INM430
Principles of Data Science

Week 02
Data Characteristics & Wrangling

Aidan Slingsby,
Constantino Carlos Reyes-Aldasoro
giCentre
CITY

Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats

Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats

Know your data

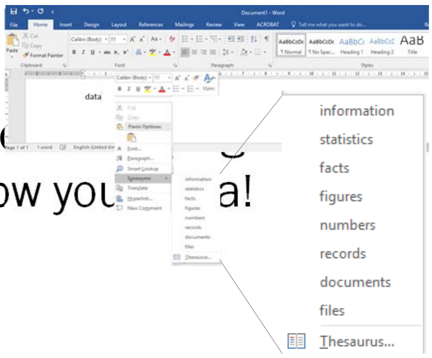
But … what is data?

Before anything else …
Know your data!

(N.B. Overlap with VA in the next slides – slight differences in vocabulary!)

Know your data

Bef
Know you a!

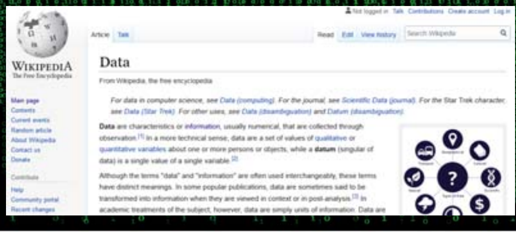


What is Data?

- Information
- Statistics
- Facts
- Figures
- Numbers
- Records
- Documents
- Knowledge
- Understanding
- Patterns

Data and Information

- Data are characteristics or information
- Is *data = information*?



Data and Information

- Imagine the following *zettabytes* of storage

00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...
00000000000000 ... 00000000 ...

Continue until you have zettabytes

Data and Information

- We could say that we have data, a lot of data, but with no information
- Can some data carry more information than other?

3 Numbers

5 Numbers

8 Numbers

– 0 1 9
– 0 1 0 3 9
– 0 1 0 3 9 12 4 9
– 0 0 0 0 0 0 0 0

8 Numbers

- But we are back to the “length” or number of symbols.
- There *must* be something else.

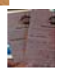
Data and Information

- Not all data conveys information.
- What is the fundamental nature of ***“information”***?
  - Today is Friday
  - London is in England
  - In England rains frequently
  - Today is raining in London
  - Tomorrow there will be heavy rains in London, flood warnings have been issued

Why is the last phrase more interesting?

Data and Information

- Information resolves **uncertainty.**
- The uncertainty of an event is measured by its **probability of occurrence.**
  - Uncertainty of flipping a coin (1 in 2)
  - Uncertainty of rolling dice (1 in 6)
  - Uncertainty of Euromillions (1 in 6,991,908)



- Information is inversely proportional to the probability of occurrence.

$$I \propto \frac{1}{p_i} \qquad p_i = 1 \rightarrow I = 0 \qquad I = \log\left(\frac{1}{p_i}\right)$$

Data and Information

- Information resolves **uncertainty.**
- The uncertainty of an event is measured by its **probability of occurrence.**
  - Uncertainty of flipping a coin (1 in 2)
  - Uncertainty of rolling dice (1 in 6)
  - Uncertainty of Euromillions (1 in 6,991,908)
- Information is inversely proportional to the probability of occurrence.

$$H = -\sum_{i=1}^{n} p_i \log(p_i)$$

Data and Information

Data, (or *Raw Data)* is a

0101100
1100110

collection/signal/record/file/matrix/ function/container/arrangement/…

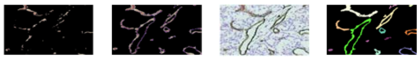
that conveys **information** about the characteristics, behaviour or attributes of some phenomenon

biological / geographical / financial / medical / cultural / meteorological / …

Information resolves **uncertainty**. To resolve the uncertainty we need to

process/analyse/visualize/transform…

the data, generally through **computational** processes.



Week 2 Schedule

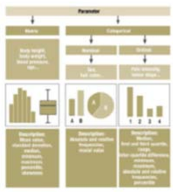
- Know your data (and your information)
- **Data attributes types**
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats

Data attribute types – How data is measured and categorised?

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
Nominal	Determination of equality	Permutation group $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	Isotonic group $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
Interval	Determination of equality of intervals or differences	General linear group $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	Similarity group $x' = ax$	Coefficient of variation

On the theory of scales and measurements, 1946 [S. Stevens]

Slide by Tamara Munzner, http://www.cs.ubc.ca/~tmm/talks/minicourse14/minicourse14-session1.pdf



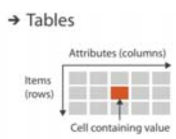
Spriestersbach, Descriptive statistics: the specification of statistical measures and their presentation in tables and graphs. Dtsch Arztebl Int, 2009

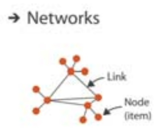
Data Type Taxonomy by Shneiderman, 96

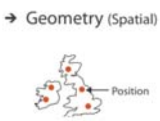
- 1D (sequences)
- Temporal
- 2D (maps)
- 3D (shapes)
- nD (relational)
- Trees (hierarchical)
- Networks (graphs)
- … ?

Shneiderman, Ben. "The eyes have it: **A task by data type taxonomy for information visualizations.**" *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996.

Dataset Taxonomy by Tamara Munzner, 2014





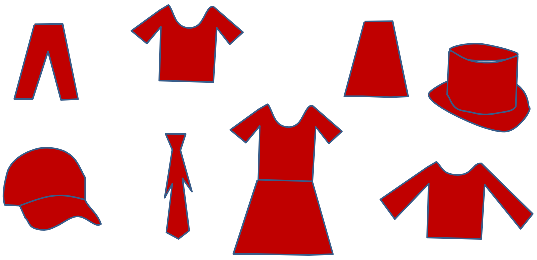


\+ Text

Talks by **Tamara Munzner**, http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

Data attribute types: Categorical

- Categorical / Nominal: related to the category, name or the label that characterises each item



Data attribute types: Categorical

- Categorical / Nominal: related to the category, name or the label that characterises each item, items may have more than one label

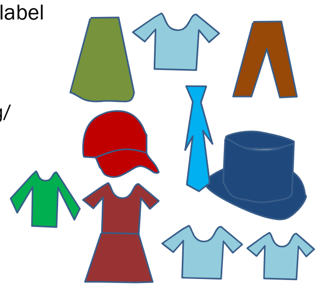
No specific rank or order

No operations like adding/ subtracting

No distance metric

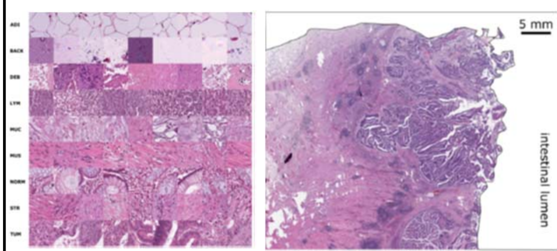
Mode / Majority

Percentage of universe



Data attribute types: Categorical

- Categorical / Nominal: allocate labels according to a certain characteristic



Kather et al. (2019) Predicting survival from colorectal cancer histology slides using deep learning. PLoS Med 16(1): e1002730.

Data attribute types: Categorical

- Categorical / Nominal: allocate labels according to a certain characteristic

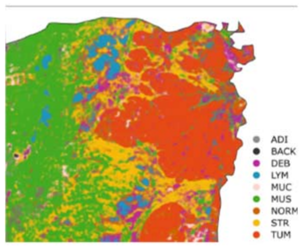
No specific rank or order

No operations like adding/ subtracting

No distance metric

Mode / Majority

Percentage of universe



Kather et al. (2019) Predicting survival from colorectal cancer histology slides using deep learning. PLoS Med 16(1): e1002730.

Small note on errors (more to come)

- Allocate labels according to a certain characteristic.
- How "good" is our allocation??????? (it depends how we define "good")
- With a "Ground Truth" (real label) we can define:
  - True Positives, True Negatives
  - False Positives, False Negatives

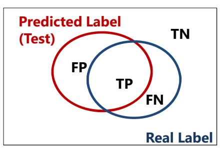
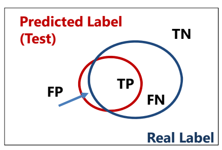


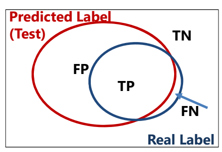
Figure 3.1 Type I and Type II errors

Paul Ellis, The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results, 2012, Cambridge University Press

Small note on errors (more to come)

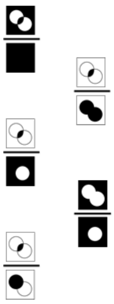
- Allocate labels according to a certain characteristic.
- How "good" is our allocation??????? (it depends how we define "good")
- With a "Ground Truth" (real label) we can define:
  - True Positives, True Negatives
  - False Positives, False Negatives





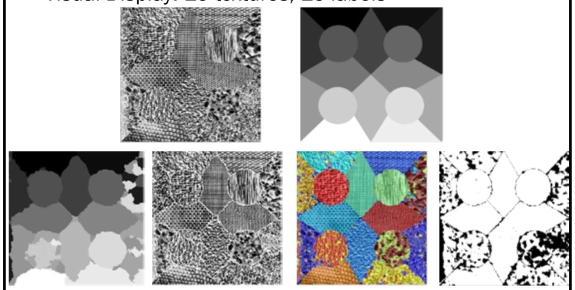
Small note on errors (more to come)

- With TP, TN, FP, FN we can define:
  - Accuracy $\frac{(TP+TN)}{(TP + FP + FN+ TN)}$
  - Jaccard $\frac{TP}{(TP + FP + FN)}$
  - Sensitivity (Recall) $\frac{TP}{(TN+ FP)}$
  - Specificity $\frac{TN}{(TN+FP)}$
  - Precision $\frac{TP}{(TP+FP)}$
  - Many many more ...



Small note on errors (more to come)

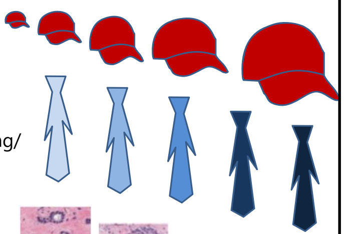
- Visual Display: 16 textures, 16 labels



Karabag et al., Texture Segmentation: An Objective Comparison between Five Traditional Algorithms and a Deep-Learning U-Net Architecture, J Imaging, 2019
Reyes-Aldasoro, Bhalero, The Bhattacharyya space for feature selection and its application to texture segmentation, Pattern Recognition, 2006

Data attribute types: Ordinal

- Order or ranks

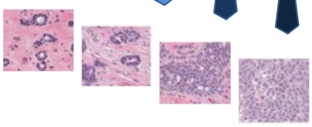


Specific rank or order

No operations like adding/ subtracting

Comparisons =, ≠, >, <

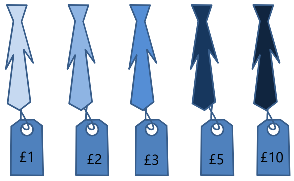
Distance metrics possible



Ortega-Ruiz et al. (2019) Morphological estimation of Cellularity on Neo-adjuvant treated breast cancer histological images. J Imaging 2020.

Data attribute types : Numerical

- Measurable quantities



Specific rank or order

Comparisons =, ≠, >, <

Operations + - (sometimes * /)

Meaningful zero = Zero denotes absence £0

No Meaningful zero = Zero does not denote absence 0 °C

Data Types and operations

- Categorical: nominal (labels)
  - Operations: =, ≠
- Categorical: Ordinal (ordered)
  - Operations: =, ≠, >, <
- Quantitative: Interval (no meaningful zero)
  - Operations: =, ≠, >, <, +, − (distance)
- Quantitative: Ratio (meaningful zero)
  - Operations: =, ≠, >, <, +, −, ×, ÷ (proportions)

Discrete
Continuous

Data Types and operations

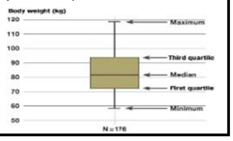
- Categorical: nominal (labels)
  - Operations: =, ≠
- Categorical: Ordinal (ordered)
  - Operations: =, ≠, >, <
- Quantitative: Interval (no meaningful zero)
  - Operations: =, ≠, >, <, +, − (distance)
- Quantitative: Ratio (meaningful zero)
  - Operations: =, ≠, >, <, +, −, ×, ÷ (proportions)

Random
Cyclical
Deterministic

Data Types and operations

- Categorical: nominal (labels)
  - Operations: =, ≠
- Categorical: Ordinal (ordered)
  - Operations: =, ≠, >, <
- Quantitative: Interval (no meaningful zero)
  - Operations: =, ≠, >, <, +, − **(distance)**
- Quantitative: Ratio (meaningful zero)
  - Operations: =, ≠, >, <, +, −, ×, ÷ **(proportions)**

Metrics: *Mean, Median, Standard Deviation, Quartiles, …*

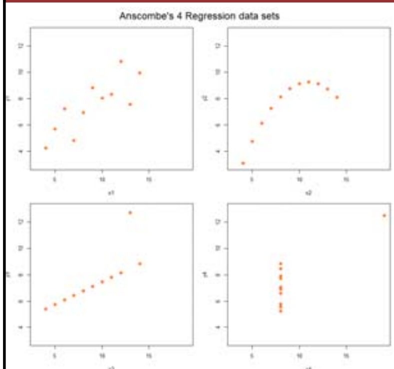


Anscombe's Quartet (warning about *metrics*)

I		II		III		IV		
x	y	x	y	x	y	x	y	
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58	
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76	
13.0	7.58	13.0	8.74	13.0	12.7	8.0	7.71	
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84	
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47	
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04	
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25	
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.5	
12.0	10.8	12.0	9.13	12.0	8.15	8.0	5.56	
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91	
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89	
9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5	Mean
3.31	2.03	3.31	2.03	3.31	2.03	3.31	2.03	Standard Deviation

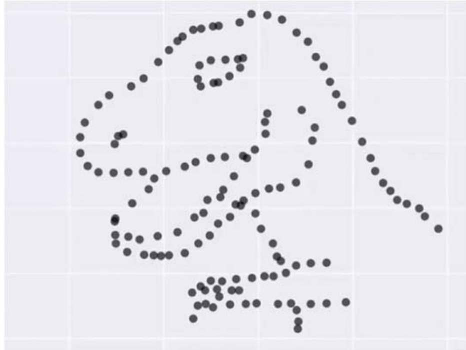
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe's Quartet



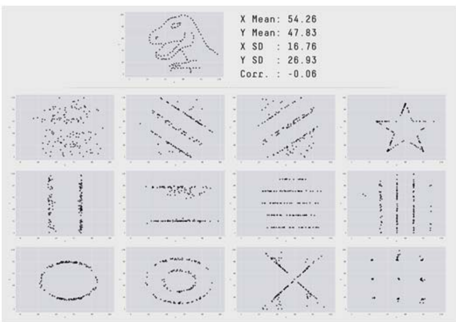
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

*The **Datasaurus Dozen***



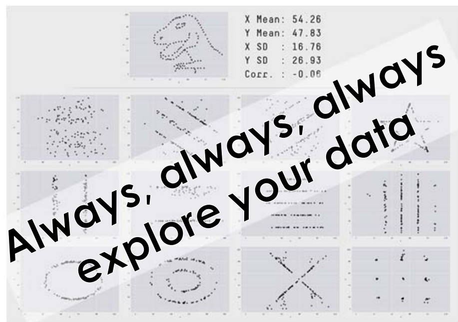
https://www.autodeskresearch.com/publications/samestats

*The **Datasaurus Dozen***



https://www.autodeskresearch.com/publications/samestats

*The **Datasaurus Dozen***



https://www.autodeskresearch.com/publications/samestats

- **Meta data** – comes in data dictionary
- Semantics of data – what it means?
- e.g., Column names in tables

Data **row**, or data **item**, or **observation**, or **sample**

country	year	e_pop_num	e_prev_100k	e_mort_exc_tbhiv	e_mort_exc_tb	e_mort_exc_tbhiv_	source_mort	e_inc_100k	source_tbhiv
Afghanistan	1990	11731193	327	7.3	72	8500	Indirect	117	Model
Afghanistan	1991	12612043	359	9.1	78	9800	Indirect	147	Model
Afghanistan	1992	13811876	387	11	83	12000	Indirect	131	Model
Afghanistan	1993	15175325	412	13	89	14000	Indirect	139	Model
Afghanistan	1994	16485018	431	15	95	16000	Indirect	147	Model
Afghanistan	1995	17586073	447	17	100	18000	Indirect	155	Model
Afghanistan	1996	18415307	461	19	106	19000	Indirect	155	Model
Afghanistan	1997	19021226	469	20	111	21000	Indirect	155	Model
Afghanistan	2012	29824536	358	15	68	20000	Indirect	156	Surveillance
Albania	1990	3446882	36	2.3	3.9	130	VR imputed	18	Surveillance
Albania	1991	3459763	35	2	3.4	120	VR imputed	18	Surveillance
Albania	1992	3446858	34	0.75	1.5	51	VR	17	Surveillance
Albania	1993	3417280	33	1.1	1.9	66	VR	18	Surveillance
Albania	1994	3384367	33	2.6	4.7	160	VR	21	Surveillance
Albania	1995	3357858	32	0.55	0.83	28	VR	20	Surveillance
Albania	1996	3341043	32	1.2	1.8	59	VR	22	Surveillance
Albania	1997	3331317	32	0.66	1.1	35	VR	21	Surveillance
Albania	1998	3325456	36	0.73	1.2	39	VR	22	Surveillance
Albania	1999	3317941	41	0.59	1	34	VR	23	Surveillance
Albania	2000	3304948	30	0.57	1.1	37	VR	19	Surveillance
Albania	2001	3286084	26	0.5	0.89	29	VR	18	Surveillance
Albania	2002	3263596	31	0.5	0.94	31	VR	19	Surveillance
Albania	2003	3239385	29	0.43	0.82	27	VR	18	Surveillance
Albania	2004	3216197	30	0.44	0.85	27	VR	18	Surveillance
Albania	2005	3196130	27	0.39	0.79	25	VR imputed	17	Surveillance
Albania	2006	3179573	25	0.36	0.75	24	VR imputed	15	Surveillance
Albania	2007	3166222	22	0.33	0.7	22	VR imputed	15	Surveillance
Albania	2008	3156608	22	0.29	0.66	21	VR imputed	14	Surveillance
Albania	2009	3151185	24	0.26	0.62	20	VR imputed	15	Surveillance

Data **column**, or data **dimension**, or **variable**, or **attribute**, or **feature**

What are the types of these columns?

country	year	population	# cases per 100k	Mortality rates
Afghanistan	1990	11731193	327	7.3
Afghanistan	1991	12612043	359	9.1
Afghanistan	1992	13811876	387	11
Afghanistan	1993	15175325	412	13
Afghanistan	1994	16485018	431	15
Afghanistan	1995	17586073	447	17
Afghanistan	1996	18415307	461	19
Afghanistan	1997	19021226	469	20
Afghanistan	2012	29824536	358	15
Albania	1990	3446882	36	2.3
Albania	1991	3459763	35	2
Albania	1992	3446858	34	0.75
Albania	1993	3417280	33	1.1
Albania	1994	3384367	33	2.6
Albania	1995	3357858	32	0.55
Albania	1996	3341043	32	1.2
Albania	1997	3331317	32	0.66
Albania	1998	3325456	36	0.73

Categorical

Quantitative Interval

Quantitative - Ratio

Quantitative - Ratio

Quantitative - Ratio

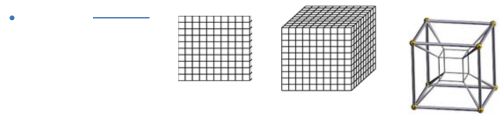
Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- **Perspectives in data types**
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats

Multidimensional data

- Without yet taking into account the *nature* of the data, or the *relationships* between dimensions we could observe the number of dimensions of a certain data set.

Point Line Square Cube *Tesseract*

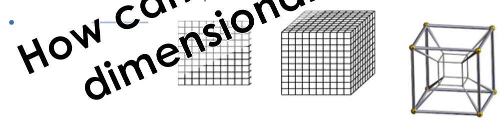


Multidimensional data

- Without yet taking into account the *nature* of the data, or the *relationships* between dimensions we could observe the number of dimensions of a certain data set.

Point Line Square Cube *Tesseract*

How can we have four dimensional data?



Multidimensional data

- Without yet taking into account the *nature* of the data, or the *relationships* between dimensions we could observe the number of dimensions of a

Multidimensional data: Examples

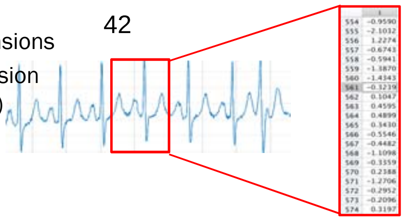
- Zero dimensions

42

Multidimensional data: Examples

- Zero dimensions
- One dimension (Univariate)

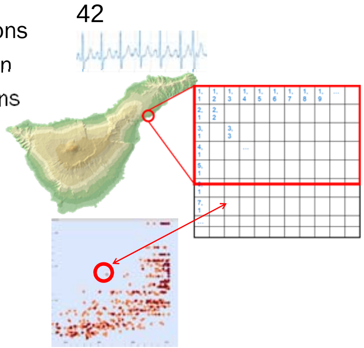
42



Multidimensional data: Examples

- Zero dimensions
- One dimension
- Two dimensions (Bivariate)

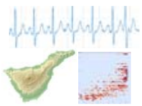
42

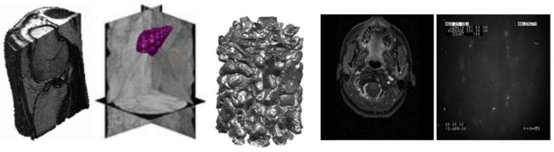


Multidimensional data: Examples

- Zero dimensions
- One dimension
- Two dimensions
- Three dimensions

42

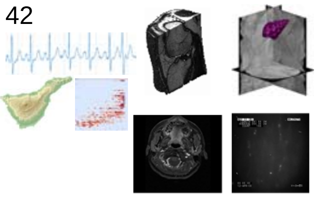


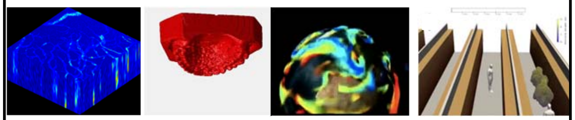


Multidimensional data: Examples

- Zero dimensions
- One dimension
- Two dimensions
- Three dimensions
- More dimensions ...

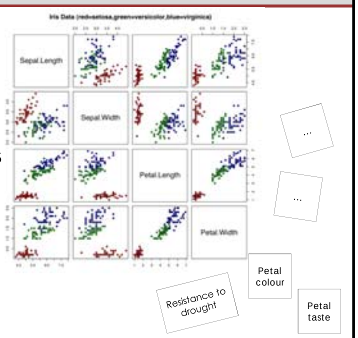
42





Multidimensional data: Examples

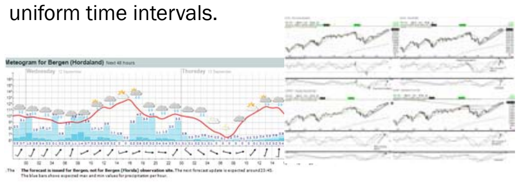
- Zero dimensions
- One dimension
- Two dimensions
- Three dimensions
- More dimensions *not geometrically related*



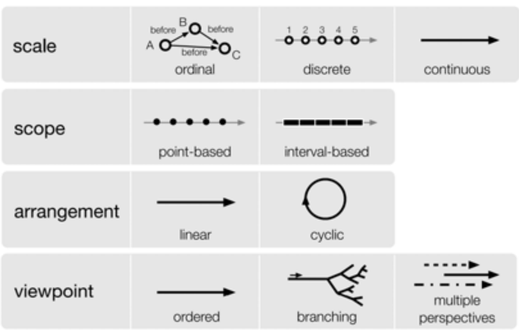
R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. 7 (2): 179–188.

Temporal data

- Data with **temporal information**
- Different names: time series data, functional data (data as a function of time), temporal data
- .. a "*sequence of data points*", measured typically at "successive time instants" spaced at uniform or non-uniform time intervals.



Considerations for temporal data



by Wolfgang Eigner et al., Visualization of Time-Oriented Data

Considerations for temporal data: scale

The **scale** along which elements of the data are given/represented

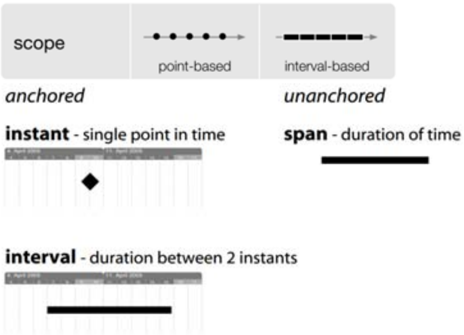
Ordinal

Discrete

Continuous
(as much as possible)

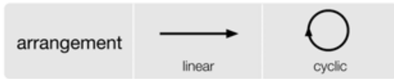
by Wolfgang Eigner et al., Visualization of Time-Oriented Data

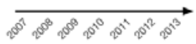
Considerations for temporal data: scope



by Wolfgang Eigner et al., Visualization of Time-Oriented Data

Considerations for temporal data: arrangement





each element of time has a unique predecessor and a unique successor

summer is before winter, but winter is also before summer

by Wolfgang Eigner et al., Visualization of Time-Oriented Data

Considerations for temporal data: view point

Viewpoint: how you decide to view/consider the temporal data in your analysis

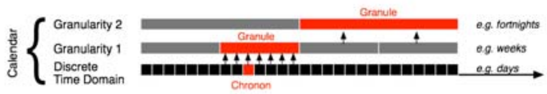
Ordered: consider things that happen one after the other

Branching: multiple strands of time branch out and allow the description and comparison of alternative scenarios (e.g., in project planning). This type of time supports decision-making processes where only one of the alternatives will actually happen.

Multiple perspectives: simultaneous (even contrary) views of time, e.g., eyewitness reports.

by Wolfgang Eigner et al., Visualization of Time-Oriented Data

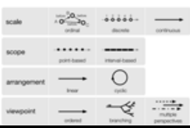
Temporal data – granularity



from Wolfgang Eigner et al., Visualization of Time-Oriented Data

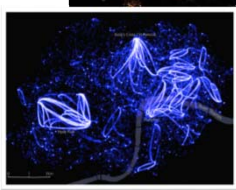
Why would these matter?

- Scale/scope : Which tools to use, how you would derive features (e.g., look at the variance of intervals)
- Arrangement: Analysis of seasonality, yearly vs. weekly cycles
- Viewpoint: How you compare multiple outcomes, e.g., several simulation runs
- Granularity: Extracting micro/macro behavior, e.g., yearly trends vs. hourly trends



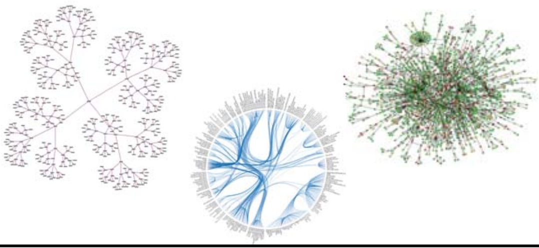
Spatial Data (more in VA)

- Data with an inherent **spatial reference**
- Several examples
  - Satellite readings
  - Phone calls, transactions
  - Land use information
  - Census enumerations
  - Social media activities
  - Photos

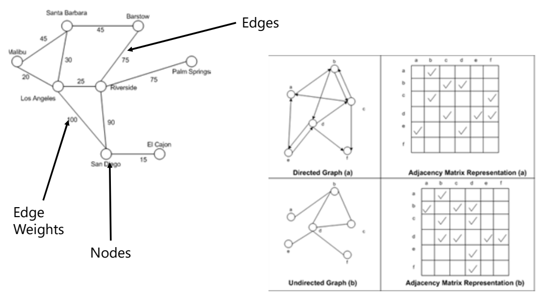


Network data (more later)

- Names: Network, graph, tree
- Encodes relations, hierarchies
- Examples: Social networks, transactions, etc…



Representing network data

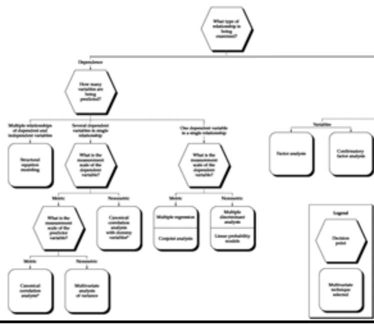


Other perspectives on data

- **Structured vs. unstructured**
  - Structured data: a certain data model, e.g., relational DBs
  - Unstructured data: no pre-defined model
    - Structure can be derived (hopefully)
    - Semi-structured forms are common, e.g., XML, JSON
    - text data (e.g. e-mail messages, word processing documents) videos, photos, audio files, presentations, WEB ! …
- **Static vs. Dynamic (streaming)**
  - Data might **stream** from sources
  - Ex: Twitter API, custom-build data sources, etc…

Why all of these are important?

- It affects the tools we choose, e.g., which multivariate?



Hair, Joseph F., et al. *Multivariate data analysis*. Vol. 6. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

Why all of these are important?

- Inferring data / structures ~ automating processes, e.g.:

https://www.automaticstatistician.com/static/abcdoutput/11-unemployment.pdf

Tableau, *Show me* Feature

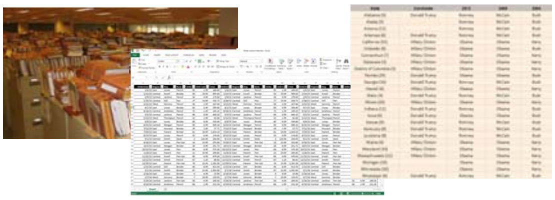
Figure 1: Chart types.

Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- **Wrangling**
- Data formats

Data: *Where? What? How?*

- How accessible are those ZB of data?
- How much can we understand of data *as it is*?
- How much of that data is relevant?
- How clean is the data?



From last week – DS Process

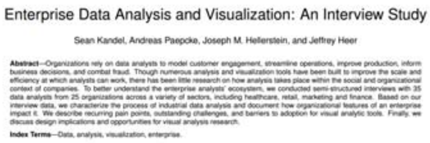
- Understand domain needs
- Collect & make data available
- Get the data ready for analysis
- Exploratively (and visually) analyse the data
- Model the phenomena (if needed)
- Evaluate findings
- ITERATE (from any stage to any other stage)!
- Communicate findings

From last week – Data wrangling & fusion

- **Getting the data ready** to be analysed
- Data is **never perfect** and it is **segregated,** i.e., multiple sources
- Many names: data **wrangling**, data **munging**, data **cleaning**, data **massaging**, data **scrubbing**, **pre-processing**, **data tidying, data curating,....**
- Data **fusion**: merging / integrating several data sources
- Handle **missing data**

On wrangling

*I spend **more than half of my time** integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis. Most of the time once you transform the data you just do an average... the insights can be scarily obvious. It's fun when you get to do something somewhat analytical.*



Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

Abstract—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

Index Terms—Data, analysis, visualization, enterprise.

Kandel, Sean, et al. "Enterprise data analysis and visualization: An interview study.", *IEEE TVCG* (2012)

Ways to cope with this

- Become a ninja wrangler!
- (Be an optimist), **remember that a by-product is that it's helping you understand the data better**
- Use application domain knowledge to only spend time on problems that will give useful results
- Experienced analysts will develop shortcuts and heuristics to know whether to invest more time

Data Quality & Usability Issues

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	errors when combining data

Usability, Credibility & Usefulness

Data is ***usable*** if it can be parsed and manipulated by computational tools. Data usability is thus defined in conjunction with the tools by which it is to be processed.

Data is ***credible*** if, according to one's subjective assessment, it is suitably representative of a phenomenon to enable productive analysis.

Data is ***useful*** if it is usable, credible, and *responsive to one's inquiry*.

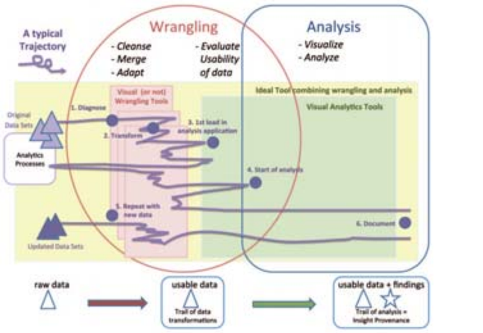
Slide by Jeff Heer

Data Wrangling

A process of iterative data exploration and transformation that enables analysis.
The goal of wrangling is to make data useful:
- Map data to a form readable by downstream tools (database, stats, visualization, …)
- Identify, document, and (where possible) address data quality issues.

Kandel, Sean, et al. "**Research directions in data wrangling**: Visualizations and transformations for usable and credible data." *Information Visualization* (2011)

Data Wrangling



Kandel, Sean, et al. "**Research directions in data wrangling**: Visualizations and transformations for usable and credible data." *Information Visualization* (2011)

Some wrangling steps

- Visualise "raw" data for detection
- Visualise missing/uncertain data
- Transform data
  - Scripts / processes to data
  - Correct errors, e.g., **missing data**
  - Statistical **data transformations**
  - Integrate / merge
- DataWrangler video: https://vimeo.com/19185801

Data Organisation perspective – Tidy data

According Wickham, in a **tidy dataset**:

- Each variable must have its own column.
- Each observation must have its own row.
- Each value must have its own cell.

[*] Wickham, H., 2014. Tidy Data. *Journal of Statistical Software*, *59*(i10).
[**] Image from http://r4ds.had.co.nz/tidy-data.html

Indications of **messy data** (from Wickham, 2014 [*])

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

[*] Wickham, H., 2014. Tidy Data. *Journal of Statistical Software*, *59*(i10)

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

- Can you identify the variables?
- Is this dataset tidy?

Discuss briefly..

from Wickham, H., 2014. Tidy Data. *Journal of Statistical Software*, *59*(i10)

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Tidy Data

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Col names: **Sex:** f-female, m-male **Age intervals:** 0-14, 15-25, 25-34, 35-44, 45-54, 55-64, unknown

- Can you identify the variables?
- Is this dataset tidy?

Discuss briefly..

from Wickham, H., 2014. Tidy Data. *Journal of Statistical Software*, *59*(i10)

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Tidy Data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Missingness mechanisms

- Missing Completely at Random (MCAR)

  an observation being missing does not depend on observed or unobserved measurements

Unit	Variables 1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

  – *Student forgot to answer the question*

Missingness mechanisms

- Missing Completely at Random (MCAR)

  an observation being missing does not depend on observed or unobserved measurements

- Missing At Random (MAR)

  the missingness mechanism

  depends on the observed data but

  not on the unobserved (missing) data

Unit	Variables 1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

  – *Men are more likely to tell you their weight/age than women (is this true????)*

Missingness mechanisms

- Missing Completely at Random (MCAR)
  an observation being missing does not depend on observed or unobserved measurements
- Missing At Random (MAR)
  the missingness mechanism depends on the observed data but not on the unobserved (missing) data

Unit	Variables 1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

- Missing not at random (MNAR)
  - the missingness mechanism depends on missing values
  - Problematic, hard to make statistics
  - *Study about students with anaemia conducted in school (but students did not attend because of anaemia)*

Missingness mechanisms

- Missing Completely at Random (MCAR)
  an observation being missing does not depend on observed or unobserved measurements
- Missing At Random (MAR)
  the missingness mechanism depends on the observed data but not on the unobserved (missing) data

Unit	Variables 1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.6
2	1	3	?	3.5	?	?

- Missing not at random (MNAR)
  - the missingness mechanism depends on missing values
  - Problematic, hard to make statistics
- ***Very hard to know which type!***

Missing data – how to deal?

- Only analyse fully available items (aka Complete Case Analysis)
  – Simple execution
  – Losing observations

Gender	Age	Score
F	32	12
F	44	10
~~M~~	~~55~~	~~?~~
~~M~~	~~?~~	~~45~~
M	13	55
M	44	63
~~F~~	~~56~~	~~?~~
~~?~~	~~?~~	~~12~~
~~F~~	~~31~~	~~?~~

Missing data – how to deal?

- Analyse columns with all available items
  – Less data lost
  – Hard to compare between analyses, samples are different
  – Suitable for aggregated analysis

Gender	Age	Score
F	32	12
F	44	10
M	55	~~?~~
M	~~?~~	45
M	13	55
M	44	63
F	56	~~?~~
~~?~~	~~?~~	12
F	31	~~?~~

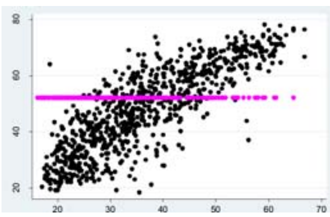
Missing data – how to deal?

- Delete a whole column
  – Only if most of the values are missing in a column
  – Avoids further problems

Gender	Age	Score
F	~~32~~	12
F	~~?~~	10
M	~~?~~	47
M	~~?~~	45
M	~~?~~	55
M	~~44~~	63
F	~~?~~	33
?	~~?~~	12
F	~~31~~	14

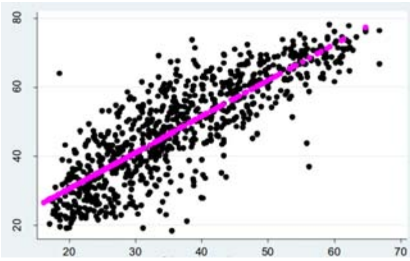
Missing value imputation

- **Mean / mode substitution**
  – Replace missing value with sample mean or mode
  – Reduces variability
  – Weakens covariance and correlation



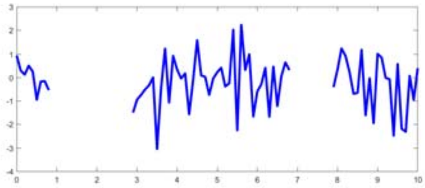
Missing value imputation

- **Regression substitution (deterministic)**
  - replaces missing values with predictions from a regression function



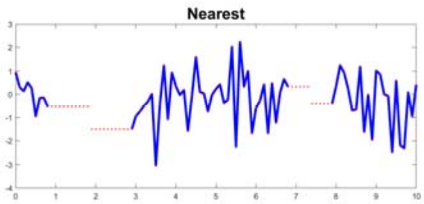
Missing value imputation

- **Interpolation**
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



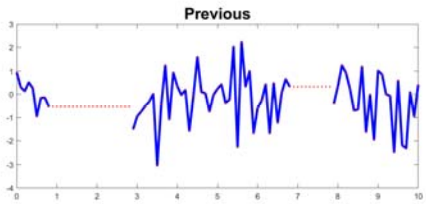
Missing value imputation

- **Interpolation**
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



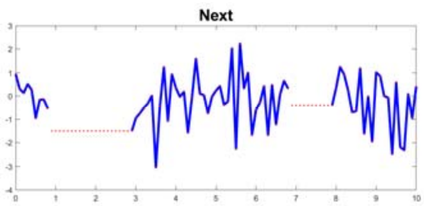
Missing value imputation

- **Interpolation**
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



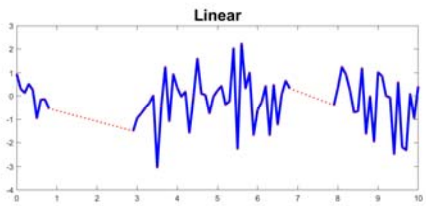
Missing value imputation

- **Interpolation**
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



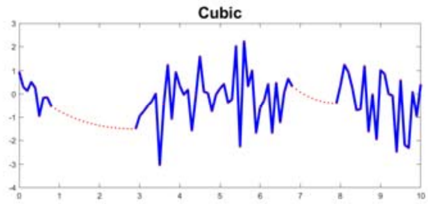
Missing value imputation

- **Interpolation**
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



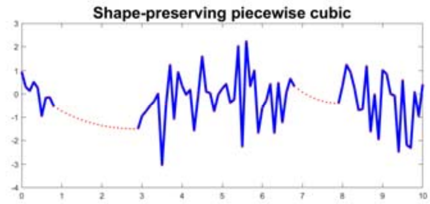
Missing value imputation

- Interpolation
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



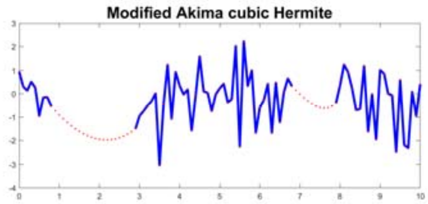
Missing value imputation

- Interpolation
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



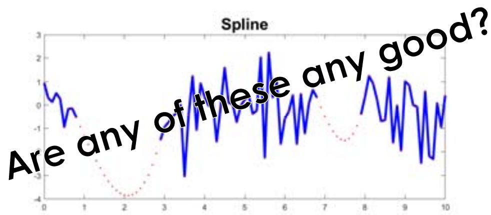
Missing value imputation

- Interpolation
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



Missing value imputation

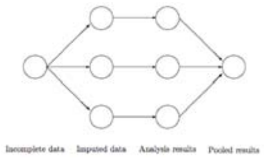
- Interpolation
  - construct new data points within the range of a discrete set of known data points with the help of a model function
  - Several different functions to interpolate how to choose?



A robust way of dealing with missing values

- Multiple Imputation -- Rubin, Donald B. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.

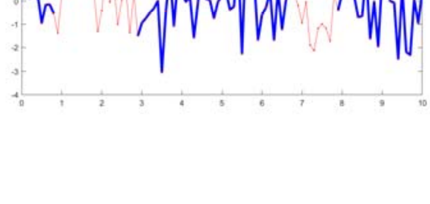
1. **Impute:** Impute missing entries *m* times, each time with a different/randomised model, you end up with m complete data sets
2. **Analyse:** Analyse the data *m* times.
3. **Pool:** Look at variations, generate "pooled" estimates



http://www.stefvanbuuren.nl/mi/MI.html

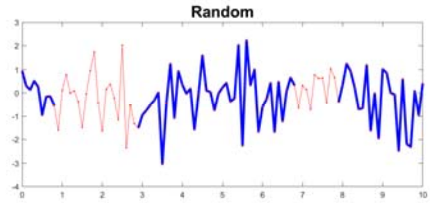
Multiple Imputation

- Visualise to observe the effects:



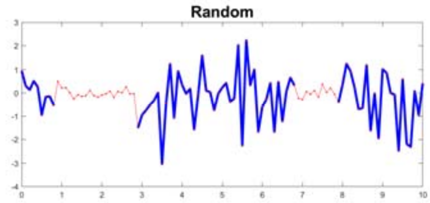
Multiple Imputation

- Visualise to observe the effects:



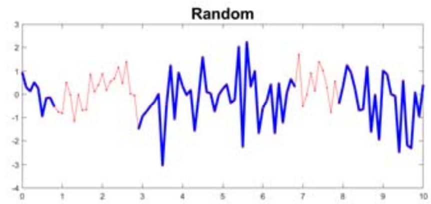
Multiple Imputation

- Visualise to observe the effects:



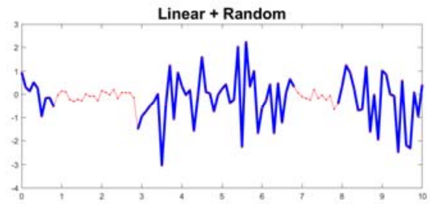
Multiple Imputation

- Visualise to observe the effects:



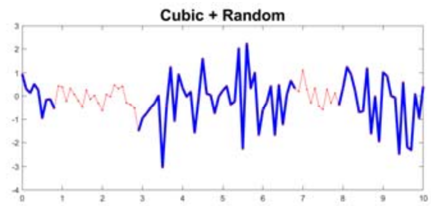
Multiple Imputation

- Visualise to observe the effects:



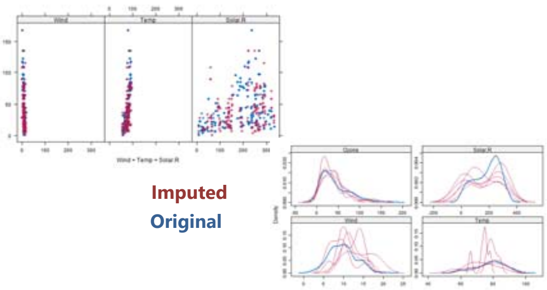
Multiple Imputation

- Visualise to observe the effects:



Multiple Imputation

- Visualise to observe the effects:



From -- http://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/

Missing value imputation

Whatever method is used, ..

Keep a record!

Analytical provenance is important

Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats

XML


```xml
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```


Text documents (structured vs. unstructured)


```
Src,Eqid,Version,Datetime,Lat,Lon,Magnitude,Depth,NST,Region
ci,14692356,1,"Tuesday, May  4, 2010 03:21:38 UTC",32.6443,-115.7605,1.6,3.20,13,"Southern California"
ci,14692348,1,"Tuesday, May  4, 2010 03:19:38 UTC",32.1998,-115.3676,2.5,6.70,12,"Baja California, Mexico"
ci,14692332,1,"Tuesday, May  4, 2010 03:16:56 UTC",32.6756,-115.8655,1.9,5.50,24,"Southern California"
ci,14692324,1,"Tuesday, May  4, 2010 03:08:47 UTC",32.6763,-115.8616,1.6,5.30,20,"Southern California"
ci,14692316,1,"Tuesday, May  4, 2010 03:08:08 UTC",32.6778,-115.8481,1.9,0.10,42,"Southern California"
ci,14692308,1,"Tuesday, May  4, 2010 03:06:20 UTC",32.7071,-116.0431,1.4,10.40,27,"Southern California"
ci,14692300,1,"Tuesday, May  4, 2010 03:01:52 UTC",32.1948,-115.3653,2.6,13.20,13,"Baja California, Mexico"
ak,10047267,1,"Tuesday, May  4, 2010 03:01:04 UTC",61.2695,-149.8942,2.3,31.20,27,"Southern Alaska"
ci,14692284,1,"Tuesday, May  4, 2010 02:58:51 UTC",32.7016,-115.8841,1.7,5.00,18,"Southern California"
ci,14692276,1,"Tuesday, May  4, 2010 02:57:46 UTC",32.6998,-115.8880,2.1,3.60,43,"Southern California"
ak,10047263,1,"Tuesday, May  4, 2010 02:56:28 UTC",63.5779,-150.8288,2.1,4.10,16,"Central Alaska"
ak,10047261,1,"Tuesday, May  4, 2010 02:52:00 UTC",60.4986,-143.0205,1.0,0.00,10,"Southern Alaska"
ci,14692268,1,"Tuesday, May  4, 2010 02:48:40 UTC",32.6813,-116.0371,1.7,10.70,40,"Southern California"
ci,14692260,1,"Tuesday, May  4, 2010 02:35:27 UTC",32.2006,-115.4625,3.0,18.20,24,"Baja California, Mexico"
nc,71392116,0,"Tuesday, May  4, 2010 02:15:24 UTC",38.8415,-122.8287,1.3,2.50,16,"Northern California"
ci,14692244,1,"Tuesday, May  4, 2010 02:05:07 UTC",33.5248,-116.4523,1.1,10.70,26,"Southern California"
ci,14692228,1,"Tuesday, May  4, 2010 01:57:08 UTC",32.6823,-115.8075,1.5,1.50,13,"Southern California"
ci,14692220,1,"Tuesday, May  4, 2010 01:53:28 UTC",32.6881,-116.0515,2.5,11.30,66,"Southern California"
ci,14692212,1,"Tuesday, May  4, 2010 01:48:53 UTC",32.6398,-115.8085,1.9,8.90,30,"Southern California"
ci,14692188,1,"Tuesday, May  4, 2010 01:26:58 UTC",32.5003,-115.6715,1.9,6.40,11,"Baja California, Mexico"
ci,14692180,1,"Tuesday, May  4, 2010 01:19:44 UTC",32.6836,-115.8438,1.6,6.90,18,"Southern California"
ci,14692172,1,"Tuesday, May  4, 2010 01:12:01 UTC",32.5321,-115.7045,1.8,2.90,18,"Baja California, Mexico"
ci,14692164,1,"Tuesday, May  4, 2010 01:08:24 UTC",32.6833,-116.0415,1.8,9.20,42,"Southern California"
```


Strobelt, Hendrik, et al. "Document cards: A top trumps visualization for documents." *Visualization and Computer Graphics, IEEE Transactions on* 15.6 (2009): 1145-1152.

JSON (JavaScript Object Notation)

- Is a lightweight data-interchange format, alternative to XML
- JSON is built on two structures:
  – A collection of **name/value pairs**
  – An ordered **list of values**
- Gaining popularity in web apps
- http://json.org/


```json
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "height_cm": 167.6,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```


http://en.wikipedia.org/wiki/JSON

JSON

XML vs. JSON

XML:


```xml
<?xml version="1.0" encoding="UTF-8"?>
<note private="true">
  <from>Alice Smith (alice@example.com)</from>
  <to>Robert Jones (roberto@example.com)</to>
  <to>Charles Dodd (cdodd@example.com)</to>
  <subject>Tomorrow's "Birthday Bash" event!</subject>
  <message language="english">
    Hey guys, don't forget to call me this weekend!
  </message>
</note>
```


JSON:


```json
{
  "private": "true",
  "from": "Alice Smith (alice@example.com)",
  "to": [
    "Robert Jones (roberto@example.com)",
    "Charles Dodd (cdodd@example.com)"
  ],
  "subject": "Tomorrow's \"Birthday Bash\" event!",
  "message": {
    "language": "english",
    "text": "Hey guys, don't forget to call me this weekend!"
  }
}
```

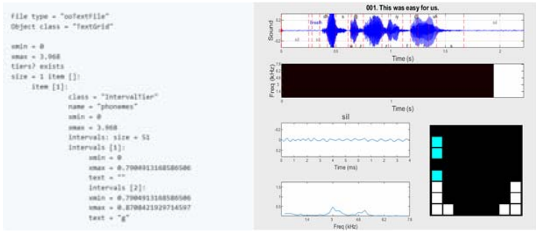
GraphML

- A file format for graphs
- http://graphml.graphdrawing.org/

```xml
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
     http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <graph id="G" edgedefault="undirected">
    <node id="n0"/>
    <node id="n1"/>
    <edge id="e1" source="n0" target="n1"/>
  </graph>
</graphml>
```

TextGrid

- A file format for phonetics
- http://www.fon.hum.uva.nl/praat/



https://github.com/reyesaldasoro/ElectroPalatography
Verhoeven, et al. Visualisation and Analysis of Speech Production with Electropalatography. J. Imaging 2019, 5(3), 40.

Some tools for Data Wrangling

- Programming yourself – Python is good!
- Open Refine (previously Google Refine)
  - Now in transition to OpenRefine
  - Runs as a local server
  - Good for also extending data
  - http://openrefine.org/index.html
- DataWrangler (now TriFacta)
  - Available online
  - Good for splitting / merging / deleting data
  - http://vis.stanford.edu/wrangler/

Collecting data – where to look?

- UK data:
  http://data.gov.uk/data/search
- About London:
  http://data.london.gov.uk/
- US Gov. data repository:
  https://www.data.gov/
- World Bank (on global indicators):
  http://data.worldbank.org/

- Biomedical Literature:
  https://pubmed.ncbi.nlm.nih.gov/
  https://www.ncbi.nlm.nih.gov/
- Biomedical Data (challenges):
  https://grand-challenge.org/challenges

Collecting data – where to look?

- British Library:
  https://data.bl.uk/bl_labs_datasets/
- An extensive collection:
  http://www.kdnuggets.com/datasets/index.html
- Public data from Google:
  http://www.google.com/publicdata/directory
- Another collection of links:
  http://blog.visual.ly/data-sources/

- Kaggle Datasets:
  https://www.kaggle.com/datasets
- Airport data:
  https://www.faa.gov/data_research/
- Rail network:
  https://datafeeds.networkrail.co.uk/

Week 2 Schedule

- Know your data (and your information)
- Data attributes types
- Perspectives in data types
  - Multidimensional data
  - Temporal data
  - Network data
- Wrangling
- Data formats