Midterm Review

Broad and Narrow Definitions of Data Mining

- Broad -> traditional statistical methods
- Narrow - > automated and heuristic methods
- Heuristics – exploratory problem-solving techniques that give a non-optimal solution (exhaustive search impractical)
  - Rule of thumb, educated guess, intuitive judgment, stereotyping, common sense
  - Computer Science – technique used when classic methods are too slow; approximate solution after exact solution not found.

Process

1. Develop understanding of application, goals
2. Create dataset
3. Data cleaning and preprocessing
4. Data reduction and projection
5. Choose data mining task
6. Choose data mining algorithms
7. Use algorithms to perform task
8. Interpret and iterate thru 1-7 *if necessary*
9. Deploy: integrate into/create new operational system

Data Mining

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- Difference between **structured and unstructured text**

IR vs. databases: Structured vs. unstructured data

- Structured data tends to refer to information in "tables"

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (for text) queries, e.g.,

Salary < 60000 AND Manager = Smith

Unstructured data

- Typically refers to free text
- Allows
  - Keyword queries including operators
  - More sophisticated “concept” queries e.g.,
    - find all web pages dealing with *drug abuse*
- Classic model for searching text documents

- Twitter, facebook, instagram, all social media

Semi-structured data

- In fact almost no data is “unstructured”
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
  - … to say nothing of linguistic structure
- Facilitates “semi-structured” search such as
  - *Title* contains data AND *Bullets* contain search
- Or even
  - *Title* is about Object Oriented Programming AND *Author* something like stro*rup
  - where * is the wild-card operator

Authoritative metadata

- AKA ‘top-down’
- Created by project team
- Formalized; focus on control
- Specialists in (at least one aspect of) the field
- Focus and coverage will depend on the requirements of the project and agency

User-created metadata

- AKA ‘bottom-up’
- Social tagging
  - May be open or within a community
- Less focused; what the “tagging public” sees
- Generally less structured, not prescriptive

Types of metadata

- **Descriptive:** Facilitates discovery and describes intellectual content
- **Administrative:** Facilitates management of digital and analog resources
- **Technical:** Describes the technical aspects
- **Structural:** Describes the relationships within object
- **Preservation:** Supports long-term retention and may overlap with technical, administrative, and structural metadata

Metadata formats

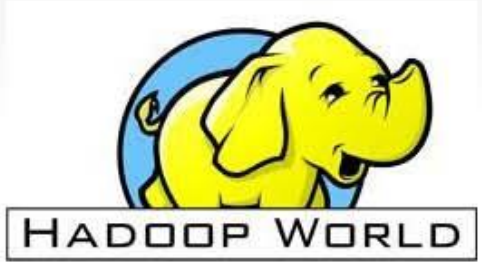
- Extensible Markup Language (XML)
  - Allows for combining and interoperability
  - XML flexibility
- Any other conceivable format
  - MS Word? PDF? Post-it notes?
  - Excel, FileMaker Pro, Access DB, CSV

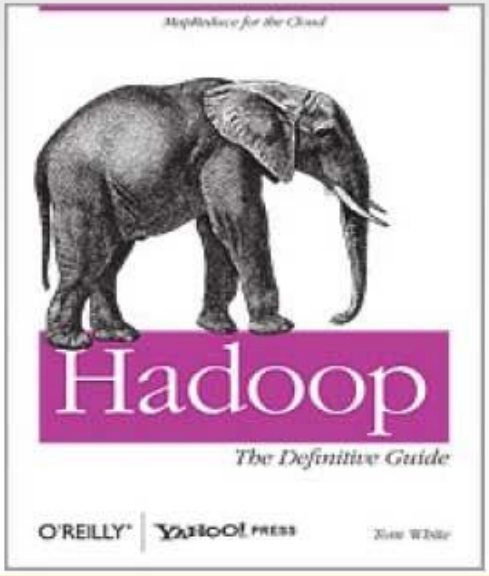
Five Stars of Data Engagement

- Be demand driven
- Put data in context
- Support conversation around data
- Build capacity, skills, and networks
- Collaborate on data as a common resource

- *http://www.opendataimpacts.net/engagement/*

The Hadoop Industry





Large scale processing on data clusters: **hadoop.apache.org**

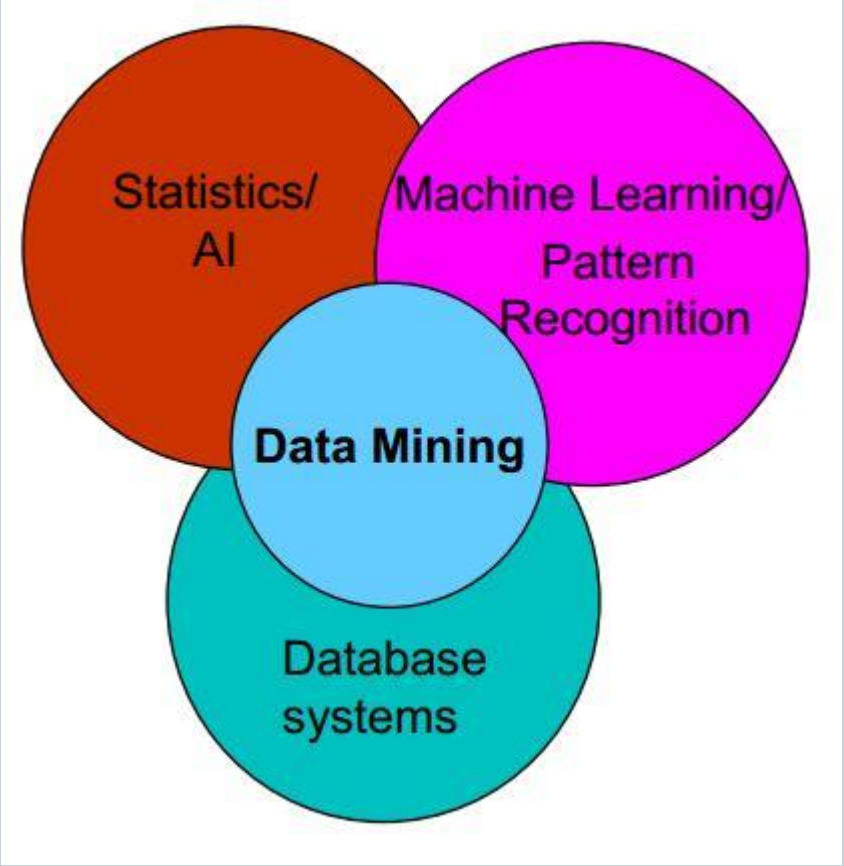
siliconangle.com/blog/2014/09/08/what-you-missed-in-big-data-hadoop-bandwagon-gets-even-more-crowded/

www.zdnet.com/teradata-acquires-hadoop-consulting-firm-think-big-analytics-7000033220/

Amazon web services

- launched 2006
- largest, most popular cloud computing platform
  - others: Rackspace, Azure, Google App Engine

Core Disciplines of Data Mining



Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.

Data Mining Tasks

- Classification *[Predictive]*
- Clustering *[Predictive]*
- Association Rule Discovery *[Descriptive]*
- Sequential Pattern Discovery *[Descriptive]*
- Natural Language Processing *[Descriptive]*
- Regression *[Predictive]*
- Deviation Detection *[Predictive]*

Continuous nature of algorithms

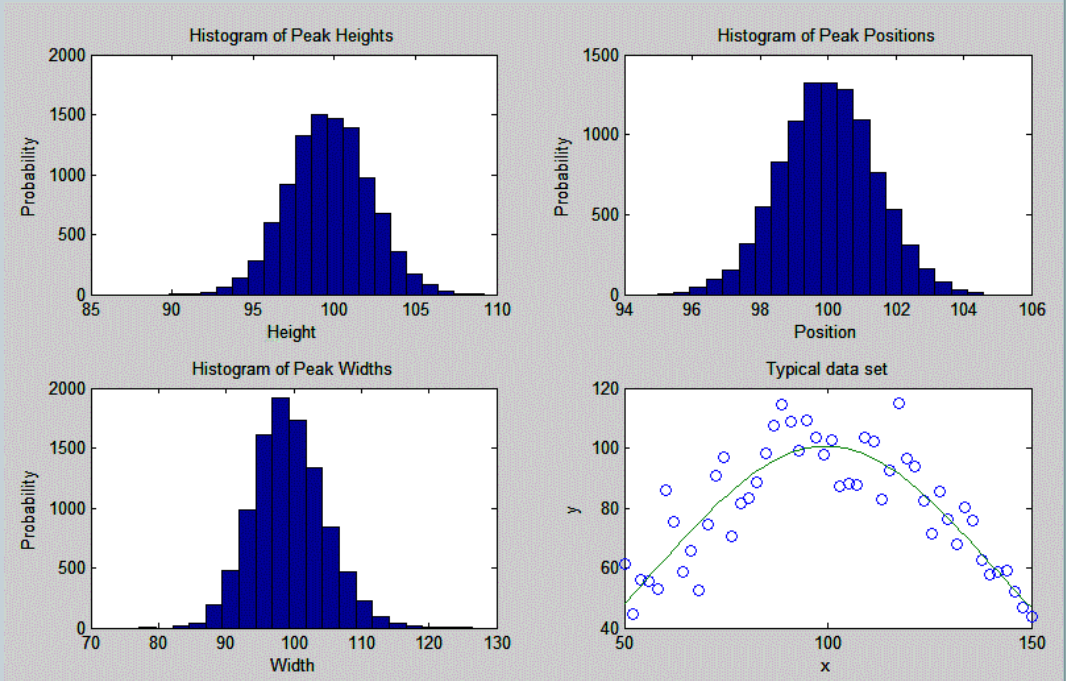
- Approximating an integral
- Solving a system of linear equations
- Finding the roots of a function
- Solving a differential equation

1. $\left(\frac{d^3y}{dx^3}\right)^4 + 2\frac{dy}{dx} = \sin x$
2. $\frac{dy}{dx} - 2xy = x^2 - x$
3. $\frac{dy}{dx} - \sin y = -x$
4. $\frac{d^2y}{dx} = 2xy$

www.analyzemath.com

Discrete nature of algorithms

- Text string
- Sorting a list of objects
- Optimal path between cities
- Simulating a random process
- Finding a point from a list which is closest to a given point



Pseudocode

- Developed out of a clear and concise way to describe an algorithm which is not language dependent
- Combination of common language and terminology (i.e., loops and conditionals)
- Does NOT contain correct syntax for any language because it is language independent

Random Sample

- A random sample is a sample in which each individual or object in the population has an equal chance of being selected.

- A *random process* – a collection of random variables – is often called a *stochastic process* in probability theory.

NO QUESTIONS

ON DATABASES

&

NATURAL LANGUAGE PROCESSING

Evaluation Data Mining Techniques

- http://ijcsi.org/papers/7-5-181-186.pdf

IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010
ISSN (Online): 1694-0814
www.IJCSI.org

A New Approach for Evaluation of Data Mining Techniques

Moawia Elfaki Yahia[1], Murtada El-mukashfi El-taher[2]

[1]College of Computer Science and IT
King Faisal University
Saudi Arabia, Alhasa 31982

[2]Faculty of Mathematical Sciences
University of Khartoum
Sudan, Khartoum 11115

Abstract

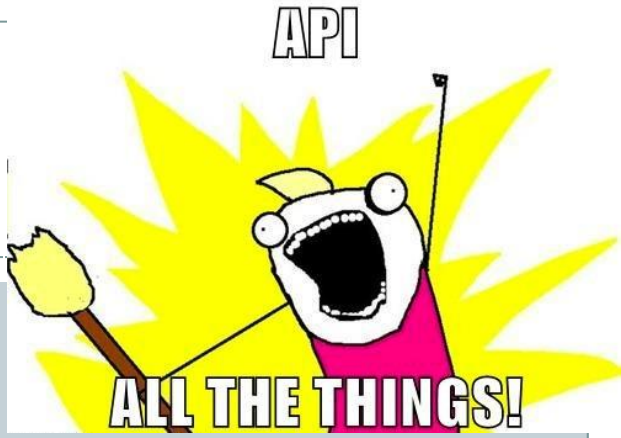
This paper tries to put a new direction for the evaluation of some techniques for solving data mining tasks such as: Statistics, Visualization, Clustering, Decision Trees, Association Rules and Neural Networks. The new approach has succeed in defining some new criteria for the evaluation process, and it has obtained valuable results based on what the technique is, the environment of using each techniques, the advantages and disadvantages of each technique, the consequences of choosing any of these techniques to extract hidden predictive information from large databases, and the methods of implementation of each technique. Finally, the paper has presented some valuable recommendations in this field.

Keywords:*Data Mining Evaluation, Statistics, Visualization, Clustering, Decision Trees, Association Rules, Neural Networks.*

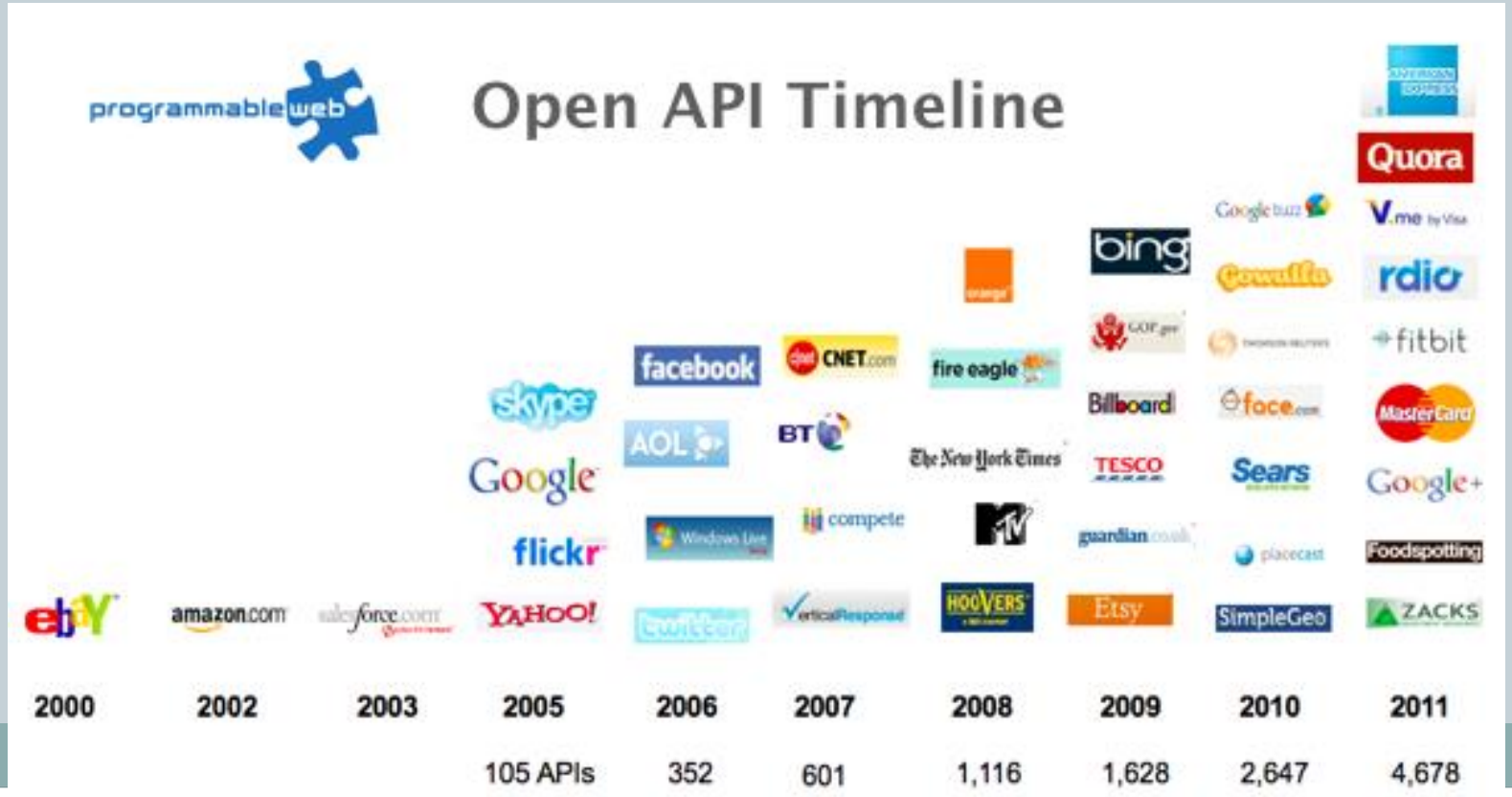
2. Data Mining Overview

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses [6]. Data mining tools predict future trends and behaviors allowing businesses to make proactive knowledge driven decisions. Data mining tools can answer business question that traditionally were too time consuming to resolve. They scour database for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Definitions



- **API**
  - Application programming interface



Human User

- Data mining algorithms often trying to extract patterns, locations that would be obvious to a human evaluator given a small scale
  - NLP
  - Geolocation
- Gold Standard = predefined evaluation
  - Rubric

Human Gold Standard

- Becomes a predefined evaluation of the results of your data mining algorithm given its performance vs a human evaluator

- Geolocation vs. Human-assigned location from context
- NLP POS-tagging vs. Human-assigned POS tagging

GISystem vs. GIScience

GISystem – software

GIScience – science behind the software

e.g. database structures, qualitative spatial reasoning, spatial web services, spatial information management and distribution……

Challenges of Big-*m*, Big-*n* Data

Increasing dimensionality (Big n) ⟶

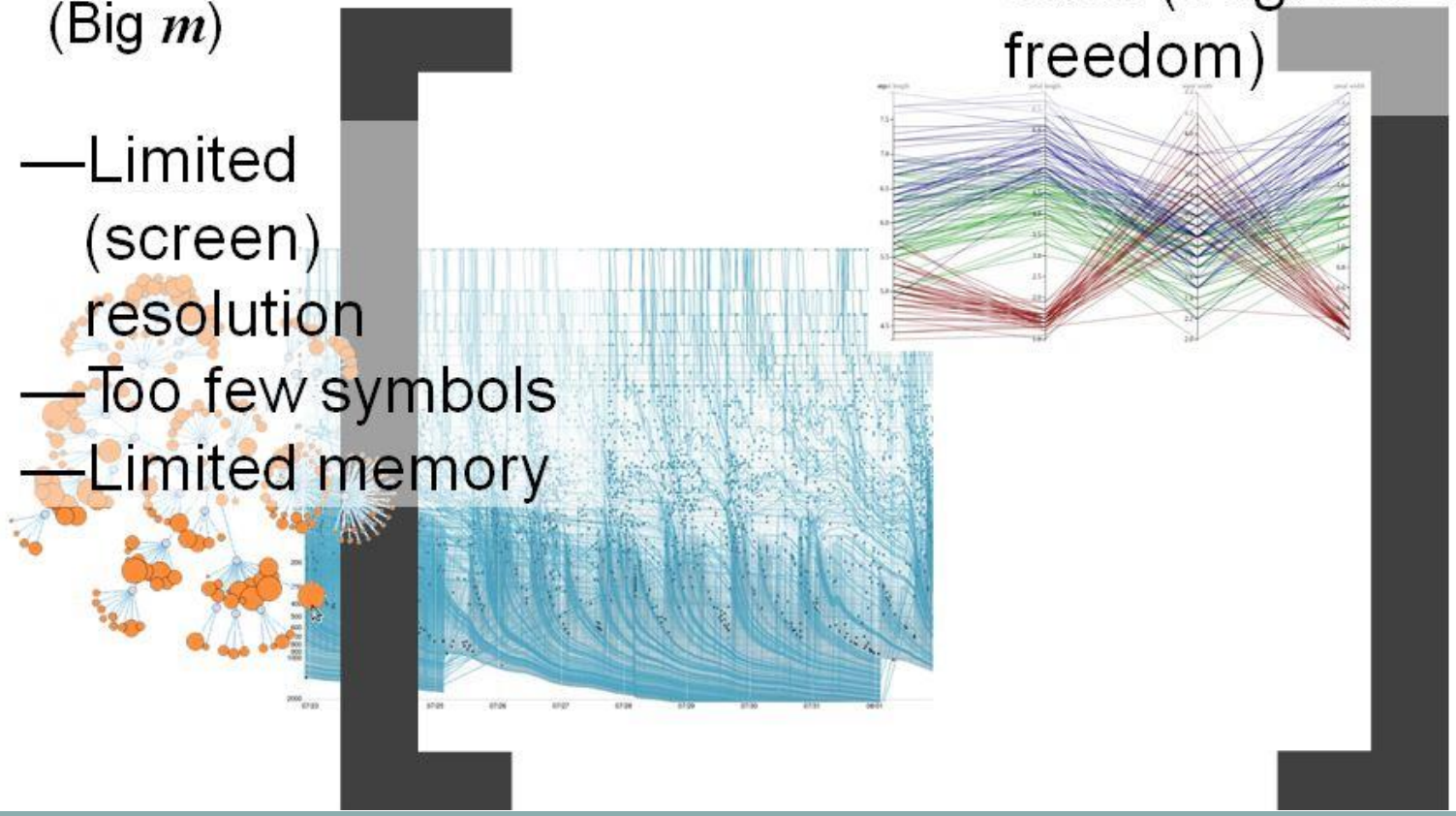
↓ Increasing resolution (Big m)

—Too few sense cues (degrees of freedom)

—Limited (screen) resolution

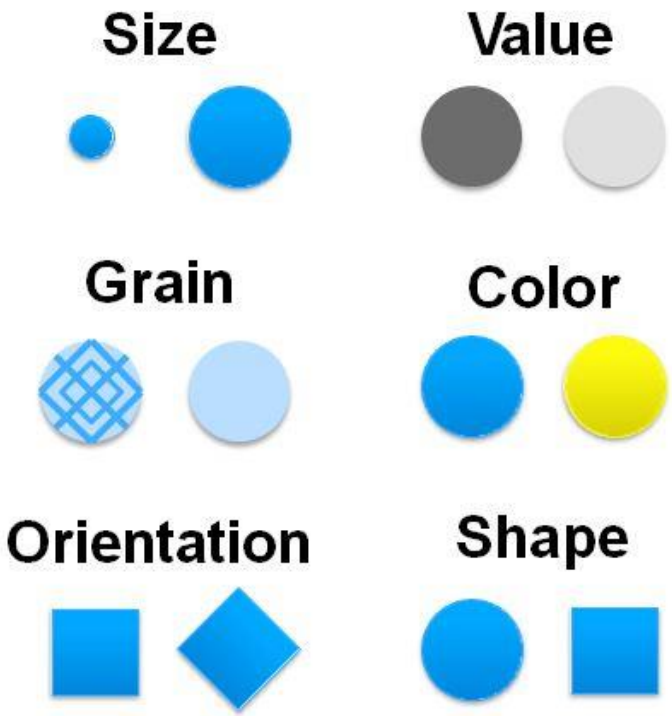
—Too few symbols

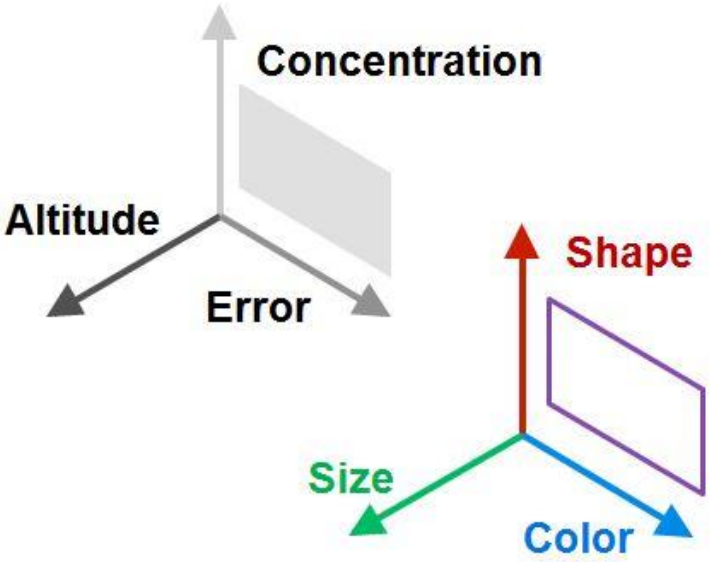
—Limited memory



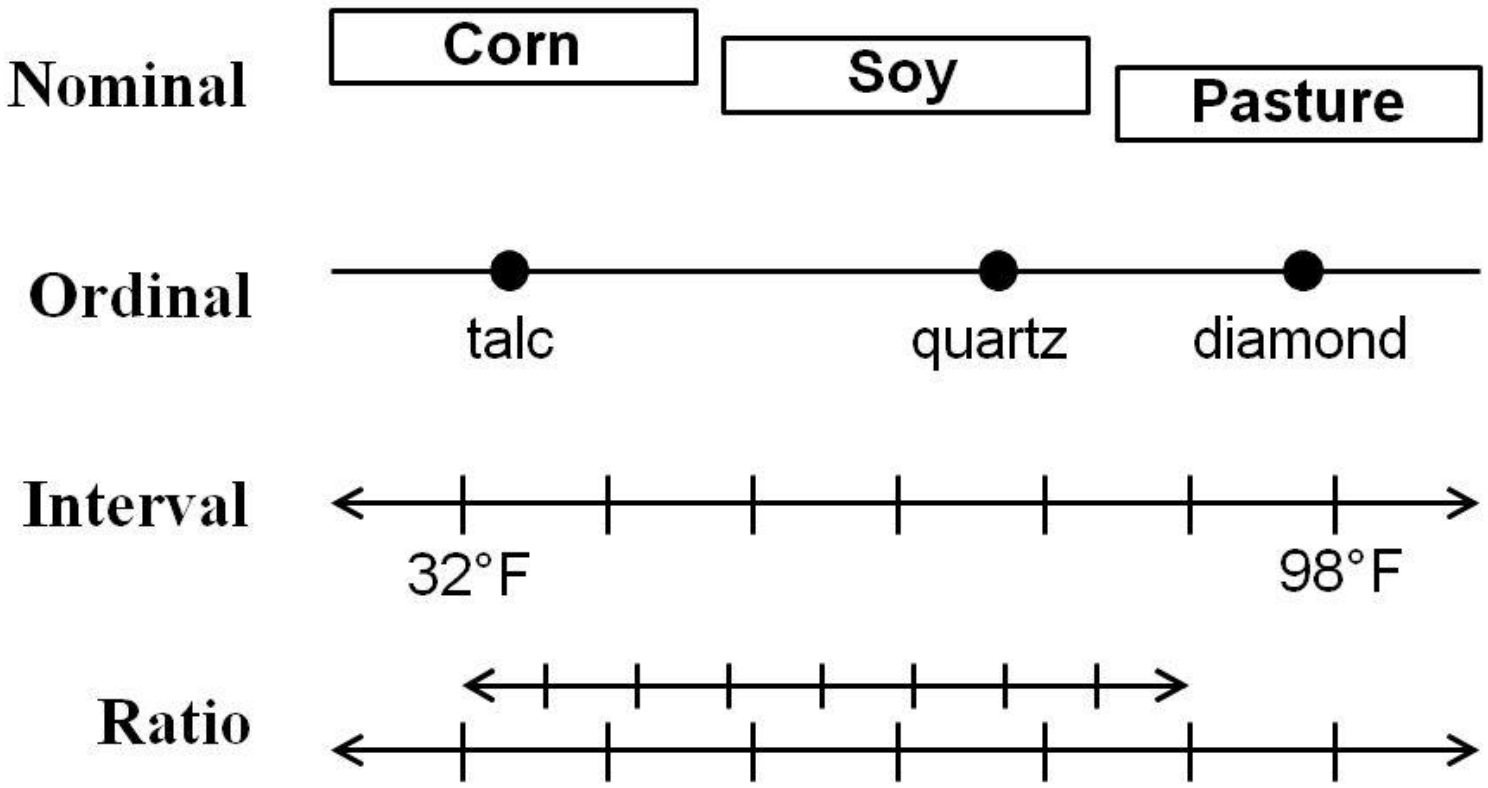
Degrees of Freedom

Bertin's **eight visual variables** (Bertin, 1983), including Cartesian coordinates





Measurement Scales (Stevens, 1946)



Measurement Scales (Stevens, 1946)

	Used for Testing…	Permitted Statistics
Nominal	*Equality?*	*Count, mode*
Ordinal	*Greater or less?*	*Median, percentiles*
Interval	*Equality of intervals?*	*Mean, std. deviation*
Ratio	*Equality of ratios?*	*Coefficient of variation*