Data Cleaning

MIE223

Winter 2025

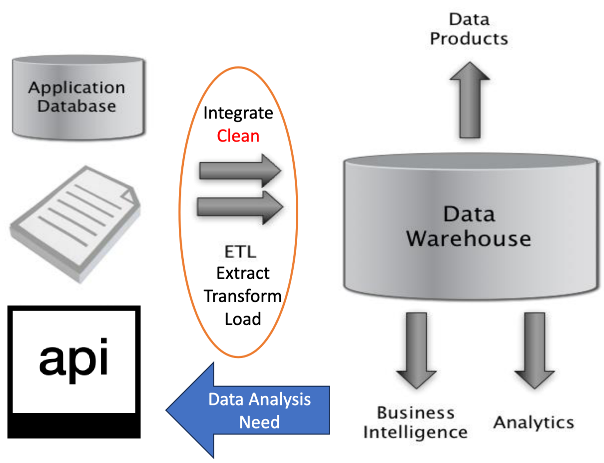
1 Why Clean Data?

1.1 Example

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View	$210
Intl. Business Machines	Armonk, NY	$200
Microsoft	Redmond, WA	$250
Sally's Lemonade Stand	Alameda,CA	$260,000

- Apple: Missing Data
- IBM: Entity Resolution / Unnormalized Naming
- Sally's Lemonade Stand: Unit Mismatch

1.2 Where is Data Cleaning in the Organization?



- Data comes in all shapres and forms
- ETL: Extract, Transform, Load
- Data science exists in business intelligence and analytics
- Data engineers work in application database

1.3 Cost of Bad Quality Data Over Time

- Cost for preventing bad data: \$1
- Cost for correcting bad data: \$10
- Cost for fixing problem resulting from bad data: \$100

2 Perspectives on Data Science and Data Cleaning

2.1 Data Science is an Intersectional Discipline

- Computer Science
- Math and Statistics
- Domain Knowledge

2.2 Dirty Data

- The Statistics View:
  - There is a process that produces data
  - Any dataset is a sample of the output of that process
  - Results are probabilistic
  - You can correct bias in your sample
- 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
- The Domain Expert's View
  - This Data Doesn't look right
  - This Answer Doesn't look right
  - What happened?
- The data scientist's view is a combination of the above.

Note 1. A Data scientist knows more programming than a statistician and more statistics than a programmer. A Data scientist intimately understands their domain!

2.3 Data Quality Problems

- Data is dirty on its own
  - Human collection or data entry (generally lazy and want to see their children)
- Data sets are clean but integration (i.e., combining them) screws them up (such as different units of measurement)
- Data sets are clean but suffer "bit rot" (lose data from cutting)
  - Constant data transformations introduce errors, noise, data loss
  - Meanings of labels change over time
- Any combination of the above... and much, much more!

2.4 Some Data Issues you will Encounter

1. Parsing text into fields (separator issues)
2. Naming conventions and entity resolution D: NYC vs New York
3. Missing required field (birthdate)
4. Gaps in time series
5. Different representations ("2" vs 2), Unicode lookalikes
6. Fields too long (get truncated)
7. Mixed data types (feet, meters)
8. Redundant Records (exact match or other)
9. Formatting issues – especially dates
10. Licensing issues/privacy/cost prevent access to all data

2.5 Key Types of Data Quality Issues

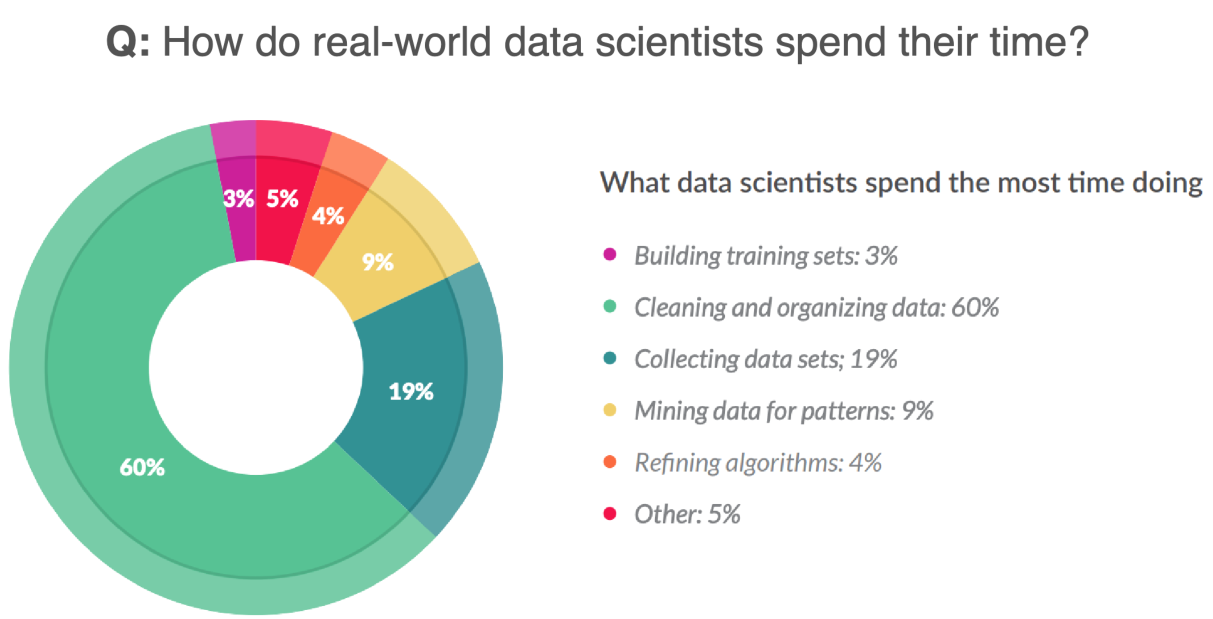
- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=" "
- **noisy**: containing errors or outliers
  - e.g., Salary="-10"
- **inconsistent**: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
  - e.g., discrepancy between duplicate records
- **redundancy**: duplicated content

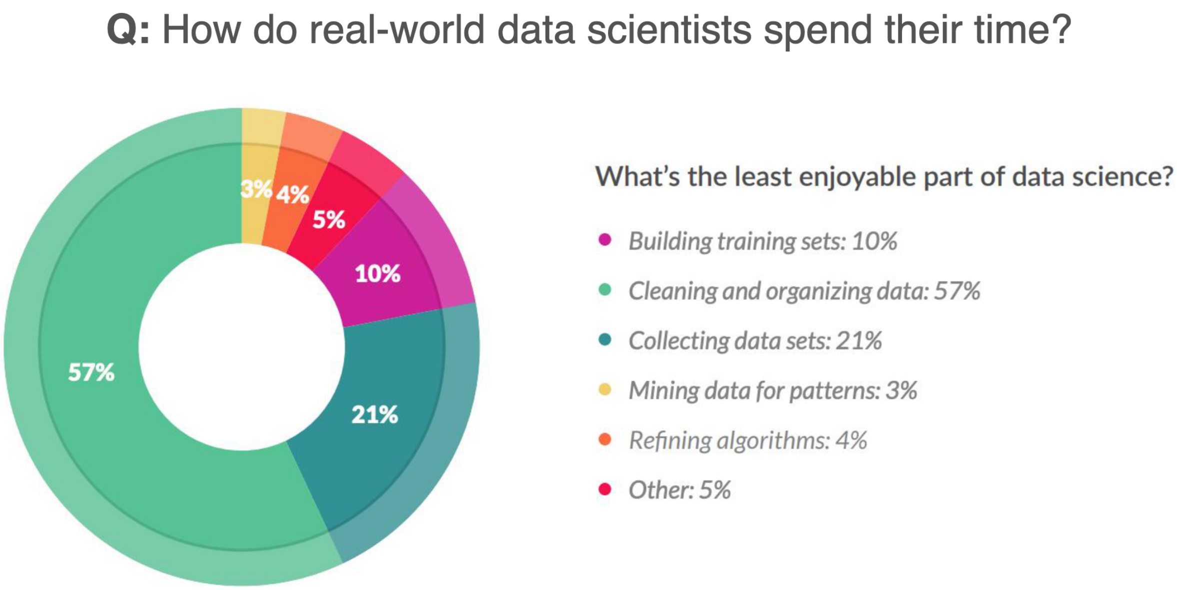
2.6 Why Is Data Dirty?

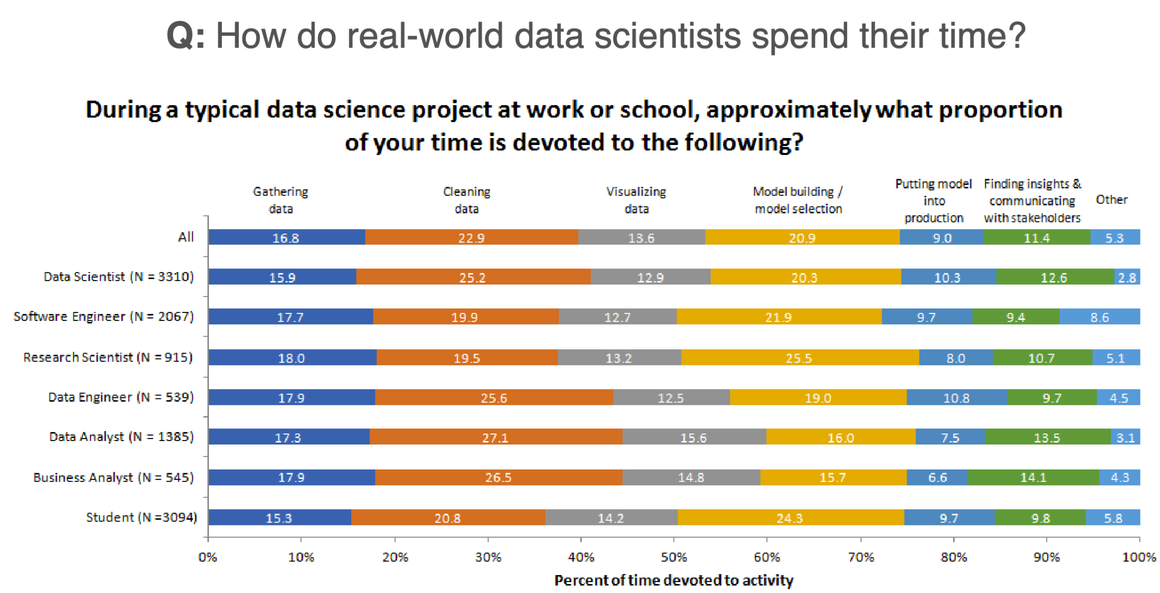
- Incomplete data may come from
  - "Not applicable" data value when collected
  - Changes in the data collected over time
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments, undocumented API changes
  - Human or computer error at data entry, UI changes!
  - Errors in data transmission, discretization, conversion (losing precision)
  - Typing errors (meters, feet, km mixed in same column)
- Inconsistent data may come from
  - Different data sources (data integration)
  - Changes in data collection practices over time
- Redundancy
  - Human error (could not find previous record), data integration

3 The Role of Data Cleaning in Data Science

3.1 Data Science in the Real World







3.2 Data Cleaning Makes Everything Okay?

Note 2. "The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning." - National Center for Atmospheric Research

The data was rejected as unreasonable by data quality control algorithms.

4 Principles of Data Cleaning

4.1 Data Cleaning Tasks

- Data Summary
- Consider filling in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy

5 Data Cleaning: Data Summary

5.1 Data Summary Procedure

- Look at data types of columns
  - Beware of ”Object” columns, indicator that you’ve mixed types (e.g., ”2” and 2)
- For high cardinality strings and categorical (e.g,. ”country”) columns
  - How many unique elements are there?
  - Are any missing?
  - Count frequencies of string
    * Look at top-10 most frequent and least frequent values
    * Look at strings near the mean and median frequency
- For low cardinality strings and categorical (e.g,. “province”) columns
  - Look at frequency of each unique value
- For ordinal numeric (integers) or floating point values
  - Compute summary statistics
  - View a histogram (but before you do this, hypothesize what you will see)

5.2 Summary Statistics and for Each Variable (Column)

- Range
  - Minimum
  - Maximum
- Central Tendency
  - Mean
  - Median
  - Mode

    * Value that occurs most frequently in discrete data
    * Value with highest probability in continuous data
    * Informally: # peaks in numeric data – unimodal, bimodal, trimodal, multimodal, ...

- If Data is non-numeric, consider frequencies below (e.g., mean frequency of jobs in "Job")
- If Data is a string, consider sorting it and looking at nearby values

5.3 Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  - Inter-quartile range: IQR = Q3 – Q1
  - Five number summary: min, Q1, M, Q3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers (min / max), and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (sample: s, population: sigma)
  - Variance: (algebraic, scalable computation)

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

  -