Data Analysis in Astronomy and Physics

Lecture 1: Introduction to data

M. Röllig

Data basics

Out[ ]=

class	age	sex	survived
1st	29	female	True
1st	1	male	True
1st	2	female	False
1st	30	male	False
1st	25	female	False
1st	48	male	True
1st	63	female	True
1st	39	male	False
1st	53	female	True
1st	71	male	False
1st	47	male	False
1st	18	female	True
1st	24	female	True
1st	26	female	True
1st	80	male	True

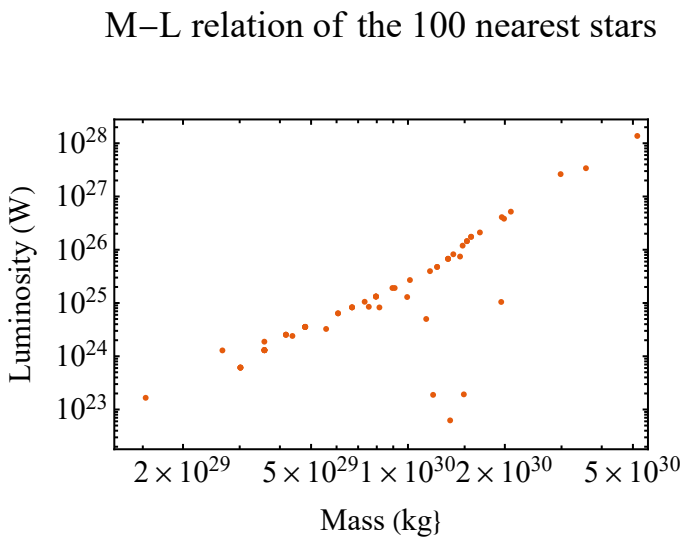
rows 1–15 of 1309

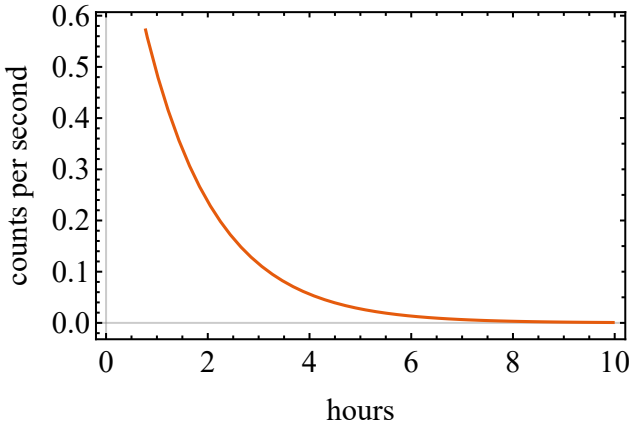
rows:	observation (case)	e.g.: 1st row	→ 1st passenger's data
columns:	variable	e.g.: 2nd column	→ passenger's age

Relationship between variables

- Two variables that show some connection with one another are called **associated** (dependent).
- Association can be **positive** or **negative**.
- If two variables are not associated, they are said to be **independent**.

Out[ ]=





Types of variables

- numerical (quantitative) variables
  - continuous
  - discrete
- categorical (qualitative) variables
  - regular categorical
  - ordinal (have inherent ordering)

Variables that show a connection are called **associated** (**positive** or **negative**) or **dependent**. Unassociated variables are called **independent**.

Sample size

Early smoking research started in the US in the 1930's.

Some smokers seemed to be affected by cigarette smoke others appeared completely unaffected.

Problems: Anecdotal evidence

"My uncle smokes three packs a day and he's in perfectly good health"

This may be true, but is based on a limited sample size that might not be representative of the population.

It was concluded that:
"smoking is a complex human behavior, by its nature difficult to study, confounded by human variability"

Sample size

Early smoking research started in the US in the 1930's.

Some smokers seemed to be affected by cigarette smoke others appeared completely unaffected.

Today: much larger sample sizes

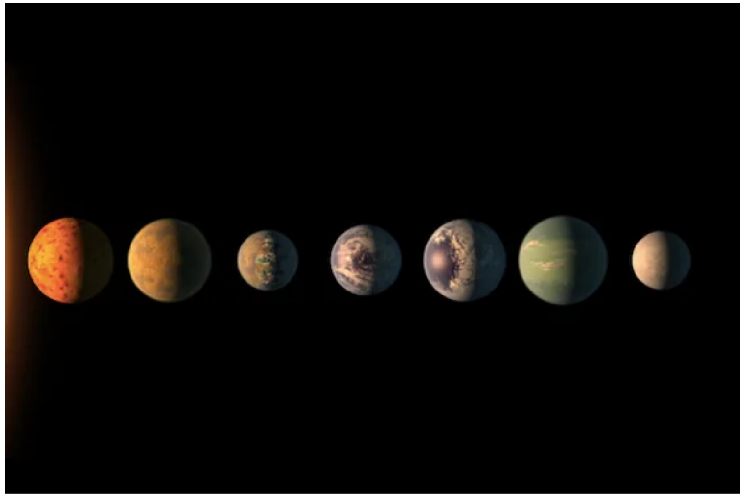
⟶ trends showing negative health impacts of smoking became much clearer

(Source: Dr. Mine Çetinkaya-Rundel, Duke University)

Representative sampling

Planets in a Kepler Multi-planet System Are Similar in Size and Regularly Spaced (2018)

Question: How special is the solar system? Is it a template for extrasolar planetary systems?
Population: All planets in Milky Way. Goal is to understand the fundamental physics of of the star- and planet formation processes.
Sample: Solar system, compared to Kepler data
Generalize to: all planetary systems



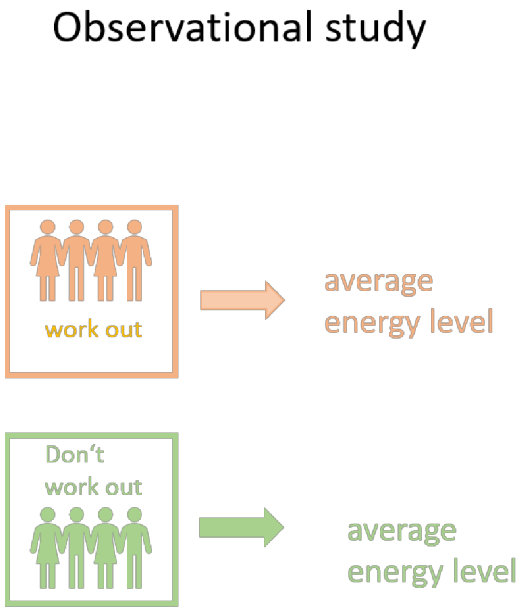
⟶ planets in the same system tend to be the same size.
⟶ the orbital distance between the first pair of planets is a good predictor of the orbital distance of the third planet, and the fourth, and so on

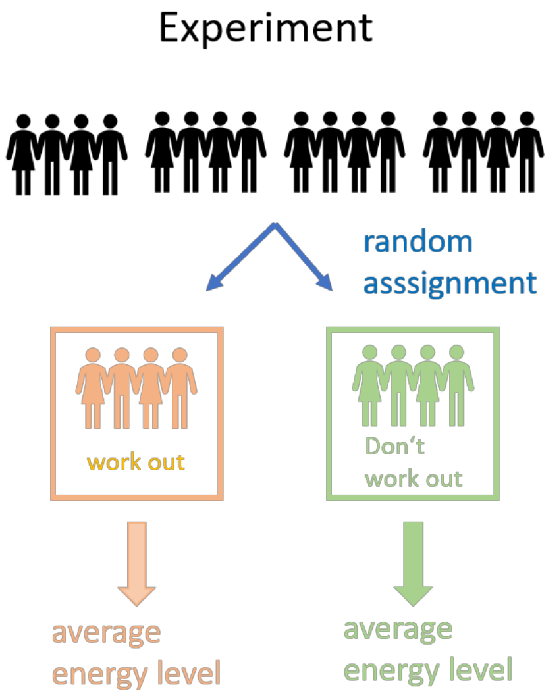
(Source: Weiss et al. (2018), ApJ, 155,1, https://arxiv.org/abs/1706.06204)

(Statistical) Studies

- observational
  - collect data in a way that does not directly interfere with how the data arise
  - only establish an association
  - **retrospective**: uses past data
  - **prospective**: data are collected throughout the study
- experiment
  - randomly assign subjects to treatment (set up an experiment with minimized bias and confound influence)
  - establish causal connections
- astronomy: difficult to set up experiments

(Statistical) Studies





Association and Causation

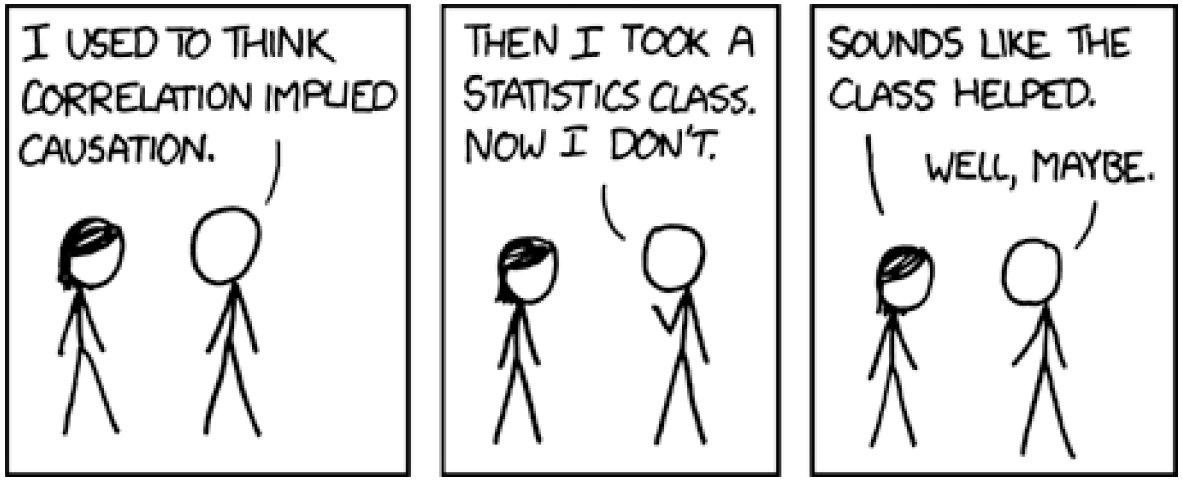
What is causation?

A relationship between two variables, the explanatory and the response, are said to be causal if the **change in the explanatory variable actually causes a change in the response variable**. The explanatory variable is usually the independent variable, and that is usually graphed on the x-axis.

Association and Causation

What is association or correlation?

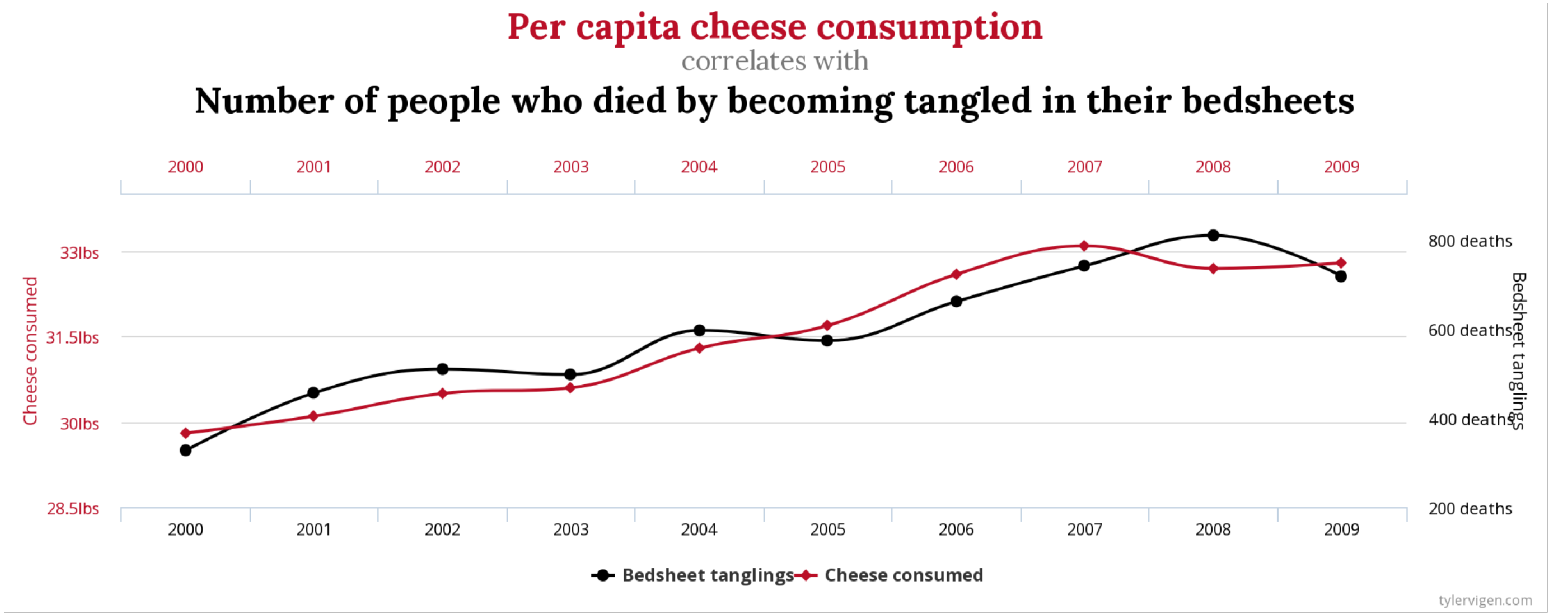
Both of these words are used to describe **relationships that exist between two or more variables**. Correlation tends to be used when describing the strength of a relationship between two numerical variables (such as height and arm span).



Explaining the association

Possible explanations for the observed association are:

- The explanatory variable is actually causing a change in the response variable.
- The response variable is actually causing a change in the explanatory variable.
- **Confounding**: there may be causation, but there are too many uncontrolled variables.
- **Common response** or **coincidence**: no causation, the association can be explained by other variables associated with both the explanatory and response variables.



Sampling & sources of bias

Census - sampling the entire population

- Some members of a population are more difficult to measure than others (sometimes those are significantly different from the rest of the population)
- often impossible or at least impractical
  - limited by resources (money, time, man-power, technology)
  - not all members accessible (e.g. visible universe limits)
  - populations are not static (evolving)

Sampling - measure a representative sample of individuals

Sampling vs. census is like tasting a spoonful of soup vs. eating the whole soup to check its taste.

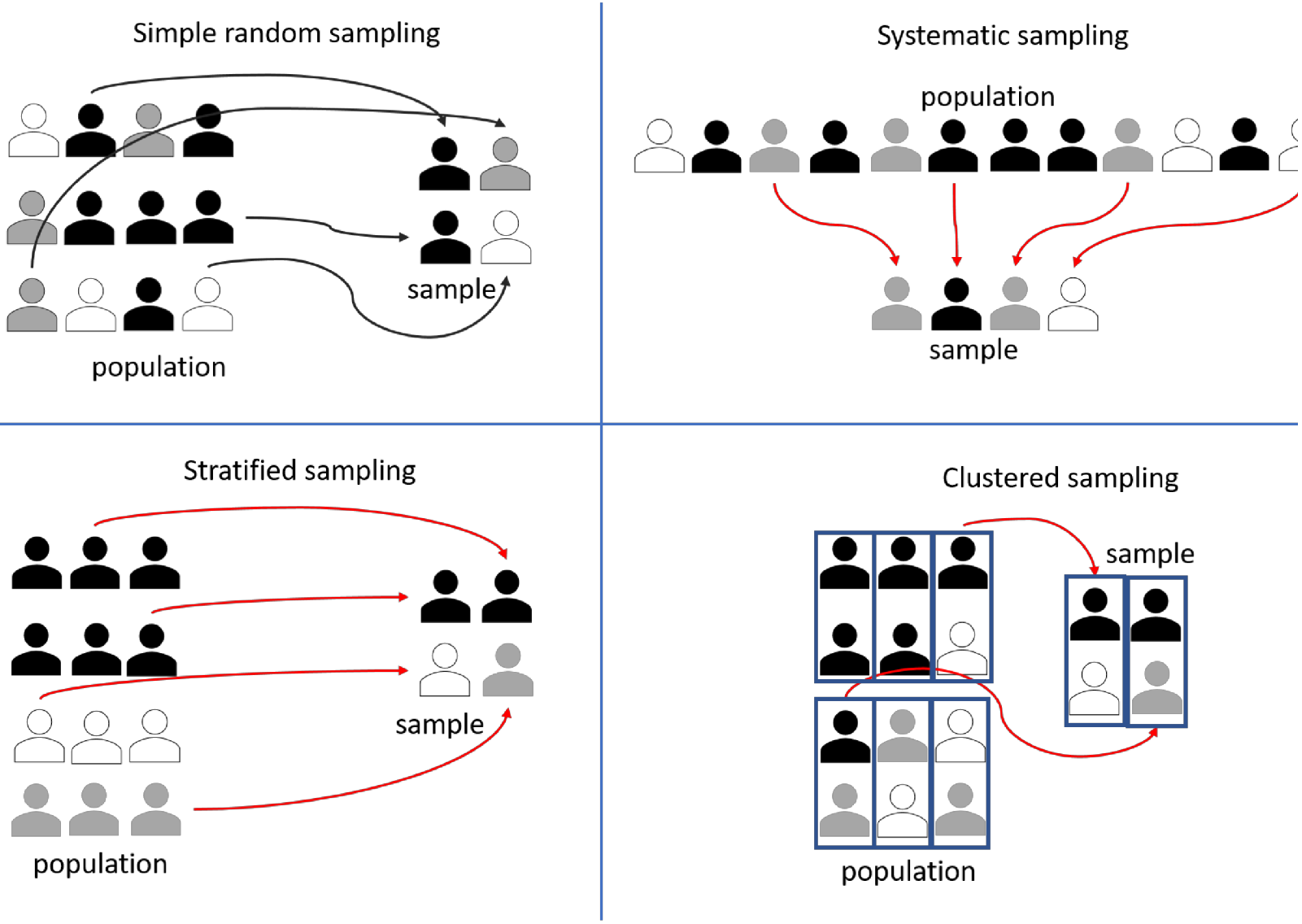
Sampling & sources of bias

- **convenience sample:** easily accessible members are more likely to be sampled (e.g. measure only nearby stars) .
- Selection from a **specific real area:** limits the applicability to the sampled area; extreme form of biased sampling, because certain members of the population are totally excluded from the sample.
- **Self-selection bias:** possible whenever the group of people being studied has any form of control over whether to participate.
- **Pre-screening** of trial participants, or **advertising** for volunteers within particular groups.
- **Exclusion** bias: results from exclusion of particular groups from the sample.
- **Caveman effect:** Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives

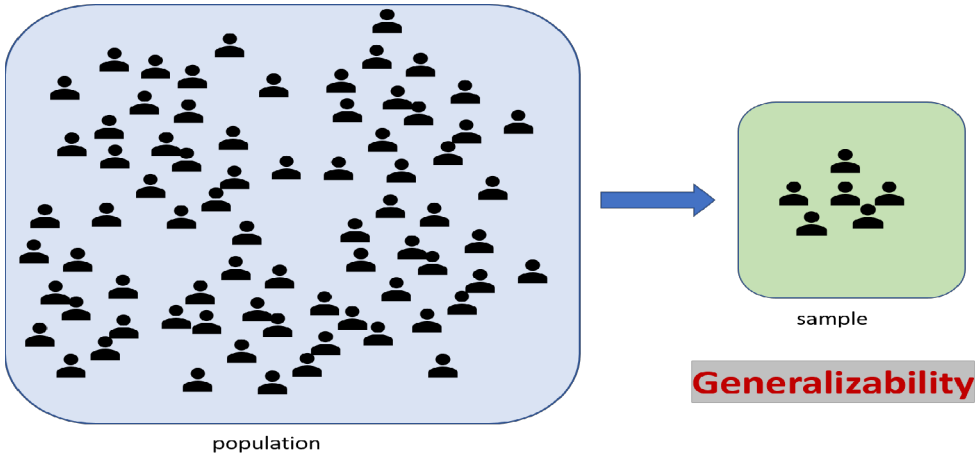
Sampling methods

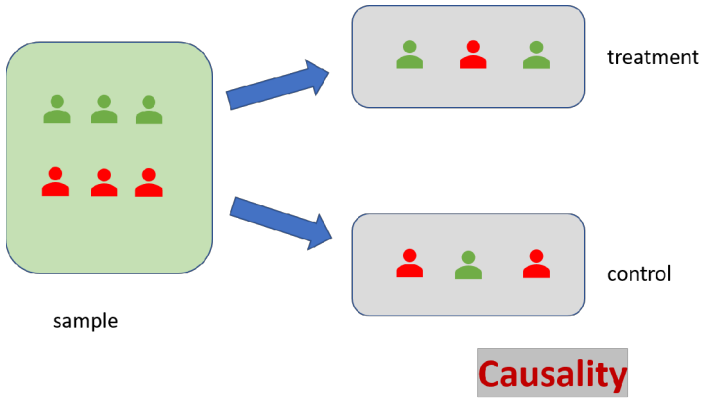
- **Simple random sampling:** Each individual is chosen randomly and entirely by chance, such that each individual has the same probability of being chosen at any stage during the sampling process. (randomness ⟶ small samples maybe not representative)
- **Systematic sampling:** arrange the study population according to some ordering scheme and then select elements at regular intervals through that ordered list. Needs a random starting point! (vulnerable to periodicity)
- **Stratified sampling:** When population splits into distinct categories,it can be organized by these categories into separate “strata.” Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.
- **Cluster sampling:** select respondents in groups (‘clusters’), e.g. clustered by geography, or by time periods.

Sampling methods


Simple random sampling
sample
population
Systematic sampling
population
sample
Stratified sampling
sample
population
Clustered sampling
sample
population

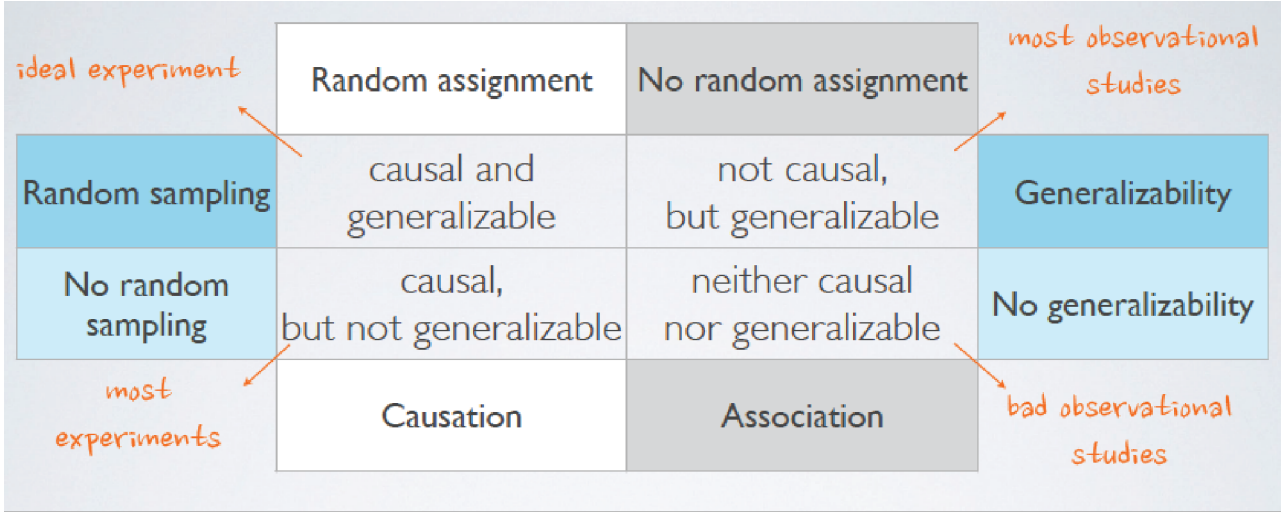
Random sampling & assignment





random sampling

random assignment



(Source : "Data Analysis and Statistical Inference", coursera)

Data File Formats

Plain Text Files - ASCII

- simplest and most intuitive data format
- human & machine readable
- storage friendly because highly compressible
- fast to process
- The simplest way to spot an ASCII file is to try and open it - if you can read the contents then it's a text file

Data File Formats

Plain Text Files - ASCII

- various incarnations:
  - CSV - comma-separated variables (python: PANDAS pd.read_csv)

Out[ ]=

```
1,"Eldon Base for stackable storage shelf, platinum",Muhammed MacIntyre,3,-213.25,38.94,35,Nunavut,Storag
2,"1.7 Cubic Foot Compact ""Cube"" Office Refrigerators",Barry French,293,457.81,208.16,68.02,Nunavut,App
3,"Cardinal Slant-D® Ring Binder, Heavy Gauge Vinyl",Barry French,293,46.71,8.69,2.99,Nunavut,Binders and
4,R380,Clay Rozendal,483,1198.97,195.99,3.99,Nunavut,Telephones and Communication,0.58
5,Holmes HEPA Air Purifier,Carlos Soltero,515,30.94,21.78,5.94,Nunavut,Appliances,0.5
6,G.E. Longer-Life Indoor Recessed Floodlight Bulbs,Carlos Soltero,515,4.43,6.64,4.95,Nunavut,Office Furn
7,"Angle-D Binders with Locking Rings, Label Holders",Carl Jackson,613,-54.04,7.3,7.72,Nunavut,Binders a
8,"SAFCO Mobile Desk Side File, Wire Frame",Carl Jackson,613,127.70,42.76,6.22,Nunavut,Storage & Organizat
9,"SAFCO Commercial Wire Shelving, Black",Monica Federle,643,-695.26,138.14,35,Nunavut,Storage & Organiza
10,Xerox 198,Dorothy Badders,678,-226.36,4.98,8.33,Nunavut,Paper,0.38
11,Xerox 1980,Neola Schneider,807,-166.85,4.28,6.18,Nunavut,Paper,0.4
12,Advantus Map Pennant Flags and Round Head Tacks,Neola Schneider,807,-14.33,3.95,2,Nunavut,Rubber Bands
13,Holmes HEPA Air Purifier,Carlos Daly,868,134.72,21.78,5.94,Nunavut,Appliances,0.5
14,"DS/HD IBM Formatted Diskettes, 200/Pack - Staples",Carlos Daly,868,114.46,47.98,3.61,Nunavut,Compute
15,"Wilson Jones 1"" Hanging DublLock® Ring Binders",Claudia Miner,933,-4.72,5.28,2.99,Nunavut,Binders a
16,Ultra Commercial Grade Dual Valve Door Closer,Neola Schneider,995,782.91,39.89,3.04,Nunavut,Office Fur
17,"#10-4 1/8"" x 9 1/2"" Premium Diagonal Seam Envelopes",Allen Rosenblatt,998,93.80,15.74,1.39,Nunavut,
18,Hon 4-Shelf Metal Bookcases,Sylvia Foulston,1154,440.72,100.98,26.22,Nunavut,Bookcases,0.6
```

  - TSV - tabulator separated variables

Data File Formats

Plain Text Files - ASCII

- fixed format, i.e. certain variables at fixed positions in the file

Out[ ]=

```
        24        29       781       781B
E STATE HAS LABEL B
VIBRATIONAL LEVELS FOR STATE E IN 1/CM FOR J= 23
    -20435.50756    -19362.69223    -18315.23192    -17293.58333
   -16298.04782    -15328.77220    -14385.79385    -13469.04461
   -12578.37473    -11713.58112    -10874.41397    -10060.61653
    -9271.90118     -8507.96930     -7768.55842     -7053.47822
    -6361.95676     -5694.93026     -5042.77653     -4431.21054
    -3836.51776     -3262.91040     -2717.73141     -2191.57359
    -1704.83974     -1246.84178      -818.16915      -439.12935
     -122.71286
PROPORTION OF STATE B:
 0.999160E+00 0.998966E+00 0.998764E+00 0.998558E+00 0.998348E+00
 0.998136E+00 0.997925E+00 0.997714E+00 0.997505E+00 0.997297E+00
 0.997091E+00 0.996884E+00 0.996674E+00 0.996451E+00 0.996184E+00
 0.995488E+00 0.995195E+00 0.995004E+00 0.731953E+00 0.994487E+00
 0.988609E+00 0.992449E+00 0.992169E+00 0.933905E+00 0.992634E+00
 0.974727E+00 0.989640E+00 0.991270E+00 0.849077E+00
PROPORTION OF STATE C:
 0.817661E-03 0.100733E-02 0.120530E-02 0.140839E-02 0.161560E-02
 0.182444E-02 0.203364E-02 0.224229E-02 0.244985E-02 0.265608E-02
 0.286109E-02 0.306741E-02 0.327714E-02 0.349935E-02 0.376584E-02
 0.446197E-02 0.475478E-02 0.494644E-02 0.267703E+00 0.546387E-02
 0.113383E-01 0.750095E-02 0.778409E-02 0.659739E-01 0.732173E-02
 0.252111E-01 0.103074E-01 0.865983E-02 0.150505E+00
PROPORTION OF STATE BP:
 0.281645E-05 0.611953E-05 0.933143E-05 0.123989E-04 0.152856E-04
 0.179634E-04 0.204160E-04 0.226358E-04 0.246203E-04 0.263712E-04
 0.278942E-04 0.291969E-04 0.302869E-04 0.311777E-04 0.318965E-04
```

Data File Formats

Plain Text Files - ASCII

- variations of the above and free format files

Out[ ]=

```
Climatic Research Unit Country File created on Thu  2 Jul 2015 18:16:18 BST, from C
RU TS run #1506241137
Country = Afghanistan          : Parameter = Precipitation                : Unit
= mm/month
Period = 1901.2014 : missing value = -999.0 : format = (i5,17f8.1)
 YEAR     JAN     FEB     MAR     APR     MAY     JUN     JUL     AUG     SEP     O
CT     NOV     DEC     MAM     JJA     SON     DJF     ANN
 1901    65.4    13.5    46.8    32.5    50.9    20.1     7.8     2.6     3.9    11
.5     7.7     7.7   130.1    30.6    23.1    48.6   270.4
 1902    19.8    21.1    41.6    33.8    14.1     5.2     2.5     3.3     1.9    19
.8    53.9    19.0    89.5    11.0    75.6   117.9   236.0
 1903    49.0    49.8    68.2    38.1    80.6     7.1     7.4     6.6     6.3     3
.0    12.0    21.4   186.8    21.1    20.2   113.8   348.4
 1904    67.8    24.6    77.2    23.6    28.2     0.3     3.0     4.1     8.6    21
.6    26.3    20.4   129.0     7.4    56.4   128.8   305.6
 1905    71.3    37.1    71.4    34.4    15.8     4.5     1.1     3.6     5.5
```

Data File Formats

Plain Text Files - ASCII

- no structured information
- difficult to include explanatory information (meta-information) without disturbing the form of the content

Data File Formats

Binary Files

Technically, every computer data file is a binary data file.

Binary files use the full byte range. Readable characters are (in most character sets) limited to the byte values 32 - 126.

If a binary file is opened in a text editor, each group of eight bits will typically be translated as a single character, and the user will see a (probably unintelligible) display of textual characters. If the file is opened in some other application, that application will have its own use for each byte.

Using all bit combinations allows a higher information density.

Data File Formats

Binary Files

- smaller file size
- faster read/write
- combination of different data types much easier possible (e.g. image & text)
- system-dependent encoding complicates exchange (e.g. Little vs. Big Endian)

Out[ ]=

C:\Users\roellig\Documents\Projekte\Teaching\Vorlesung\Data Analysis\2019\Notes\lena512.bmp

```
            0  1  2  3  4  5  6  7  8  9  a  b  c  d  e  f
00000000h: 42 4D 36 04 04 00 00 00 00 00 36 04 00 00 28 00 ; BM6.......6...(.
00000010h: 00 00 00 02 00 00 00 02 00 00 01 00 08 00 00 00 ; ................
00000020h: 00 00 00 00 04 00 00 00 00 00 00 00 00 00 00 01 ; ................
00000030h: 00 00 00 00 00 00 00 00 00 00 01 01 01 00 02 02 ; ................
00000040h: 02 00 03 03 03 00 04 04 04 00 05 05 05 00 06 06 ; ................
00000050h: 06 00 07 07 07 00 08 08 08 00 09 09 09 00 0A 0A ; ................
00000060h: 0A 00 0B 0B 0B 00 0C 0C 0C 00 0D 0D 0D 00 0E 0E ; ................
00000070h: 0E 00 0F 0F 0F 00 10 10 10 00 11 11 11 00 12 12 ; ................
00000080h: 12 00 13 13 13 00 14 14 14 00 15 15 15 00 16 16 ; ................
00000090h: 16 00 17 17 17 00 18 18 18 00 19 19 19 00 1A 1A ; ................
000000a0h: 1A 00 1B 1B 1B 00 1C 1C 1C 00 1D 1D 1D 00 1E 1E ; ................
000000b0h: 1E 00 1F 1F 1F 00 20 20 20 00 21 21 21 00 22 22 ; ......   .!!!."
000000c0h: 22 00 23 23 23 00 24 24 24 00 25 25 25 00 26 26 ; ".###.$$$.%%%.&&
000000d0h: 26 00 27 27 27 00 28 28 28 00 29 29 29 00 2A 2A ; &.'''.(((.))).**
000000e0h: 2A 00 2B 2B 2B 00 2C 2C 2C 00 2D 2D 2D 00 2E 2E ; *.+++.,,,.---...
000000f0h: 2E 00 2F 2F 2F 00 30 30 30 00 31 31 31 00 32 32 ; ../// .000.111.22
00000100h: 32 00 33 33 33 00 34 34 34 00 35 35 35 00 36 36 ; 2.333.444.555.66
00000110h: 36 00 37 37 37 00 38 38 38 00 39 39 39 00 3A 3A ; 6.777.888.999.::
00000120h: 3A 00 3B 3B 3B 00 3C 3C 3C 00 3D 3D 3D 00 3E 3E ; :.;;;.<<<.===.>>
00000130h: 3E 00 3F 3F 3F 00 40 40 40 00 41 41 41 00 42 42 ; >.???.@@@.AAA.BB
00000140h: 42 00 43 43 43 00 44 44 44 00 45 45 45 00 46 46 ; B.CCC.DDD.EEE.FF
```

Specialized Data Formats

FITS - Flexible Image Transport System

- file extensions: .fits, .fit, .fts
- Developed for astronomical data (Wells, D. C.; Greisen, E. W.; Harten, R. H. (June 1981). "FITS: A Flexible Image Transport System". Astronomy and Astrophysics Supplement Series. 44: 363–370, 1981A&AS...44..363W)
- open standard defining a digital file format useful for storage, transmission and processing of data: formatted as N-dimensional arrays (for example a 2D image), or tables.

Specialized Data Formats

FITS - Flexible Image Transport System

- A major feature of the FITS format is that image metadata is stored in a human-readable ASCII header, so that an interested user can examine the headers to investigate a file of unknown provenance.
- FITS is also often used to store non-image data, such as spectra, photon lists, data cubes, or even structured data such as multi-table databases.
- FITS image headers can contain information about one or more scientific coordinate systems that are overlaid on the image itself.

FITS viewer: https://heasarc.gsfc.nasa.gov/ftools/fv/

Specialized Data Formats

HDF5 - Hierarchical Data Format 5

- file extensions: .hdf5, .h5
- free and open source - The HDFGroup: https://www.hdfgroup.org/
- supports n-dimensional datasets and each element in the dataset may itself be a complex object.
- is a self-describing file format, meaning everything all data and metadata can be passed along in one file.
- is high-performance I/O with a rich set of integrated performance features that allow for access time and storage space optimizations.

Specialized Data Formats

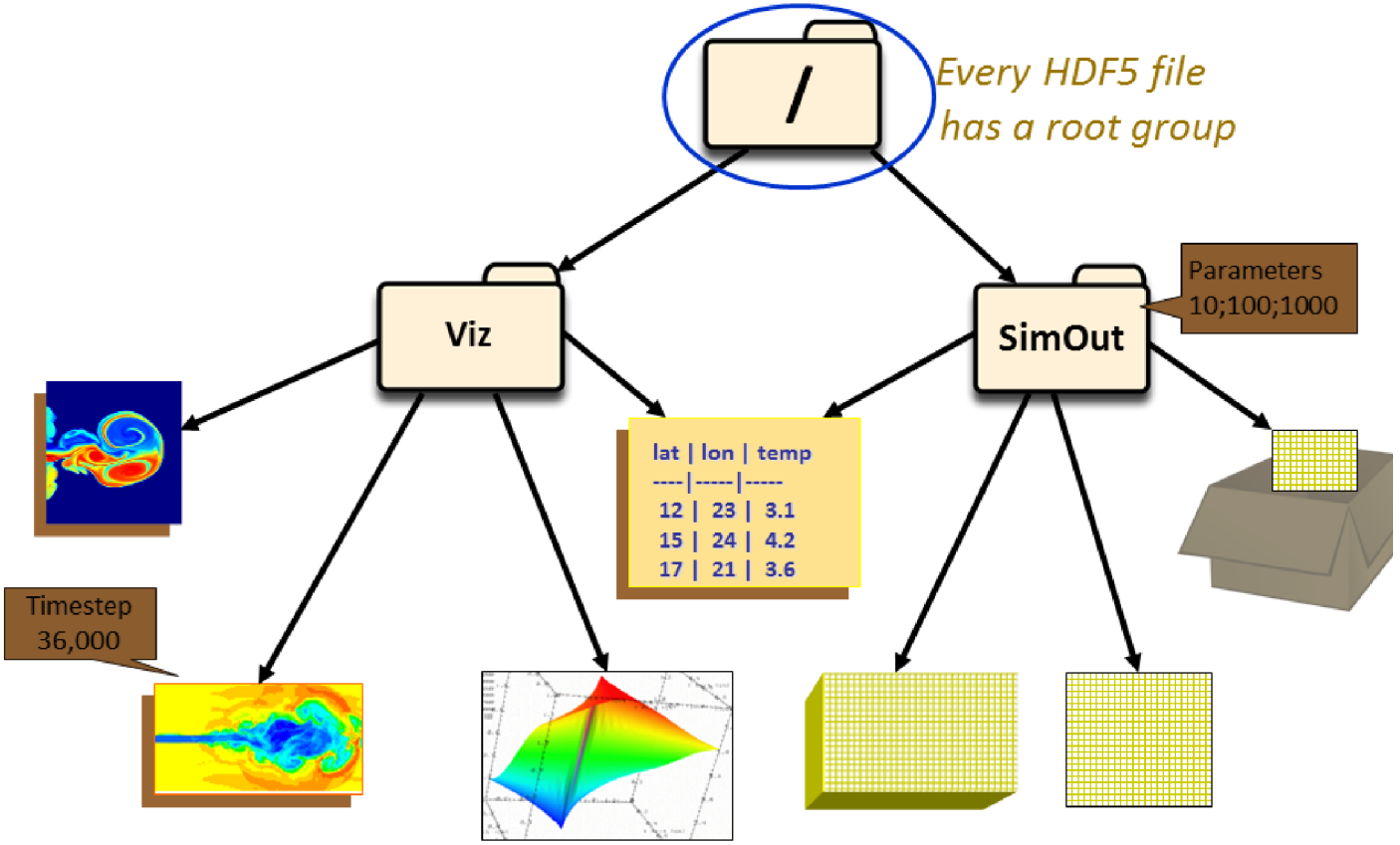
HDF5 - Hierarchical Data Format 5

- has no limit on the number or size of data objects in the collection, giving great flexibility for big data.
- allows you to keep the metadata with the data, streamlining data lifecycles and pipelines.

Specialized Data Formats

HDF5 - Hierarchical Data Format 5

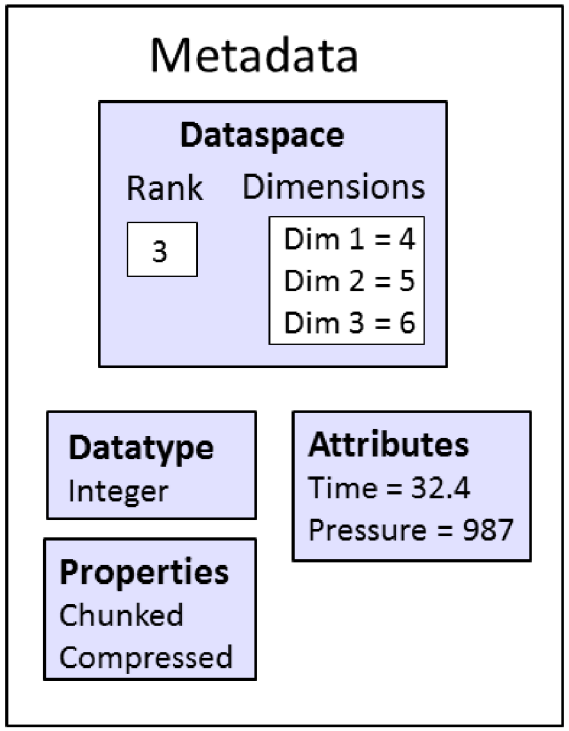
- Every HDF5 file contains a root group that can contain other groups or be linked to objects in other files.

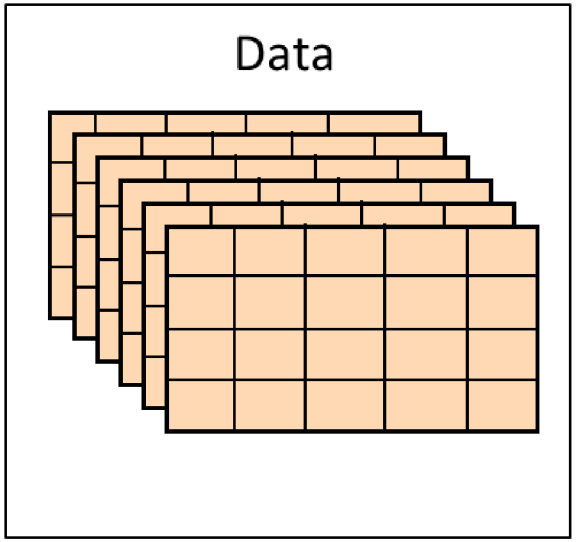


Specialized Data Formats

HDF5 - Hierarchical Data Format 5

- HDF5 datasets organize and contain the “raw” data values. A dataset consists of metadata that describes the data, in addition to the data itself:





Specialized Data Formats

Many other formats

- HDF4 is the older version of the format
- NetCDF (Network Common Data Form)
- JSON (JavaScript Object Notation)
- XML (Extensible Markup Language)
- HDFS (Hadoop Distributed File System)
- …

Init