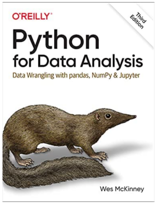
7.2: Data Wrangling

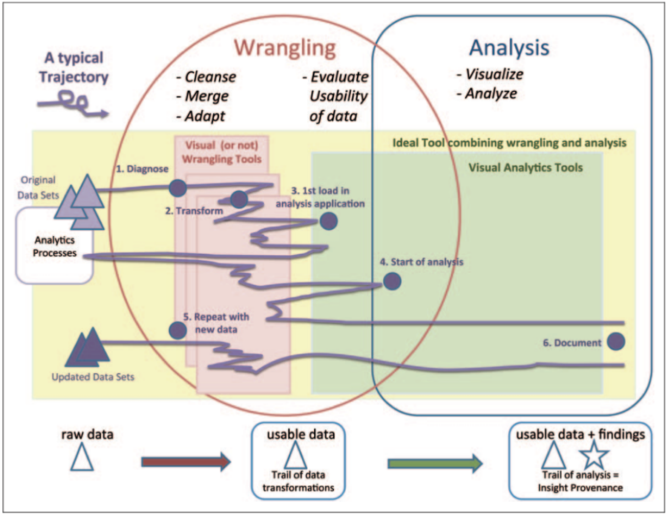


- **Instructor**: Dr. GP Saggese - gsaggese@umd.edu
- **Reference**
  - Pandas tutorial
  - Class project
  - Web
    - https://pandas.pydata.org
    - Onslaught of free resources
  - Mastery
    - https://wesmckinney.com/book
    - Read cover-to-cover and execute all examples 2-3x time to really *master*

Overview

- **Data wrangling**
  - Aka “data preparation”, “data munging”, “data curation”
  - Structure data for analysis
  - Majority of time (80-90%) spent here
- **Key steps**
  - *Scraping*: extract info from sources (e.g., webpages)
  - *Data cleaning*: remove inconsistencies/errors
  - *Data transformation*: structure data correctly
  - *Data integration*: combine data from multiple sources
  - *Information extraction*: extract structured info from unstructured/text sources

Overview

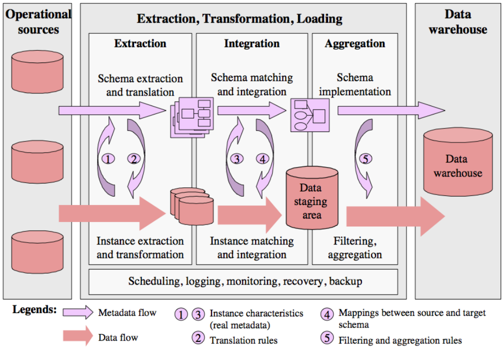


igure 1. The iterative process of wrangling and analysis. One or more initial data sets may be used and new versions nay come later. The wrangling and analysis phases overlap. While wrangling tools tend to be separated from the visual inalysis tools, the ideal system would provide integrated tools (light yellow). The purple line illustrates a typical iterative rocess with multiple back and forth steps. Much wrangling may need to take place before the data can be loaded within isualization and analysis tools, which typically immediately reveals new problems with the data. Wrangling might take lace at all the stages of analysis as users sort out interesting insights from dirty data, or new data become available or eeded. At the bottom we illustrate how the data evolves from raw data to usable data that leads to new insights.

Overview

- Data wrangling problems are hard to formalize, with little research, e.g.,
  - Data cleaning: statistics, outlier detection, imputation
  - Data transformation: structuring data (e.g., tidy data)
  - Information extraction: feature computation, domain-specific
- Other aspects studied in depth, e.g.,
  - Schema mapping
  - Data integration
- In ETL process
  - Data extraction is E step
  - Data wrangling is T step

Overview



Operational sources
Extraction, Transformation, Loading
Data warehouse
Extraction
Integration
Aggregation
Schema extraction and translation
Schema matching and integration
Schema implementation
1
2
3
4
5
Data staging area
Data warehouse
Instance extraction and transformation
Instance matching and integration
Filtering, aggregation
Scheduling, logging, monitoring, recovery, backup
Legends:
Metadata flow
Data flow
1 3 Instance characteristics (real metadata)
2 Translation rules
4 Mappings between source and target schema
5 Filtering and aggregation rules

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

- From Data Cleaning: Problems and Current Approaches
  - Paper old: data from structured sources
  - Today unstructured/semi-structured equally important

SCIENCE ACADEMY

Data Extraction

- Data resides in various sources
  - Files (CSV, JSON, XML)
  - Databases
  - Spreadsheets
  - AWS S3 buckets
  -
  - Analytical tools support data import through adapters
- Web scraping
  - Use APIs or scrape data explicitly
  - Scraping is challenging
    - Fragile
    - Throttling
    - Cat-and-mouse with websites
  - Set up pipelines for periodic scraping
  - Tools for automated scraping
    - E.g., import.io, portia,

Tidy Data

- Tidy data, Wickham, 2014
  - Each variable forms a column
  - Each observation forms a row
- Wide vs long format

Wide format

type	date	clicks	conversions	impressions
0	2020-01-01	1.0	NaN	18.0
1	2020-01-02	2.0	NaN	19.0
2	2020-01-03	1.0	1.0	14.0
3	2020-01-04	NaN	NaN	5.0
4	2020-01-05	1.0	NaN	8.0
5	2020-01-06	1.0	1.0	15.0
6	2020-01-07	2.0	NaN	8.0

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

“Messy” data

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Tidy data

	date	type	count
0	2020-01-01	impressions	18.0
1	2020-01-02	impressions	19.0

SCIENCE ACADEMY

Data Quality Problems

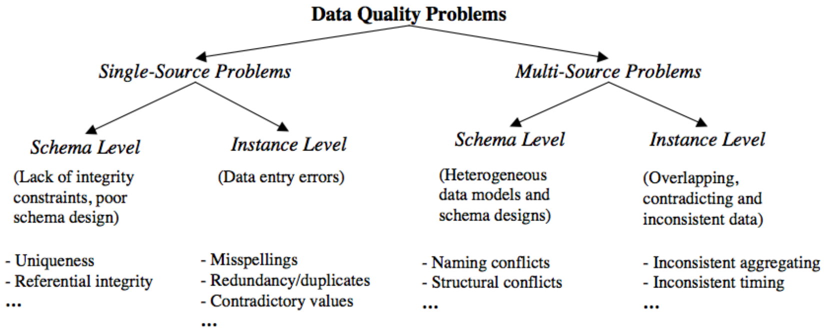
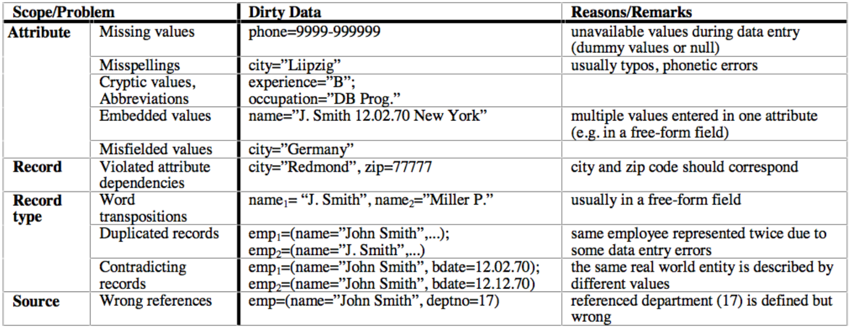


Figure 2. Classification of data quality problems in data sources

Single-Source Problems

- Depends largely on source
- Databases enforce constraints
- Data from spreadsheets is often "clean"
    - Schema exists
- Logs are messy
- Data from web-pages is messier
- Types of problems:
    - Ill-formatted data
    - Missing/illegal values, misspellings, wrong fields, extraction issues
    - Duplicated records, contradicting info, referential integrity violations
    - Unclear default/missing values
    - Evolving schemas/classification schemes (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

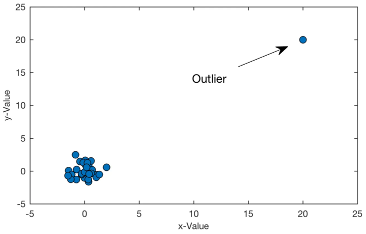
- Different data sources:
    - Developed separately
    - Maintained by different people
    - Stored in different systems
- Schema mapping / transformation:
    - Map information across sources
    - Naming conflicts: same name for different objects, different names for same objects
    - Structural conflicts: different representations across sources
- Entity resolution:
    - Match entities across sources
- Data quality issues:
    - Contradicting information
    - Mismatched information

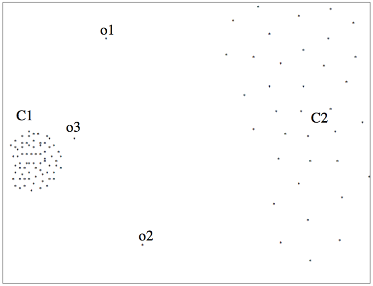
Data Cleaning: Outlier Detection

- Quantitative Data Cleaning for Large Databases, Hellerstein, 2008
  - Focuses on numerical data (integers/floats measuring quantities)
- Sources of errors in data
  - Data entry errors: arbitrary values entered
  - Measurement errors: sensor data inaccuracies
  - Distillation errors: processing and summarization issues
  - Data integration errors: inconsistencies across combined sources

Univariate Outlier Detection

- A set of values characterized by metrics:
  - Center (e.g., mean)
  - Dispersion (e.g., standard deviation)
  - Higher momenta (e.g., skew, kurtosis)
- Use statistics to identify outliers
  - Watch for "masking": one extreme outlier may alter metrics, masking others
  - Robust statistics: minimize effect of corrupted data
  - Robust center metrics:
    - Median
    - k%-trimmed mean (discard lowest and highest k% values)
  - Robust dispersion:
    - Median absolute deviation (MAD)
    - Median distance from median value





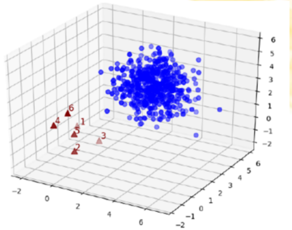
SCIENCE ACADEMY

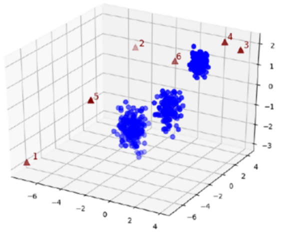
Outlier Detection

- For Gaussian data
  - Data points 1.4826x MAD from median
  - Eyeball data (e.g., plot histogram) to confirm
- For non-Gaussian data
  - Estimate generating distribution (parametric)
  - Distance-based methods: find points with few neighbors
  - Density-based methods:
    - Define *density* as average distance to k nearest neighbors
    - *Relative density* = density of node/average density of neighbors
    - Use relative density to identify outliers
- Techniques break down as data dimensionality increases
  - *Curse of dimensionality*
    - Need O(eˆn) points with n dimensions to estimate
    - "In high dimensional spaces, data is sparse"
  - Project data into lower-dimensional space to find outliers
    - Not straightforward
- Wikipedia article on Outliers

Multivariate Outliers

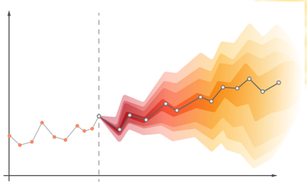
- One set of techniques *multivariate Gaussian distribution* data
  - Defined by *mean* μ and *covariance matrix* Σ
- Mean/covariance not robust (sensitive to outliers)
- Robust statistics analogous to univariate case
- Iterative approach
  - Mahalanobis distance: square root of $(x-\mu)'\Sigma^{-1}(x-\mu)$
  - Measures distance of point x from multivariate normal distribution
  - Outliers: points too far away by Mahalanobis distance
  - Remove outliers
  - Recompute mean and covariance
- Often data volume too large
  - Use approximation techniques
- Try different techniques based on data





Time Series Outliers

- Often data is in the form of a time series
- A **time series** is a sequence of data points recorded at regular intervals
  - Stock prices
  - Sales revenue
  - Website traffic
  - Inventory levels
  - Energy consumption
  - Market demand
  - Social media engagement
  - Hourly energy usage
  - Customer satisfaction ratings
  - Weekly retail foot traffic
  - ...

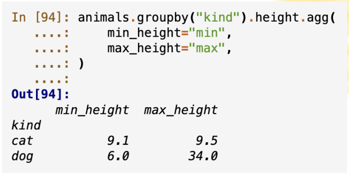


- Rich literature on *forecasting* in time series data
- Use historical patterns to flag outliers
  - Rolling MAD (median absolute variation)

SCIENCE ACADEMY

Split-Apply-Combine

- The Split-Apply-Combine Strategy for Data Analysis, Wickam, 2011
- Common data analysis pattern
  - Split: break data into smaller pieces
  - Apply: operate on each piece independently
  - Combine: combine pieces back together
- Pros
  - Code is compact
  - Easy to parallelize
- E.g.,
  - group-wise ranking
  - group vars (sums, means, counts)
  - create new models per group
- Supported by many languages
  - Pandas
  - SQL GROUP BY operator
  - Map-Reduce



```
In [94]: animals.groupby("kind").height.agg(
   ....:     min_height="min",
   ....:     max_height="max",
   ....: )
   ....:
Out[94]:
      min_height  max_height
kind
cat          9.1         9.5
dog          6.0        34.0
```

SCIENCE ACADEMY