Data Mining and Analysis

Data Cleaning

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

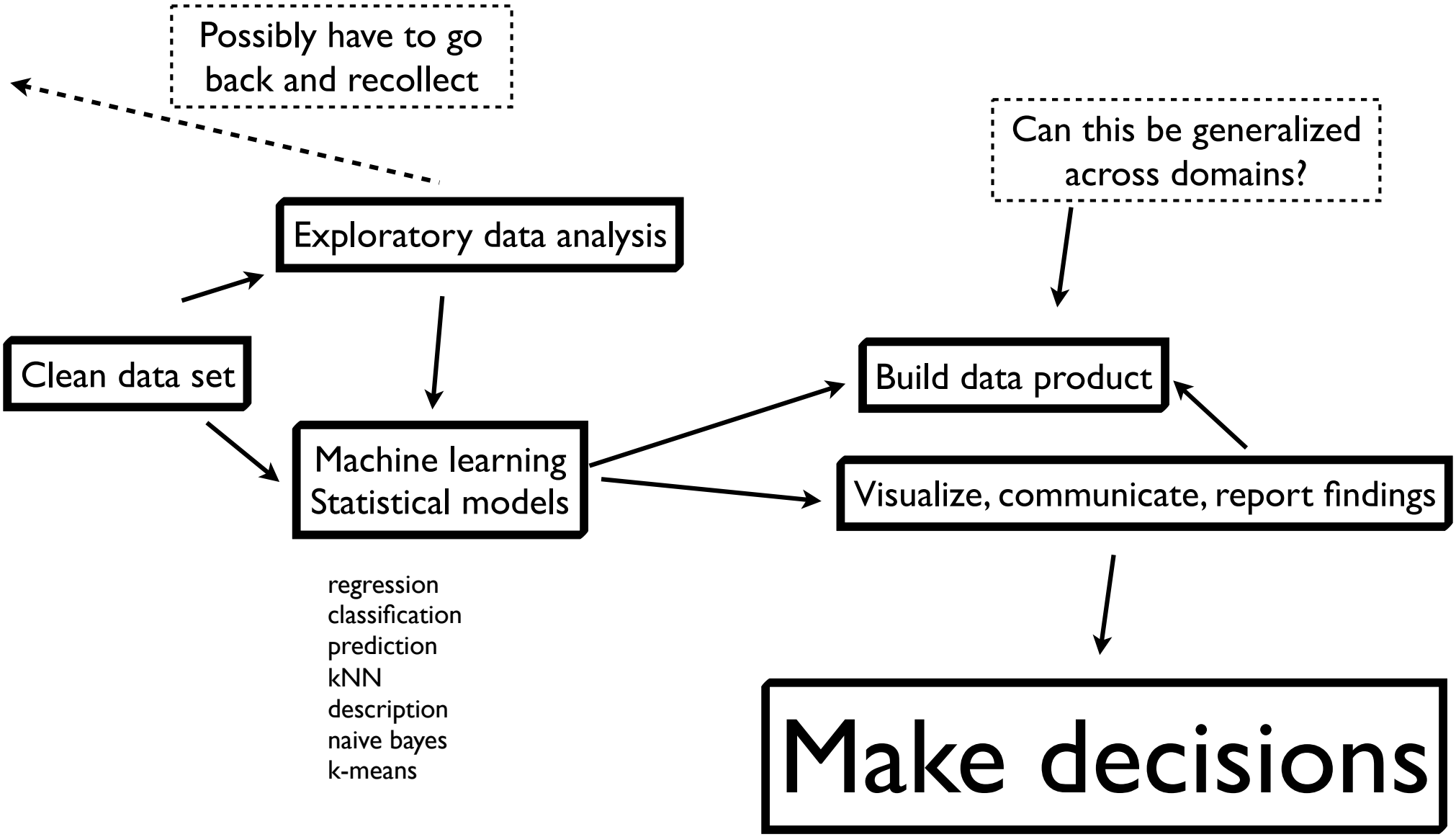
Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

Data set

"clean"
table

http://columbiadatascience.com/2012/09/24/reflections-after-jakes-lecture/


Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning
Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Real World Data is a Mess

Incomplete

Missing some attribute values, lacking the attributes you care about, or not at the granularity you want

Noisy

Errors and outliers

Inconsistent

Discrepancies in how values are coded / entered

MO
California
YR
NULL

Mashable VIDEO ENTERTAINMENT CULTURE TECH SCIENCE SOCIAL GOOD SHOP MORE

He tried to prank the DMV. Then his vanity license plate backfired big time.

https://mashable.com/article/dmv-vanity-license-plate-def-con-backfire/

18125814	NULL	CA	$37.00
18125813	NULL	CA	$37.00
17129283	NULL	CA	$74.00
17129309	NULL	CA	$74.00
12127575	NULL	CA	$37.00
11128431	NULL	CA	$37.00
12127632	NULL	CA	$37.00
12127660	NULL	CA	$37.00
12127772	NULL	CA	$37.00
19129664	NULL	CA	$37.00
		Total Due:	$12049.00

Our overarching goal: Extract Value from Data

To get there, we will:

- remove errors
- fill in missing information
- transform units and formats
- map and align columns
- remove duplicate records
- fix integrity constraint violations

Some slides from: SIGMOD 2016 Tutorial: Data Cleaning: Overview and Emerging Challenges

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

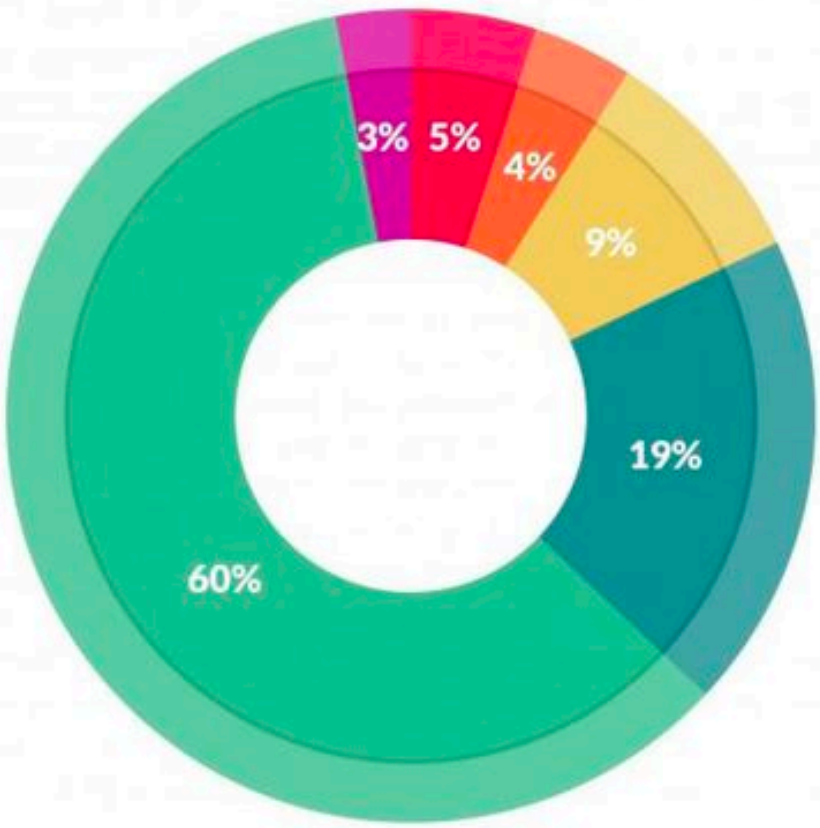
Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

"Data wrangling is a huge — and surprisingly so — part of the job," said Monica Rogati, vice president for data science at Jawbone, whose sensor-filled wristband and software track activity, sleep and food consumption, and suggest dietary and health tips based on the numbers. "It's something that is not appreciated by data civilians. At times, it feels like everything we do."

61,440 views | Mar 23, 2016, 09:33am

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press Contributor
Enterprise & Cloud
I write about technology, entrepreneurs and innovation.



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Real World Data is a Mess

Incomplete

Missing some attribute values, lacking the attributes you care about, or not at the granularity you want

e.g., Occupation = “” (missing)

Noisy

Errors and outliers

e.g., Salary = -100 (error)

Inconsistent

Discrepancies in how values are coded / entered

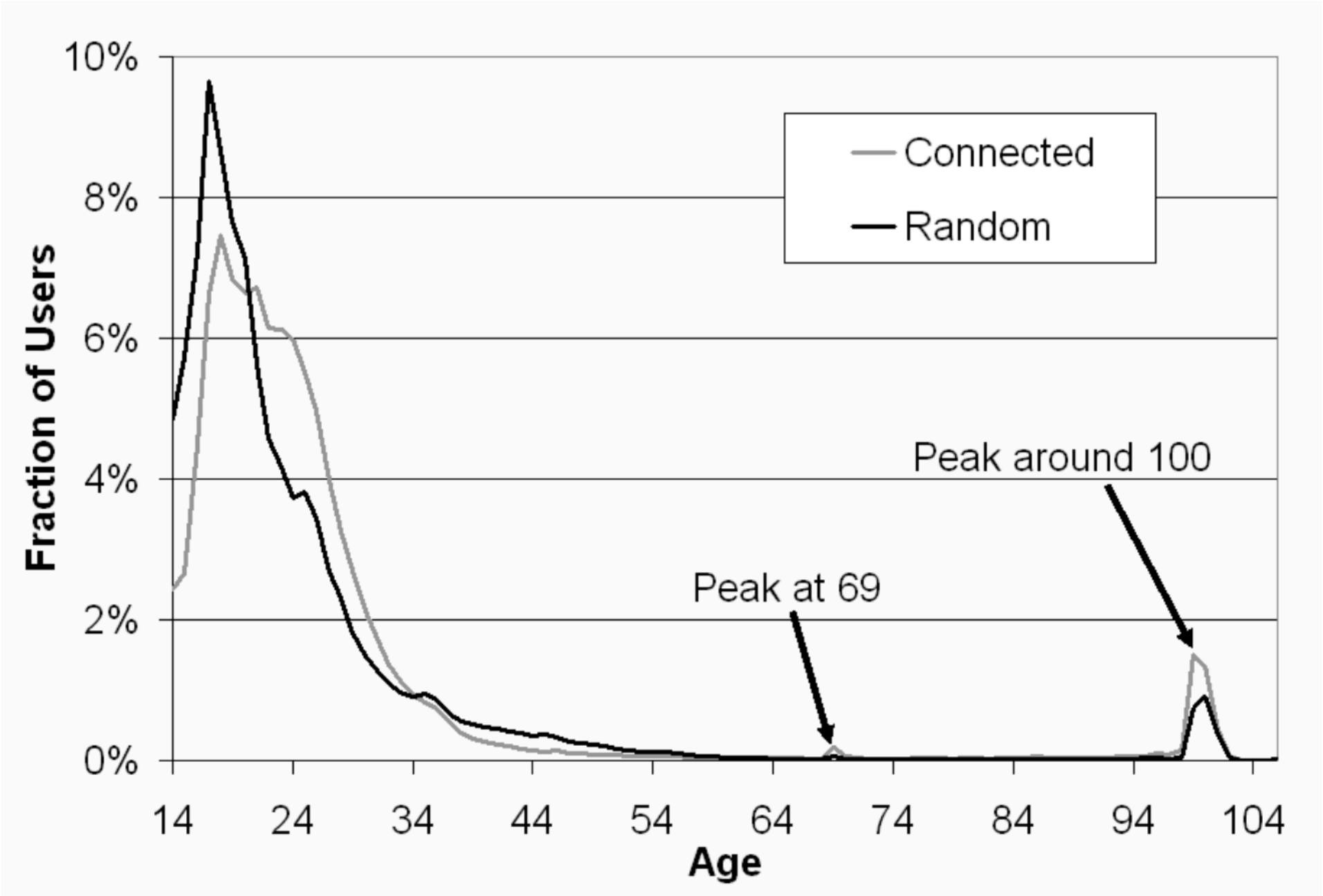
e.g., Age = 45 vs. Birthday = June 1974

Intentional

Data intentionally entered incorrectly

e.g., MySpace ages …

Self-Reported Ages on MySpace



Group Activity!
(form groups; find examples of data + identify “cleaning” need)

Examples

Amazon Reviews —> fake reviews, fake ratings

Mouse tracking —> user idles may skew heatmap

Eyetracking —> random looks away, noise

Fitness tracker —> sampling rate mismatch may disrupt activity recognition

Weather sensing —> random sensing errors, noise, errors due to "exogenous factors", e.g., person holding a match next to a sensor

Social media posts —> polluted by bots, trolls, propaganda, plus some posts may be missing due to censorship

Shipping addresses —> mistakes between street address and ZIP code

Instagram ads —> ad/click fraud (intentional errors)

Let’s walk through some typical data cleaning tasks

For each, ask yourself:

- What is it?
- Why does it occur?
- How to deal with?
- What impact will this have on our analysis?

Task 1: Missing Values

Most real data collected from sensors, surveys, agents, have a high percentage of N/A or nulls, special values (99999) etc.

Solutions?

Missing Values

What do we do? What strategies?

- Ignore the tuple (throw the baby out with the bathwater)
- Make a guess (imputation):
  - if numeric, fill in with special value (e.g., 1000 for age), mean, median, or mode
  - if categorical: ?
  - if ordinal: ?

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Task 2: Duplicates

Similarity measures: how to define? Exact match or soft match? (e.g., James Caverlee vs Jim Shmaverlee)

Machine learning for classifying pairs as duplicates or not (unsupervised, supervised, and active)

Clustering and handling of transitivity

Merging and consolidation of records

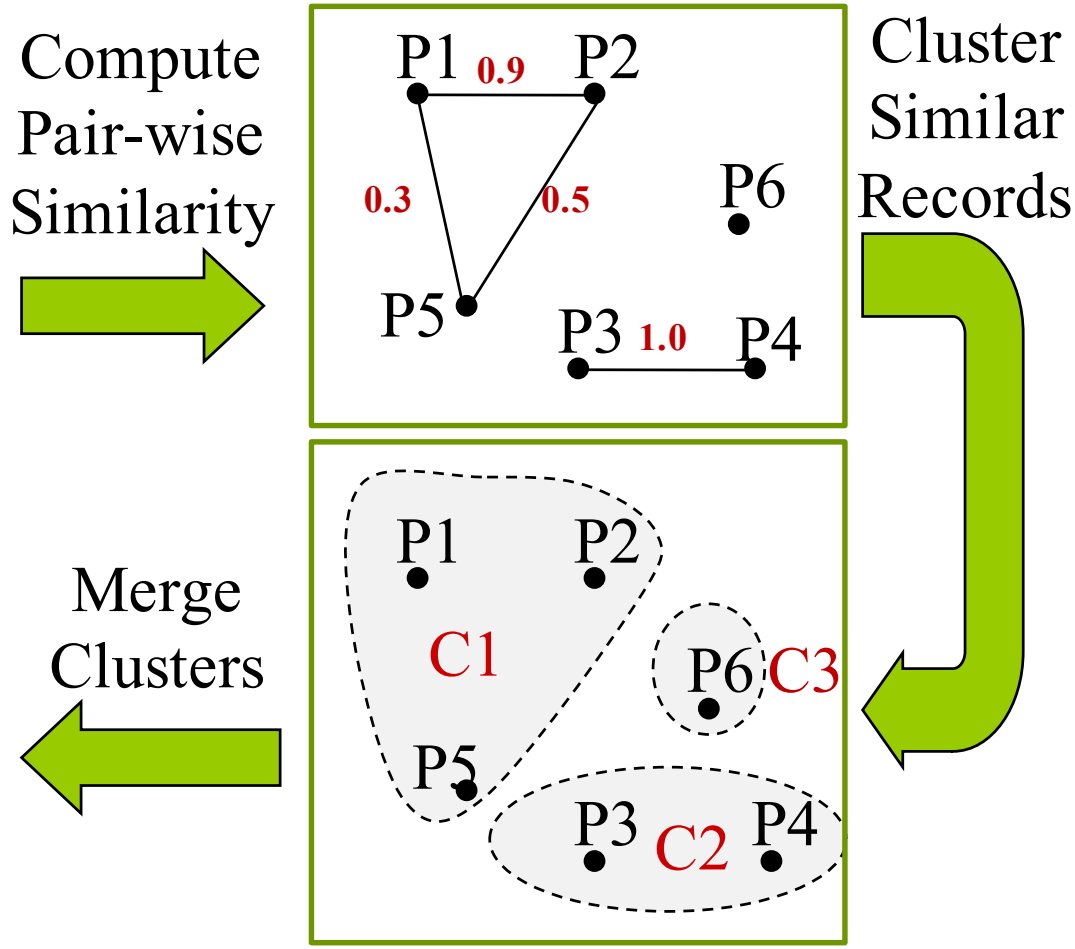
Example: Data Deduplication

Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k



Task 3: Integrity Constraints

	ID	FN	LN	ROLE	CITY	ST	SAL
t_1	105	Anne	Nash	M	NYC	NY	110
t_2	211	Mark	White	E	SJ	CA	80
t_3	386	Mark	Lee	E	NYC	AZ	75
t_4	235	John	Smith	M	NYC	NY	1200

Employee Table

Functional dependency:

$$City \rightarrow ST$$

Integrity Constraints

	ID	FN	LN	ROLE	CITY	ST	SAL
t_1	105	Anne	Nash	M	NYC	NY	110
t_2	211	Mark	White	E	SJ	CA	80
t_3	386	Mark	Lee	E	NYC	AZ	75
t_4	235	John	Smith	M	NYC	NY	1200

Employee Table

Business Rule:
Two employees of the same role in the same city, their salary difference cannot be greater than 100

Task 4: Noise / Oddities

Data Cleaning Sherlock:

- Use exploratory data analysis to investigate: Unique values, Most frequent values, Highest and lowest values, Location and dispersion, Quartiles

Consider “suspicious data” like outliers

Task 5: Combining Data Sources

aka “Data integration”:

- Combines data from multiple sources into a coherent store

Schema integration: e.g., A.cust-id B.cust-#

- Integrate metadata from different sources

Entity identification problem:

- Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

Detecting and resolving data value conflicts

- For the same real world entity, attribute values from different sources are different
- Possible reasons: different representations, different scales, e.g., metric vs. British units

Task 6: Data Reduction

Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

We'll revisit this topic later in the semester with:

- Clustering
- SVD
- PCA

Task 7: Data Massaging

A typical pre-processing step before we apply our “data mining” toolkit

Topics include: min-max normalization, z-score normalization, one-hot encoding

Tools

http://openrefine.org/

https://www.trifacta.com/start-wrangling/

…

OpenRefine

A free, open source, powerful tool for working with messy data

Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

OpenRefine is available in English, Chinese, Spanish, French, Russian, Portuguese (Brazil), German, Japanese, Italian, Hungarian, Hebrew, Filipino, Cebuano, Tagalog

OpenRefine was supported in 2018 by:

Google News Initiative

Introduction to OpenRefine

1. Explore Data

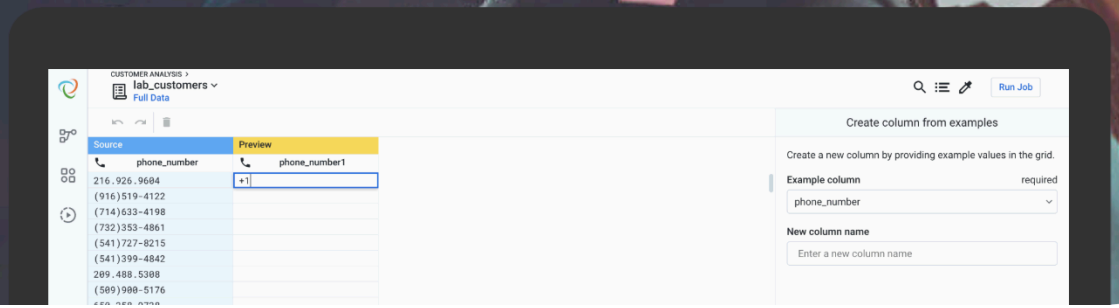
OpenRefine can help you explore large data sets with ease. You can find out more about this functionality by watching the video below.

TRIFACTA

Login

Get Trifacta Wrangler

Speed up data preparation and cleaning



Start Wrangling

Wrangle up to 100MB files!

SIGN UP

System Requirements:

- Chrome 70 or above

Scaling and Normalization

Features have different scales

- GPA vs. Age vs. Height

Map to a common space

- Z-score
- Min-max

z-score normalization

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

z-score of 0 is right on the mean

z-score of 1 is one standard deviation above mean

z-score of -2 is two standard deviations below mean

Min-max Scaling

$$z_i = \frac{x_i - \min}{\max - \min}$$

Map to [0,1]

Sensitive to outliers

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Dealing with Categorical Data

Suppose we've got four majors:

[English, History, Math, CS]

Many of our downstream analyses will only understand data as a number (integer, float, etc.)

[English, History, Math, CS] —> [0, 1, 2, 3]

But that indicates order, i.e., CS > English

One alternative: One Hot Encoding

One Hot Encoding

[English, History, Math, CS]

becomes

[1, 0, 0, 0]

[0, 1, 0, 0]

[0, 0, 1, 0]

[0, 0, 0, 1]

Arguments against Data Cleaning?

Thoughts?