CSE 590
Data Science Fundamentals

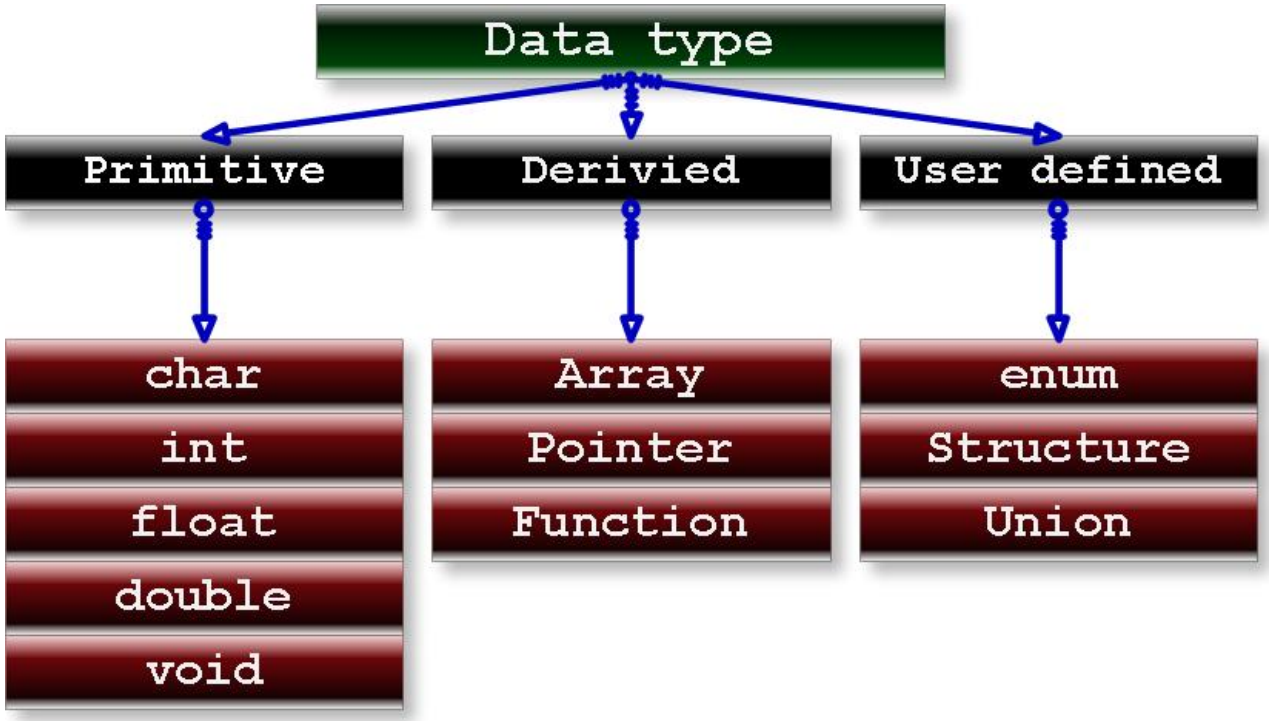
Data Types

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Data Science components and tasks	
3	Data types	Project #1 out
4	Introduction to R, statistics foundations	
5	Introduction to D3, visual analytics	
6	Data preparation and reduction	
7	Data preparation and reduction	Project #1 due
8	Similarity and distances	Project #2 out
9	Similarity and distances	
10	Cluster analysis	
11	Cluster analysis	
12	Pattern miming	Project #2 due
13	Pattern mining	
14	Outlier analysis	
15	Outlier analysis	Final Project proposal due
16	Classifiers	
17	Midterm	
18	Classifiers	
19	Optimization and model fitting	
20	Optimization and model fitting	
21	Causal modeling	
22	Streaming data	Final Project preliminary report due
23	Text data	
24	Time series data	
25	Graph data	
26	Scalability and data engineering	
27	Data journalism	
	Final project presentation	Final Project slides and final report due

Data Types Every CS Person Knows



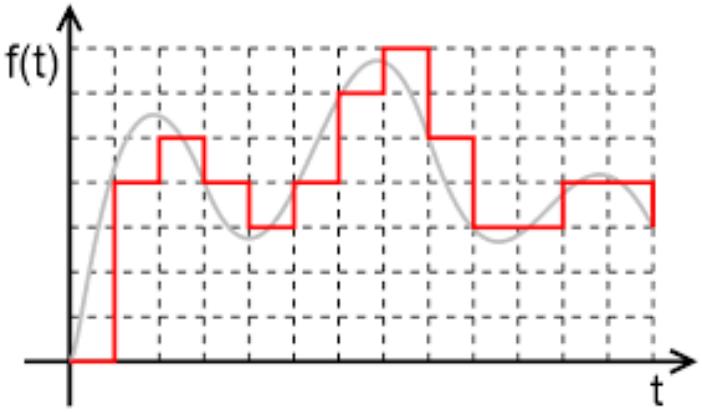
Data Types in data Science

Source data type
Numeric
Categorical
Text
Time series
Time series
Discrete sequence
Spatial
Graphs
Any type

Variables in Statistics

Numeric variables

- measure a **quantity** as a number
- like: ‘how many' or 'how much'
- can be continuous (grey curve)
- or discrete (red steps)



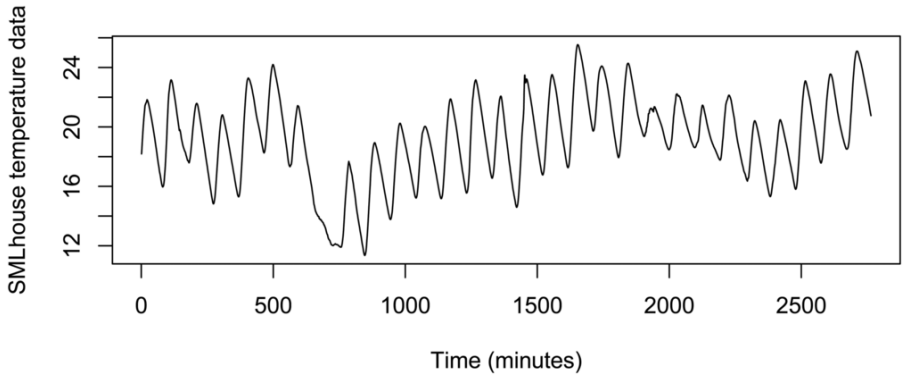
Categorical variables

- describe a **quality** or characteristic
- like: 'what type' or 'which category’
- can be ordinal = ordered, ranked (distances need not be equal)
  - clothing size, academic grades, levels of agreement
- or nominal = not organized into a logical sequence
  - gender, business type, eye color, brand

Sensor data

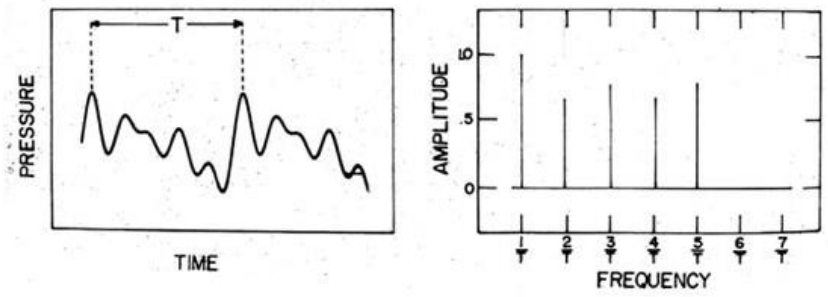
Characteristics

- often large scale
- time series

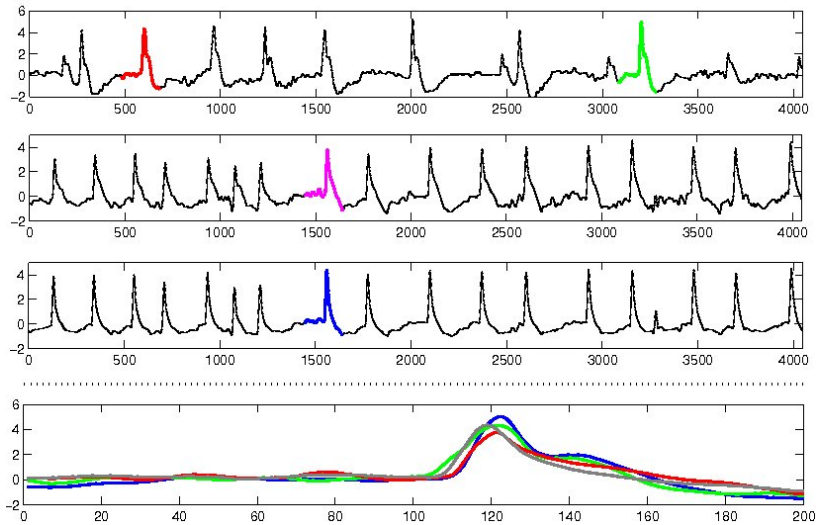


Feature Analysis

- Fourier transform (FT, FFT)
- Wavelet transform (WT, FWT)
- Motif discovery



Fourier transform



Motif discovery

Image data

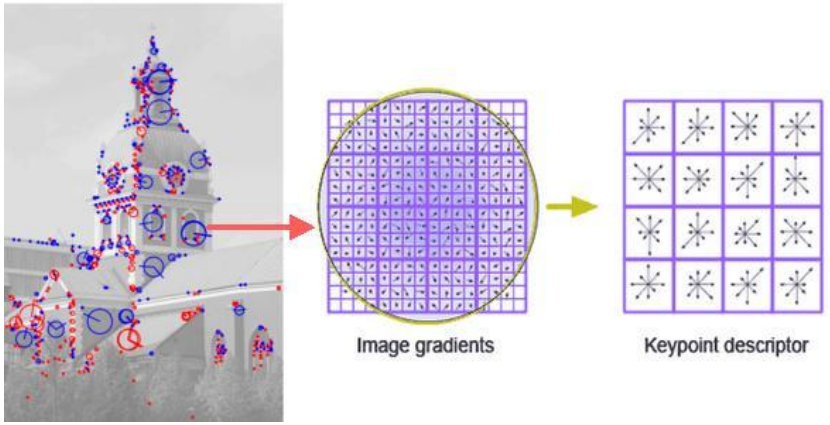
Characteristics

- array of pixels

Feature Analysis

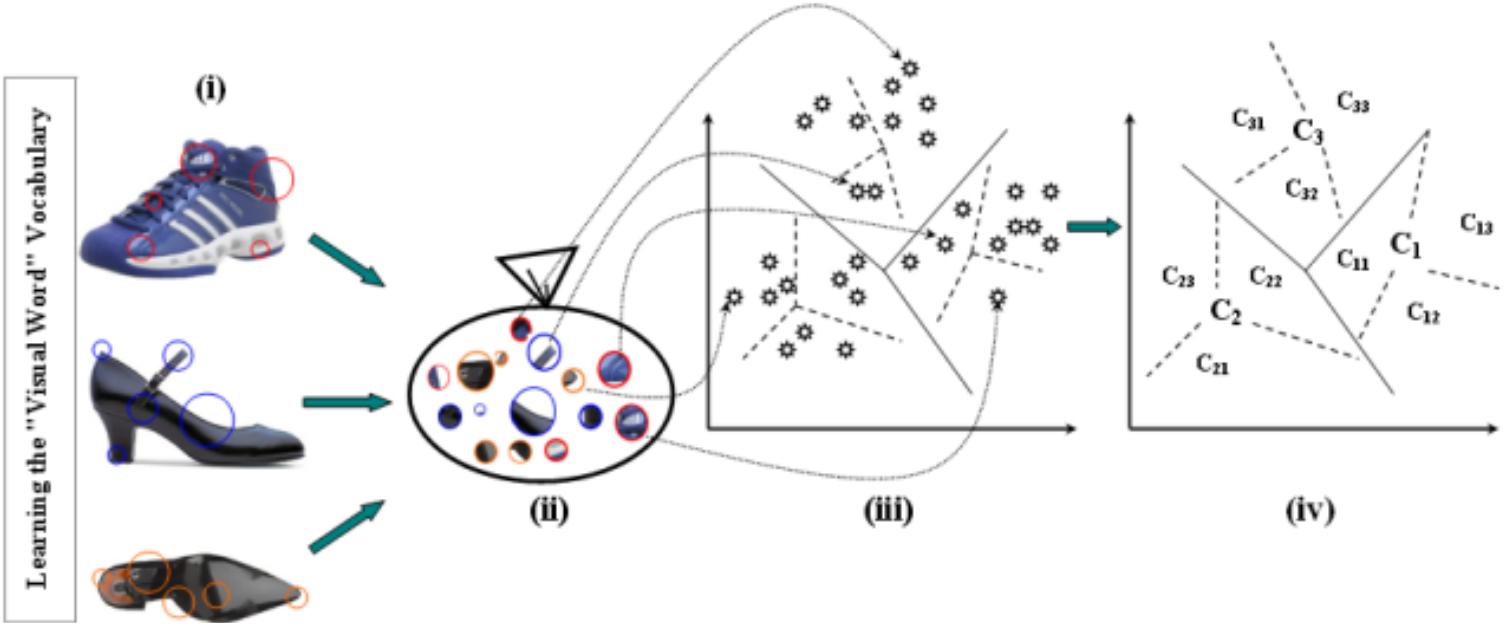
- histograms
  - values
  - gradients
- FFT, FWT
- Scale Invariant Feature Transform (SIFT)
- Bag of Features (BoF)
- visual words

histograms



SIFT

Bag of Features (BoF)



Bag of Features (BoF)

1. Obtain the set of bags of features
   (i) Select a large set of images
   (ii) Extract the SIFT feature points of all the images in the set and obtain the SIFT descriptor for each feature point extracted from each image
   (iii) Cluster the set of feature descriptors for the amount of bags we defined and train the bags with clustered feature descriptors
   (iv) Obtain the visual vocabulary

2. Obtain the BoF descriptor for a given image/video frame
   (v) Extract SIFT feature points of the given image
   (vi) Obtain SIFT descriptor for each feature point
   (vii) Match the feature descriptors with the vocabulary we created in the first step
   (viii) Build the histogram

More information

Video data

Characteristics

- essentially a time series of images

Feature Analysis

- many of the above techniques apply albeit extension is non-trivial

Other data

Weblogs

- typically represented as text strings in a pre-specified format
- this makes it easy to convert them into multidimensional representation of categorical and numeric attributes

Network traffic

- characteristics of the network packets are used to analyze intrusions or other interesting activity
- a variety of features may be extracted from these packets
  - the number of bytes transferred
  - the network protocol used
  - IP ports used

Text Data

Characteristics

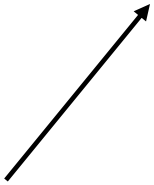
- often raw and unstructured

Feature analysis

- first step is to remove stop words and stem the data
- perform **named-entity recognition** to gain atomic elements
  - identify names, locations, actions, numeric quantities, relations
  - understand the structure of the sentence and complex events
- example:
  - Jim bought 300 shares of Acme Corp. in 2006.
  - $[\text{Jim}]_{\text{Person}}$ bought $[\text{300 shares}]_{\text{Quantity}}$ of $[\text{Acme Corp.}]_{\text{Organiz.}}$ in $[\text{2006}]_{\text{Time}}$
- distinguish between
  - application of grammar rules (old style, need experienced linguists)
  - statistical models (Google etc., need big data to build)

Data Types in data Science

Source data type	Destination data type	Methods
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent semantic analysis (*LSA*)
Time series	Discrete sequence	*SAX*
Time series	Numeric multidimensional	*DWT*, *DFT*
Discrete sequence	Numeric multidimensional	*DWT*, *DFT*
Spatial	Numeric multidimensional	2-d *DWT*
Graphs	Numeric multidimensional	*MDS*, spectral
Any type	Graphs	Similarity graph (Restricted applicability)



which you may need to convert into

using these methods

Numeric to Categorical Data: Discretization

Goal

- divide the ranges of the numeric attribute into φ ranges
- examples:
  - age: [0, 10], [11, 20], [21, 30], … → 0, 1, 2, 3, …
  - salary: [40,000, 80,000], …. [1,040,000, 1,080,000] → 1, …, 26,…
- what is lost here?
  - distribution information within group
  - data might be distributed unevenly (less for age, more for salary)
- what can we do?
  - store a value that characterizes the distribution
  - use log scaling (when the attribute shows an exponential distribution)
  - use equi-depth range – store the same number of records per bin
  - any disadvantages?

Categorical to Numeric Data: Binarization

If a categorical attribute has φ different values, then φ different binary attributes are created

This can have disadvantages

- ordinal data are not spaced by actual distance, whatever the metric
- nominal data have no order and spacing at all
- there are techniques based on correlation (stay tuned)

Text to Numeric Data

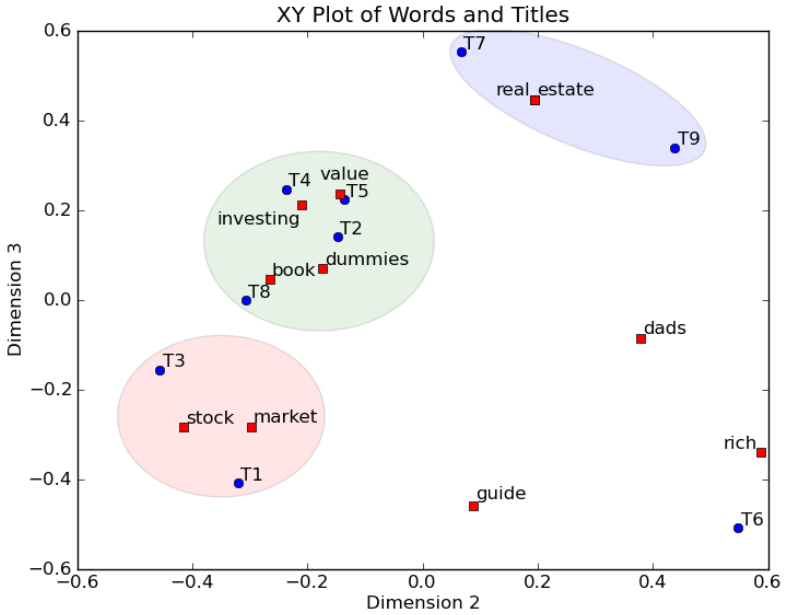
Use Latent Semantic Analysis (LSA)

- also known as Latent Semantic Indexing (LSI)
- turns text into a high-dimensional vector which can be compared

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

Count Matrix

LSA

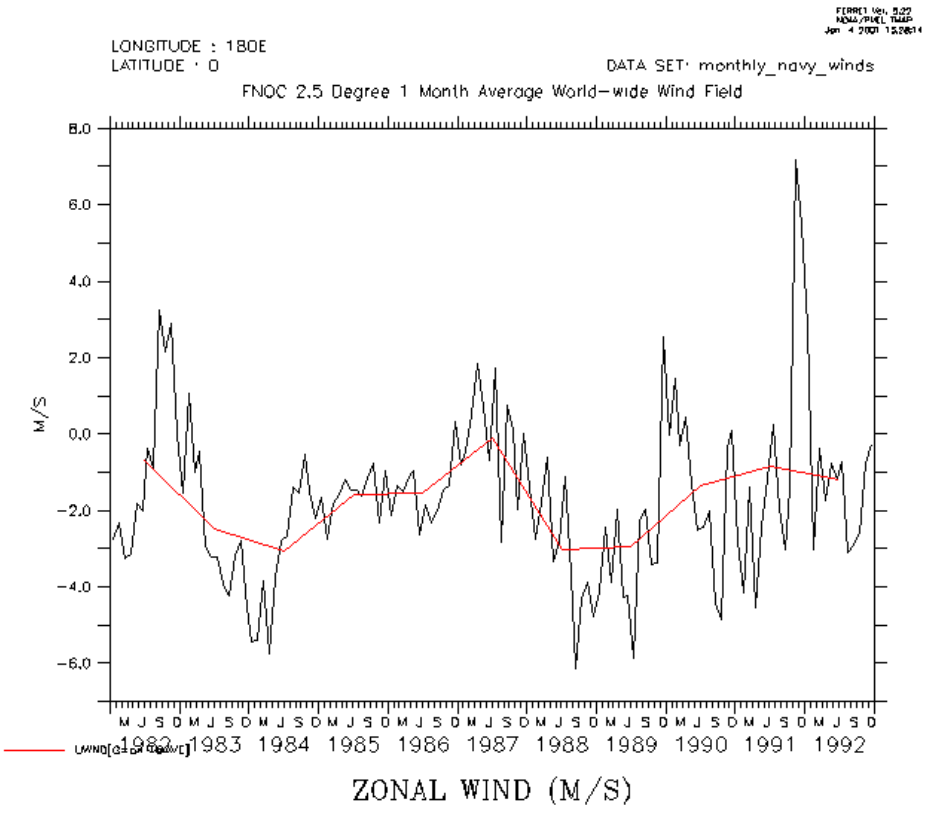


Word/document cluster

Sequence Data

Time series to discrete sequence data

- window based averaging
  - eliminates noise in the data
- value-based discretization
  - collapse series into a smaller number of intervals
  - usually done after averaging



Time series to numeric data

- FFT, DFT
- also works for discrete series and spatial data

Project #1

Dataset of NYC bike sharing service

- ~1M records each for July 2013, 2014, 2015
- will link from the course website (under *Labs*)

Attributes:

- Trip Duration
- Start Time & Date, Stop Time & Date
- Start/End Station Name
- Station ID
- Station Latitude/Longitude
- Bike ID
- User Type (24-hour pass. 7-day pass, Annual)
- User Gender, Year of Birth

Project #1 (cont'd)

Tasks:

- make interesting observations about the data
- keep in mind the nature of the data
- what might be important to the city, provider, users, investors?

Strategies

- use R (next lecture) to analyze the data
- R has visualization tools to present the findings
- can also use D3 (lecture after next)

Deliverables

- comprehensive report with findings (numbers, plots, text narration)
- submit to instructor by email by 9/22