PRINCIPLES OF DATA SCIENCE

JOHN P DICKERSON

Lecture #6 – 10/3/2018

CMSC641
Wednesdays
7:00pm – 9:30pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

ANNOUNCEMENTS

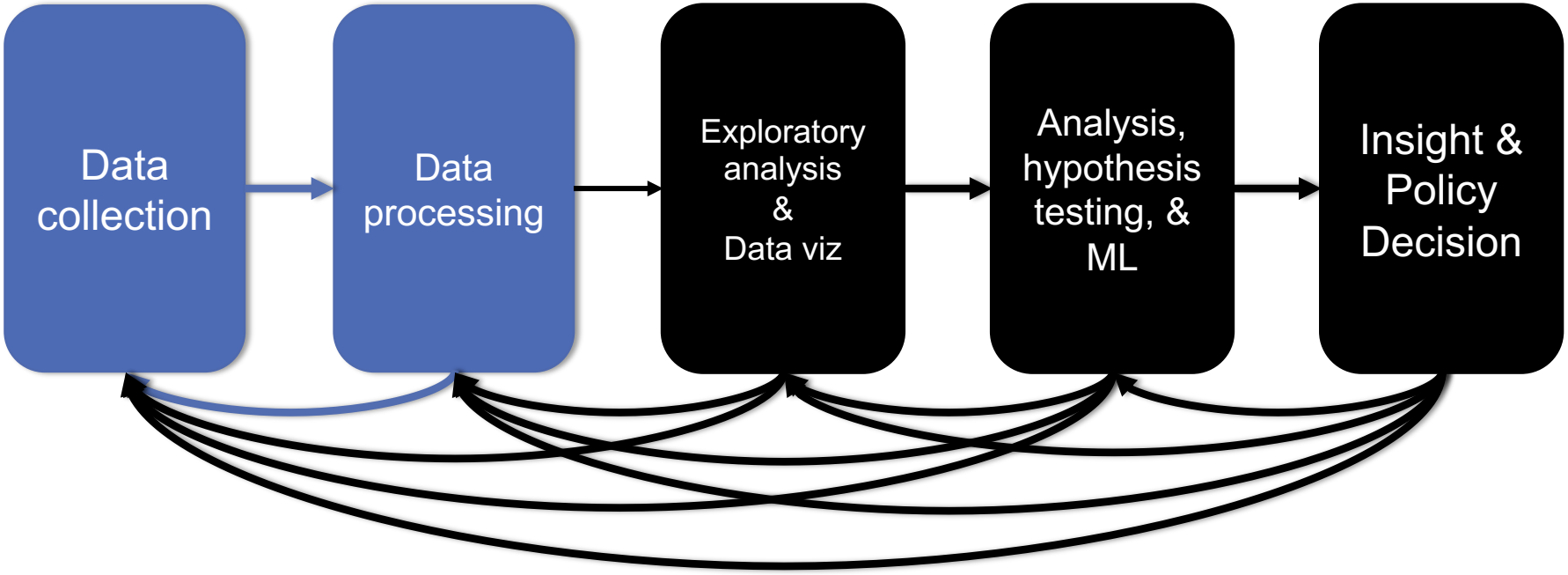
Mini-Project #1 was due today!

- It is linked to from ELMS; also available at: https://github.com/umddb/cmsc641-fall2018/tree/master/project1
- Deliverable is a .ipynb file submitted to ELMS

Mini-Project #2 is out!

- It is linked to from ELMS; also available at: https://github.com/umddb/cmsc641-fall2018/tree/master/project2
- Deliverable is a .ipynb file submitted to ELMS
- Due Wednesday, October 24th

REST OF TODAY'S LECTURE



Continue with the general topic of data wrangling and cleaning

Many slides from Amol Deshpande (UMD)

OVERVIEW

Goal: get data into a structured form suitable for analysis

- Variously called: data preparation, data munging, data curation
- Also often called ETL (Extract-Transform-Load) process

Often the step where majority of time (80-90%) is spent

Key steps:

- Scraping: extracting information from sources, e.g., webpages, spreadsheets
- Data transformation: to get it into the right structure
- Data integration: combine information from multiple sources
- Information extraction: extracting structured information from unstructured/text sources
- Data cleaning: remove inconsistencies/errors

OVERVIEW

Goal: get data into a structured form suitable for analysis

- Variously called: data preparation, data munging, data curation
- Also often called ETL (Extract-Transform-Load) process

Often the step where majority of time (80-90%) is spent

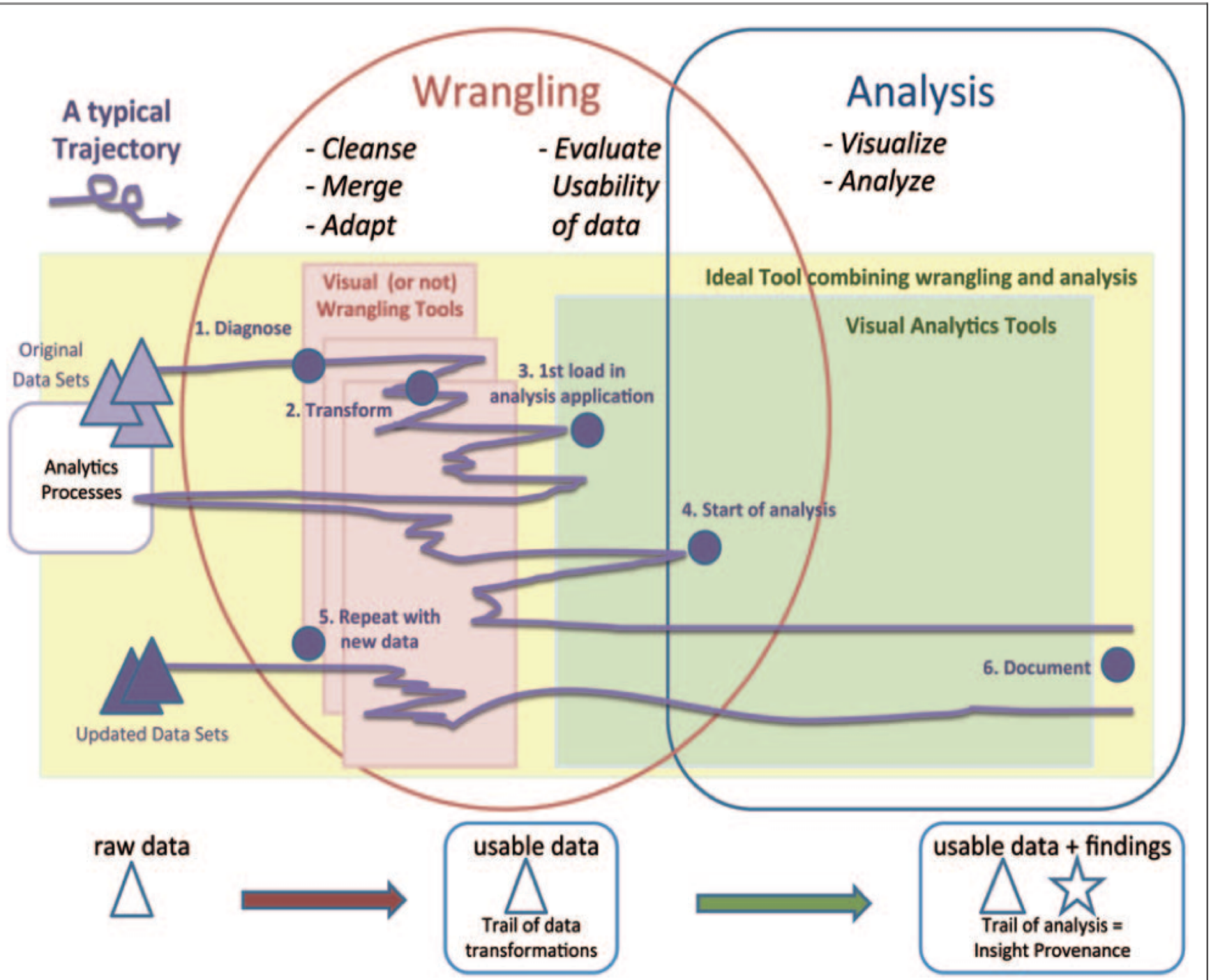
Key steps:

Already covered

- Scraping: extracting information from sources, e.g., webpages, spreadsheets
- Data transformation: to get it into the right structure
- Information extraction: extracting structured information from unstructured/text sources

In a few classes

- **Data integration: combine information from multiple sources**
- **Data cleaning: remove inconsistencies/errors**


A typical Trajectory
Wrangling
- Cleanse
- Merge
- Adapt
- Evaluate Usability of data
Analysis
- Visualize
- Analyze
Ideal Tool combining wrangling and analysis
Visual (or not) Wrangling Tools
Visual Analytics Tools
Original Data Sets
1. Diagnose
2. Transform
3. 1st load in analysis application
Analytics Processes
4. Start of analysis
5. Repeat with new data
6. Document
Updated Data Sets
raw data
usable data
Trail of data transformations
usable data + findings
Trail of analysis = Insight Provenance

OVERVIEW

Many of the problems are not easy to formalize, and have seen little work

- E.g., Cleaning
- Others aspects of integration, e.g., schema mapping, have been studied in depth

A mish-mash of tools typically used

- Visual (e.g., Trifacta), or not (UNIX grep/sed/awk, Pandas)
- Ad hoc programs for cleaning data, depending on the exact type of errors
- Different types of transformation tools
- Visualization and exploratory data analysis to understand and remove outliers/noise
- Several tools for setting up the actual pipelines, assuming the individual steps are setup (e.g., Talend, AWS Glue)

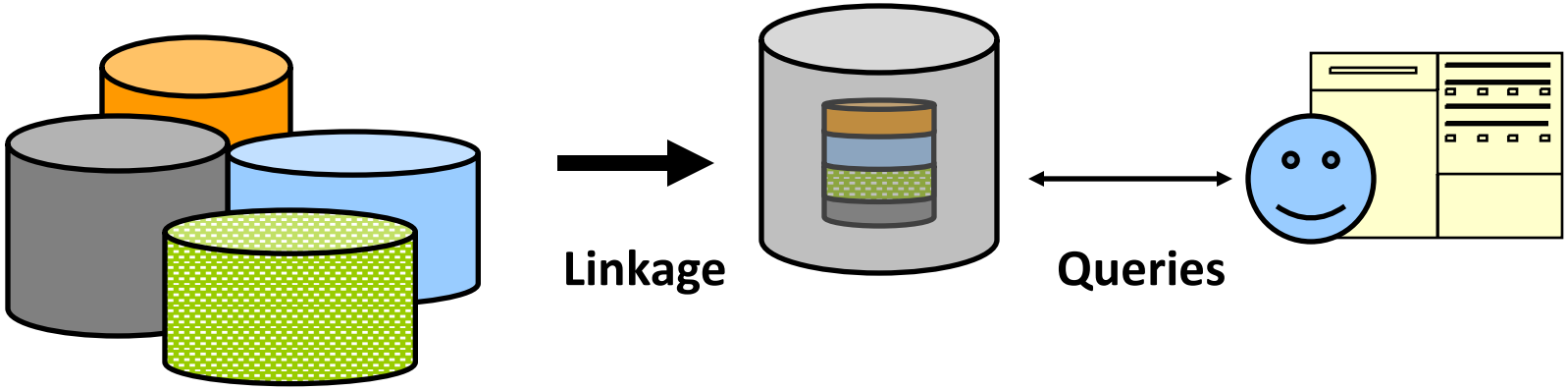
OUTLINE

- Data Integration
- Data Quality Issues
- Data Cleaning
- Entity Resolution

OUTLINE

- Data Integration
- Data Quality Issues
- Data Cleaning
- Entity Resolution

DATA INTEGRATION



- **Discovering** information sources (e.g. deep web modeling, schema learning, …)
- **Gathering** data (e.g., wrapper learning & information extraction, federated search, …)

- **Cleaning** data (e.g., de-duping and **linking records**) to form a single [virtual] database

- **Querying** integrated information sources (e.g. queries to views, execution of web-based queries, …)
- **Data mining & analyzing** integrated information (e.g., collaborative filtering/classification learning using extracted data, …)

DATA INTEGRATION

Goal: Combine data residing in different sources and provide users with a unified view of these data for querying or analysis

- Each data source has its own schema called **local schemas** (much work assumes relational schemas, but some work on XML as well)
- The unified schema is often called **mediated schema** or **global schema**

Two different setups:

1. Bring the data together into a single repository (often called data warehousing)
2. Keep the data where it is, and send queries back and forth

1. DATA WAREHOUSING

From Data Cleaning: Problems and Current Approaches

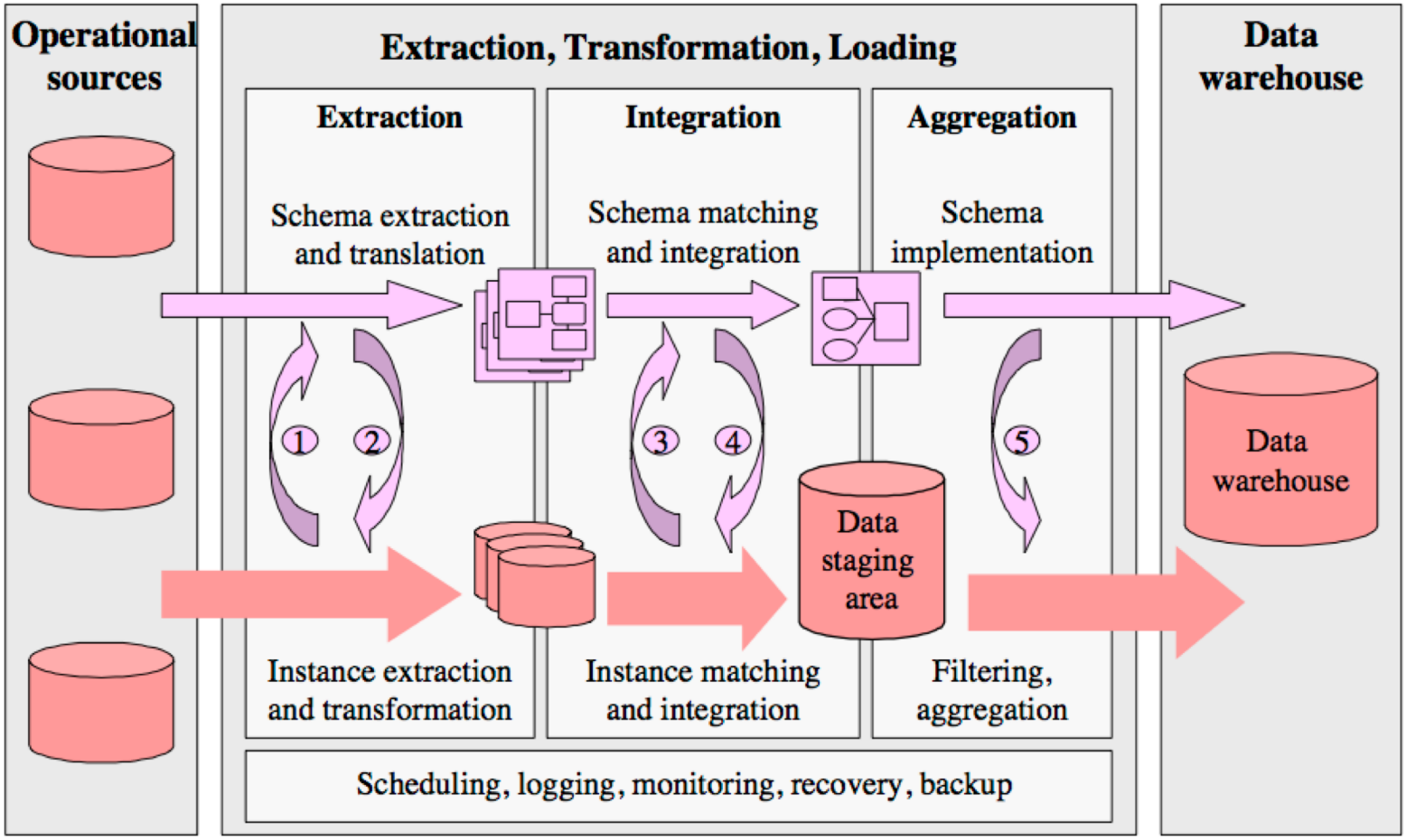
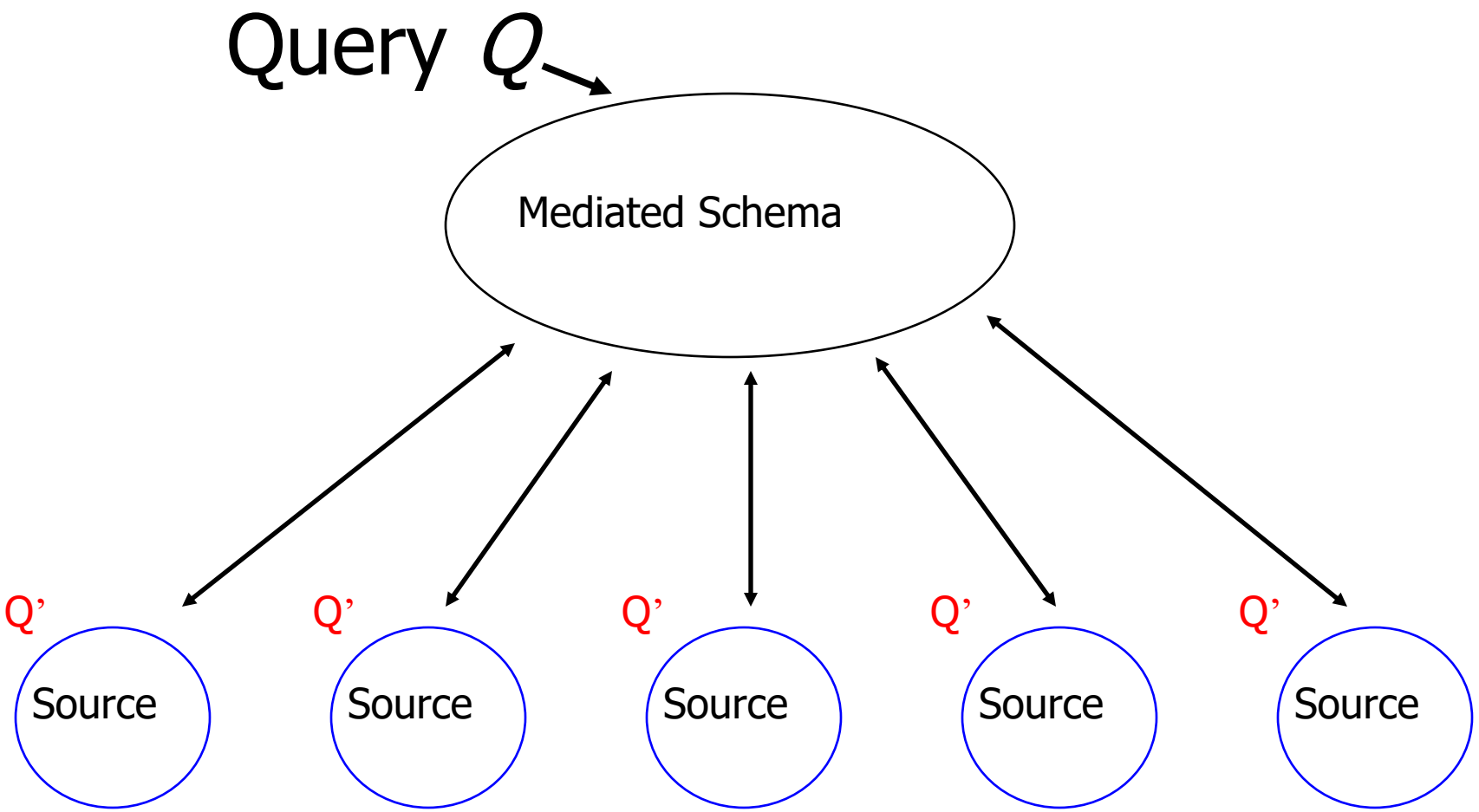


Figure 1. Steps of building a data warehouse: the ETL process

2. IN-PLACE INTEGRATION



DATA INTEGRATION

Two different setups:

1. Bring the data together into a single repository (often called data warehousing)
   - Relatively easier problem - only need one-way-mappings
   - Query performance predictable and under your control

2. Keep the data where it is, and send queries back and forth
   - Need two-way mappings -- a query on the mediated schema needs to be translated into queries over data source schemas
   - Not as efficient and clean as data warehousing, but a better fit for dynamic data
   - Or when data warehousing is not feasible

DATA INTEGRATION: KEY CHALLENGES

Data extraction, reconciliation, and cleaning

- Get the data from each source in a structured form
- Often need to use wrappers to extract data from web sources
- May need to define a schema

Schema alignment and mapping

- Decide on the best mediated schema
- Figure out mappings and matchings between the local schemas and the global schema

Answer queries over the global schema

- In the second scenario, need to figure out how to map a query on global schema onto queries over local schemas
- Also need to decide which sources contain relevant data

Limitations in mechanisms for accessing sources

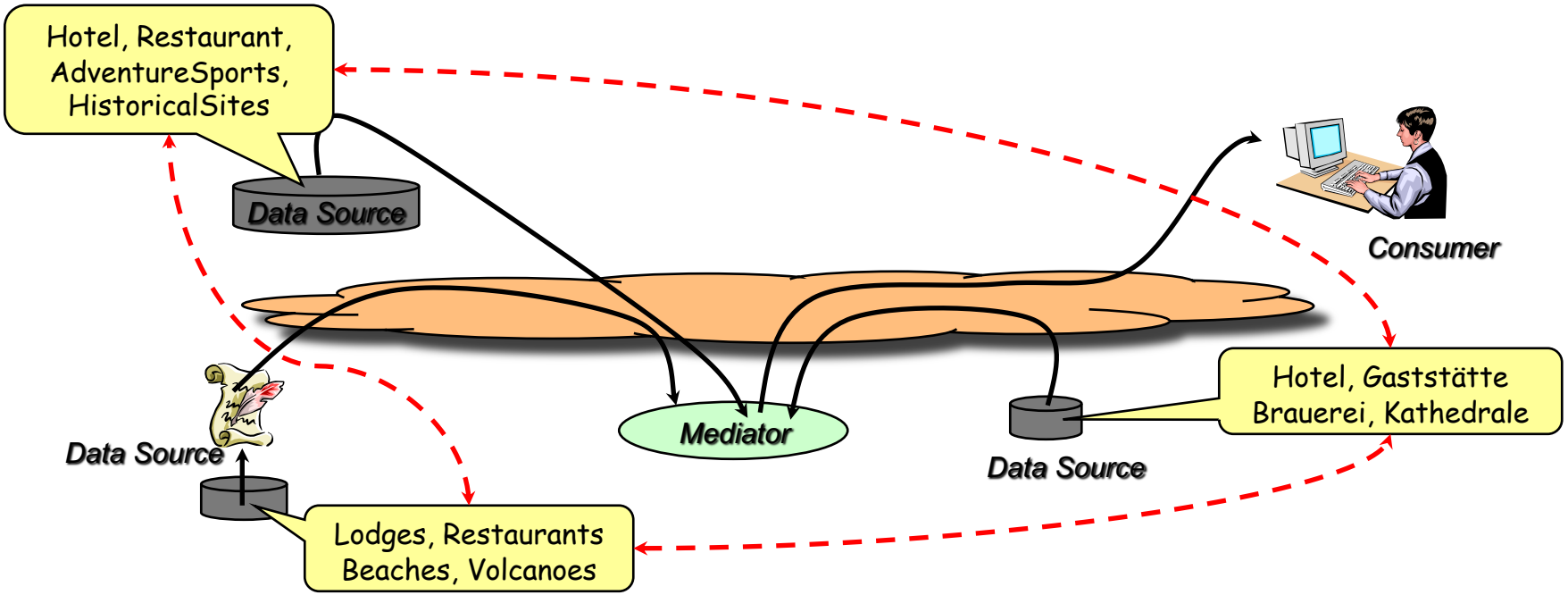
- Many sources have limits on how you can access them
- Limits on the number of queries you can issues (say 100 per min)
- Limits on the types of queries (e.g., must enter a zipcode to get information from a web source)

Flashbacks to Project 1 …

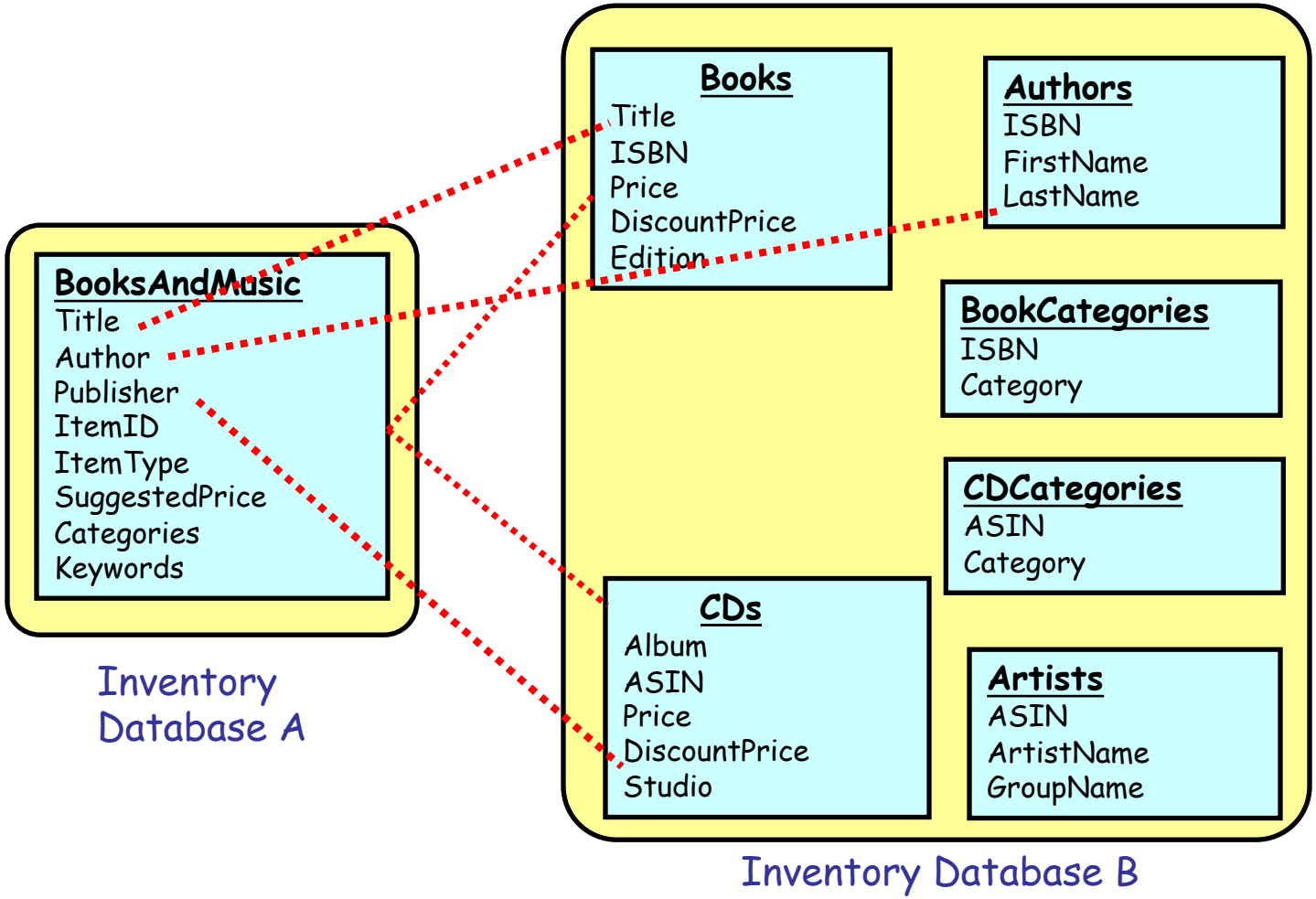
SCHEMA MATCHING OR ALIGNMENT

Goal: Identify corresponding elements in two schemas

- As a first step toward constructing a global schema
- Schema heterogeneity is a key roadblock
  - Different data sources speak their own schema



SCHEMA MATCHING OR ALIGNMENT



SUMMARY

- **Data integration continues to be a very active area in research and increasingly industry**
- **Solutions still somewhat ad hoc and manual, although tools beginning to emerge**
- **Need to minimize the time needed to integrate a new data source**
  - Crucial opportunities may be lost otherwise
  - Can take weeks to do it properly
- **Dealing with changes to the data sources a major headache**
  - Especially for data sources not under your control

OUTLINE

- Data Integration
- Data Quality Issues
- Data Cleaning
- Entity Resolution

DATA QUALITY PROBLEMS

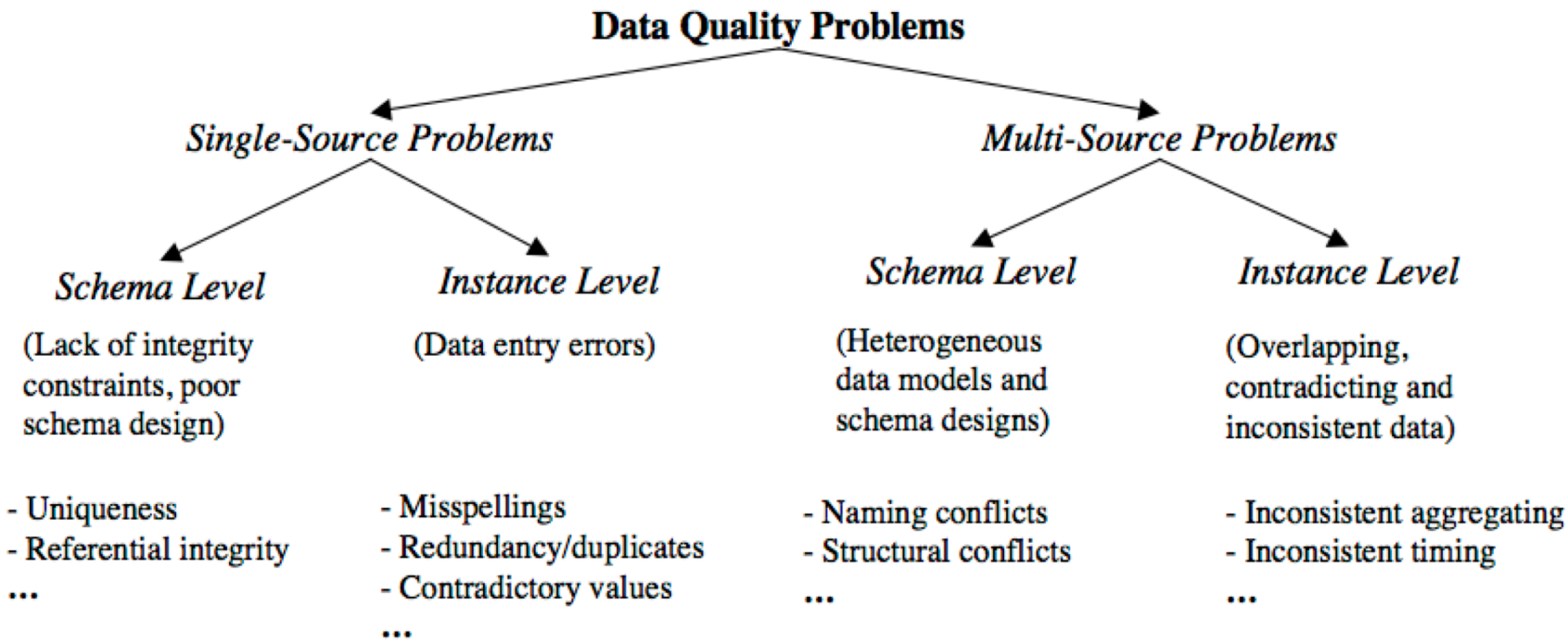


Figure 2. Classification of data quality problems in data sources

From Data Cleaning: Problems and Current Approaches

SINGLE-SOURCE PROBLEMS

Depends largely on the source

Databases can enforce constraints, whereas data extracted from files or spreadsheets, or scraped from webpages is much more messy

Types of problems:

- Ill-formatted data, especially from webpages or files or spreadsheets
- Missing or illegal values, Misspellings, Use of wrong fields, Extraction issues (not easy to separate out different fields)
- Duplicated records, Contradicting Information, Referential Integrity Violations
- Unclear default values (e.g., data entry software needs something)
- Evolving schemas or classification schemes (for categorical attributes)
- Outliers

DATA QUALITY PROBLEMS

Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

Table 2. Examples for single-source problems at instance level

MULTI-SOURCE PROBLEMS

Different sources are developed separately, and maintained by different people

Issue 1: Mapping information across sources (schema mapping/transformation)

- Naming conflicts: same name used for different objects
- Structural conflicts: different representations across sources
- We will cover this later

Issue 2: Entity Resolution: Matching entities across sources

Issue 3: Data quality issues

- Contradicting information, Mismatched information, etc.

OUTLINE

- **Data Integration**
- **Data Quality Issues**
- **Data Cleaning**
  - Outlier Detection
  - Constraint-based Cleaning
  - Entity Resolution

UNIVARIATE OUTLIERS

A set of values can be characterized by metrics such as center (e.g., mean), dispersion (e.g., standard deviation), and skew

Can be used to identify outliers

- Must watch out for "masking": one extreme outlier may alter the metrics sufficiently to mask other outliers
- Should use **robust statistics**: considers effect of corrupted data values on distributions – we will talk about this in depth later
- Robust center metrics: median, k% trimmed mean (discard lowest and highest k% values)
- Robust dispersion:
  - Median Absolute Deviation (MAD): median distance of all the values from the median value

A reasonable approach to find outliers: any data points 1.4826x MAD away from median

- The above assumes that data follows a **normal** distribution
- May need to eyeball the data (e.g., plot a histogram) to decide if this is true

UNIVARIATE OUTLIERS

Wikipedia Article on Outliers lists several other normality-based tests for outliers

If data appears to be not normally distributed:

- Distance-based methods: look for data points that do not have many neighbors
- Density-based methods:
  - Define *density* to be average distance to *k* nearest neighbors
  - *Relative density* = density of node/average density of its neighbors
  - Use relative density to decide if a node is an outlier

Most of these techniques start breaking down as the dimensionality of the data increases

- *Curse of dimensionality*
- Can project data into lower-dimensional space and look for outliers there
  - Not as straightforward

OTHER OUTLIERS

Timeseries outliers

- Often the data is in the form of a timeseries
- Can use the historical values/patterns in the data to flag outliers
- Rich literature on *forecasting* in timeseries data

Frequency-based outliers

- An item is considered a "heavy hitter" if it is much more frequent than other items
- In relational tables, can be found using a simple *groupby-count*
- Often the volume of data may be too much (e.g., internet routers)
    - Approximation techniques often used
    - To be discussed sometime later in the class

Things generally not as straightforward with other types of data

- Outlier detection continues to be a major research area

WRAP-UP

Data wrangling/cleaning are a key component of data science pipeline

Still largely ad hoc although much tooling in recent years

Specifically, we covered:

- Schema mapping and matching
- Outliers

Next up:

- Constraint-based Cleaning
- Entity Resolution/Record Linkage/Data Matching

DATA CLEANING: OUTLIER RESOLUTION

From: Entity Resolution Tutorial

Identify different manifestations of the same real world object

- Also called: identity reconciliation, record linkage, deduplication, fuzzy matching, Object consolidation, Coreference resolution, and several others

Motivating examples: ??????????????

- Postal addresses
- Entity recognition in NLP/Information Extraction
- Identifying companies in financial records
- Comparison shopping
- Author disambiguation in citation data
- Connecting up accounts on online networks
- Crime/Fraud Detection
- Census
- ...

DATA CLEANING: OUTLIER RESOLUTION

Important to correctly identify references

- Often actions taken based on extracted data
- Cleaning up data by entity resolution can show structure that may not be apparent before

Challenges

- Such data is naturally ambiguous (e.g., names, postal addresses)
- Abbreviations/data truncation
- Data entry errors, Missing values, Data formatting issues complicate the problem
- Heterogeneous data from many diverse sources

No magic bullet here !!

- Approaches fairly domain-specific
- Be prepared to do a fair amount of manual work

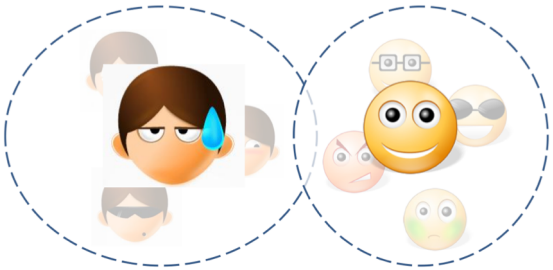
ENTITY RESOLUTION: THREE SLIGHTLY DIFFERENT PROBLEMS

Setup:

- Real world: there are entities (people, addresses, businesses)
- We have a large collection of noisy, ambiguous "references" to those entities (also called "mentions")
- Somewhat different techniques, but a lot of similarities

Deduplication

- Cluster records/mentions that correspond to the same entity
- Choose/construct a cluster representative
    - This is in itself a non-trivial task (e.g., averaging may work for numerical attributes, but what about string attributes?)



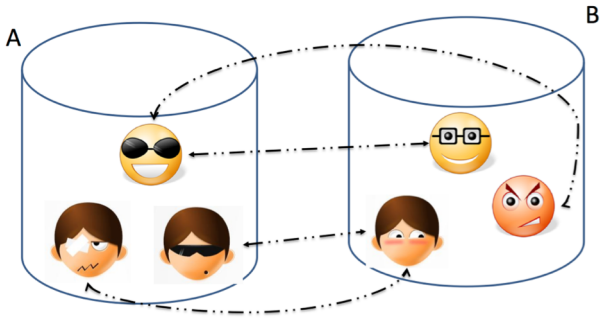
ENTITY RESOLUTION: THREE SLIGHTLY DIFFERENT PROBLEMS

Setup:

- Real world: there are entities (people, addresses, businesses)
- We have a large collection of noisy, ambiguous "references" to those entities (also called "mentions")
- Somewhat different techniques, but a lot of similarities

Record Linkage

- Match records across two different databases (e.g., two social networks, or financial records w/ campaign donations)
- Typically assume that the two databases are fairly clean



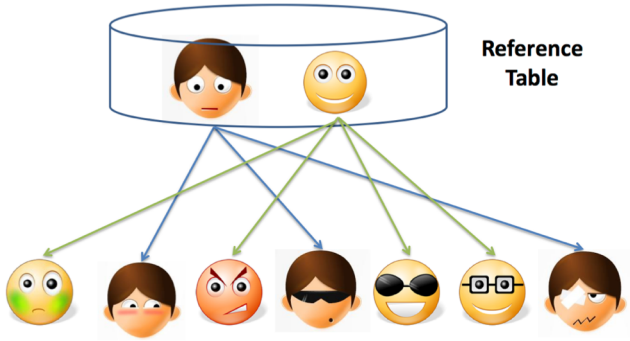
ENTITY RESOLUTION: THREE SLIGHTLY DIFFERENT PROBLEMS

Setup:

- Real world: there are entities (people, addresses, businesses)
- We have a large collection of noisy, ambiguous "references" to those entities (also called "mentions")
- Somewhat different techniques, but a lot of similarities

Reference Matching

- Match "references" to clean records in a reference table
- Commonly comes up in "entity recognition" (e.g., matching newspaper article mentions to names of people)



ENTITY RESOLUTION: DATA MATCHING

Comprehensive treatment: Data Matching; P. Christen; 2012 (Springer Books -- not available for free)

One of the key issues is finding similarities between two references

- What similarity function to use?

Edit Distance Functions

- Levenstein: min number of changes to go from one reference to another
  - A change is defined to be: a single character insertion or deletion or substitution
  - May add transposition
- Many adjustments to the basic idea proposed (e.g., higher weights to changes at the start)
- Not cheap to compute, especially for millions of pairs

Set Similarity

- Some function of intersection size and union size
- E.g., Jaccard distance = size of intersection/size of union
- Much faster to compute

Vector Similarity

- Cosine similarity – we'll talk about this much more in NLP lectures

ENTITY RESOLUTION: DATA MATCHING

Q-Grams

- Find all length-q substrings in each string
- Use set/vector similarity on the resulting set

Several approaches that combine the above (especially q-grams and edit distance, e.g., Jaro-Winkler)

Soundex: Phonetic Similarity Metric

- Homophones should be encoded to the same representation so spelling errors can be handled
- Robert and Rupert get assigned the same code (R163), but Rubin yields R150

May need to use Translation Tables

- To handle abbreviations, nicknames, other synonyms

Different types of data requires more domain-specific functions

- E.g., geographical locations, postal addresses
- Also much work on computing distances between XML documents etc.

ENTITY RESOLUTION: ALGORITHMS

Simple threshold method

- If the distance below some number, the two references are assumed to be equal
- May review borderline matches manually

Can be generalized to rule-based:

- Example from Christen, 2012

$$(s(\text{GivenName})[r_i, r_j] \geq 0.9) \wedge (s(\text{Surname})[r_i, r_j] = 1.0) \wedge (s(\text{BMonth})[r_i, r_j] = 1.0) \wedge (s(\text{BYear})[r_i, r_j] = 1.0) \Rightarrow [r_i, r_j] \rightarrow \text{Match}$$

$$(s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \wedge (s(\text{BDay})[r_i, r_j] = 1.0) \wedge s(\text{BMonth})[r_i, r_j] = 1.0) \wedge (s(\text{BYear})[r_i, r_j] = 1.0) \Rightarrow [r_i, r_j] \rightarrow \text{Match}$$

$$(s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \wedge (s(\text{StrName})[r_i, r_j] \geq 0.8) \wedge (s(\text{Suburb})[r_i, r_j] \geq 0.8) \Rightarrow [r_i, r_j] \rightarrow \text{Match}$$

$$(s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \wedge (s(\text{BDay})[r_i, r_j] \leq 0.5) \wedge (s(\text{BMonth})[r_i, r_j] \leq 0.5) \wedge (s(\text{BYear})[r_i, r_j] \leq 0.5) \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match}$$

$$(s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \wedge (s(\text{StrName})[r_i, r_j] \leq 0.6) \wedge (s(\text{Suburb})[r_i, r_j] \leq 0.6) \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match}$$

ENTITY RESOLUTION: ALGORITHMS

May want to give more weight to matches involving rarer words

- More naturally applicable to record linkage problem
- If two records match on a rare name like "Machanavajjhala", they are likely to be a match
- Can formalize this as "probabilistic record linkage"

Constraints: May need to be satisfied, but can also be used to find matches

- Often have constraints on the matching possibilities
- Transitivity: M1 and M2 match, and M2 and M3 match, and M1 and M3 must match
- Exclusivity: M1 and M2 match --> M3 cannot match with M2
- Other types of constraints:
    - E.g., if two papers match, their venues must match

ENTITY RESOLUTION: ALGORITHMS

Clustering-based ER Techniques:

- Deduplication is basically a clustering problem
- Can use clustering algorithms for this purpose
- But most clusters are very small (in fact of size = 1)
- Some clustering algorithms are better suited for this, especially Agglomerative Clustering
    - Unlikely K-Means would work here

ENTITY RESOLUTION: ALGORITHMS

Crowdsourcing

- Humans are often better at this task
- Can use one of the crowdsourcing mechanisms (e.g., Mechanical Turk) for getting human input on the difficult pairs
- Quite heavily used commercially (e.g., to disambiguate products, restaurants, etc.)

ENTITY RESOLUTION: SCALING TO BIG DATA

One immediate problem

- There are $O(N^2)$ possible matches
- Must reduce the search space

Use some easy-to-evaluate criterion to restrict the pairs considered further

- May lead to false negative (i.e., missed matches) depending on how noisy the data is

Much work on this problem as well, but domain-specific knowledge likely to be more useful in practice

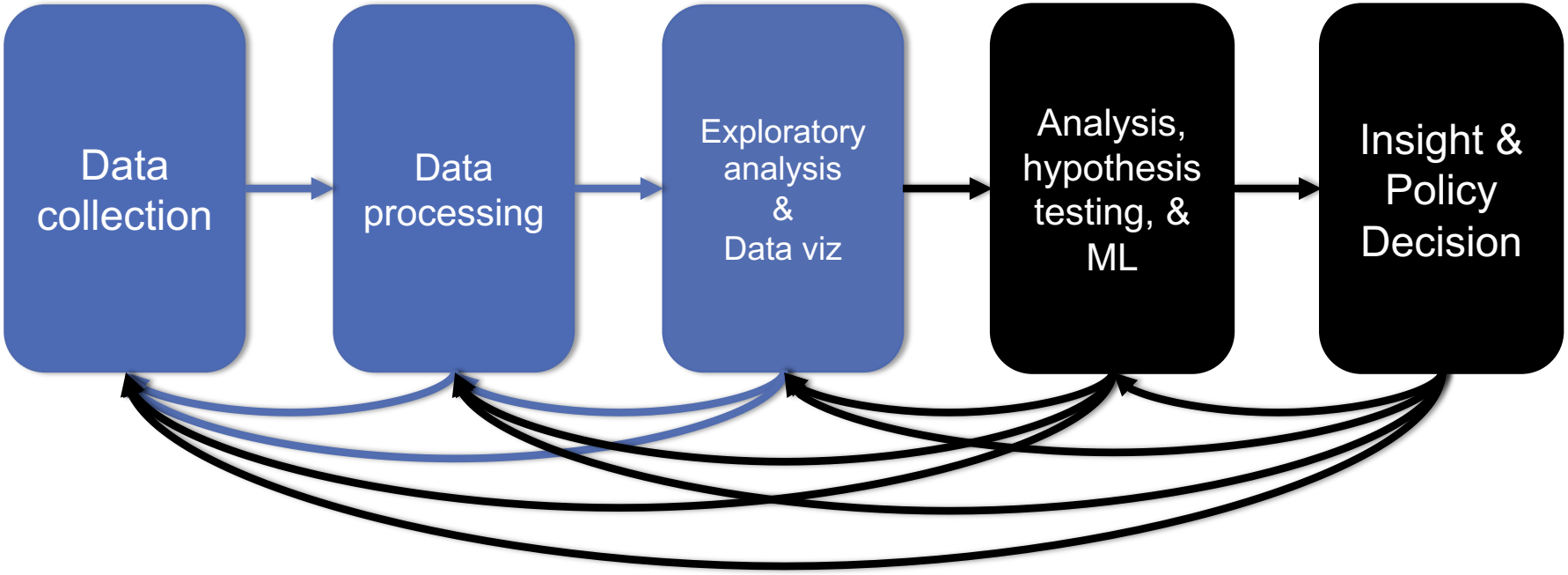
One useful technique to know: min-hash signatures

- Can quickly find potentially overlapping sets
- Turns up to be very useful in many domains (beyond ER)

NEXT UP:

SUMMARY STATISTICS &VISUALIZATION

TODAY'S LECTURE



EXPLORATORY DATA ANALYSIS

Seen so far:

- Manipulations that prepare datasets into tidy form
- Join tables and compute summaries
- Form relationships between different entities

EDA is the last step before Big Time Statistics and ML™:

- Want to quickly "get a feel" for the data through summary statistics, visualization, et cetera
- Spot nuances like skew, how distributed the data is, trends, how pairs of variables interact, problems
- Suggests which Stats/ML assumptions to make and approaches to take

LAST WEEK'S LESSON

Having a really big sample does not assure you of an accurate result.

It may assure you of a really solid, really bad (inaccurate) result.

Not all randomness is create equal when it comes to random sampling of a population:

- **Ask why data are missing! MCAR, MAR, MNAR.**
- **Ask how the data were collected.**

TODAY'S LESSON: SUMMARY STATISTICS

Part of descriptive statistics, used to summarize data:

- Convey lots of information with extreme simplicity

Descriptive statistics for a variable:

- Measures of location: mean, median, mode
- Measure of dispersion: variance, standard deviation

Measuring correlation of two variables:

- Understanding correlation
- Measuring correlation
- Scatter plots and regression

Thanks to William Green (NYU Stern IOMS) and Hector Corrada Bravo (UMD)

MEASURES OF LOCATION

These are 30 hours of average defect data on sets of circuit boards. Roughly what is the typical value?

1.45	1.65	1.50	2.25	1.65	1.60	2.30	2.20	2.70	1.70
2.35	1.70	1.90	1.45	1.40	2.60	2.05	1.70	1.05	2.35
1.90	1.55	1.95	1.60	2.05	2.05	1.70	2.30	1.30	2.35

Location and central tendency

- There exists a distribution of values
- We are interested in the "center" of the distribution

Two measures are the sample mean and the sample median

They look similar, and measure the same thing

They differ systematically (and predictably) when the data are not symmetric

[WG]

THE MEAN OF AGGREGATE DATA

State	Listing	IncomePC	State	Listing	IncomePC	State	Listing	IncomePC
Hawaii	896800	24057	Rhode Island	432534	22251	Texas	266388	19857
California	713864	22493	Delaware	420845	22828	Mississippi	255774	15838
New York	668578	25999	Oregon	417551	20419	Tennessee	255064	19482
Connecticut	654859	29402	Idaho	415885	18231	Wisconsin	243006	21019
Dist.Columbia	577921	31136	Illinois	377683	23784	Michigan	241107	22333
Nevada	549187	24023	New Hampshire	361691	23434	Missouri	221773	20717
New Jersey	529201	23038	New Mexico	358369	17106	South Dakota	220708	19577
Massachusetts	521769	25616	Vermont	346469	20224	West Virginia	219275	17208
Wyoming	499674	20436	South Carolina	340066	17695	Arkansas	217659	16898
Maryland	480578	24933	North Carolina	330432	19669	Ohio	209189	20928
Utah	475060	17043	Georgia	326699	20251	Kentucky	208391	17807
Colorado	467979	22333	Alaska	324774	23788	Oklahoma	203926	17744
Arizona	448791	19001	Minnesota	306009	22453	Kansas	201389	20896
Florida	447698	21677	Maine	299796	19663	Indiana	200683	20378
Montana	446584	17865	Pennsylvania	295133	22324	Iowa	184999	20265
Virginia	443618	22594	Louisiana	280631	17651	North Dakota	173977	18546
Washington	440542	22610	Alabama	269135	18010	Nebraska	164326	20488

Average list price:
1/51 ($898,800 + $713,864 + … + $164,326) = $369,687

[WG]

AVERAGING AVERAGES?

Hawaii's average listing	= $896,800
Hawaii's population	= 1,275,194
Illinois' average listing	= $377,683
Illinois' population	= 12,763,371

Illinois and Hawaii each get an equal weight of 1/51 = .019607 when the mean is computed.

Looks like Hawaii is getting too much influence …

[WG]

WEIGHTED AVERAGE

$$\text{Simple average} = \overline{\text{Listing}} = \sum_{\text{States}} \text{Weight}_{\text{State}}\ \text{Listing}_{\text{State}}$$

$$\text{Weight} = \frac{1}{51} = .019607$$

Illinois is 10 times as big as Hawaii. Suppose we use weights that are in proportion to the state's population. (The weights sum to 1.0.)

$\text{Weight}_{\text{State}}$ varies from .001717 for Wyoming to .121899 for California

New average is $409,234 compared to $369,687 without weights, an error of 11%

Sometimes an unequal weighting of the observations is necessary

State population data: http://www.factmonster.com/ipka/A0004986.html

AVERAGES & TIME SERIES

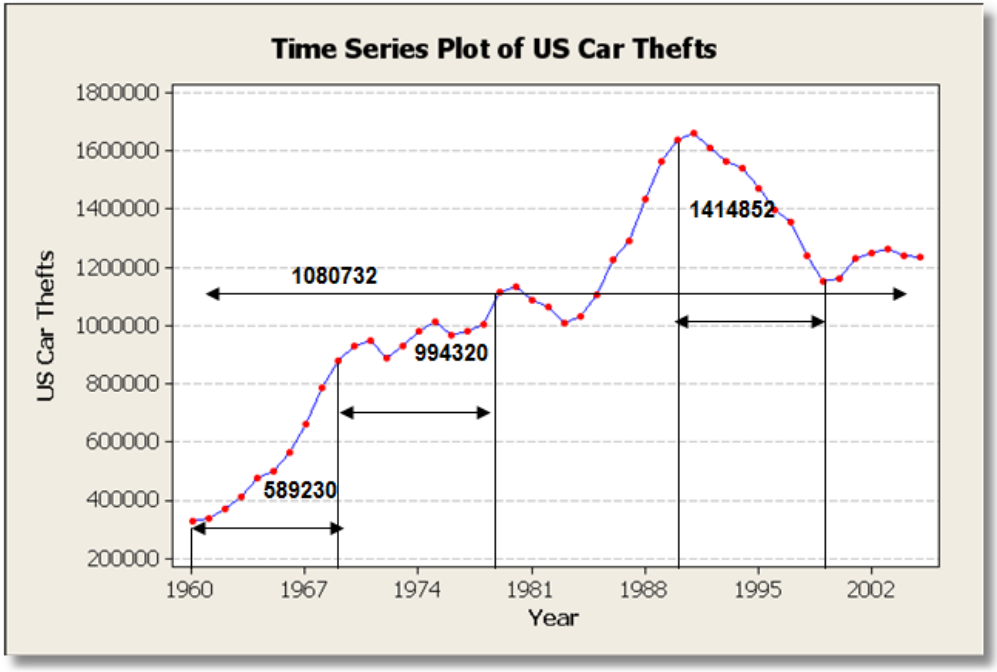
Averaging trending time series is usually not helpful

Mean changes completely depending on time interval

What about periodic time series data ??????????

Ask yourself:

- **Does the mean over the entire observation period mean anything?**
- **Does it estimate anything meaningful?**



THE SAMPLE MEDIAN

Median:

- Sort the data
- Take the middle point*

Odd number:

- Central observation: Med[1,2,4,6,8,9,17]

Even number:

- Midpoint between the two central observations
  Med[1,2,4,6,8,9,14,17] = (6+8)/2=7

* There are faster ways, e.g., to find the median in linear time!

WHAT IS THE CENTER?

The mean and median measure the central tendency of data

Generally, the center of of a dataset is a point in its range that is close to the data.

Close? Need a distance metric between two points x and x_2

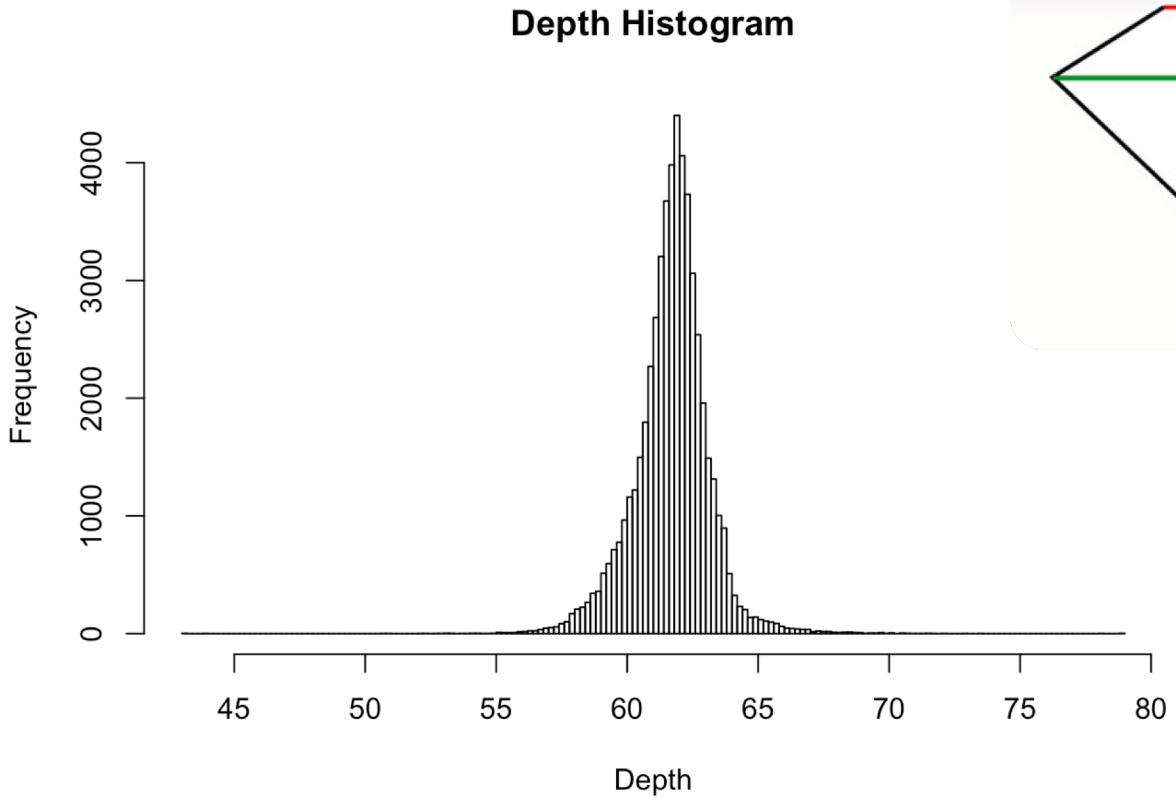
We've talked about some already!

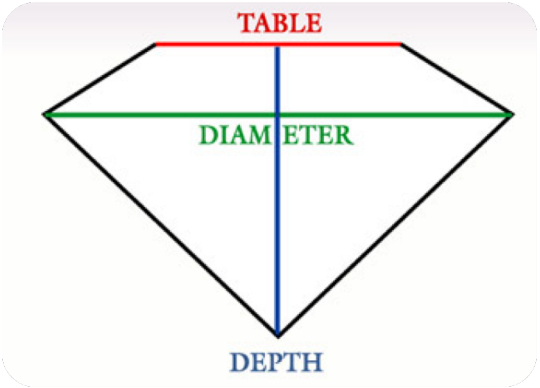
- Absolute deviation: $| x_1 - x_2 |$
- Squared deviation: $(x_1 - x_2)^2$

We'll define the center based on these metrics

DATASET FOR THIS PART

53,940 measurements of diamonds





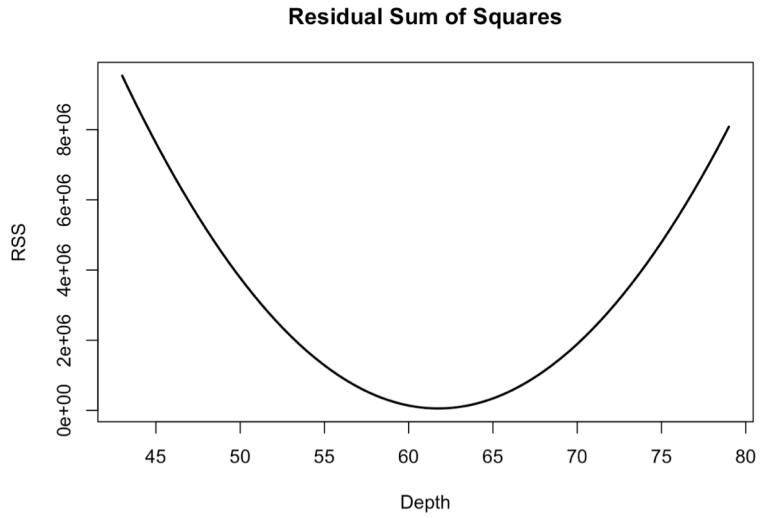
More info: https://en.wikipedia.org/wiki/Diamond_(gemstone)#Gemological_characteristics

THE MEAN REVISITED

Define a center point μ based on some function of the distance from each data point to that center point

- Residual sum of squares (RSS) for a point μ:

$$RSS(\mu) = \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$



So what should our estimate of the "center" of this dataset be, based on the RSS metric?
?????????????

THE MEAN REVISITED

Want the point μ that minimizes the RSS ??????????

- Find the derivative of RSS and set it to zero, solve for μ!

$$\frac{\partial}{\partial \mu} \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (x_i - \mu)^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} 2(x_i - \mu) \times (-1)$$

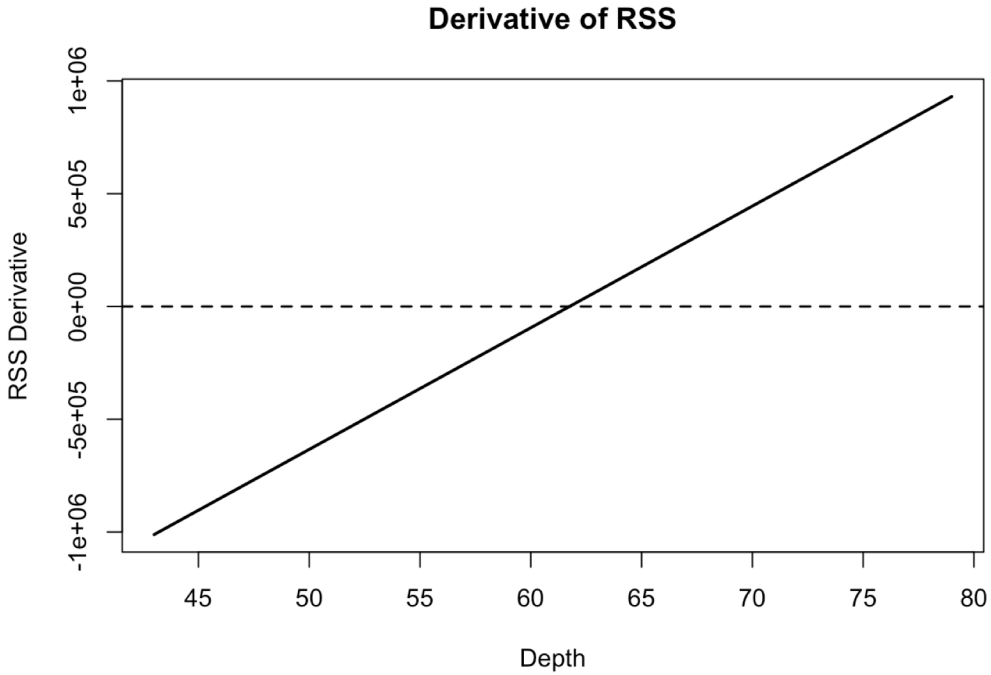
THE MEAN REVISITED

$$= \frac{1}{2} \sum_{i=1}^{n} 2(x_i - \mu) \times (-1)$$

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

$$= \sum_{i=1}^{n} \mu - \sum_{i=1}^{n} x_i$$

$$= n\mu - \sum_{i=1}^{n} x_i$$



THE MEAN REVISITED

Set the derivative to zero and solve for μ:

$$\frac{\partial}{\partial \mu} = 0$$

$$n\mu - \sum_{i=1}^{n} x_i = 0$$

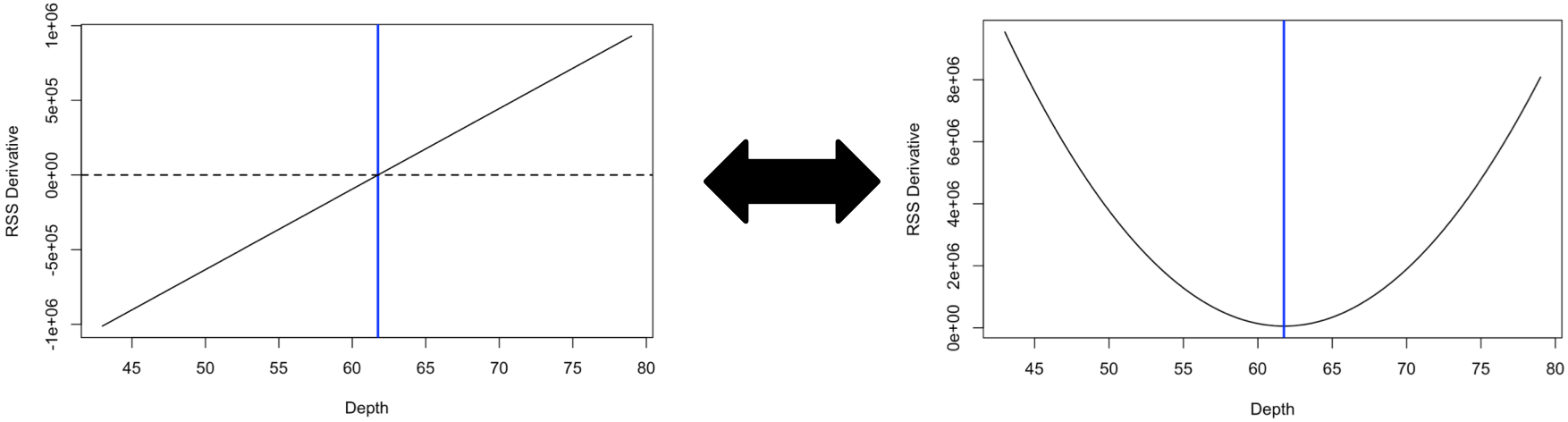
$$n\mu = \sum_{i=1}^{n} x_i$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The mean is the point μ that minimizes the RSS for a dataset.

THE MEAN REVISITED

What about a weighted average ???????



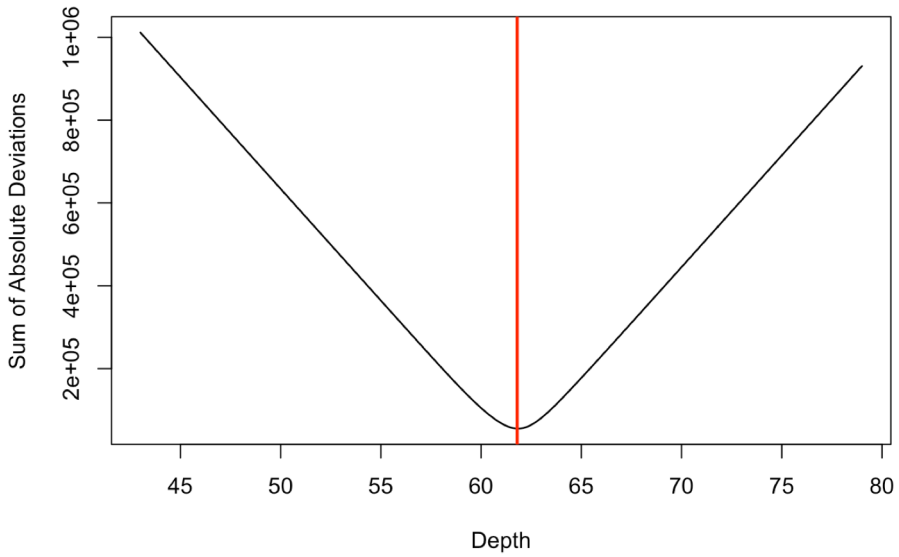
The mean is the point μ that minimizes the RSS for a dataset.

THE MEDIAN REVISITED

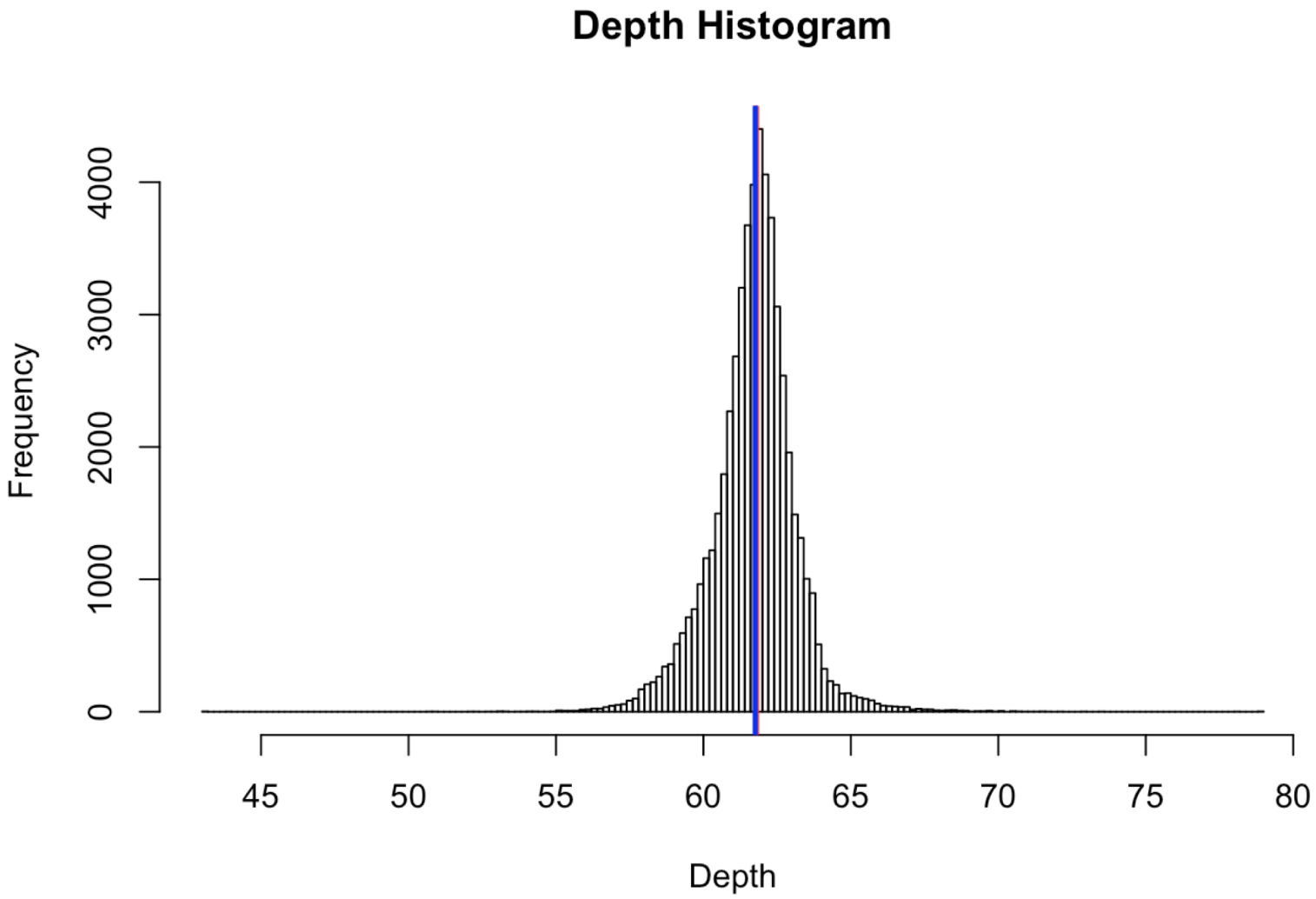
Define a center point *m* based on some function of the distance from each data point to that center point

- **The median *m* minimizes the sum of absolute differences:**

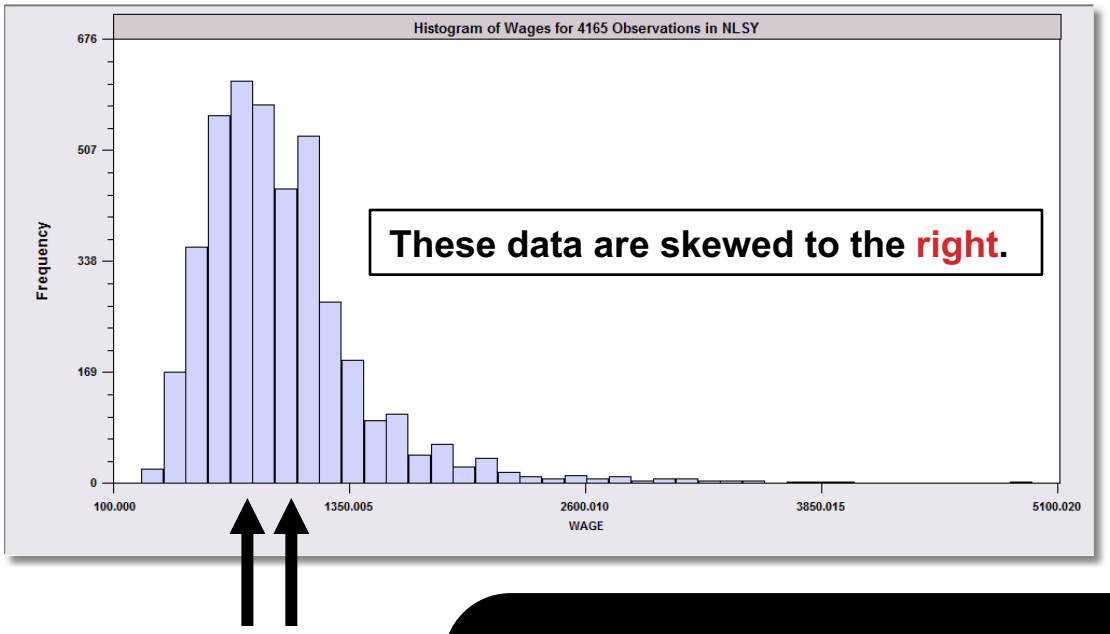
$$\sum_{i=1}^{n} |x_i - m|$$



MEAN != MEDIAN


Depth Histogram
Frequency
0
1000
2000
3000
4000
45
50
55
60
65
70
75
80
Depth

SKEWED DATA



Monthly Earnings
N = 595,
Median = 800
Mean = 883

The mean will exceed the median when the distribution is skewed to the right.

Skewness is in the direction of the long tail

SKEWNESS

Extreme observations distort means but not medians.

Outlying observations distort the mean:

- Med [1,2,4,6,8,9,17] = 6
- Mean[1,2,4,6,8,9,17] = 6.714
- Med [1,2,4,6,8,9,17000] = 6 (still)
- Mean[1,2,4,6,8,9,17000] = 2432.8 (!)

Typically occurs when there are some outlying observations, such as in cross sections of income or wealth and/or when the sample is not very large.

The New York Times

Business Day

DATAPOINTS

Income Gap Grows Wider (and Faster)

By ANNA BERNASEK
Published: August 31, 2013

INCOME inequality in the United States has been growing for decades, but the trend appears to have accelerated during the Obama administration. One measure of this is the relationship between median and average wages.

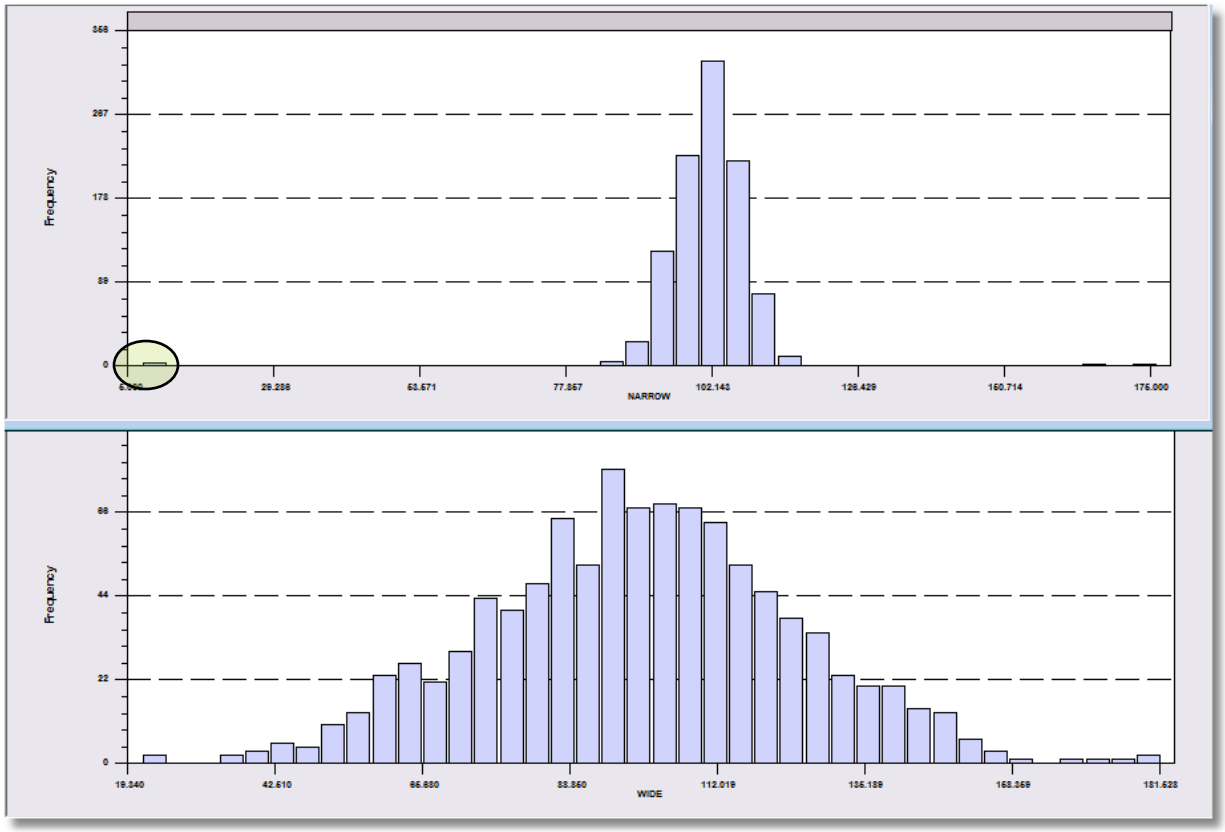
1.7%
Increase in **median** annual wage

3.9%
Increase in **average** annual wage
2009 through 2011

The median wage is straightforward: it's the midpoint of everyone's wages. Interpreting the average, though, can be tricky. If the income of a handful of people soars while everyone else's remains the same, the entire group's average may still rise substantially. So when average wages grow faster than the median, as happened from 2009 through 2011, it means that lower earners are falling further behind those at the top.

One way to see the acceleration in inequality is to look at the ratio of average to median annual wages. From 2001 through 2008, during the George W. Bush administration, that ratio grew at 0.28 percentage point per year. From 2009 through 2011, the latest year for which the data is available, the ratio increased 1.14 percentage points annually, or roughly four times faster.

MORE INFORMATION NEEDED!

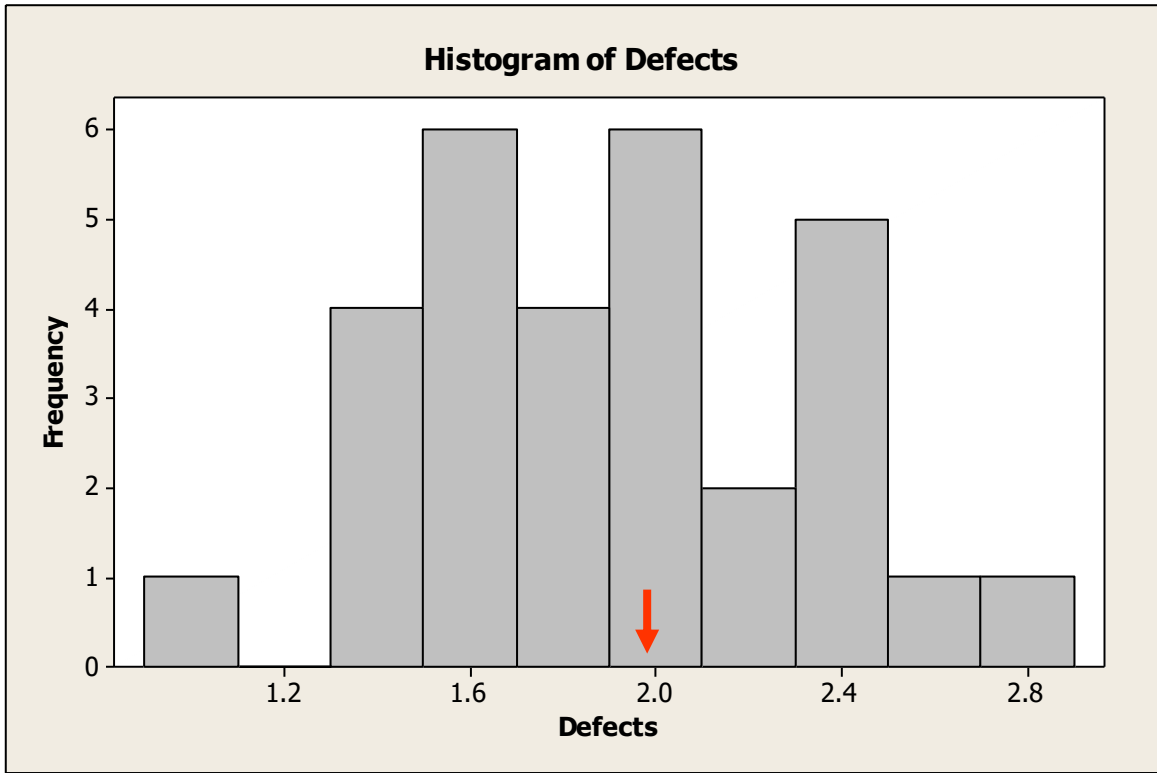


Both data sets have a mean of about 100.

DISPERSION OF THE OBSERVATIONS

30 hours of average defect data on sets of circuit boards.

1.45	1.65	1.50	2.25	1.65	1.60	2.30	2.20	2.70	1.70
2.35	1.70	1.90	1.45	1.40	2.60	2.05	1.70	1.05	2.35
1.90	1.55	1.95	1.60	2.05	2.05	1.70	2.30	1.30	2.35



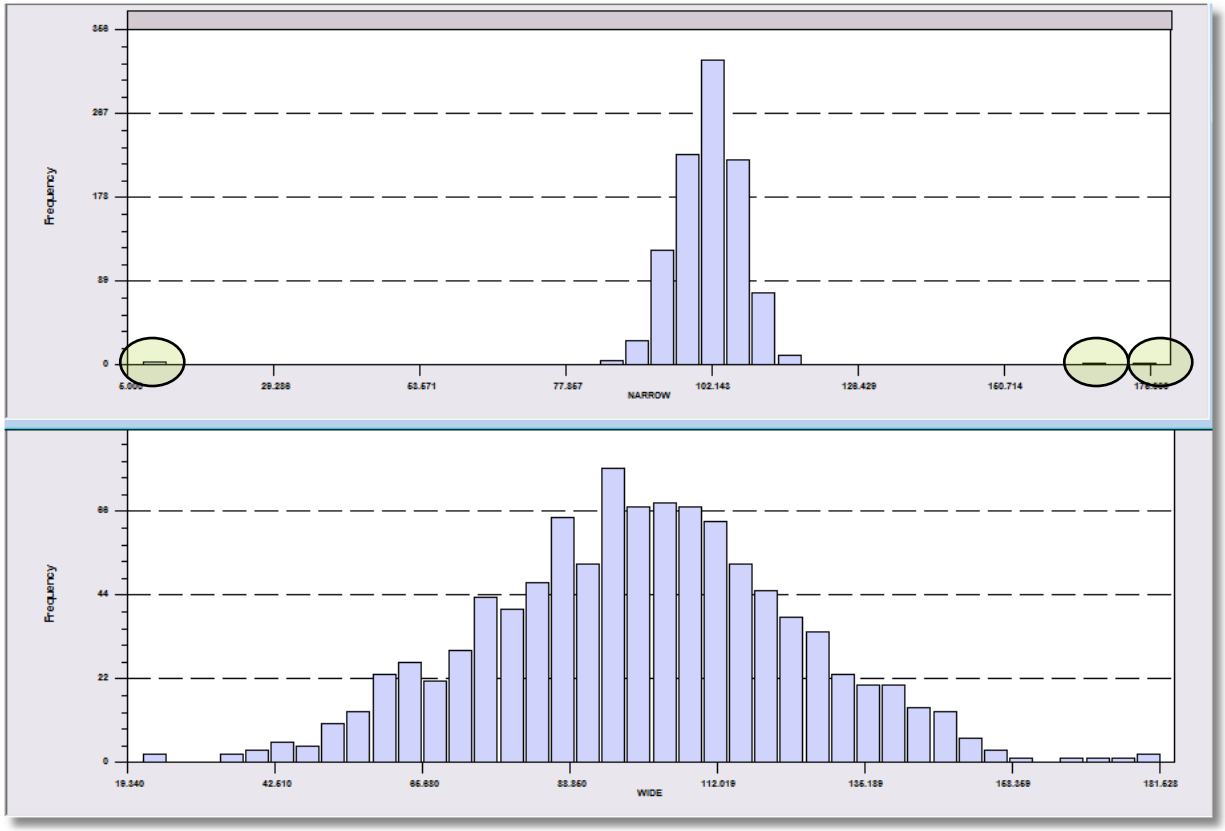
We quantify the variation of the values around the mean.

Note the range is from 1.05 to 2.70. This gives an idea where the data lie.

The mean plus a measure of the variation do the same job.

RANGE AS A MEASURE OF DISPERSION

Problems ????????



These two data sets both have 1,000 observations that range from about 10 to about 180.

VARIANCE & STDEV: MEASURES OF DISPERSION

Variance = s_X^2 = $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ or $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

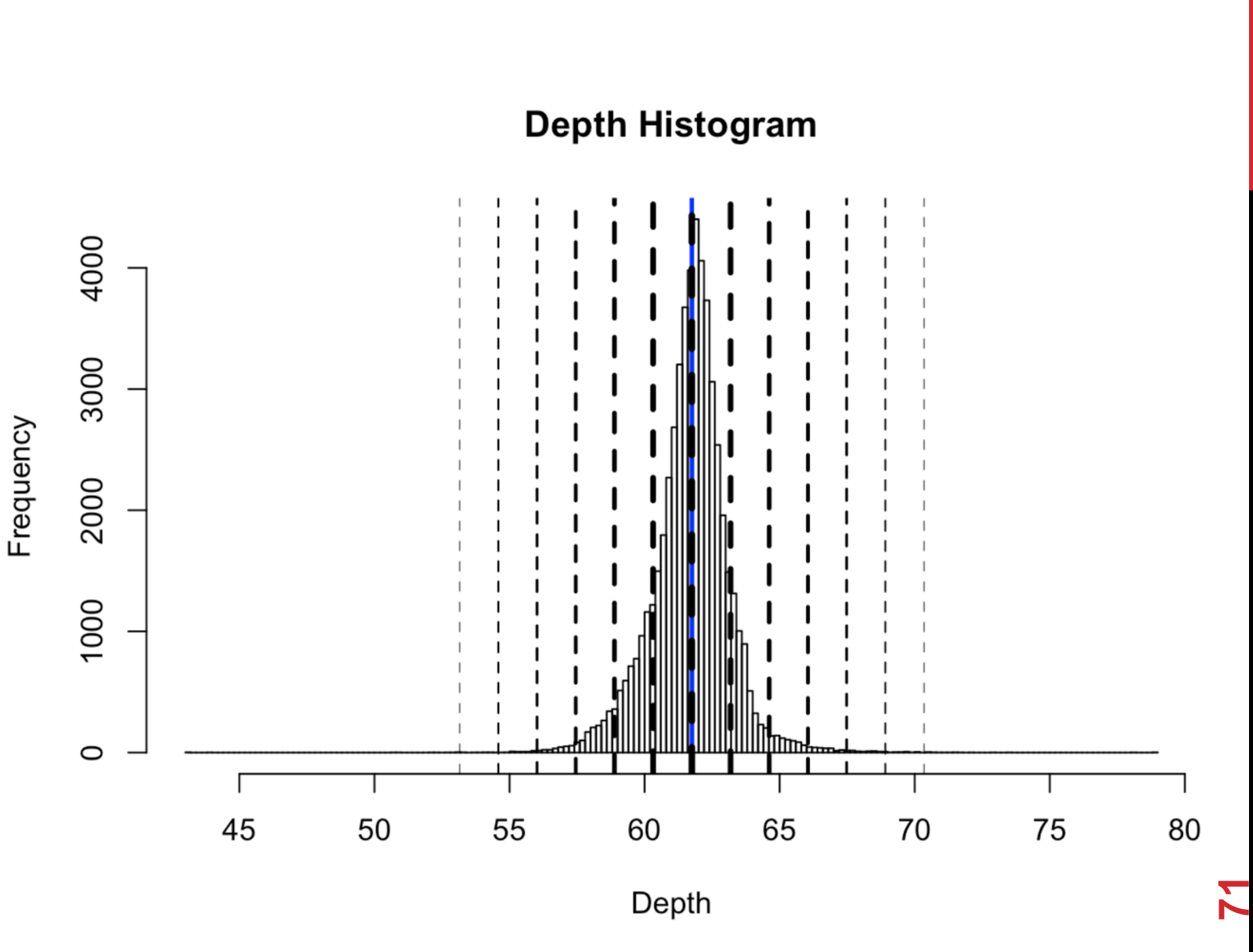
Standard deviation = s_X = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

The variance is commonly used statistic for spread

- What are the units of the variance ??????????

Standard deviation "fixes this," can be used as an interpretable unit of measurement

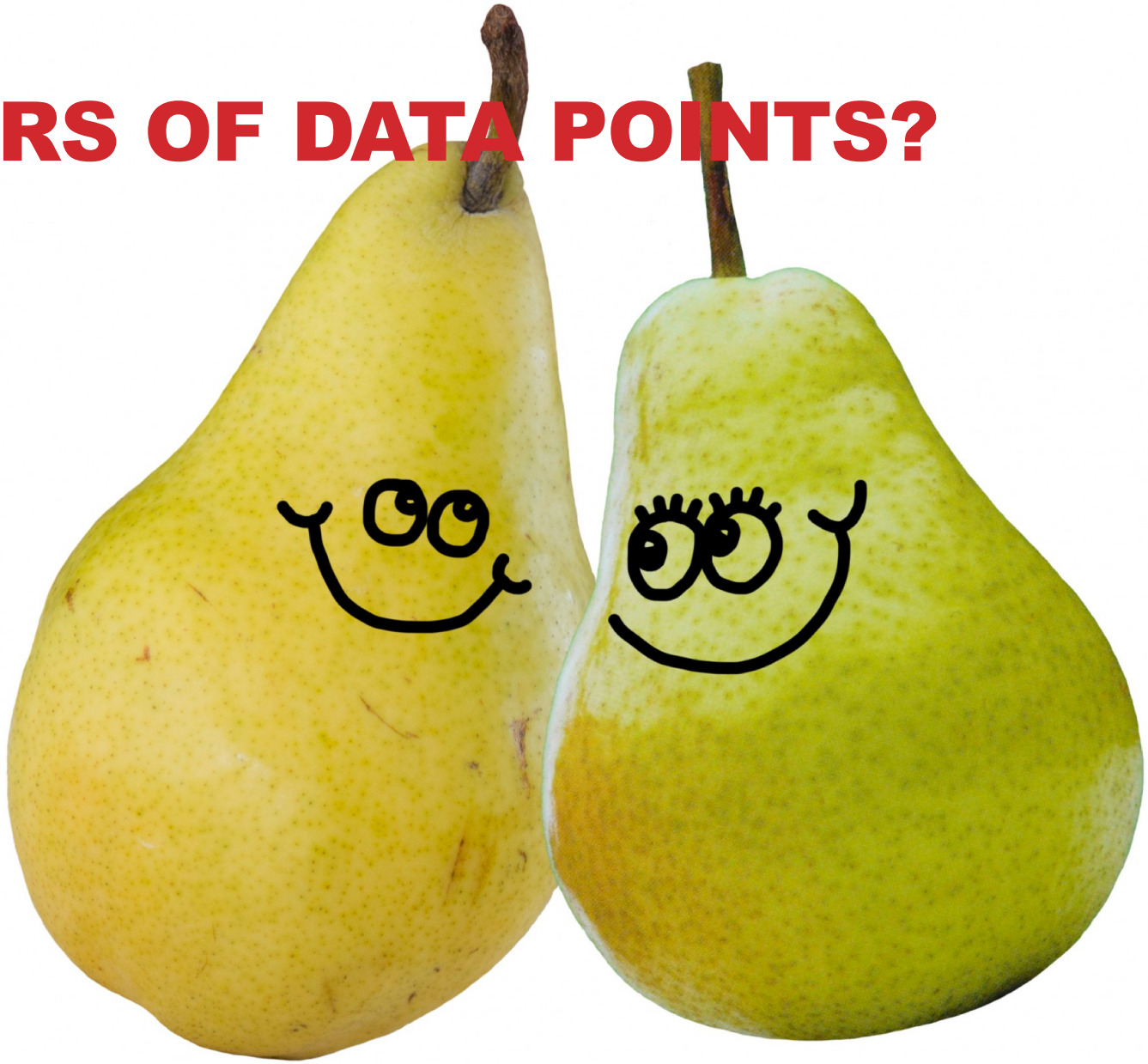
Why n-1? http://nebula.deanza.edu/~bloom/math10/m10divideby_nminus1.pdf


Depth Histogram
Frequency
0
1000
2000
3000
4000
45
50
55
60
65
70
75
80
Depth

USING "STANDARD DEVIATIONS FROM THE MEAN" AS A UNIT

SDs	Proportion	Interpretation
1	0.68	68% of the data is within ± 1 sds
2	0.95	95% of the data is within ± 2 sds
3	0.9973	99.73% of the data is within ± 3 sds
4	0.999937	99.9937% of the data is within ± 4 sds
5	0.9999994	99.999943% of the data is within ± 5 sds
6	1	99.9999998% of the data is within ± 6 sds

PAIRS OF DATA POINTS?



CORRELATION

Variables Y and X vary together

Causality vs. correlation: Does movement in X "cause" movement in Y in some metaphysical sense?

Correlation

- Simultaneous movement through a statistical relationship
- Simultaneous variation "induced" by the variation of a common third effect

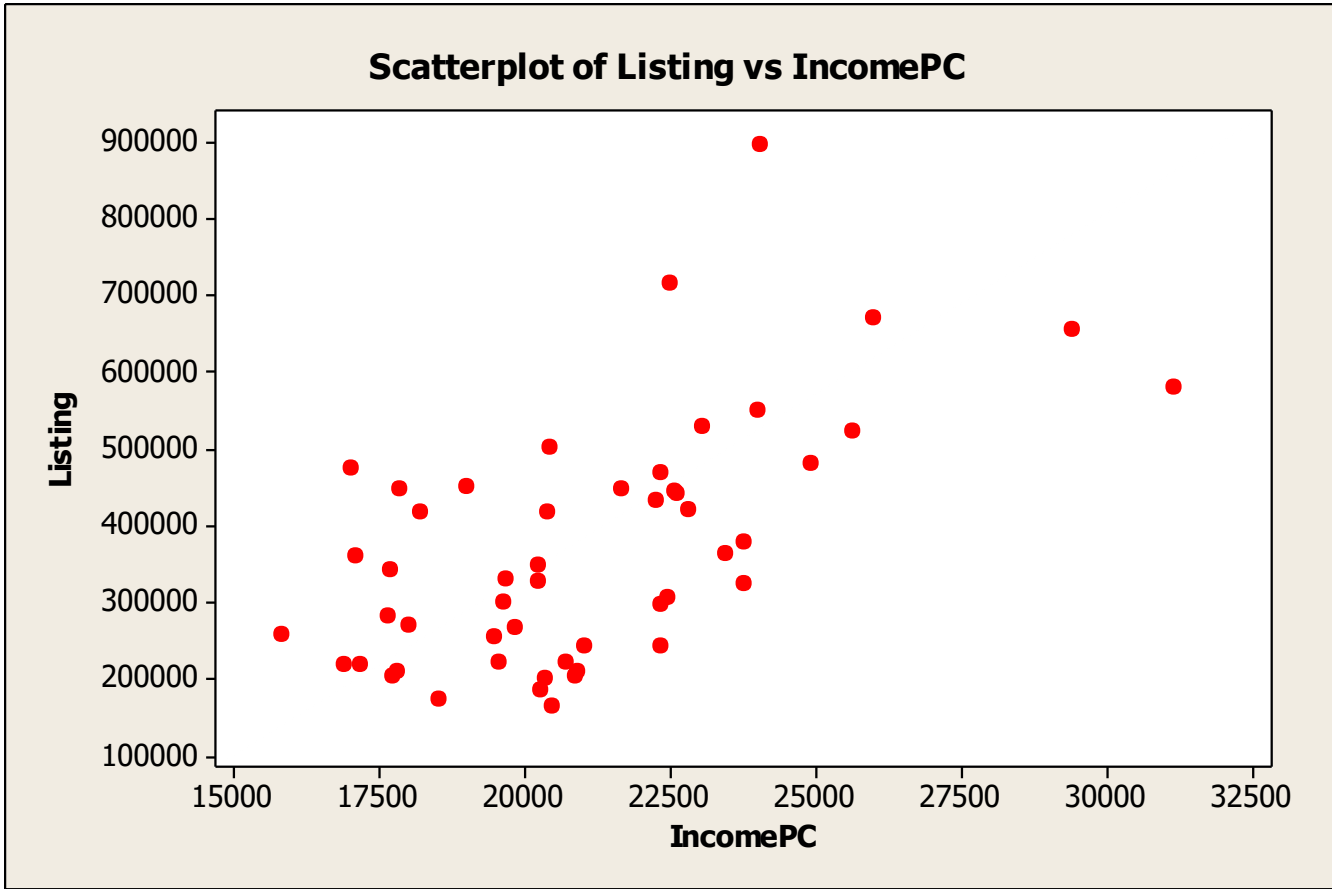
HOUSE PRICES & PER CAPITA INCOME

State	Listing	IncomePC
Hawaii	896800	24057
California	713864	22493
New York	668578	25999
Connecticut	654859	29402
Dist.Columbia	577921	31136
Nevada	549187	24023
New Jersey	529201	23038
Massachusetts	521769	25616
Wyoming	499674	20436
Maryland	480578	24933
Utah	475060	17043
Colorado	467979	22333
Arizona	448791	19001
Florida	447698	21677
Montana	446584	17865
Virginia	443618	22594
Washington	440542	22610

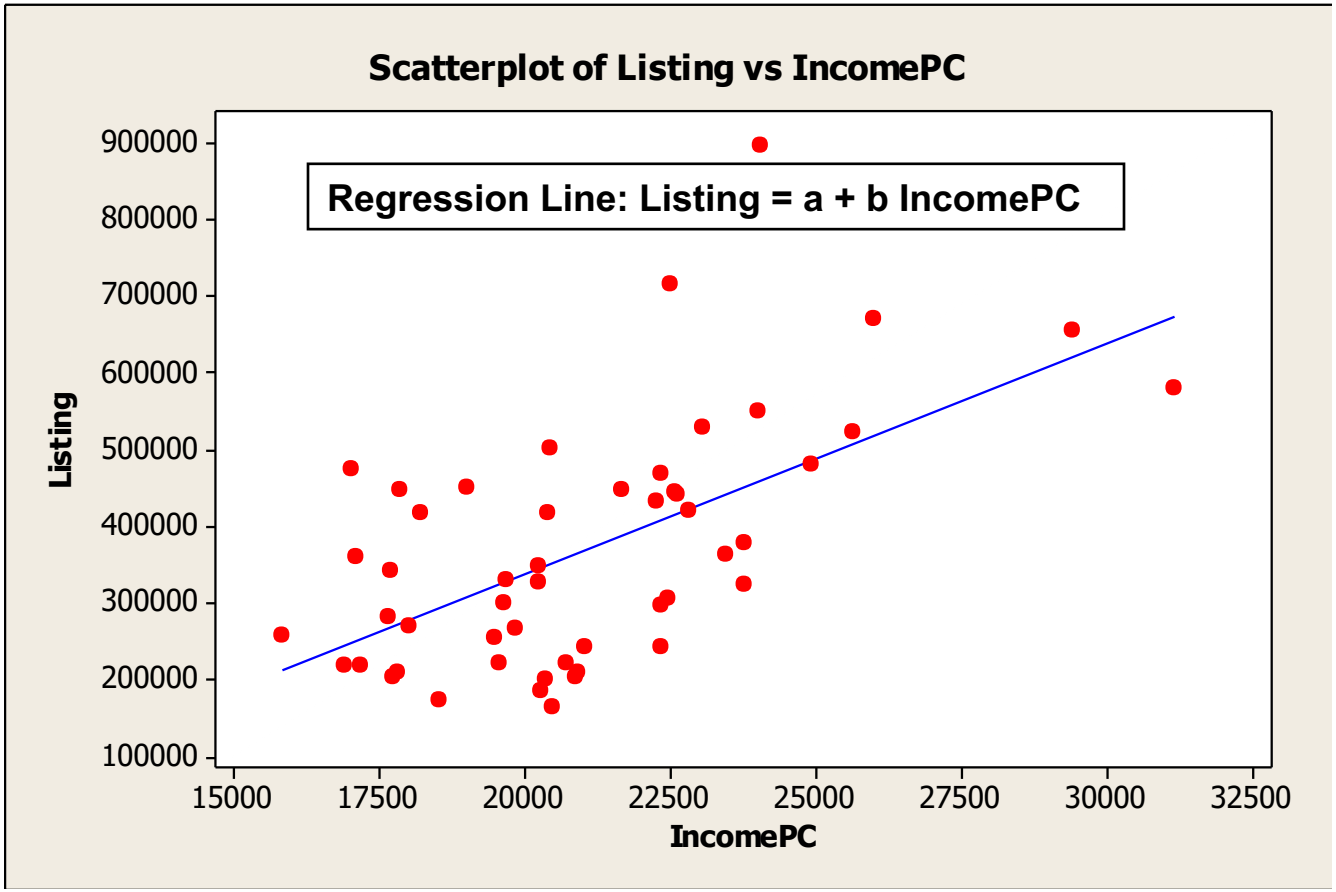
State	Listing	IncomePC
Rhode Island	432534	22251
Delaware	420845	22828
Oregon	417551	20419
Idaho	415885	18231
Illinois	377683	23784
New Hampshire	361691	23434
New Mexico	358369	17106
Vermont	346469	20224
South Carolina	340066	17695
North Carolina	330432	19669
Georgia	326699	20251
Alaska	324774	23788
Minnesota	306009	22453
Maine	299796	19663
Pennsylvania	295133	22324
Louisiana	280631	17651
Alabama	269135	18010

State	Listing	IncomePC
Texas	266388	19857
Mississippi	255774	15838
Tennessee	255064	19482
Wisconsin	243006	21019
Michigan	241107	22333
Missouri	221773	20717
South Dakota	220708	19577
West Virginia	219275	17208
Arkansas	217659	16898
Ohio	209189	20928
Kentucky	208391	17807
Oklahoma	203926	17744
Kansas	201389	20896
Indiana	200683	20378
Iowa	184999	20265
North Dakota	173977	18546
Nebraska	164326	20488

SCATTER PLOT SUGGESTS POSITIVE CORRELATION

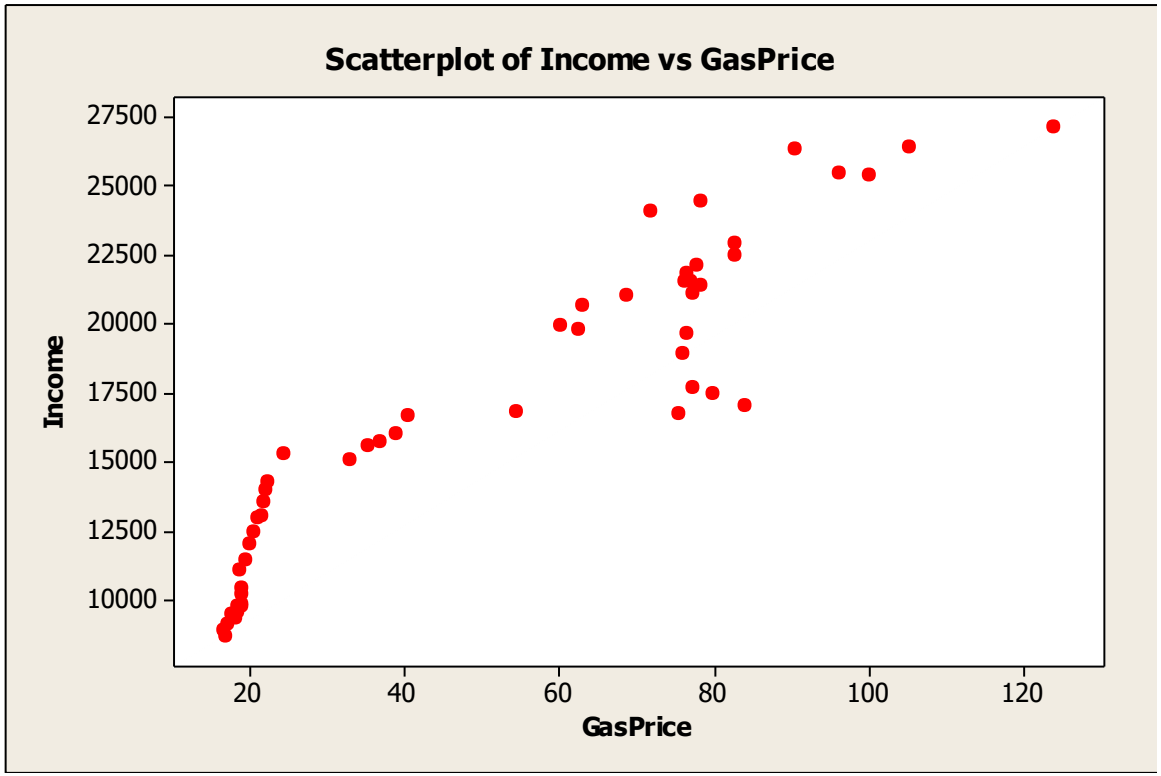


LINEAR REGRESSION MEASURES CORRELATION



CORRELATION IS NOT CAUSATION

Price and income seem to be positively correlated.

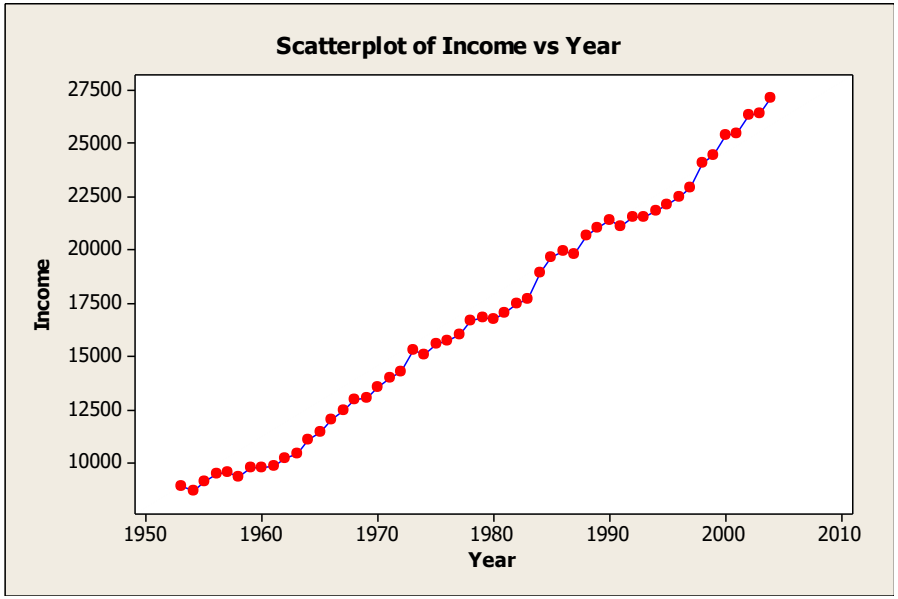


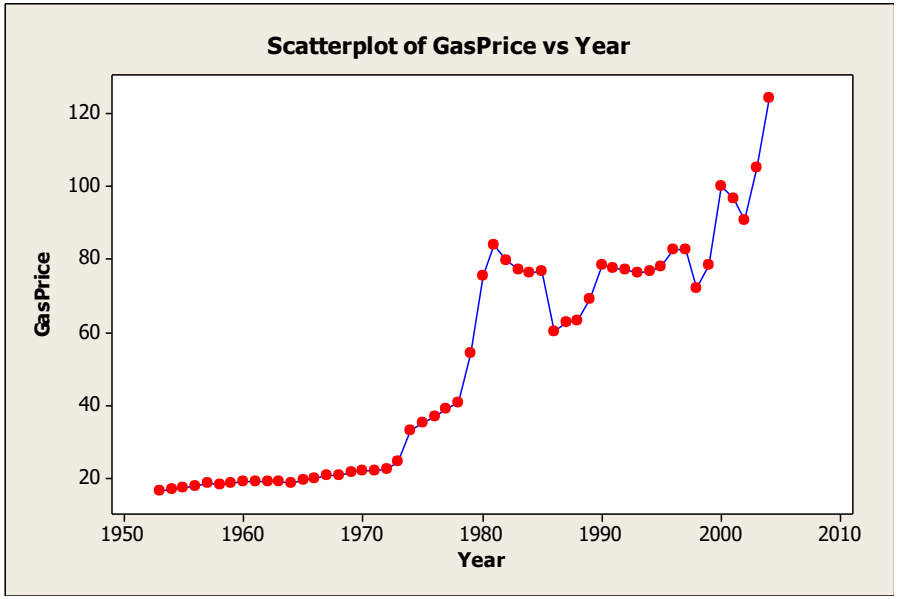
Does a rise in income cause a rise in gas prices ??????????????

US gasoline prices, 1953-2004, plotted against per-capita US income

A HIDDEN RELATIONSHIP

Not positively "related" to each other;
both positively related to "time."





"RELATED" …?

Want to capture: some variable X varies in the same direction and at the same scale as some other variable Y

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

What happens if:

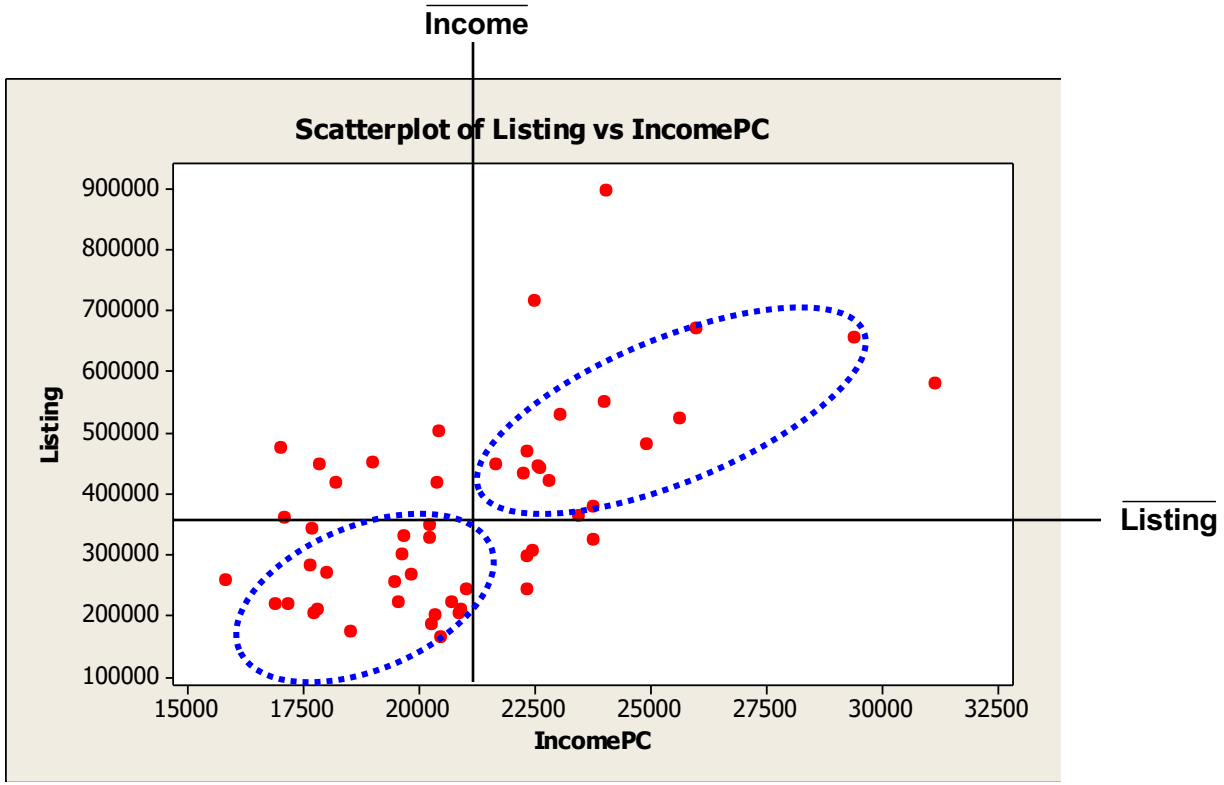
- X varies in the opposite direction as Y ????????
- X varies in the same direction as Y ????????

What are the units of the covariance ????????

Pearson's correlation coefficient is unitless in [-1,+1]:

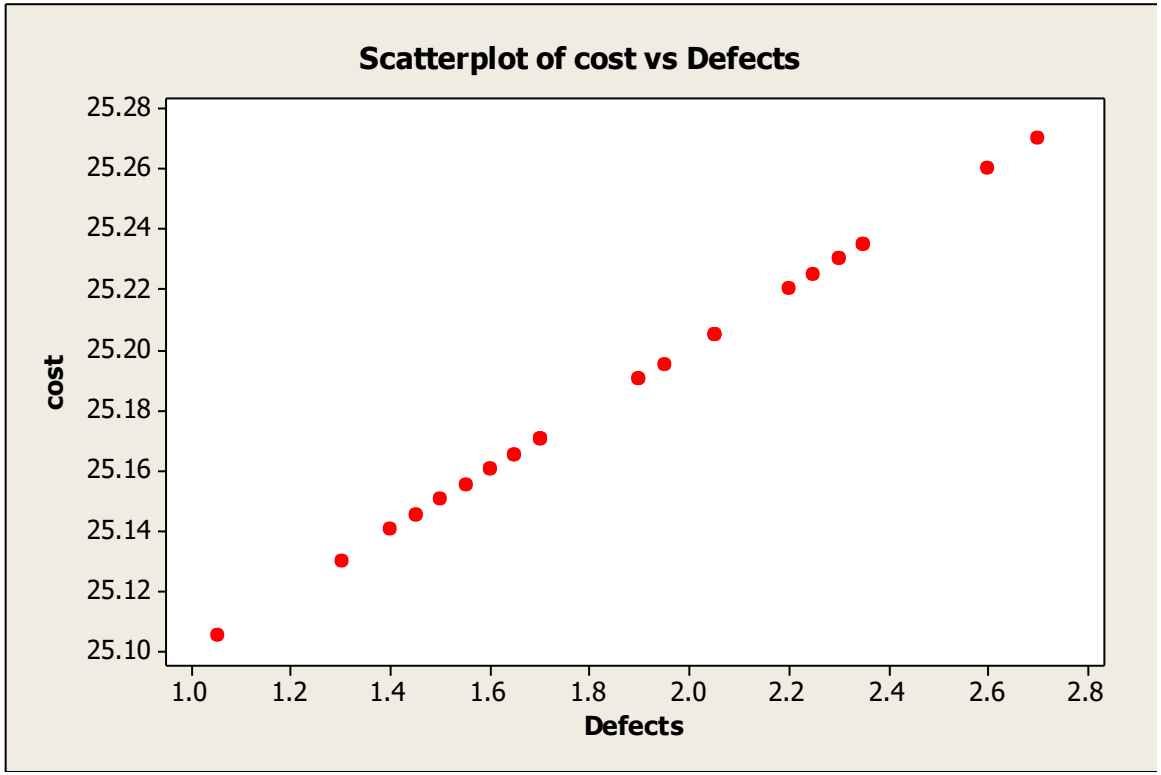
$$cor(x,y) = \frac{cov(x,y)}{sd(x)sd(y)}$$

CORRELATION

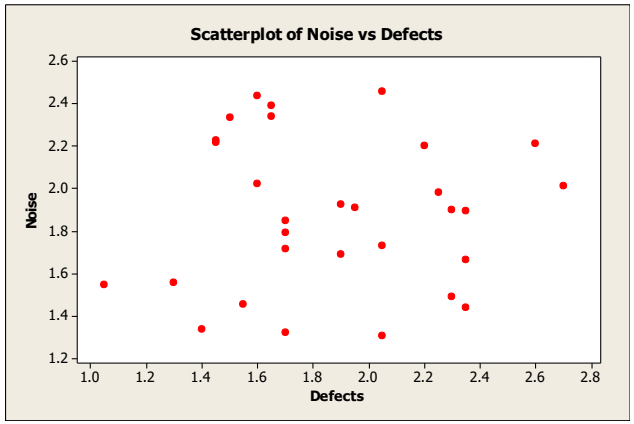


$r_{Income,Listing} = +0.591$

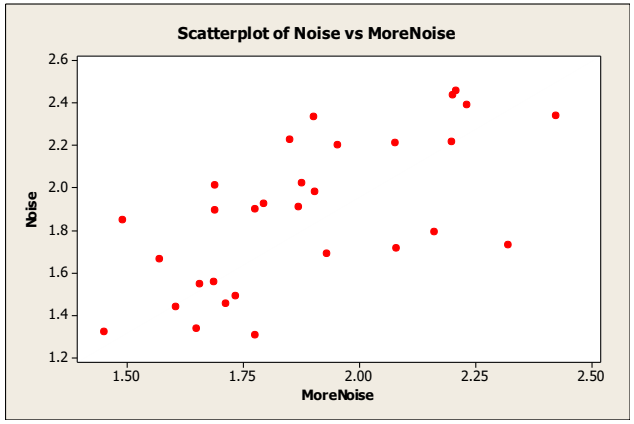
CORRELATIONS



r = +1.0

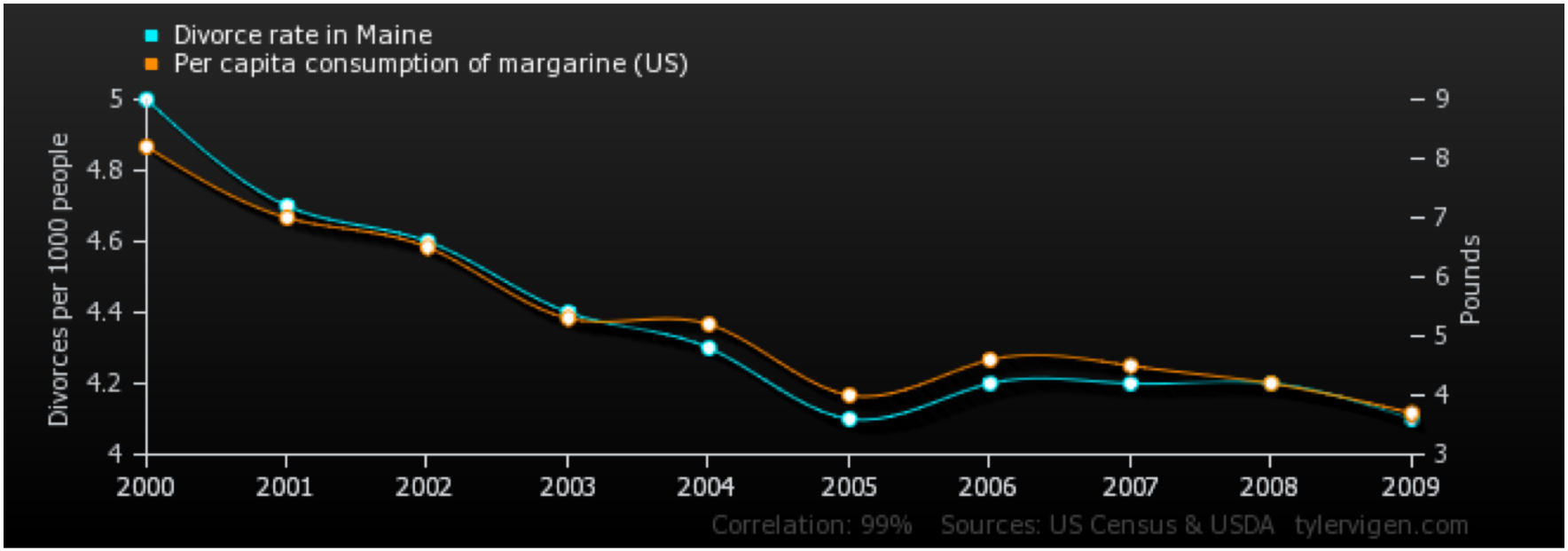


r = 0.0



r = +0.5

CORRELATION IS NOT CAUSATION!!!



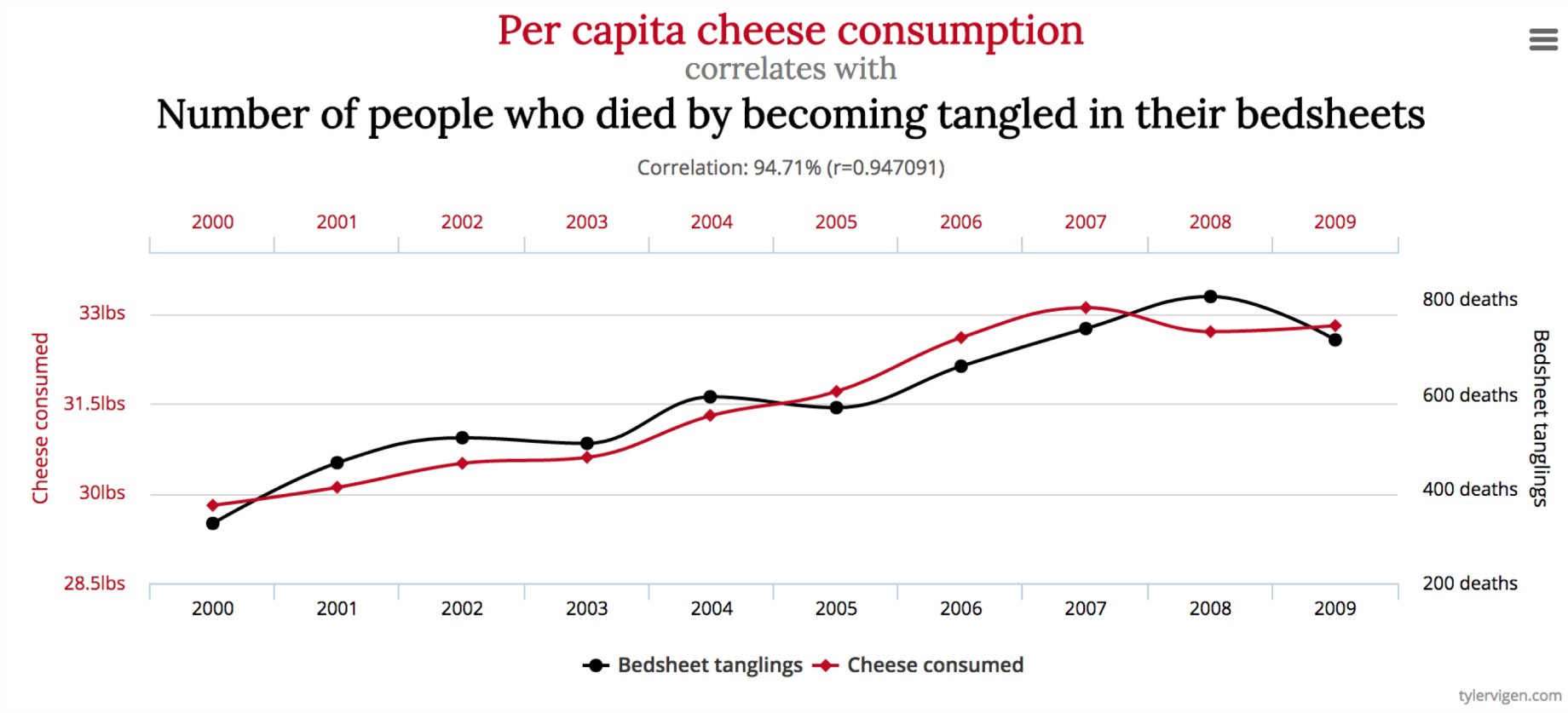
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine Divorces per 1000 people (US Census)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) Pounds (USDA)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

r=0.993

?????????

http://tylervigen.com/spurious-correlations

JUST TO DRIVE THE POINT HOME …



TRANSFORMATIONS

TRANSFORMATIONS

So, you've figured out that your data are:

- Skewed
- Have vastly different ranges across datasets and/or different units

What do you do?

Transform the variables to:

- ease the validity and interpretation of data analyses
- change or ease the type of Stat/ML models you can use

STANDARDIZATION

Transforming the variable to a comparable metric

- known unit
- known mean
- known standard deviation
- known range

Three ways of standardizing:

- P-standardization (percentile scores)
- Z-standardization (z-scores)
- D-standardization (dichotomize a variable)

Slides adapted from Maarten Buis

WHEN YOU SHOULD ALWAYS STANDARDIZE

When averaging multiple variables, e.g. when creating a socioeconomic status variable out of income and education.

When comparing the effects of variables with unequal units, e.g. does age or education have a larger effect on income?

Slides adapted from Maarten Buis

P-STANDARDIZATION

Every observation is assigned a number between 0 and 100, indicating the percentage of observation beneath it.

Can be read from the cumulative distribution

In case of knots: assign midpoints

The median, quartiles, quintiles, and deciles are special cases of P-scores.

Slides adapted from Maarten Buis

	rent	cum %	percentile
room 1	175	5,3%	5,3%
room 2	180	10,5%	10,5%
room 3	185	15,8%	15,8%
room 4	190	21,1%	21,1%
room 5	200	26,3%	26,3%
room 6	**210**	31,6%	**36,8%**
room 7	**210**	36,8%	**36,8%**
room 8	**210**	42,1%	**36,8%**
room 9	230	47,4%	47,4%
room 10	**240**	52,6%	**55,3%**
room 11	**240**	57,9%	**55,3%**
room 12	**250**	63,2%	**65,8%**
room 13	**250**	68,4%	**65,8%**
room 14	280	73,7%	73,7%
room 15	**300**	78,9%	**81,6%**
room 16	**300**	84,2%	**81,6%**
room 17	310	89,5%	89,5%
room 18	325	94,7%	94,7%
room 19	620	100,0%	100,0%

Slides adapted from Maarten Buis

P-STANDARDIZATION

Turns the variable into a ranking, i.e. it turns the variable into a ordinal variable.

It is a non-linear transformation: relative distances change

Results in a fixed mean, range, and standard deviation; M=50, SD=28.6, This can change slightly due to knots

A histogram of a P-standardized variable approximates a uniform distribution

Slides adapted from Maarten Buis

CENTERING AND SCALING

Transform your data into a unitless scale

- Put data into "standard deviations from the mean" units
- This is called standardizing a variable, into standard units

Given data points $x = x_1, x_2, \ldots, x_n$:

$$z_i = \frac{(x_i - \bar{x})}{\mathrm{sd}(x)}$$

Translates *x* into a scaled and centered variable *z*

What is the mean of *z* ??????????

What is the standard deviation of z ??????????

CENTERING OR SCALING

Maybe you just want to center the data:

$$z_i = (x_i - \bar{x})$$

What is the mean of *z* ??????????

What is the standard deviation of z ??????????

Maybe you just want to scale the data:

$$z_i = \frac{x_i}{\mathrm{sd}(x_i)}$$

What is the mean of *z* ??????????

What is the standard deviation of z ??????????

DISCRETE TO CONTINUOUS VARIABLES

Some models only work on continuous numeric data

Convert a binary variable to a number ???????????

- health_insurance = {“yes”, “no”} → {1, 0}

Why not {-1, +1} or {-10, +14}?

- 0/1 encoding lets us say things like “if a person has healthcare then their income increases by $X.”
- Might need {-1,+1} for certain ML algorithms (e.g., SVM)

DISCRETE TO CONTINUOUS VARIABLES

What about non-binary variables?

My main transportation is a {BMW, Bicycle, Hovercraft}

One option: { BMW → 1, Bicycle → 2, Hovercraft → 3 }

- Problems ??????????

One-hot encoding: convert a categorical variable with N values into a N-bit vector:

- BMW → [1, 0, 0]; Bicycle → [0, 1, 0]; Hovercraft → [0, 0, 1]

```
# Converts dtype=category to one-hot-encoded cols
cols = ['my_transportation']
df = df.get_dummies( columns = cols )
```

CONTINUOUS TO DISCRETE VARIABLES

Do doctors prescribe a certain medication to older kids more often? Is there a difference in wage based on age?

Pick a discrete set of bins, then put values into the bins

Equal-length bins:

- Bins have an equal-length range and skewed membership
- Good/Bad ????????

Equal-sized bins:

- Bins have variable-length ranges but equal membership
- Good/Bad ????????

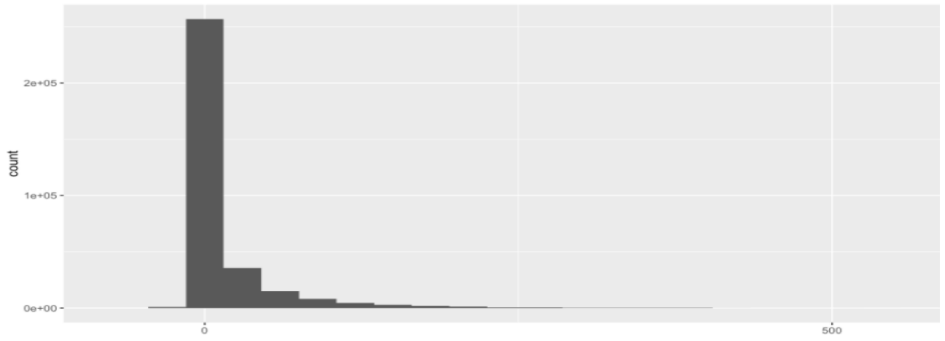
SKEWED DATA

Skewed data often arises in multiplicative processes:

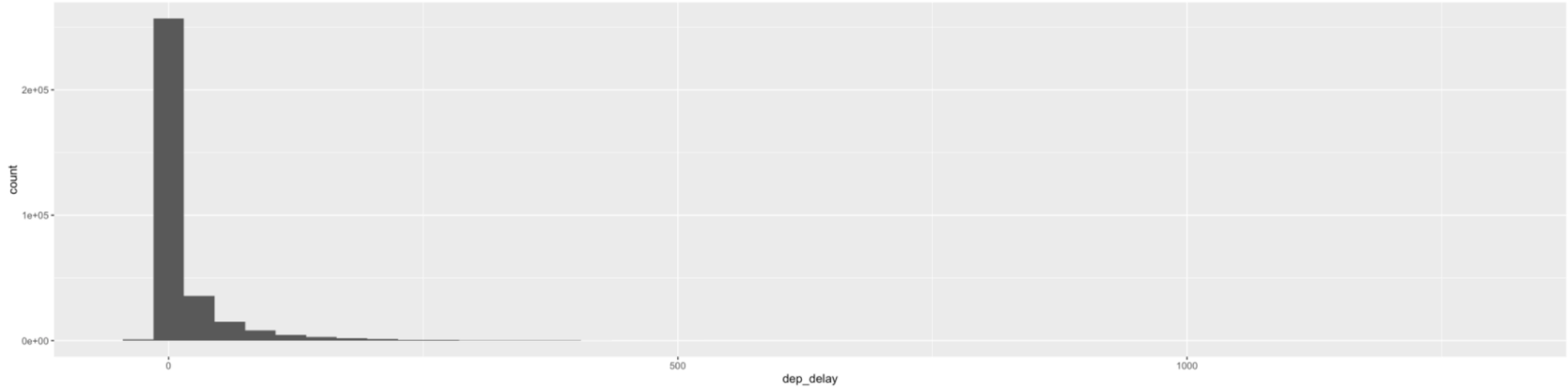
- Some points float around 1, but one unlucky draw → 0

Logarithmic transforms reduce skew:

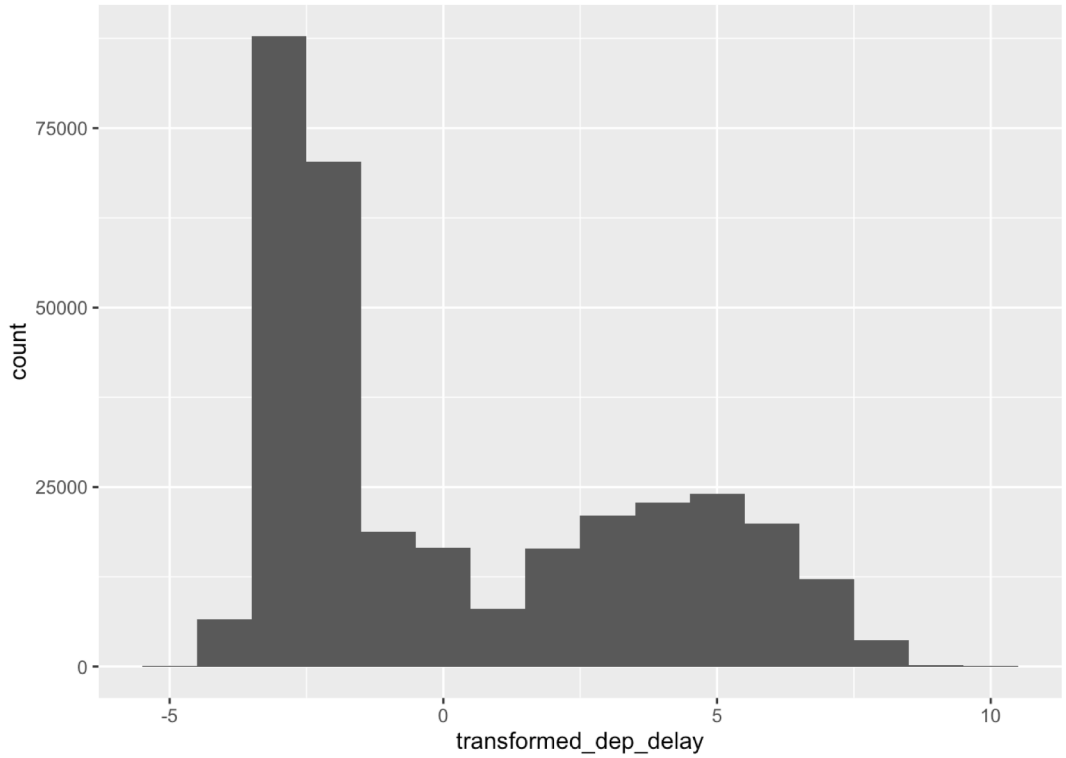
- If values are all positive, apply $\log_2$ transform
- If some values are negative:
  - Shift all values so they are positive, apply $\log_2$
  - Signed log: sign(x) * $\log_2$(|x| + 1)



SKEWED DATA



$\log_2$ transform on airline takeoff delays



NEXT CLASS:

VISUALIZATION, GRAPHS, & NETWORKS