DATA SCI 415: Overview

Snigdha Panigrahi

Department of Statistics, University of Michigan

What's in the name?

- Our course is called "Data mining and Statistical Learning"
- Our book is called "Statistical Learning"
- Other names you hear: machine learning, artificial intelligence, data science, analytics... and of course statistics!
- The differences are mostly historical and cultural; by and large, these fields solve the same problems, but may sometimes differ in their focus.

Statistical Learning from Big Data

- Fact: The amount of data collected and stored is exponentially increasing, due to advances in data collection, computerization of many aspects of life and breakthroughs in technology.

  Last time: data generated by GAI

- Consequence: Data analysis problems have dramatically increased in size and complexity.

- Your future job: know how to make sense of all these data!

Statistical Learning from Big Data

- Your future job: make sense of all these data!
- Identify patterns and trends: uncover "interesting" relationships among the variables and/or the observations

  Replication crisis: Not all of them are real!!
- Predict future behavior
- Attach uncertainties to patterns and make inferences
- Your future job: know how to make sense of all these data! Convert uncovered patterns to knowledge

- Technology helps
  - Faster computers, more storage $\Rightarrow$ more flexible and thus more powerful techniques $\Rightarrow$ fewer modeling assumptions
  - New visualisation capabilities (a picture is worth a thousand words...)
- But not always
  - Some problems are inherently computationally intractable
  - "Easy" black-box data analysis can lead to flexible modeling: a lot of misuse and misunderstanding

    A famous example: Google Flu (predict flu prevalence)

    GF failed at the peak of the 2013 flu season by 140 percent

    Why??

- Technology helps
  - Faster computers, more storage $\Rightarrow$ more flexible and thus more powerful techniques $\Rightarrow$ fewer modeling assumptions
  - New visualisation capabilities (a picture is worth a thousand words...)
- But not always
  - Some problems are inherently computationally intractable
  - "Easy" black-box data analysis can lead to flexible modeling: a lot of misuse and misunderstanding

    A famous example: Google Flu (predict flu prevalence)

    GF failed at the peak of the 2013 flu season by 140 percent

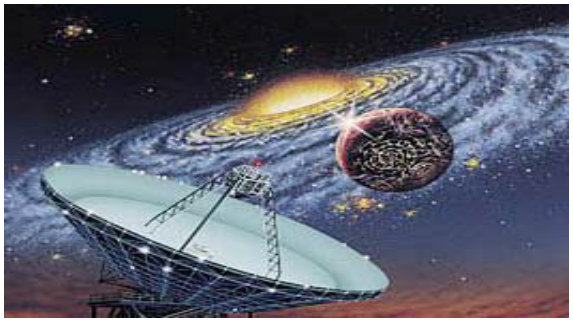
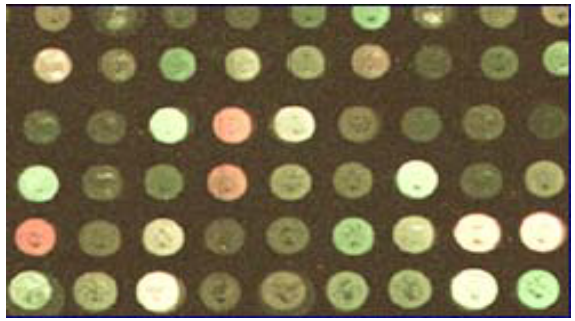
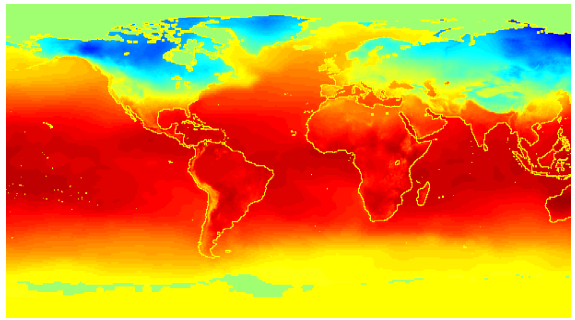
    Why?? Flexible models can overfit (too much of a good thing)

Understanding underlying assumptions and interpreting conclusions correctly remains as important as ever

Statistical Learning from Big Data: Motivation

- There is often "hidden" information in the data that is not readily evident
- Human analysts without large-scale algorithms may never discover that useful information
- Much of the data available are never analyzed at all

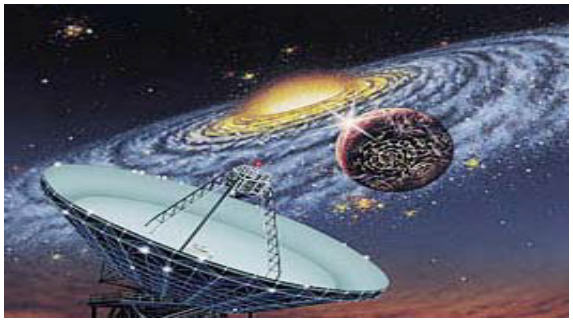
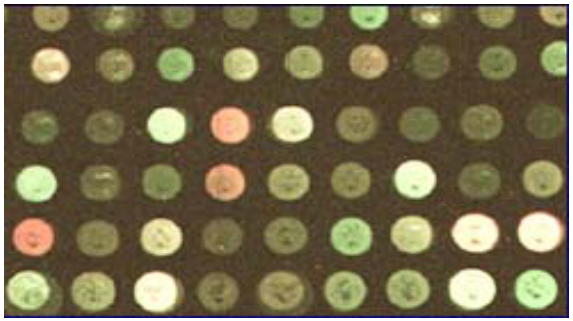
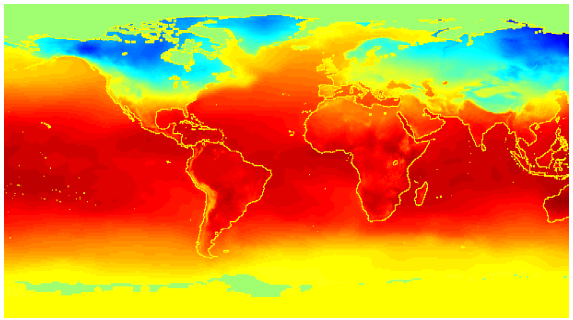
Statistical Learning from data: benefits to science

Data are collected and stored at enormous speeds

- remote sensors on a satellite (NASA)
- telescopes scanning the skies (SDSS)
- microarrays generating gene expression data (MEDLINE)
- synthetic data by generative AI

Statistical Learning from data: benefits to science

Statistical learning helps scientists with

- classifying and clustering data
- formulating hypotheses/theories
- validating hypotheses/theories
- predicting future behavior

Statistical Learning from data: benefits to business

Almost any commercial transaction generates data

- Web searches, social connections (Google, Facebook, Twitter, etc)
- Purchases both online and at stores (Amazon, eBay, Walmart)
- Bank/credit card transactions (Bank of America, Visa, Mastercard)

Statistical Learning from data: benefits to business

- Customized services and successful advertizing give competitive edge
- Have to balance useful services vs annoyance and privacy concerns: a fine line!
- Fair to customers: ensure inclusive search and recommendations

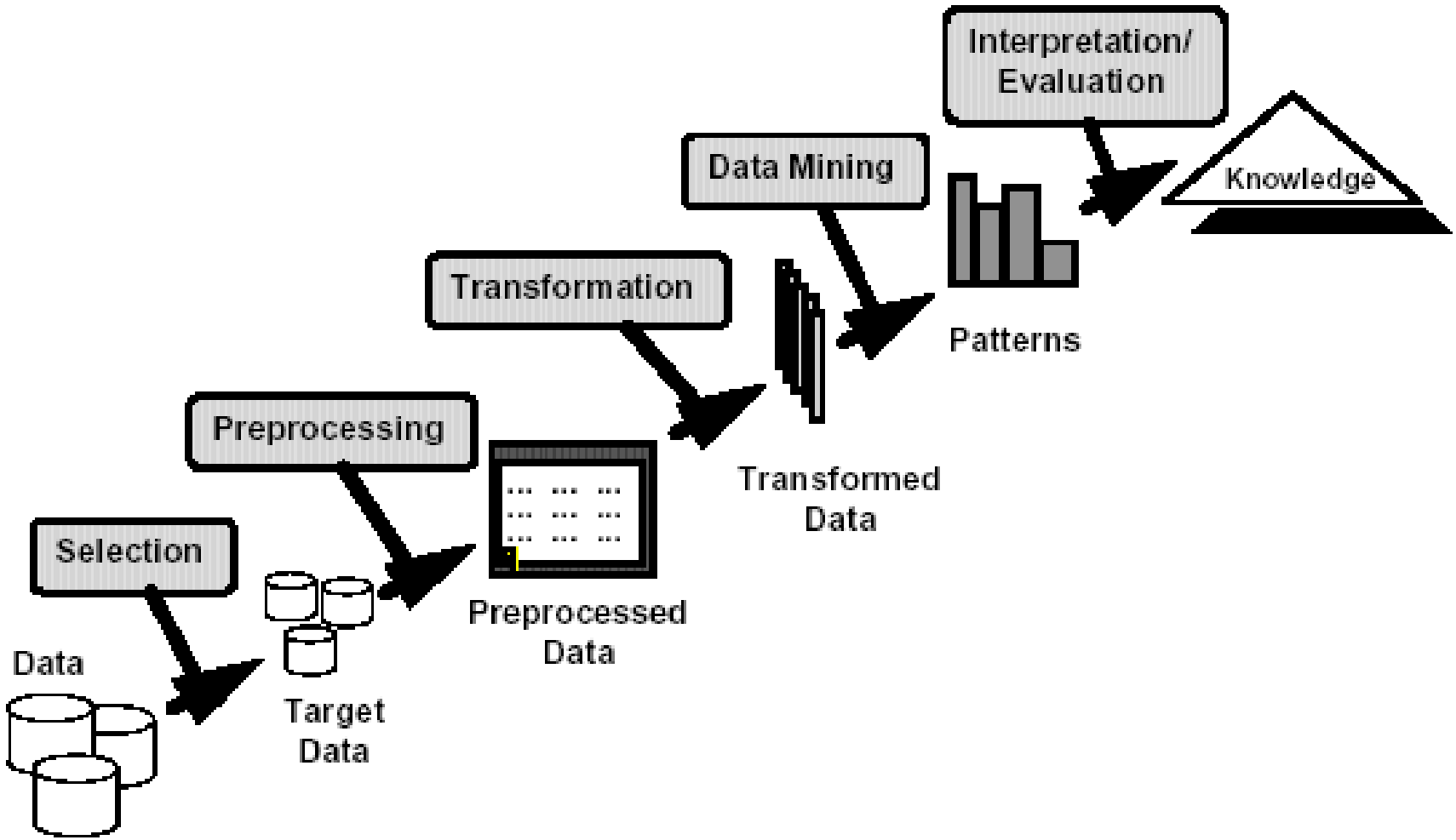
Statistical Learning from data: benefits to YOU

- We see/read massive amounts of data
- We generate massive amounts of data ourselves (everything on your phone is data!)
- Human brains are great at spotting and recognizing patterns, but also easy to trick
- Big data can help you (fitness trackers, movie recommendations)
- Big data can harm you (link between social media and depression, fake news)

Statistical Learning from data: benefits to YOU

Learning to critically think about data is one of the most imporant skills in the modern world

The process of learning from data



Some standard notation

- n: the number of observations (cases, data points)
- p: the number of variables (features, predictors, predictor variables)
- Variables can be quantitative, ordinal, categorical, or a mix
- Data matrix: $n \times p$ matrix X

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

- (Optional) Response: Y, which can be one or several variables for each observation – a special variable of interest. Typically, with one response, Y is stored as an $n \times 1$ vector.

Two main types of Statistical Learning

Supervised learning: X and Y are observed

- Goal: understand/summarize/visualize the relationships between X and Y, and/or learn to predict Y from X

Unsupervised learning: only X is observed

- Goal: understand/summarize/visualize the relationships between the variables in X
- Examples: dimension reduction, e.g., principal components analysis, clustering

Examples

- Visualization (applicable to both supervised and unsupervised tasks, often with different plots)
- Classification (supervised; Y is a categorical variable)
- Regression (supervised; Y is a continuous variable)
- ANOVA (supervised; categorical X, continuous Y)
- Clustering (unsupervised)

Classification: Definition

- Given a collection of data points (training set) $(x_1, y_1), \ldots, (x_n, y_n)$, where y is a class label (categorical)
- Find a model or an algorithm that outputs the class label y as a function of the values of variables x
- Goal: previously unseen data points x should be assigned a class label y as accurately as possible
- A test set (previously unseen) is used to determine the accuracy of the model

Classification example: Customer scoring

- A bank has a database of 1M past customers, 10% of whom took out mortgages with the bank
- Task: predict whether a current customer will take out a mortgage or not, based on the customer's data
  - History of transactions with the bank
  - Other credit data
  - Demographic data

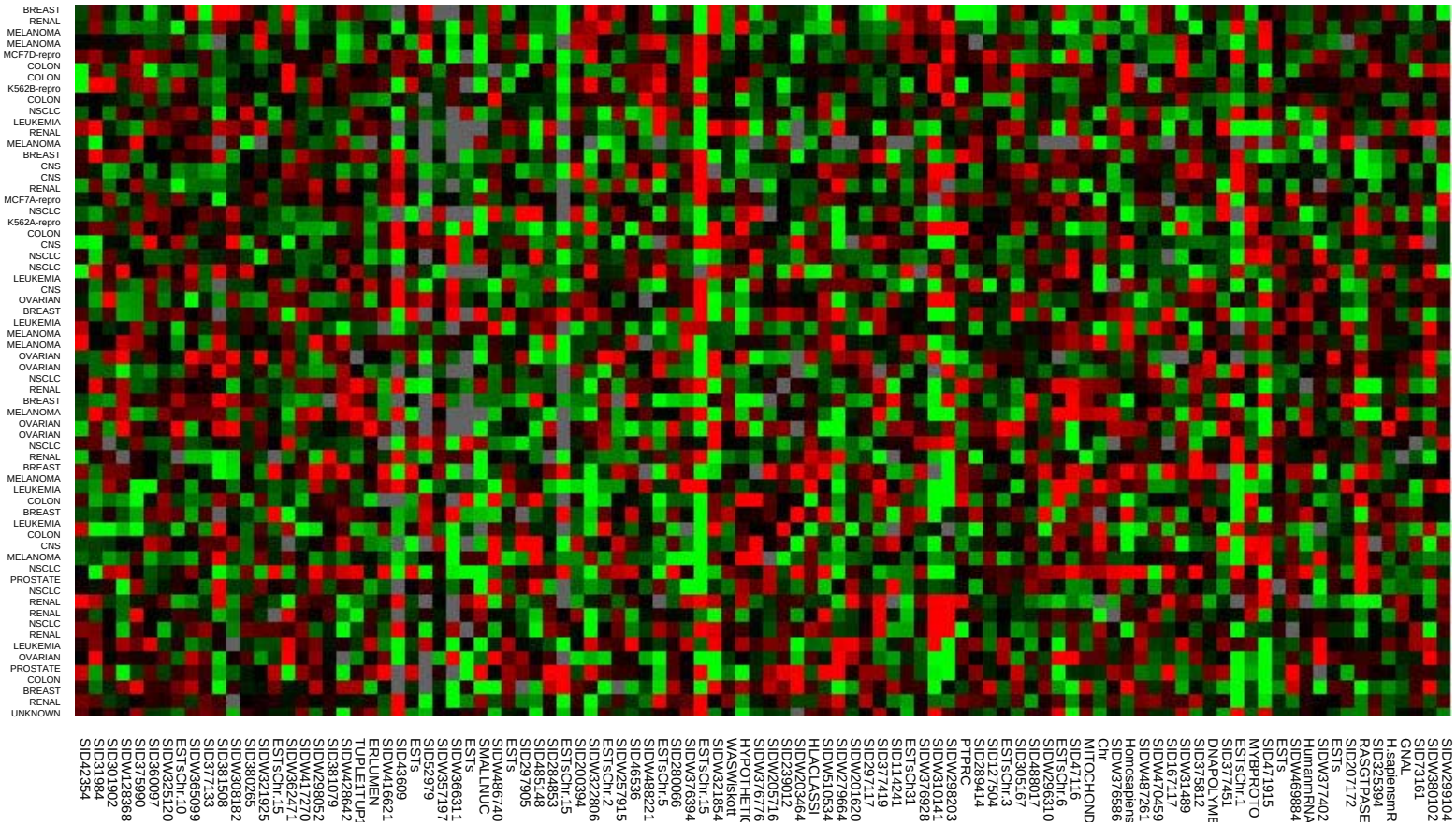
Classification example: Spam filter

- Task: customize an email spam detection system for an individual user.
- Features: relative frequencies of words and punctuation
- Most commonly occuring words:

	george	you	your	hp	free	re	remove
spam	0.01	2.26	1.38	0.02	0.52	0.13	0.28
email	1.27	1.27	0.44	0.90	0.07	0.42	0.01

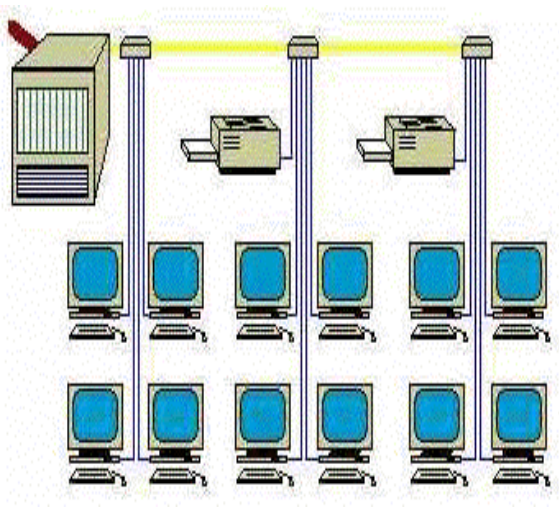
Classification example: Microarray gene expression data

Task: predict risk of developing cancer based on genetic profile



Classification example: Anomaly detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit card fraud
  - Network intrusion



Classification example: Credit card fraud detection

- Credit card losses in the US are over 1 billion $ per year
- Roughly 1 in 50k transactions are fraudulent
- Fair-Issac's fraud detection software based on neural networks, led to reported fraud decreases of 30-50%
- Challenge: false alarm rate vs missed detection

Regression: Definition

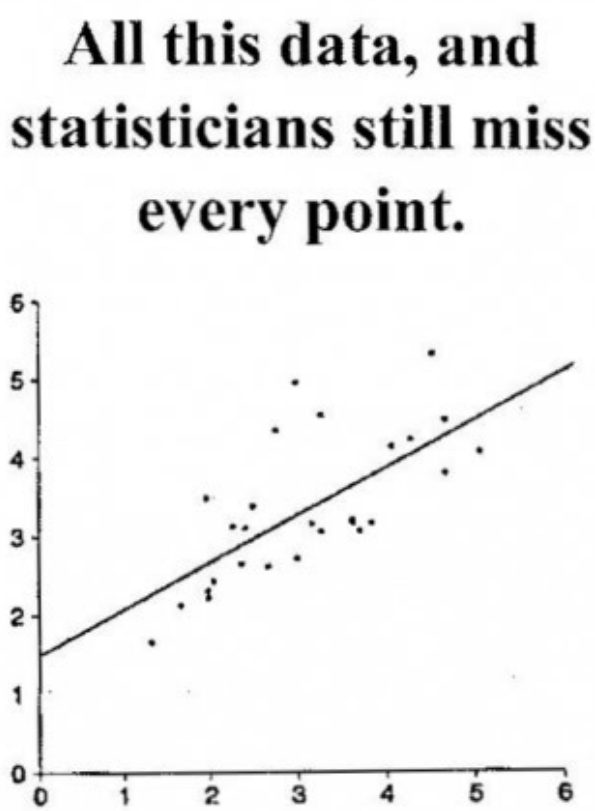
- Predict a value of a continuous-valued response variable based on the values of other variables
- Linear regression: predict Y from a linear combination of X
- There are many other tools for regression: nonlinear functions, trees, neural networks, etc

Regression: Examples

- Predicting sales amounts of new product based on advertising expenditure and product characteristics
- Predicting wind velocity as a function of temperature, humidity, air pressure, etc
- Predict a student's freshman year GPA based on high school grades and SATs

Universal trade-offs

- Occam's razor ("Less is More"): accuracy vs. interpretability
- Bias (accuracy) vs variance (replicability)
- Overfitting vs underfitting



Prediction vs inference

- Prediction: the goal is to predict Y from X. The predictor could be a black box as long as it's accurate.
- Inference: the goal is to understand how Y is connected to X and find explanatory value in X

Get good predictions, but fail to make inferences

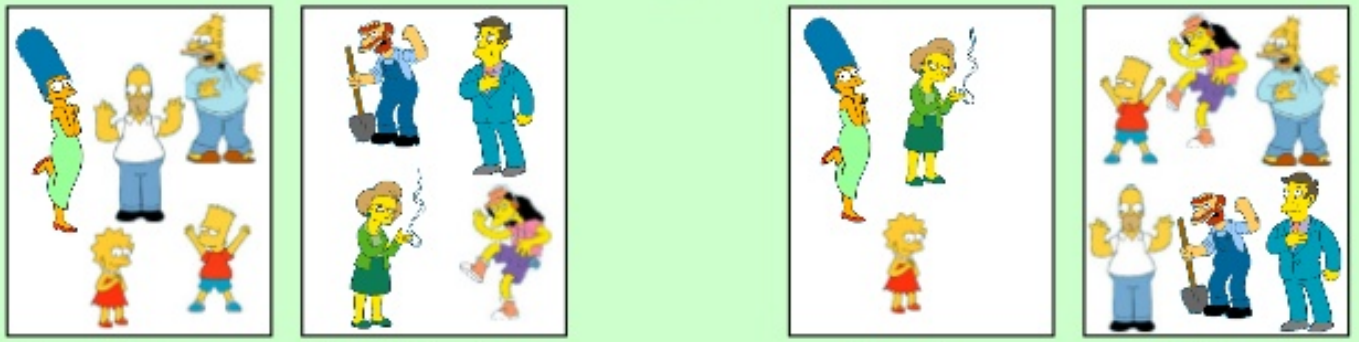
Clustering: Definition

- Given a set of data points, each having a set of variables, find clusters such that
  - data points in the same cluster are " more similar" to one another, and
  - data points in different clusters are "less similar" to one another.
- Similarity measures
  - Euclidean distance if variables are continuous
  - Other problem-specific measures

Importance of similarity measures

What is a natural grouping among these objects?

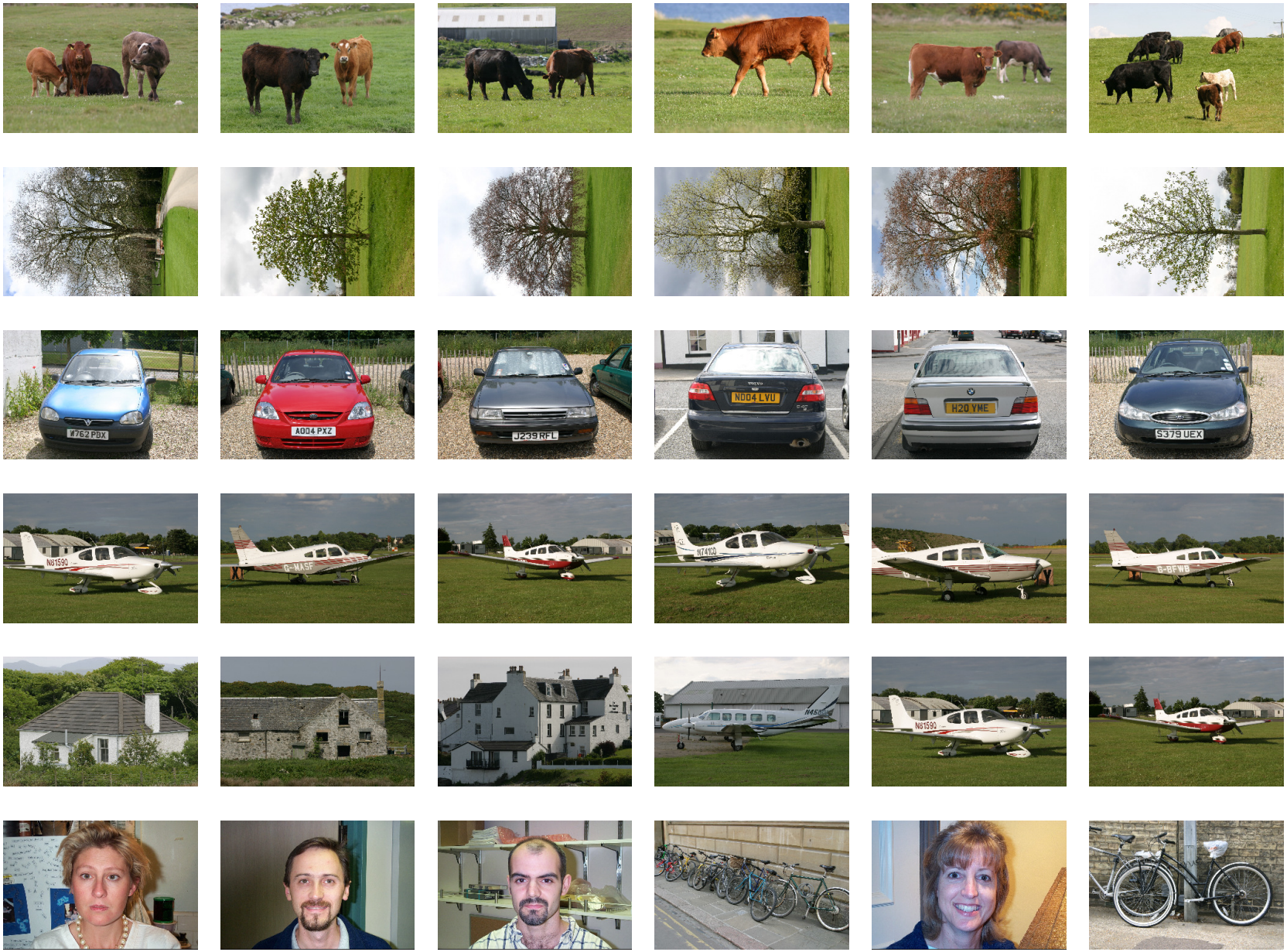
Clustering is subjective



Clustering example: Market segmentation

- Goal: subdivide a market into distinct subsets of customers for targeted marketing
- Collect different variables on customers (age, gender, marital status, education; geographical location; lifestyle, hobbies, etc)
- Find clusters of similar customers
- Here we rely on the assumption similar customers will like the same kind of marketing, but have no previous data on how they actually respond to particular marketing strategies

Clustering example: Images



More clustering examples

- Cluster documents that are similar to each other based on the important terms appearing in them, for example to organize news articles
- Cluster patients by symptoms and medical history, to develop personalized treatments
- Cluster stocks based on their movements every day, to find patterns in the market

Modern challenges in Statistical Learning

- Hype: people often expect more than is realistic
- Data snooping and fishing: finding spurious structure that is not replicable (Topic that is close to my heart!!)
- Irreplicable analysis
- Trade-offs:
  - Prediction vs inference: may get great performance from a black box, and a more interpretable simpler model may not predict as well
  - Bias vs variance: flexibility vs overfitting
  - Balancing false alarms against missed detection (Type 1 vs Type 2 error)

Data science/ Statistics in the age of AI

The future of data science is intimately linked to the future of AI.

As AI advances, it will create new opportunities and challenges for data scientists.

Data science/ Statistics in the age of AI

Google's chief economist Hal Varian, 2009:

The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that's going to be a hugely important skill in the next decades, not only at the professional level but even at educational levels...
Because now we really do have essentially free and ubiquitous data. So the complementary scarce factor is the ability to understand that data and extract value from it.