DATA PREPROCESSING

PLEASE SIT IN YOUR PROJECT GROUPS.

1. INTRODUCTION

2. VISUAL ENCODING

3. BASIC CHART TYPES

Basics

4. INTERACTION

5. VISUALIZATION DESIGN

6. DATA PREPROCESSING

Visualization Building Blocks & Processes

7. RECAP 1st Half

8. MULTIVARIATE DATAVIS

9. TEMPORAL DATAVIS

10. GEOSPATIAL DATAVIS

11. GRAPH DATAVIS

Visualization Techniques

12. 3D DATAVIS

13. VISUAL ANALYTICS

Visualization Applications

14. RECAP 2nd Half

WHY TALK ABOUT PREPROCESSING? 1/3

Issue 1: Data Characteristics are Unknown

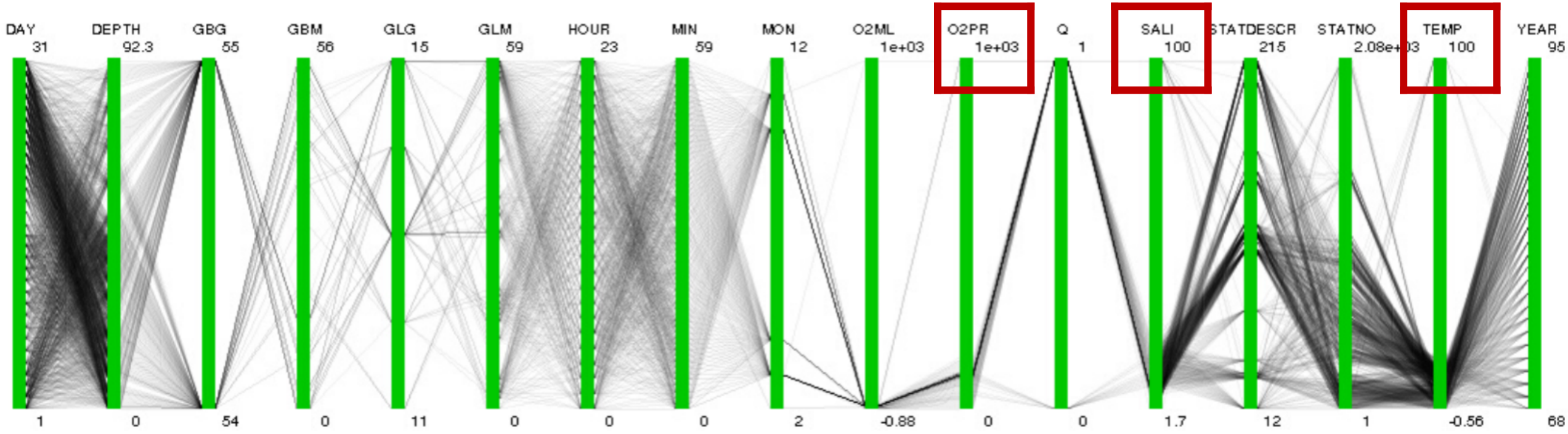


Image Source: [Schulz et al. 2017]

WHY TALK ABOUT PREPROCESSING? 2/3

Issue 2: Poor Data Quality

Data were entered by humans

Human data entry is such a common problem that symptoms of it are mentioned in at least 10 of the other issues described here. There is no worse way to screw up data than to let a single human type it in, without validation. For example, I once acquired the complete dog licensing database for Cook County, Illinois. Instead of requiring the person registering their dog to choose a breed from a list, the creators of the system had simply given them a text field to type into. As a result this database contained at least 250 spellings of `Chihuahua`. Even with the best tools available, data this messy can't be saved. They are effectively meaningless. This is not that important with dog data, but you don't want it happening with wounded soldiers or stock tickers. Beware human-entered data.

Source: The Quartz Guide to Bad Data

WHY TALK ABOUT PREPROCESSING? 3/3

Issue 3: Too many data

Image Source: [Zheng et al. 2021]

Crimes in Philadephia (2006-2013): Full dataset on the left (700k data points),
Sample on the right (5,3k data points)

OVERVIEW

1. **Data Profiling**, e.g. -
   - Determine access modalities
   - Establish / validate data characteristics

2. **Data Wrangling**, e.g. -
   - Check data for correctness and fix data where possible
   - Fuse data from multiple sources

3. **Data Transformation / Reduction**, e.g. -
   - Sampling data
   - Aggregating data

3 Stages of Data Preprocessing

Adapted from [Kandel et al. 2011]

DATA PROFILING

DATA PROFILING

Data profiling := the process of diagnosing a new or otherwise unknown dataset for its access modalities and data space characteristics

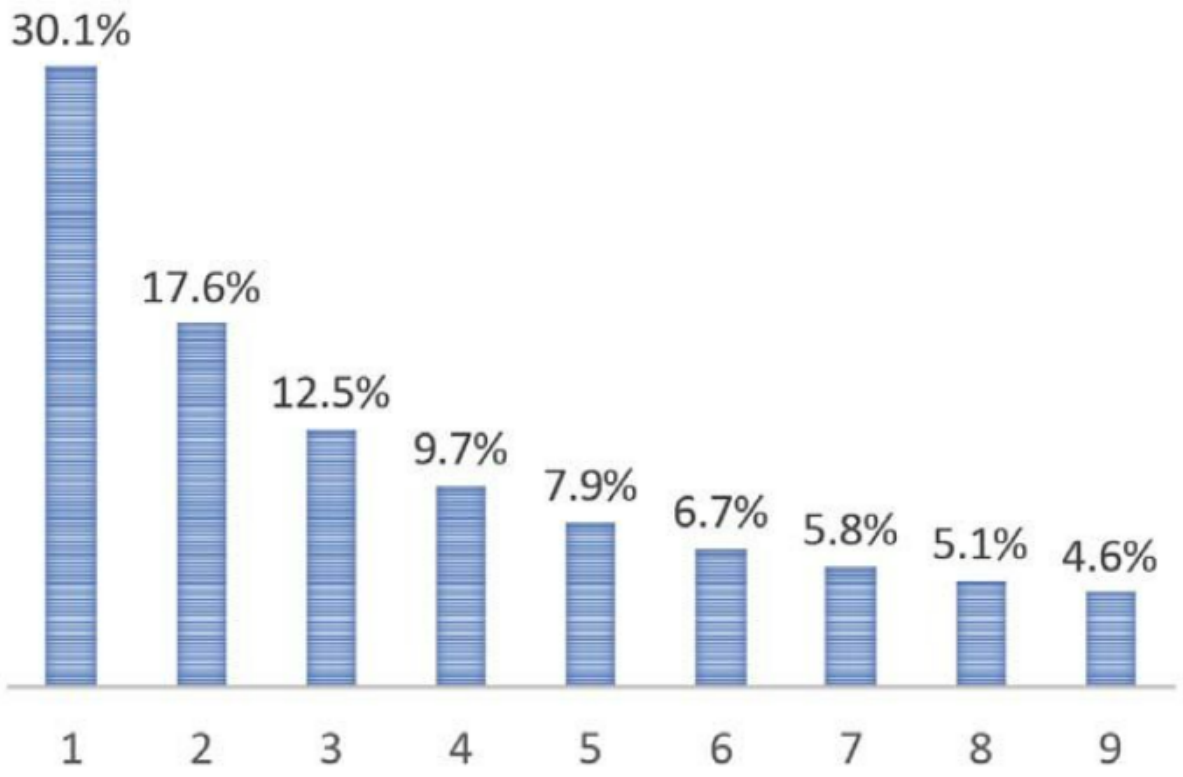
Access modalities: how often, how much, how detailed data can be retrieved (GDPR!)

- Different character encodings (ASCII vs. UTF-8 vs. ISO-8859 …)
- Different number formats (10,000.00 vs. 10000,00 vs. 1E+4 …)
- Different date formats (09-NOV-2011 vs. 09.11.2011 vs. 11/09/11…)
- Different time formats (16.20 vs. 4.20pm vs. 04:20:00 …)
- Different time zones / standards (UTC vs. GMT vs. EPOCH …)
- Different coordinate formats (DMS vs. Decimal DMS vs. UTM …)

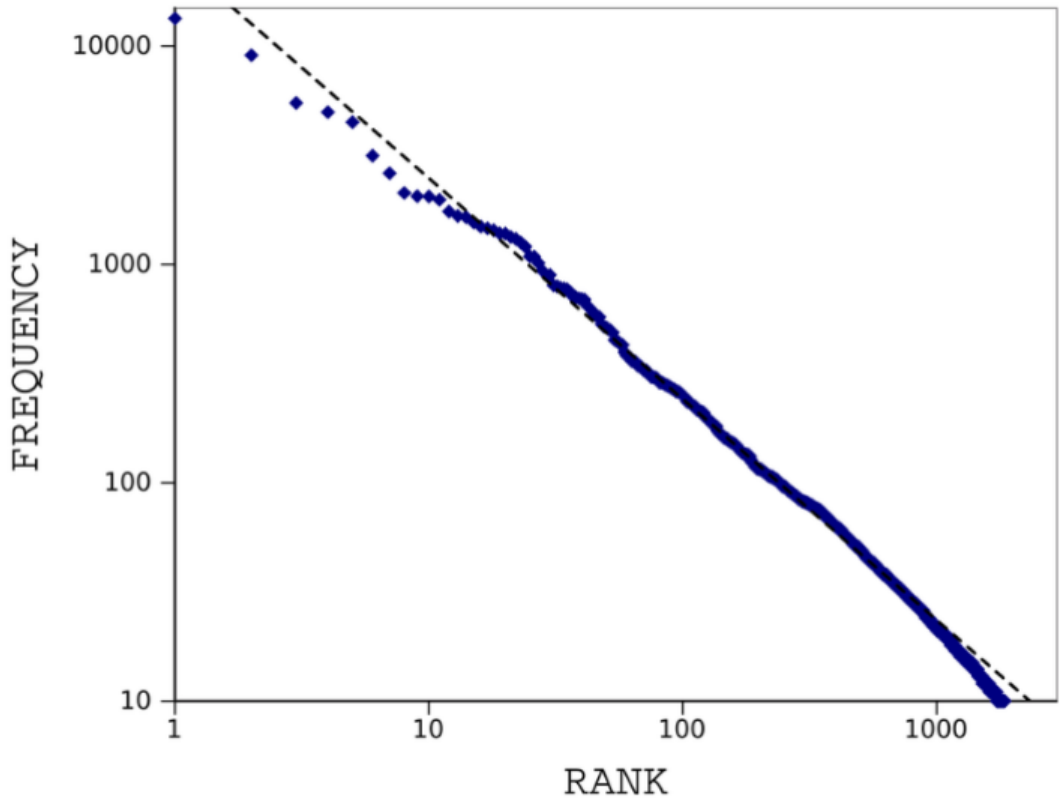
DATA SPACE CHARACTERISTICS

Determine what are you looking at and whether it is plausible

- incl. data formatting (e.g., IP addresses, lat/lon coordinates, license plates, CPR#,...)
- value ranges (e.g., lat: -90°...+90°, lon: -180°...+180°)
- units (e.g., Celsius (-273.15°C...), Fahrenheit (-459.67°F...) or Kelvin (0K...))
- placeholders (e.g., std::numeric_limits<float>::max())
- consistency (e.g., noisy sensor but whole data values, time running backwards)
- topology of data structure (e.g., if tree structure, check #edges +1 = #nodes)
- distribution of numbers (Benford's law), words (Zipf's law)



Benford's Law



Zipf's Law (Darwin's Origin of Species)

atmosphere

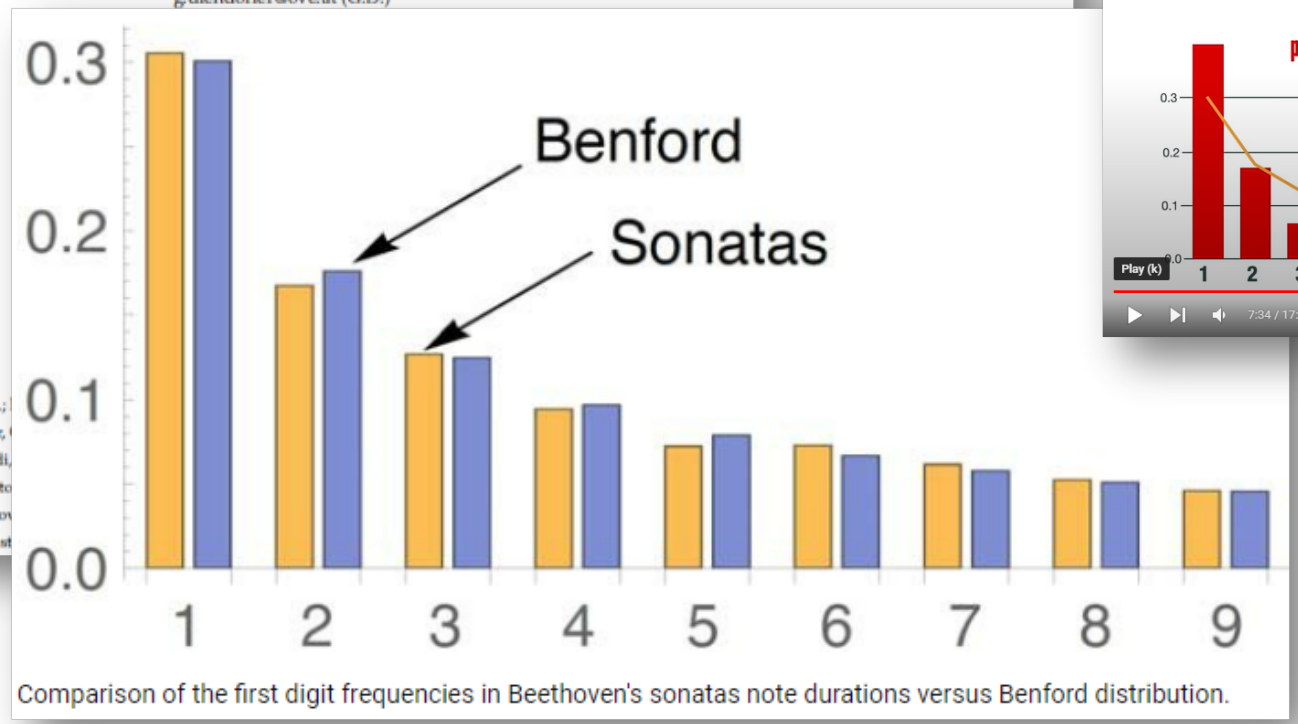
MDPI

Article

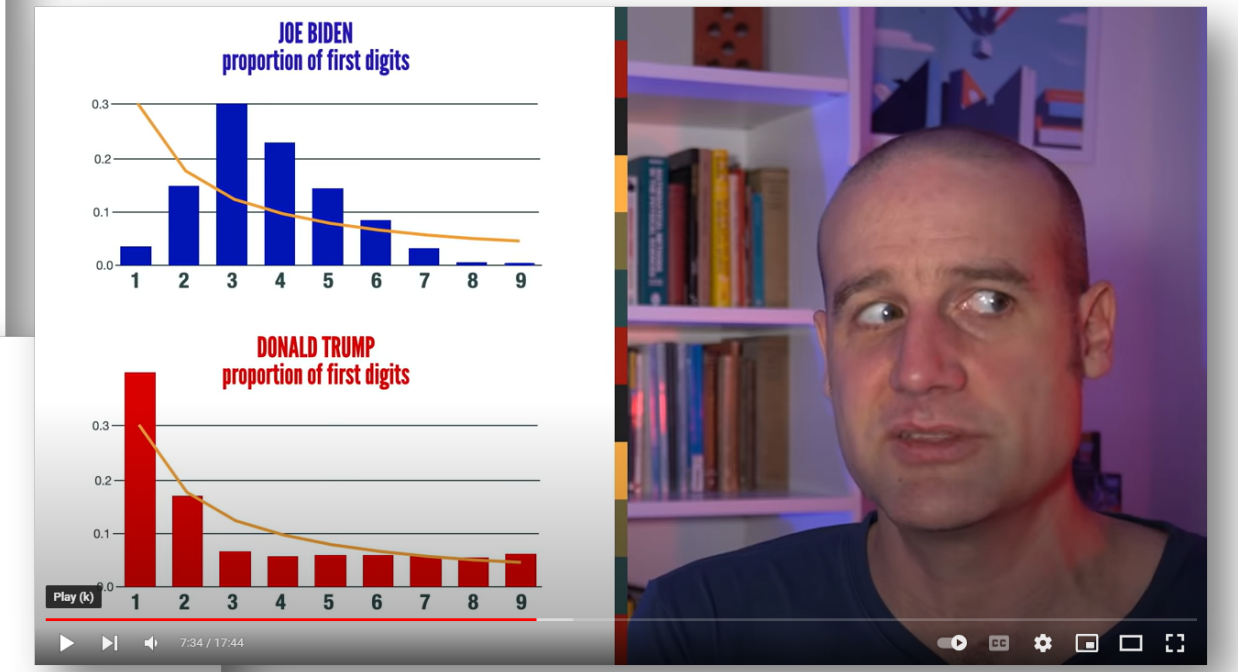
On the Use of Benford's Law to Assess the Quality of the Data Provided by Lightning Locating Systems

Ehsan Mansouri [1], Amirhosein Mostajabi [1], Wolfgang Schulz [2], Gerhard Diendorfer [2], Marcos Rubinstein [3] and Farhad Rachidi [1,*]

[1] Electromagnetic Compatibility Laboratory, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland; ehsan.mansouri@epfl.ch (E.M.); amirhossein.mostajabi@epfl.ch (A.M.)
[2] Department of ALDIS, OVE Service GmbH, 1010 Vienna, Austria; w.schulz@ove.at (W.S.); g.diendorfer@ove.at (G.D.)



Comparison of the first digit frequencies in Beethoven's sonatas note durations versus Benford distribution.



https://youtu.be/etx0k1nLn78

MAKE USE OF KNOWN VALUE RANGES!

Sensible value ranges as simple guards against “bad” data:

e.g., Temperature

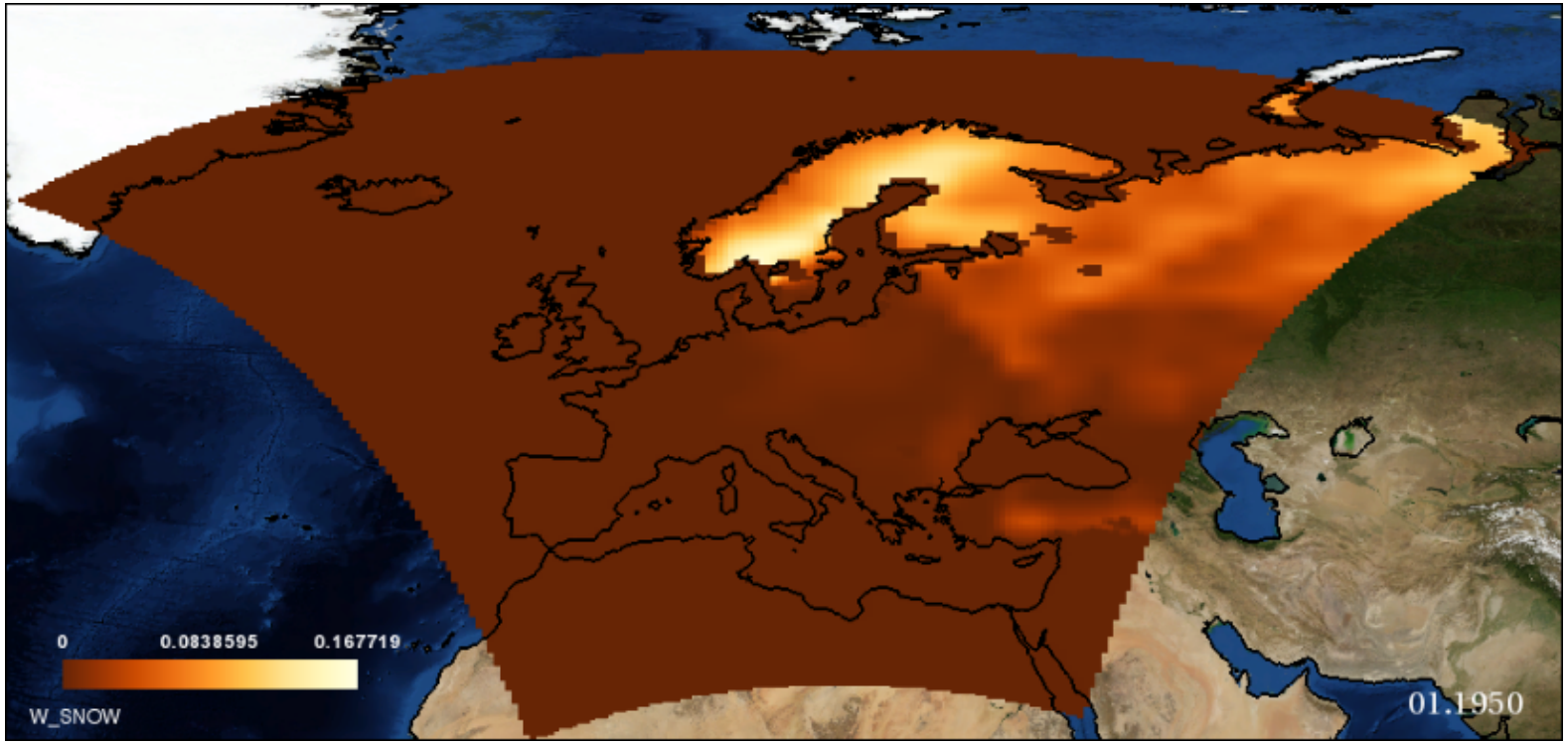
- **Body Temperature:** 10°C … 50°C
  min record that was survived: 14.2°C
  max record that was survived: 46.5°C
- **Room Temperature:** 0°C … 60°C
- **Surface Temperature:** -100°C … +100°
  min record: Vostok Station in Antarctica with -89.2°C on July 21, 1983
  max record: Death Valley in California with +56.7°C on July 10, 1913
  -> surface temperature: Lut Desert in Iran with +70.7°C in 2004/2005
- **General Temperature:** -273.15°C …

THE PROBLEM(S) WITH "STANDARDS"

- Competing Standards
- Different versions of the same standard
- Flexibility in interpreting the standard
- Incomplete implementations of the standard
- Standardized data still requires validation
- Not all required information may be part of the standard
- ...

[Schulz et al. 2017]

STANDARD ISSUES



Snow Cover Dataset visualized in the Software Avizo

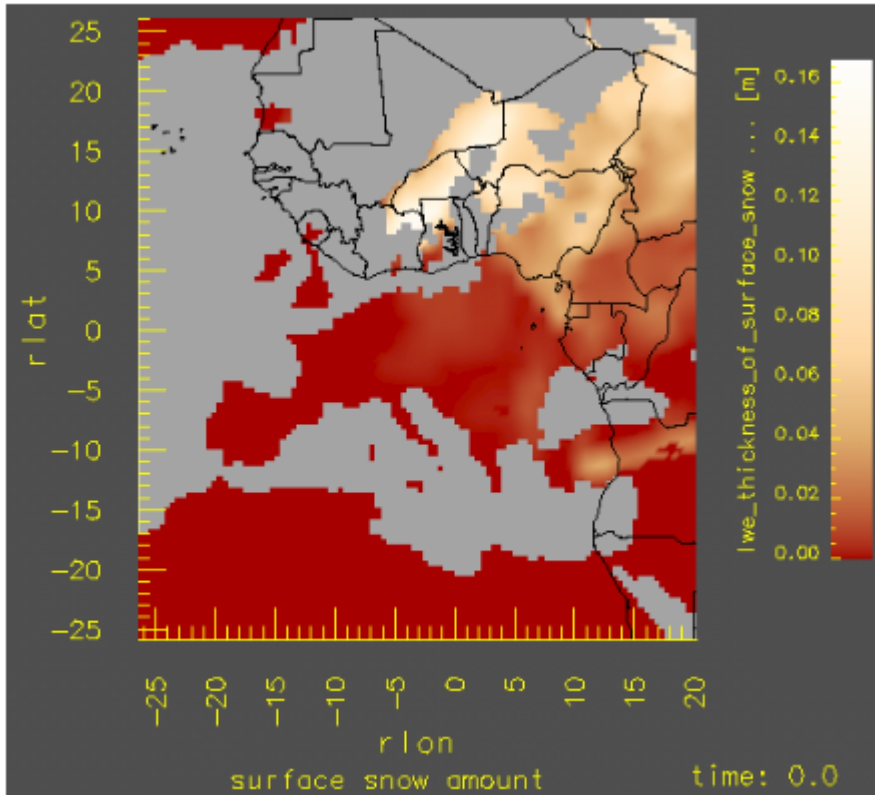
```
float W_SNOW(time, rlat, rlon);
  W_SNOW:standard_name =
    "lwe_thickness_of_surface_snow_amount";
  W_SNOW:long_name =
    "surface snow amount";
  W_SNOW:units = "m";
  W_SNOW:grid_mapping = "rotated_pole";
  W_SNOW:coordinates = "lon lat";
  W_SNOW:_FillValue = -1.e+20f;
```

```
char rotated_pole;
  rotated_pole:grid_mapping_name =
    "rotated_latitude_longitude";
  rotated_pole:grid_north_pole_latitude =
    39.25f;
  rotated_pole:grid_north_pole_longitude =
    -162.f;
  float rlon(rlon);
    rlon:axis = "X";
    rlon:standard_name = "grid_longitude";
    rlon:long_name = "rotated longitude";
    rlon:units = "degrees";
  float rlat(rlat);
    rlat:axis = "Y";
    rlat:standard_name = "grid_latitude";
    rlat:long_name = "rotated latitude";
    rlat:units = "degrees";
```

Dataset Description in NetCDF

Image Source: [Schulz et al. 2017]

STANDARD ISSUES



Snow Cover Dataset visualized in the Software OpenDX

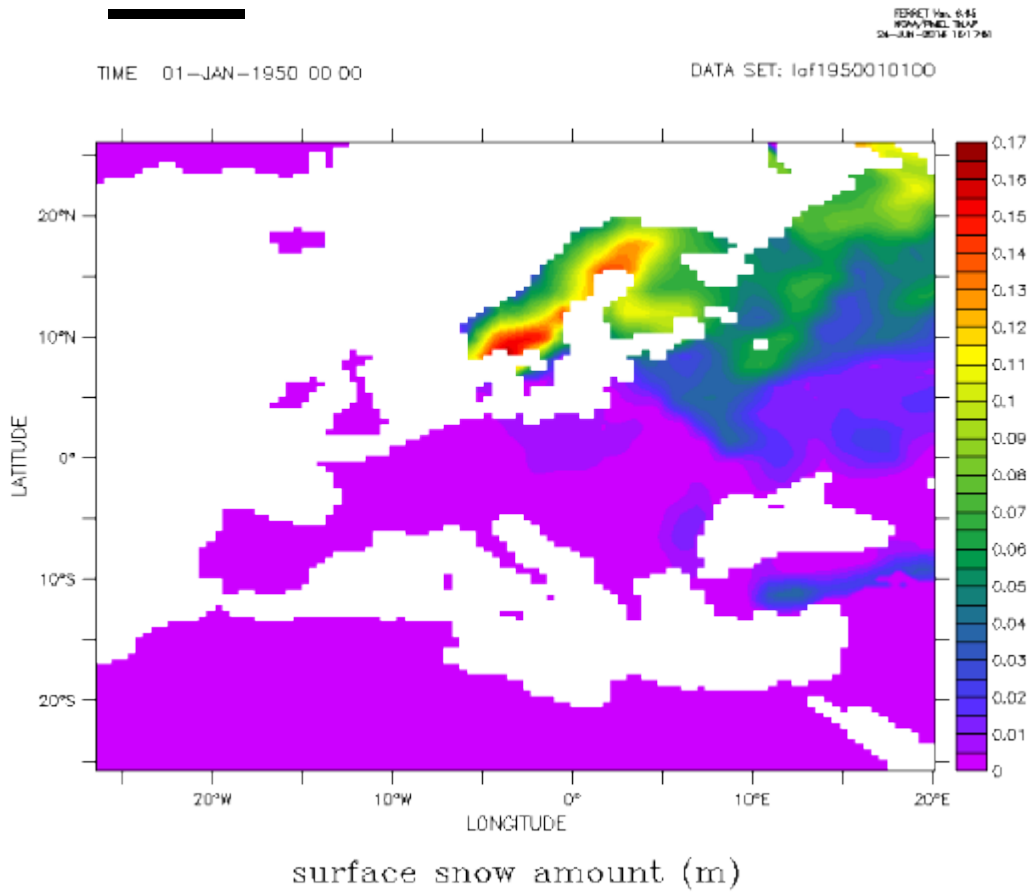
```
float W_SNOW(time, rlat, rlon);
  W_SNOW:standard_name =
    "lwe_thickness_of_surface_snow_amount";
  W_SNOW:long_name =
    "surface snow amount";
  W_SNOW:units = "m";
  W_SNOW:grid_mapping = "rotated_pole";
  W_SNOW:coordinates = "lon lat";
  W_SNOW:_FillValue = -1.e+20f;
```

```
char rotated_pole;
  rotated_pole:grid_mapping_name =
    "rotated_latitude_longitude";
  rotated_pole:grid_north_pole_latitude =
    39.25f;
  rotated_pole:grid_north_pole_longitude =
    -162.f;
  float rlon(rlon);
    rlon:axis = "X";
    rlon:standard_name = "grid_longitude";
    rlon:long_name = "rotated longitude";
    rlon:units = "degrees";
  float rlat(rlat);
    rlat:axis = "Y";
    rlat:standard_name = "grid_latitude";
    rlat:long_name = "rotated latitude";
    rlat:units = "degrees";
```

Dataset Description in NetCDF

Image Source: [Schulz et al. 2017]

STANDARD ISSUES



Snow Cover Dataset visualized in the Software Ferret

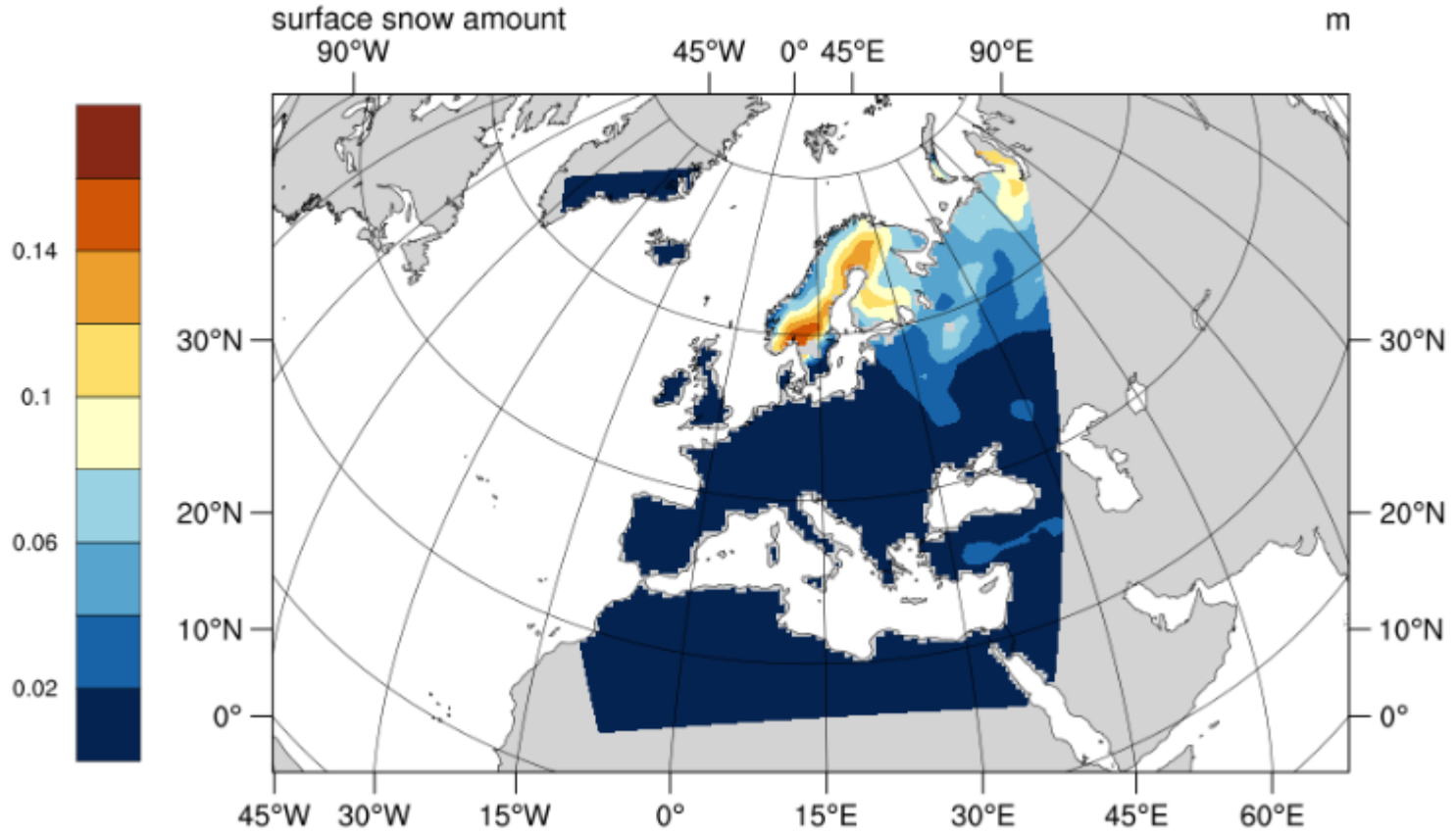
```
float W_SNOW(time, rlat, rlon);
  W_SNOW:standard_name =
    "lwe_thickness_of_surface_snow_amount";
  W_SNOW:long_name =
    "surface snow amount";
  W_SNOW:units = "m";
  W_SNOW:grid_mapping = "rotated_pole";
  W_SNOW:coordinates = "lon lat";
  W_SNOW:_FillValue = -1.e+20f;
```

```
char rotated_pole;
  rotated_pole:grid_mapping_name =
    "rotated_latitude_longitude";
  rotated_pole:grid_north_pole_latitude =
    39.25f;
  rotated_pole:grid_north_pole_longitude =
    -162.f;
  float rlon(rlon);
    rlon:axis = "X";
    rlon:standard_name = "grid_longitude";
    rlon:long_name = "rotated longitude";
    rlon:units = "degrees";
  float rlat(rlat);
    rlat:axis = "Y";
    rlat:standard_name = "grid_latitude";
    rlat:long_name = "rotated latitude";
    rlat:units = "degrees";
```

Dataset Description in NetCDF

Image Source: [Schulz et al. 2017]

STANDARD ISSUES



Snow Cover Dataset visualized in the Software NCL (NCAR Command Language)

Image Source: [Schulz et al. 2017]

```
float W_SNOW(time, rlat, rlon);
  W_SNOW:standard_name =
    "lwe_thickness_of_surface_snow_amount";
  W_SNOW:long_name =
    "surface snow amount";
  W_SNOW:units = "m";
  W_SNOW:grid_mapping = "rotated_pole";
  W_SNOW:coordinates = "lon lat";
  W_SNOW:_FillValue = -1.e+20f;
```

```
char rotated_pole;
  rotated_pole:grid_mapping_name =
    "rotated_latitude_longitude";
  rotated_pole:grid_north_pole_latitude =
    39.25f;
  rotated_pole:grid_north_pole_longitude =
    -162.f;
  float rlon(rlon);
    rlon:axis = "X";
    rlon:standard_name = "grid_longitude";
    rlon:long_name = "rotated longitude";
    rlon:units = "degrees";
  float rlat(rlat);
    rlat:axis = "Y";
    rlat:standard_name = "grid_latitude";
    rlat:long_name = "rotated latitude";
    rlat:units = "degrees";
```

Dataset Description in NetCDF

DATA WRANGLING

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

HANS-JÖRG SCHULZ
ASSOCIATE PROFESSOR

DATA WRANGLING

Data wrangling := the process of making any raw dataset useful by identifying and treating missing values, duplicate and possibly contradicting entries, formatting issues, and other problems of data quality

Missing values: imputation vs. amputation

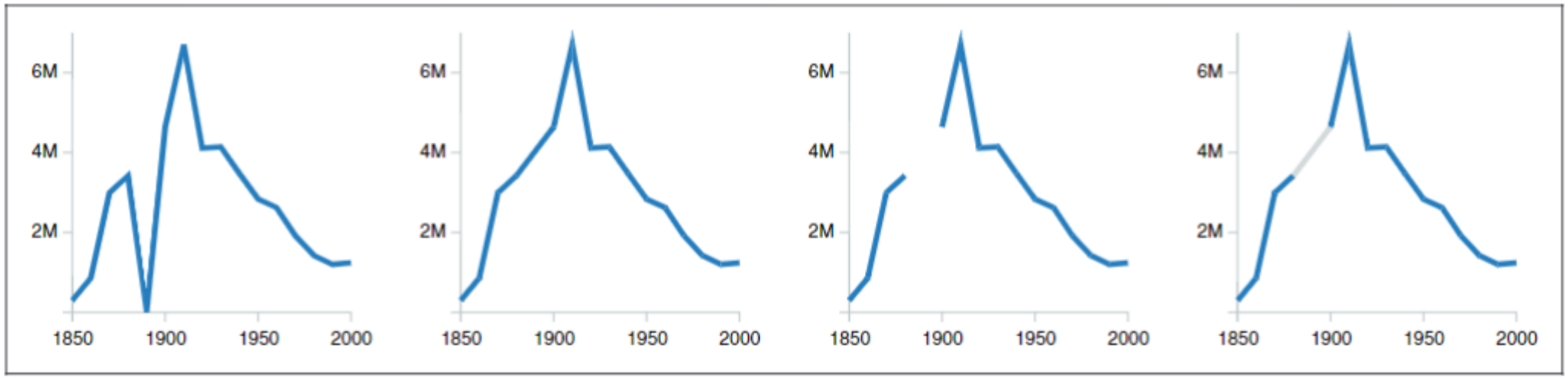
Duplicate entries: de-duplication by deleting or merging

Formatting issues: file format, version, interpretation, implementation, validation

Data quality: uncertainty, accuracy,...

HANDLING MISSING DATA VALUES

Number of Farm Laborers in the US, records for 1890 missing due to fire [Source: Kandel 2011]



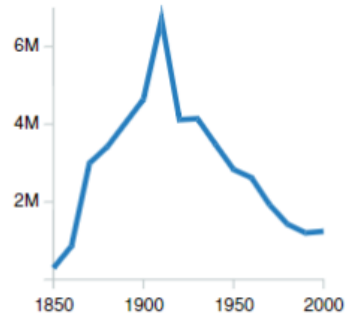
use placeholder (0) as is

data imputation using linear interpolation

data amputation

data imputation and explicit encoding

DIFFERENT FORMS OF IMPUTATION



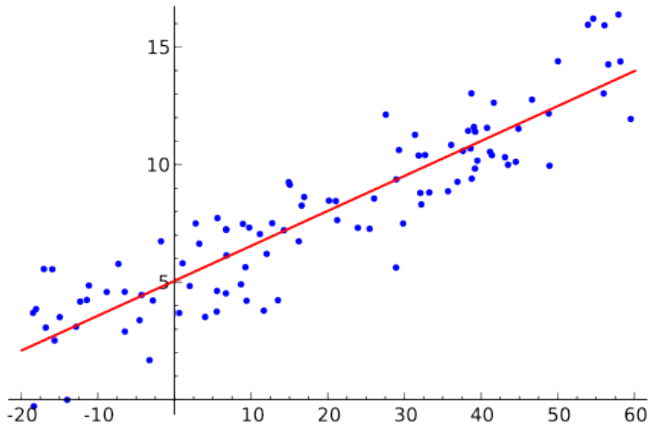
- Last Observation Carried Forward

- Mean Value Imputation (Average Value)
- Median Value Imputation (Center Value)
- Mode Value Imputation (Most Frequent Value)

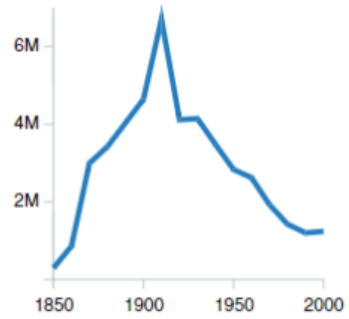
⇒ for Continuous Data
⇒ for Ordinal Data
⇒ for Categorical Data

- Linear Interpolation

- Linear Regression for quantitative values



DIFFERENT FORMS OF IMPUTATION

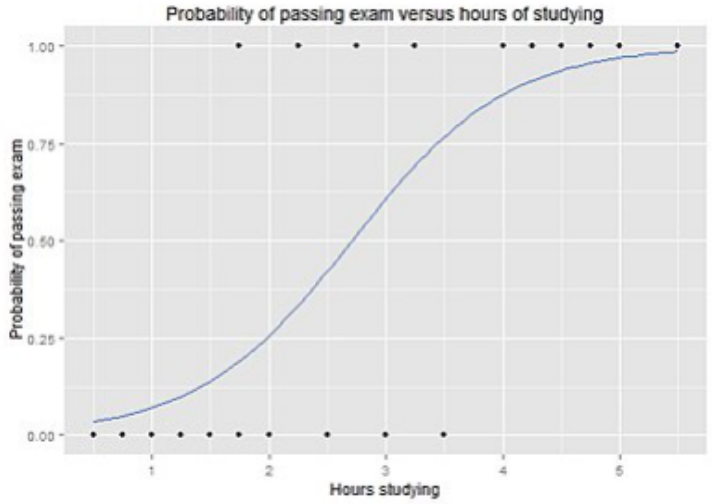


- Last Observation Carried Forward

- Mean Value Imputation (Average Value)
- Median Value Imputation (Center Value)
- Mode Value Imputation (Most Frequent Value)

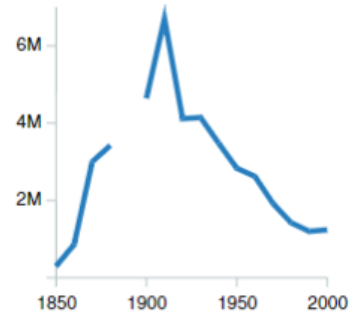
⇒ for Continuous Data
⇒ for Ordinal Data
⇒ for Categorical Data

- Linear Interpolation



- Linear Regression for quantitative values
- Logistic Regression for qualitative values

DIFFERENT FORMS OF AMPUTATION



- Row-wise Deletion => if missing data <5% of dataset
- Column-wise Deletion => if missing data >40% of dataset

} Rules of Thumb!

- Imputation => if missing data >5% but <40%
  (i.e., too large to ignore rows, but to small to discard column)

- Pair-wise Deletion := remove items only from analyses over missing values

  => Problem: analyses over different variables may be based on different item sets

WHY RULES OF THUMB?

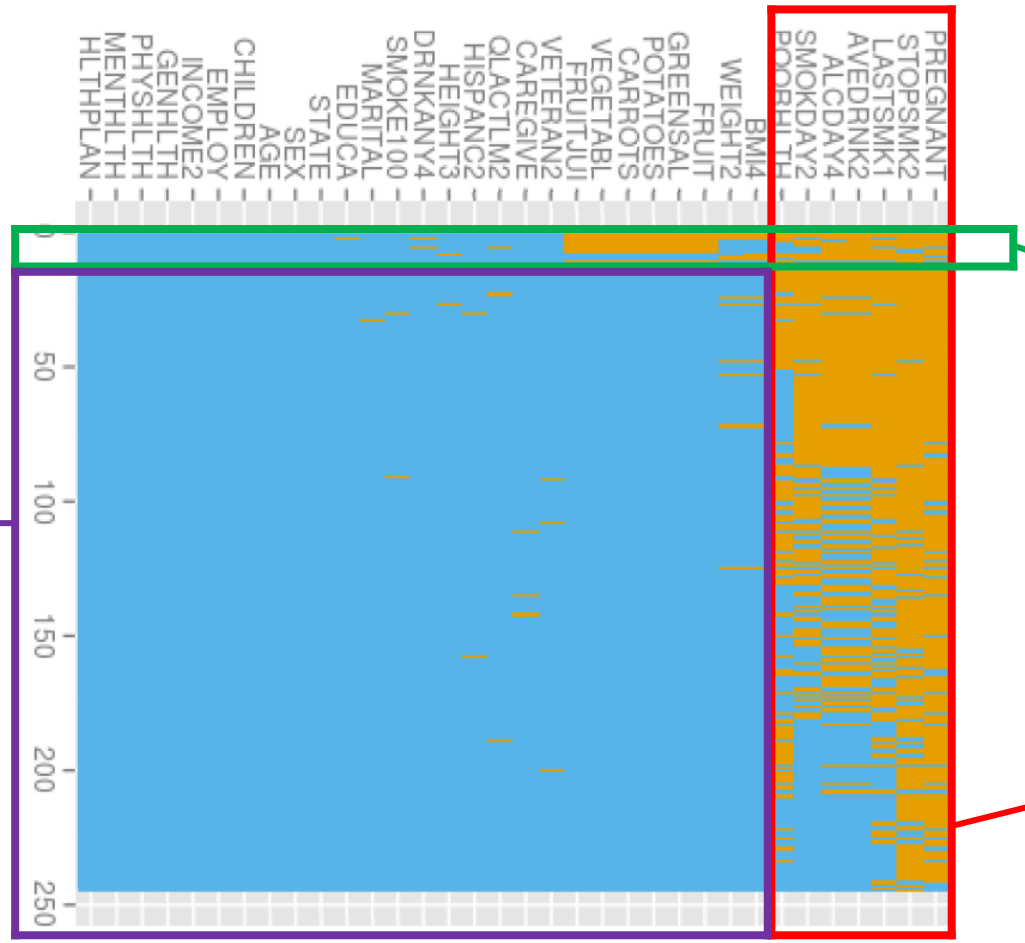
For example: Column-wise Deletion => if missing data >40% of dataset

Even if 90% of your data is missing, you may still be able to gain useful results from the remaining 10% -> **Under which circumstances?**

- **Data is missing randomly** -> doesn't introduce a bias in the remaining data, which can then be treated as a 10% sample
- **Dataset is large enough** -> so that one can still draw statistically significant conclusions from the remaining 10%

VIS SUPPORT: MISSINGNESS MAPS

Source: [Cheng et al. 2015]



Remaining data subset

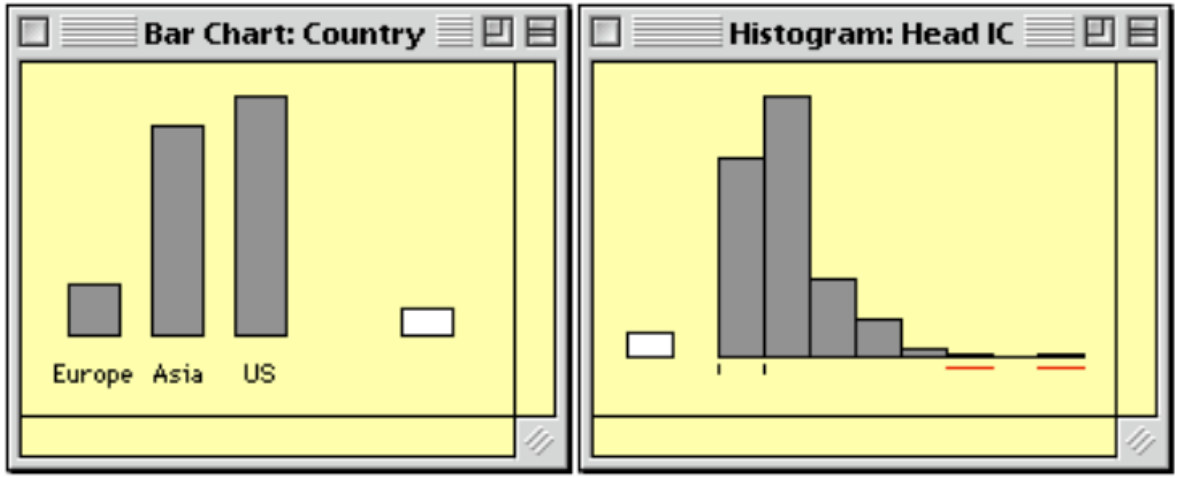
Candidates for pair-wise deletion or imputation

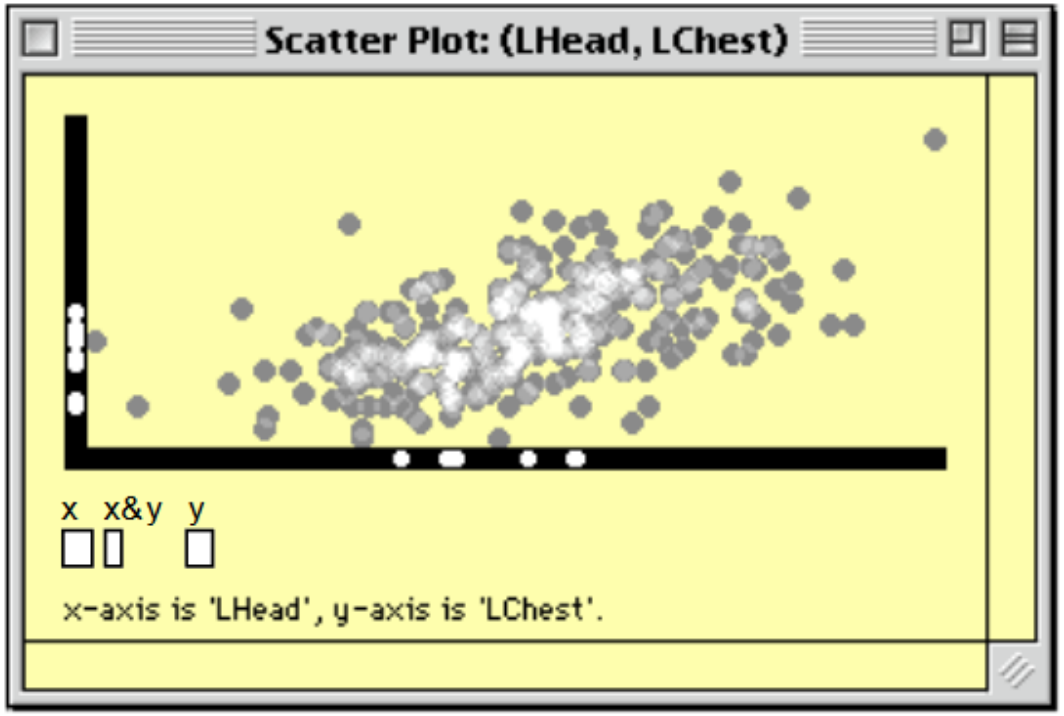
Candidates for row-wise deletion

Candidates for column-wise deletion

VIS SUPPORT: EXPLICIT ENCODING

Source: [Theus et al. 1997]





DE-DUPLICATION

SSN	Name	Date of Birth	Sex	City of Residence
000956723	William J. Smith	1/2/73	Male	Berkeley, California
000005555	Robert Jones	1942/08/14	Male	Seattle, WA
000956723	Smith, W. J.	1973.1.2		Berkeley, CA
	Bill Smith	Jan 2, 1973	Male	Berkeley, Calif.
123001234	Sue, Mary	November 19, 1972	Female	
	Jones, Bob	14/08/1942	Male	Seattle, WA

1. Use unique identifiers where they are available (e.g., SSN/CPR#)
2. Use quasi-identifiers where not (Name, DoB, Sex)

STRING MATCHING ALGORITHMS

Entity Resolution based on the individual data value:

- Fuzzy Matching => e.g., bitap algorithm: https://en.wikipedia.org/wiki/Bitap_algorithm
- Phonetic Matching => e.g., Metaphone algorithm: https://en.wikipedia.org/wiki/Metaphone

1. Drop duplicate adjacent letters, except for C.
2. If the word begins with 'KN', 'GN', 'PN', 'AE', 'WR', drop the first letter.
3. Drop 'B' if after 'M' at the end of the word.
4. 'C' transforms to 'X' if followed by 'IA' or 'H' (unless in latter case, it is part of '-SCH-', in which case it transforms to 'K'). 'C' transforms to 'S' if followed by 'I', 'E', or 'Y'. Otherwise, 'C' transforms to 'K'.
5. 'D' transforms to 'J' if followed by 'GE', 'GY', or 'GI'. Otherwise, 'D' transforms to 'T'.
6. Drop 'G' if followed by 'H' and 'H' is not at the end or before a vowel. Drop 'G' if followed by 'N' or 'NED' and is at the end.
7. 'G' transforms to 'J' if before 'I', 'E', or 'Y', and it is not in 'GG'. Otherwise, 'G' transforms to 'K'.
8. Drop 'H' if after vowel and not before a vowel.
9. 'CK' transforms to 'K'.
10. 'PH' transforms to 'F'.
11. 'Q' transforms to 'K'.
12. 'S' transforms to 'X' if followed by 'H', 'IO', or 'IA'.
13. 'T' transforms to 'X' if followed by 'IA' or 'IO'. 'TH' transforms to '0'. Drop 'T' if followed by 'CH'.
14. 'V' transforms to 'F'.
15. 'WH' transforms to 'W' if at the beginning. Drop 'W' if not followed by a vowel.
16. 'X' transforms to 'S' if at the beginning. Otherwise, 'X' transforms to 'KS'.
17. Drop 'Y' if not followed by a vowel.
18. 'Z' transforms to 'S'.
19. Drop all vowels unless it is the beginning.

Original Metaphone (Source: Wikipedia)
Double Metaphone (http://aspell.net/metaphone/dmetaph.cpp)
Metaphone 3 (commercial licenses starting at 240$)

STRING MATCHING ALGORITHMS

Entity Resolution based on the individual data value:

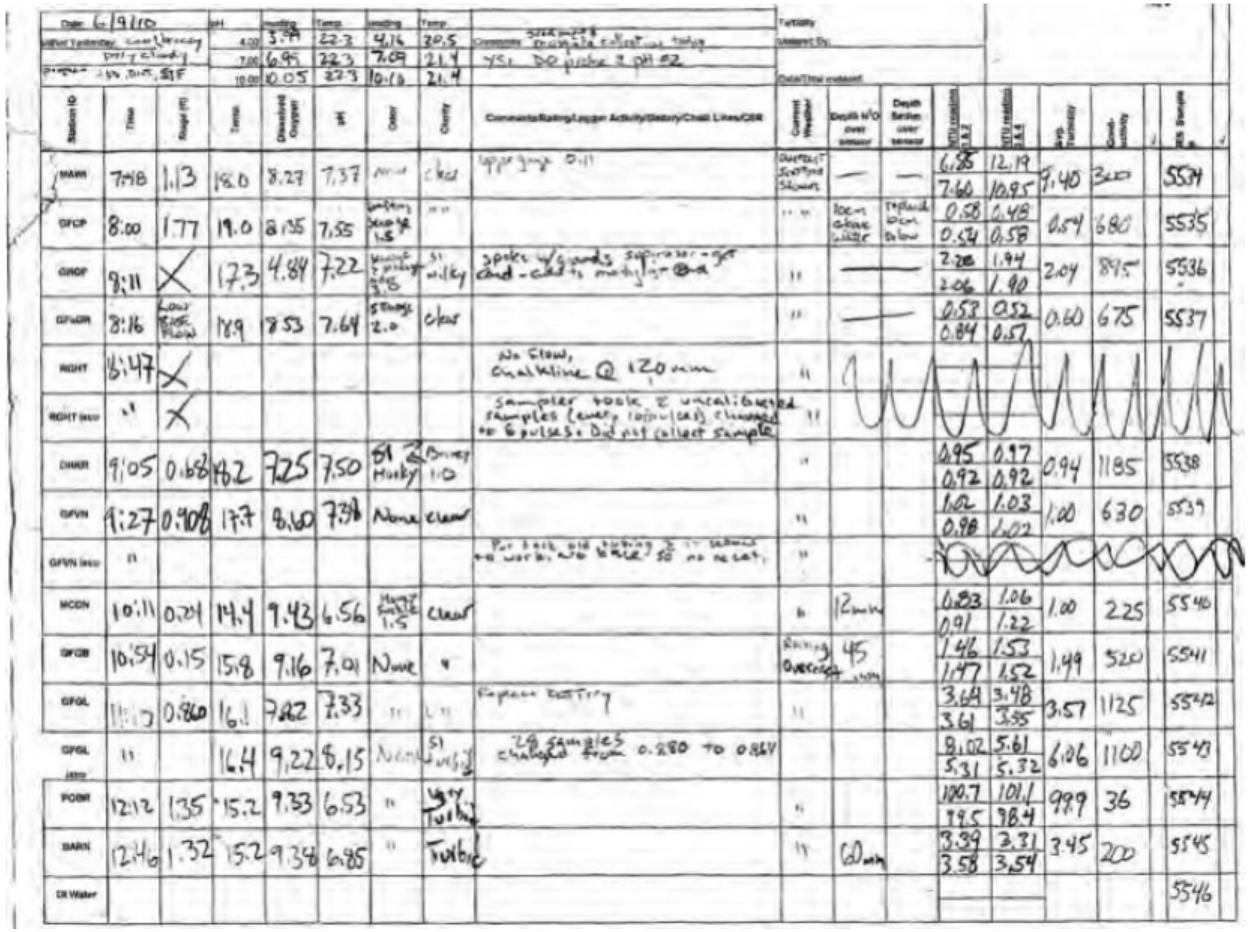
- Fuzzy Matching => e.g., bitap algorithm: https://en.wikipedia.org/wiki/Bitap_algorithm
- Phonetic Matching => e.g., Metaphone algorithm: https://en.wikipedia.org/wiki/Metaphone

  -> Alternatives: Caverphone: https://en.wikipedia.org/wiki/Caverphone
  Soundex: https://en.wikipedia.org/wiki/Soundex
  Daitch–Mokotoff Soundex / Beider-Morse Soundex: https://en.wikipedia.org/wiki/Daitch%E2%80%93Mokotoff_Soundex
- https://medium.com/@ievgenii.shulitskyi/phonetic-matching-algorithms-50165e684526
- https://stackabuse.com/phonetic-similarity-of-words-a-vectorized-approach-in-python/

DATA QUALITY: UNCERTAINTY

Uncertainty may be introduced when generating the data – e.g. through

- **technical / environmental sources:**
  measurement errors,
  biased measurements
- **human sources:**
  manual data collection
- **inherent sources:**
  stochastic processes,
  ensemble simulation



Field Sheet – Source: [Ribes & Jackson 2013: Data bite man]

DATA TRANSFORMATION

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

HANS-JÖRG SCHULZ
ASSOCIATE PROFESSOR

DATA REDUCTION – SAMPLING

- **Random Sampling**

- **Stratified Sampling:** random sampling proportionally across strata/sub-groups
  Example: sampling 1000 AU students, proportionally across faculties

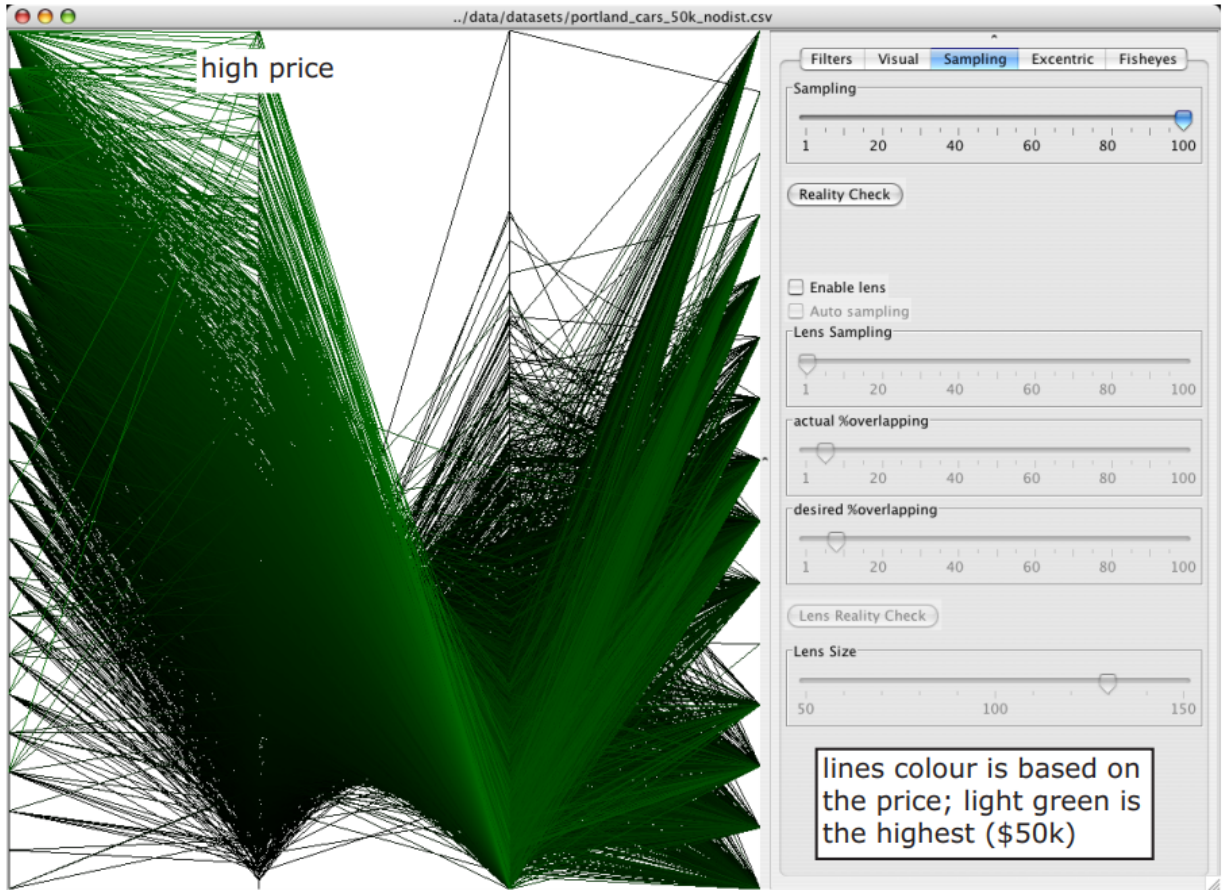
- **Quota Sampling:** non-random sampling proportionally across strata/sub-groups
  Example: sampling 1000 AU students (500 male/500 female), proportionally across faculties

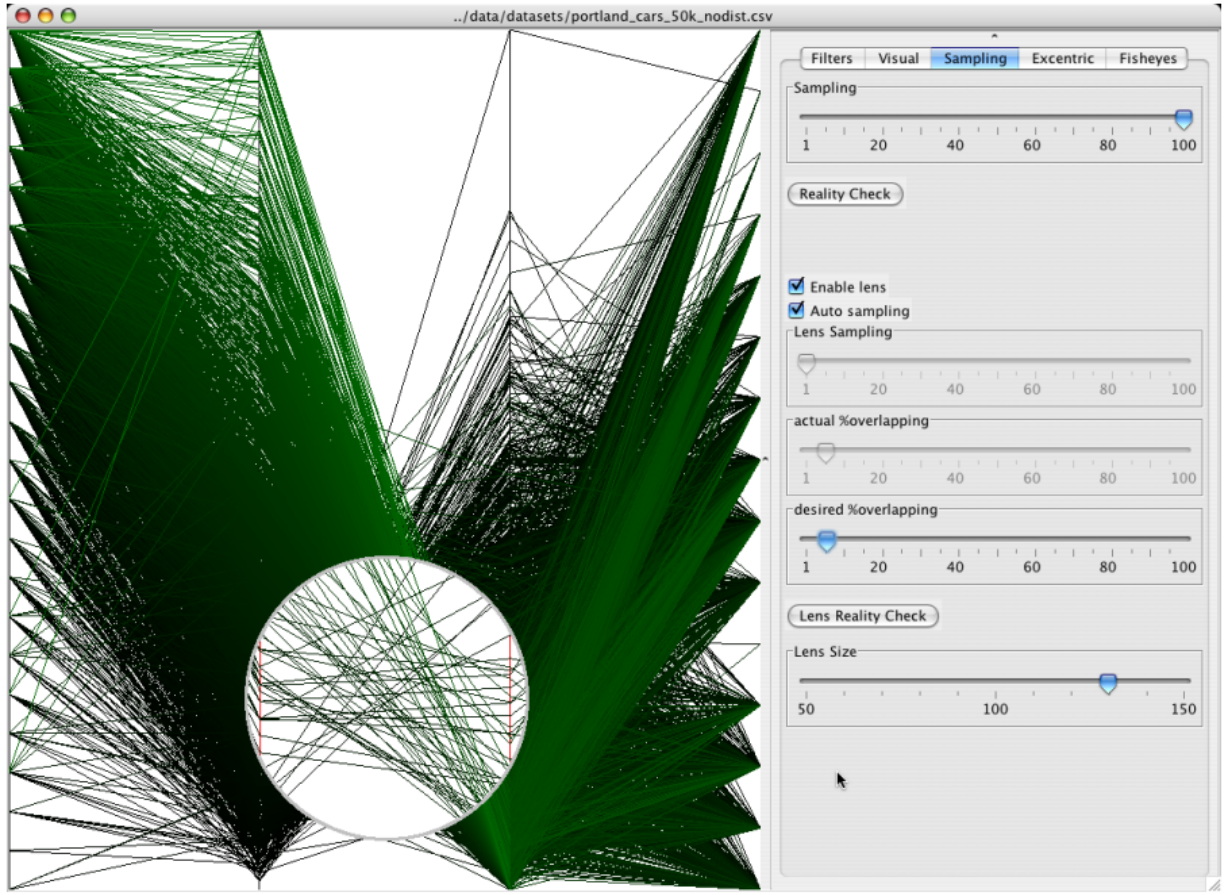
SAMPLING FOR DOT MAPS

Crimes in Philadephia (2006-2013): Full dataset on the left (700k data points),
Sample on the right (5,3k data points)

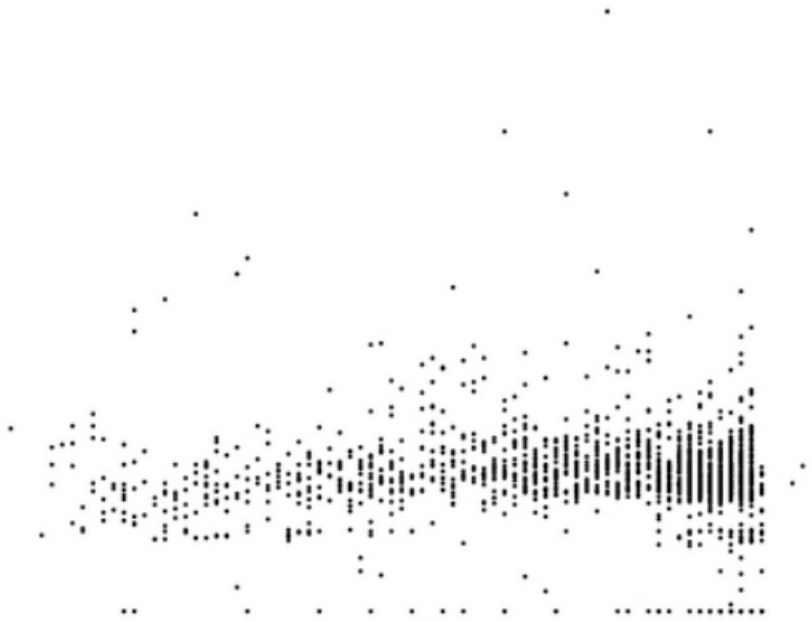
THE SAMPLING LENS

[Ellis et al. 2005]

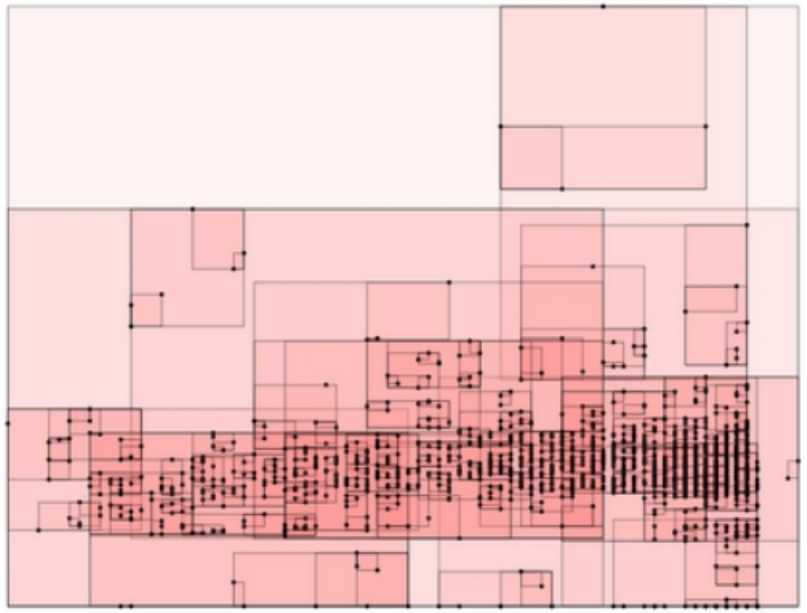




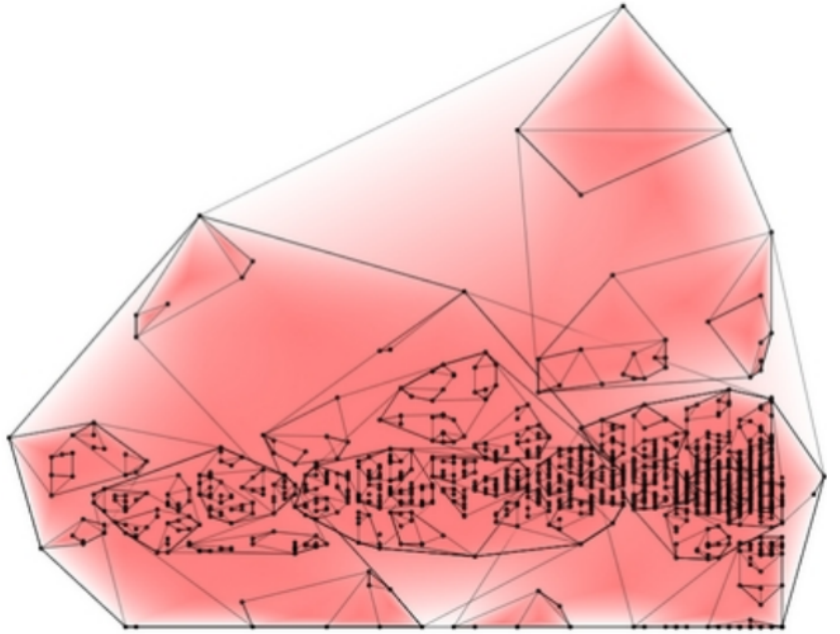
AGGREGATING DATA



(a) 2D scatterplot visualization.

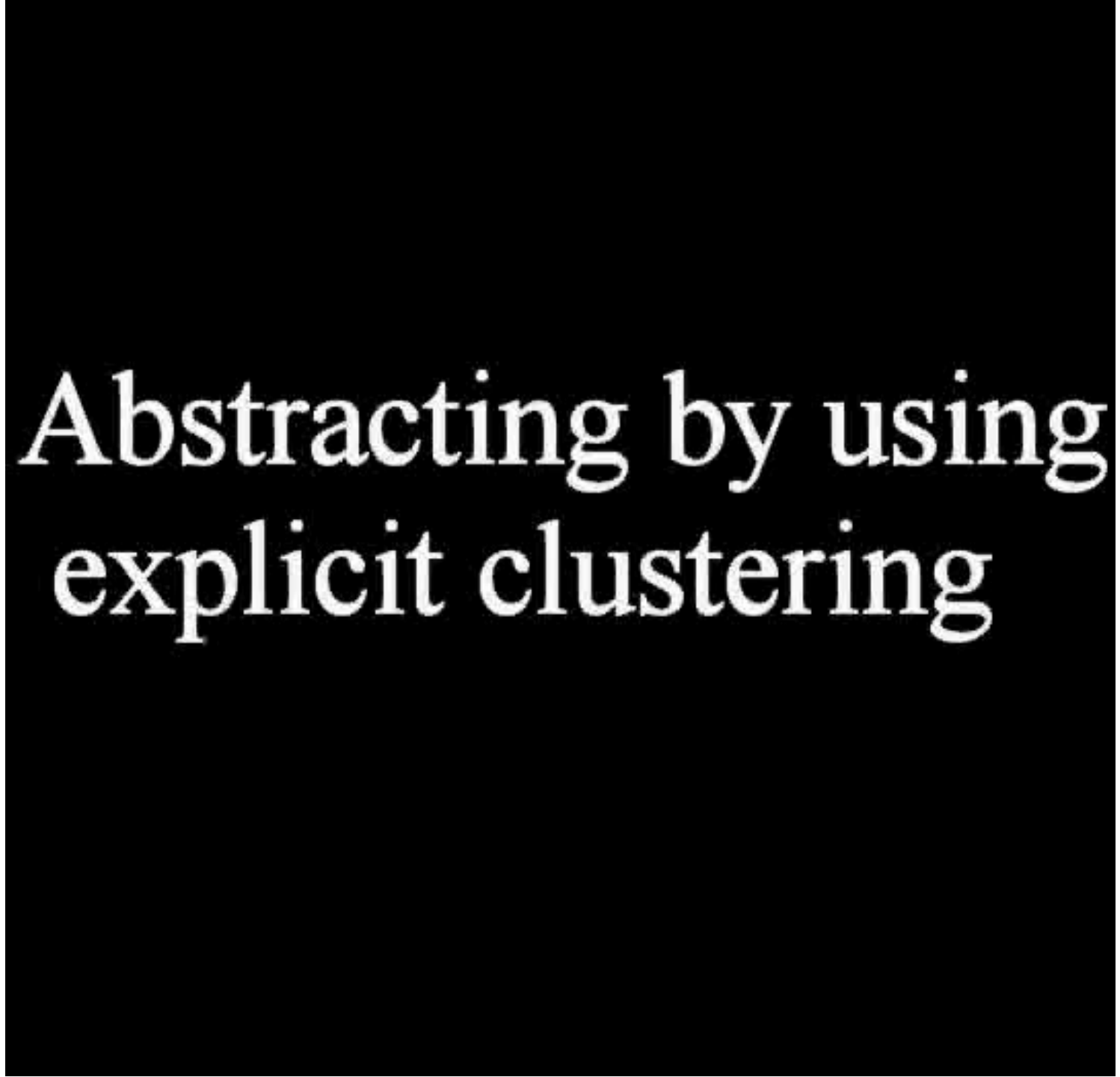


(b) 2D bounding box aggregation.



(c) 2D convex hull aggregation.

Image source: [Elmqvist & Fekete 2010]



Source: v.Ham & Wijk 2004

READING RECOMMENDATION

The Quartz guide to bad data

https://github.com/Quartz/bad-data-guide

Spreadsheet has dates in 1900 or 1904

For reasons beyond obscure, Excel's default date from which it counts all other dates is `January 1st, 1900`, *unless* you're using Excel on a Mac, in which case it's `January 1st, 1904`. There are a variety of ways in which data in Excel can be entered or calculated incorrectly and end up as one of these two dates. If you spot them in your data, it's probably an issue.

Inflation skews the data

Currency inflation means that over time money changes in value. There is no way to tell if numbers have been "inflation adjusted" just by looking at them. If you get data and you aren't sure if they have been adjusted then check with your source. If they haven't you'll likely want to perform the adjustment. This inflation adjuster is a good place to start.

https://timeseriesreasoning.com/contents/inflation-adjustment/
https://www.worldbank.org/en/research/brief/inflation-database

READING

CHAPTER 2

Criteria, Factors, and Models

CONTENTS

INTERACTIVE VISUAL data analysis is highly context-dependent. We will need different techniques for analyzing time-series data than for graph data. We will want to use completely different visual representations for getting an overview of the overall data distribution than for inspecting individual patterns and trends. And we will most likely interact differently when working

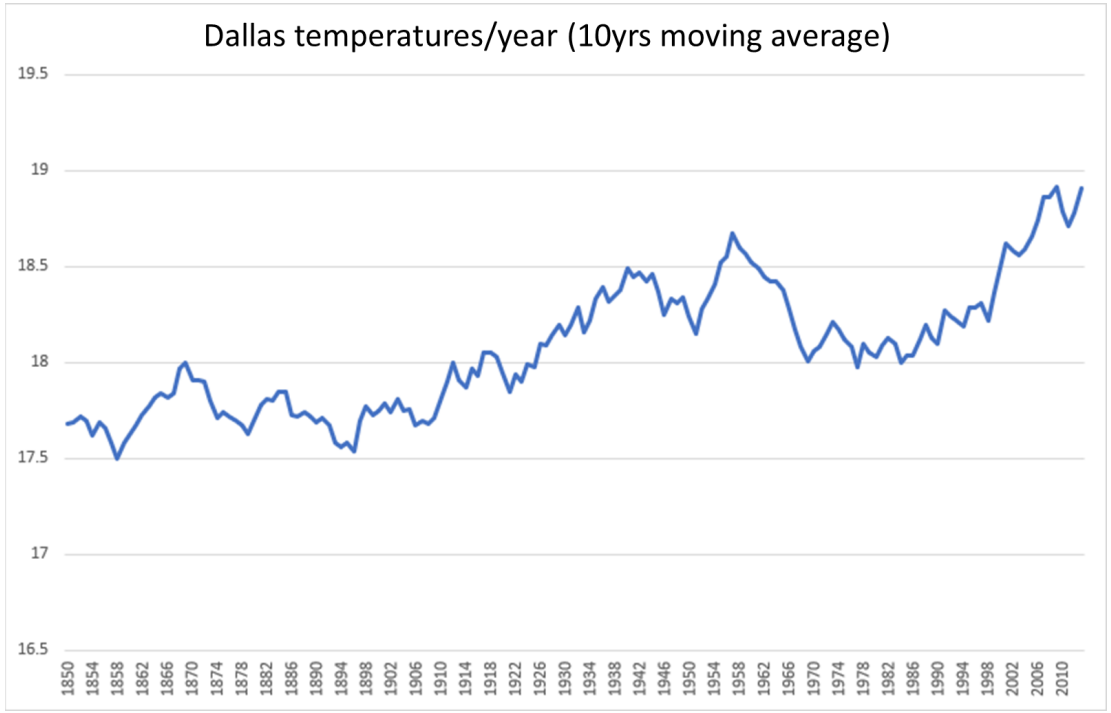
LIST OF LITERATURE SOURCES

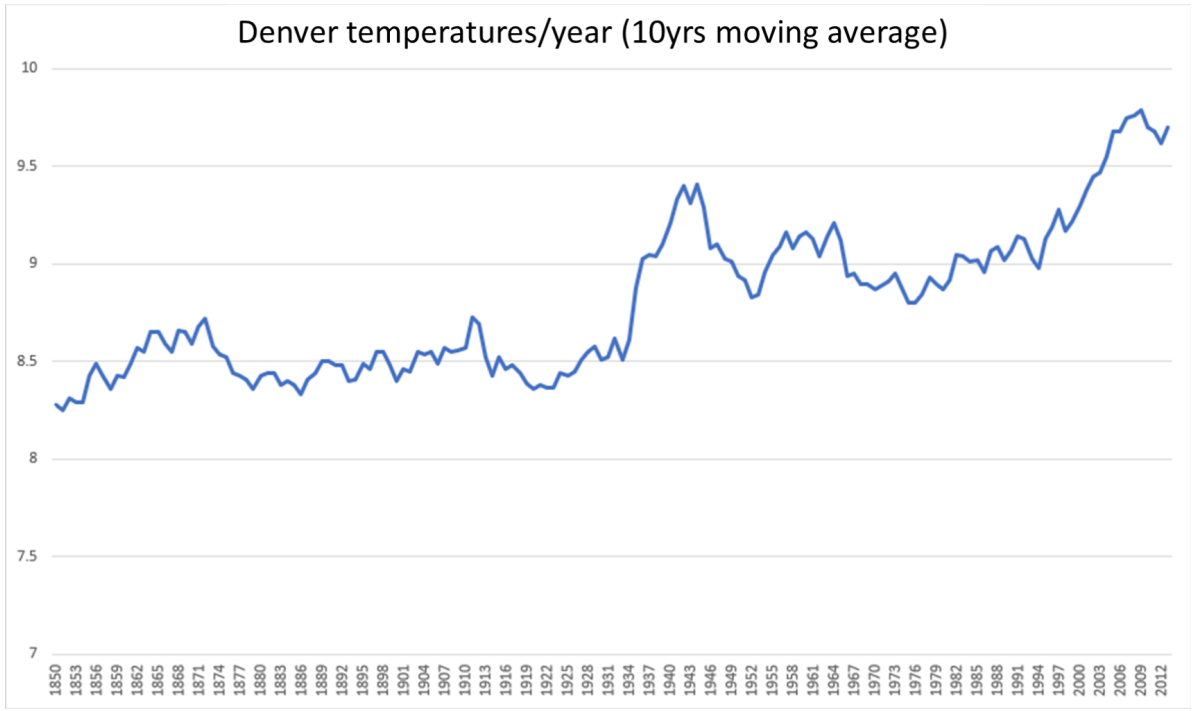
- Schulz et al. 2017: https://doi.org/10.1177/1473871616667767
- The Quartz Guide to Bad Data: https://github.com/Quartz/bad-data-guide
- Zheng et al. 2021: https://doi.org/10.1109/tbdata.2019.2913655
- Kandel et al. 2011: https://doi.org/10.1177/1473871611415994
- Mansouri et al. 2022: https://doi.org/10.3390/atmos13040552
- Cheng et al. 2015: https://doi.org/10.18637/jss.v068.i06
- Theus et al. 1997: book chapter in "New Techniques and Technologies for Statistics II", IOS Press, pp.247-259
- Ribes & Jackson 2013: https://doi.org/10.7551/mitpress/9302.003.0010
- Ellis et al. 2005: https://doi.org/10.1145/1056808.1056914
- Elmqvist & Fekete 2010: https://doi.org/10.1109/TVCG.2009.84
- Jonathan Taylor 2010 – updated version: https://web.stanford.edu/class/stats202/notes/Unsupervised/Clustering.html
- V.Ham & v.Wijk 2004: https://doi.org/10.1109/INFVIS.2004.43

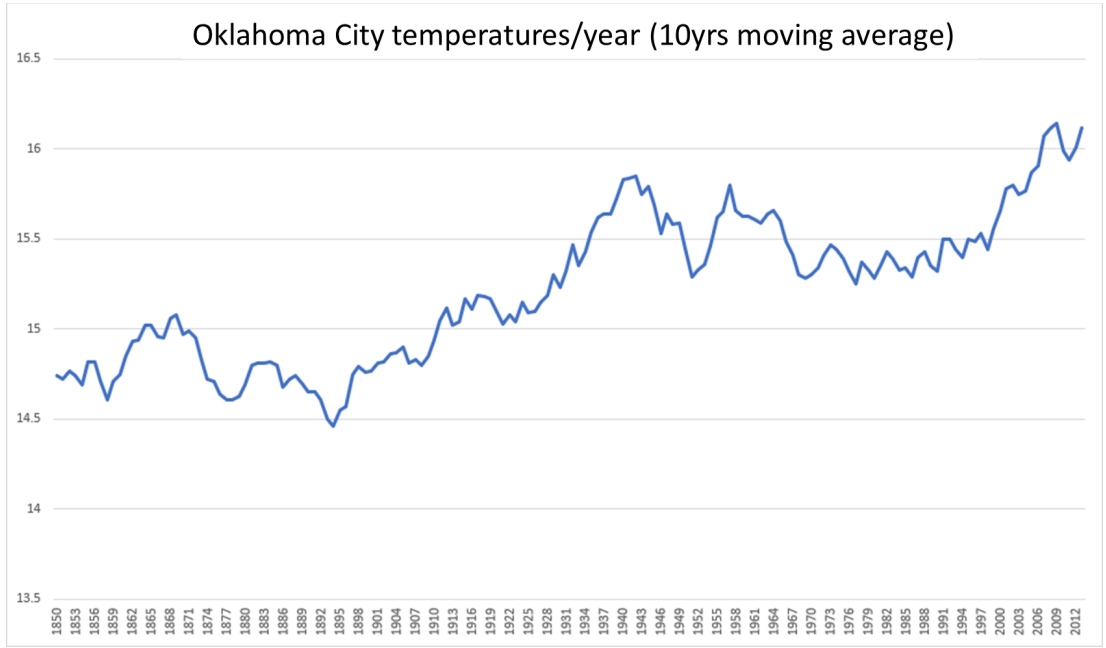
BLACK HAT VISUALIZATION:
Preparing an intentionally misleading visualization

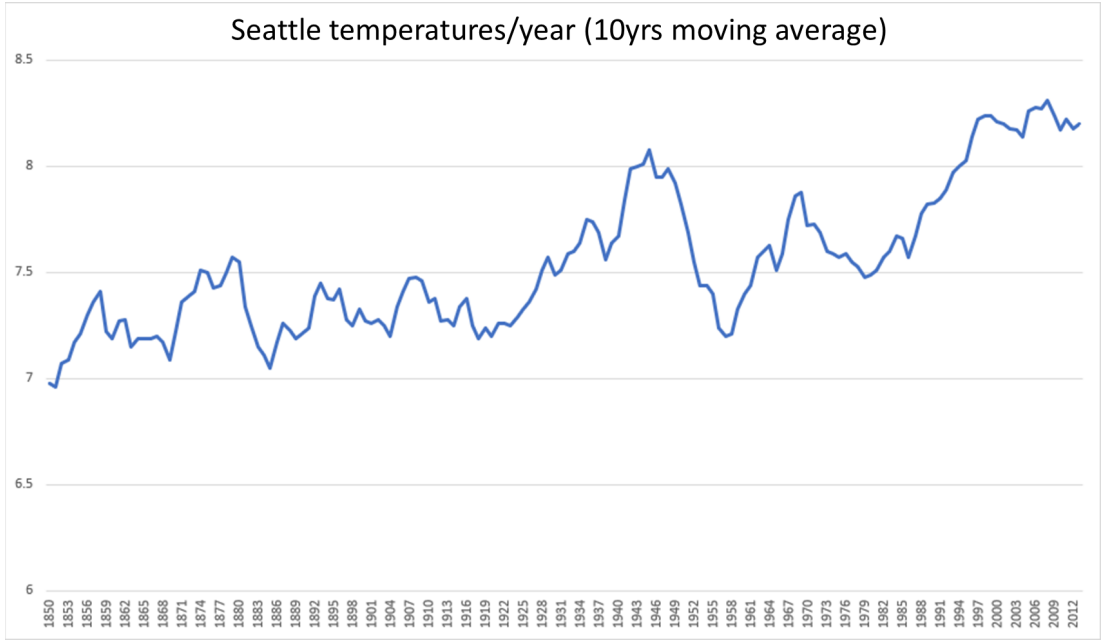
AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

HANS-JÖRG SCHULZ
ASSOCIATE PROFESSOR


Dallas temperatures/year (10yrs moving average)
19.5
19
18.5
18
17.5
17
16.5
1850 1854 1858 1862 1866 1870 1874 1878 1882 1886 1890 1894 1898 1902 1906 1910 1914 1918 1922 1926 1930 1934 1938 1942 1946 1950 1954 1958 1962 1966 1970 1974 1978 1982 1986 1990 1994 1998 2002 2006 2010


Denver temperatures/year (10yrs moving average)
10
9.5
9
8.5
8
7.5
7
1850 1853 1856 1859 1862 1865 1868 1871 1874 1877 1880 1883 1886 1889 1892 1895 1898 1901 1904 1907 1910 1913 1916 1919 1922 1925 1928 1931 1934 1937 1940 1943 1946 1949 1952 1955 1958 1961 1964 1967 1970 1973 1976 1979 1982 1985 1988 1991 1994 1997 2000 2003 2006 2009 2012


Oklahoma City temperatures/year (10yrs moving average)
16.5
16
15.5
15
14.5
14
13.5
1850 1853 1856 1859 1862 1865 1868 1871 1874 1877 1880 1883 1886 1889 1892 1895 1898 1901 1904 1907 1910 1913 1916 1919 1922 1925 1928 1931 1934 1937 1940 1943 1946 1949 1952 1955 1958 1961 1964 1967 1970 1973 1976 1979 1982 1985 1988 1991 1994 1997 2000 2003 2006 2009 2012


Seattle temperatures/year (10yrs moving average)
8.5
8
7.5
7
6.5
6
1850 1853 1856 1859 1862 1865 1868 1871 1874 1877 1880 1883 1886 1889 1892 1895 1898 1901 1904 1907 1910 1913 1916 1919 1922 1925 1928 1931 1934 1937 1940 1943 1946 1949 1952 1955 1958 1961 1964 1967 1970 1973 1976 1979 1982 1985 1988 1991 1994 1997 2000 2003 2006 2009 2012

DETAILS

Data: Brightspace – Content – 1st Half of Semester – Week 40

Task: Create an intentionally misleading visualization of that data.

1 Rule: You may not falsify data.

Process / Tooling: Entirely up to you.
(Recommendation: LibreOffice Calc + Inkscape for Postprocessing)

Hand-in: As a PDF/PNG/JPG in a separate thread in your project group's discussion board by 8pm on Tuesday (30-SEP-2025)

Results/Discussion: During the TA session on Wednesday (01-OCT)

EVALUATION CATEGORIES

1. **The "wrongest":** the group that managed to squeeze the most visualization mistakes into one chart (by count)

2. **The ugliest:** the group that made the ugliest visualization (by popular vote)

3. **The most unethical:** the group whose visualization most convincingly conveyed the exact opposite of what's in the data (by popular vote)

WHAT COUNTS AS A VISUALIZATION DESIGN MISTAKE

Examples of what counts:

- cut-off axes
- glass slippers
- perspective distortion
- bad color scale
- (ab)using gestalt principles
- …

Examples of what doesn't count:

- overplotting / too small
- using subpar channels
- mapping data randomly

- multiple effects due to a single design decision will only count as one (example: 3D)


SOLIDUM PETIT IN PROFUNDIS
UNIVERSITAS ARHUSIENSIS