NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

NC STATE UNIVERSITY

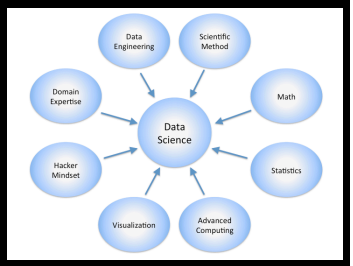
Logistics

- Waitlist cleared
- Keep an eye on Moodle
  - Lecture notes/slides
  - Additional resources
- TA Office: EBII-1229B. T(11.15-12.30pm)/Th (3.45-5.00pm)
- Final Exam
  - 12/8/15, 1:00-4:00pm.

NC STATE UNIVERSITY

Key points from 8/20

- At a high level, *data science* is a set of fundamental principles that guide the extraction of knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles. -- Foster Provost and Tom Fawcett



Source: WIKIBOOKS on Data Science

NC STATE UNIVERSITY

Today

- Data Exploration
- Different types of data
- Data dimensions
- Data reduction
- Introduction to "R"

NC STATE UNIVERSITY

Workflow

Does Streptomycin cure Tuberculosis?

- Design a sampling scheme
- Decide on relevant attributes (measurements)

- Data types, Plots
- Missing data, anomalies, …
- Filtering, transformation

- Build, Fit, and Validate the model

NC STATE UNIVERSITY

Once data is collected, …

- We start with exploring the data

Data Type Taxonomy

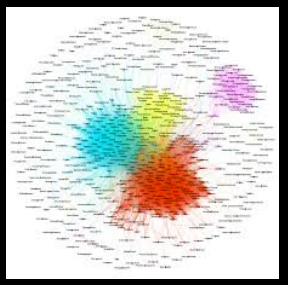
- 1D, 2D, nD
  - Sequences (1D)
  - Maps (2D)
  - Relational (nD)
- Temporal data
- Trees (hierarchical data)
- Networks (graphs)
- Others

Ref: Shneiderman, 96. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization

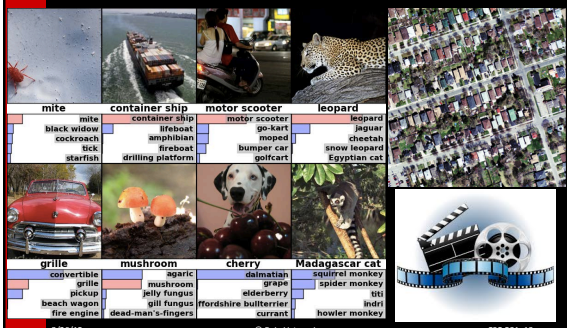
Example Datasets: Tables

Day	Date	Project	Billable Hours	Non-Billable Hours	Total
Friday	4/1/16	Project 1	0.20		$12.50
Sunday	4/3/16	Project 2	0.50		$25.00
Tuesday	4/5/16	Project 2	80.00		$4,000.00
Thursday	4/7/16	Project 1	50.00		$2,500.00

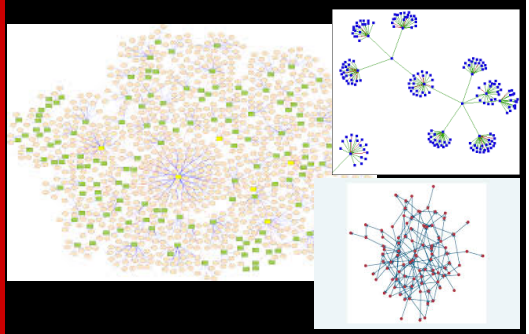
Example Datasets: Text



Example Datasets: Images



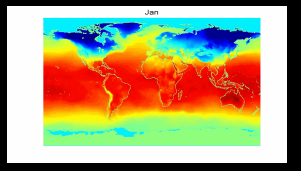
Example Datasets: Graphs



Example Datasets: Ordered Data

- SpatioTemporal
- Genomic Sequences

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```



Data Semantics Vs. Data Types

- Semantics
  - Real-world meaning
  - E.g., “Name”, “Height”, “Temperature”
- Types
  - Interpretation in terms of scales of measurements
  - E.g., quantity or category, data structures, …

Data Objects Vs. Attributes

- Datasets are made up of data objects
- Data object represents an entity
- Examples
  - Sales database: customers, stores items, …
  - Medical database: patients, treatments, ..
  - University database: students, professors, …
- Objects are also called as: samples, instances, data points, tuples
- Database Rows -> Data Objects
- Database Columns -> Attributes

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Ratio

- 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. (location of zero is arbitrary)
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts (location of zero is fixed)

Properties

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Differences are meaningful : + -
  - Ratios are meaningful * /
- Question: Which of these prosperities are applicable
  - Nominal attribute:
  - Ordinal attribute:
  - Interval attribute:
  - Ratio attribute:

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Source

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative esti-

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A CLASSIFICATION OF SCALES OF MEASUREMENT

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?
- Consider measuring the height above average
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bob?
  - Is this situation analogous to that of temperature?

Allowed Transformations

	Attribute Type	Transformation	Comments
Categorical Qualitative	Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
	Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function	An attribute encompassing the notion of good, better, best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Numeric Quantitative	Interval	*new_value = a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

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.

Asymmetric Attributes

- Only **presence** (a non-zero attribute value) is regarded as important
  - Words present in documents
  - Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."
- Symmetric binary: both outcomes are equally important
  - E.g., gender
- Asymmetric binary: outcomes are not equally important
  - E.g., medical test (positive vs. negative)

Things to remember

- The types of operations an analysis you choose should be "meaningful" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not there
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - Many times what is meaningful is measured by statistical significance
  - But in the end, what is meaningful is measured by the domain

Today

- Different types of data
- Data dimensions
- Data reduction
- Exploratory analysis
- Introduction to "R"

Important Characteristics of Data

- Dimensionality (number of attributes)
  - Curse of Dimensionality
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale

Data Quality

- Poor data quality negatively affects many data processing efforts

  "The most important point is that poor data quality is an unfolding disaster."
  - Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

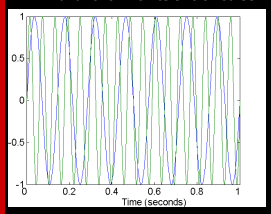
    Thomas C. Redman, DM Review, August 2004
- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality

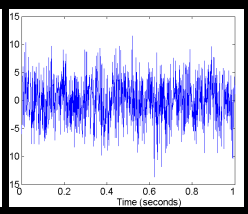
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



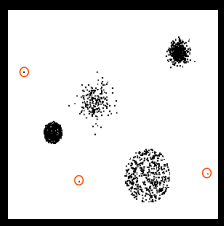
Two Sine Waves



Two Sine Waves + Noise

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection
- Causes?



Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

Duplicate Data

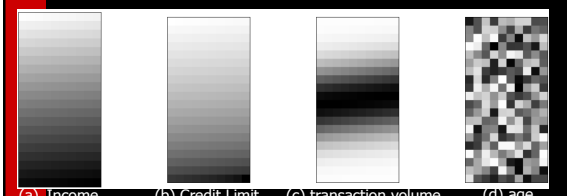
- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

Visual Exploration

- Why data visualization?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Vis. Techniques

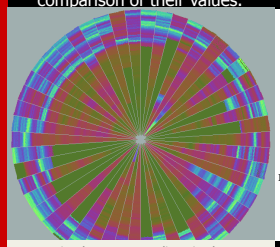
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- Pixel-oriented visualization techniques map each attribute value of the data to a single colored pixel
- The colors of the pixels reflect the corresponding values



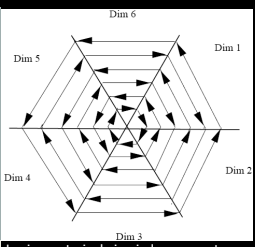
(a) Income (b) Credit Limit (c) transaction volume (d) age

The Circle Segments Technique

- The rationale of this approach is that close to the center all attributes are close to each other enhancing the visual comparison of their values.



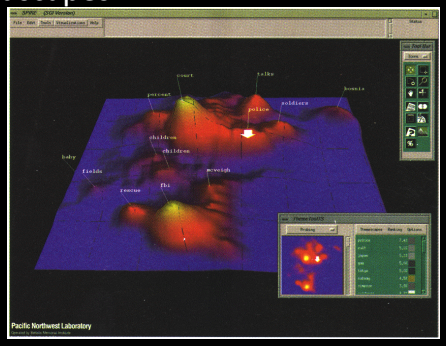
Representing about 265,000 50-dimensional Data Items with the 'Circle Segments' Technique



Laying out pixels in circle segment

Ref: Ankerst, Vis. Data Mining

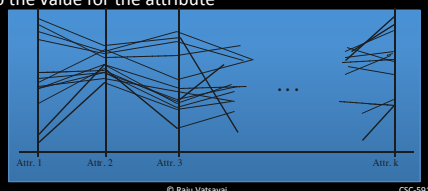
Landscapes



NC STATE UNIVERSITY

Parallel Coordinates

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



NC STATE UNIVERSITY

Icon-Based Vis Techniques

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)

NC STATE UNIVERSITY

Acknowledgements

- Vipin Kumar (Minnesota)
- Jiawei Han (UIUC)
- Hanspeter Pfister (Harvard)