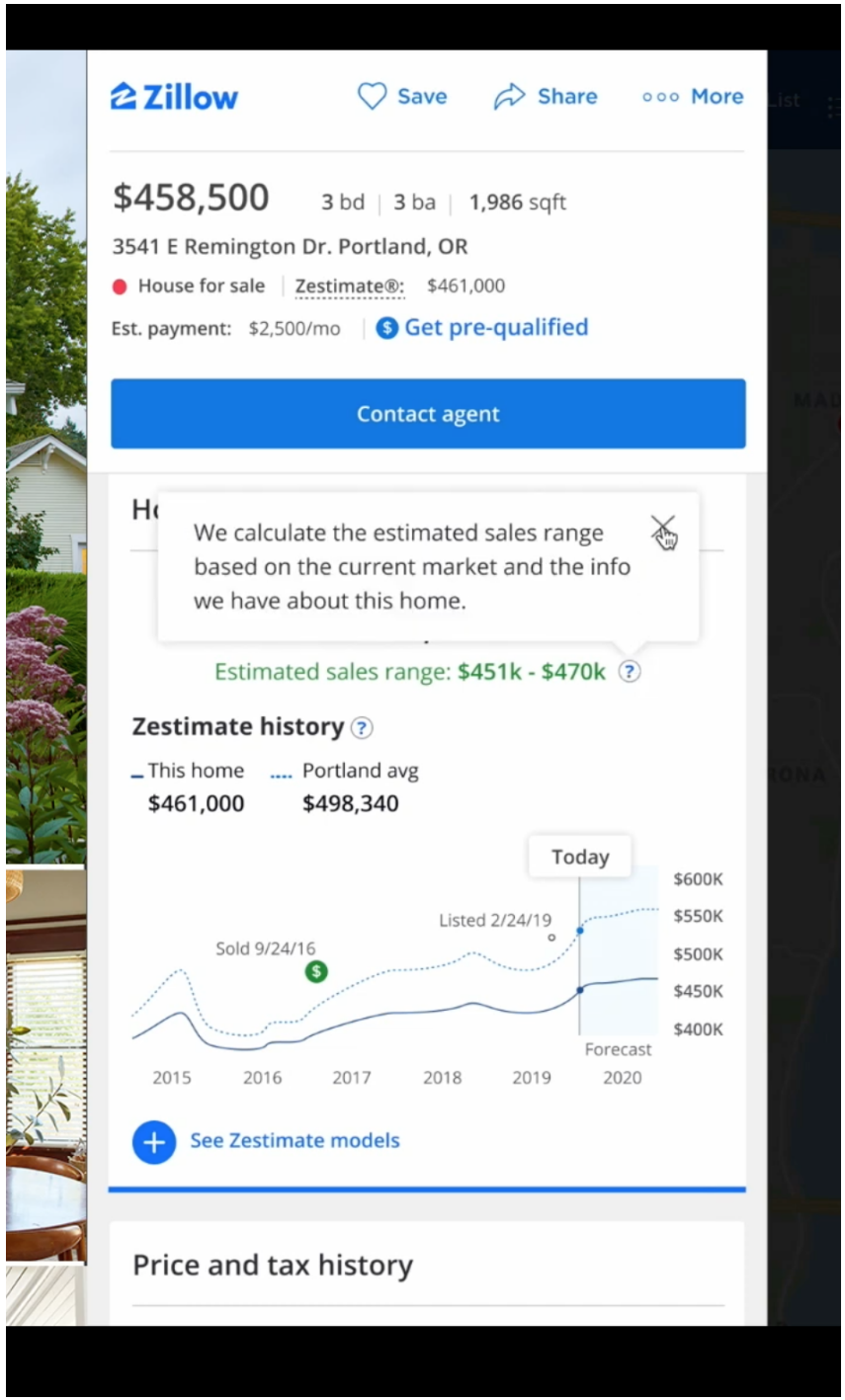
Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation



Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation



Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation

Press Releases

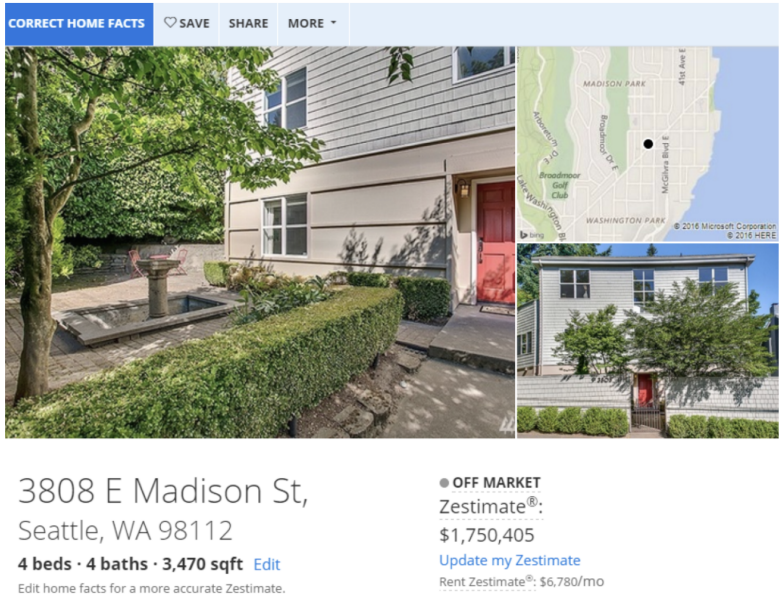
Zillow Starts Making Cash Offers For the Zestimate

15 years after the creation of the Zestimate, latest evolution has Zillow standing behind its trusted home valuation tool to guide initial cash offers on qualifying homes

Feb 25, 2021

Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation



Screen shot shows $1.75 million Zestimate of property formerly owned by Spencer Rascoff the day after the home sold for $1.05 million

Business Applications

- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation

After Zillow's Home-Flipping Fiasco, Think Twice About Trusting 'Zestimate' Home Values

By Samantha Sharf
November 3, 2021

SHARE

Business Applications

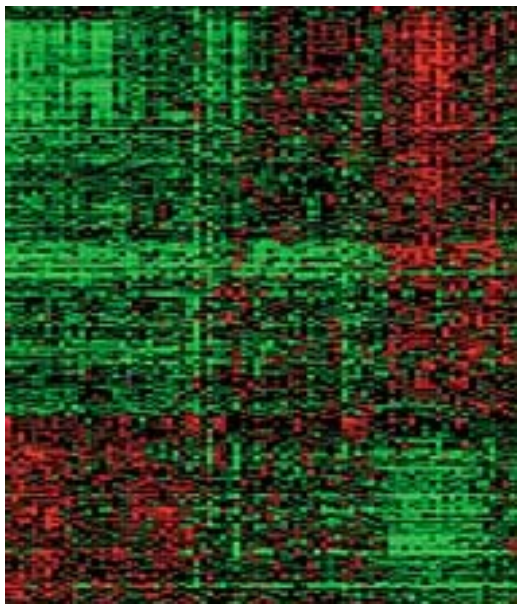
- Targeted marketing
  - Identifying likely purchasers
- Evaluating risk of credit card applicants
  - How likely to go 90 days past due on credit card?
- Market segmentation
  - Finding groups of customers with similar purchasing habits
- Home valuation

Zillow **announced** Tuesday that it is leaving the home-flipping business after three years. The decision raises questions not only about the company's future and the state of the housing market, but also about how good the company is at predicting home prices. The critical question for homebuyers and sellers: Can you trust the Zestimate?

Identification of Cancer Genes

Patient

Gene

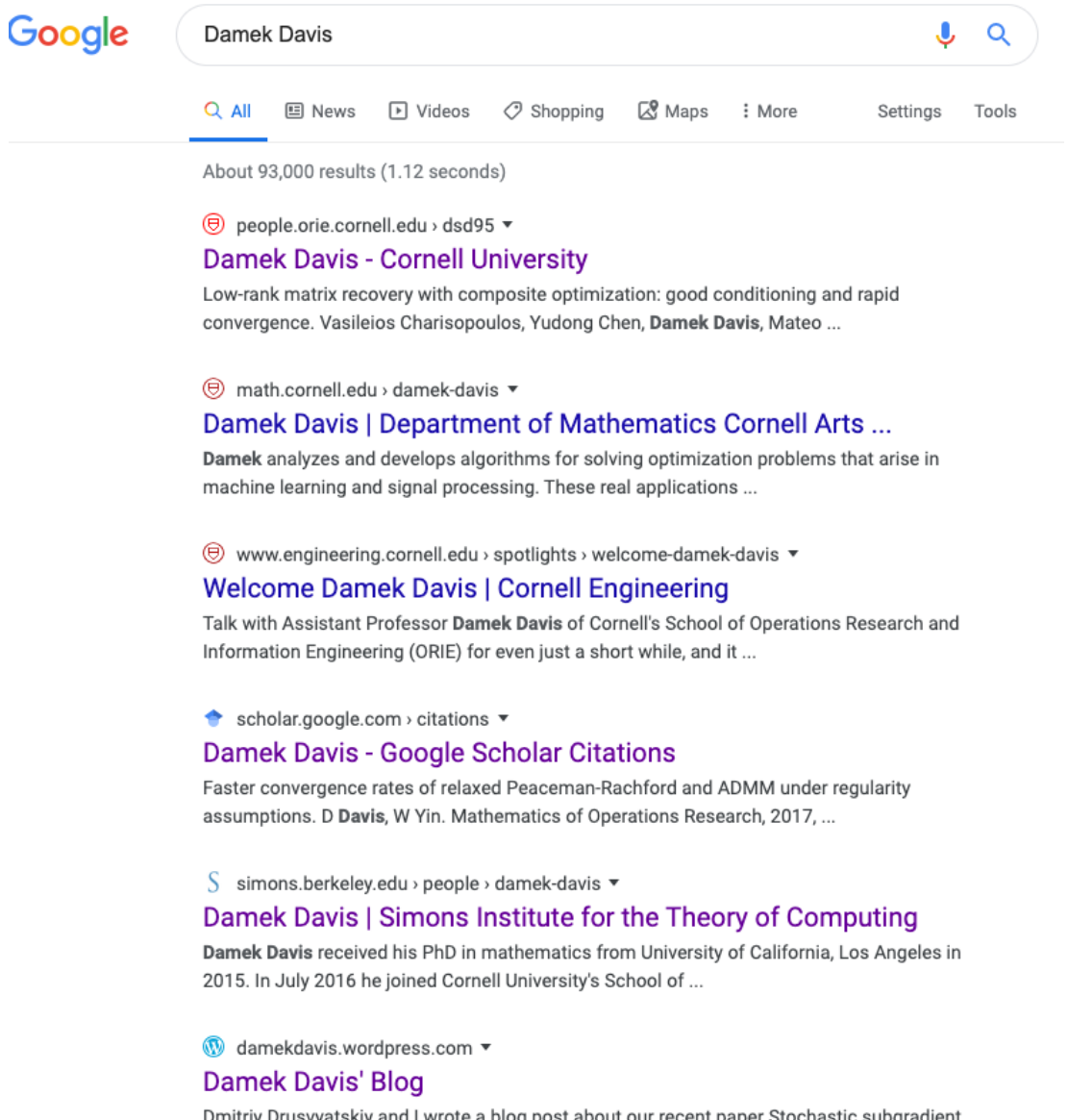


- Patient has a breast cancer
- High gene activity
- Low gene activity

- **(Similarity of columns)** which patients most resemble each other?
- **(Similarity of rows)** which genes most resemble each other?
- **(Feature selection)** for a given cancer, which genes have high expression? low expression?

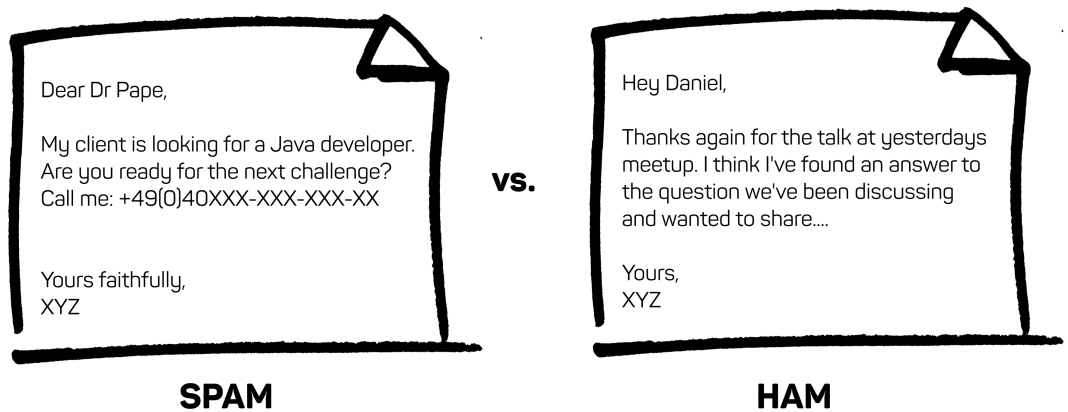
Information Tech. Applications

- Search engine
- Identifying spam
- Personalized online recommendation
- Smart web browsers that identify ads on web pages



Information Tech. Applications

- Search engine
- Identifying spam
- Personalized online recommendation
- Smart web browsers that identify ads on web pages



Information Tech. Applications

- Search engine
- Identifying spam
- Personalized online recommendation
- Smart web browsers that identify ads on web pages

Information Tech. Applications

- Search engine
- Identifying spam
- Personalized online recommendation
- Smart web browsers that identify ads on web pages

Supervised Learning

- Learn a rule for predicting the value of an outcome/response variable based on the value of some set of predictor variables.
  - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of "training" samples for which the predictors and outcome values are known.
- Example: classification
  - Outcome variable is a class (category).
  - E.g., predicting whether or not a Skype account is fraudulent based on the number of contact requests issued, number of contact requests accepted, and number of contact requests received.
  - Learn a good "classification rule" from the training data
  - Apply it to new data to predict the class.

Supervised Learning

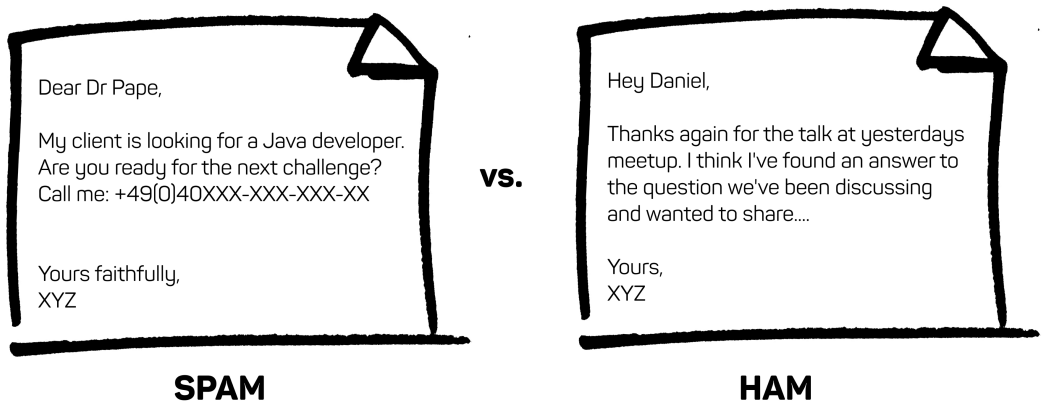
- Learn a rule for predicting the value of an outcome/response variable based on the value of some set of predictor variables.
  - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of "training" samples for which the predictors and outcome values are known.
- Example: classification
  - Outcome variable is a class (category).
  - E.g., predicting whether or not a Skype account is fraudulent based on the number of contact requests issued, number of contact requests accepted, and number of contact requests received.
  - Learn a good "classification rule" from the training data
  - Apply it to new data to predict the class.

Supervised Learning

- Learn a rule for predicting the value of an outcome/response variable based on the value of some set of predictor variables.
  - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of "training" samples for which the predictors and outcome values are known.
- Example: classification
  - Outcome variable is a class (category).
  - E.g., predicting whether or not a Skype account is fraudulent based on the number of contact requests issued, number of contact requests accepted, and number of contact requests received.
  - Learn a good "classification rule" from the training data
  - Apply it to new data to predict the class.

Supervised Learning

- Learn a rule for predicting the value of an outcome/response variable based on the value of some set of predictor variables.
  - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of "training" samples for which the predictors and outcome values are known.
- Example: classification



- Outcome variable is a class (category).
- E.g., predicting whether or not a Skype account is fraudulent based on the

Supervised Learning

- Learn a rule for predicting the value of an outcome/response variable based on the value of some set of predictor variables.
  - e.g. predicting house sale price using square footage, number of BRs, etc.
- Have a set of "training" samples for which the predictors and outcome values are known.
- Example: classification
  - Outcome variable is a class (category).
  - E.g., predicting whether or not a Skype account is fraudulent based on the number of contact requests issued, number of contact requests accepted, and number of contact requests received.
  - Learn a good "classification rule" from the training data
  - Apply it to new data to predict the class.

Supervised Learning: Evaluating Accuracy

- How would you evaluate the predictive accuracy of a classification rule on new (test) data?

Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers
- How might one evaluate accuracy for this type of decision rule?

Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers

- How might one evaluate accuracy for this type of decision rule?

Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers
- How might one evaluate accuracy for this type of decision rule?

Unsupervised Learning

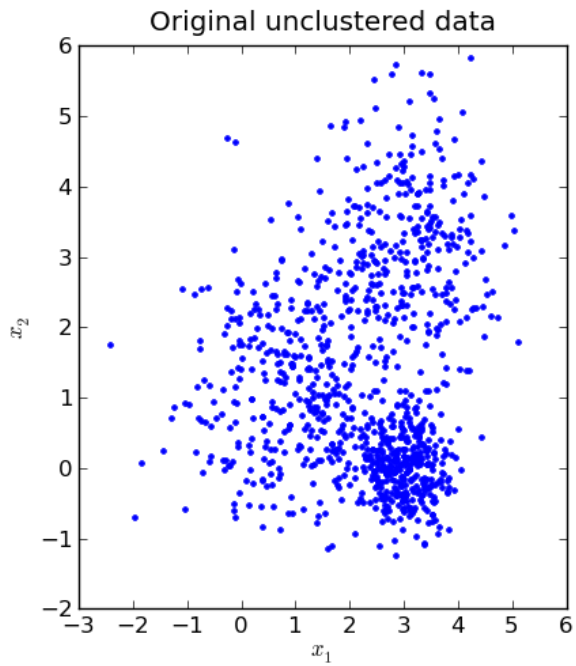
Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers
- How might one evaluate accuracy for this type of decision rule?

Unsupervised Learning

Unsupervised Learning:

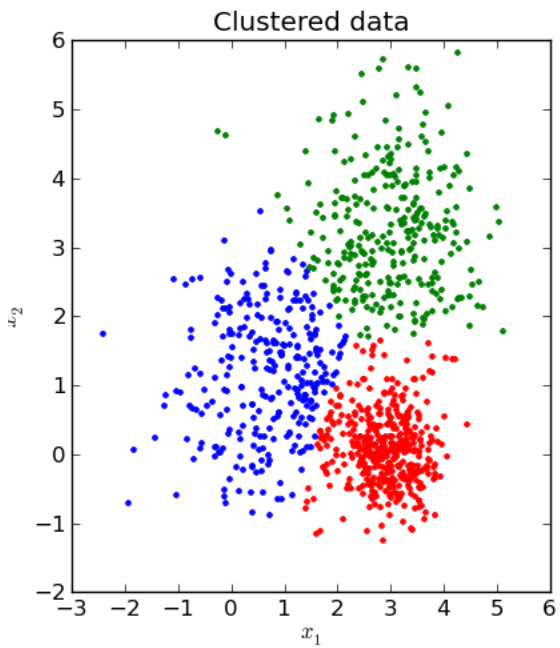
- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers



Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers



Unsupervised Learning

Unsupervised Learning:

- Sometimes the outcome variable is unknown for the training data
- E.g., Categorizing customers into groups with similar purchasing habits
  - The training data is the purchasing data of a set of customers
  - Learn a decision rule for categorizing new customers
- How might one evaluate accuracy for this type of decision rule?

Supervised or Unsupervised?

A child points at a bike and its parent points and says “bike." The child then repeats “bike."

Choose one:

A. Supervised learning

B. Unsupervised learning

Supervised or Unsupervised?

A child looks through photographs. There are pictures of houses and pictures of bikes. The child makes two stacks of photos, organized by how they look.

Choose one:

A. Supervised learning

B. Unsupervised learning

Supervised or Unsupervised?

Say we want to estimate the amount of $ an individual spends annually on fabric softener, based on their demographics (age, income, location of residence, etc.). The training data consist of measurements of the demographic variables and self-reported amount of $ spent annually on fabric softener, for 6,000 individuals.

Choose one:

A. Supervised learning

B. Unsupervised learning

Supervised or Unsupervised?

We want to reduce the dimension of our data, while preserving as much information as possible. E.g., want to create a single numeric summary of financial status, based on the variables: income, savings, potential future earnings, and credit rating (reduces dimension from 4 to 1). The training data consist of these 4 financial measurements for 200,000 customers.

Choose one:

A. Supervised learning

B. Unsupervised learning

Supervised or Unsupervised?

We want to group power generation facilities into groups of similar facilities, using a set of operational variables:

- power output per year
- generation capacity
- % time fully utilized
- type of fuel.

The training data consist of these operational measurements, for 1,700 facilities.

Choose one:

A. Supervised learning

B. Unsupervised learning

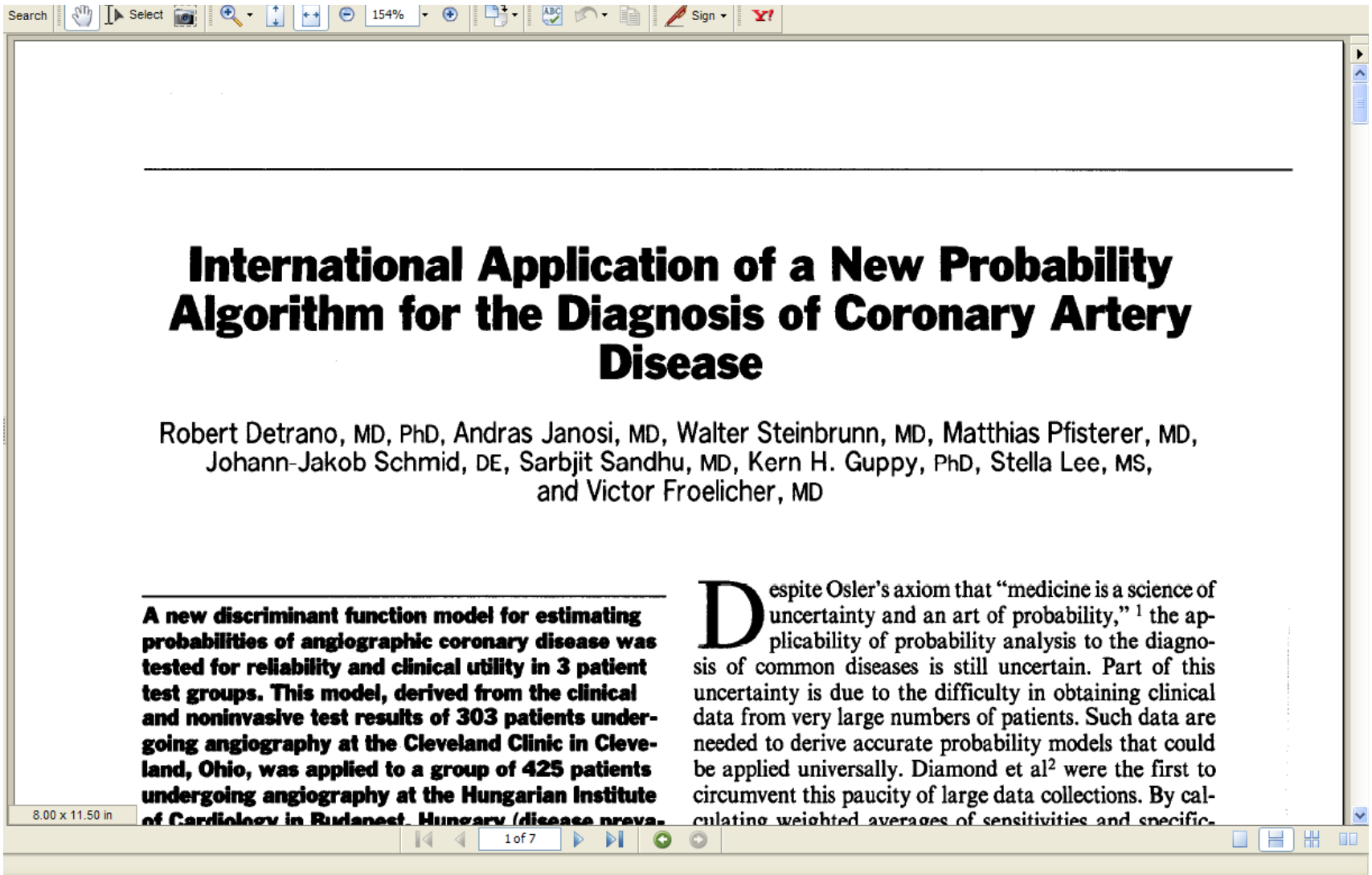
Heart Disease Detection

- Can patients be effectively screened for the presence of heart disease (CAD) without the use of angiography?
- Angiography is an invasive and expensive procedure where a tube is inserted into the artery of concern.

Heart Disease Detection

- Can patients be effectively screened for the presence of heart disease (CAD) without the use of angiography?
- Angiography is an invasive and expensive procedure where a tube is inserted into the artery of concern.

Heart Disease Detection



Search Select 154% Sign

International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease

Robert Detrano, MD, PhD, Andras Janosi, MD, Walter Steinbrunn, MD, Matthias Pfisterer, MD, Johann-Jakob Schmid, DE, Sarbjit Sandhu, MD, Kern H. Guppy, PhD, Stella Lee, MS, and Victor Froelicher, MD

A new discriminant function model for estimating probabilities of angiographic coronary disease was tested for reliability and clinical utility in 3 patient test groups. This model, derived from the clinical and noninvasive test results of 303 patients undergoing angiography at the Cleveland Clinic in Cleveland, Ohio, was applied to a group of 425 patients undergoing angiography at the Hungarian Institute of Cardiology in Budapest, Hungary (disease preva-

Despite Osler's axiom that "medicine is a science of uncertainty and an art of probability," [1] the applicability of probability analysis to the diagnosis of common diseases is still uncertain. Part of this uncertainty is due to the difficulty in obtaining clinical data from very large numbers of patients. Such data are needed to derive accurate probability models that could be applied universally. Diamond et al[2] were the first to circumvent this paucity of large data collections. By calculating weighted averages of sensitivities and specific-

8.00 x 11.50 in 1 of 7

Heart Disease Detection

- The authors use data from the Cleveland Clinic. The data has non-invasive clinical test results as well as the actual presence/absence of the heart disease (CAD) for 303 patients.
- They learn a classification rule for predicting CAD based on the non-invasive test results.
- This classification rule uses logistic regression.
- They check the predictive accuracy of their classification rule on data from patients in Hungary (74%) and California (77%).

Heart Disease Detection

- The authors use data from the Cleveland Clinic. The data has non-invasive clinical test results as well as the actual presence/absence of the heart disease (CAD) for 303 patients.
- They learn a classification rule for predicting CAD based on the non-invasive test results.
- This classification rule uses logistic regression.
- They check the predictive accuracy of their classification rule on data from patients in Hungary (74%) and California (77%).

Heart Disease Detection

- The authors use data from the Cleveland Clinic. The data has non-invasive clinical test results as well as the actual presence/absence of the heart disease (CAD) for 303 patients.
- They learn a classification rule for predicting CAD based on the non-invasive test results.
- This classification rule uses logistic regression.
- They check the predictive accuracy of their classification rule on data from patients in Hungary (74%) and California (77%).

Heart Disease Detection

- The authors use data from the Cleveland Clinic. The data has non-invasive clinical test results as well as the actual presence/absence of the heart disease (CAD) for 303 patients.
- They learn a classification rule for predicting CAD based on the non-invasive test results.
- This classification rule uses logistic regression.
- They check the predictive accuracy of their classification rule on data from patients in Hungary (74%) and California (77%).

Heart Disease Detection

What type of learning is this?

A. Supervised

B. Unsupervised

Heart Disease Detection

- The goal is to learn a good classification rule to predict the presence / absence of CAD from the 13 predictors in the data set:
  - age
  - assigned sex
  - chest pain type
  - blood pressure
  - Number of vessels showing calcium on fluoroscopy
  - exercise thallium scintigraphic defects (fixed, reversible, none)
  - electrocardiogram results
  - exercise-induced angina (presence / absence)
  - etc.

Heart Disease Detection

The Cleveland data:

```
70.0 1.0 4.0 130.0 322.0 0.0 2.0 109.0 0.0 2.4 2.0 3.0 3.0 2
67.0 0.0 3.0 115.0 564.0 0.0 2.0 160.0 0.0 1.6 2.0 0.0 7.0 1
57.0 1.0 2.0 124.0 261.0 0.0 0.0 141.0 0.0 0.3 1.0 0.0 7.0 2
64.0 1.0 4.0 128.0 263.0 0.0 0.0 105.0 1.0 0.2 2.0 1.0 7.0 1
74.0 0.0 2.0 120.0 269.0 0.0 2.0 121.0 1.0 0.2 1.0 1.0 3.0 1
65.0 1.0 4.0 120.0 177.0 0.0 0.0 140.0 0.0 0.4 1.0 0.0 7.0 1
56.0 1.0 3.0 130.0 256.0 1.0 2.0 142.0 1.0 0.6 2.0 1.0 6.0 2
59.0 1.0 4.0 110.0 239.0 0.0 2.0 142.0 1.0 1.2 2.0 1.0 7.0 2
60.0 1.0 4.0 140.0 293.0 0.0 2.0 170.0 0.0 1.2 2.0 2.0 7.0 2
63.0 0.0 4.0 150.0 407.0 0.0 2.0 154.0 0.0 4.0 2.0 3.0 7.0 2
59.0 1.0 4.0 135.0 234.0 0.0 0.0 161.0 0.0 0.5 2.0 0.0 7.0 1
53.0 1.0 4.0 142.0 226.0 0.0 2.0 111.0 1.0 0.0 1.0 0.0 7.0 1
44.0 1.0 3.0 140.0 235.0 0.0 2.0 180.0 0.0 0.0 1.0 0.0 3.0 1
61.0 1.0 1.0 134.0 234.0 0.0 0.0 145.0 0.0 2.6 2.0 2.0 3.0 2
57.0 0.0 4.0 128.0 303.0 0.0 2.0 159.0 0.0 0.0 1.0 1.0 3.0 1
71.0 0.0 4.0 112.0 149.0 0.0 0.0 125.0 0.0 1.6 2.0 0.0 3.0 1
46.0 1.0 4.0 140.0 311.0 0.0 0.0 120.0 1.0 1.8 2.0 2.0 7.0 2
53.0 1.0 4.0 140.0 203.0 1.0 2.0 155.0 1.0 3.1 3.0 0.0 7.0 2
64.0 1.0 1.0 110.0 211.0 0.0 2.0 144.0 1.0 1.8 2.0 0.0 3.0 1
40.0 1.0 1.0 140.0 199.0 0.0 0.0 178.0 1.0 1.4 1.0 0.0 7.0 1
```

Heart Disease Detection

The meta-data:

```
-------------------------
      -- 1. age
      -- 2. sex
      -- 3. chest pain type  (4 values)
      -- 4. resting blood pressure
      -- 5. serum cholestoral in mg/dl
      -- 6. fasting blood sugar > 120 mg/dl
      -- 7. resting electrocardiographic results  (values 0,1,2)
      -- 8. maximum heart rate achieved
      -- 9. exercise induced angina
      -- 10. oldpeak = ST depression induced by exercise relative to res
      -- 11. the slope of the peak exercise ST segment
      -- 12. number of major vessels (0-3) colored by flourosopy
      -- 13.  thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Attributes types
----------------

Real: 1,4,5,8,10,12
Ordered:11,
Binary: 2,6,9
Nominal:7,3,13
```

Heart Disease Detection

- We'll come up with our own classification rule for this problem

Heart Disease Detection

For example, we may use a classification tree: