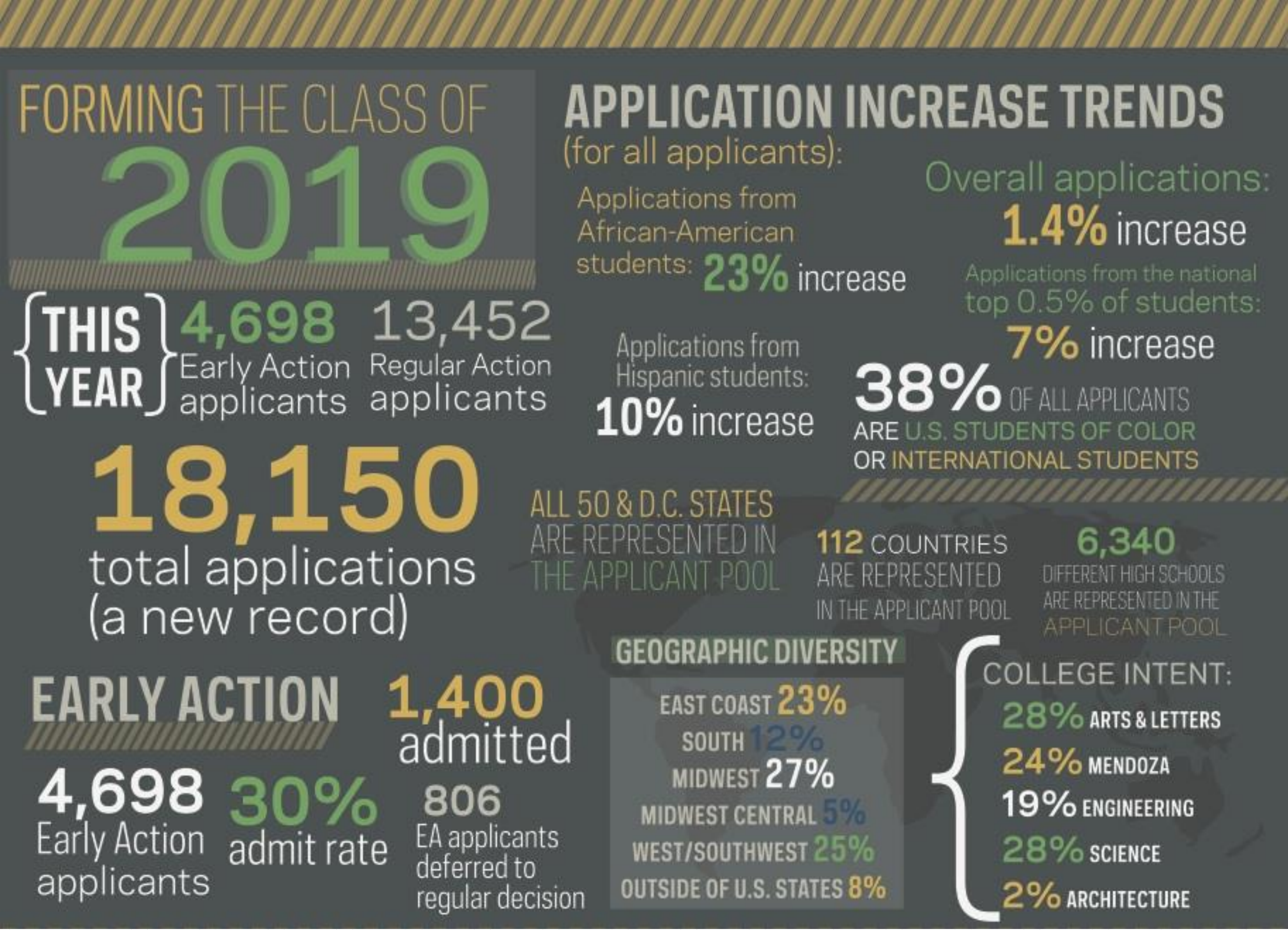
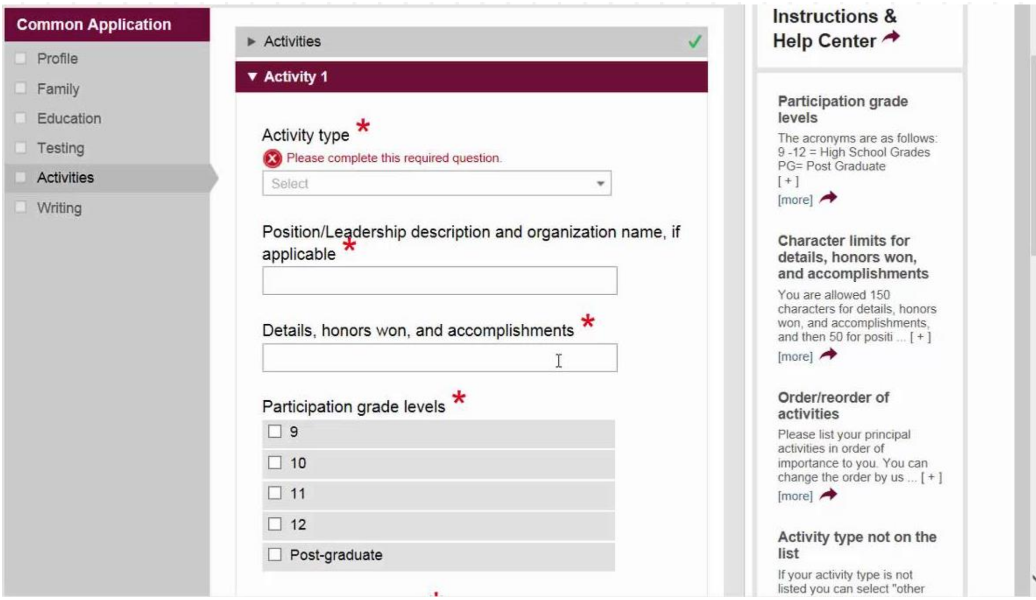
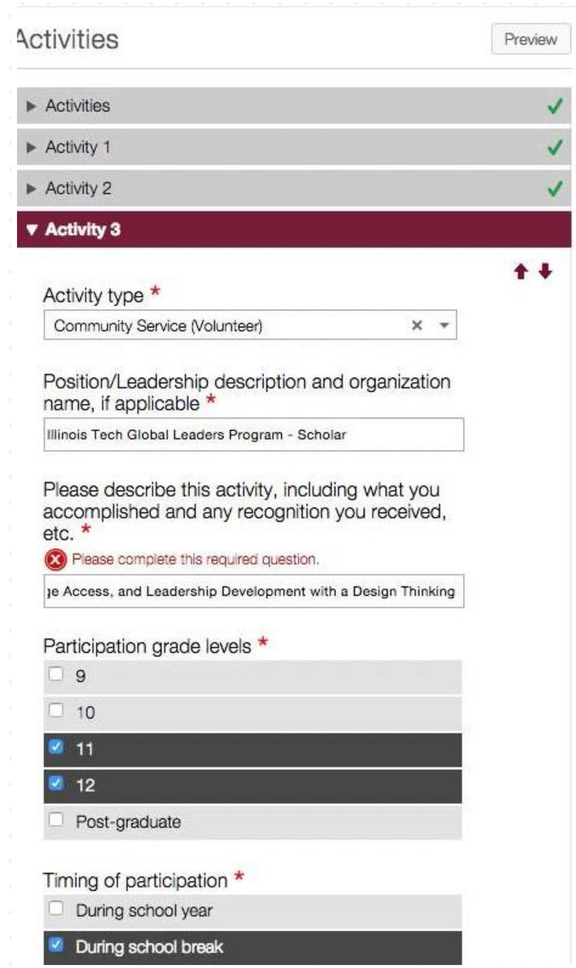
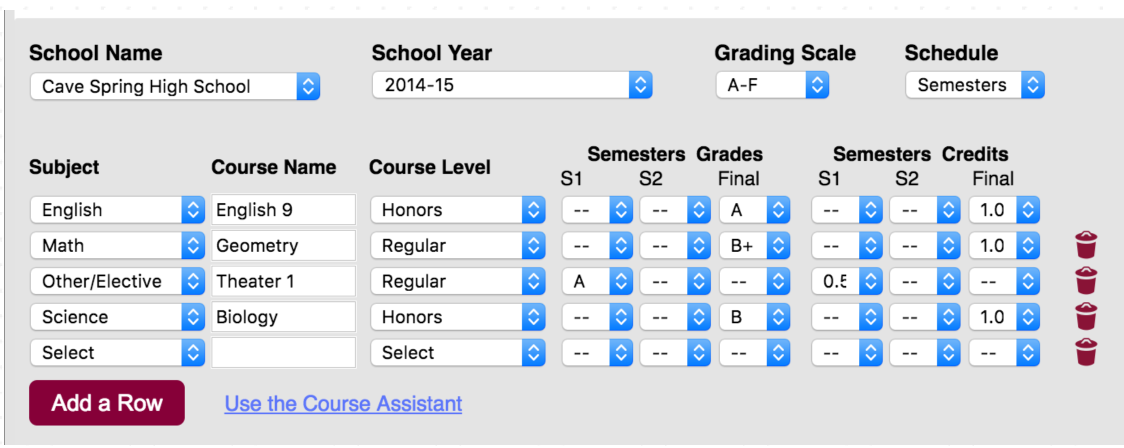


FORMING THE CLASS OF 2019
{THIS YEAR} 4,698 Early Action applicants 13,452 Regular Action applicants
18,150 total applications (a new record)
EARLY ACTION
1,400 admitted
4,698 Early Action applicants
30% admit rate
806 EA applicants deferred to regular decision
APPLICATION INCREASE TRENDS
(for all applicants):
Applications from African-American students: 23% increase
Overall applications: 1.4% increase
Applications from the national top 0.5% of students: 7% increase
Applications from Hispanic students: 10% increase
38% OF ALL APPLICANTS ARE U.S. STUDENTS OF COLOR OR INTERNATIONAL STUDENTS
ALL 50 & D.C. STATES ARE REPRESENTED IN THE APPLICANT POOL
112 COUNTRIES ARE REPRESENTED IN THE APPLICANT POOL
6,340 DIFFERENT HIGH SCHOOLS ARE REPRESENTED IN THE APPLICANT POOL
GEOGRAPHIC DIVERSITY
EAST COAST 23%
SOUTH 12%
MIDWEST 27%
MIDWEST CENTRAL 5%
WEST/SOUTHWEST 25%
OUTSIDE OF U.S. STATES 8%
COLLEGE INTENT:
28% ARTS & LETTERS
24% MENDOZA
19% ENGINEERING
28% SCIENCE
2% ARCHITECTURE

From Data To Knowledge

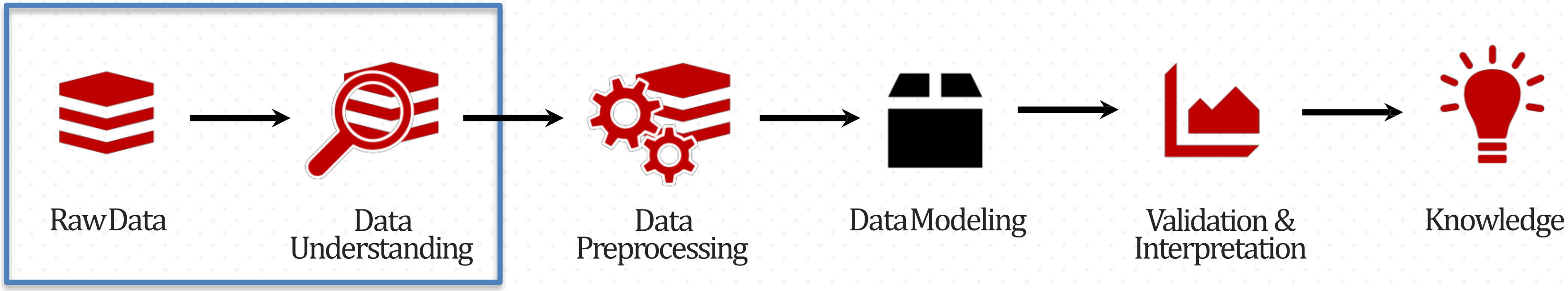






Data Science Pipeline

Chapter 2. Getting to Know Your Data



Data Acquisition

- What data do I need? What's available?
- Identify a domain expert, if available
  - Identify relevance of data
- Is the data sufficient?
  - Are there enough instances for each class?
- Do I have all relevant features?
  - Get a data dictionary

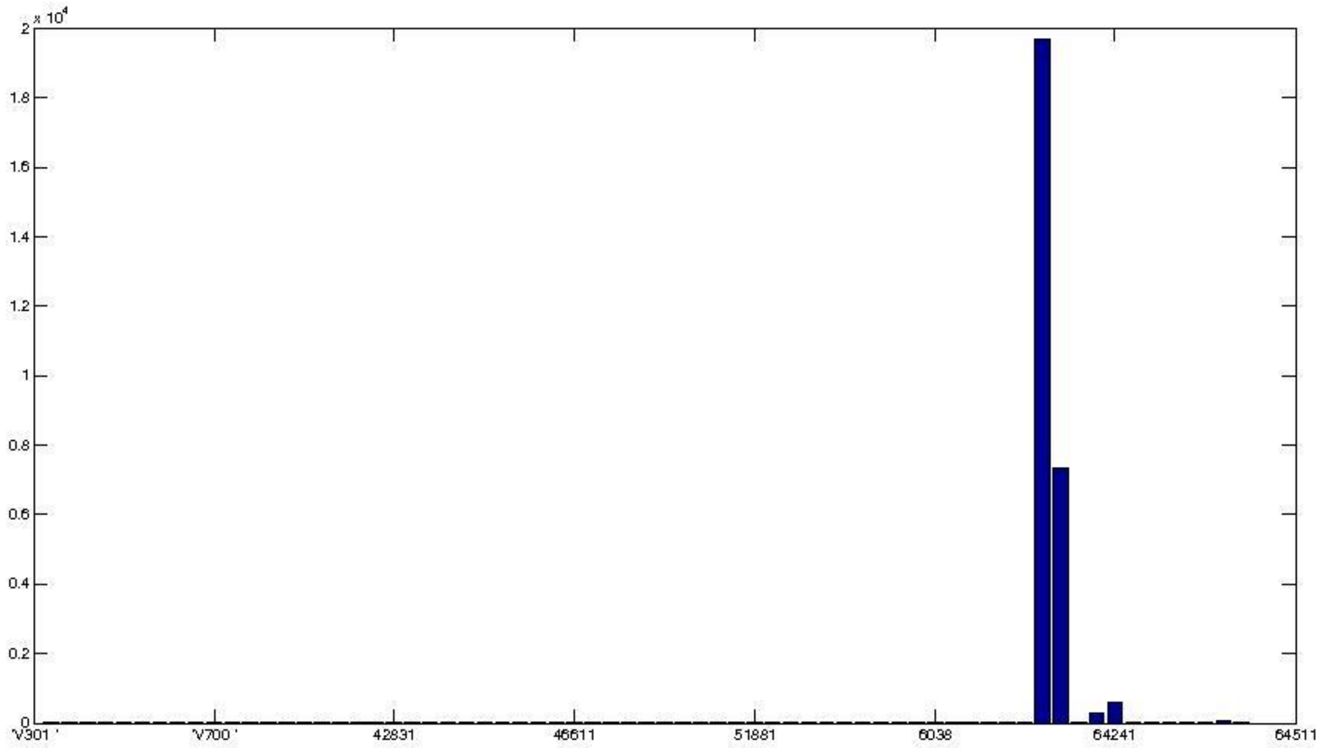
Data Acquisition

- **What data do I need? What's available?**
- Identify a domain expert, if available
  - Identify relevance of data
- Is the data sufficient?
  - Are there enough instances for each class?
- Do I have all relevant features?
  - Get a data dictionary

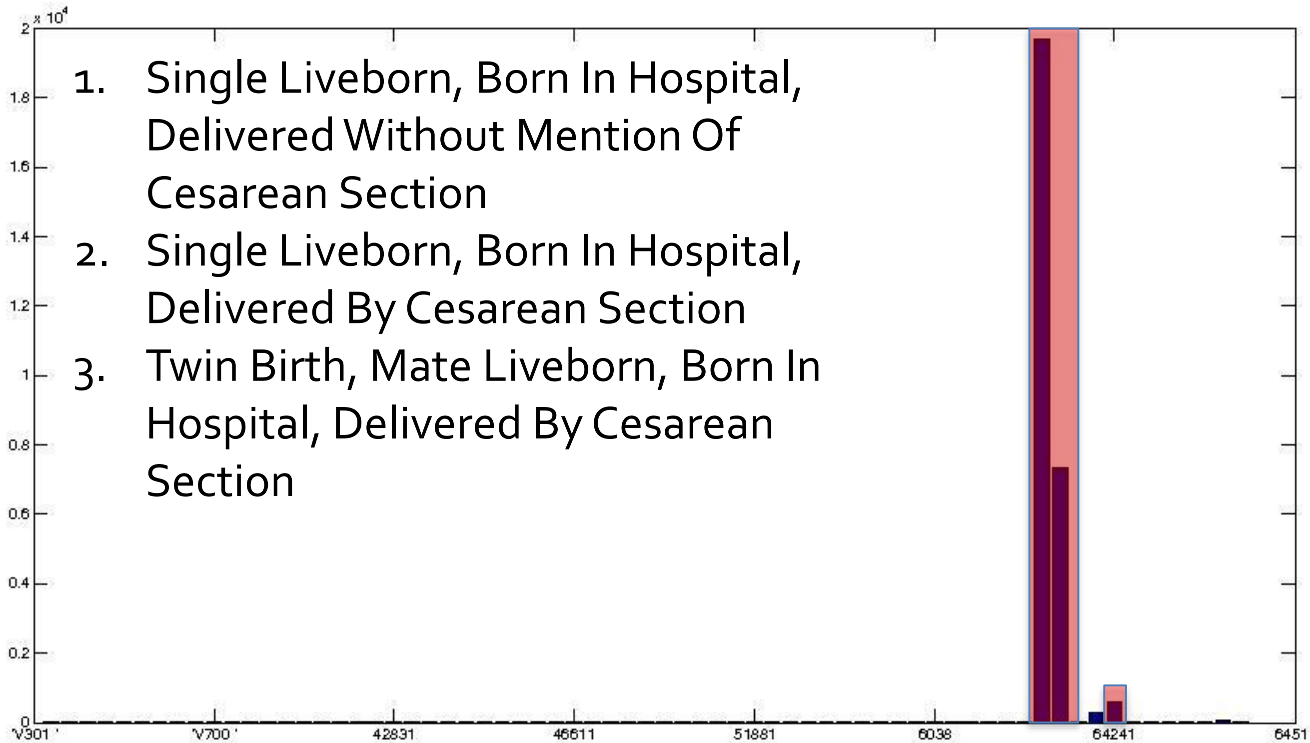
Data Acquisition

- What data do I need? What's available?
- **Identify a domain expert, if available**
  - **Identify relevance of data**
- Is the data sufficient?
  - Are there enough instances for each class?
- Do I have all relevant features?
  - Get a data dictionary

Primary Diagnosis NICU

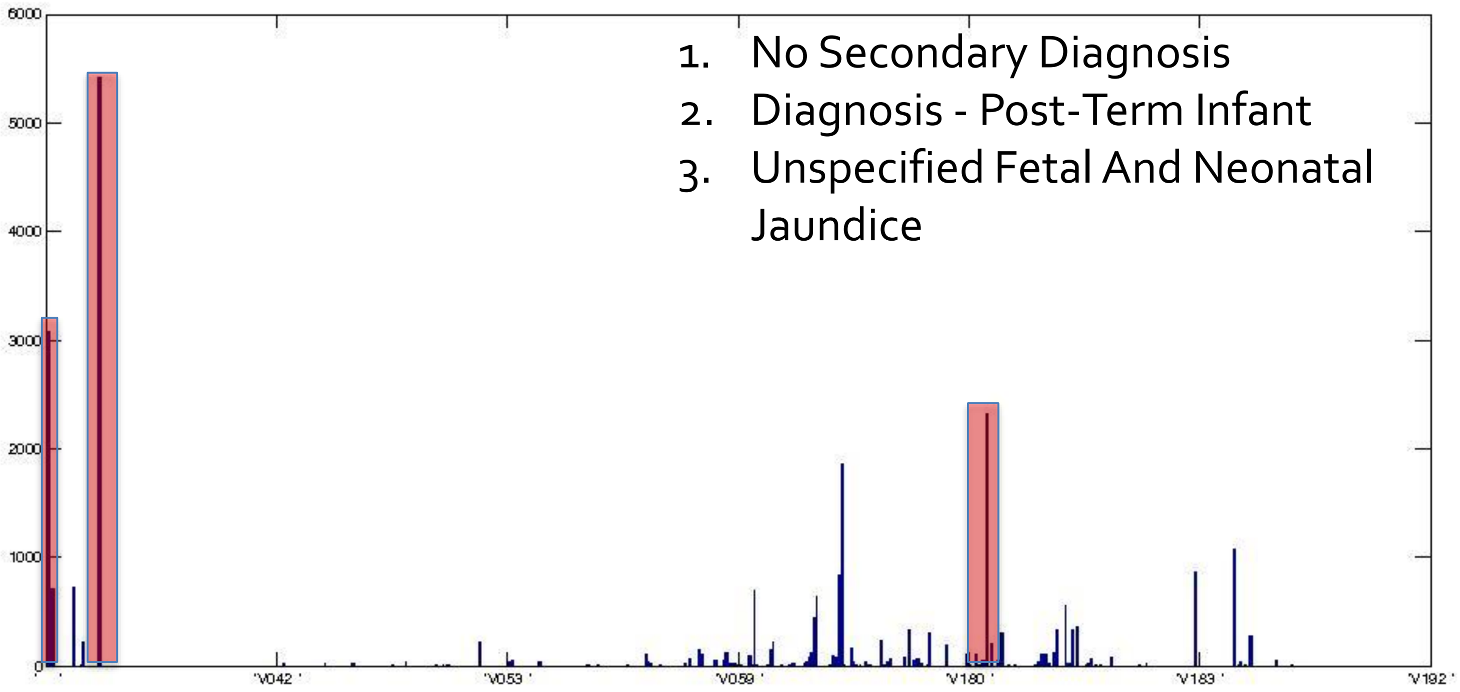


Primary Diagnosis



1. Single Liveborn, Born In Hospital, Delivered Without Mention Of Cesarean Section
2. Single Liveborn, Born In Hospital, Delivered By Cesarean Section
3. Twin Birth, Mate Liveborn, Born In Hospital, Delivered By Cesarean Section

Secondary Diagnosis

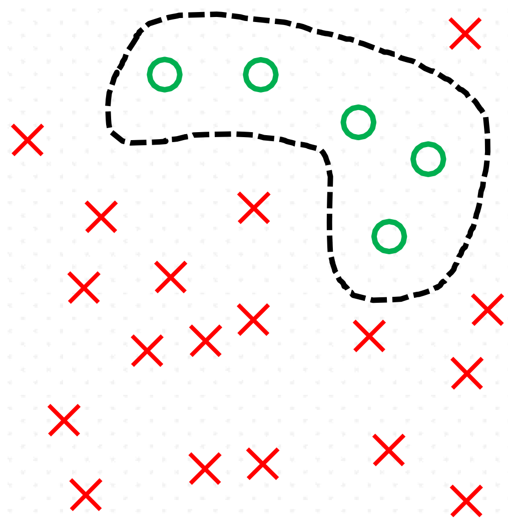


Data Acquisition

- What data do I need? What's available?
- Identify a domain expert, if available
  – Identify relevance of data
- **Is the data sufficient?**
  – **Are there enough instances for each class?**
- Do I have all relevant features?
  – Get a data dictionary

Quick Preview

Asymmetric / Imbalanced Classes



Data Acquisition

- What data do I need? What's available?
- Identify a domain expert, if available
  - Identify relevance of data
- Is the data sufficient?
  - Are there enough instances for each class?
- **Do I have all relevant features?**
  - **Get a data dictionary**

Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions
- Data Visualization
- Measuring Data Similarity and Dissimilarity

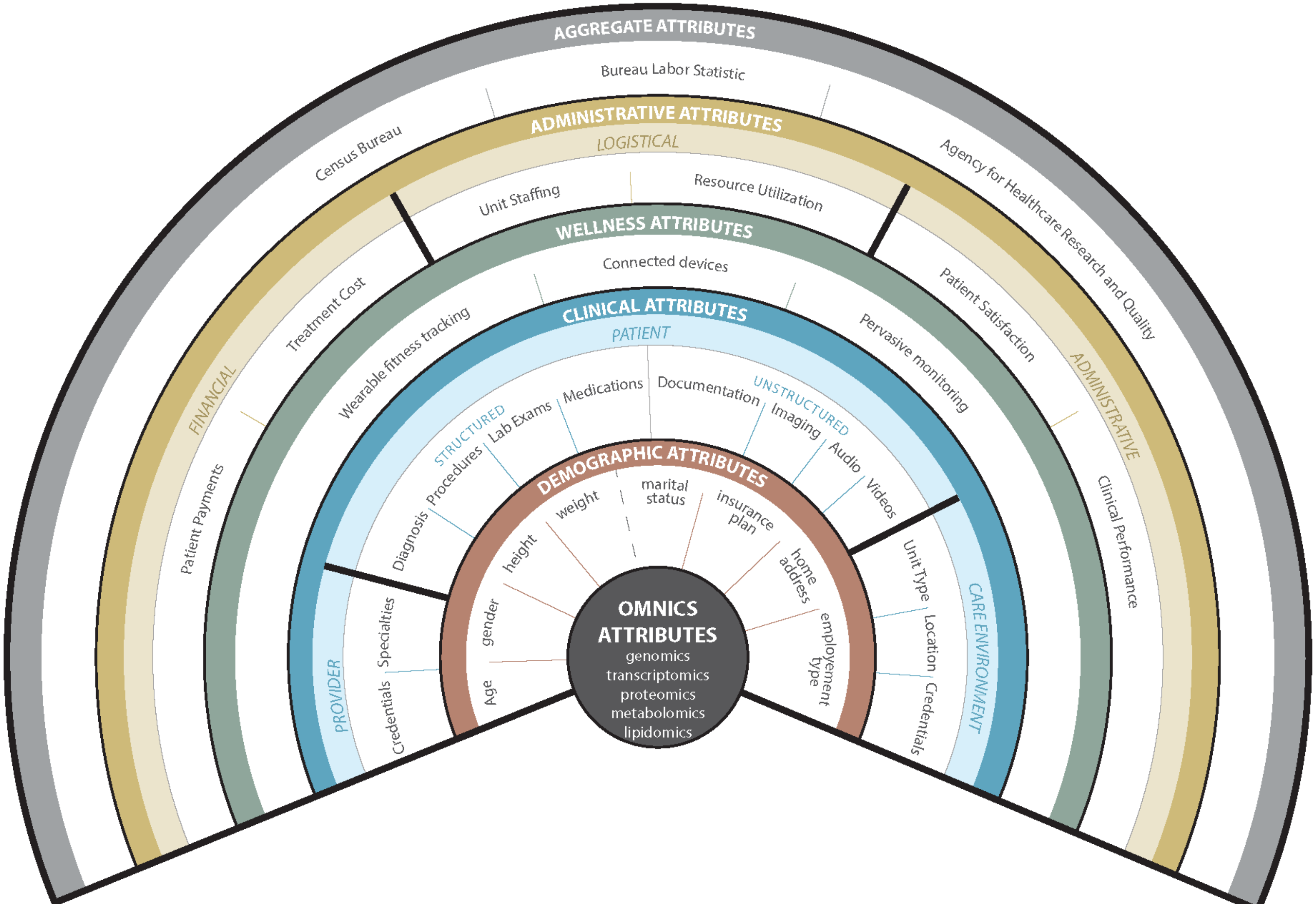
Chapter 2. Getting to Know Your Data

- **Data Objects and Attribute Types**
- **Basic Statistical Descriptions**
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Types of Data

Structured Data Sources **and** **Unstructured Data Sources**

Types of Data



Types of Data Sets: (1) Record Data

- Relational records in relational tables: highly structured
- Transaction data
- Document data: Term-frequency matrix of text documents

HOME TEAM: Notre Dame 26-9

			TOT-FG	3-PT		REBOUNDS									
##	Player Name		FG-FGA	FG-FGA	FT-FTA	OF	DE	TOT	PF	TP	A	TO	BLK	S	MIN
03	VJ Beachem..........	f	1-9	0-3	0-0	0	6	6	1	2	3	0	0	1	37
35	Bonzie Colson.......	f	6-13	0-1	6-10	2	5	7	2	18	2	0	2	1	31
00	Rex Pflueger........	g	2-3	0-0	0-0	0	2	2	2	4	0	1	0	0	28
05	Matt Farrell........	g	6-9	3-5	1-3	0	4	4	2	16	4	3	0	2	36
32	Steve Vasturia......	g	3-12	1-2	3-4	3	5	8	0	10	1	0	0	0	37
01	Austin Torres.......		0-1	0-0	0-0	1	0	1	0	0	0	1	1	0	7
02	TJ Gibbs............		0-1	0-0	2-2	0	2	2	1	2	0	0	0	0	13
04	Matt Ryan...........		2-3	0-0	2-2	0	2	2	0	6	0	0	0	0	9
23	Martinas Geben......		1-1	0-0	0-0	1	0	1	1	2	0	1	0	0	2
	TEAM................					2	1	3							
	Totals..............		21-52	4-11	14-21	9	27	36	9	60	10	6	3	4	200

TOTAL FG%	1st Half:	14-30	46.7%	2nd Half:	7-22	31.8%	Game: 40.4%	DEADB
3-Pt. FG%	1st Half:	2-5	40.0%	2nd Half:	2-6	33.3%	Game: 36.4%	REBS
F Throw %	1st Half:	6-8	75.0%	2nd Half:	8-13	61.5%	Game: 66.7%	3

Representation

Input Features: X
(attributes)

Class: Y

Data Objects (Instances)

	x_1	x_2	..	x_m	y
1					
2					
.					
.					
n					

Examples

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

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

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - Sales database: customers, store items, sales.
  - Medical database: patients, treatments.
  - University database: students, professors, courses.
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples.*
- Data objects are described by **attributes.**
- Database: (often) rows → data objects; columns → attributes.

Attributes

- **Attribute** (or features, variables)
  - A data field, representing a characteristic or feature of a data object
- Types:
  - Nominal (e.g., red, blue)
  - Binary (e.g., {true, false})
  - Ordinal (e.g., {freshman, sophomore, junior, senior})
  - Numeric: quantitative

Nominal Attributes

- Qualitative features.
  - Enough information to distinguish one object from another.
- Has only a reasonable set of values.
  - Thumb-rule: count with your fingers.
  - Can be many more 1000's ICD-9 Codes
- Often represented as integer variables.
  - For example: 0 for red; 1 for blue; etc.

Nominal Attributes – Special Cases

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., ________
  - Asymmetric binary: outcomes not equally important.
    - e.g., ________, ________
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - *Size* = {*small*, *medium*, *large*}, ________, ________

Attribute Types

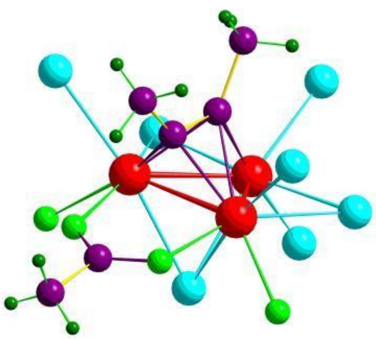
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

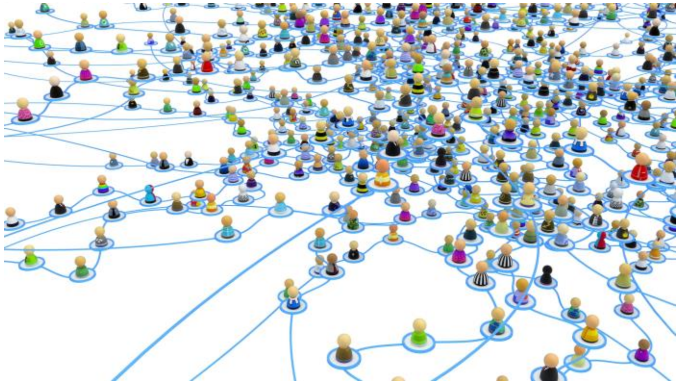
Continuous Features

- 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.

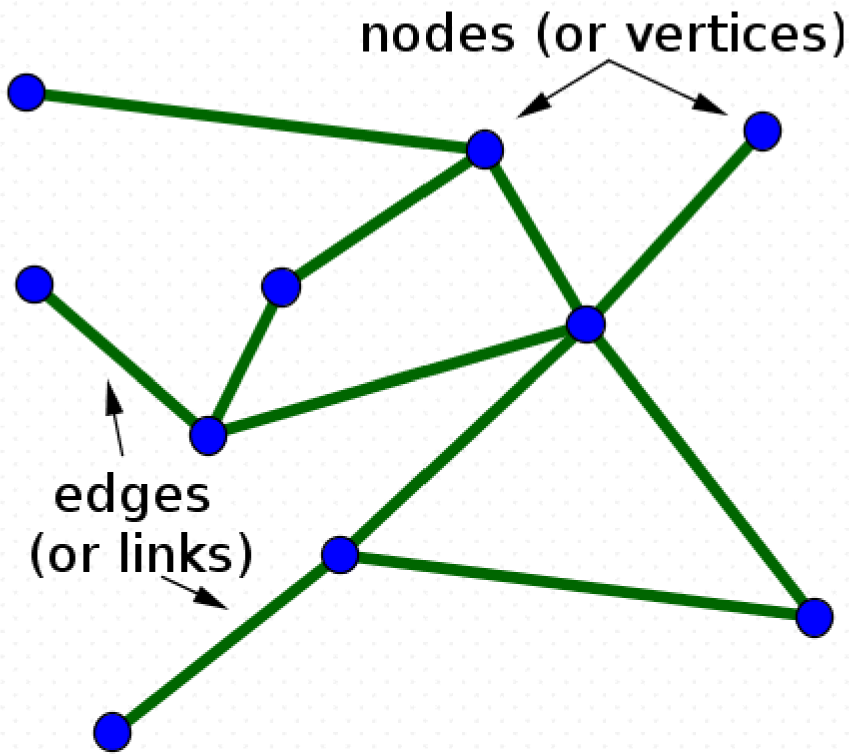
Types of Data Sets: (2) Graphs and Networks

- Transportation networks
- World Wide Web
- Molecular structures
- Social or information networks



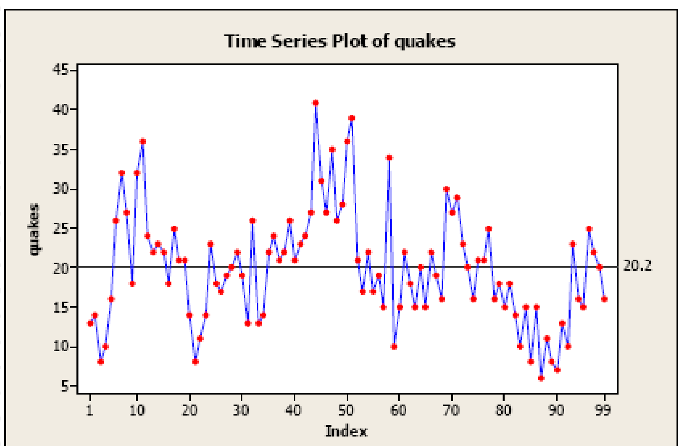


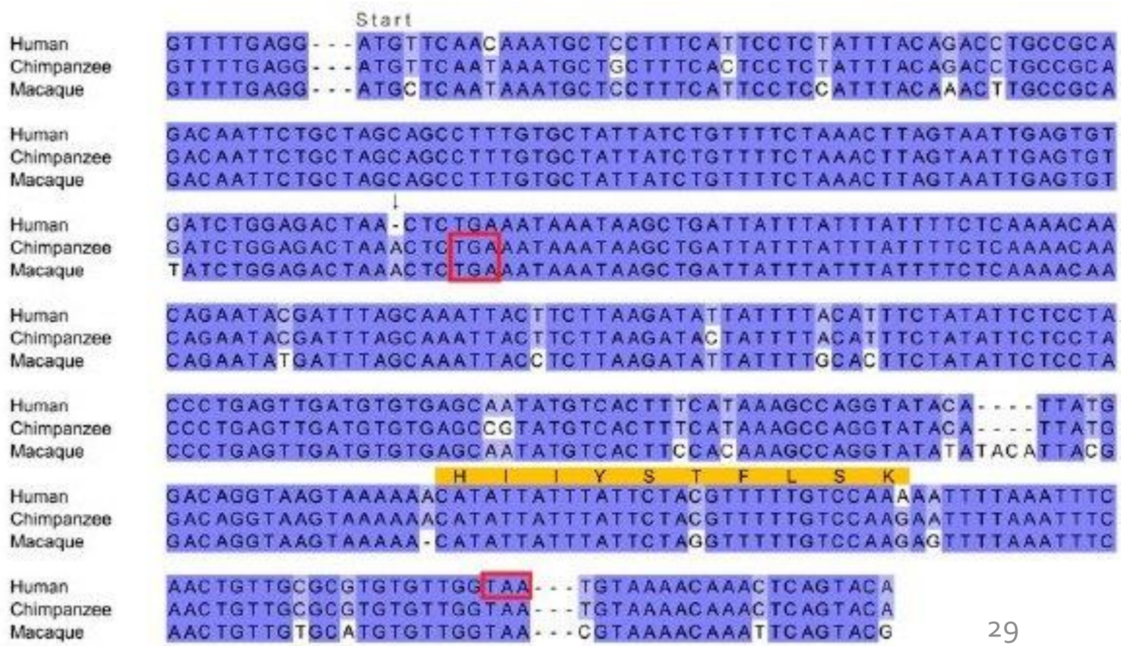
Representation



Types of Data Sets: (3) Ordered Data

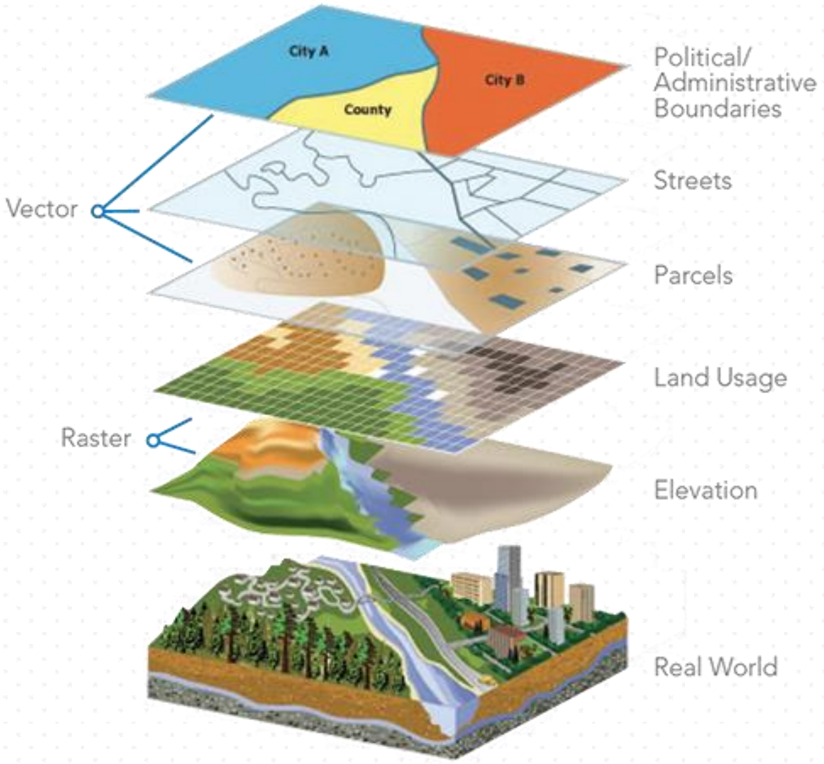
- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data





Other Types of Data Sets

- Spatial data
- Image and multimedia data



Break From the Slides

Chapter 2. Getting to Know Your Data

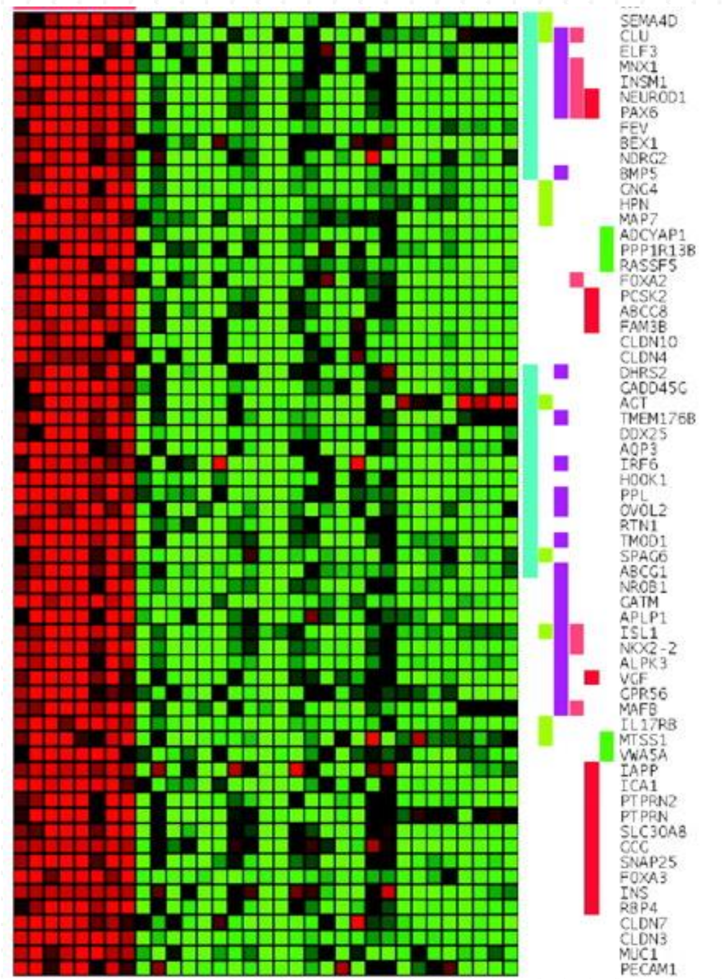
- Data Objects and Attribute Types
- **Basic Statistical Descriptions**
- Data Visualization
- Measuring Data Similarity and Dissimilarity

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)

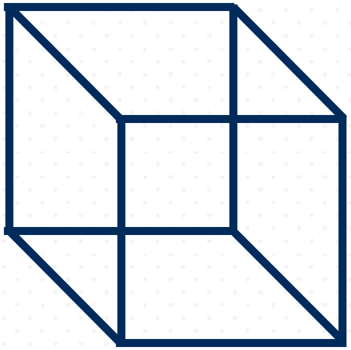
Class: Y

Data Objects (Instances)

	x_1	x_2	..	x_m	y
1					
2					
.					
.					
n					

Curse of Dimensionality

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



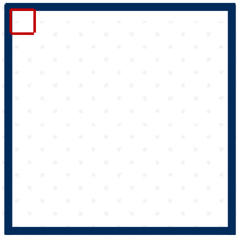
Curse of Dimensionality

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

The short line is 0.01 of the length of the long line.

Curse of Dimensionality

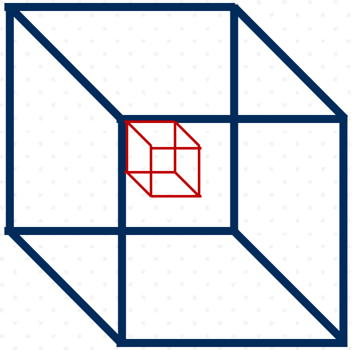
- In 2 dimensions, we must go a distance of $\sqrt{0.01} = 0.1$ on average to reach our nearest neighbor.



The small square contains 0.01 of the volume of the large square.

Curse of Dimensionality

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



The small cube contains 0.01 the volume of the large cube.

Curse of Dimensionality

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

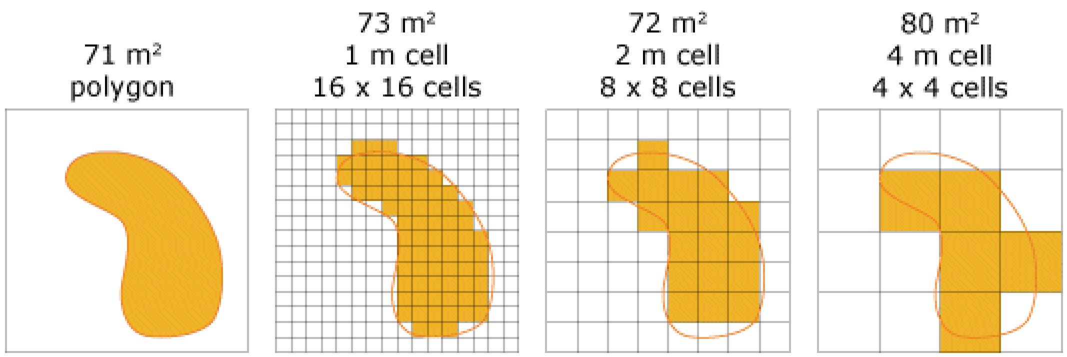
Data Sparsity

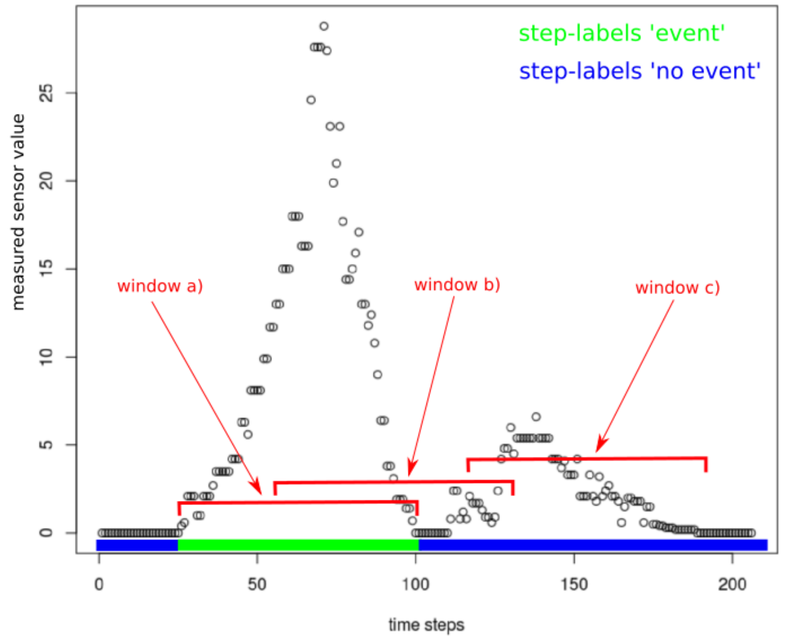
- For some datasets, most features have values of 0.
- Can be a problem for many methods.
  - Can create statistical bias due to small samples.
  - Can reduce the meaningfulness of distance calculations.
- Can also be an advantage.
  - Requires less storage.

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





Attributes

Are all attributes the same?

Are all attributes collected as raw data?

Engineering Activity

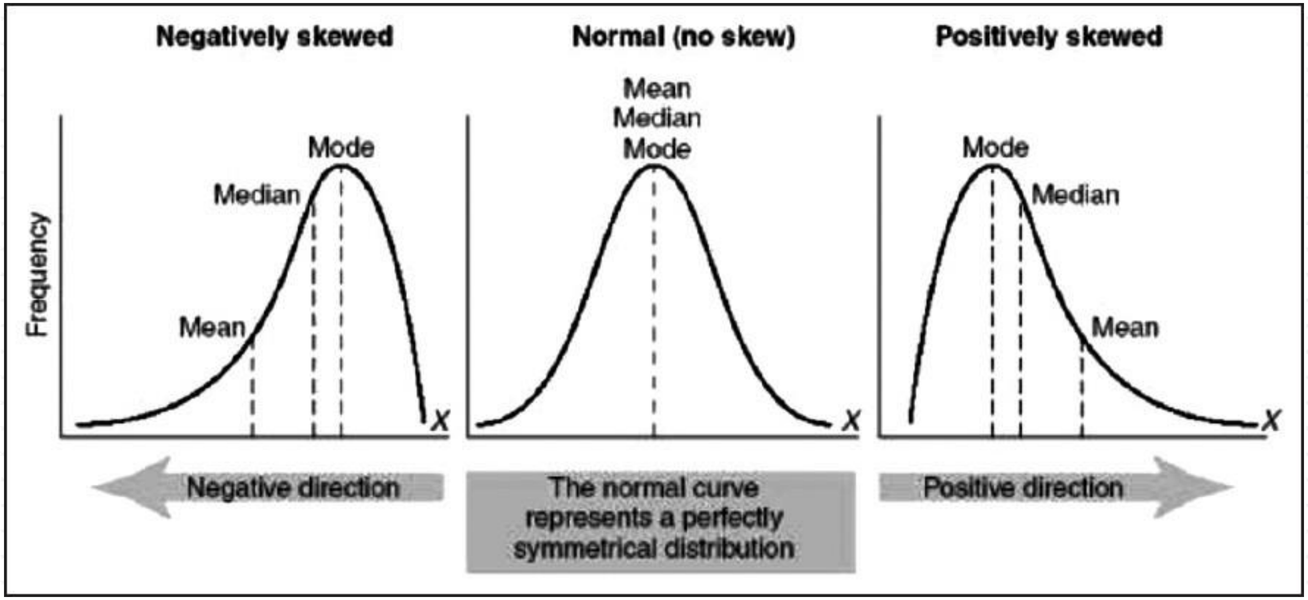
Lat 1	Long 1	Lat 2	Long 2	Walk
48.8715	2.354	48.8721	2.3549	Yes
48.87211	2.3549	44.597	-123.24	No
48.872232	2.354211	48.872	2.3549	Yes
44.597422	-123.248367	48.872232	2.354211	No

Engineering Activity

Lat 1	Long 1	Lat 2	Long 2	Distance	Walk
48.8715	2.354	48.8721	2.3549	2	Yes
48.87211	2.3549	44.597	-123.24	9059	No
48.872232	2.354211	48.872	2.3549	5	Yes
44.597422	-123.248367	48.872232	2.354211	9056	No

Basic Statistical Descriptions of Data

- Motivation: to better understand the data
- Data characteristics
  - Central Tendency: Mean, median, mode
  - Spread : Variance, standard deviation, max, min, Z-score



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 $x_{50\%}$

Measuring the Central Tendency: (1) Mean and (2) Median

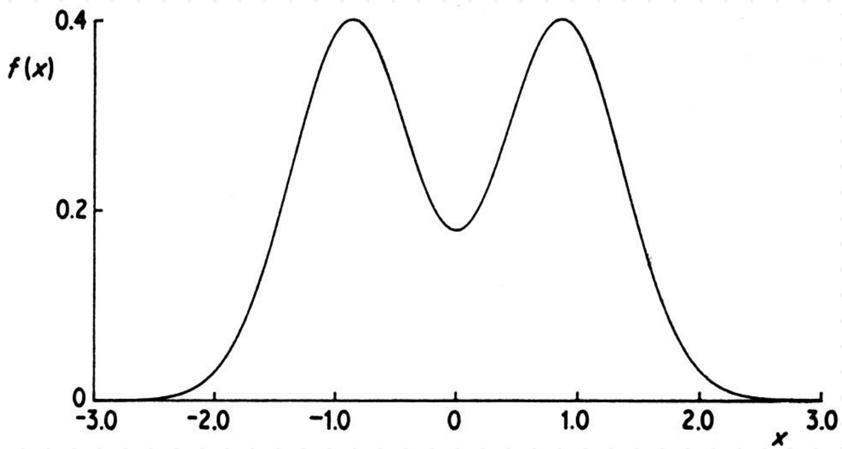
- Mean (sample vs. population):
  - Note: n is **sample** size and N is **population** size.

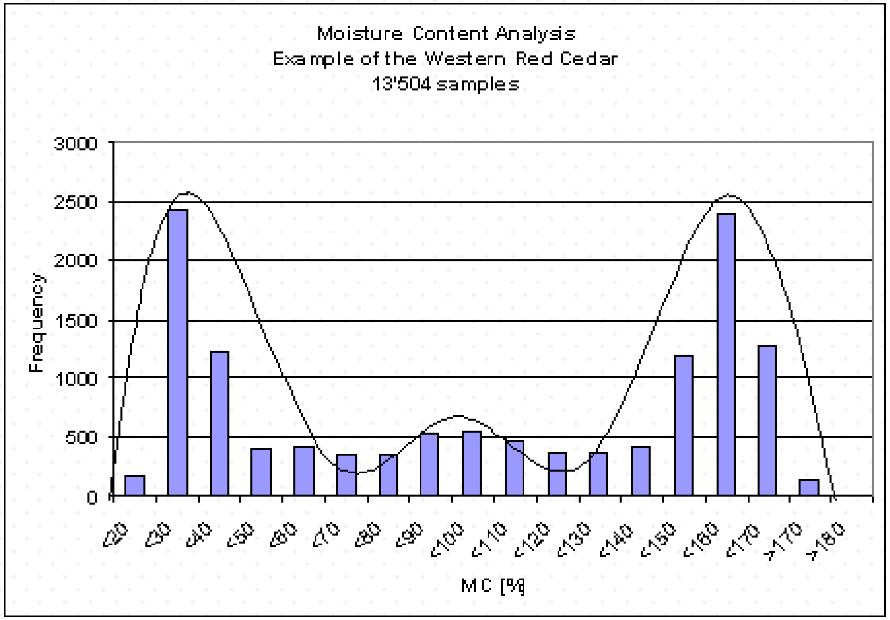
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Trimmed mean: Chopping extreme values
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise

Measuring the Central Tendency: (3) Mode

- Mode: Value that occurs most frequently in the data
- Multi-modal
  - Bimodal
  - Trimodal





Frequency

- 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 notions of frequency and mode are typically used with categorical data.

Variance and Standard Deviation

- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Standard deviation s (or σ) is **square root** of variance s^2 (or σ^2)

Back to iPython

Measuring the Outlierness: Variance and Standard Deviation

- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Why?

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Standard deviation s (or σ) is **square root** of variance s^2 (or σ^2)

Bias in Population Estimates

Population Mean = 4

- Consider we have samples 2, 5, 11

Bias in Population Estimates

Population Mean = 4

- Consider we have samples 2, 5, 11
  - Mean = 6
  - Median = 5

Bias in Population Estimates

Population Mean = 4

- Consider we have samples 2, 5, 11
  - Mean = 6
  - **Median = 5**

Bias in Population Estimates

Population Mean = 4

- Consider we have samples 2, 6, 7
  - **Mean = 5**
  - Median = 6

How About Variance?

Suppose we have 3 cards in a bag

0 2 4

$$\mu = \frac{0+2+4}{3} = 2$$

$$\sigma^2 = \frac{(0-2)^2 + (2-2)^2 + (4-2)^2}{3} = \frac{8}{3}$$

Sample Variance (Unbiased)

$$\bar{x} = \frac{\sum x}{n} \qquad S^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$(0,0) \quad \frac{0+0}{2} = 0 \quad \frac{(0-0)^2 + (0-0)^2}{1} = 0$$

$$(0,2) \quad \frac{0+2}{2} = 1 \quad \frac{(0-1)^2 + (2-1)^2}{1} = 2$$

$$(0,4) \quad \frac{0+4}{2} = 2 \quad \frac{(0-2)^2 + (4-2)^2}{1} = 8$$

$$(2,0) \quad \frac{2+0}{2} = 1 \quad \frac{(2-1)^2 + (0-1)^2}{1} = 2$$

$$(2,2) \quad \frac{2+2}{2} = 2 \quad \frac{(2-2)^2 + (2-2)^2}{1} = 0$$

$$(2,4) \quad \frac{2+4}{2} = 3 \quad \frac{(2-3)^2 + (4-3)^2}{1} = 2$$

$$(4,0) \quad \frac{4+0}{2} = 2 \quad \frac{(4-2)^2 + (0-2)^2}{1} = 8$$

$$(4,2) \quad \frac{4+2}{2} = 3 \quad \frac{(4-3)^2 + (2-3)^2}{1} = 2$$

$$(4,4) \quad \frac{4+4}{2} = 4 \quad \frac{(4-4)^2 + (4-4)^2}{1} = 0$$

Sample Variance (Unbiased)

Sample Mean

$$\frac{0+1+2+1+2+3+2+3+4}{9} = 2$$

Sample Variance (Unbiased)

$$\frac{0+2+8+2+0+2+8+2+0}{9} = \frac{8}{3}$$

$$(0,0) \quad \frac{0+0}{2} = 0 \quad \frac{(0-0)^2+(0-0)^2}{1} = 0$$

$$(0,2) \quad \frac{0+2}{2} = 1 \quad \frac{(0-1)^2+(2-1)^2}{1} = 2$$

$$(0,4) \quad \frac{0+4}{2} = 2 \quad \frac{(0-2)^2+(4-2)^2}{1} = 8$$

$$(2,0) \quad \frac{2+0}{2} = 1 \quad \frac{(2-1)^2+(0-1)^2}{1} = 2$$

$$(2,2) \quad \frac{2+2}{2} = 2 \quad \frac{(2-2)^2+(2-2)^2}{1} = 0$$

$$(2,4) \quad \frac{2+4}{2} = 3 \quad \frac{(2-3)^2+(4-3)^2}{1} = 2$$

$$(4,0) \quad \frac{4+0}{2} = 2 \quad \frac{(4-2)^2+(0-2)^2}{1} = 8$$

$$(4,2) \quad \frac{4+2}{2} = 3 \quad \frac{(4-3)^2+(2-3)^2}{1} = 2$$

$$(4,4) \quad \frac{4+4}{2} = 4 \quad \frac{(4-4)^2+(4-4)^2}{1} = 0$$

Sample Variance

$$\bar{x} = \frac{\sum x}{n} \qquad S^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$(0,0) \quad \frac{0+0}{2} = 0 \quad \frac{(0-0)^2 + (0-0)^2}{2} = 0$$

$$(0,2) \quad \frac{0+2}{2} = 1 \quad \frac{(0-1)^2 + (2-1)^2}{2} = 1$$

$$(0,4) \quad \frac{0+4}{2} = 2 \quad \frac{(0-2)^2 + (4-2)^2}{2} = 4$$

$$(2,0) \quad \frac{2+0}{2} = 1 \quad \frac{(2-1)^2 + (0-1)^2}{2} = 1$$

$$(2,2) \quad \frac{2+2}{2} = 2 \quad \frac{(2-2)^2 + (2-2)^2}{2} = 0$$

$$(2,4) \quad \frac{2+4}{2} = 3 \quad \frac{(2-3)^2 + (4-3)^2}{2} = 1$$

$$(4,0) \quad \frac{4+0}{2} = 2 \quad \frac{(4-2)^2 + (0-2)^2}{2} = 4$$

$$(4,2) \quad \frac{4+2}{2} = 3 \quad \frac{(4-3)^2 + (2-3)^2}{2} = 1$$

$$(4,4) \quad \frac{4+4}{2} = 4 \quad \frac{(4-4)^2 + (4-4)^2}{2} = 0$$

Sample Variance

Sample Mean

$$\frac{0+1+2+1+2+3+2+3+4}{9} = 2$$

Sample Variance (Unbiased)

$$\frac{0+1+4+1+0+1+4+1+0}{9} = \frac{4}{3}$$

$$(0,0) \quad \frac{0+0}{2} = 0 \quad \frac{(0-0)^2+(0-0)^2}{2} = 0$$

$$(0,2) \quad \frac{0+2}{2} = 1 \quad \frac{(0-1)^2+(2-1)^2}{2} = 1$$

$$(0,4) \quad \frac{0+4}{2} = 2 \quad \frac{(0-2)^2+(4-2)^2}{2} = 4$$

$$(2,0) \quad \frac{2+0}{2} = 1 \quad \frac{(2-1)^2+(0-1)^2}{2} = 1$$

$$(2,2) \quad \frac{2+2}{2} = 2 \quad \frac{(2-2)^2+(2-2)^2}{2} = 0$$

$$(2,4) \quad \frac{2+4}{2} = 3 \quad \frac{(2-3)^2+(4-3)^2}{2} = 1$$

$$(4,0) \quad \frac{4+0}{2} = 2 \quad \frac{(4-2)^2+(0-2)^2}{2} = 4$$

$$(4,2) \quad \frac{4+2}{2} = 3 \quad \frac{(4-3)^2+(2-3)^2}{2} = 1$$

$$(4,4) \quad \frac{4+4}{2} = 4 \quad \frac{(4-4)^2+(4-4)^2}{2} = 0$$

Biased Sample Variance

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu) + (\mu - \bar{X})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + 2(\bar{X} - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\mu - \bar{X})^2
\end{aligned}$$

Biased

Unbiased

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Bessel's Correction: 3 alternative proofs of correctness

Thinking Ahead

- Variance and standard deviation (sample: s, population: σ)
  - Variance: (algebraic, scalable computation)
    - **Q: Can you compute it incrementally and efficiently?**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Standard deviation s (or σ) is **square root** of variance s^2 (or σ^2)

Multivariate Measures

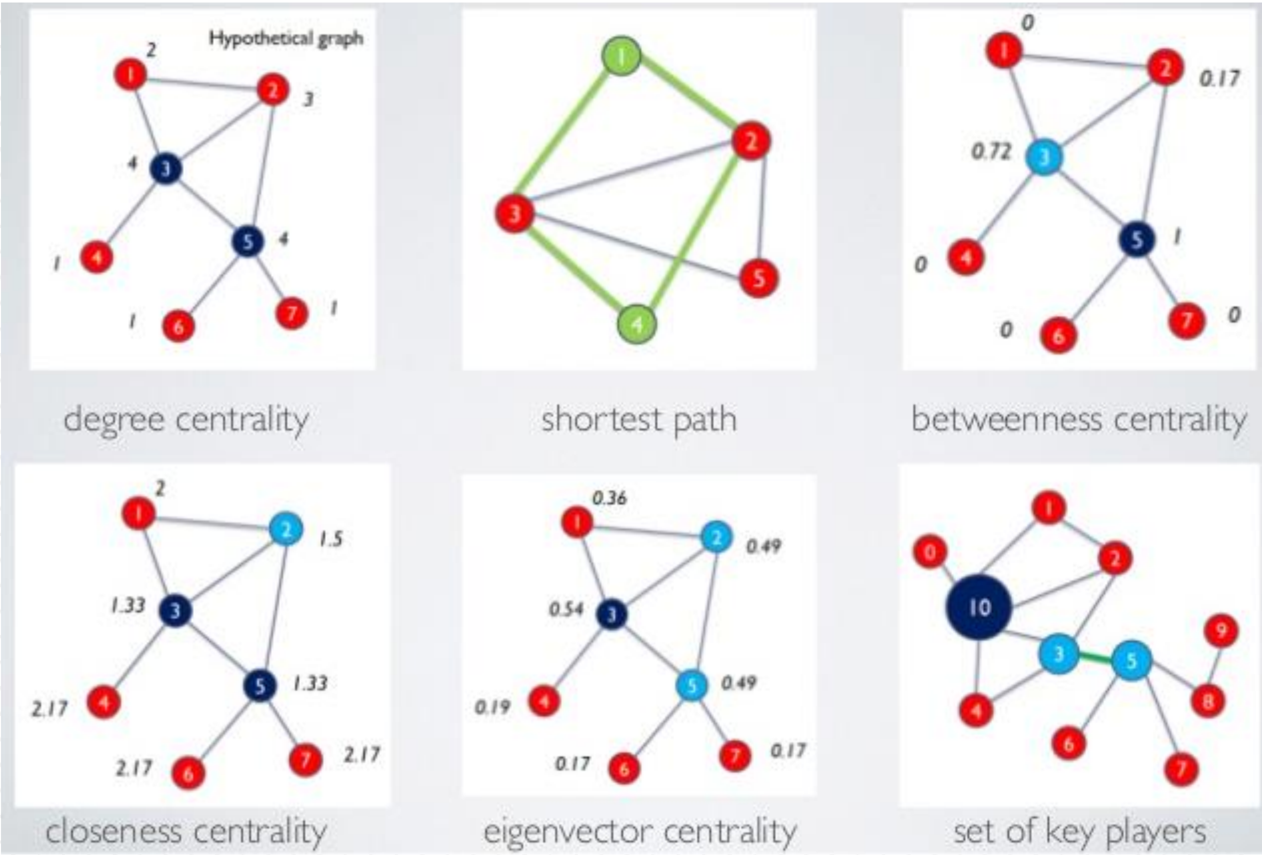
- The **covariance** is a measure of the degree to which two variables vary together, and is given by:

$$\text{Covariance}(x_i, x_j) = \frac{1}{m-1}\sum_{i=1}^{m}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

- Where x_{ki} and x_{kj} are the values of the i^{th} and j^{th} features for the k^{th} object

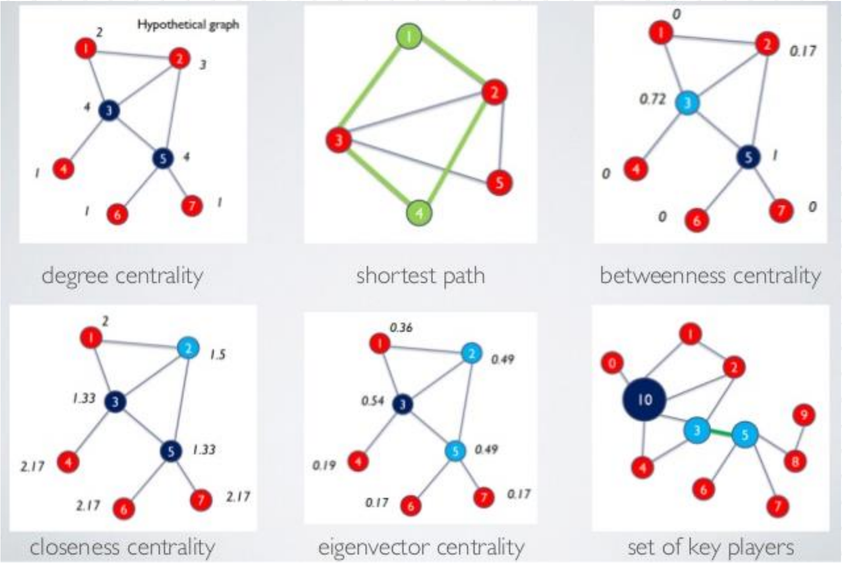
Additional Representation Metrics

Probably not required, but interesting!



Additional Representation Metrics

Probably not required, but interesting!



Degree: How many people can this person reach directly

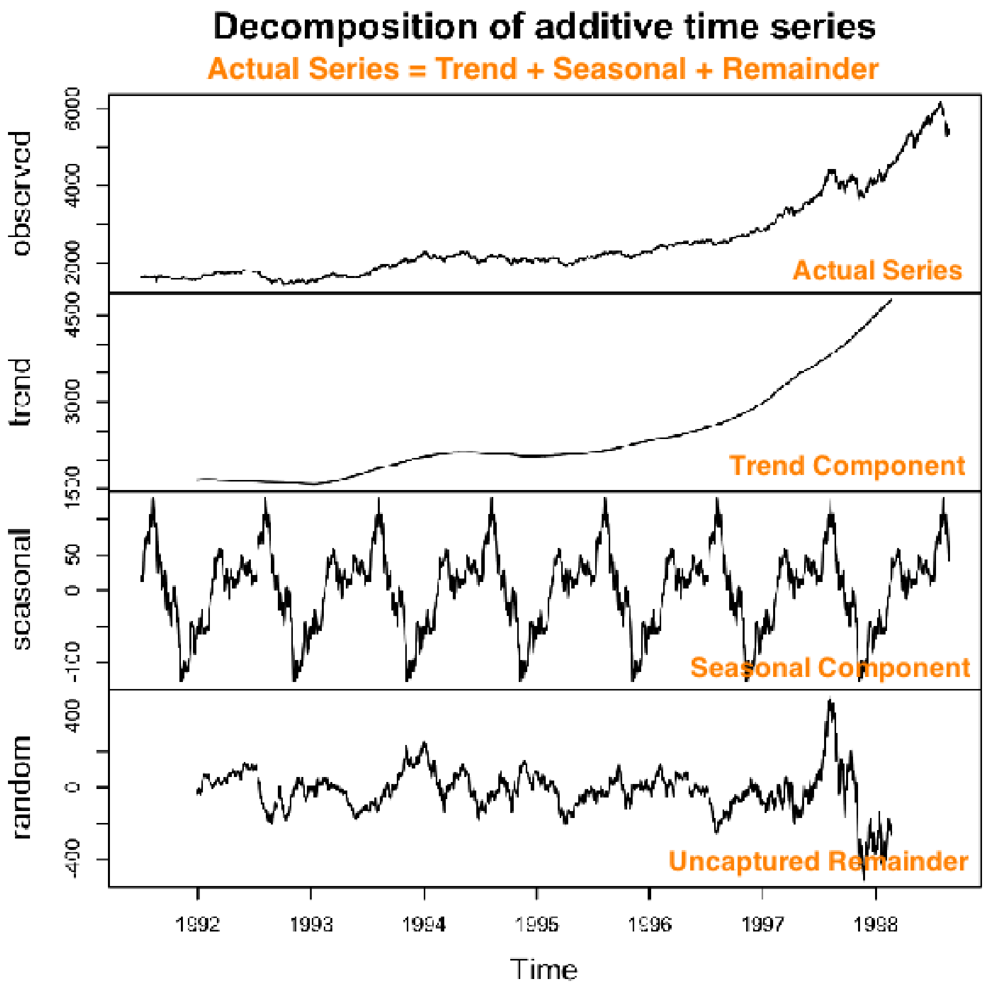
Betweeness: How likely is this person to be the most direct route between two people in the network?

Closeness: How fast can this person reach everyone in the network?

Eigenvector: How well is this person connected to other well-connected people

Time Series

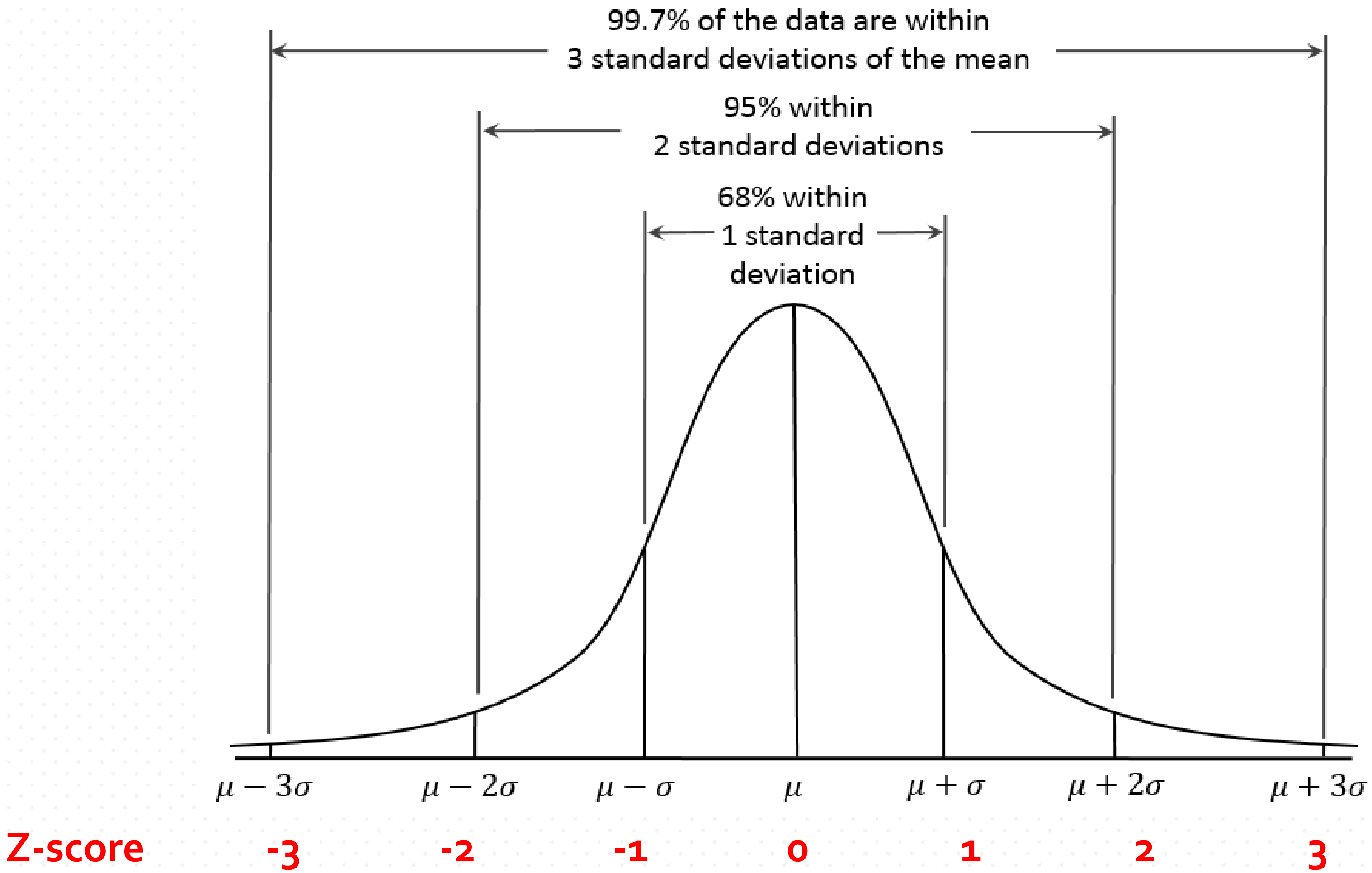
Probably not required, but interesting!



Distributional Analysis

Measuring the Outlierness: Properties of Normal Distribution Curve

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation



Discussion

- Can you use Z-score to automatically find phrases?
  - If we have 1,000 “matrix” and 1,000 “factorization” in 1,000,000 words, and we assume independency, we should have only one “matrix factorization” (expected).
  - But actually we have more! - Outlierness

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han. “**Automated Phrase Mining from Massive Text Corpora**”. Submitted to Transactions on Knowledge and Data Engineering.

Normalization

The goal of normalization is to make an entire set of values have a particular property.

Data Transformation: Normalization

- Normalization is often performed on data to remove amplitude variation and only focus on the underlying distribution shape.

- Makes training less sensitive to the scale of features:
  - Consider a regression problem where you're given features of an apartment and are required to predict the price of the apartment. Let's say there are 2 features — no. of bedrooms and the area of the apartment. Now, the no. of bedrooms will be in the range 1–4 typically, while the area will be in the range 100–200m^2. Modelling the task as linear regression you want to solve for coefficients w_1 and w_2 corresponding to no. of bedrooms and area. Now, because of the scale of the features, a small change in w_2 will change the prediction by a lot compared to the same change in w_1, to the point that setting w_2 correctly might dominate the optimization process.

- Sometimes used in order to speed up the convergence.

Data Transformation: Normalization

- Min-max normalization
- Z-score normalization
- Normalization by decimal scaling

Min-Max Normalization

Transform the data from measured units to a new interval from new_min_F to new_max_F for feature F:

$$v' = \frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$$

where v is the current value of feature F.

Min-Max Normalization: Example

Suppose that the minimum and maximum values for the feature income are $120,000 and $98,000, respectively. We would like to map income to the range [0.0,1.0]. By min-max normalization, a value of $73,600 for income is transformed to:

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0.0) + 0 = 0.716$$

Z-score (zero-mean) Normalization

Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, v, of A is normalized to v' by computing:

$$v' = \frac{v - F}{\sigma_F}$$

where F and σ_F are the mean and standard deviation of feature F, respectively.

Z-score Normalization

- The normalized value of X_i is calculated as:

$$Z_i = \frac{X_i - \bar{X}}{s}$$

$$\mathbf{y} = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix}$$

$$\begin{aligned} s &= \sqrt{\frac{(35-51)^2 + (36-51)^2 + (46-51)^2 + (68-51)^2 + (70-51)^2}{5-1}} \\ &= \frac{1}{2}\sqrt{(-16)^2 + -15^2 + (-5)^2 + 17^2 + 19^2} \\ &= 17\,. \end{aligned}$$

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}$$

vs. Min-Max Normalization:

[0, 1/35, 11/35, 33/35, 1] = [0, 0.0286, 0.3143, 0.9429, 1.0]

Decimal Scaling Normalization

Transform the data by moving the decimal points of values of feature F. The number of decimal points moved depends on the maximum absolute value of F. A value v of F is normalized to $v^{'}$ by computing :

$$v^{'} = \frac{v}{10},$$

where j is the smallest integer such that $Max\,(|v^{'}|) < 1$.

Decimal Scaling Normalization

- Suppose that the recorded values of F range from – 986 to 917. The maximum absolute value of F is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that −986 normalizes to −0.986 and 917 normalizes to 0.917.

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