BROWN
Computer Science

CS1951A: Data Science

Lecture 5: Data Cleaning

Lorenzo De Stefani

Spring 2022

Overiew

- Dirty data
- Where does the “dirt” come from?
- Dirty Data problems
- Data Quality Continuum
- Data Quality Metrics

Lets look as some data

ID	Name	Street	City	State	Zip	Hours
1	N Aldroubi	123 University Ave	Providenc e	RI	98106	42
2	Sunny Deng	245 3rd St	Pawtucket	RI	98052-1234	30
3	Nam Do	345 Broadway	PVD	Rhode Island	98101	19
4	S Deng	245 Third Street	Pawtucket	NULL	98052	299
5	Do Nam	345 Broadway St	Providnce	Rhode Island	98101	19
6	Nazem Aldroubi	123 Univ Ave	PVD	Rhode Island	NULL	41
7	Nam D	123 University Ave	Providenc e	Guyana	94305	NULL

…

Data is imperfect

- Most Data is *intrinsically dirty* at its source
- Transformations corrupt the data (complexity of software pipelines)
- Individual data sets are clean but integration (i.e., combining them) introduces problems (e.g., incoherent records, duplication, non-consistent representation)
  - "Rare" errors can become frequent after transformation or integration.
- Data sets are clean but suffer "bit rot"
  - Old data loses its value/accuracy over time
- Outliers may skew data analysis
  - Hard to distinguish true outliers from rare values
- Any combination of the above

The impact of dirty data

TAS

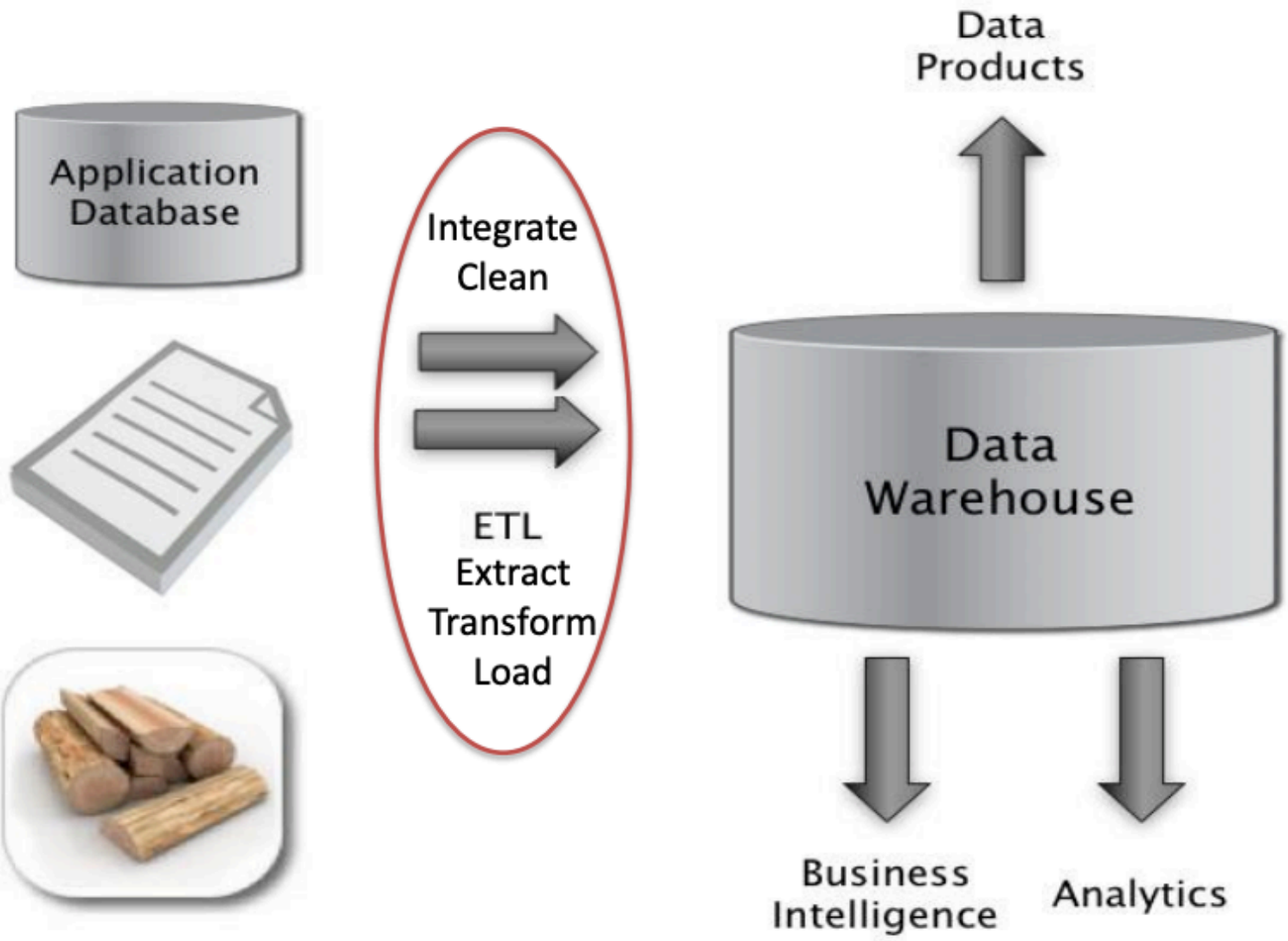
ID	Name	City	State	Hours
1	N Aldroubi	Providence	RI	42
2	Sunny Deng	Pawtucket	RI	30
3	Nam Do	PVD	Rhode Island	19
4	S Deng	Pawtucket	NULL	300
5	Do Nam	Providence	Rhode Island	19
6	Nazem Aldroubi	PVD	Rhode Island	42
7	J-P Champa	Warwick	RI	NULL

How many TAS are there?

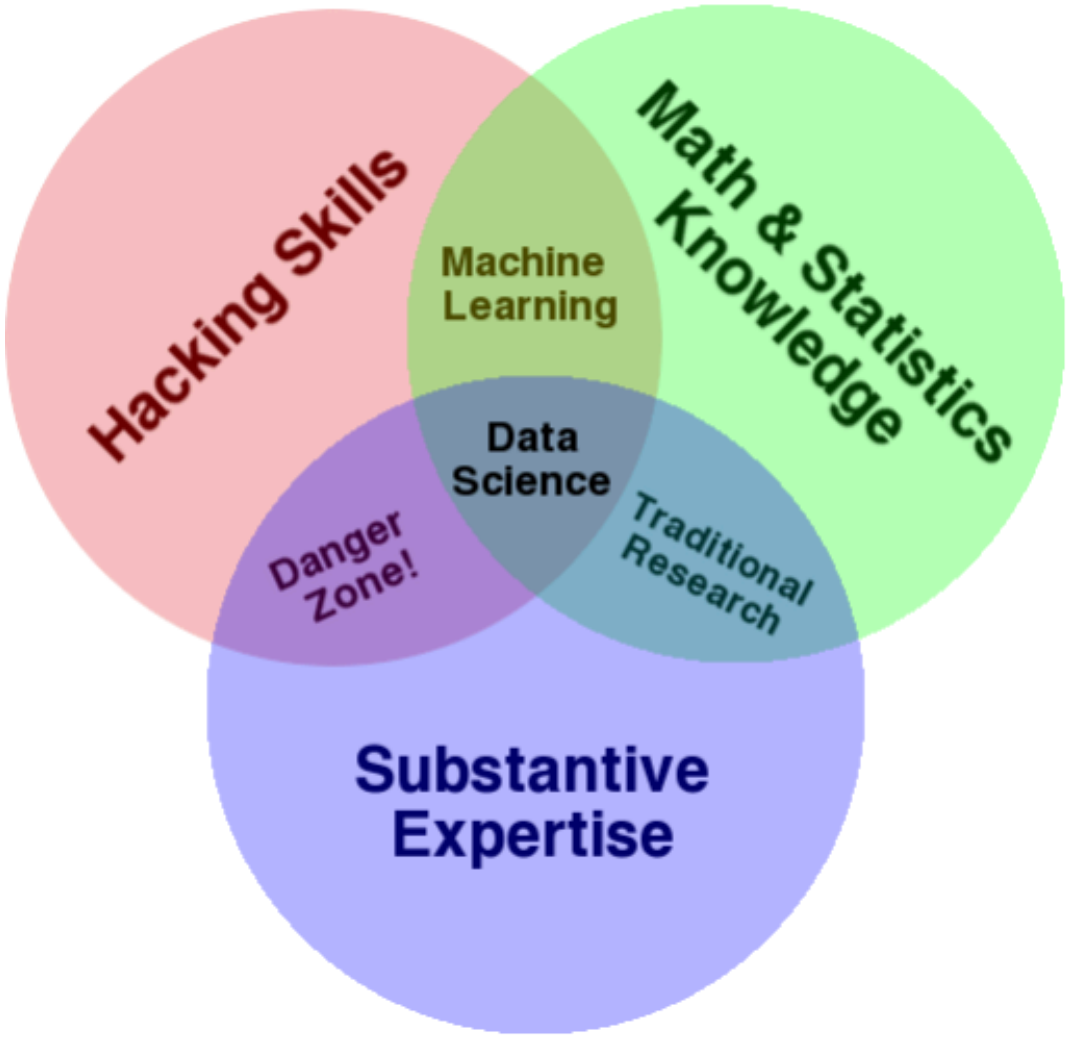
```
SELECT COUNT(*)
FROM TAS
```

- What is the impact of dirty data?

Where does the “dirt” come from?



Data Science is intersectional



Why data is "dirty": the Statistics view

- There is a stochastic process that produces data
- We want to model ideal samples of that process, but in practice we have non-ideal samples
  - Distortion – some samples are corrupted by a process
  - Selection Bias - likelihood of a sample depends on its value
  - Left and right censorship - users come and go from our scrutiny
  - Dependence – samples are supposed to be independent, but are not (e.g. social networks)
- You can add new models for each type of imperfection, but you can't model everything
- What's the best trade-off between accuracy and simplicity?

Why data is “dirty”: the Database view

- I got my hands on this data set
- Some of the values are missing, corrupted, wrong, duplicated
- Results are absolute (relational model)
- You get a better answer by improving the quality of the values in your dataset

Why data is “dirty”: the domain experts view

- This data does not look right
- This answer doesn’t look right
- What happened?
- Domain experts have an implicit model of the data that they can test against…
- Domain experts can be very powerful in removing corrupted data but inherently “expensive”

Why data is “dirty”: the data scientists view

- All of the previous!

Dirty data problems

1) Parsing text into fields (separator issues)
2) Naming conventions (e.g., NYC vs New York)
3) Missing required field (e.g., key field)
4) Different representations (e.g., 2 vs Two)
5) Fields too long (get truncated)
6) Primary key violation (from un- to structured or during integration
7) Redundant Records (exact match or other)
8) Formatting issues – especially dates
9) Licensing issues/Privacy/ keep you from using the data as you would like?

Conventional Definition of Data quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Consistency
  - The data agrees with itself.
- Timeliness
  - The data is kept up to date.
  - Time consistency.

Challenges

- Difficult to measure
  - Accuracy and completeness are extremely difficult, perhaps impossible to measure.
- Context independent
  - No accounting for what is important.
  - E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.
- Incomplete
  - What about interpretability, accessibility, metadata, analysis, etc.
- Vague
  - The conventional definitions provide no guidance towards practical improvements of the data.

A modern definition for data quality

- We need a definition of data quality that:
  - Accounts for the intended use of the data
  - Leads to improvements in processes
  - Is measurable (we can define metrics to evaluate it)
- First, we need a better understanding of how and where data quality problems occur
  - The data quality continuum

Meaning of Data Quality

- There are many types of data, which have different uses and typical quality problems
  - Federated data
    - Data obtained by merging together multiple sources
  - High dimensional data
  - Descriptive data
    - Summarize description of statistical data properties
  - Longitudinal data
    - Sometimes referred to as panel data, track the same sample at different points in time
  - Streaming data
  - Web (scraped) data
  - Numeric vs. categorical vs. text data
- Different types of data require opportune methods to ensure quality
- Know your data/domain!

Meaning of Data Quality

- There are many uses of data
  - Operations
  - Aggregate analysis
  - Customer relations …
- Data Interpretation: the data is useless if we don't know all of the rules behind the data.
- Data Suitability : Can you get the answer from the available data?
  - Use of proxy data
  - Relevant data is missing

The Data Quality Continuum

Data and information is not static, it flows in a data collection and usage process

- Data gathering
- Data delivery
- Data storage
- Data integration
- Data retrieval
- Data mining/analysis

Data Gathering

- How does the data enter the system?
- Sources of problems:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (duplicates)
  - Approximations, surrogates
  - SW/HW constraints
  - Measurement or sensor errors

Data Gathering: Potential Solutions

Potential Solutions:

- Preemptive:
  - Process architecture (build in integrity checks)
  - Process management (reward accurate data entry, data sharing, data stewards)
  - Set reasonable default values
- Retrospective:
  - Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
  - Diagnostic focus (automated detection of glitches).

Data Delivery

- Destroying or mutilating information by inappropriate pre-processing
  - Inappropriate aggregation
  - Nulls converted to default values
- Loss of data:
  - Transmission problems
  - Buffer overflows
  - No checks

Data Delivery: Solutions

- Build reliable transmission protocols
  - Use a relay server
- Verification
  - Checksums, verification parser
  - Do the uploaded files fit an expected pattern?
- Interface agreements
  - Data quality commitment/contract from the data stream supplier

Data Storage

- Problems in physical storage
  - Can be an issue, but terabytes are cheap ☺
- Problems in logical storage
  - Poor metadata.
    - Data feeds are often derived from application programs or legacy data sources.
- Inappropriate data models
    - Missing timestamps, incorrect normalization, etc.
    - Default values can help

Data Storage: Solutions

- Metadata
  - Document and publish data specifications.
- Planning
  - Assume that everything bad will happen.
  - Can be very difficult and time-expensive.
- Data exploration
  - Use data browsing and data mining tools to examine the data.
    - Does it meet the specifications you assumed?
    - Has something changed?

Data Quality Constraints

- Many data quality problems can be captured by static constraints based on the schema.
  - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to problems in workflow, and can be captured by dynamic constraints
  - E.g., orders above $200 are processed by Biller 2
- The constraints follow an 80-20 rule
  - A few constraints capture most cases, thousands of constraints to capture the last few cases.

Data Quality Metrics

- We want a measurable quantity
  - Indicates what is wrong and how to improve
  - Realize that Data Quality is a messy problem, no set of numbers will be perfect
- Types of metrics
  - Static vs. dynamic constraints
  - Operational vs. diagnostic
- Metrics should be directionally correct with an improvement in use of the data.
- A very large number metrics are possible
  - Choose the most important ones
  - Chose metrics appropriate for your project

Examples of Data Quality Metrics

- Conformance to schema
  - Evaluate constraints on a snapshot of the database
- Conformance to business rules
  - Evaluate constraints on changes in the database
- Accuracy
  - Perform inventory (expensive)
  - Use proxy (track complaints)
  - Audit samples
- Accessibility
- Interpretability
- Successful completion of end-to-end process

Technical approaches

- We need a multi-disciplinary approach to attack data quality problems
  - No one approach solves all problem
- Process management
  - Ensure proper procedures
- Statistics
  - Focus on analysis: find and repair anomalies in data.
- Database
  - Focus on relationships: ensure consistency.
- Metadata / domain expertise
  - What does it mean? Interpretation
  - Set reasonable default values

Example: duplicate detections via string comparison

- Comparing strings values for some attributes may be useful to detect error such as typos/erroneous formatting/repetitions (or possible repartition)
- We set a similarity threshold such that strings which are similar enough are considered the same
- CAREFUL: this is very domain-dependent!
  - In some cases even minor differences are significant
    - E.g., BannerID, Phone numbers, ZIP codes, SSN,...
- Multiple possible notions of string similarity

String similarity: Edit Distance

- Minimum number of edits (insertions, deletions, substitutions) needed to transform String A into String B
- Can be computed efficiently using a dynamic algorithm

$$d_{i0} = \sum_{k=1}^{i} w_{\text{del}}(b_k), \qquad \text{for } 1 \le i \le m$$

$$d_{0j} = \sum_{k=1}^{j} w_{\text{ins}}(a_k), \qquad \text{for } 1 \le j \le n$$

$$d_{ij} = \begin{cases} d_{i-1,j-1} & \text{for } a_j = b_i \\ \min \begin{cases} d_{i-1,j} + w_{\text{del}}(b_i) \\ d_{i,j-1} + w_{\text{ins}}(a_j) \\ d_{i-1,j-1} + w_{\text{sub}}(a_j, b_i) \end{cases} & \text{for } a_j \ne b_i \end{cases} \qquad \text{for } 1 \le i \le m, 1 \le j \le n.$$

- Example:
  - 148th Ave NE, Redmond, WA
  - NE 148th Ave, Redmond, WA
  - edit distance =?

Jaccard distance

- The Jaccard distance is a measure of similarity between sets
- Given sets A,B we have $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$
- Can be used to compare strings if these are seen as set of words or characters
  - As set of words

```
148th Ave NE, Redmond, WA
140th Ave NE, Redmond, WA
```

J(A,B)=4/6

  - As sets of characters

```
148th Ave NE, Redmond, WA
140th Ave NE, Redmond, WA
```

J(A,B) = 16/18

- Captures different information than edit distance

```
148th Ave NE, Redmond, WA
NE 148th Ave, Redmond, WA
```

- Edit distance = 6
- J(A,B) = 1 (i.e., no difference)

- Symmetric: J(A,B)=J(B,A)

Quiz 1

- What is the Jaccard similarity of these two strings (w.r.t. words)?

```
A = iPad Two 16GB WiFi White
B = iPad 2nd generation 16GB WiFi White
```

Weighted Jaccard Similarity

- We can assign weights to words and then compare them using the Jaccard approach:
  - $J(A,B) = \frac{\sum_{w_i \in A \cap B} w_i}{\sum_{w_j \in A \cup B} w_j}$
  - Weights can be used to highlight the relative importance of some words
- E.g.:
  - Unweighted

    A = Michigan State University
    B = Ohio State University

    J(A,B)=2/4=0.5
  - Weighted

    A = Michigan (3) State (1) University (1)
    B = Ohio (3) State (1) University (1)

    $J_w(A,B)=2/8 = 0.25$

Similarity between vectors

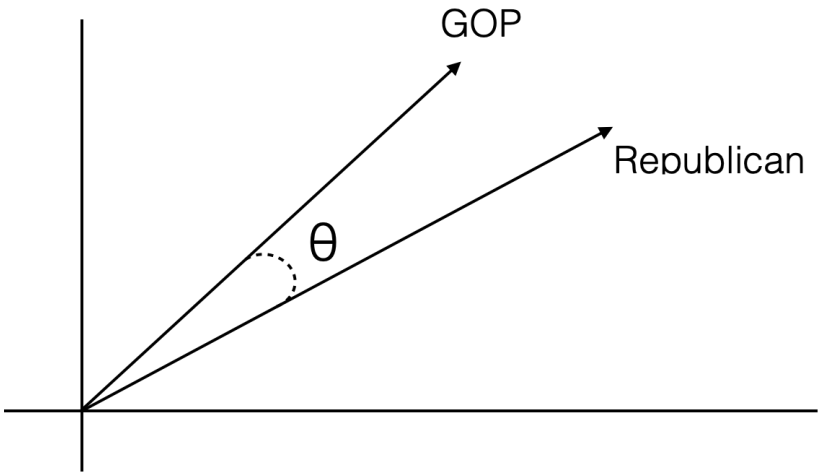
- How can we compare two vectors of values?
- Cosine similarity of vector A and B

$$\cos(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

- Cosine similarity takes value between 1 (identical vectors), to 0 maximally different vectors
- Can be useful to set a threshold to distinguish quasi-duplicates and remove them, or to normalize

Cosine similarity

	Senator	Washington	announced	party	primary	chairman
GOP	1002	41	502	700	400	3
Republican	800	35	521	698	423	10



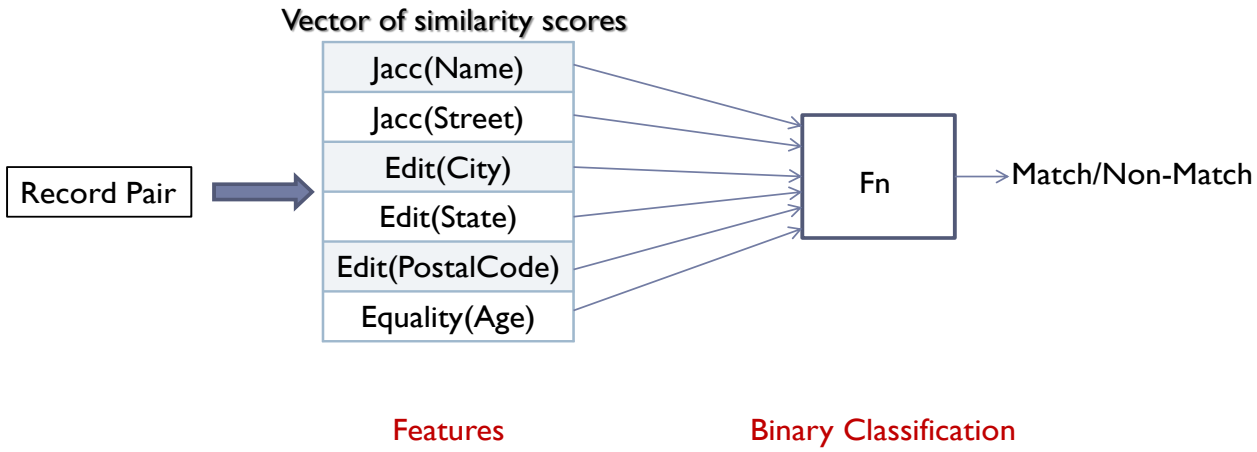
- Can be used to compare strings seen as vectors where each feature corresponds to a character and the value to the number of occurrences

How to set threshold

- How do we set opportune thresholds?
  - If threshold is too low many duplicates may not be detected
  - If threshold is too high different records may be erroneously considered duplicates
- Domain experts may provide the such values and adapt them to the specific case

How to set threshold

- Machine learning can be used to set these parameters
  - Use a subsample of the records in the database
  - Compute the values that optimize duplicate detection
  - Use these values for the rest of the database
  - Criteria can be arbitrarily complex
    - Risk of overfitting
  - Can be very effective but must handled with care!



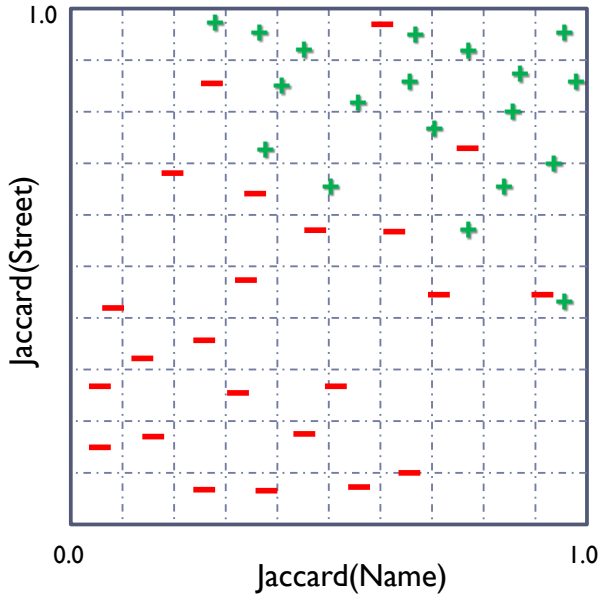
How to set threshold

- Machine learning can be used to set these parameters
  - Use a subsample of the records in the database
  - Compute the values that optimize duplicate detection
  - Use these values for the rest of the database
  - Criteria can be arbitrarily complex
    - Risk of overfitting
  - Can be very effective but must handled with care!

Bob Wilson	345 Broadway	Seattle	Washington	98101	19	Match
Robert Wilson	345 Broadway St	Seattle	WA	98101	19	
B Wilson	123 Broadway	Boise	Idaho	83712	19	Non-Match
Robert Wilson	345 Broadway St	Seattle	WA	98101	19	
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30	Match
M Jones	245 Third Street	Redmond	NULL	98052	299	
Mary Jones	245 3rd St	Redmond	WA	98052-1234	30	Non-Match
Robert Wilson	345 Broadway St	Seattle	WA	98101	19	

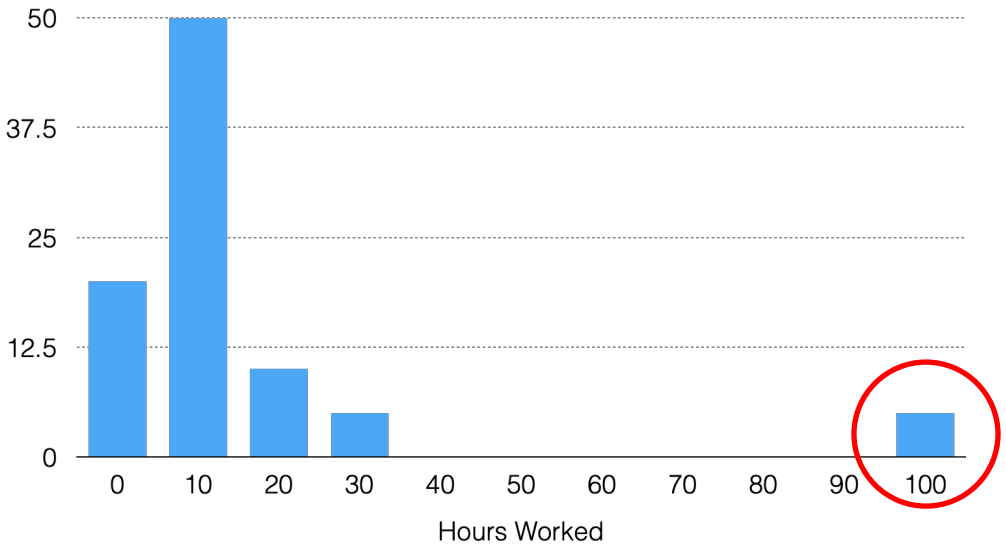
How to set threshold

- Machine learning can be used to set these parameters
  - Use a subsample of the records in the database
  - Compute the values that optimize duplicate detection
  - Use these values for the rest of the database
  - Criteria can be arbitrarily complex
    - Risk of overfitting
  - Can be very effective but must handled with care!



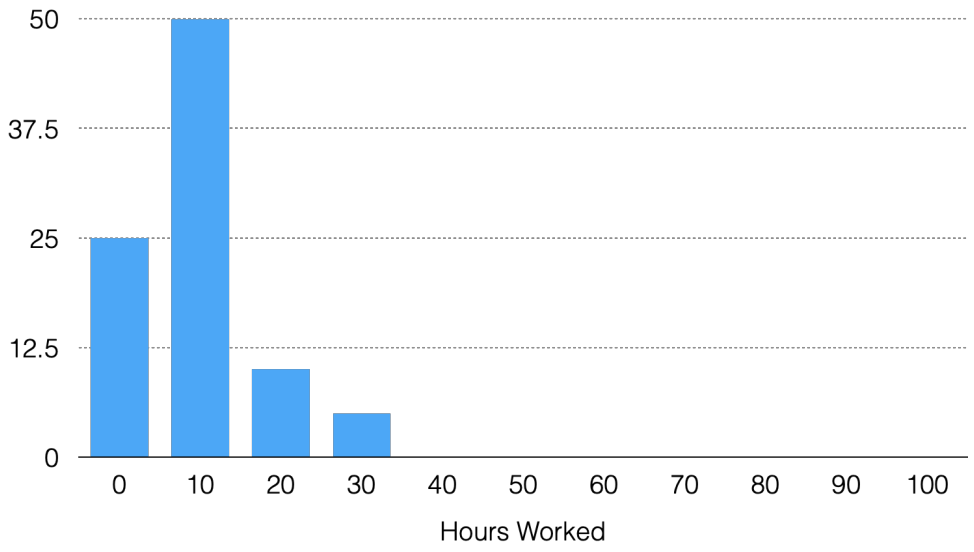
Outliers

- Data points that exhibit extreme behavior with respect to the majority of observed data, or with respect to the priori belief of the statistical properties of the data
  - May be due to errors in data acquisition/manipulation
  - May be due to some random sporadic occurrence
  - May be representative of a rare but still valid phenomenon



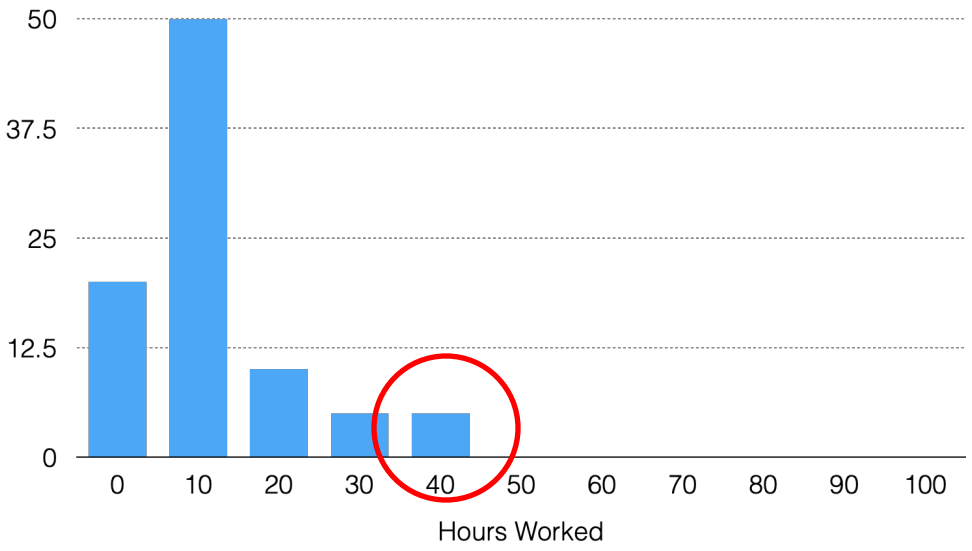
How to handle outliers

- Always understand the nature of the outlier!
  - Domain experts can be essential
- Depending on the confidence according to which we can determinate the nature of an outlier, different actions may appropriate
  - Delete



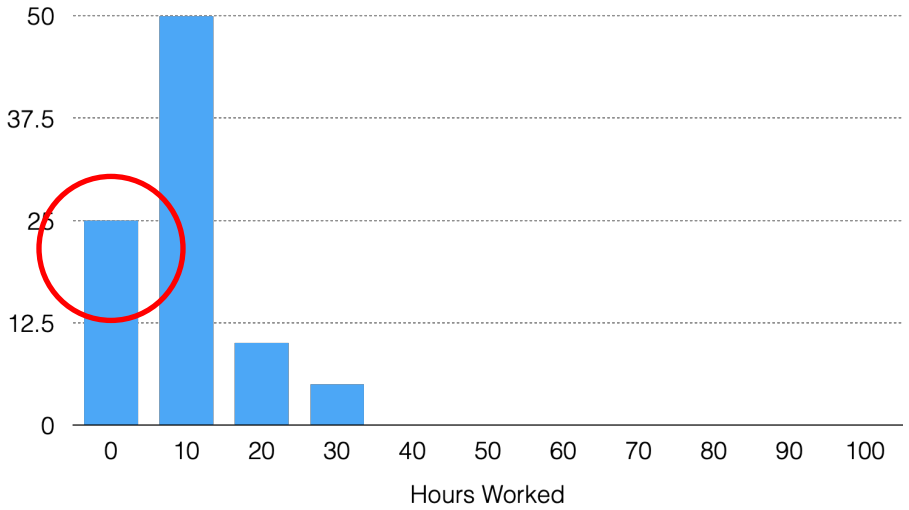
How to handle outliers

- Always understand the nature of the outlier!
  - Domain experts can be essential
- Depending on the confidence according to which we can determinate the nature of an outlier, different actions may appropriate
  - Delete
  - Set to a default value



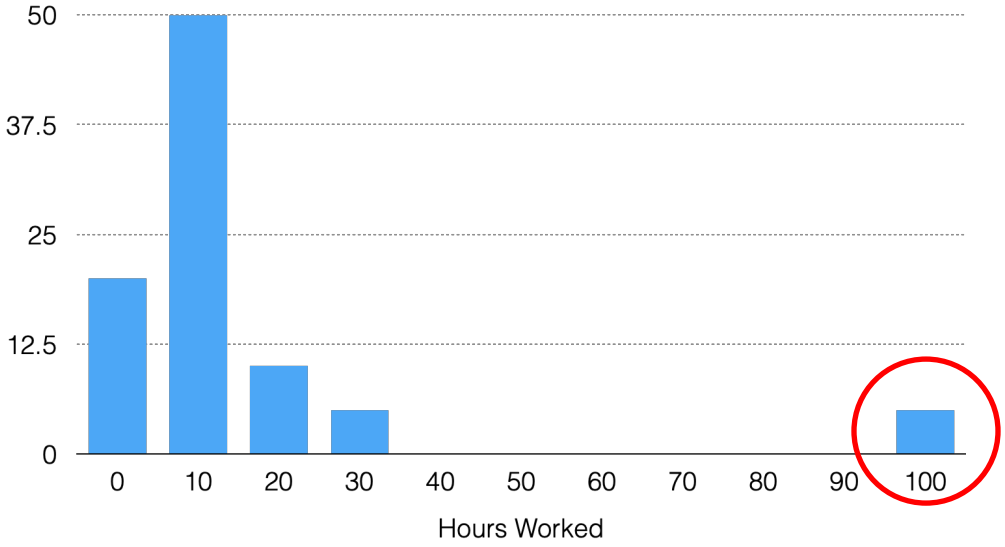
How to handle outliers

- Always understand the nature of the outlier!
  - Domain experts can be essential
- Depending on the confidence according to which we can determinate the nature of an outlier, different actions may appropriate
  - Delete
  - Set to a default value

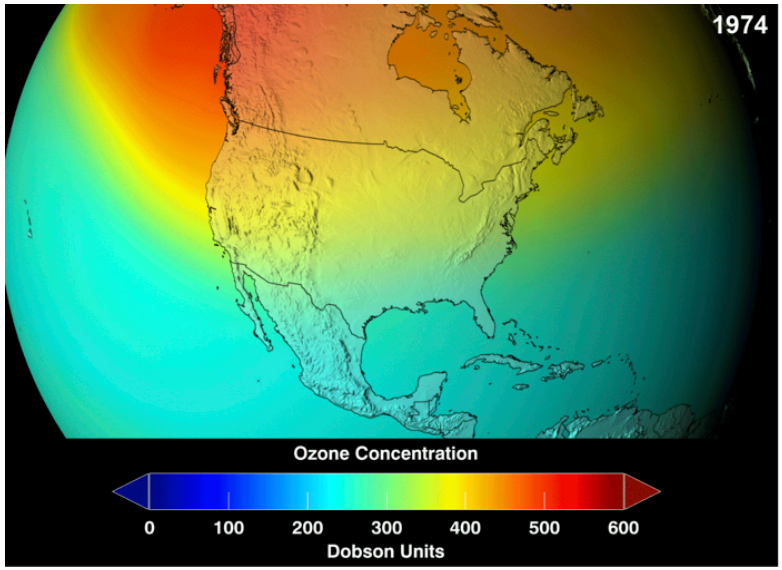


How to handle outliers

- Always understand the nature of the outlier!
  - Domain experts can be essential
- Depending on the confidence according to which we can determinate the nature of an outlier, different actions may appropriate
  - Delete
  - Set to a default value
  - Take it at face value



Outliers and rare occurrences



The discovery of the Antarctic "ozone hole" by British Antarctic Survey scientists Farman, Gardiner and Shanklin…came as a shock to the scientific community…[The data] were initially rejected as unreasonable by data quality control algorithms (they were filtered out as errors since the values were unexpectedly low); the ozone hole was detected only in satellite data when the raw data was reprocessed following evidence of ozone depletion in in situ observations. When the software was rerun without the flags, the ozone hole was seen as far back as 1976.

https://en.wikipedia.org/wiki/Ozone_depletion#Antarctic_ozone_hole

Outliers and rare occurrences

- Handling of outliers is very challenging!
- Hard to distinguish errors from rare but still "valid" phenomena
- Rare signals are often of high interest
  - We do not want our preprocessing to eliminate them
- Heuristics and domain experts may be useful
  - Still no guarantee
  - Hard to have expertise on rare, or unexplored, phenomena
- Statistics and Machine Learning provide some powerful tools
  - Still, assumptions are needed
- Know and observe your data!