What *is* Data Mining?

What Do the Textbooks Say?

"Data mining is defined as the process of discovering patterns in data."

—Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition (Witten, Frank & Hall)

"Data mining refers to extracting or 'mining' knowledge from large amounts of data."

—Data Mining: Concepts and Techniques, 2nd Edition (Han & Kambler)

What Do More Textbooks Say?

"Data mining is the process of automatically discovering useful information in large repositories of data.

—Introduction to Data Mining (Tan, Kumar, & Steinbach)

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

—Principles of Data Mining (Hand, Mannila, & Smyth)

Even More Textbooks?

"Data mining is the set of methods and techniques for exploring and analysing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies; special systems output the essentials of the useful information while reducing the quantity of data."

—Data Mining and Statistics for Decision Making (Tuffery)

Our Definition

Data mining is the art of extracting knowledge from large bodies of structured data.

Two Key Features

1

Extracting Knowledge

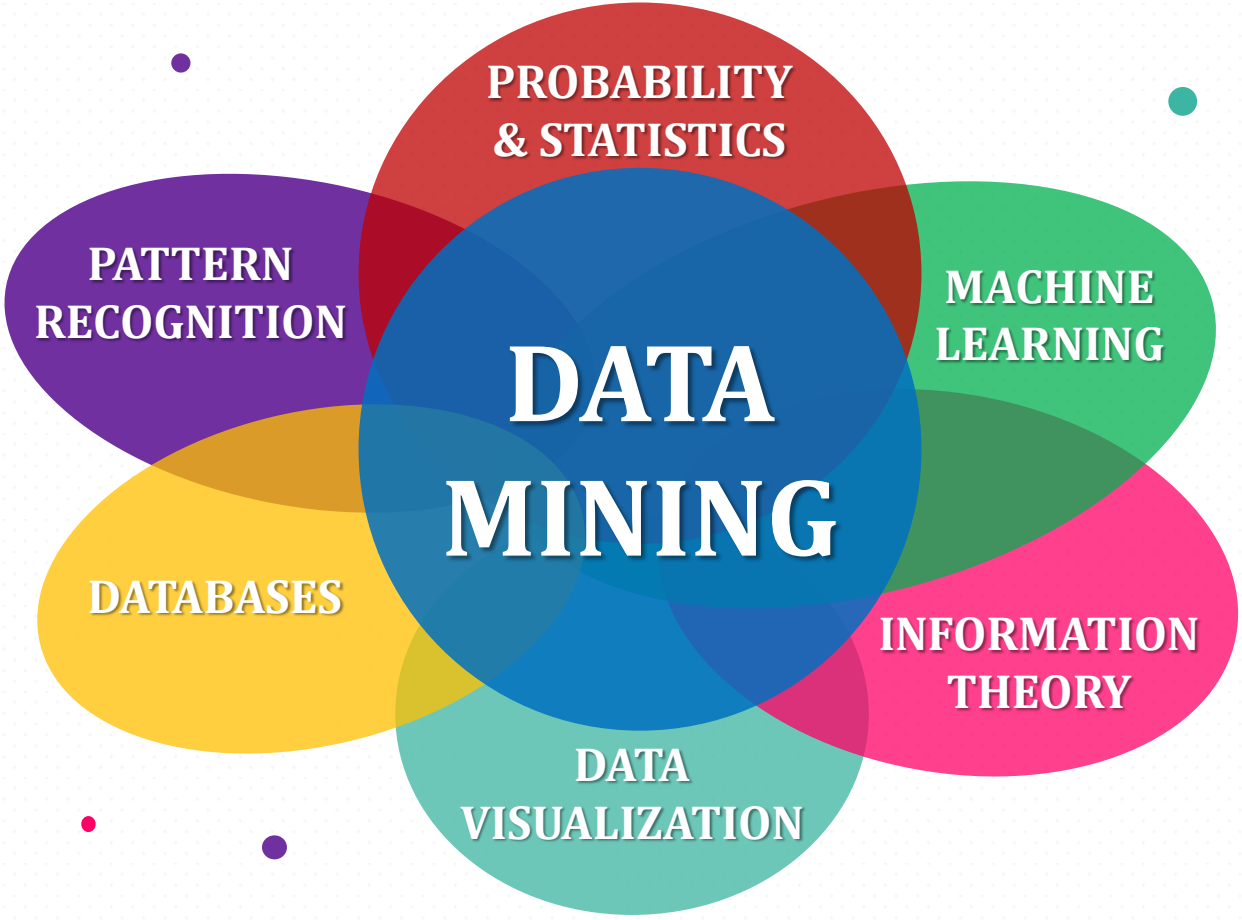
Previously unknown patterns, descriptions, or relations—potentially useful information—are being extracted from data. Discovering this knowledge often requires some form of learning.

2

Large Bodies of Data

Datasets are structured, often as a database. They are also typically larger than those encountered other domains. In fact, the datasets are so large that the process of extracting knowledge must be automated—or at least augmented—by computer.

It's a Confluence of Many Domains



What Isn't Data Mining?

Looking up a record in a database.

No pattern is revealed by this lookup.

Searching for a term on Google.

This is simply a "match" or "non-match".

Testing a two-sample hypothesis in a clinical trial.

The dataset is often not large.

What's Almost Data Mining

Noting that some last names occur in certain geographical areas.

Taking all query results from Google and discovering that they can be grouped or categorized.

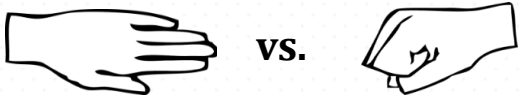
When doing multiple tests across many different genes, identifying very strongly significant genes.

Let’s look at a basic example of data mining.

Rock, Paper, Scissors

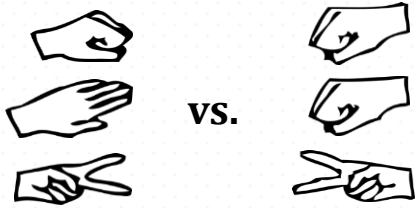
1 Rock-it:

Males have the tendency to produce rock on their first throw. If you are playing against one, try using paper.



2 Double on the Rocks:

When you see a two-Rock run, it is highly likely that your opponent's next move will be Scissors or Paper. People dislike being predicable. Counter with rock.



3 Paper Please:

Paper is thrown the least in a match. Use it as an unexpected options.

Paper is thrown 29.6% of the time.

Rock is thrown 35.4% of the time.

Scissors is thrown 35% of the time.

4 Spock & Roll:

When in doubt, and all seems lost, go for the Spock. It is unexpected and highly illegal, but also impossible to counter.

Okay, this part is not really data mining.

Based on: http://flowingdata.com/2010/07/30/how-to-win-rock-paper-scissors-every-time/

So maybe you're thinking, "That's *kind* of neat.
But what about some *real* examples."

FROM POLITICS TO BIOLOGY TO BIG DATA, THE EFFECTS OF DATA MINING ARE *EVERYWHERE*

ONE WAY DATA MINING CAN INFLUENCE OUR LIVES IS THROUGH PREDICTIONS

GENOMIC DATA

THE HUMAN SIDE OF DATA MINING

CHROMOSOME 5

Expansive gene ‘desert’

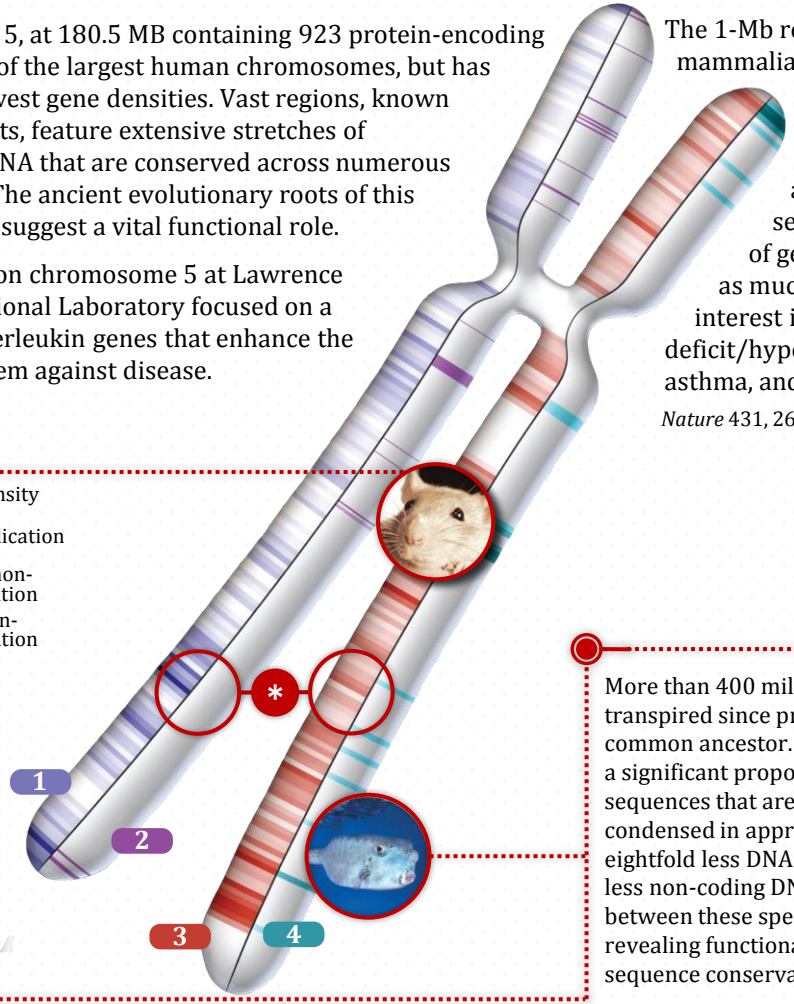
Chromosome 5, at 180.5 MB containing 923 protein-encoding genes, is one of the largest human chromosomes, but has one of the lowest gene densities. Vast regions, known as gene deserts, feature extensive stretches of non-coding DNA that are conserved across numerous vertebrates. The ancient evolutionary roots of this genetic motif suggest a vital functional role.

Pilot studies on chromosome 5 at Lawrence Berkeley National Laboratory focused on a cluster of interleukin genes that enhance the immune system against disease.

The 1-Mb region * illustrates how multi-mammalian sequence comparisons have led to the identification of non-coding elements possessing gene regulatory activities. These gene deserts seem to influence the regulation of genes separated by distances of as much as 12 Mb or more. Genes of interest include ADHD (attention deficit/hyperactivity disorder), obesity, asthma, and colorectal cancer.

Nature 431, 268-274 (2004)

CHROMOSOMES ARE PACKAGES OF DNA AND PROTEIN, LOCATED IN THE NUCLEI OF HUMAN AND OTHER EUKARYOTIC CELLS



Only 2% of the human genome encodes protein, leaving the vast proportion of non-coding DNA sequences as one of the mysteries still being deciphered. Comparative genomics between human and other vertebrates—rodents and the puffer fish, *Fugu rubripes*—serves to distinguish functional regions based on conservation between species. Mouse-human non-coding conservation is depicted here in red. With genomes of comparable size (three billion bases), it has been 80 million years since the last human-mouse common ancestor. Human-mouse comparisons have provided clues to such functional non-coding sequences as those regulating immune response.

More than 400 million years of evolution have transpired since primates and fish last shared a common ancestor. Although the *Fugu* genome contains a significant proportion of genes and regulatory sequences that are similar to a human, they are condensed in approximately 400 Mb, or nearly eightfold less DNA than the human genome. With far less non-coding DNA to sift through, conserved regions between these species have proved successful in revealing functionality. *Fugu*-human non-coding sequence conservation is depicted in turquoise.

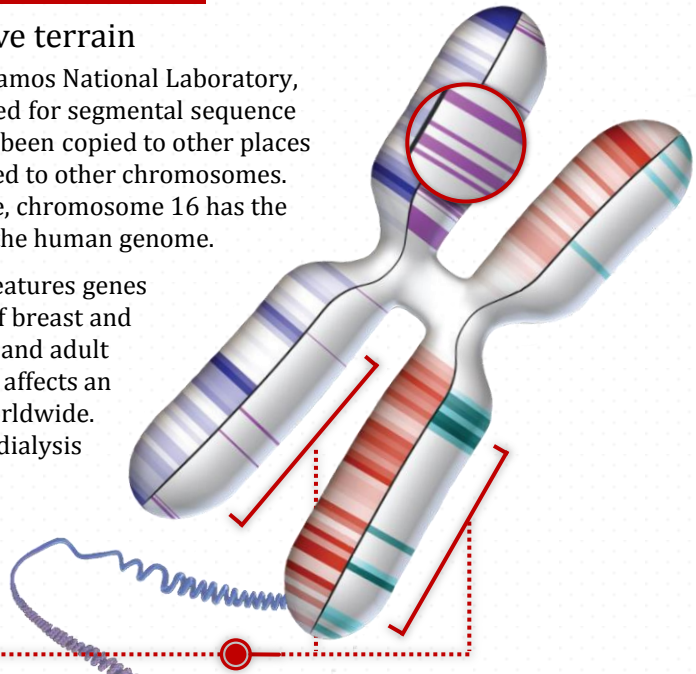
Source: DOE Contribution to Sequencing the Human Genome | US Department of Energy

Highly repetitive terrain

A focus of early studies at Los Alamos National Laboratory, chromosome 16 is highly enriched for segmental sequence duplications—regions that have been copied to other places within the chromosome, or copied to other chromosomes. Excluding the Y sex chromosome, chromosome 16 has the most segmental duplications in the human genome.

At 88.7 Mb, it has 880 genes. It features genes implicated in the development of breast and prostate cancer, Crohn's disease and adult polycystic kidney disease, which affects an estimated five million people worldwide. Half the affected people require dialysis or kidney transplant.

Nature 432, doi:10.1038/nature03187 (2004)



The chromosome figures depicted here are sister chromatids; that is, two identical copies of a single chromosome that are connected by a centromere. Sister chromatids are created during the interphase period of the cell cycle.

Histones are the proteins around which DNA coils to form chromatin, which organizes into chromosomes. The human genome consists of 23 pairs of chromosomes, with one set donated by each parent.

The number assigned to chromosomes is inversely proportional to their size, so that, excluding the sex chromosomes X and Y, chromosome 1 is the largest and chromosome 22 the smallest.

Each chromosome contains a single molecule of the DNA double helix arranged in long stretches of nucleotides, one of four different nitrogenous bases; adenine (A), Thymine (T), cytosine (C), and guanine (G). The order of these bases along a strand of DNA is the genome sequence.

Data mining has provided massive insights to understanding the human genome, illuminating the causes of ADHD (attention deficit/hyperactivity disorder), obesity, asthma, colorectal cancer, Crohn's disease, and providing clues to factors that have influenced human evolution.

EXPLODING DATA

THE POTENTIAL OF BIG DATA

WHAT IS BIG DATA?

"Big Data" refers to sets of data whose size surpasses that of what data storage tools can typically handle. It's something that grows concurrently with the development of technology and something that helps continued innovation.

The amount of digital data in our world has been exponentially growing in just a few short years. Big data has the potential to become the next frontier for innovation, competition, and profit.

A GROWING TORRENT

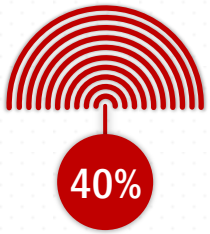
Just how big is big data? Huge; and with the potential to expand even more in the future.

4 Billion

PIECES OF CONTENT ARE SHARED EVERYDAY ON FACEBOOK

235 Terabytes

OF DATA HAS BEEN COLLECTED BY THE US LIBRARY OF CONGRESS IN APRIL 2011



THE PROJECTED GROWTH OF GLOBAL DATA PER YEAR

THE PROJECTED GROWTH OF GLOBAL IT SPENDING PER YEAR

DATA SECTORS

15 OUT OF 17 SECTORS IN THE UNITED STATES HAVE MORE DATA PER COMPANY THAN THE US LIBRARY OF CONGRESS

AMOUNT OF STORED STATE BY SECTOR:
(IN PETABYTES, 2009)

- 966 DISCRETE MANUFACTURING
- 848 GOVERNMENT
- 715 COMMUNICATIONS
- 694 PROCESS MANUFACTURING
- 619 BANK
- 434 HEALTHCARE
- 429 SECURITIES & INVESTMENTS
- 411 PROFESSIONAL SERVICES
- 364 RETAIL
- 269 EDUCATION
- 243 INSURANCE
- 227 TRANSPORTATION
- 202 WHOLESALE
- 194 UTILITIES
- 116 RESOURCE INDUSTRIES
- 106 CONSUMER RECREATIONAL
- 51 CONSTRUCTION

PROCESS MANUFACTURING

3.8 PETABYTES

ON AVERAGE, A SECURITIES AND INVESTMENT FIRM WITH UNDER 1,000 EMPLOYEES WILL HAVE 3.8 PETABYTES OF DATA STORED.

WHAT IS THIS DATA WORTH?

Simply knowing that all this data is out there is one thing. Utilizing this data to turn a profit is another issue. Using big data has the potential to increase profitability across all sectors. However there are five sectors that stand to benefit the most.

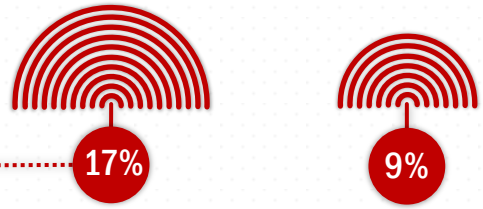
HEALTH

Healthcare has lagged behind many other industries when it comes to improving operational performance and adopting technology-enabled process improvements.

$300 Billion

IN TEN YEARS, THE IMPLEMENTATION OF BIG DATA IN THE HEALTH INDUSTRY COULD CAPTURE $300 BILLION ANNUALLY

AND BRING CURRENT HEALTHCARE EXPENDITURES BY THE US GOVERNMENT DOWN 8%



HEALTHCARE EXPENDITURES AS A PERCENTAGE OF THE GDP

AREAS WHERE BIG DATA COULD BE UTILIZED

POTENTIAL WORTH OF DATA

R&D
Research and development; clinical trail design; personalized medicine

$108 BILLION

CLINICAL
Transparency in clinical data and clinical decision support

$65 BILLION

ACCOUNTING
Advanced fraud detection; performance-based drug pricing

$47 BILLION

PUBLIC HEALTH
Public health surveillance and response systems

$9 BILLION

NEW BUSINESS MODEL
Aggregation of patient records; online platforms and communities.

$5 BILLION

GOVERNMENT

Governments in many parts of the world need to increase their productivity through digital means. Examining the public sector of the European Union, we can where the utilization of big data can create value through efficiency.

€300 BILLION

UTILIZING BIG DATA, EUROPE'S PUBLIC SECTOR COULD REDUCE COSTS BY 20% OR 300 BILLION EUROS

AREAS WHERE BIG DATA COULD BE UTILIZED

OPERATIONAL EFFICIENCY SAVINGS

€200 BILLION

REDUCTION IN FRAUD AND ERROR

€30 BILLION

INCREASE IN TAX COLLECTION

€110 BILLION

POTENTIAL WORTH OF DATA

RETAIL

The use of technology and digital data in the retail industry has allowed for a boost in profitability and productivity over the decade. The continued adoption of big data has the potential for further profitability.

60%

THE POTENTIAL INCREASE IN RETAILERS' OPERATING MARGINS FROM BIG DATA COULD BE 60%

MANUFACTURING

The manufacturing sector has adopted data in the use of information technology and automation. Continued adoption of big data could lead to increased production and decreased costs.

50%

BIG DATA HAS THE POTENTIAL TO CUT OPERATING COSTS BY NEARLY 50% ACROSS ALL SECTORS OF MANUFACTURING

PERSONAL LOCATION TECHNOLOGY

Personal location data volumes has increased rapidly with the adoption of mobile phones. The potential for this data is far greater than any other, because it is not confined to a single sector, but cuts across all industries.

$600 BILLION

THE POTENTIAL ANNUAL CONSUMER SURPLUS FROM GLOBAL PERSONAL LOCATION DATA IS $600 BILLION

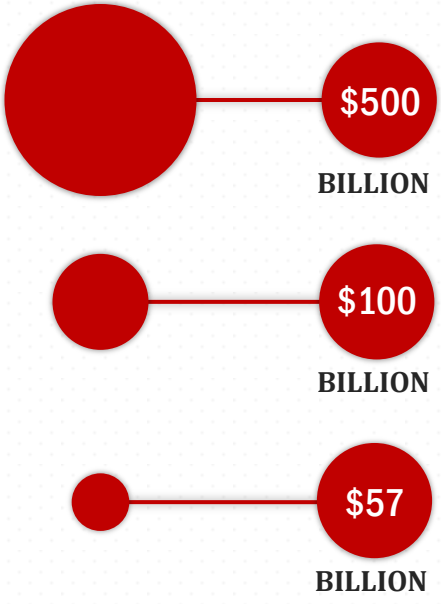
1 Petabyte

AREAS WHERE BIG DATA COULD BE UTILIZED

GPS
Navigation including smart routing based on real-time traffic

MARKETING
Geo-targeted mobile advertising (advertising platform providers)

SOCIAL
People tracking location sharing and entertainment



CAN WE CAPTURE THE FULL POTENTIAL OF BIG DATA?

There are several issues that must be dealt with before all industry sectors can access the full potential of this big data.

DATA POLICIES

Big data raises several legal issues due to the fact that data is fundamentally different from other assets. Since data can be so easily copied, intellectual property becomes an urgent consideration to policy makers. Also, there is an issue of culpability. Who is liable for an inaccurate piece of data when it leads to negative consequences?

TECHNOLOGY

Local systems and inferior standards/formats prevents the integration of big data in many sectors, especially the public sector. Making use of large datasets requires both adequate storage and compatible technology. These investment costs are sometimes far too large.

DATA ACCESS

In order to create the most broad data available, companies will need to increasingly rely on third-party data sources and integrate external information with their own. Currently, there are not completely efficient markets that allow for this transfer and sharing of data..

TALENT

In many instances, there is a lack of skilled personnel needed to mine big data, create the necessary structures, and make use of big data through informed decisions.

Sources:

- "Big Data: The Next Frontier for Innovation, Competition and Productivity." US Bureau of Labor Statistics | McKinsey Global Institute Analysis
- Column Five Infographic

Data Mining Tasks

1

Descriptive Tasks

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that are able to summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

2

Predictive Tasks

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the **target** or **dependent variable**, while the attributes used for making the prediction are known as the **explanatory** or **independent variables**.

Course Topics

1

Data Understanding

Types of data; information and uncertainty; classes and attributes; interactions among attributes; relative distributions; summary statistics; data visualization

2

Data Preprocessing

Standardizing data; sampling data; using principal components to eliminate attributes

3

Clustering and Association

Dissimilarity and scatter; *k*-means clustering; hierarchical clustering; determining the number of clusters

Course Topics

4

Classification and Regression

Nearest neighbor; decision trees; Bayesian learners; regression

5

Validation and Interpretation

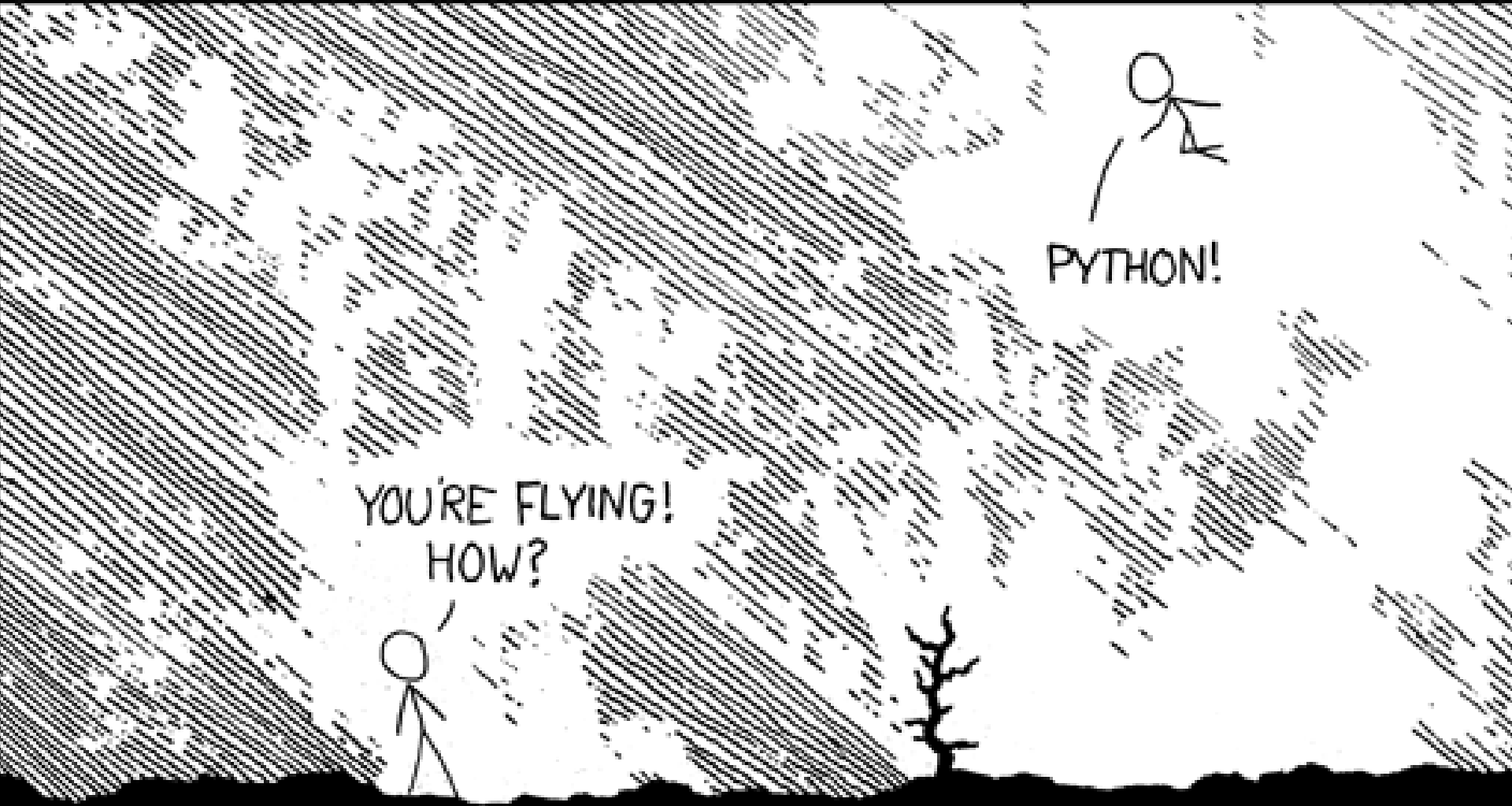
Concept of validation and testing data; comparing benchmarking techniques; performance metrics (error, ROC curves, lift curves)

Tools and Modules

SciPy

IP[y]: Notebook

[Static Demo]



Source: xkcd

Data Mining

The art of extracting knowledge from large bodies of structured data.

What does the process look like?

Data Mining Tasks

1

Descriptive Tasks

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that are able to summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require postprocessing techniques to validate and explain the results.

2

Predictive Tasks

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the **target** or **dependent variable**, while the attributes used for making the prediction are known as the **explanatory** or **independent variables**.

Data Mining Tasks

- Anomaly Detection: the task of detecting unusual deviations.
- Association Analysis: the task of discovering patterns that describe relationships.
- Clustering: the task of discovering groups and structures
- Classification: the task of assigning (discrete) target variables to one of several predefined categories.
- Regression: the task of finding a function that models (continuous) target variables.
- Collaborative Filtering: the task of filtering patterns for an unknown user based on patterns for known users.

Defining a Data Mining Task

- Generate a problem statement.
- Utilize background knowledge.
- Posit the right question.
- Understand the data.
- Implement one or more modeling approaches.
- Identify performance measurement criteria.
- Interpret the model(s).
- Visualize and present the results.

The Goal

Using the knowledge discovery process to turn data into knowledge.

Knowledge Discovery Process

Let's look at what we mean by *data*.

What's in Data?

What's in Data?

- Instance:
  - Thing to be classified, associated, or clustered.
  - Individual, independent example of target concept.
  - Characterized by a predetermined set of features or attributes.
- Input to learning scheme: set of instances
  - Usually represented as a single relation.
  - Traditional form for data mining.
  - Advanced methods now exist for relational data.

What's in an Instance?

- Each instance is described by a set of "features."
- A feature is a property or characteristic of an instance.
- A feature can take several values (feature values).
  - Can be categorical (nominal or ordinal)
  - Can be numeric (interval or ratio)
- Features can discrete or continuous.

Discrete Features

- Qualitative features.
  - Enough information to distinguish one object from another.
- Has only a reasonable set of values.
  - Thumb-rule: count with your fingers.
- Often represented as integer variables.
  - For example: 0 for red; 1 for blue; etc.
- Note: binary attributes are a special case of discrete features.

Continuous Features

- Most numeric properties hold.
- Can be integer or real number.
- Examples: temperature, height, weight, age, counts.
- Practically, real values can only be measured and represented using a finite number of digits.

Supervised Learning

- Data in the form of tuples or input-output pairs (x_i, y_i) that comes from a deterministic mapping of $X \rightarrow Y$.

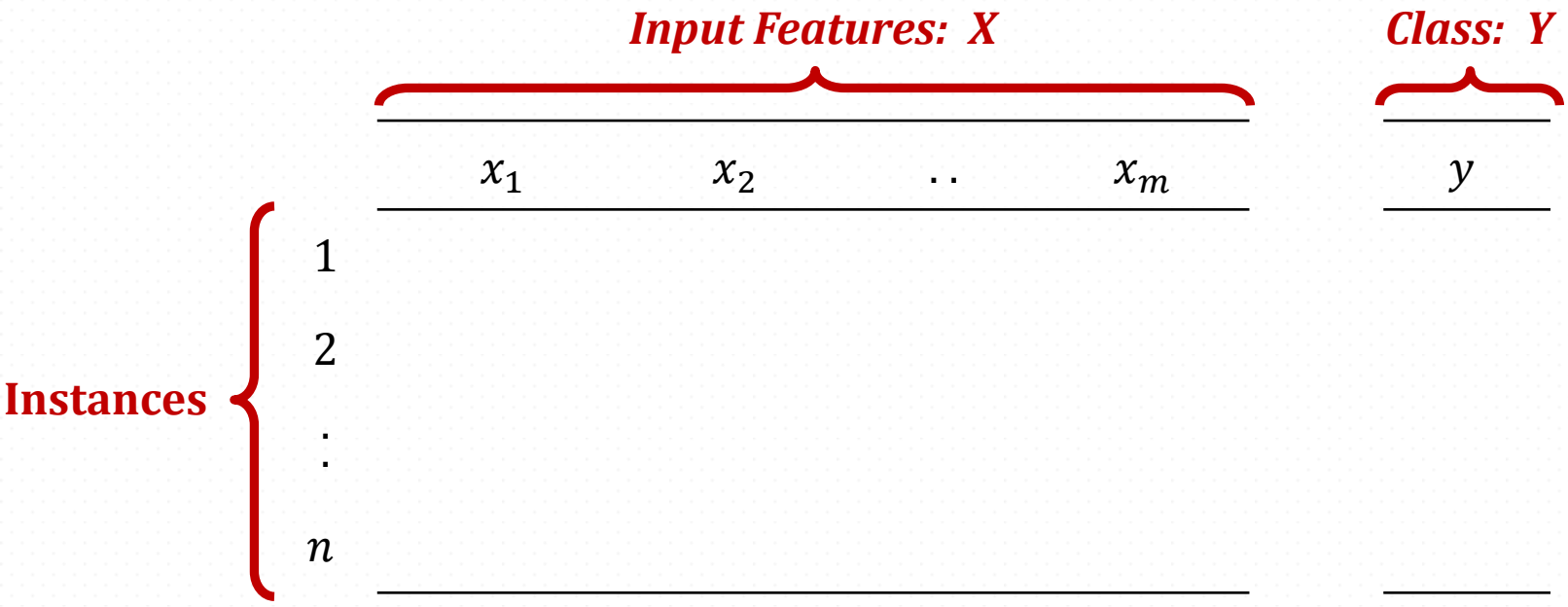
$$\forall \vec{x}_i \in X$$

$$\vec{x} \equiv \{x_1, x_2 \ldots, x_n\}$$

$$\forall y_i \in Y \ (set\ of\ classes/concepts\ c(x))$$

  - An example of "supervised" learning.
- Develop an approximation to the mapping that is "consistent" with the data and "not too complex."
  - Learn a function of the hypothesis.

Instances, Features, and Classes



Instances, Features, and Classes

Make	Cylinders	Length	Weight	Style
Honda	Four	150	1956	Hatchback
Toyota	Four	167.9	2280	Wagon
BMW	Six	176.8	2765	Sedan

Given car make, cylinders, length, and weight, learn a function for the body style.

Instances, Features, and Classes

Temperature	Wind Speed	Decision
80°	Low	Bike Day
40°	Low	Couch Day
60°	Medium	Couch Day
80°	High	Bike Day

Go out and bike or laze on the couch.

Features and Classes

1. If wind-speed = high, then Bike Day.
2. If wind-speed = medium, then Couch Day.
3. If wind-speed = low and temp ≤ 40, then Couch Day.
4. If wind-speed = low and temp > 40, then Bike Day.

Now Consider this Problem

- An advertising company wants to group customers based on similarities. There are no predefined labels for this group, and based on the groups on demographics and past buying behavior, they will have targeted marketing and advertising initiatives.
- What is this?
  - An example of unsupervised learning.
  - No predefinition of groups, a.k.a. classes.
  - Find similarities in data based on features.
  - This is the simplistic view of clustering.

Unsupervised Learning

- Given data Points X:
  - Develop a model or representation of the data such that "important" structure or "regularities" (and irregularities)" are captured.
  - Organizing instances into groups that share similar features.
  - Model, for example, can be probability distribution estimation: $p(x)$ of the entire X.

Examples of Kinds of Data

- Financial transactions
- Genetic sequence data
- Documents
- WWW
- Molecular structures
- Medical data
- Geographical data

Knowledge Discovery Process

So that's *data*. The process also often involves the concept of *learning*.

What is Machine Learning?

Definition of Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E."

—Tom Mitchell/Machine Learning

Definition of Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E."

—Tom Mitchell/Machine Learning

- Improve over class of tasks *T*
- With respect to performance measure *P*
- Based on experience *E*

We need to be able to formulate *T*, *P*, and *E*.

Learning to Play Checkers

- **Task:** Playing checkers.
- **Performance:** Percent of games won in world tournament.
- What training experience?
- What exactly should be learned? (*Target function*)
- How to represent the target function?
- What specific algorithm to learn it?

Formalizing the Learning Task

- Training experience → training data.
- Task → target function required.
  - What is the outcome or what is to be predicted?
- Identify the objective or learning function required to fit the data.
  - For example, rules or decision trees.
- Performance measurement criteria → evaluate the learning function on the testing data.
- How accurate is the function?

Concept Learning

- Acquire general concepts from a set of training examples.
- A concept can describe some objects or events.
  - People, continually, attach "description" to objects or events.
    - I will enjoy sports today if the sky is sunny, air temperature is warm, humidity is normal, and wind is not strong.
- Typically, inferring a boolean-valued function.
  - True or false.
- Definition of concept learning: approximate a boolean valued function from training examples.

Instances, Features, and Classes

Make	Cylinders	Length	Weight	Style
Honda	Four	150	1956	Hatchback
Toyota	Four	167.9	2280	Wagon
BMW	Six	176.8	2765	Sedan

Given car make, cylinders, length, and weight, learn a function for the body style.

The Weather Problem

Outlook	Temperature	Humidity	Windy	Play
sunny	85°	85	false	no
sunny	80°	90	true	no
overcast	83°	86	false	yes
rainy	70°	96	false	yes

Given past data, can you come up with the rules for determining the value of Play?

Formalizing Concept Learning

- Given:
  - Instances X: Possible days, each described by the attributes Outlook, Temperature, Humidity, Windy.
  - Target function c: *PlayGolf*: $X \rightarrow \{0,1\}$.
  - Hypothesis set H: Conjunction of attributes.
  - Training examples D: Positive and negative examples of the target function $\langle x_1, c(x_1), \ldots x_n, c(x_n) \rangle$.
- Determine a hypothesis h in H such that $h(x) = c(x)$ for all x in D.

Consistency

"A hypothesis h is consistent with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\big(x, c(x)\big)$ in D."

$$Consistent(h, D) \equiv (\forall \langle x, c(x) \rangle \in D) h(x) = c(x)$$

The Weather Problem

Conditions for playing sport:

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	normal	false	yes
⋮	⋮	⋮	⋮	⋮

If Outlook = sunny and Humidity = high, then Play = no.
If Outlook = rainy and Windy = true, then Play = no.
If Outlook = overcast, then Play = yes.
If Humidity = normal, then Play = yes.
If none of the above, then Play = yes.

The Weather Problem (Mixed Features)

Conditions for playing sport (with some numeric attributes):

Outlook	Temperature	Humidity	Windy	Play
sunny	85°	85	false	no
sunny	80°	90	true	no
overcast	83°	86	false	yes
rainy	70°	96	false	yes
⋮	⋮	⋮	⋮	⋮

If Outlook = sunny and Humidity > 83, then Play = no.
If Outlook = rainy and Windy = true, then Play = no.
If Outlook = overcast, then Play = yes.
If Humidity < 85, then Play = yes.
If none of the above, then Play = yes.

The Premise of Learning

- Given a training set, the (concept) learning algorithm has to estimate the function f or hypothesize about it.
- We have just formed a premise behind inductive learning.
  - What is inductive learning?

Inductive Learning Hypothesis

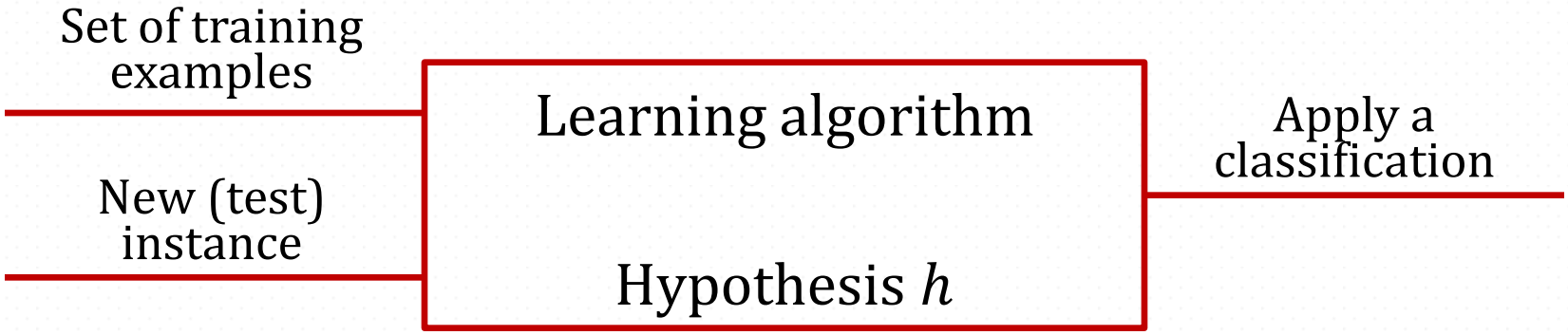
"Any hypothesis found to approximate the target function well over a large set of training examples will also approximate the target function well over other unobserved examples."

—Tom Mitchell/Machine Learning

Inductive Learning Hypothesis

- In other words:
  - Given a training set (known information), at best we can build (induce) a hypothesis around it.
  - Think about studying for an exam.
  - Finding a solution that is having a different “inductive” bias.

Inductive Learning Hypothesis



Use training instances to formulate or induce or discover a theory.

Go from specific to general.

Types of Learning

- Supervised learning
  - Given the value of an input vector X and $c(x)$, predict c on future unseen x's.
  - ex., classification, regression

- Unsupervised learning
  - Given X, automatically discover the structure, representations, etc.
  - ex., clustering

Supervised Learning

- Classification
  - The predictions or outputs, $c(x)$ are categorical while x can take any set of values (real or categorical). The goal is select correct class for a new instance.
- Regression
  - Given the value of an input X, the output Y denoted by $\ddot{y}$ belongs to the set of real values $\mathbb{R}$. Goal is to predict output accurately for new input.

Supervised Learning

- Time series prediction
  - Data is in the form of a moving time series. The goal is to perform classification/regression on future time series based on data known so far.

Classification

- Find ways to separate data items into pre-defined groups.
  - We know X and Y belong together, find other things in same group.
- Requires “training data”: data items where group is known.

Unsupervised Learning

- Anomaly detection
  - X can be anything, goal is to detect deviations from normal behavior.
- Association rules
  - Find joint values of the variables $X = \{x_1, x_2, \dots, x_n\}$, that tend to appear more frequently in the database.

Unsupervised Learning

- Clustering
  - X is provided $c(x)$ or Y is unknown. Grouping a collection of objects into “clusters” such that objects within a cluster are more closely related than those in different clusters.
- Density estimation
  - Describing your data.

Anomaly Detection

- Find unexpected values and outliers.

Association Rules

- Identify dependencies in the data.
  - X makes Y likely
- Indicate significance of each dependency.

Clustering

- Find groups of similar data items.
- Statistical techniques require some definition of “distance” (e.g., between travel profiles) while conceptual techniques use background concepts and logical descriptions.

Inductive Learning Bias

- Consider:
  - Concept learning algorithm L.
  - Instances X, target concept c.
  - Training examples $D_c = \{\langle x, c(x)\rangle\}$.
  - Let $L(x_i, D_c)$ denote the classification assigned to the instance x_i by L after training on data D_c.
  - The inductive bias of L is any minimal set of assertions B such that for any target concept c and corresponding training examples D_c.

$$(\forall x_i \in X)[(B \wedge D_c \wedge x_i) \longmapsto L(x_i, D_c)]$$

The Simpler the Better

- If there rare two hypotheses describing a data, one complex and one simple, which one to take?
- William of Occam in the year 1320 said, “Prefer the simplest hypothesis that fits the data.”
  - “One should not increase beyond what is necessary, the number of entities required to explain anything.”
  - Solid theory in machine learning behind it.
- Remember our set of hypotheses H.
  - Given a set of data, there are multiple ways to model it, a set of H.
  - Choose the simpler h of H that fits the data.

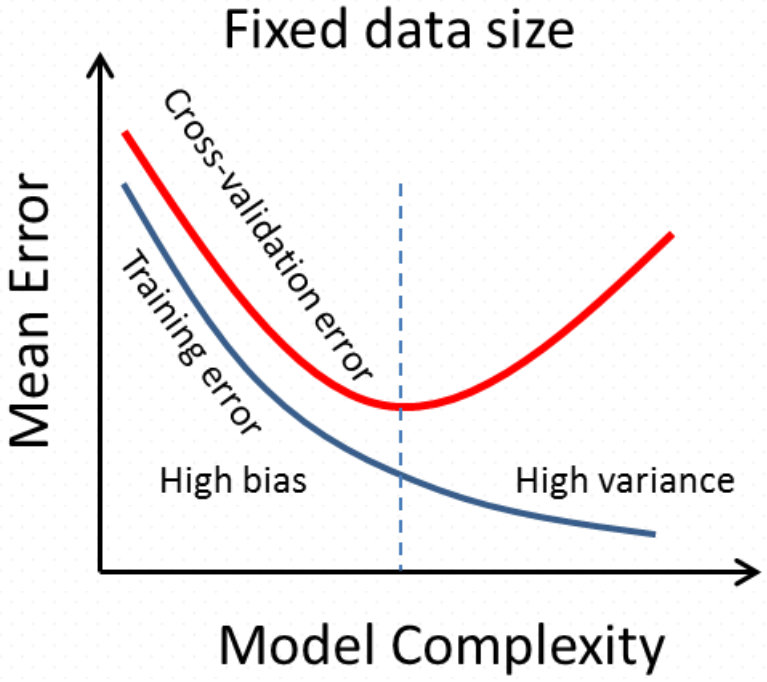
Inductive Learning Issues

- Along the way, we also learned that the learning algorithm should be able to generalize well.
- The algorithm should be able to induce knowledge of a domain from given training examples, and not merely and completely "overfit" on the training examples.

Overfitting

- Overfitting means doing very well on training data but poorly on the test data, lest test data exactly mimics the training data
  - Counterintuitive?
  - Think about preparing for an exam. What is better: rote memorization or understanding a concept?
  - Learner could not induce a function that generalizes well.
- Generalization is often a tenet of machine learning.

Generalization Behavior



Summarizing the Basics of Learning

- Define the “domain” (training data and testing data); What are we trying to learn? What is our data source? What characteristics of data are relevant?
- Define the “learner.” Select what is appropriate (no free lunch).
- Do we have any prior knowledge about the domain?
- Define the objective function, Φ. Adjust the learning to minimize the objective function.
- How do we define success? What is the output? How does the learner demonstrate that it has learned something? What decisions can I take based on what the learner tells me?

Summarizing the “Preliminaries”

Summarizing the "Preliminaries"

- Data are composed of *instances*.
- Each instance is described by a set of *features*.
- Sets of instances are used to perform a *task*.
- Data mining tasks can be *descriptive* or *predictive*.
- Often, these data mining tasks involve *learning*.
- Learning involves a tradeoff of *bias* and *variance*.

Understanding the Data

“Data, data everywhere, but not a thought to think.”
—Jesse H. Shera

“Data, data everywhere, but not a thought to think.”
—Jesse H. Shera

So let’s starting thinking about the data.

Data Understanding

The preliminary investigation of the data in order to better understand their specific characteristics.

Data Understanding

The preliminary investigation of the data in order to better understand their specific characteristics.

This process should be applied to *all* data.

Data Understanding Steps

1. Collection of initial data
2. Description of data
3. Exploration of data
4. Identifying data quality problems

Data Understanding Steps

1. Collection of initial data

Let's start by looking at this step.

2. Description of data
3. Exploration of data
4. Identifying data quality problems

Data Collection

- What data do I need and is available?
- Identify a domain expert, if available.
- Identify the relevance of the data.
- Is this data sufficient?
  - Enough instances?
  - All the relevant features?

Data Understanding Steps

1. Collection of initial data
2. **Description of data**

Now, let's look at this step.

3. Exploration of data
4. Identifying data quality problems

Describing Data

- Dimensionality
- Sparsity
- Resolution

Data Dimensionality

- The dimensionality of a dataset is the number of features that the objects in the dataset possesses.
- Data with a small number of dimensions tend to be qualitatively different than moderate or high-dimensional data.

Data Dimensionality

More data, more features or dimensions, is always a good thing, right?

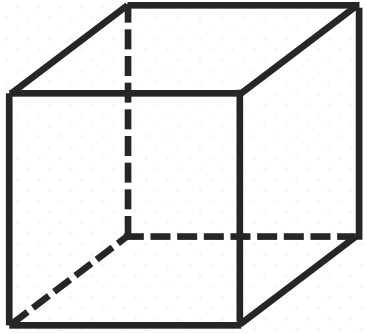
Data Dimensionality

More data, more features or dimensions, is always a good thing, right?

It's actually a blessing *and* a curse.

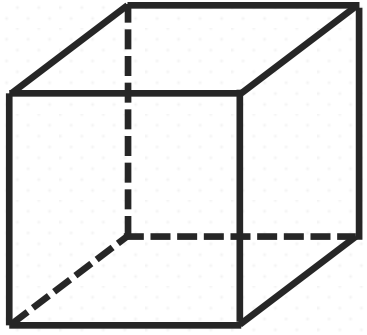
Curse of Dimensionality

- Suppose we have 100 instances uniformly distributed in a unit hypercube.



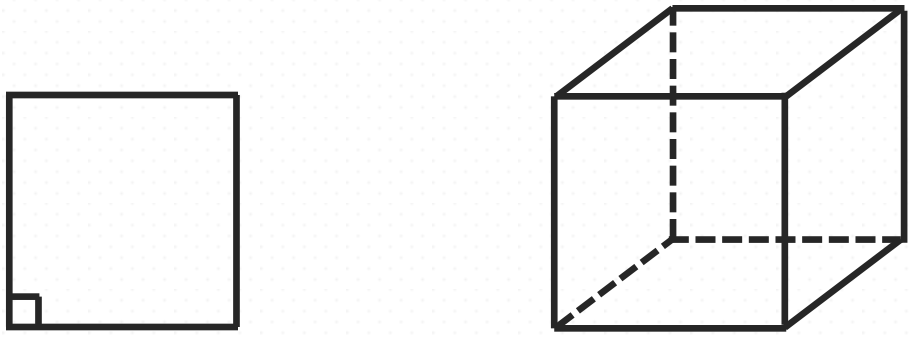
Curse of Dimensionality

- In 1 dimension, we must go a distance of $1/100 = 0.01$ on the average to reach our nearest neighbor.



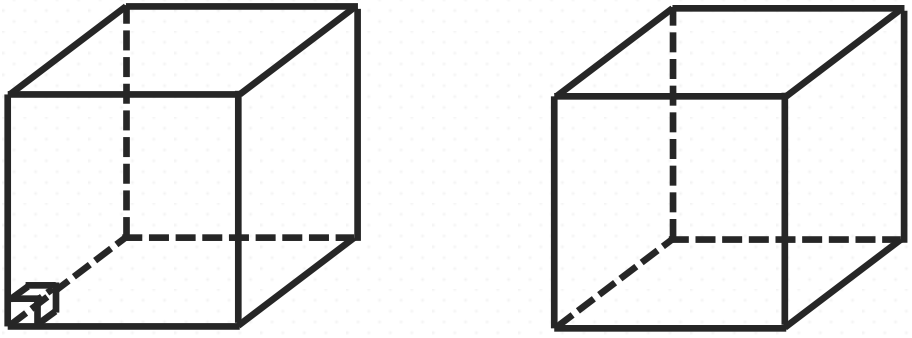
Curse of Dimensionality

- In 2 dimensions, we must go a distance of $\sqrt{0.01}$ to get a square that contains 0.01 of the volume.



Curse of Dimensionality

- In d dimensions, we must go $(0.01)^{1/d}$.



Curse of Dimensionality

The volume of the feature space increases so quickly that the available data become sparse.

This can often be a problem!

Data Sparsity

- For some datasets, most features have values of 0.
- Can be a problem for many methods, often statistical ones.
  - Can create a statistical bias due to small samples.
- Can also be an advantage, because less storage may be needed.

Data Resolution

- Different resolutions reveal different patterns.
- If the resolution is too fine, a pattern may be buried in noise.
- If the resolution is too coarse, the pattern may disappear.

Data Understanding Steps

1. Collection of initial data
2. Description of data
3. **Exploration of data**

Now, let's look at this step.

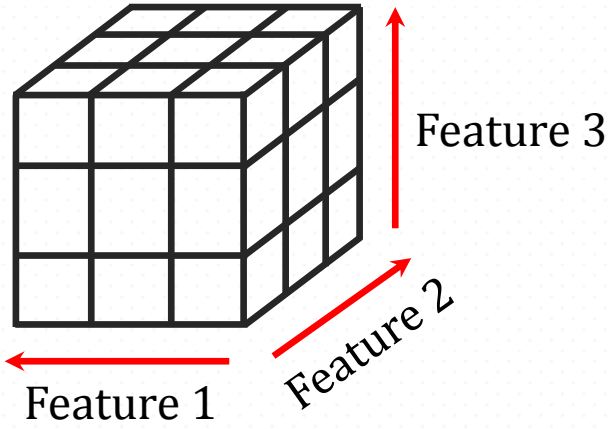
4. Identifying data quality problems

Data Understanding

- Gather domain knowledge.
- Analyze the data.
- Calculate summary statistics.

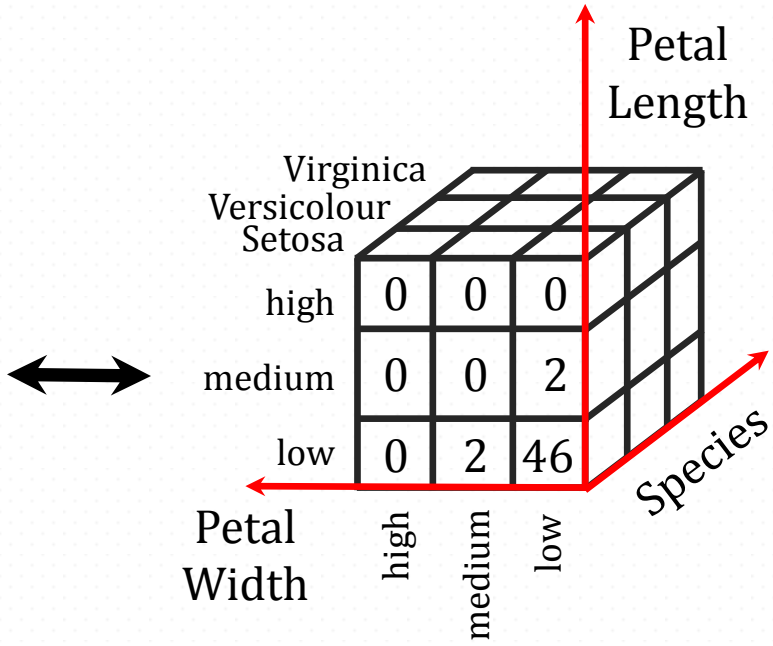
Data/OLAP Cubes

- Most datasets can be represented as a table.
- It is often possible to also represent data as a multidimensional array.



Data/OLAP Cubes: An Example

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44



A table (left) and multidimensional representation (right) for the Iris dataset. Each serve different purposes.

Data/OLAP Cubes

- Can perform several operations on these cubes.
  - Slice is the act of selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.
  - Dice involves selecting a subset of cells by specifying a range of feature values.
  - Roll-up and drill-down refer to aggregating (roll-up) or splitting (drill-down) features that have levels of granularity.
    - e.g., aggregating sales across all dates in a month (roll-up) or splitting monthly sales totals into daily sales totals (drill-down).

Summary Statistics

Summary statistics are the numbers that summarize properties of the data.

Summary Statistics

- Summarized properties include frequency, location, and spread.
- Most summary statistics can be calculated in a single pass through the data.

Frequency and Mode

- The frequency of a feature value is the percentage of time the value occurs in the dataset.
  - For example, given the feature ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of a feature is the most frequent feature value.
- The notions of frequency and mode are typically used with categorical data.

Percentiles

- For continuous data, the notion of a percentile is more useful.
- Given an ordinal or continuous feature x and a number p between 0 and 100, the pth percentile is a value x_p of x such that $p\%$ of the observed values of x are less than x_p.
  - For example, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$.

Measures of Location

- The mean is the most common measure of the location of a set of points, though is very sensitive to outliers:

$$mean(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Measures of Location

- The median, which is more resistance to outliers, is also commonly used:

$$median(x) = \begin{cases} x_{(r+1)} & if\ m\ is\ odd, m = 2r + 1 \\ 1/2\left(x_{(r)} + x_{(r+1)}\right) & if\ m\ is\ even, m = 2r \end{cases}$$

Measures of Spread

- The range is the difference between the maximum and minimum values.

Measures of Spread

- The variance or standard deviation is the most common measure of the spread of a set of points:

$$variance(x) = {s_x}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

- Additional measures, less sensitive to outliers, are also used.

Multivariate Measures

- The covariance is a measure of the degree to which two variables vary together, and is given by:

$$covariance(x_i, x_j) = \frac{1}{m-1}\sum_{i=1}^{m}(x_{ki} - \overline{x_i})(x_{kj} - \overline{x_j})$$

where x_{ki} and x_{kj} are the values of the ith and jth features for the kth object.

Data Understanding Steps

1. Collection of initial data
2. Description of data
3. Exploration of data
4. **Identifying data quality problems**

Finally, let's look at this step.

Data Quality

- Missing data
- Noise and artifacts
- Outliers
- Inconsistent data
- Duplicate data

Missing Data

- Various reasons: changes in experiment, measurement not possible, human error, combining various datasets, human bias.
- Key is to know how and why data is missing.
- Missing values can have a meaning.
  - Incorporate this in the model development.
  - Example: absence of a medical test indicates a particular prognosis (value of the missing feature).

Missing Data

Three types of missing data (or “missingness”):

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

Missing Completely at Random

- Missingness does not depend on any values of any variables in the dataset.
- Missingness depends on neither the values of the observed variables, nor on those of unobserved variables.

Missing at Random

- Missingness does not depend on the values of any of the missing or unobserved variables, but might depend on values of the observed variables.
- The pattern of missing values is identifiable.

Missing Not at Random

- Missingness depends on the values of the missing or unobserved variables.
- Pattern is non-random, non-ignorable, and typically arises due to the variable on which the data is missing.

MCAR Example

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	100,000
Customer 3	25	*Missing*
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	*Missing*

$$P(Balance\ Missing|Age = 25) = P(Balance\ Missing|Age = 60)$$

MAR Example

Customer	Age	Account Balance
Customer 1	25	*Missing*
Customer 2	25	100,000
Customer 3	25	*Missing*
Customer 4	60	50,000
Customer 5	60	120,000
Customer 6	60	150,000

The account balance is primarily observed only for $Age = 60$, thus the missingness can be modeled on age.

MNAR Example

Customer	Age	Account Balance
Customer 1	25	20,000
Customer 2	25	*Missing*
Customer 3	25	15,000
Customer 4	60	50,000
Customer 5	60	*Missing*
Customer 6	60	*Missing*

$$P(Balance\ Missing|Balance < 100{,}000) = 0$$
$$P(Balance\ Missing|Balance > 100{,}000) = 1$$

Noise and Artifacts

- Noise is the random component of a measurement error.
- The elimination of noise is frequently difficult.
- An important property of an algorithm is its "robustness to noise."
  - This is the stability of the algorithm on noisy data.
- Robust algorithms are often key to producing acceptable results even when noise is present.

Outliers

- Outliers are either:
  1. Data that have characteristics that are different from most of the other data.
  2. Values of a feature that are unusual with respect to the typical values for that feature.
- Unlike noise, outliers can be legitimate data or values.

Inconsistent Data

- Data can contain inconsistent values.
  - i.e., An address field with both ZIP code and city, but where the specified ZIP code area is not in the specified city.
- Some inconsistencies are easy to detect.
- Some inconsistencies may require consulting an external source.
- The correction of an inconsistency requires additional or redundant information.

Duplicate Data

- A dataset may include completely or partially duplicated data.
- Occasionally, two or more objects are identical with respect to the features measured by the database, but still represent different objects.

Noise and Artifacts

- Noise is the random component of a measurement error.
- The elimination of noise is frequently difficult.
- An important property of an algorithm is its "robustness to noise."
  - This is the stability of the algorithm on noisy data.
- Robust algorithms are often key to producing acceptable results even when noise is present.

Outliers

- Outliers are either:
  1. Data that have characteristics that are different from most of the other data.
  2. Values of a feature that are unusual with respect to the typical values for that feature.
- Unlike noise, outliers can be legitimate data or values.

Inconsistent Data

- Data can contain inconsistent values.
  - i.e., An address field with both ZIP code and city, but where the specified ZIP code area is not in the specified city.
- Some inconsistencies are easy to detect.
- Some inconsistencies may require consulting an external source.
- The correction of an inconsistency requires additional or redundant information.

Duplicate Data

- A dataset may include completely or partially duplicated data.
- Occasionally, two or more objects are identical with respect to the features measured by the database, but still represent different objects.

The Mantra

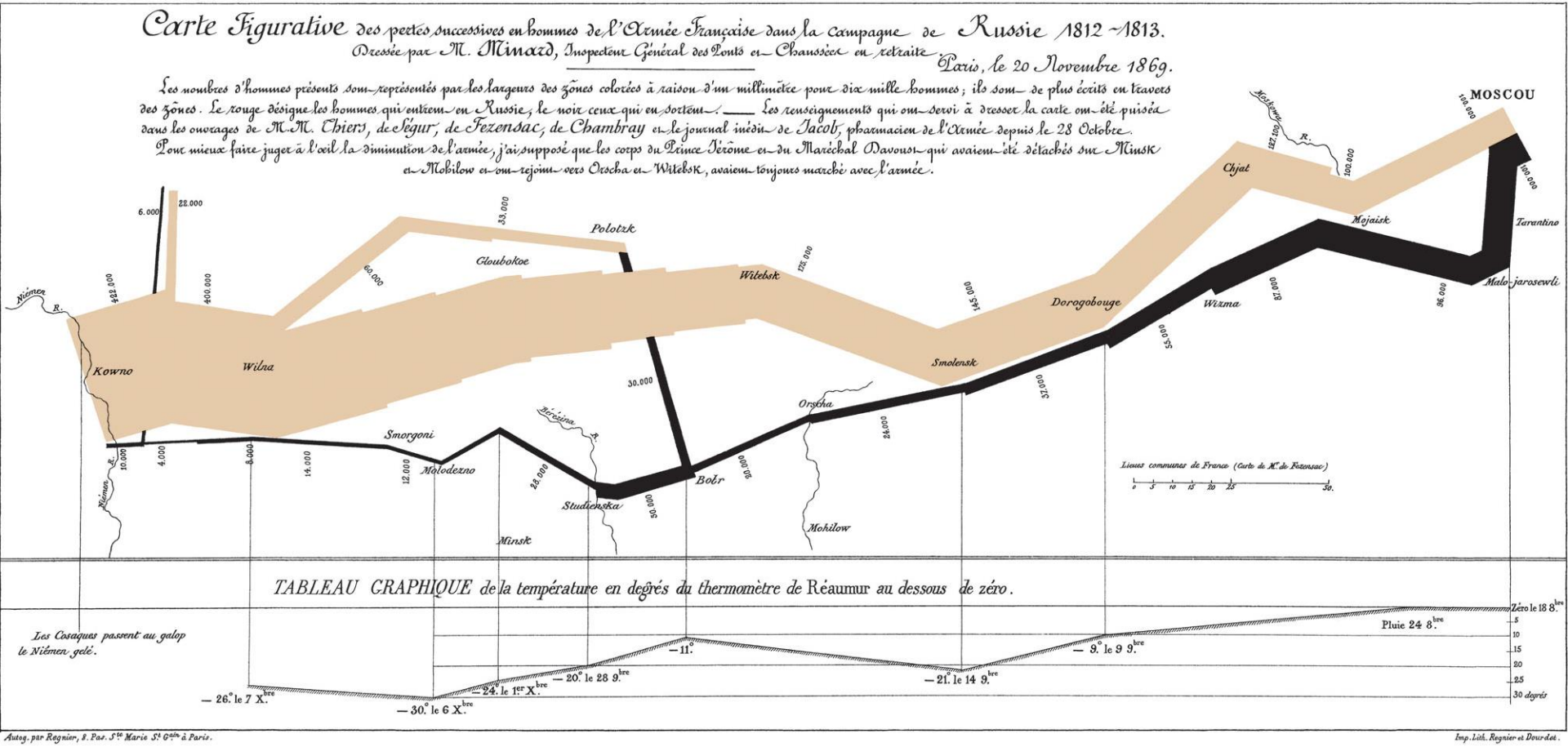
GIGO: Garbage in, Garbage Out

Successful data mining depends upon the data.

Exploring Data through Visualization

Data Understanding Steps

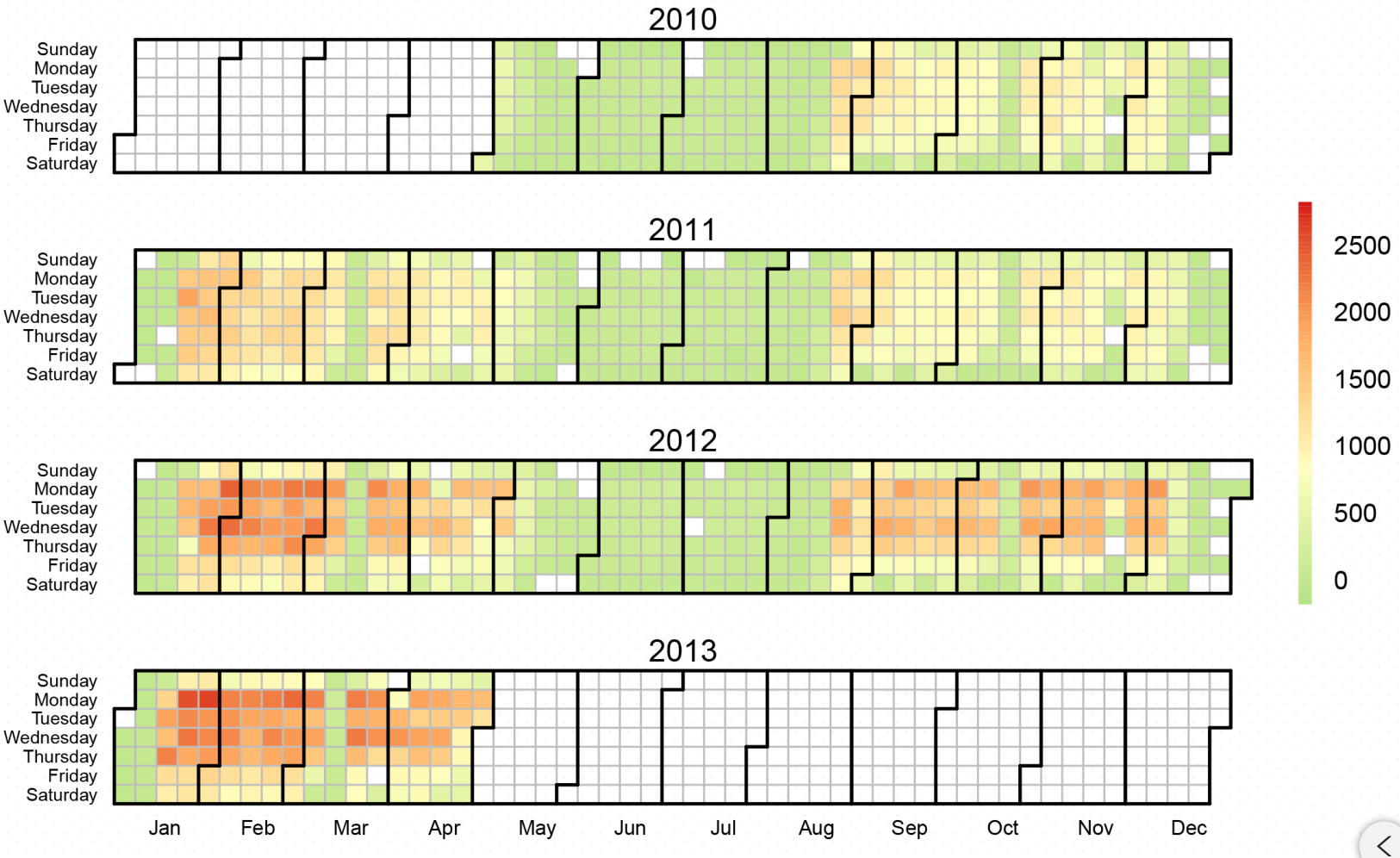
1. Collection of initial data
2. Description of data
3. Exploration of data
4. Identifying data quality problems



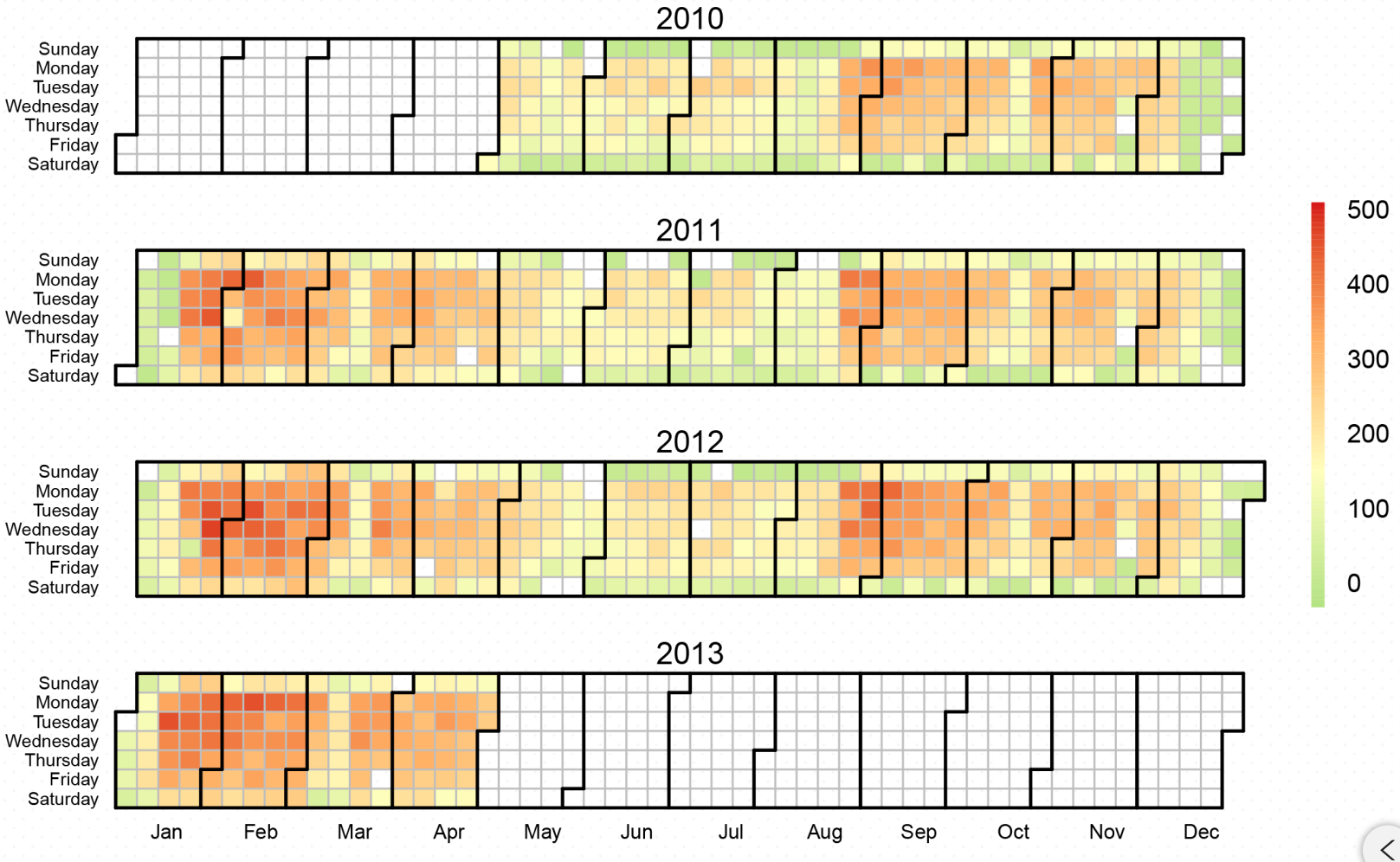
Napoleon's March to Moscow — Joseph Minard

Wikipedia.org. <http://en.wikipedia.org/wiki/File:Minard.png>

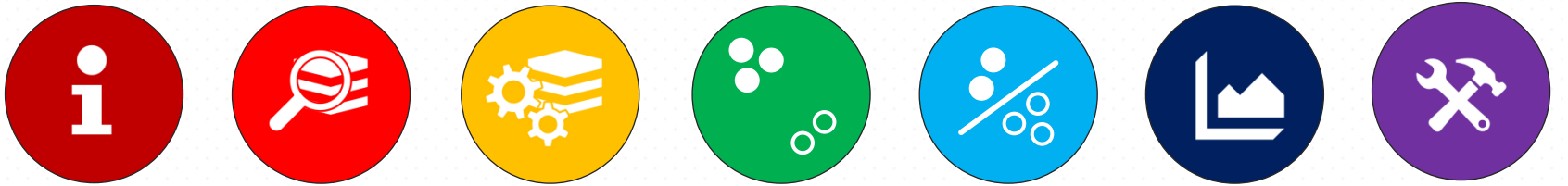
Calendar Heat Map of Undergraduate Student Visits to RecSports Facilities



Calendar Heat Map of Graduate Student Visits to RecSports Facilities



Course Topics

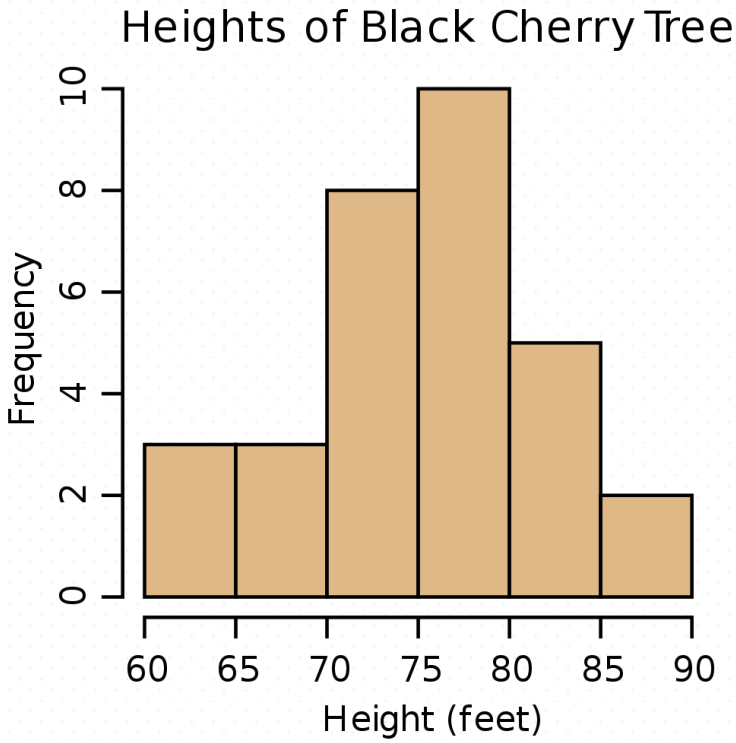


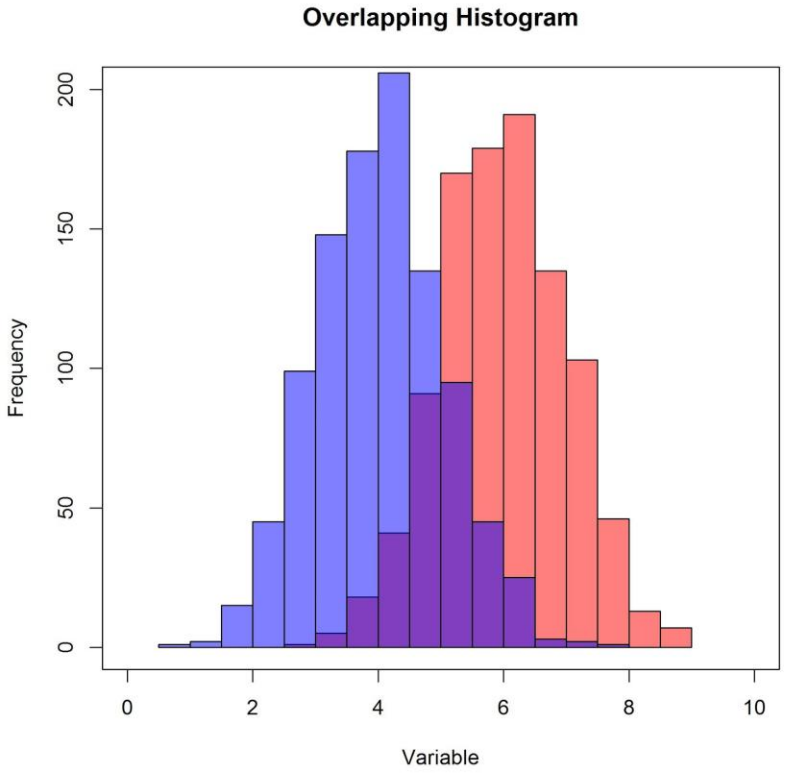
We are still here

Histograms

- Usually shows the distribution of values of a single variable of objects in each bin
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

Histograms


Heights of Black Cherry Trees
Frequency
0
2
4
6
8
10
60
65
70
75
80
85
90
Height (feet)

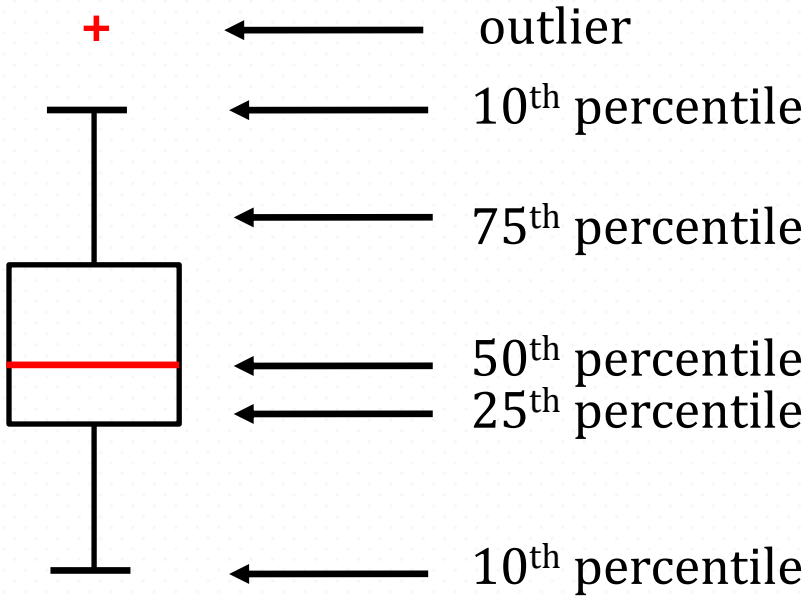

Overlapping Histogram
Frequency
0
50
100
150
200
0
2
4
6
8
10
Variable

Histograms Address Things Like

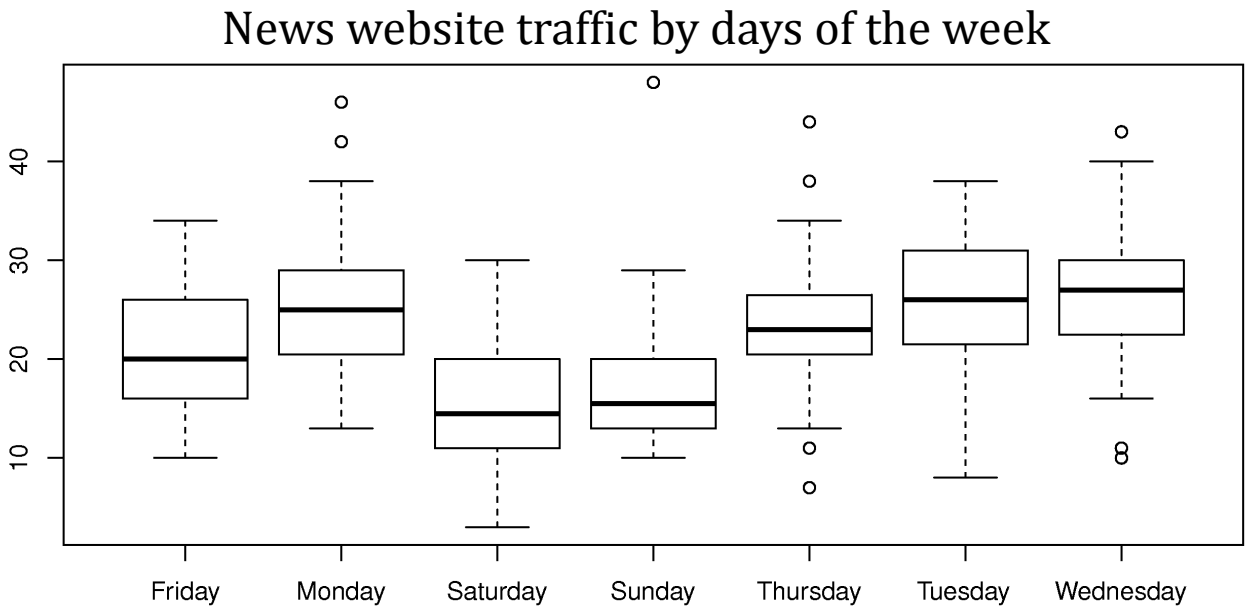
- What kind of population distribution do the data come from?
- Where are the data located?
- How spread out are the data?
- Are the data symmetric or skewed?
- Are the outliers in the data?

Box Plots

- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot



Box Plots


News website traffic by days of the week
10
20
30
40
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

Box Plots Address Things Like

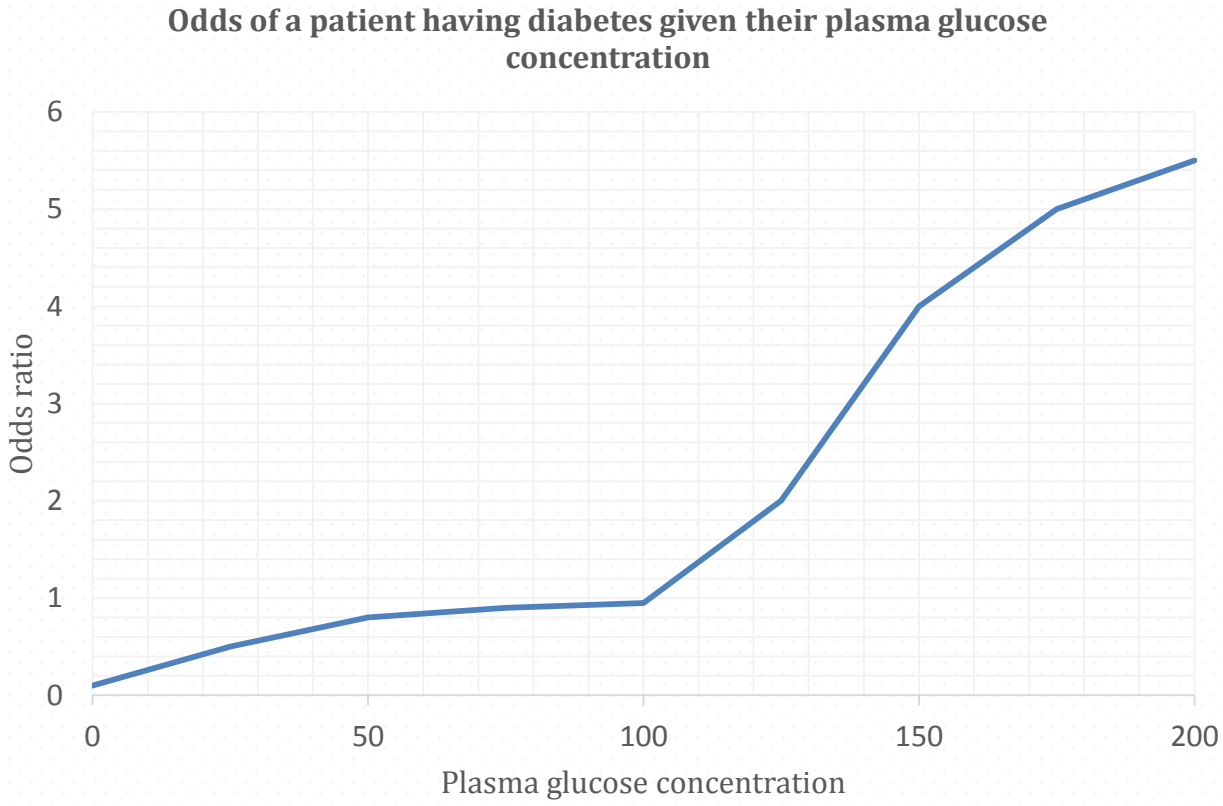
- Is a factor significant?
- Does the location differ between subgroups?
- Does the variation differ between subgroups?
- Are there any outliers?

Odds Plot

- An odds plot depicts the ratio:

$$OddsRatio(x_i, y) = \sum_{j \in K} \frac{p(x_{ij}|y = 1)}{p(x_{ij}|y = 0)}$$

Odds Plot



1 = patient has diabetes

0 = patient does not have diabetes

Scatter Plot

- Attribute values determine the position.
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots.
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects.
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes.

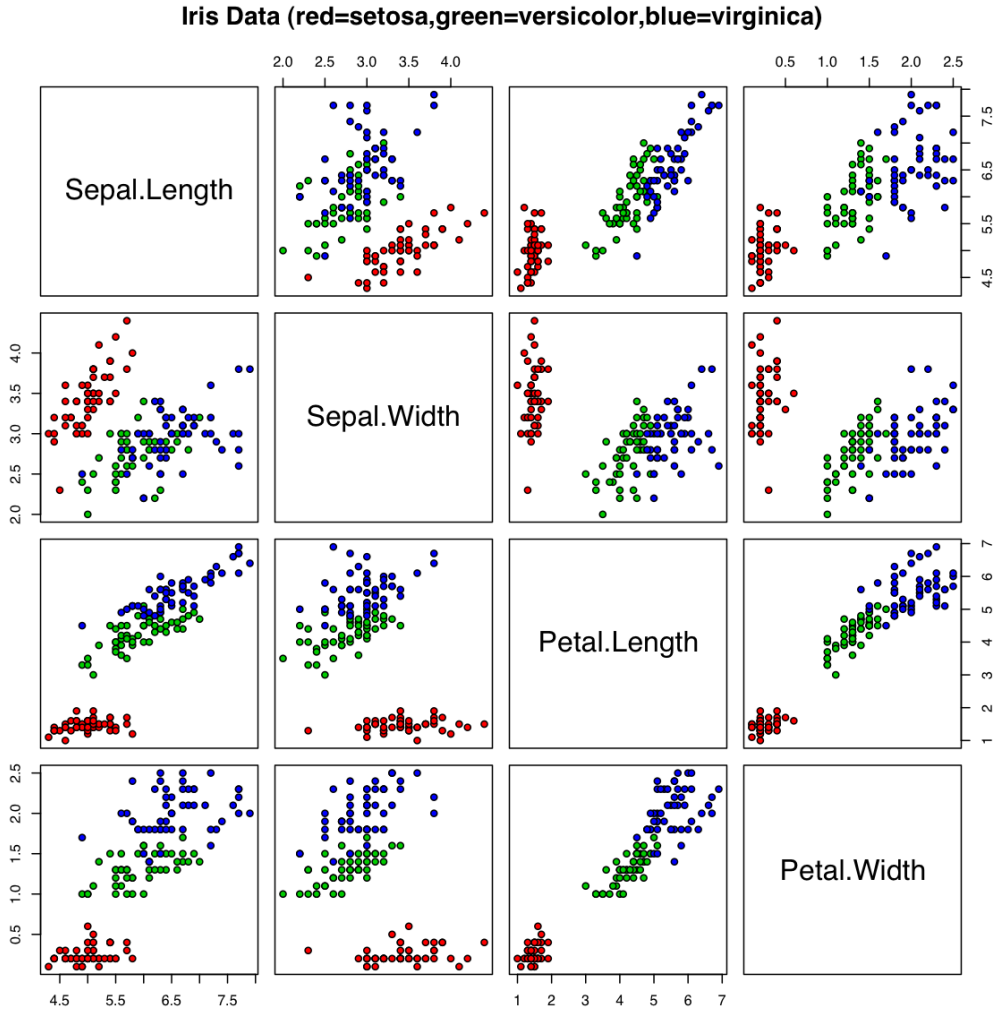
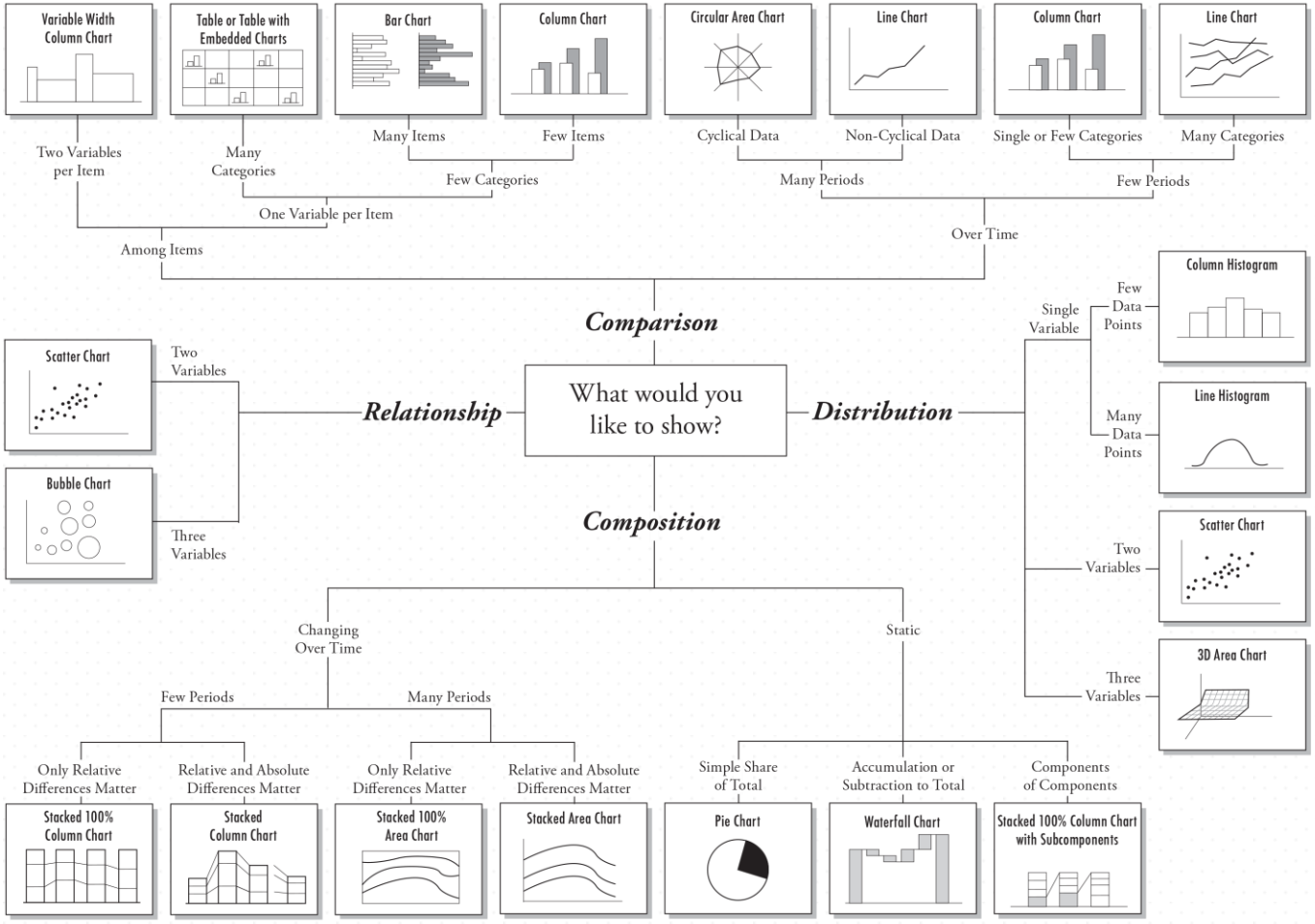
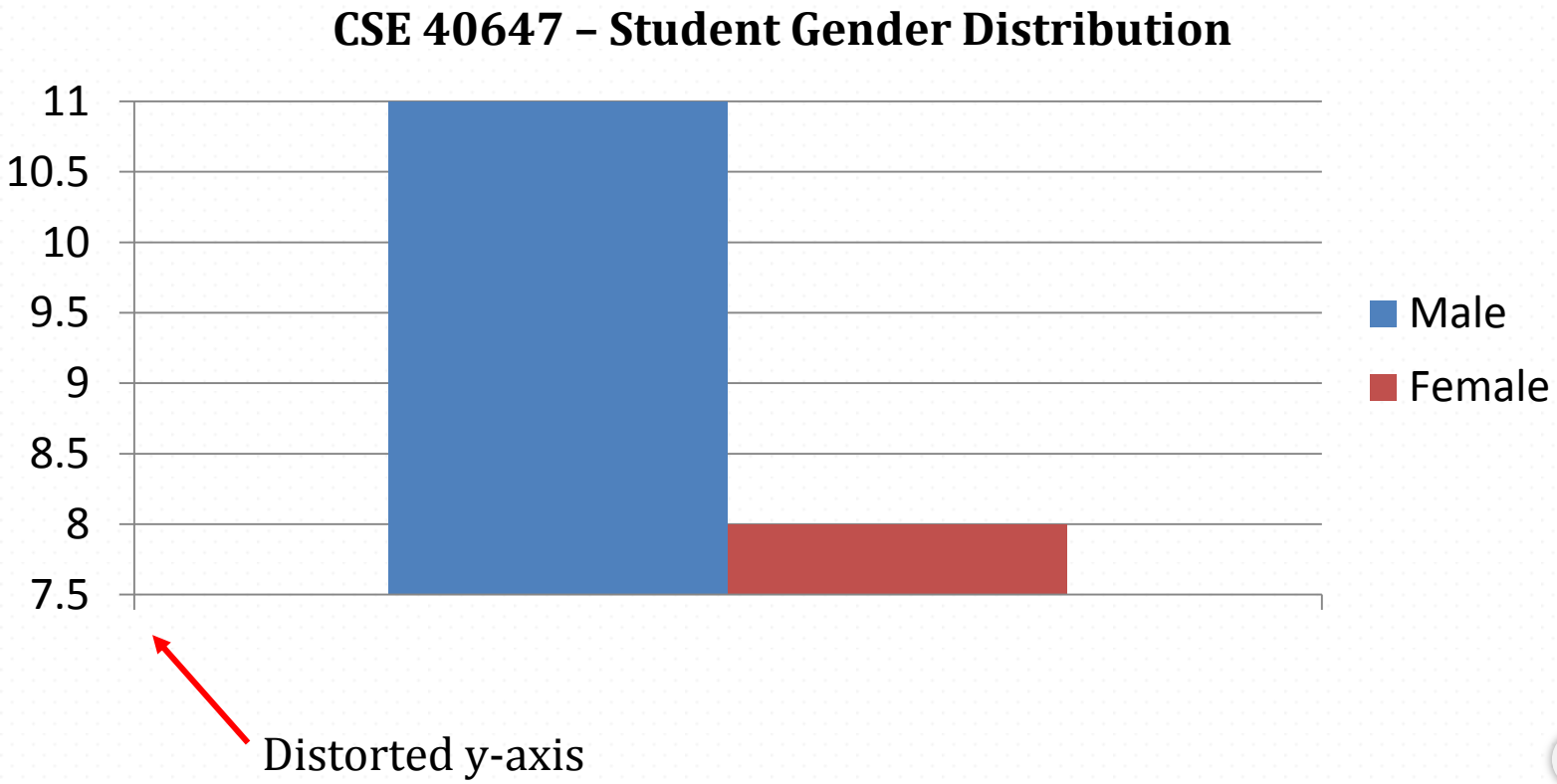

Iris Data (red=setosa,green=versicolor,blue=virginica)
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

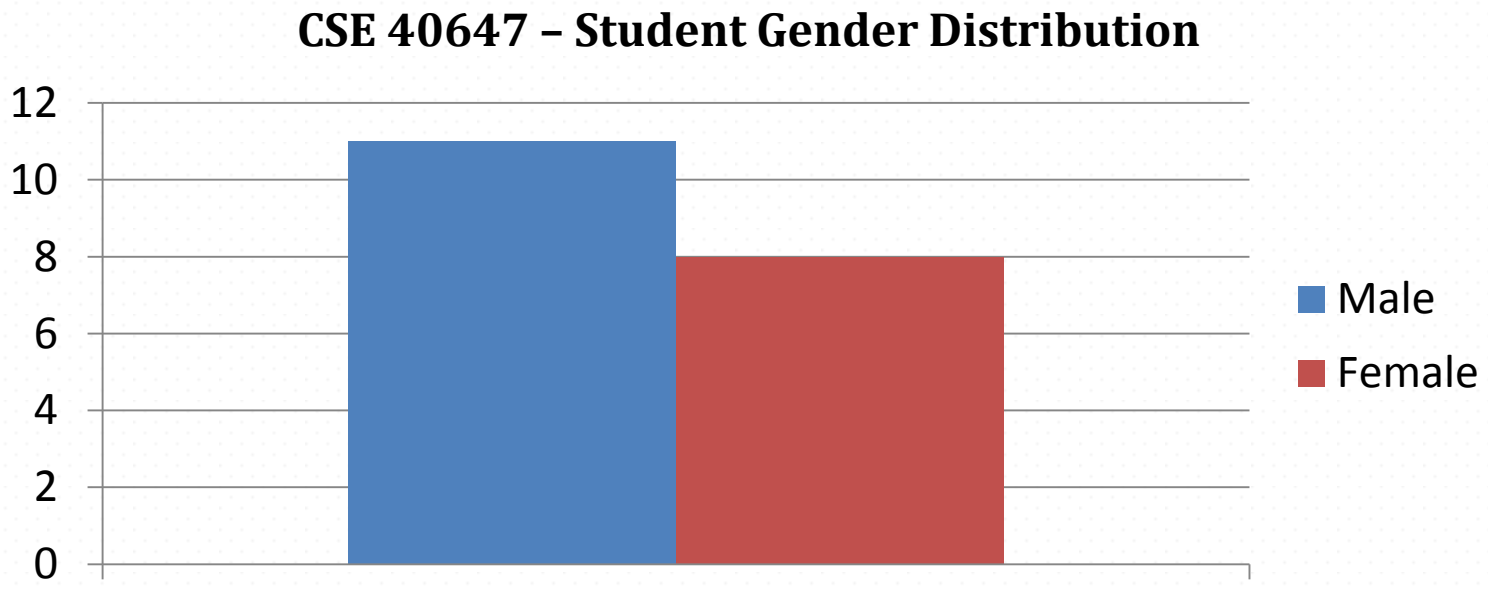
Chart Suggestions—A Thought-Starter



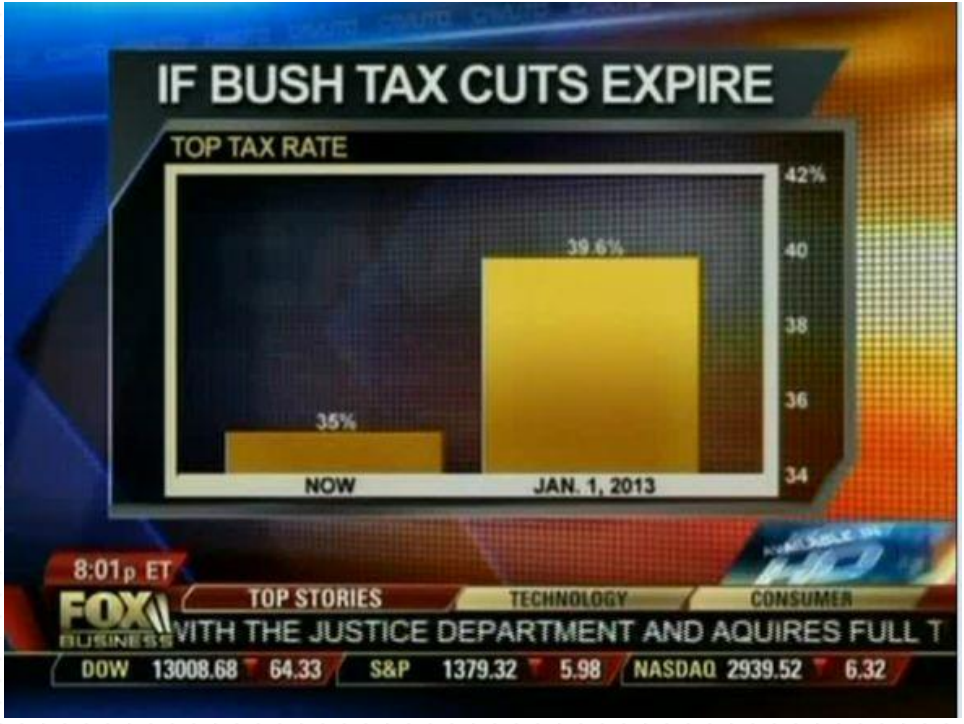
Design Principles: Graphical Integrity

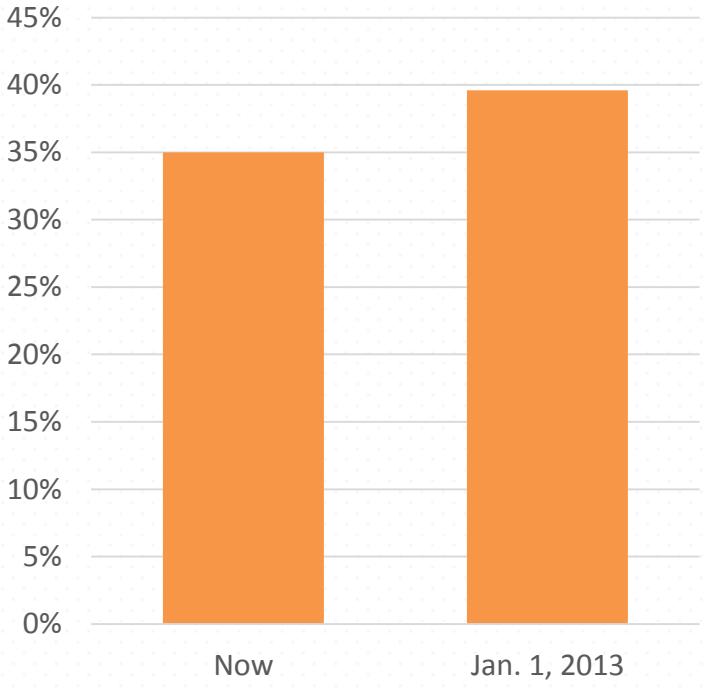


Design Principles: Graphical Integrity

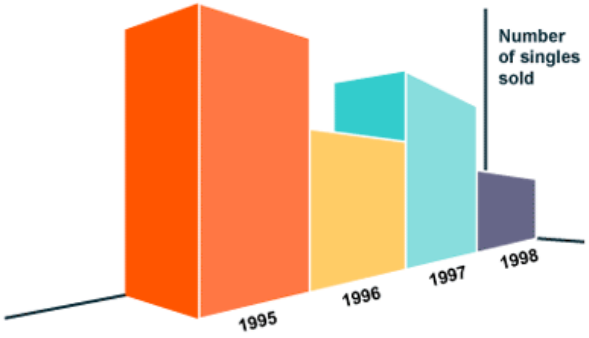


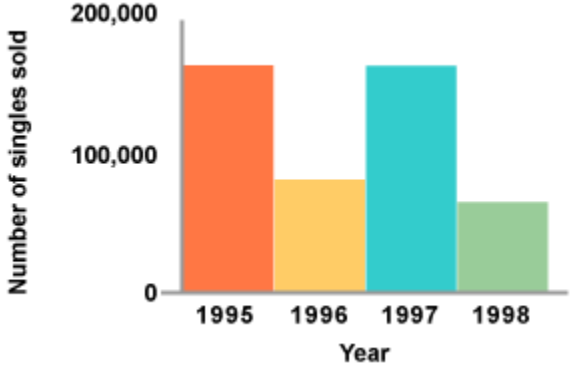
Design Principles: Graphical Integrity





Design Principles: Graphical Integrity

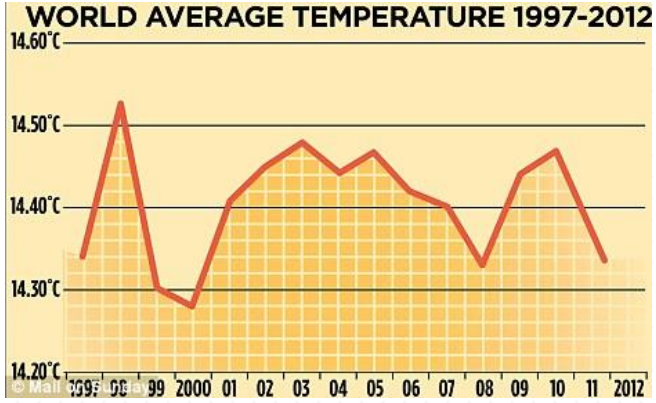


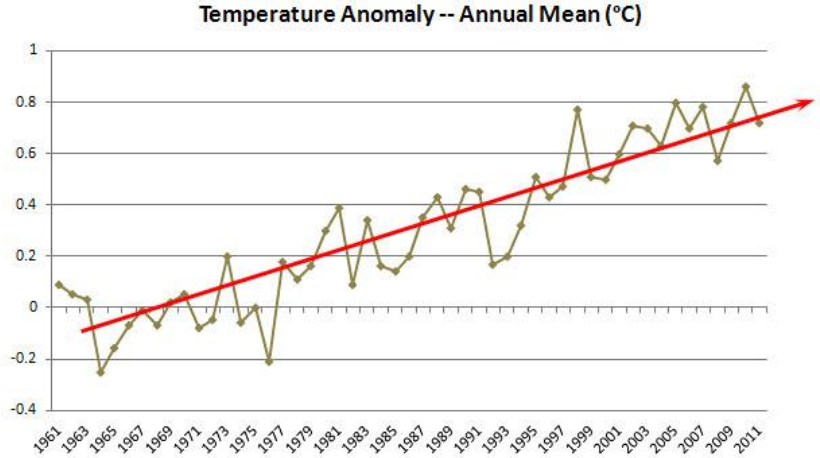


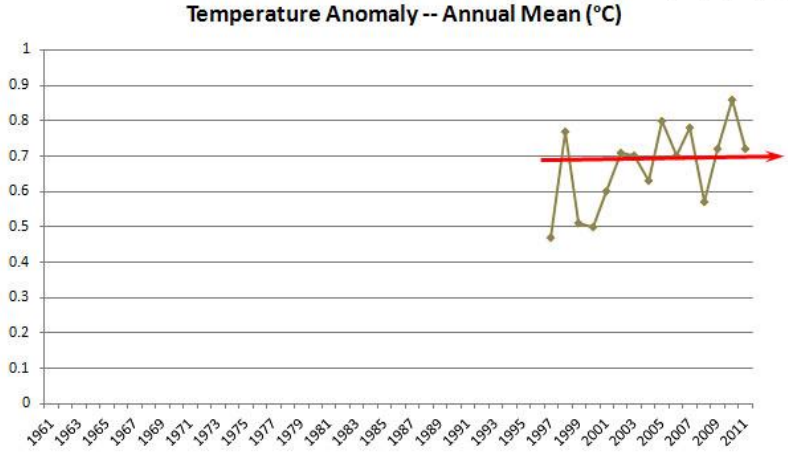
Design Principles: Graphical Integrity

Design Principles: Graphical Integrity

Global warming?

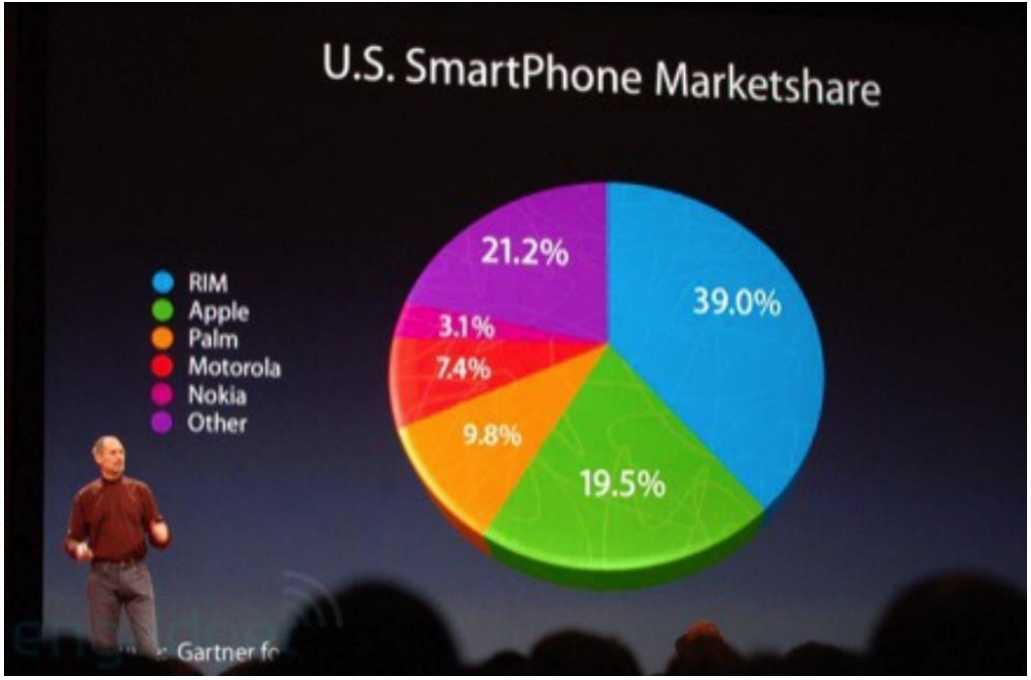


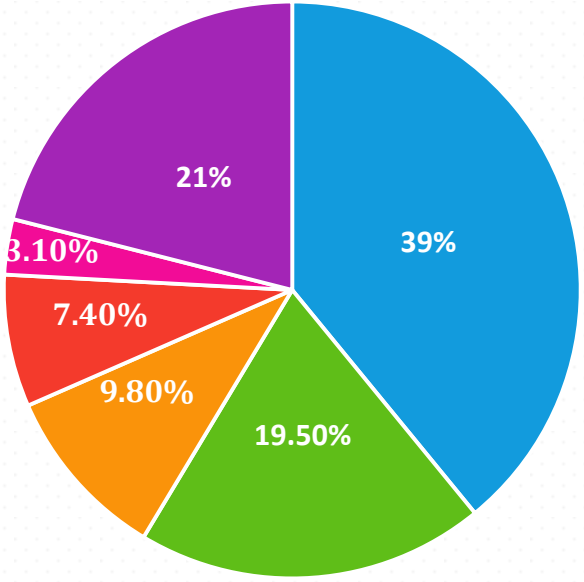




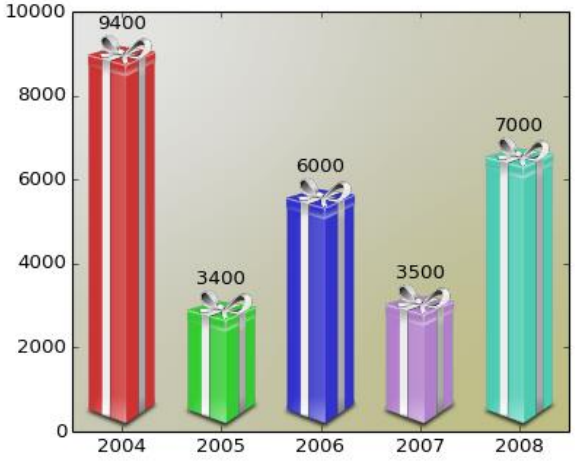
Global warming!
☹

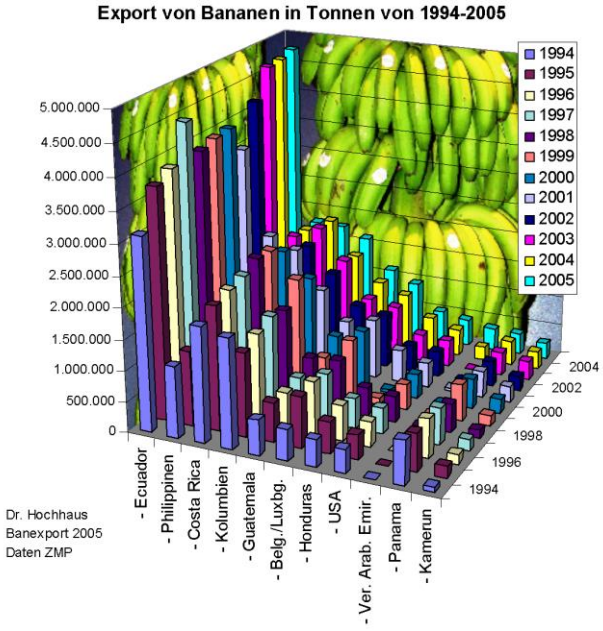
Design Principles: Graphical Integrity

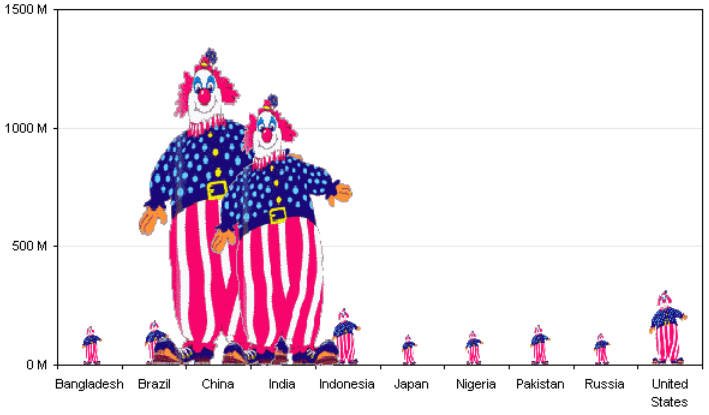




Design Principles: Don't Do It!







Design Principles: The Lie Factor

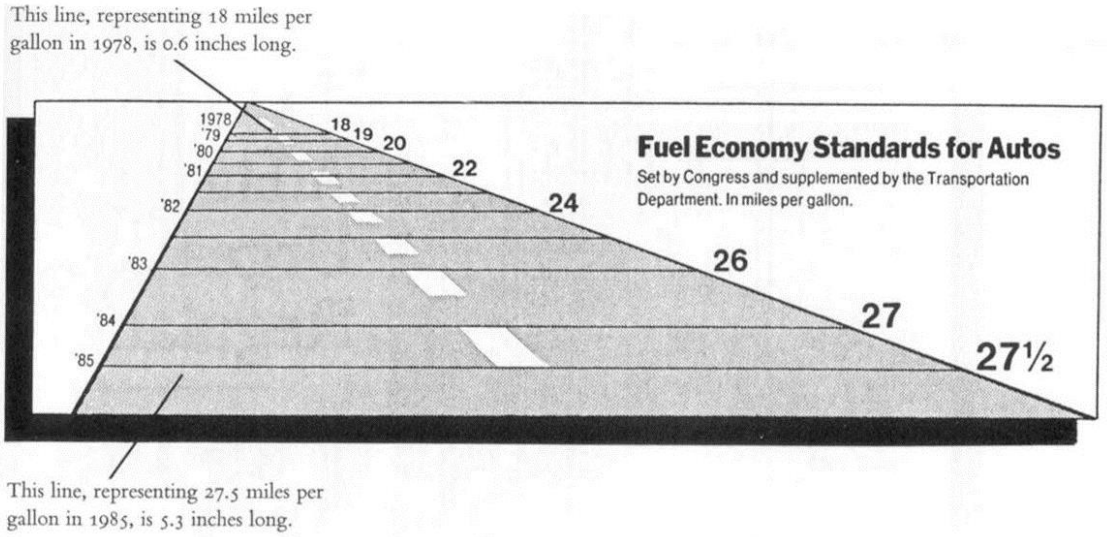
- Coined by Eduard Tufte, the *Lie Factor* is defined to be a measure of the amount of "distortion" in a graph. That is:

$$\text{Lie Factor} = \frac{\text{size of effect shown in graph}}{\text{size of effect shown in data}}, \text{where}$$

$$\text{size of effect} = \frac{|\text{second value} - \text{first value}|}{\text{first value}}$$

- If the lie factor is greater than 1, the graph is exaggerating the size of the effect.

Design Principles: The Lie Factor



New York Times, August 9, 1978, p. D-2.

$$\text{Lie Factor} = \frac{\frac{5.3 - 0.6}{0.6}}{\frac{27.5 - 18}{18}} = \mathbf{14.8}$$

Design Principles

- In summary you should:
  - NOT try to deceive your audience
  - Avoid 3D
  - Keep “chartjunk” to a minimum to prevent distractions

Summarizing Data Understanding

Summarizing Data Understanding

- Data Understanding consists of the collection, description, exploration of data, culminating in the identification of data quality problems.
- Data can be described by characteristics such as dimensionality, sparsity, and resolution.
- Data exploration involves statistical analysis of data.
- Data quality concerns include missing data, noise and artifacts, outliers and anomalies, inconsistent data, and duplicate data.

And Now...

Let's see some real-time visualizations!

Preprocessing the Data

Data Preprocessing

The process of making the data more suitable for data mining.

Data Preprocessing

The process of making the data more suitable for data mining.

The tasks employed in this process are informed by the process of data understanding.

Data Preprocessing Tasks

1. Data Cleaning
2. Data Transformation
3. Data Reduction
4. Data Discretization

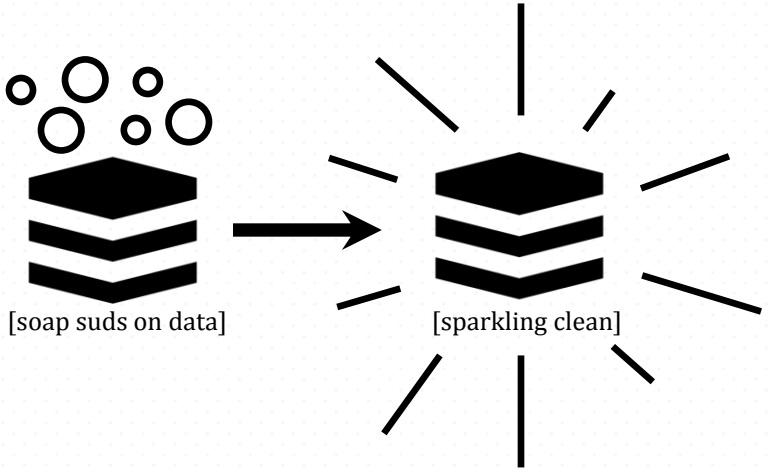
Data Preprocessing Tasks

1. **Data Cleaning**
   Let's start by looking at this task.
2. Data Transformation
3. Data Reduction
4. Data Discretization

Data Cleaning

Data cleaning involves the correction of data quality problems. These tasks include:

- Filling in missing data
- Smoothing-out noisy data
- Removing outliers and artifacts
- Correcting inconsistent data
- Removing duplicate data



Filling-in Missing Data

Ignore the instance: often not very effective, especially when few features are missing.

Fill in the missing value manually: tedious and typically infeasible.

Use a global constant to fill in the missing value: e.g., "unknown". May be mistaken for concept.

Imputation: fill in the missing value using the feature mean or the most probable value.

Imputing Missing Data

- Delete missing observations
  - Can lead to serious biases.
  - If missing data is relatively small, may be okay.
- Cold-deck imputation
- Hot-deck imputation
- Distribution-based imputation
- Statistical imputation
- Predictive imputation

Cold-Deck Imputation

- Fill in the data using means or other analysis of the variable to fill in the value.
- Measure of central tendency (mean, median, mode)

Hot-Deck Imputation

- Identify the most similar case to the case with a missing value and substitute the most similar case's value for the missing case's value.
- Advantages: simplicity, maintains level of measurement, complete data at the end.
- Disadvantage: can identify more than one similar case and randomly select or use average.

Distribution-based Imputation

- Assign value based on the probability distribution of the non-missing data.
- Tries to capture the “observed” empirical distribution of data.

EM Imputation

- Expectation—Maximization (EM)
- Iterative, 2 steps:
  - E step: estimate distributions of all missing variables using a guessed parameter estimate
  - M step: using those distributions, compute a new ML estimator of the parameter
  - Repeat until convergence obtained
- Does not need to estimate actual data points.

Statistical Imputation

- Build a regressor to classify the input value
  - Consider the “missing” value as the “output” and the rest of the features as input
- Imputes the value based on other features

Predictive Imputation

- Let a classifier model the underpinnings of the missingness mechanism.

Smoothing-out Noisy Data

- Noise: Random error or variance in a measured variable.
- **Binning**: Smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of *bins*.
- **Clustering:** Detect and remove outliers.
- **Regression:** Smooth by fitting the data into regression functions.

Binning: Simple Discretization Methods

Equal-width (distance) partitioning:

- It divides the range into N intervals of equal size
- If A and B are the lowest and highest values of the attribute, the width of intervals with be: $W = (B - A)/N$.
- The most straight-forward
- But outliers may dominate presentation
- Skewed data is not handled well.

Binning: Simple Discretization Methods

Equal-depth (frequency) partitioning:

- It divides the range into *N* intervals, each containing approximately the same number of samples
- Good data scaling.
- Managing categorical features can be tricky.

Binning Methods for Data Smoothing

Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into (equal-depth) bins:

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

Binning Methods for Data Smoothing

Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Smoothing by bin means:

Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

Bin 3: 29, 29, 29, 29

Binning Methods for Data Smoothing

Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Smoothing by bin boundaries:

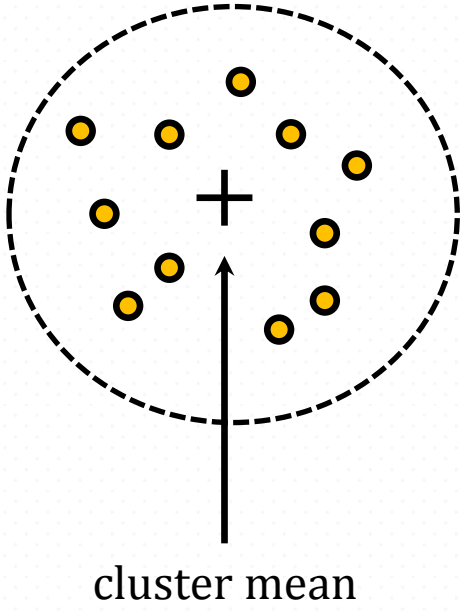
Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

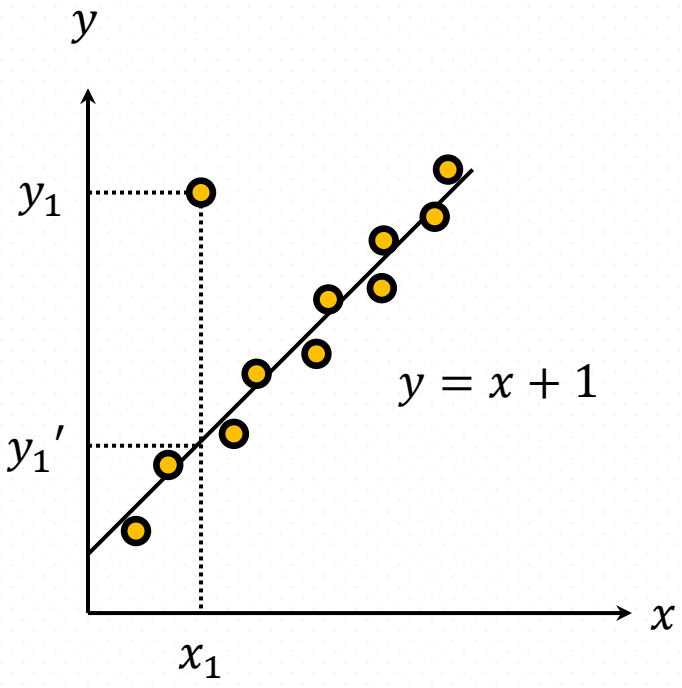
Clustering for Data Smoothing

- Cluster the data and use properties of the clusters to represent the instances constituting those clusters.



Regression for Data Smoothing

- Data can be smoothed by fitting the data to a function, such as with regression.



Removing Outliers and Artifacts

- **Proximity-based Techniques**: It is often possible to define a proximity measure between objects, with outliers being distant from most of the other data.
- **Density-based Techniques:** An outlier has a local density significantly less than that of most of its neighbors.

Correcting Inconsistent Data

- Some types of inconsistences are easy to detect.
  - e.g., a person's height should not be negative
- In other cases, it can be necessary to consult an external source of information

Removing Duplicate Data

- Removing duplicate data raises two issues:
  1. If there are two objects that actually represent a single object, then the values of corresponding features may differ, and these inconsistent values must be resolved.
  2. Care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.
- The term deduplication is often used to refer to the process of dealing with these issues.

And Now...

Let's clean some data!