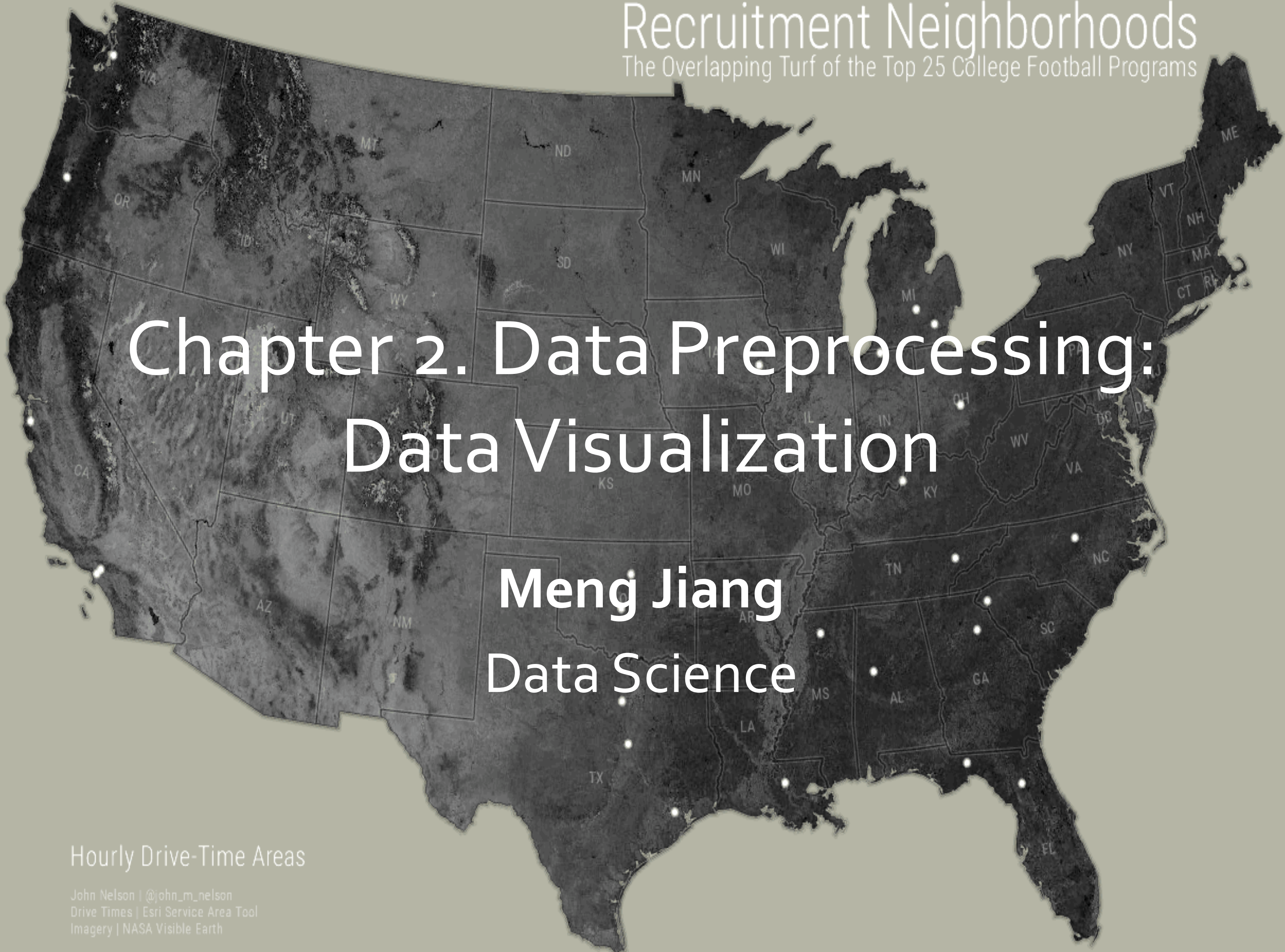
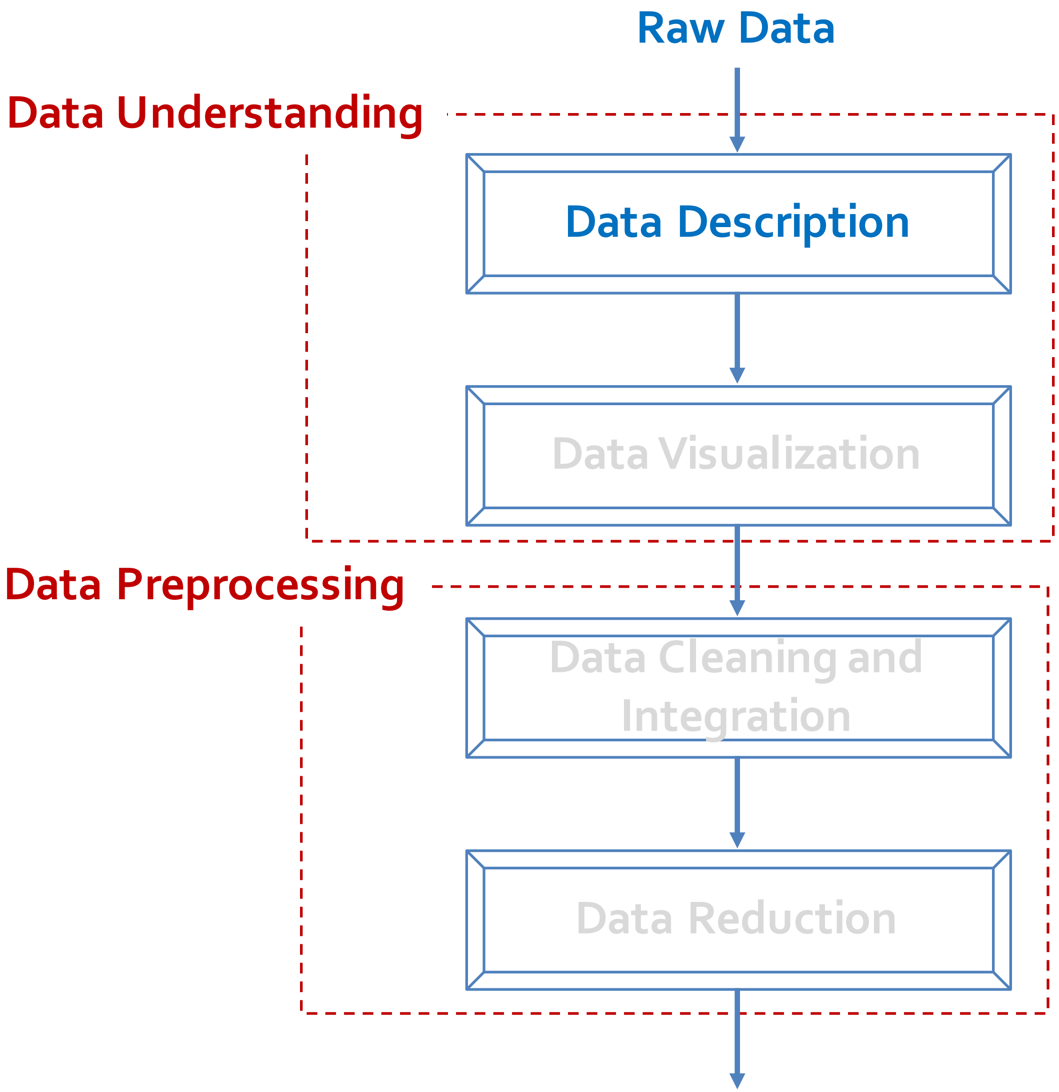


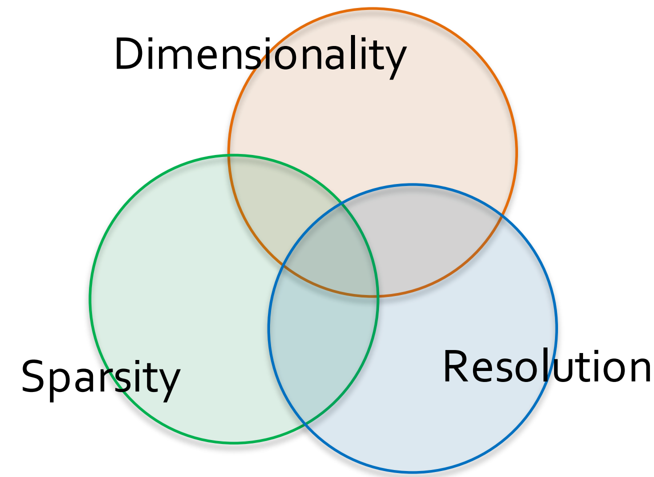
Chapter 2. Data Preprocessing: Data Visualization

Meng Jiang

Data Science



Describing data:

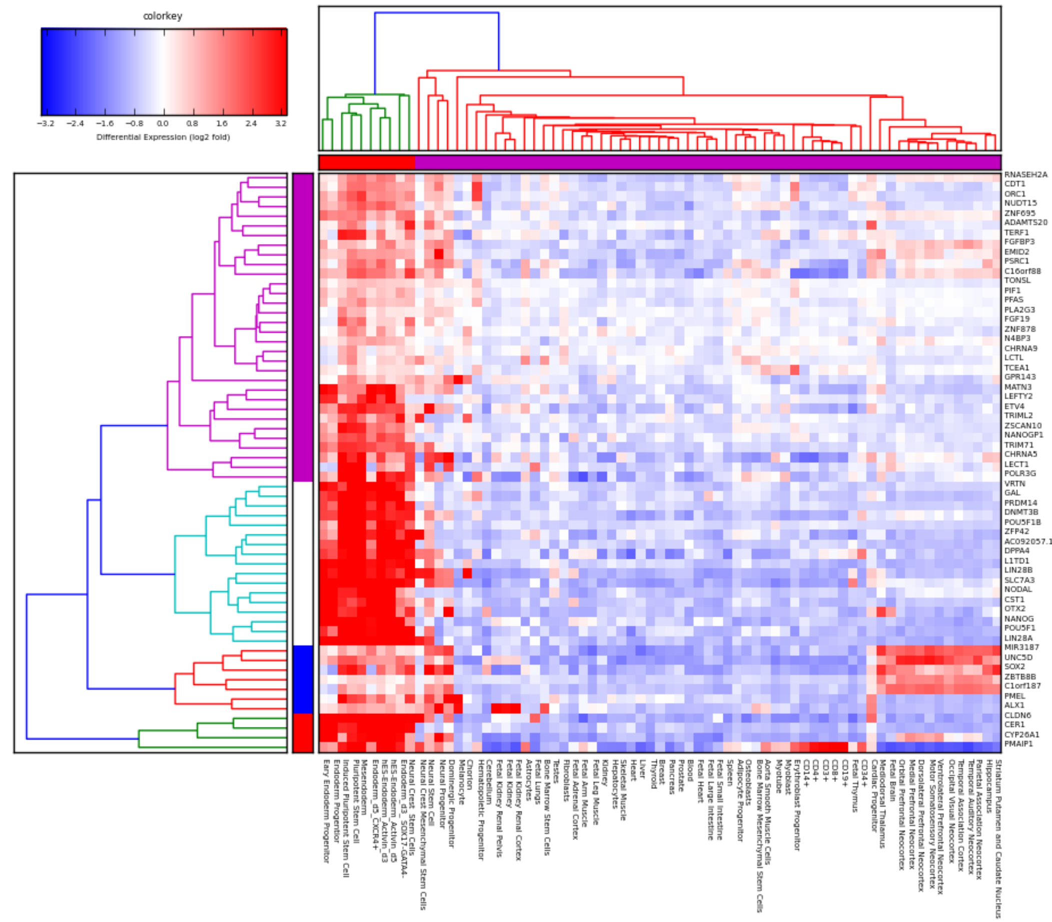


Describing Data

- Dimensionality
  - How many features are there in the data?
- Sparsity
  - Does the data contain many empty values?
- Resolution
  - Is the data granular or coarse?

Dimensionality

- The number of features that the entities or objects in the dataset possesses.
- Datasets with few dimensions tend to be qualitatively different than those with many dimensions.



A Quick Aside

Are more dimensions (i.e., features) always helpful?

Input Features: X
(attributes)

Class: Y
(label)

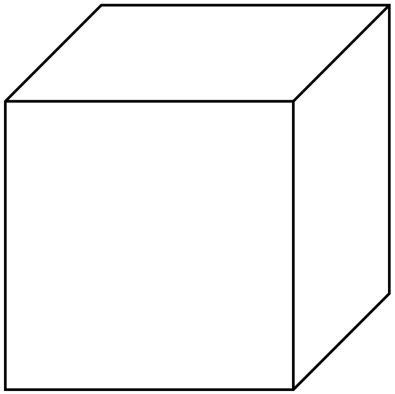
Data Objects
(instances)

	x_1	x_2	…	x_m
1				
2				
…				
n				

y

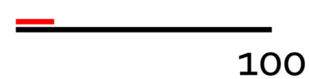
Curse of Dimensionality

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



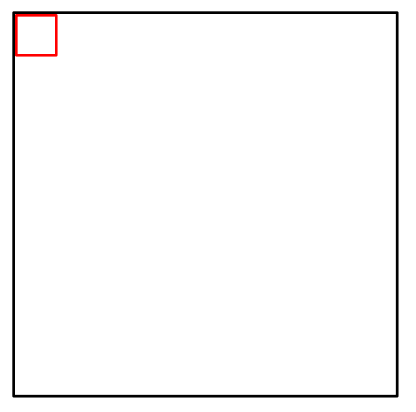
Curse of Dimensionality

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



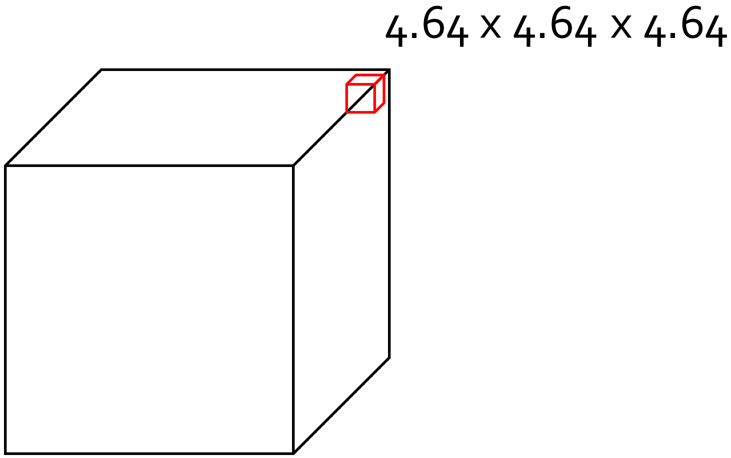
Curse of Dimensionality

- In 2 dimensions, we must go a distance of $(0.01)^{1/2} = 0.1$ on average to reach our nearest neighbor.



Curse of Dimensionality

- In 3 dimensions, we must go a distance of $(0.01)^{1/3} \approx 0.215$ on average to reach our nearest neighbor.



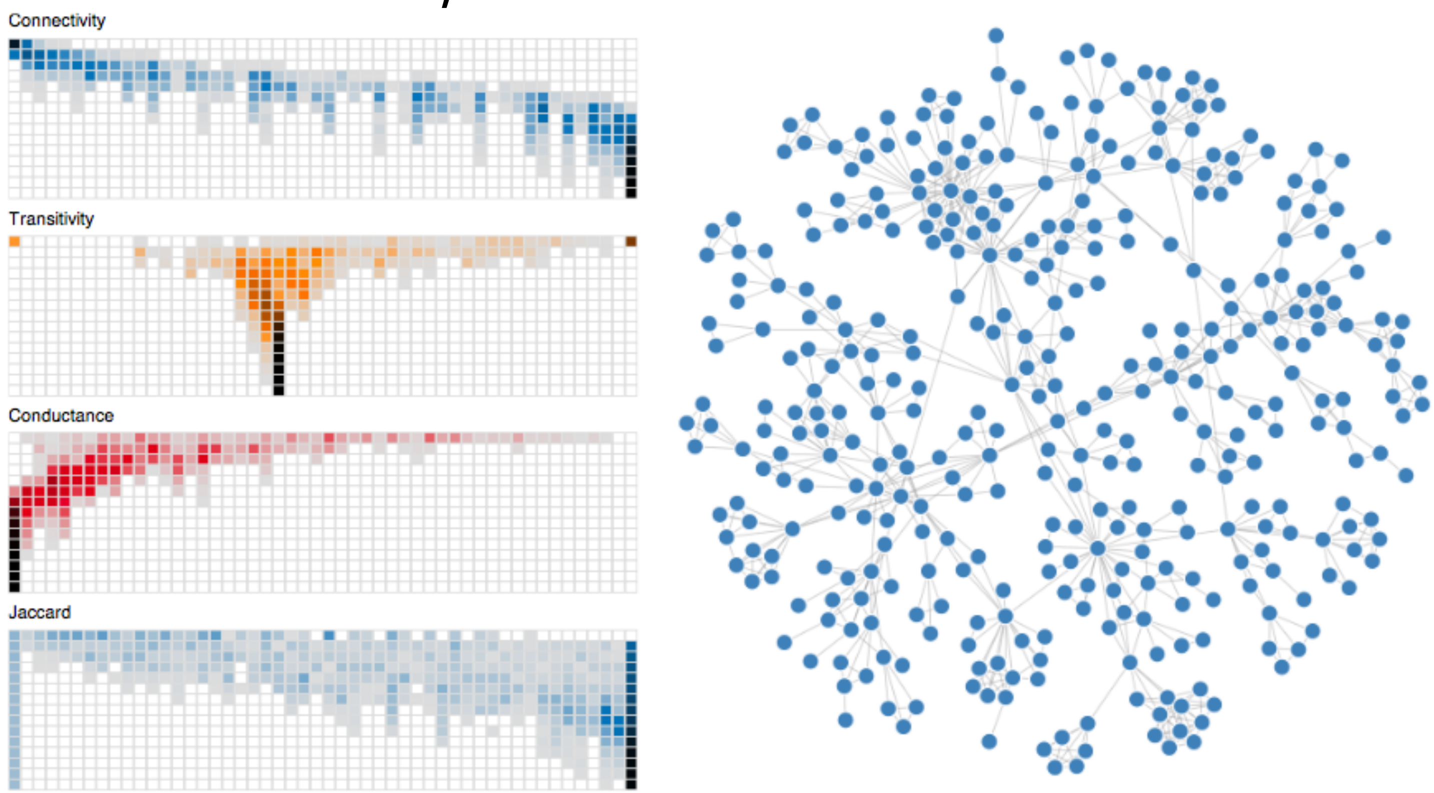
Curse of Dimensionality

- In *d* dimensions, we must go a distance of $(0.01)^{1/d}$ on average to reach our nearest neighbor.
- As *d* increases, this distance approaches 1 (the entire length of the hypercube)!

Dimension Reduction

Data Sparsity

- When the distance between the data becomes large, we call the data **sparse**.
- For some datasets, most features have values of zero.



Data Sparsity

- Can be a problem for many methods
  - Can create statistical bias due to small samples
  - Can reduce the meaningfulness of distance calculations
- Can also be an advantage
  - Requires less storage

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	▸ Carnegie Mellon University ◕	18.5	150
4	2	▸ Massachusetts Institute of Technology ◕	12.2	82
5	3	▸ Stanford University ◕	10.9	54
6	3	▸ University of California - Berkeley ◕	10.9	81
7	5	▸ Univ. of Illinois at Urbana-Champaign ◕	9.9	84
8	6	▸ Cornell University ◕	8.7	68
9	7	▸ University of Michigan ◕	8.6	63
10	8	▸ University of Washington ◕	8.3	56
11	9	▸ University of California - San Diego ◕	6.9	54
12	10	▸ Georgia Instit		
13	11	▸ University of		
14	12	▸ Columbia Uni		
15	13	▸ University of		
16	14	▸ University of		
17	15	▸ Princeton Uni		
18	16	▸ University of Texas at Austin ◕	5.2	42
19	16	▸ University of Maryland - College Park ◕	5.2	44
20	18	▸ University of California - Los Angeles ◕	5	37
21	19	▸ Northeastern University ◕	4.8	54
22	19	▸ Purdue University ◕	4.8	51
23	21	▸ University of Massachusetts Amherst ◕	4.7	50
24	22	▸ New York University ◕	4.5	47
25	23	▸ Harvard University ◕	4.2	29
26	23	▸ University of California - Irvine ◕	4.2	54
27	25	▸ Rutgers University ◕	3.9	43
28	26	▸ University of California - Santa Barbara ◕	3.5	25
29	27	▸ University of Utah ◕	3.4	39
30	27	▸ Pennsylvania State University ◕	3.4	31
31	29	▸ Stony Brook University ◕	3.3	41
32	30	▸ University of California - Davis ◕	3.2	29

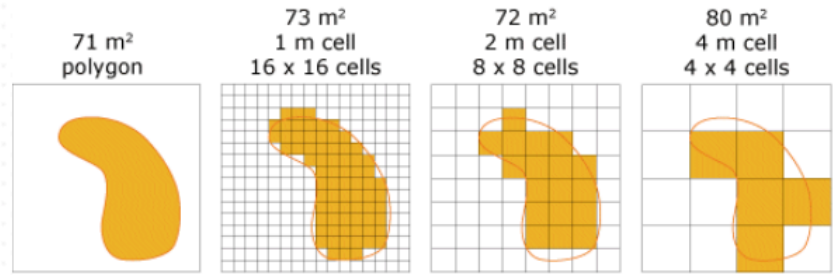
126 institutions

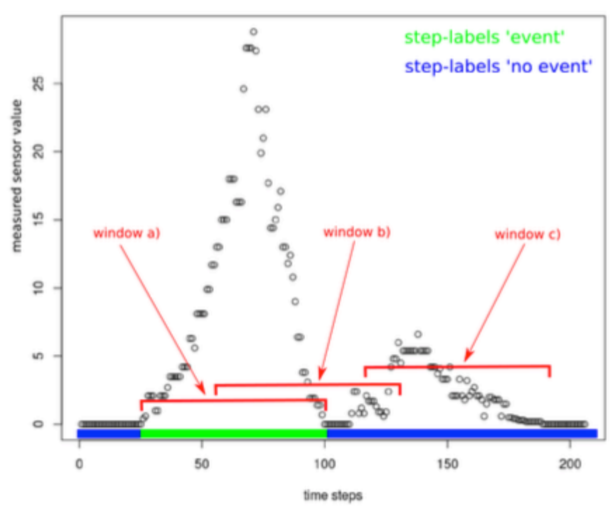
▸ Cornell University ◕	8.7	68
Cornell University - Ithaca, NY		

Data Cleaning and Integration

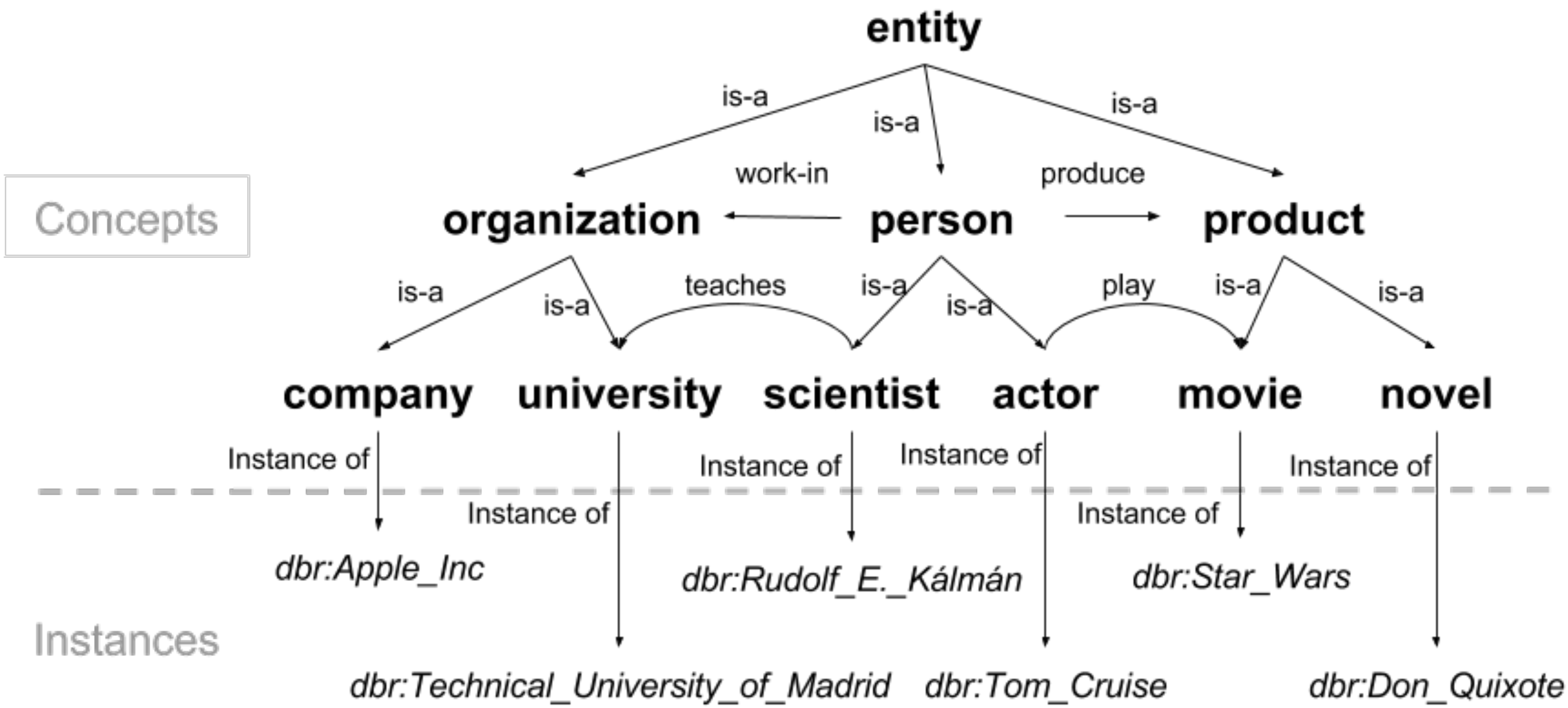
Data Resolution

- 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.

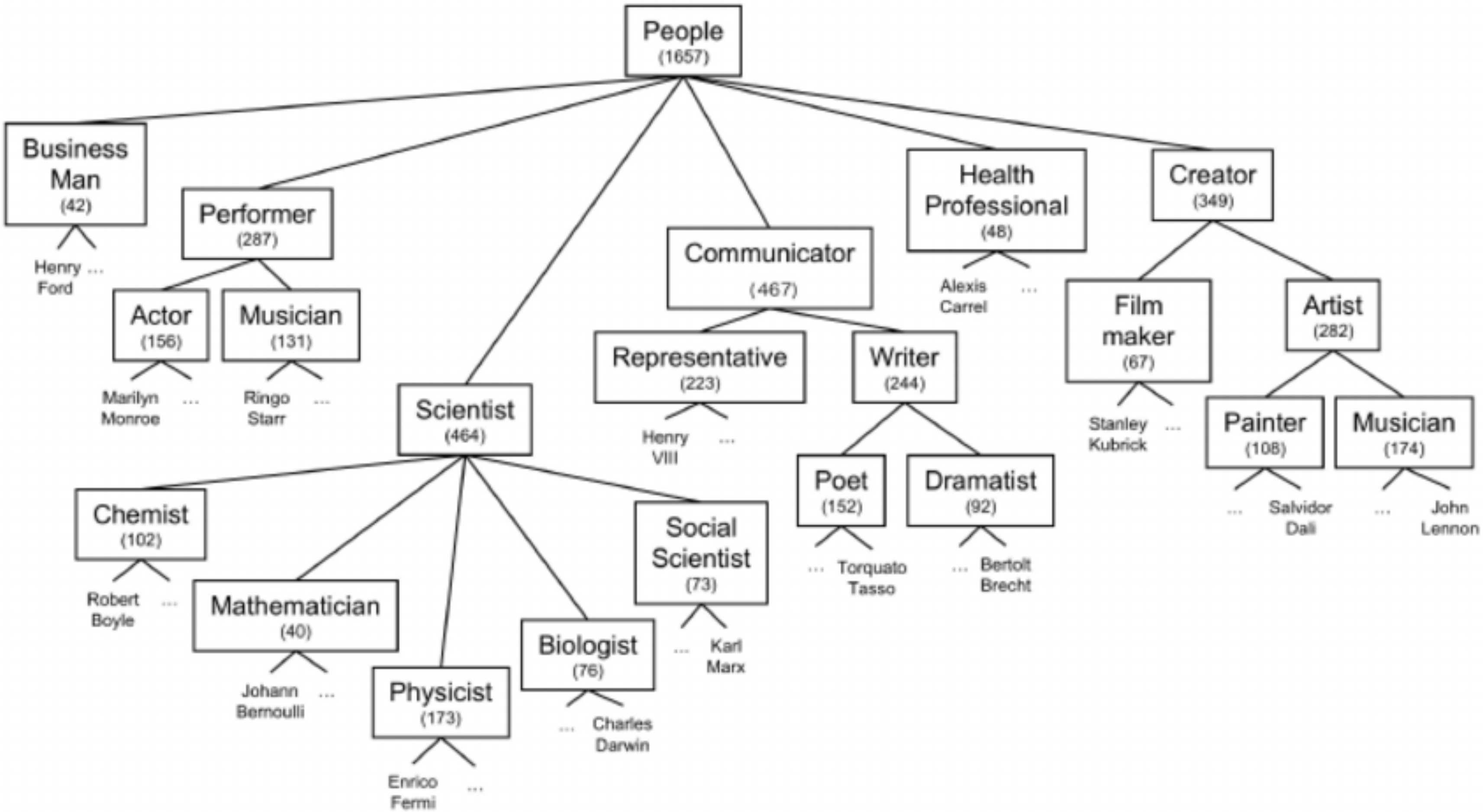




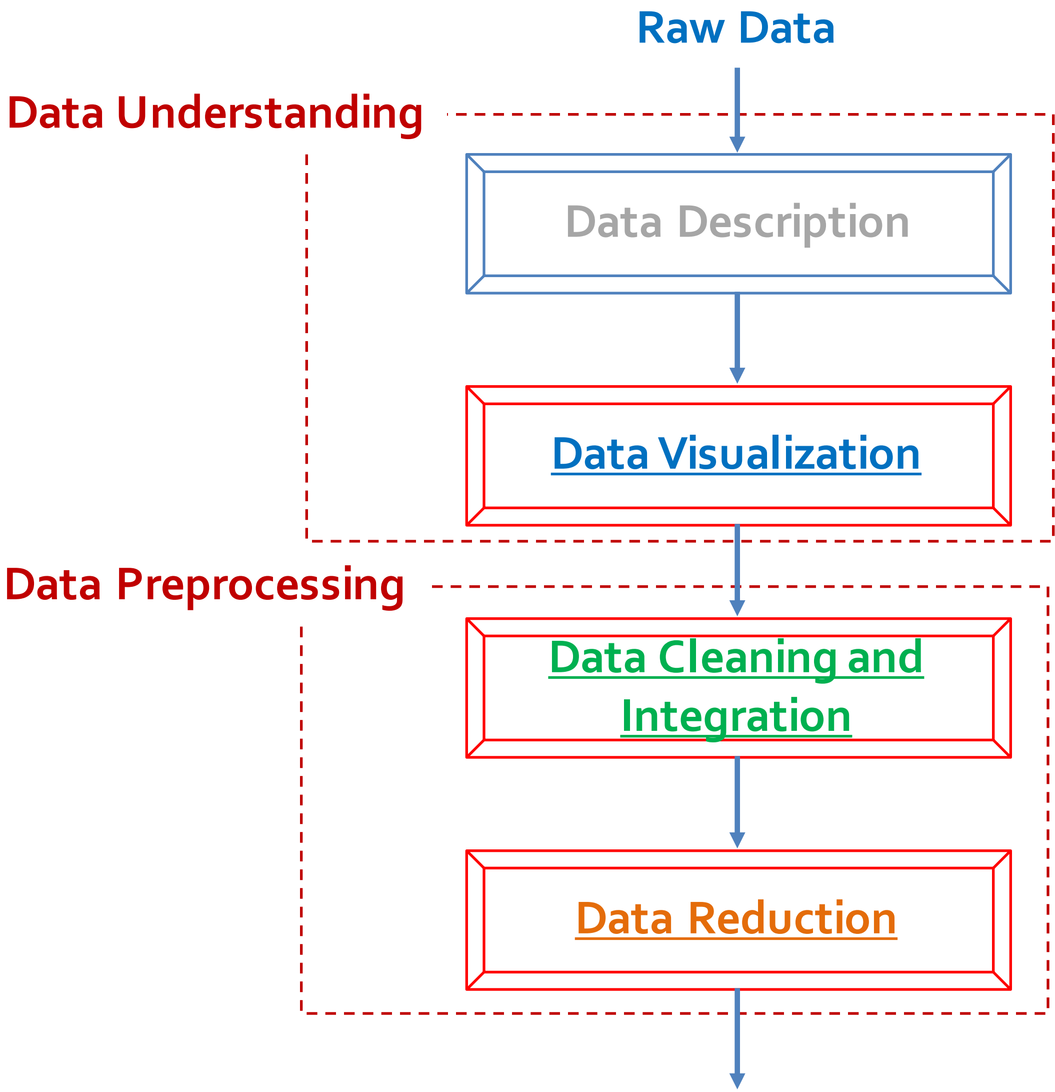
Data Resolution



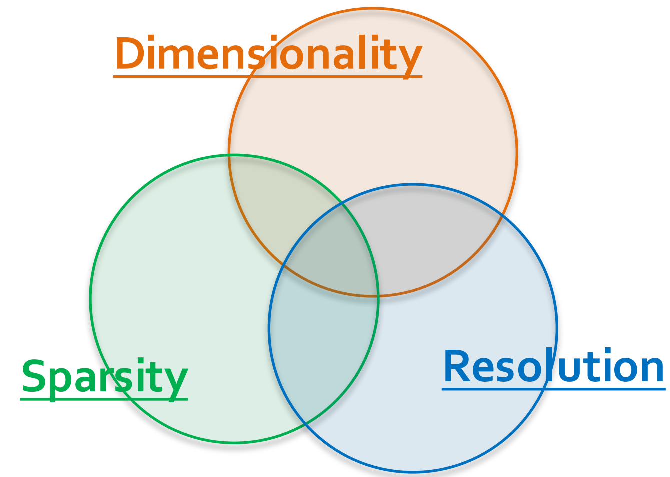
Data Resolution



[Data Visualization]()



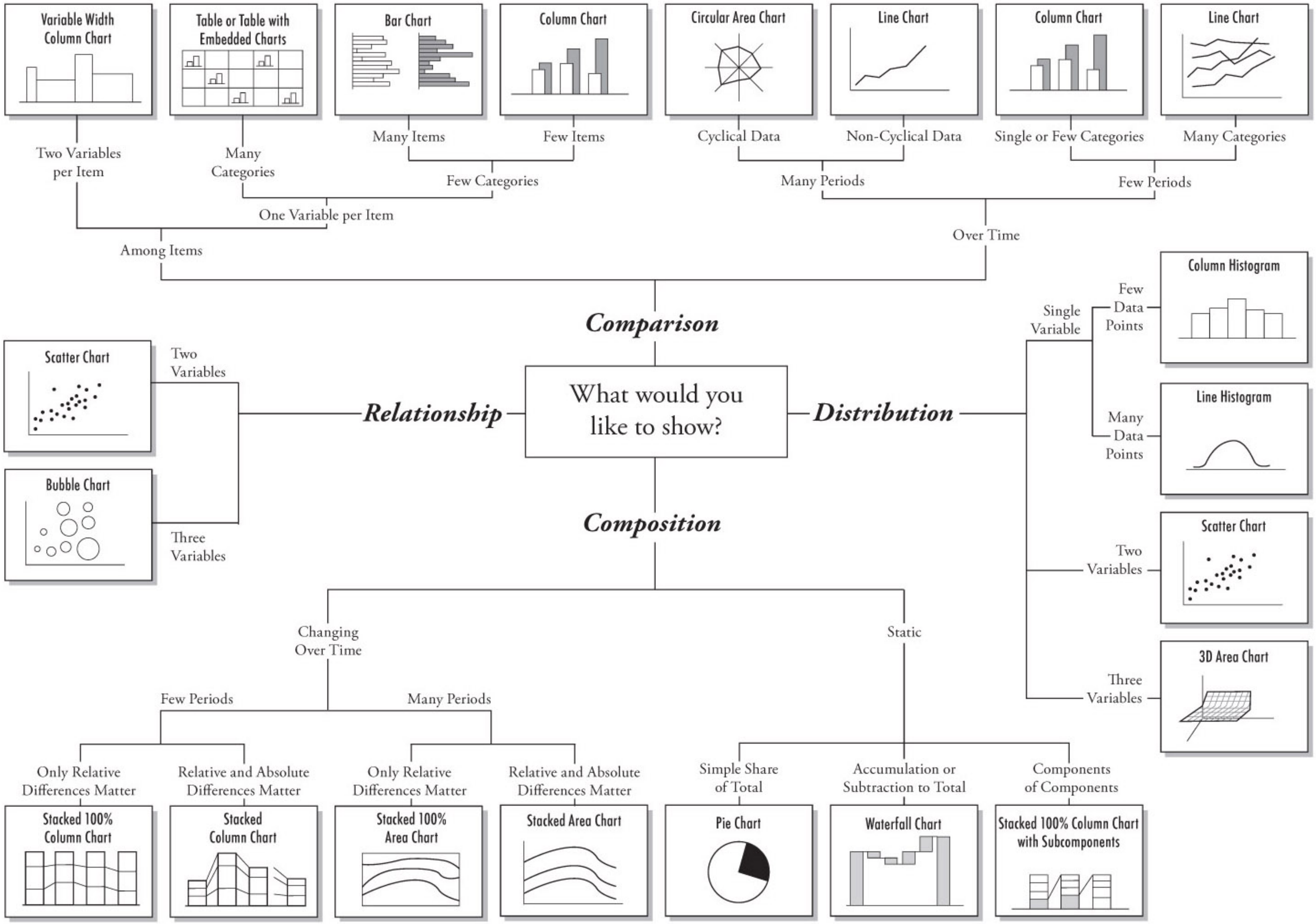
Describing data:



Today: Data Visualization

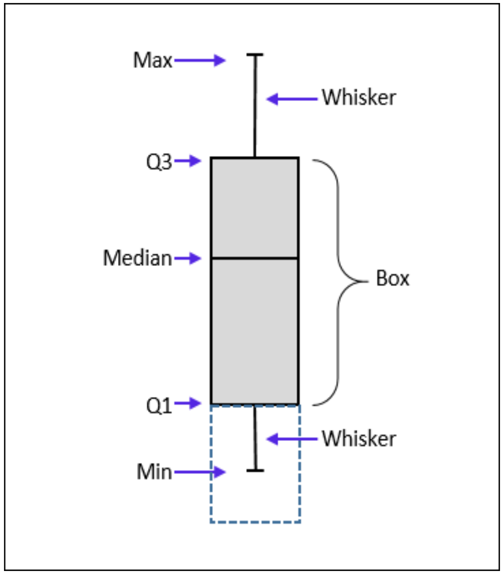
- Understand dimensionality, sparsity, and resolution
- Describe and generate plots
  - **Boxplot, histogram, bar chart**
  - **Quantile plot, Q-Q plot**
  - **Scatter plot**
- Understand graphical integrity
- Describe data proximity
  - Describe **data dissimilarity**: Calculate three Minkowski distance measures
  - Describe **data similarity**: Calculate Jaccard similarity, cosine similarity, and KL divergence

Chart Suggestions—A Thought-Starter



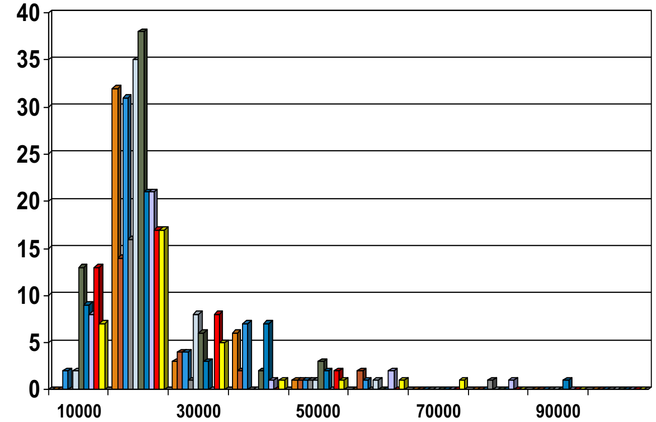
Quartiles & Boxplots

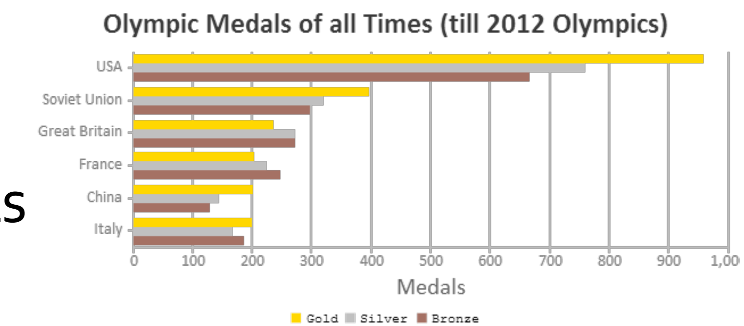
- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: $IQR = Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Boxplot**: Data is represented with a box
  - Q_1, Q_3, IQR: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - Median (Q_2) is marked by a line within the box
  - Whiskers: Two lines outside the box extended to Minimum and Maximum



Histograms

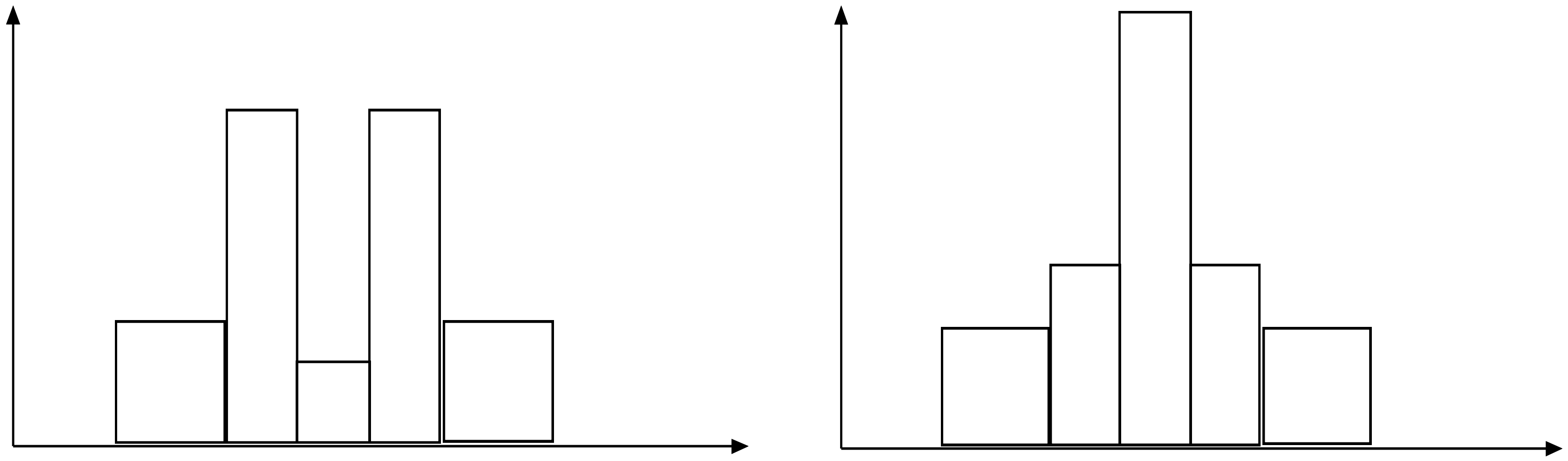
- **Histogram:** Graph display of tabulated frequencies, shown as bars
- Between histograms and **bar charts**
  - Histograms are used to show distributions of variables while bar charts are used to compare variables
  - Histograms plot binned quantitative data while bar charts plot categorical data
  - Bars can be reordered in bar charts but not in histograms





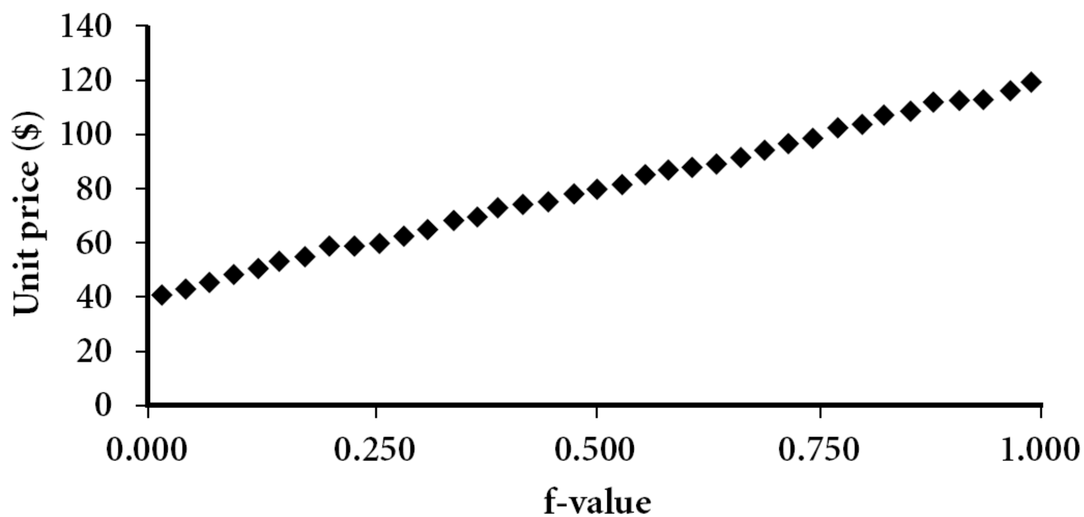
Histograms Often Tell More than Boxplots

- The two histograms may have the same boxplot
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



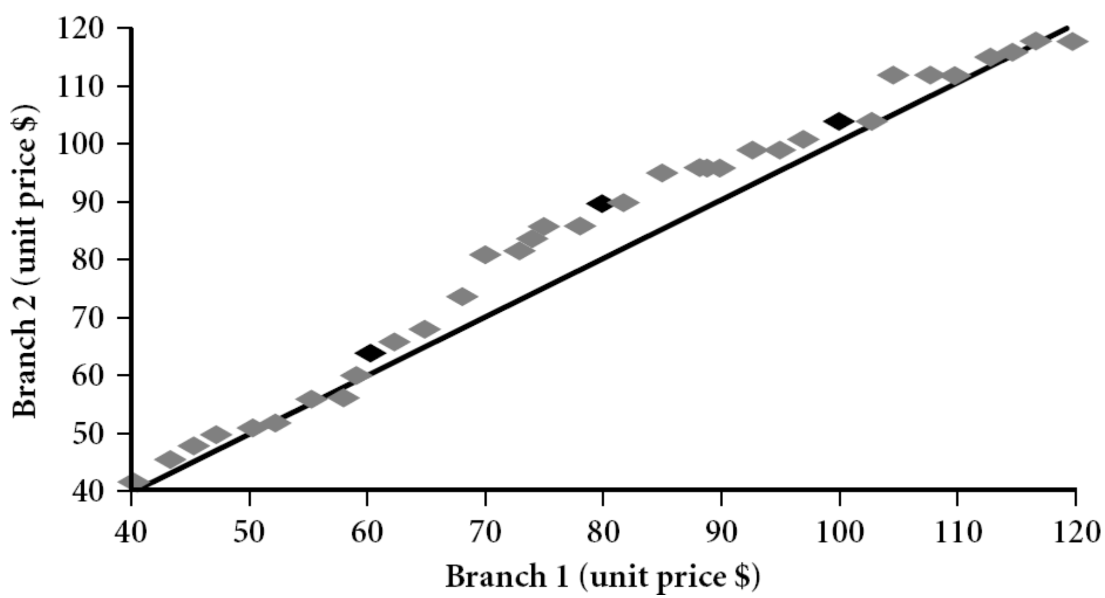
Quantile Plot

- Plots **quantile** information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i



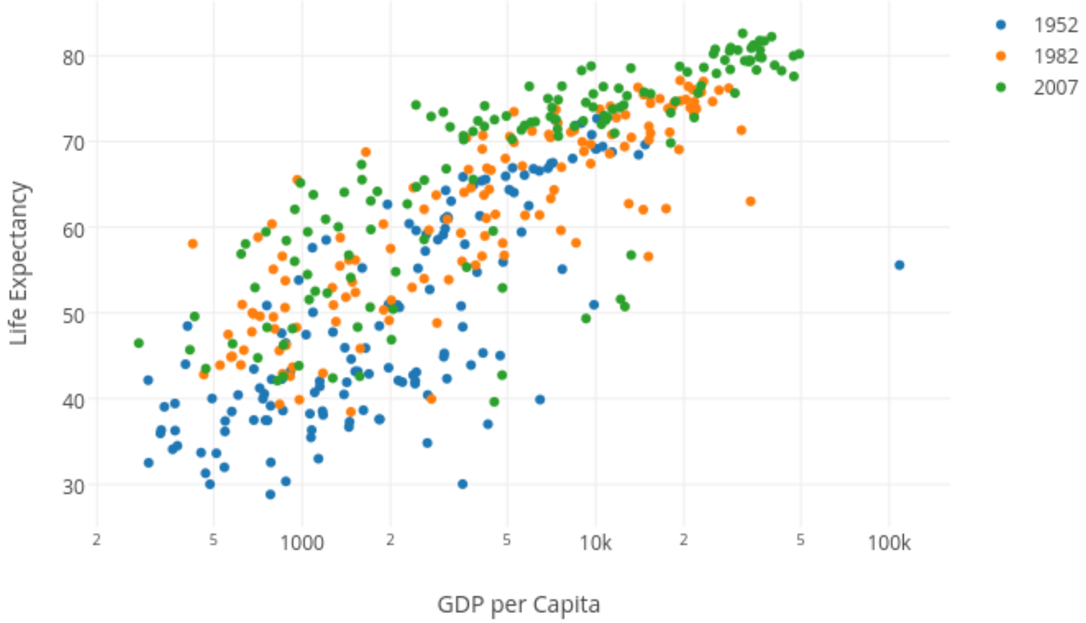
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2



Scatter plot

- Provides a first look at data to see clusters of points, outliers
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

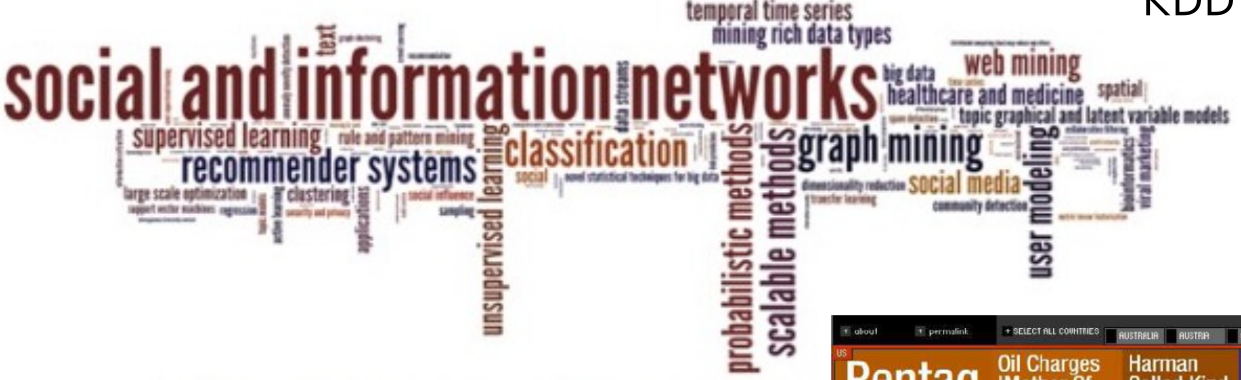


Summary: Graphic Displays of Basic Statistical Descriptions

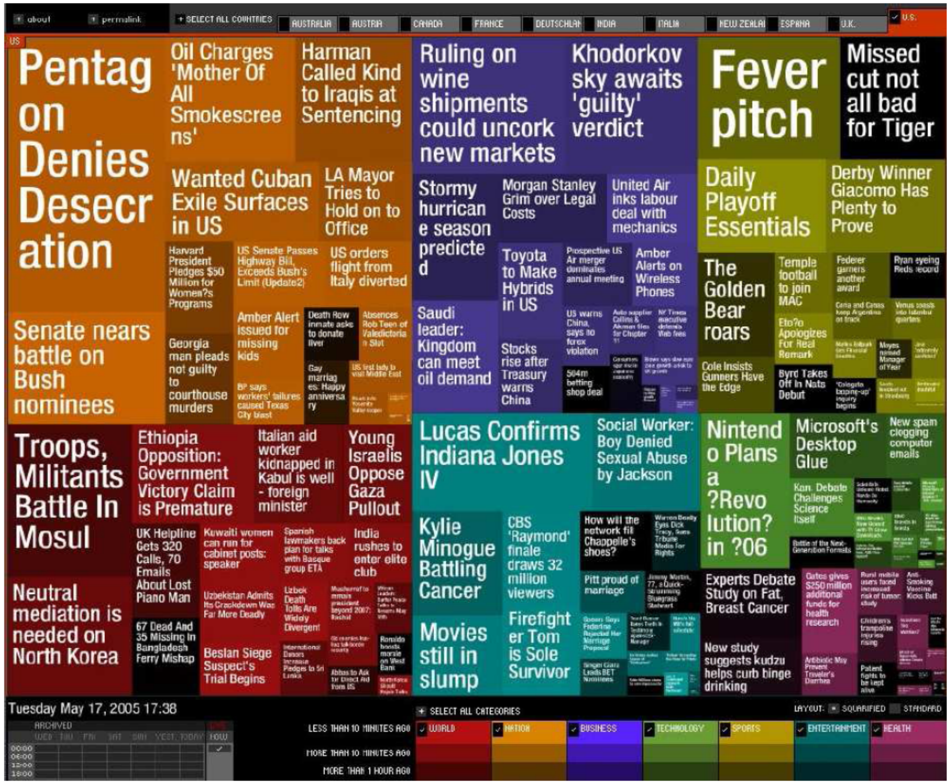
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis representative frequencies
- **Quantile plot**: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-Quantile (Q-Q) plot**: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

Other Visualization: Tag Cloud

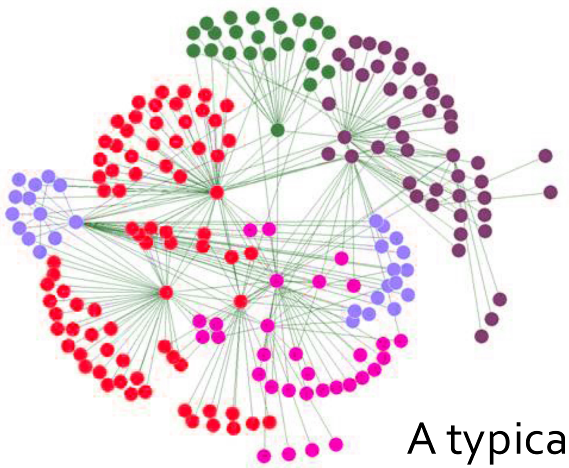
KDD 2013 Research Paper Title



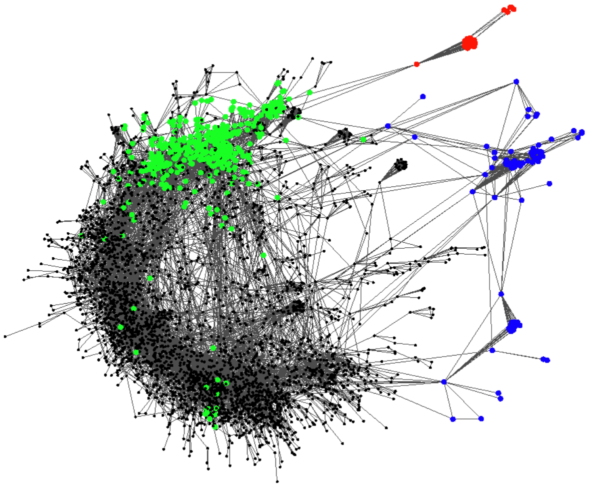
Newsmap: Google News Stories in 2005



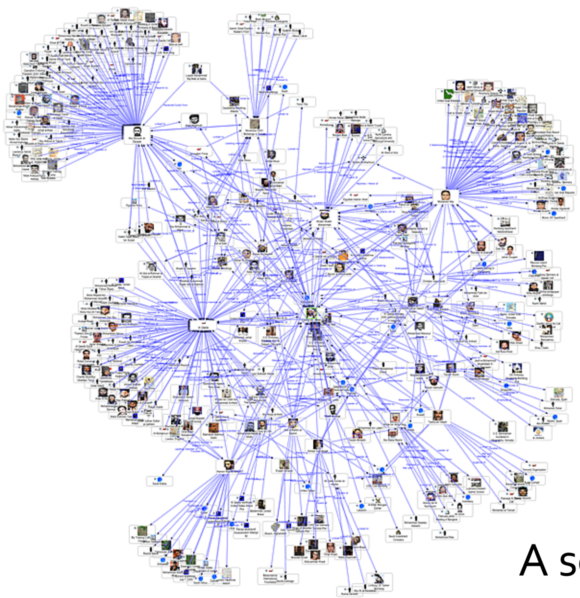
Other Visualization: Networks



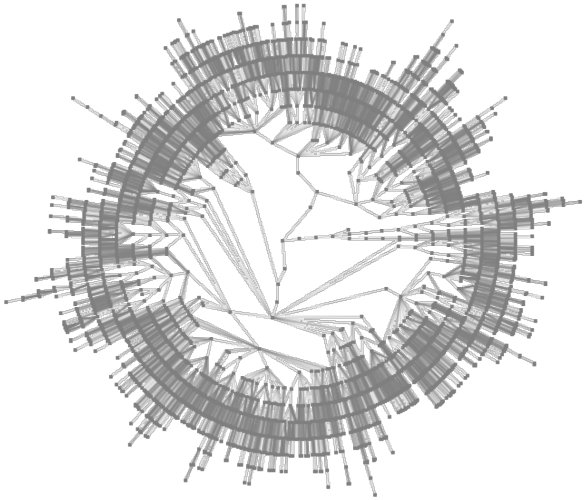
A typical network structure



organizing information networks



A social network

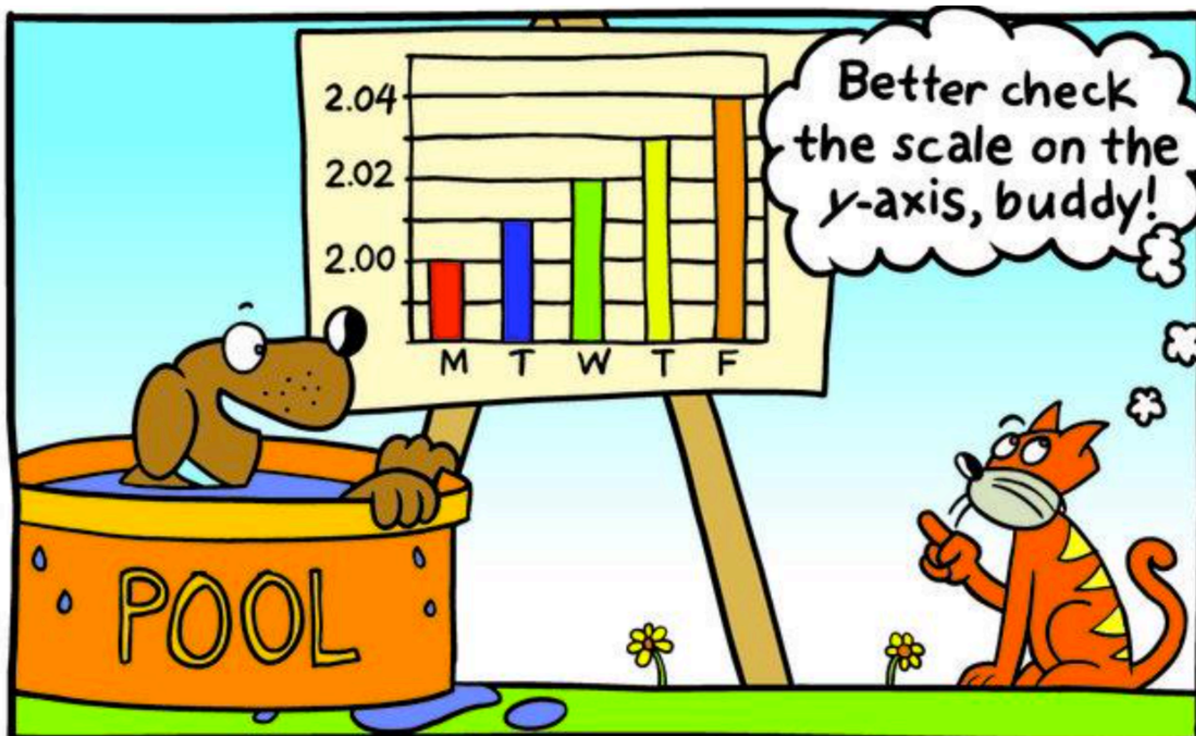


iPython (I)

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	▸ Carnegie Mellon University	18.5	150
4	2	▸ Massachusetts Institute of Technology	12.2	82
5	3	▸ Stanford University	10.9	54
6	3	▸ University of California - Berkeley	10.9	81
7	5	▸ Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	▸ Cornell University	8.7	68
9	7	▸ University of Michigan	8.6	63
10	8	▸ University of Washington	8.3	56
11	9	▸ University of California - San Diego	6.9	54
12	10	▸ Georgia Institute of Technology	6.8	75
13	11	▸ University of Wisconsin - Madison	5.9	47
14	12	▸ Columbia University	5.8	47
15	13	▸ University of Pennsylvania	5.6	46
16	14	▸ University of Southern California	5.5	49
17	15	▸ Princeton University	5.3	51
18	16	▸ University of Texas at Austin	5.2	42
19	16	▸ University of Maryland - College Park	5.2	44
20	18	▸ University of California - Los Angeles	5	37
21	19	▸ Northeastern University	4.8	54
22	19	▸ Purdue University	4.8	51
23	21	▸ University of Massachusetts Amherst	4.7	50
24	22	▸ New York University	4.5	47
25	23	▸ Harvard University	4.2	29
26	23	▸ University of California - Irvine	4.2	54
27	25	▸ Rutgers University	3.9	43
28	26	▸ University of California - Santa Barbara	3.5	25
29	27	▸ University of Utah	3.4	39
30	27	▸ Pennsylvania State University	3.4	31
31	29	▸ Stony Brook University	3.3	41
32	30	▸ University of California - Davis	3.2	29

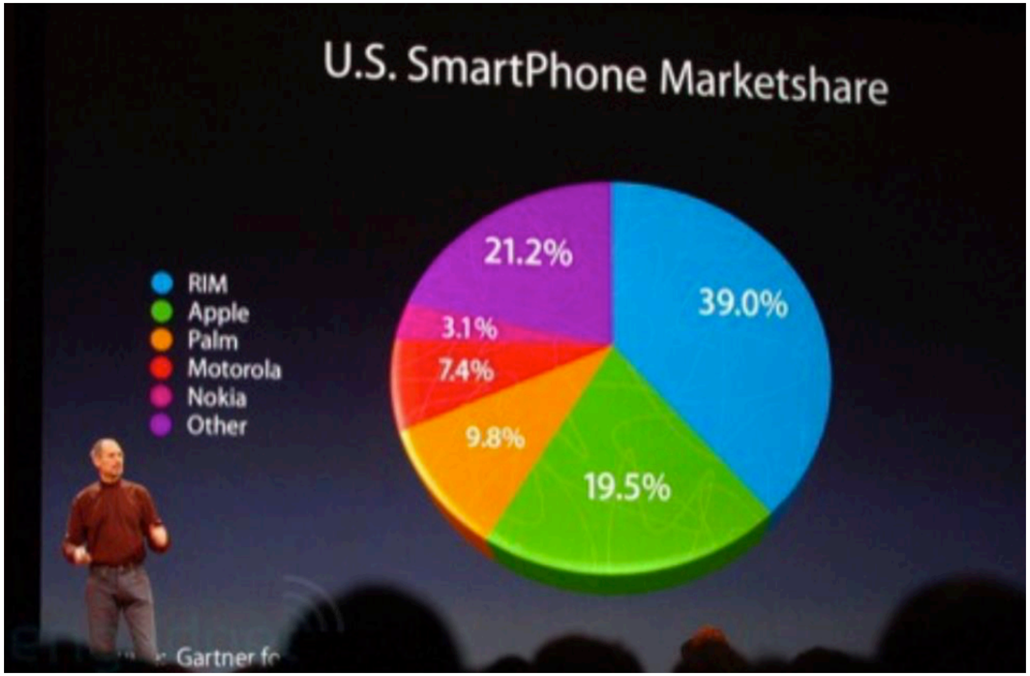
126 institutions

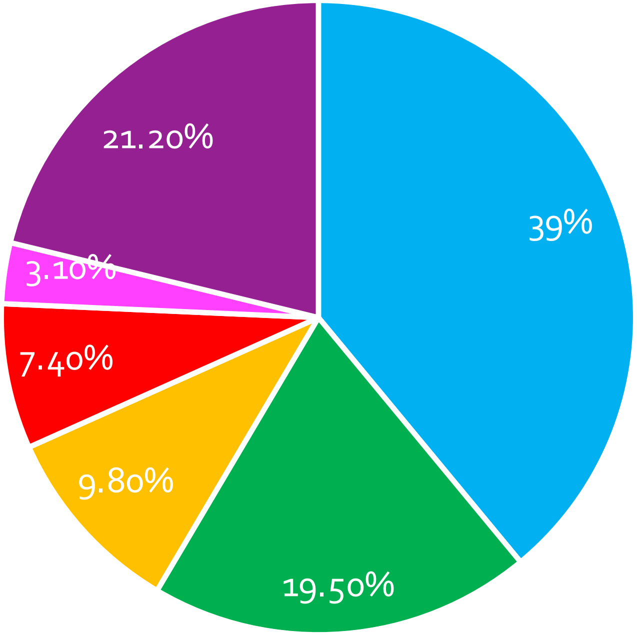
Graphical Integrity



"Wow. The number of minutes I can dog paddle is growing like crazy!"

Graphical Integrity





The Lie Factor

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

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

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

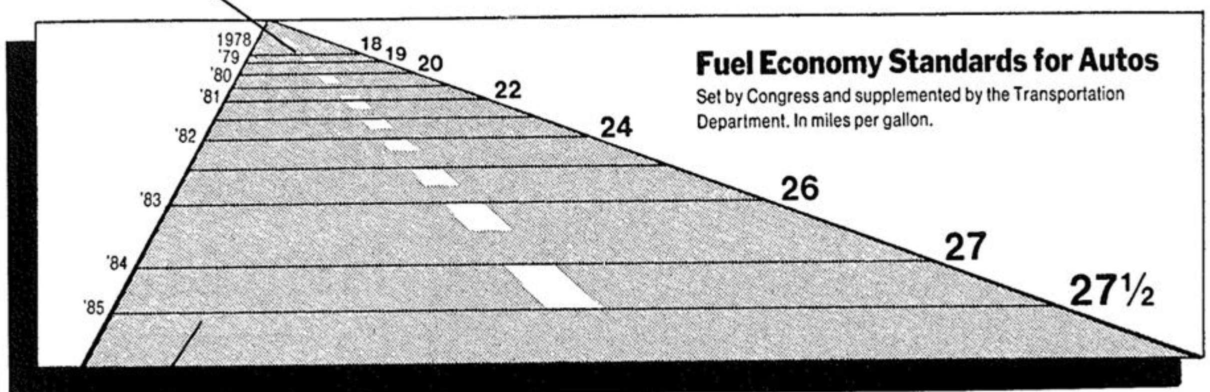
- If the lie factor is greater than 1, the graph is exaggerating the size of the effect.

The Lie Factor

Lie factor: 14.8

(5.3-0.6)/0.6 / (27.5-18)/18 = 14.8

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Line increase: 783%
Actual increase: 53%

New York Times, **9th August 1978, p D-2**

Data Proximity

- Understand dimensionality, sparsity, and resolution.
- Describe and generate plots
  - Boxplot, histogram, bar chart
  - Quantile plot, Q-Q plot
  - Scatter plot
- Understand graphical integrity
- Describe data proximity
  - Describe **data dissimilarity**: Calculate three Minkowski distance measures
  - Describe **data similarity**: Calculate Jaccard similarity, cosine similarity, and KL divergence

Similarity and Dissimilarity

- Similarity measure or similarity function
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- Dissimilarity (or distance) measure
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)
  - Range [0, 1] or [0, ∞) , depending on the definition

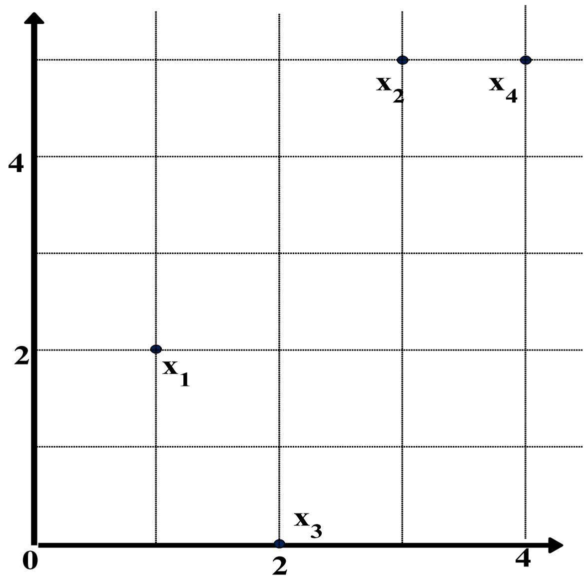
Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with *l* dimensions
- Dissimilarity (distance) matrix
  - n data points, but registers only the distance d(i, j)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal variables
  - Weights can be associated with different variables based on applications and data semantics

$$D = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1l} \\ x_{21} & x_{22} & \ldots & x_{2l} \\ & \ldots & & \\ x_{n1} & x_{n2} & \ldots & x_{nl} \end{pmatrix}$$

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ & \ldots & & \\ d(n,1) & d(n,2) & \ldots & 0 \end{pmatrix}$$

Example: Euclidean Distance



Data Matrix

point	attribute	attribute
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

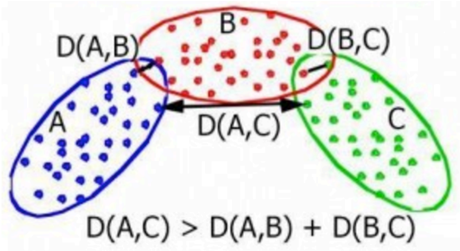
Minkowski Distance

- Minkowski distance: A popular distance measure

$$d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ are two l-dimensional data objects, and p is the order (the distance so defined is called **L-*p* norm**)

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positivity)
  - $d(i, j) = d(j, i)$ (**Symmetry**)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (**Triangle Inequality**)



- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., *set difference*

Special Cases of Minkowski Distance

- p = 1: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{il} - x_{jl}|$$

- p = 2: (L_2 norm) Euclidean distance

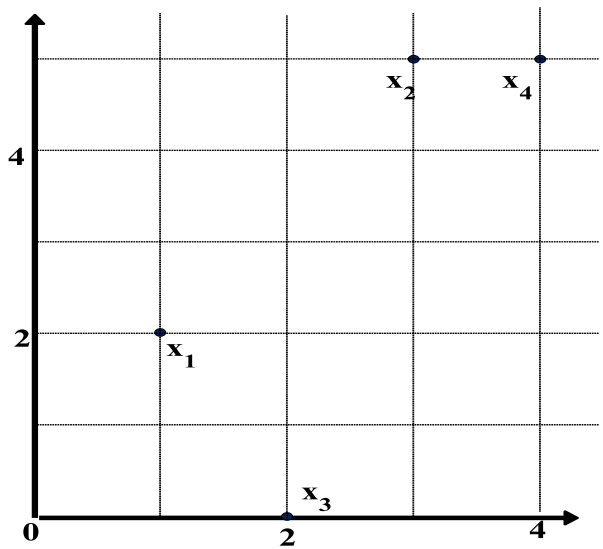
$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2}$$

- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{if}|$$

Example: Minkowski Distance at Special Cases

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum (L_∞)

L∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity Measure for Binary Attributes

- A contingency table for binary data

Object j

Object i

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Distance measure for *symmetric* binary variables:

$$d(i, j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for *asymmetric* binary variables:
  - *If **t** is very large*

$$d(i, j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (similarity measure for *asymmetric* binary variables):

$$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

- Note: Jaccard coefficient is the same as "coherence":

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q+r) + (q+s) - q}$$

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.
- Method 1: Simple matching
  - m: # of matches, p: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
  - Creating a new binary attribute for each of the M nominal states

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
- Can be treated like interval-scaled
  - Replace an ordinal variable value by its rank and map the range of each variable onto [0, 1]:
    - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
      - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3
  - Compute the dissimilarity using methods for interval-scaled variables

Cosine Similarity of Two Vectors

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- **Cosine measure:** If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: the length of vector d

Matrix Computation for Cosine Similarity

A =

	team	coach	hockey	baseball	soccer	penalty
D1	5	0	3	0	2	0
D2	3	0	2	0	1	1
D3	0	7	0	2	1	0
D4	0	1	0	0	1	3

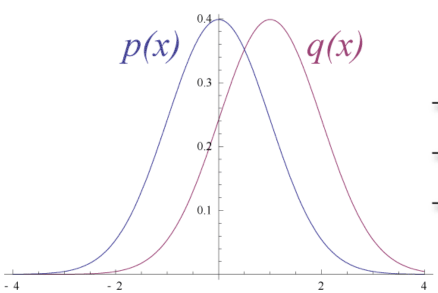
Normalize **A**

Calculate $\mathbf{AA}^T$

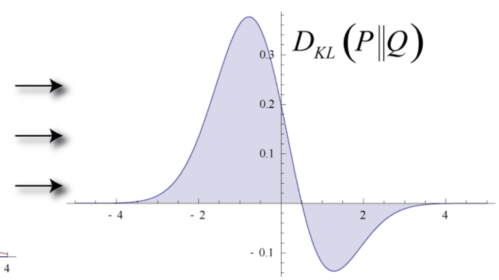
	D1	D2	D3	D4
D1	1.0	?	?	?
D2	?	1.0	?	?
D3	?	?	1.0	?
D4	?	?	?	1.0

KL Divergence: Comparing Two Probability Distributions

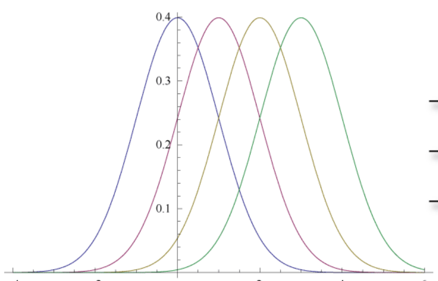
- *The Kullback-Leibler (KL) divergence:* Measure the **difference** between two probability distributions over the same variable *x*
  - From information theory, closely related to *relative entropy*, *information divergence*, and *information for discrimination*
- $D_{KL}(p(x) \,||\, q(x))$: divergence of $q(x)$ from $p(x)$, measuring the **information lost when $q(x)$ is used to approximate $p(x)$**

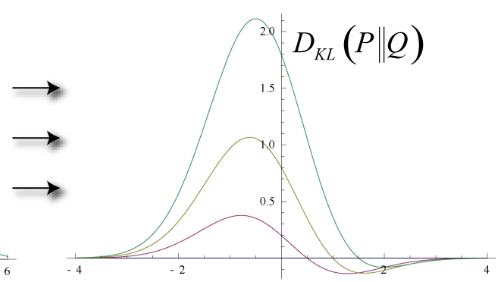


Original Gaussian PDF's



KL Area to be Integrated





Discrete form

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) \ln \frac{p(x)}{q(x)}$$

Continuous form

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

iPython (II)

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	▸ Carnegie Mellon University	18.5	150
4	2	▸ Massachusetts Institute of Technology	12.2	82
5	3	▸ Stanford University	10.9	54
6	3	▸ University of California - Berkeley	10.9	81
7	5	▸ Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	▸ Cornell University	8.7	68
9	7	▸ University of Michigan	8.6	63
10	8	▸ University of Washington	8.3	56
11	9	▸ University of California - San Diego	6.9	54
12	10	▸ Georgia Institute of Technology	6.8	75
13	11	▸ University of Wisconsin - Madison	5.9	47
14	12	▸ Columbia University	5.8	47
15	13	▸ University of Pennsylvania	5.6	46
16	14	▸ University of Southern California	5.5	49
17	15	▸ Princeton University	5.3	51
18	16	▸ University of Texas at Austin	5.2	42
19	16	▸ University of Maryland - College Park	5.2	44
20	18	▸ University of California - Los Angeles	5	37
21	19	▸ Northeastern University	4.8	54
22	19	▸ Purdue University	4.8	51
23	21	▸ University of Massachusetts Amherst	4.7	50
24	22	▸ New York University	4.5	47
25	23	▸ Harvard University	4.2	29
26	23	▸ University of California - Irvine	4.2	54
27	25	▸ Rutgers University	3.9	43
28	26	▸ University of California - Santa Barbara	3.5	25
29	27	▸ University of Utah	3.4	39
30	27	▸ Pennsylvania State University	3.4	31
31	29	▸ Stony Brook University	3.3	41
32	30	▸ University of California - Davis	3.2	29

126 institutions

Summary: Data Visualization

- Understand dimensionality, sparsity, and resolution
- Describe and generate plots
  - Boxplot, histogram, bar chart
  - Quantile plot, Q-Q plot
  - Scatter plot
- Understand graphical integrity
- Describe data proximity
  - Describe data dissimilarity: Calculate three Minkowski distance measures
  - Describe data similarity: Calculate Jaccard similarity, cosine similarity, and KL divergence

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009