

«Data is the new oil»



```
186 #define XNU_NO_UTUN_IFACE_NAME -3
187 #define XNU_UTUN_UNSUPPORTED -2
188 #define XNU_UTUN_IN_USE -1
189
190
191 static int xnu_open_utun(int utunnum)
192 {
193     int fd;
194     struct sockaddr_ctl sc;
195     struct ctl_info ctlInfo;
196
197     if (strncpy(ctlInfo.ctl_name, UTUN_CONTROL_NAME, sizeof(ctlInfo.ctl_name)) >=
198         sizeof(ctlInfo.ctl_name))
199     {
200         printf("Opening utun: UTUN_CONTROL_NAME too long\n");
201         return XNU_UTUN_UNSUPPORTED;
202     }
203
204     //open socket
205     fd = socket(PF_SYSTEM, SOCK_DGRAM, SYSPROTO_CONTROL);
206     if (fd < 0) {
207         printf("Opening utun (%s): %s", "socket(SYSPROTO_CONTROL)", strerror (errno));
208     }
209
210     : %s", "ioctl(CTLIOCGINFO)", strerror (errno)
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
```

Interactive Visualization: Data Literacy

ZHdK BA Interaction Design
4/8/15. November 2019

Timo Grossenbacher



- BSc/MSc in **Geography** UZH, Minor **Computer Science**
- **Since November 2014 with SRF Data** as Data Journalist
- **Lecturer** @UZH / @ZHdK / @SfGZ / @DataCamp
- Twitter: @grssnbchr
- Website: **timogrossenbacher.ch**



Now, what about you?

Why this course?

Data Literacy: Course contents

	This morning (830-1030): Introduction to Data, Information and Knowledge
Program	<ul style="list-style-type: none">● Introduction● What exactly <i>are</i> data?● Digital vs. analog data
Readings (links in Wiki)	Economist, T. (2017). The world's most valuable resource is no longer oil, but data. <i>The Economist: New York, NY, USA.</i>

Data Literacy: Course contents

	This friday (900-1600): Raw, refined, structured, unstructured data Digression: Data journalism	Next friday (900-1600): Data sources & data quality Digression: AI
Program	<ul style="list-style-type: none">● “Raw” vs. “refined” data● Structured vs. unstructured data● Digression: data journalism● The data processing pipeline● ...	<ul style="list-style-type: none">● Different types of data (sources)● The “data openness pyramid”● Data formats / data types● Data quality and data quality issues● Digression: Data as the fuel of the “AI revolution”● ...
Readings (links in Wiki)	Nick Barrowman (2018): Why data is never raw (The New Atlantis, Summer/Fall edition: 129-135)	Quartz (2018): The Quartz Guide to Bad Data

Data Literacy: Mentorship

- Thu 7.11. (morning 9-11)
- Thu 14.11. (morning 9-12)
- Thu 21.11. (morning 9-12)
- Tue 26.11. (morning 9-12)

Questions?

Data: *what is given*

Capta: *what is taken*

P. Checkland and S. Holwell (1998). *Information, Systems, and Information Systems: Making Sense of the Field*. Chichester, West Sussex: John Wiley & Sons

Data = Information?

- Data consist of structured **characters** from a well-defined **character set** (e.g. {A-Z}, {0|1}, {🚫|😡})
- Data are always tied to a **medium** (“Datenträger”), e.g. a hard disk, but also a piece of paper.

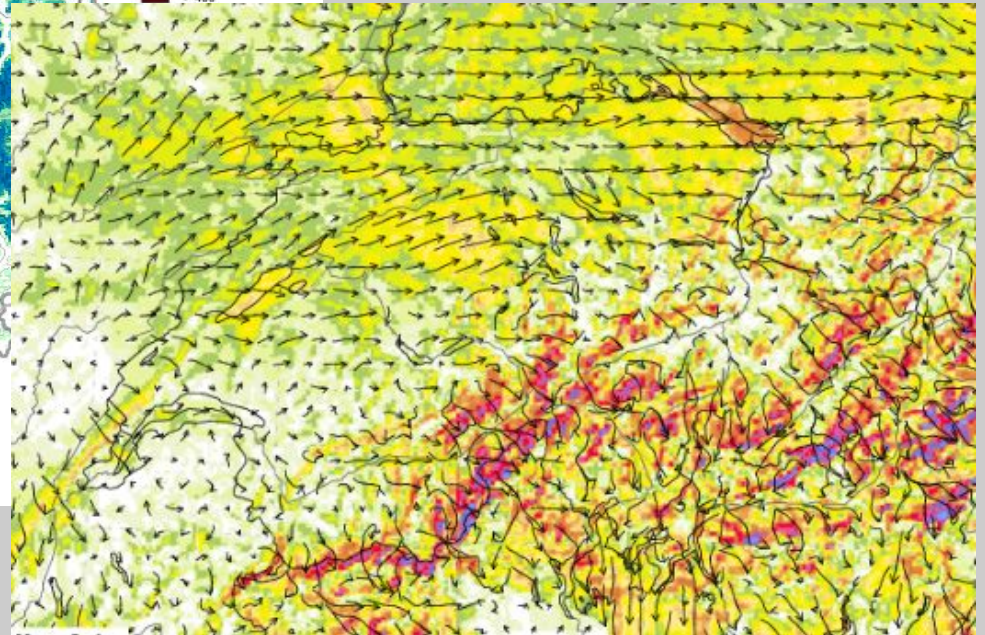
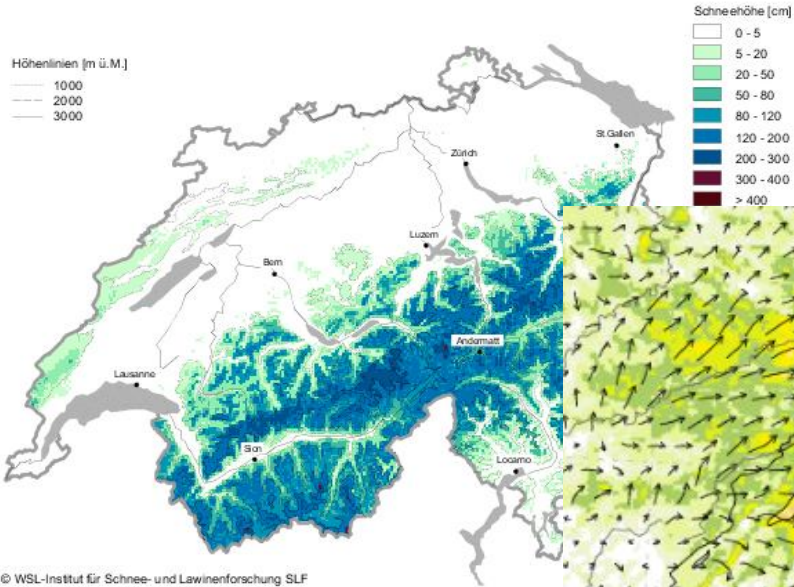
- Once data are used within any **context** (e.g. read by someone, used in a calculation, etc.), we speak of **information**
- Once information is **linked** and **intellectually embedded**, we speak of **knowledge**.

**(Characters) > Data >
Information > Knowledge >
(Wisdom)**

 **BUT: THERE EXIST NO
PRECISE DEFINITIONS OF
THESE CONCEPTS!!!** 

Schneehöhe

aktualisiert am 10.01.2013, 09:12



Lawinengefahr

Sonntag, 17. Januar 2016

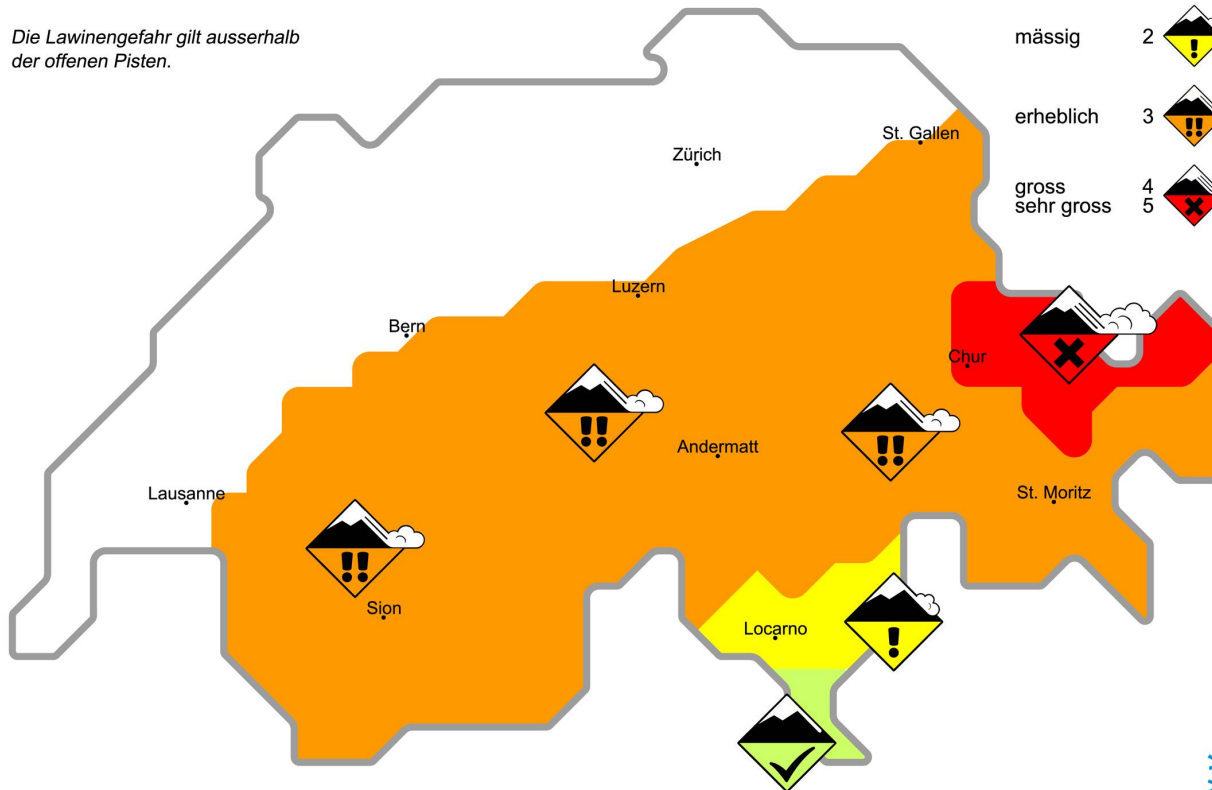
Höhenlinie
10
20
30

Die Lawinengefahr gilt ausserhalb
der offenen Pisten.

- | | | |
|---------------------|--------|--|
| gering | 1 | |
| mässig | 2 | |
| erheblich | 3 | |
| gross
sehr gross | 4
5 | |

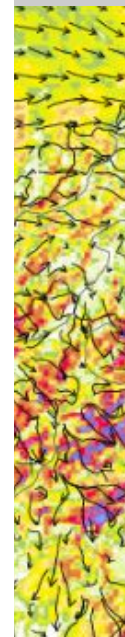


© WSL-Inst



© WSL-Institut für Schnee- und Lawinenforschung SLF

Lawinenbulletin: www.slf.ch



Lawinengefahr

Sonntag, 17. Januar 2016

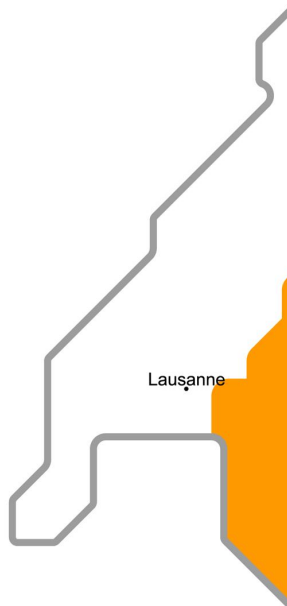
Höhenlinie

- 10
- 20
- 30

Die Lawinengefahr gilt ausserhalb der offenen Pisten.



© WSL-Inst



© WSL-Institut für Schnee- und Law



s





- 1
- 2
- 3
- 4
- 5



© WSL-Institut für Schnee- und Law

Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

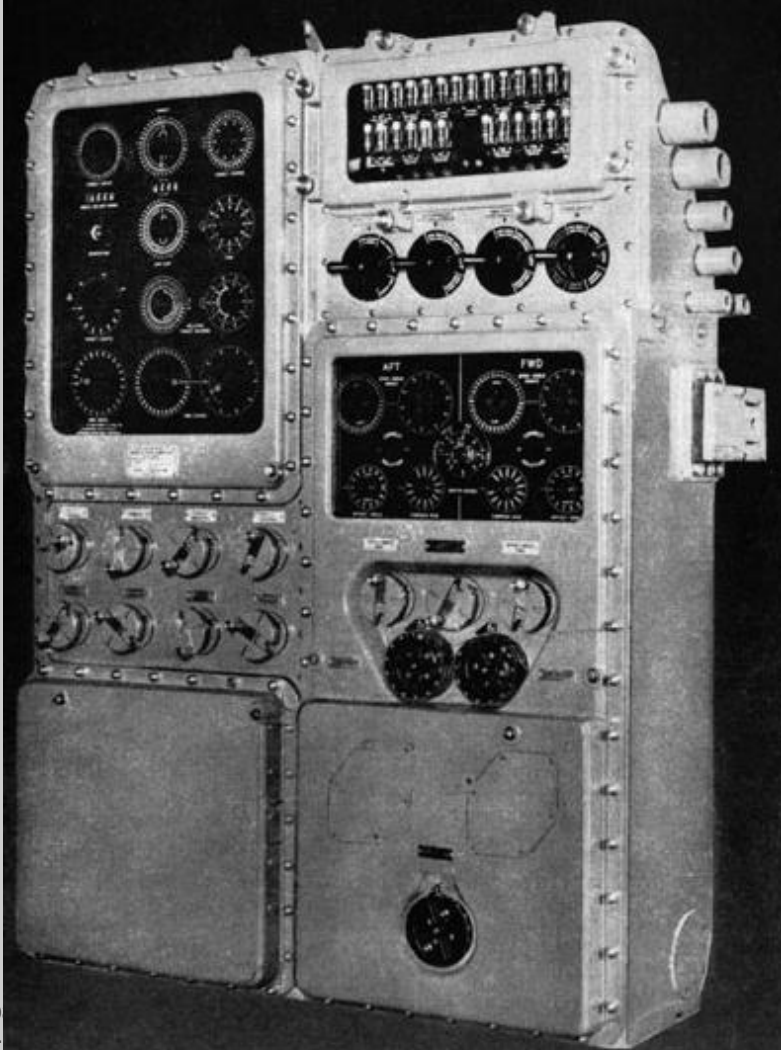
**Do you think this example
makes sense?**

Data storage:

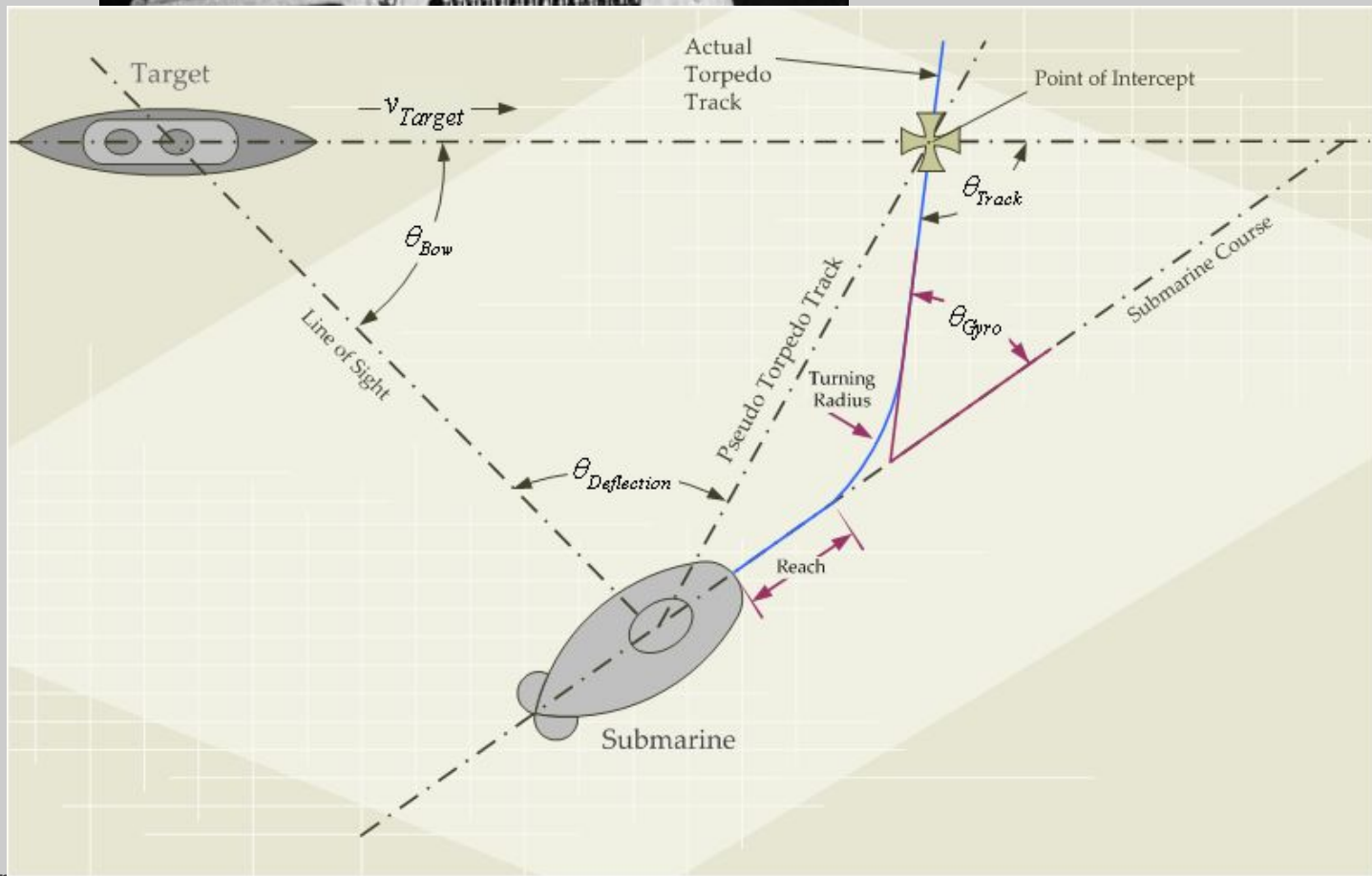
Analog vs. digital

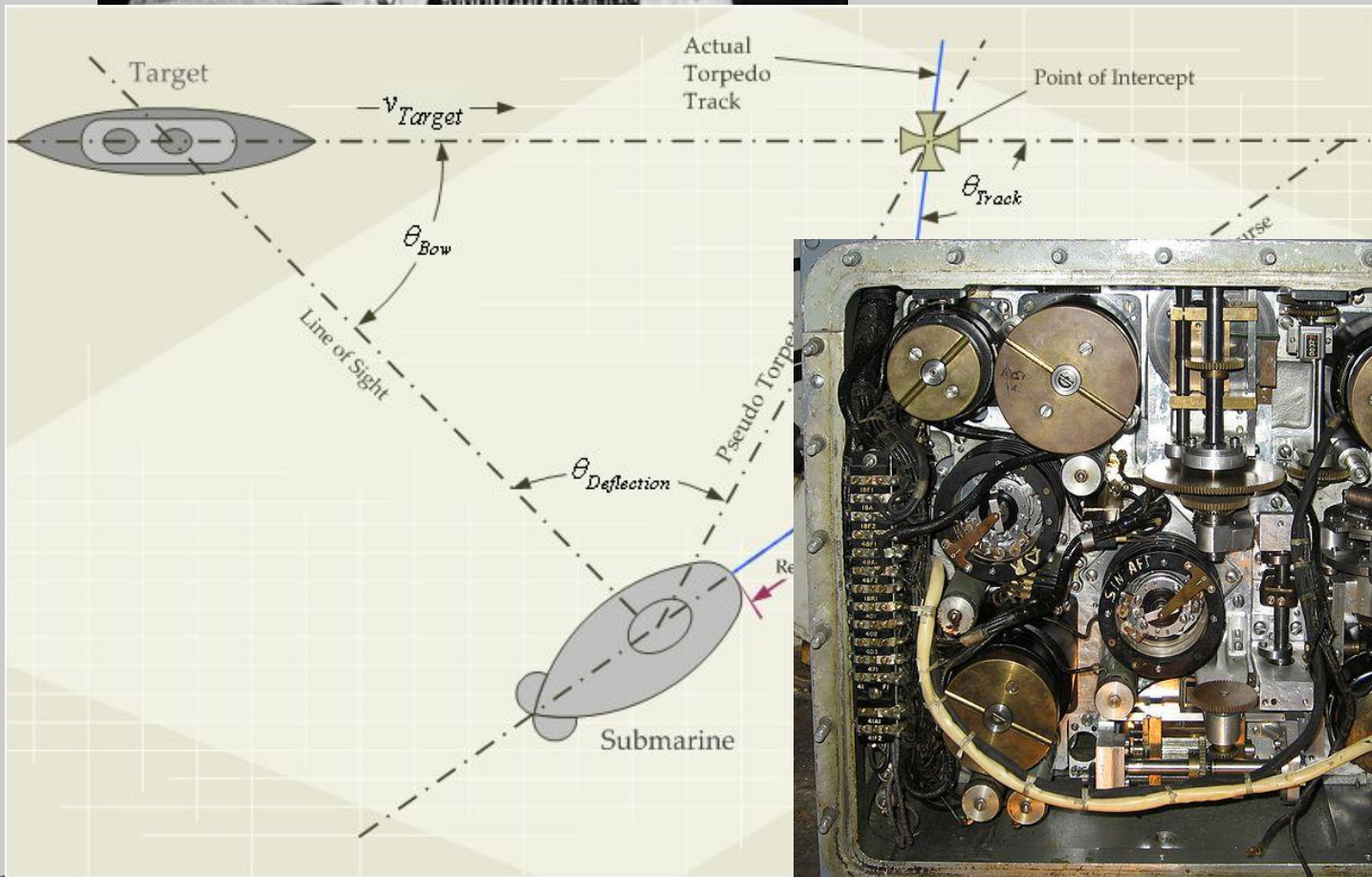


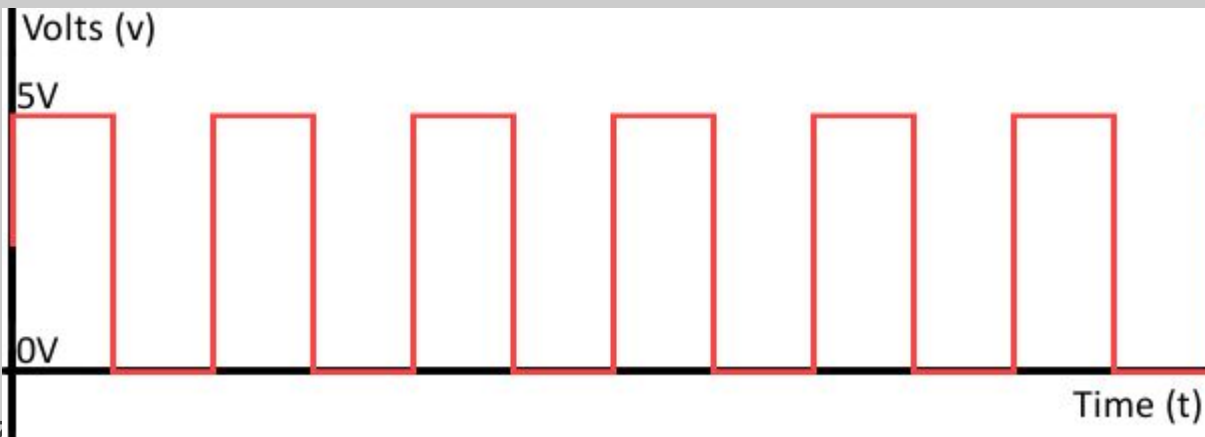
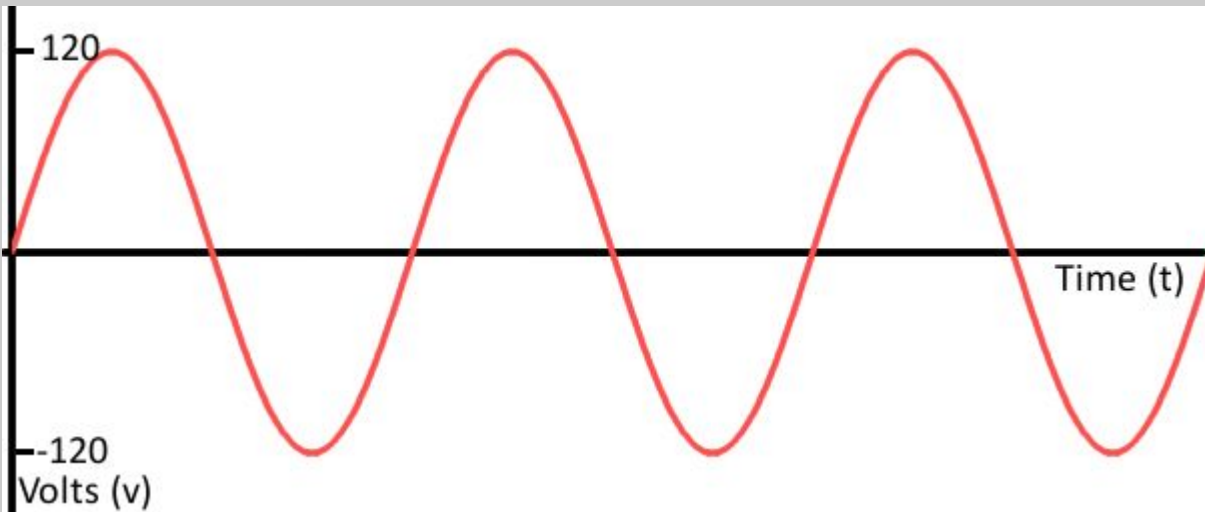
Timo Grossenbacher, ZHd
@grssnbchr // timo@timo@



Timo Grossenb
@grssnbchr // t









binarytranslator.com

One character = 8 Bit = 1 Byte

D = 01000100 = 68

a = 01100001 = 97

...



Monochrome (1-bit)



2-bit Grayscale



4-bit Grayscale



8-bit Grayscale

1	0	1
0	0	1
1	1	1

$$2^1 = 2$$



Monochrome (1-bit)



2-bit Grayscale



4-bit Grayscale



8-bit Grayscale

$$2^2 = 4$$

10	00	01
10	00	11
10	01	11



Monochrome (1-bit)



2-bit Grayscale

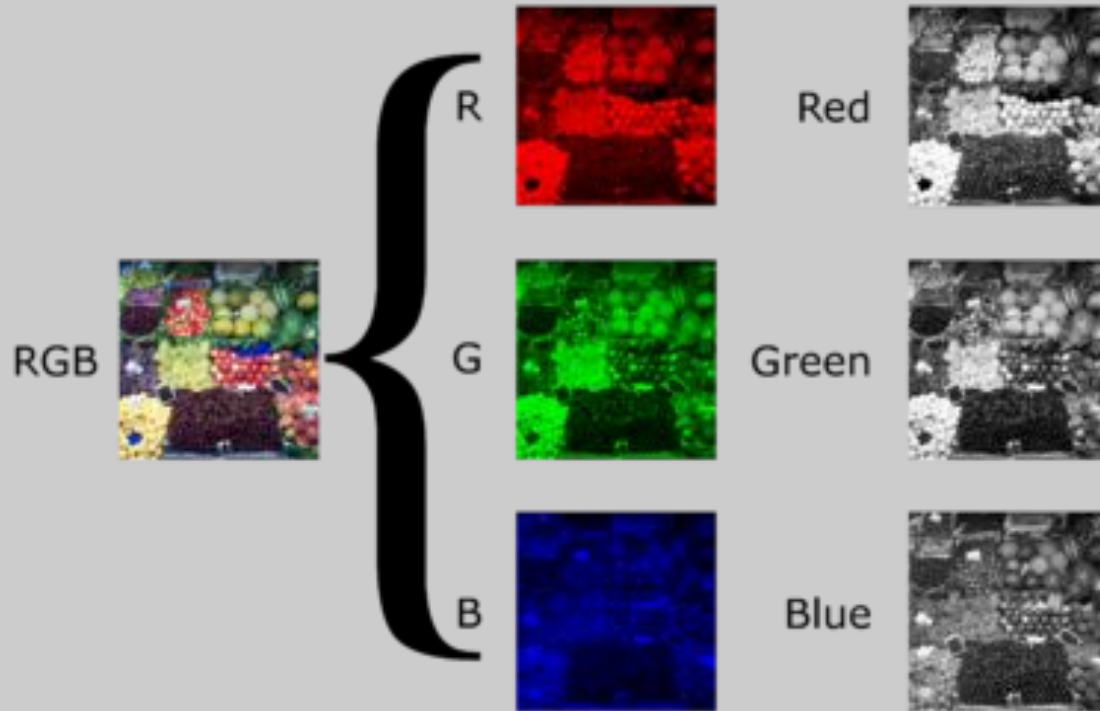
$2^8 = 256$ Shades of Grey!



4-bit Grayscale



8-bit Grayscale

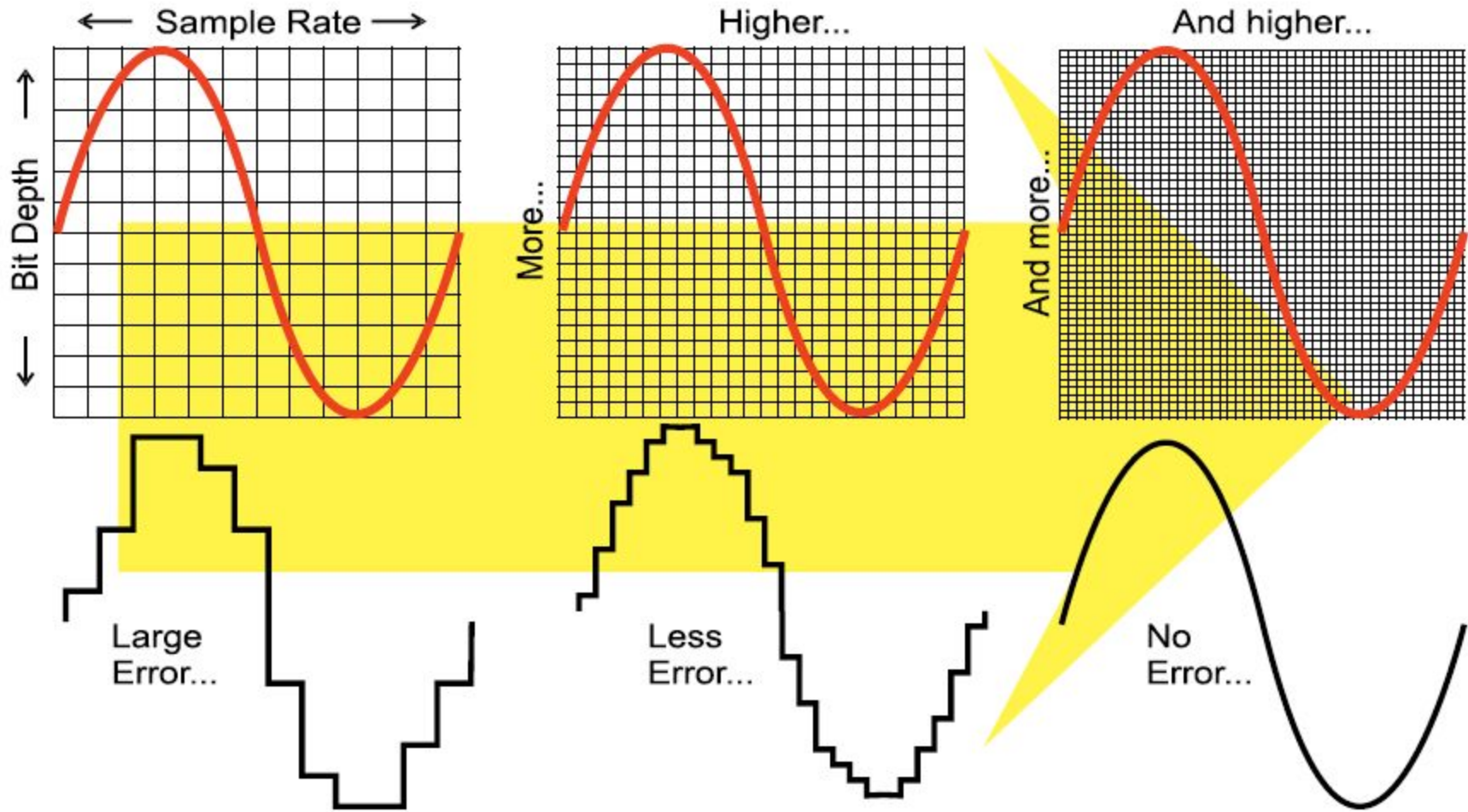




Timo Grossenbacher, ZHdK Fall 2
@grssnbchr // timo@timogrosser

Audio:

Sample rate, bit depth, bit rate



Bit Depth	Sample Rate	Bit Rate	File Size of one stereo minute
16	44,100	1.35 Mbit/sec	10.1 megabytes
16	48,000	1.46 Mbit/sec	11.0 megabytes
24	96,000	4.39 Mbit/sec	33.0 megabytes
mp3 file	128 k/bit rate	0.13 Mbit/Sec	0.94 megabytes

192KBS

Digital data is all about: *discretization / quantization*

Summary of today

- Data vs. capta: Think about it!
- Data vs. information vs. knowledge
- Analog vs. digital data: We usually mean the latter when we talk of data (in this course)

Raw vs. refined data

Nick Barrowman (2018): Why data is
never raw (The New Atlantis,
Summer/Fall edition: 129-135)

Illustrative example



Key point 0

“‘Raw data’ is used as ‘ground truth’, as objective, non-disputable facts”

Key point 1

“Collection is also the result of human decisions”

Key point 2

“Collection means leaving out,
narrowing down and places
limits on inference”

Key point 3

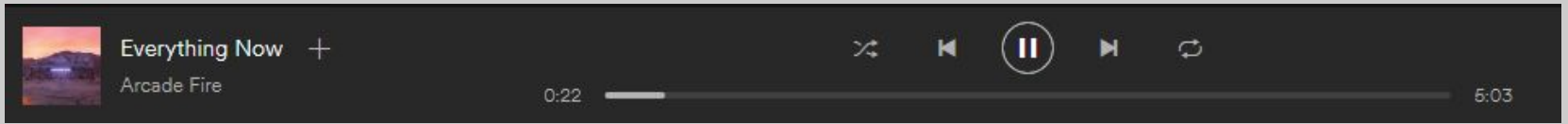
“DIKW turned upside down,
nowadays”

Key point 4

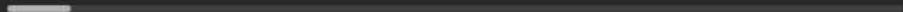
“Consequence: the importance for scrutiny regarding underlying values and assumptions only grows”

“Data is always the product of cognitive, cultural, and institutional processes that determine what to collect and how to collect it.”

Structured vs. unstructured data



Everything Now +
Arcade Fire

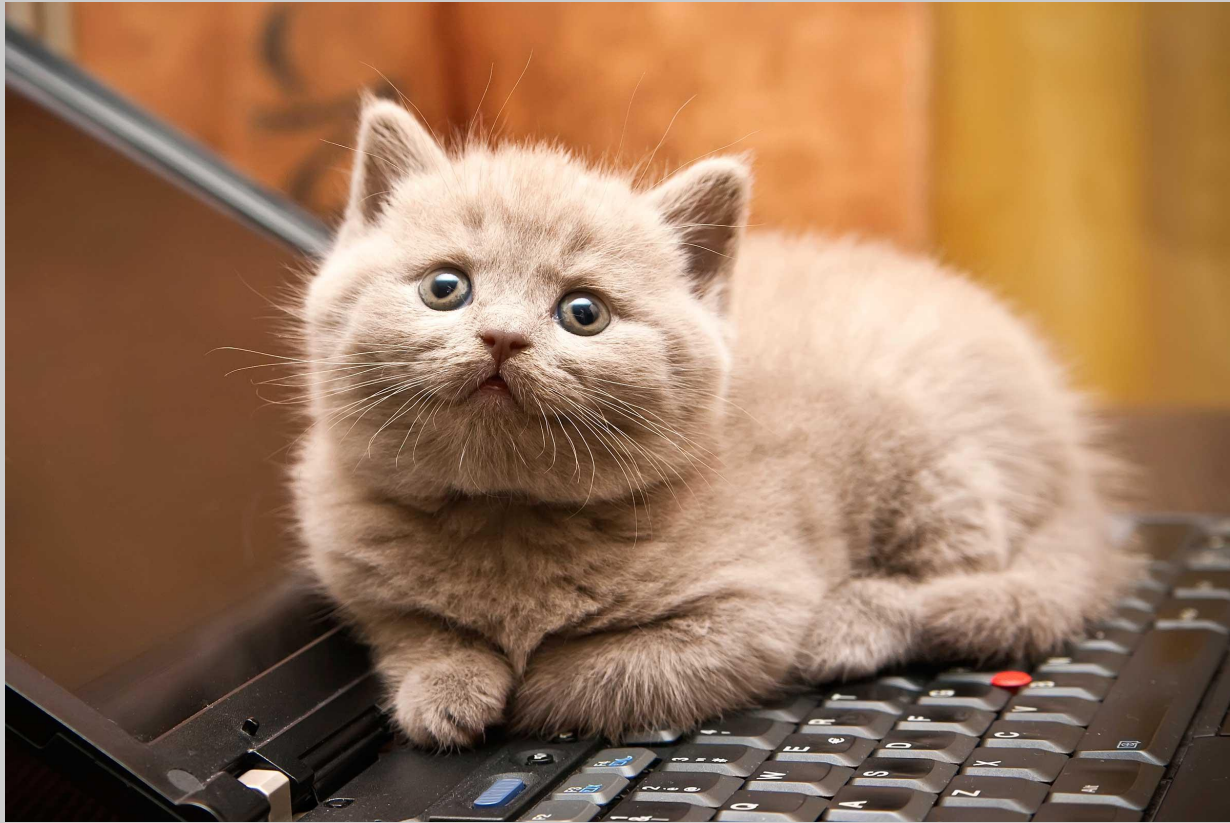
0:22  5:03

⌂ ⏮ ⏸ ⏭ ↺

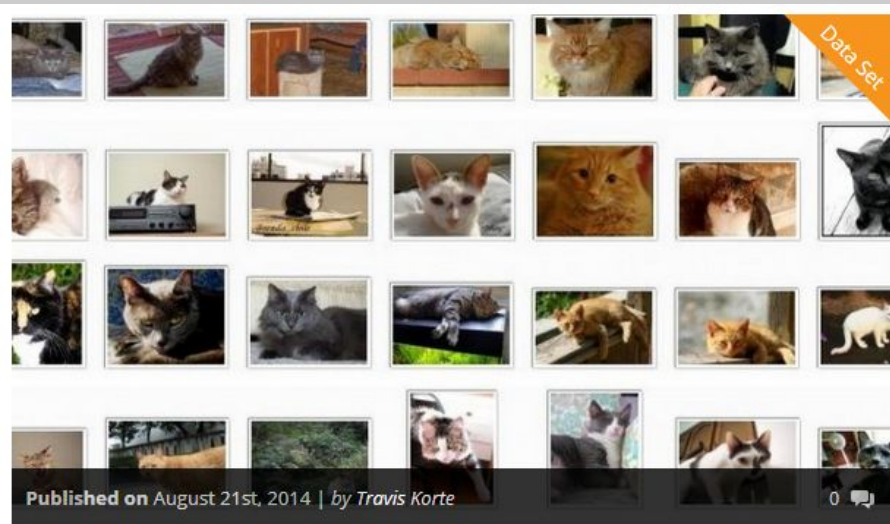
The image shows a dark-themed YouTube music player interface. On the left, there is a square album cover for 'Everything Now' by Arcade Fire. To its right, the song title 'Everything Now' and the artist name 'Arcade Fire' are displayed. A plus sign is to the right of the title. Below the title and artist name is a progress bar. The current time is 0:22 and the total duration is 5:03. To the right of the progress bar are five control icons: a home icon, a previous track icon, a play/pause icon (which is highlighted with a white circle), a next track icon, and a refresh icon.

```
curl -X GET "https://api.spotify.com/v1/audio-features/06AKEBrKUckW0KREUWRnvT" -H  
"Authorization: Bearer {your access token}"
```

```
{  
  "duration_ms" : 255349,  
  "key" : 5,  
  "mode" : 0,  
  "time_signature" : 4,  
  "acousticness" : 0.514,  
  "danceability" : 0.735,  
  "energy" : 0.578,  
  "instrumentalness" : 0.0902,  
  "liveness" : 0.159,  
  "loudness" : -11.840,  
  "speechiness" : 0.0461,  
  "valence" : 0.624,  
  "tempo" : 98.002,  
  "id" : "06AKEBrKUckW0KREUWRnvT",  
  "uri" : "spotify:track:06AKEBrKUckW0KREUWRnvT",  
  "track_href" : "https://api.spotify.com/v1/tracks/  
/06AKEBrKUckW0KREUWRnvT",  
  "analysis_url" : "https://api.spotify.com/v1/audio-analysis/  
/06AKEBrKUckW0KREUWRnvT",  
  "type" : "audio_features"  
}
```

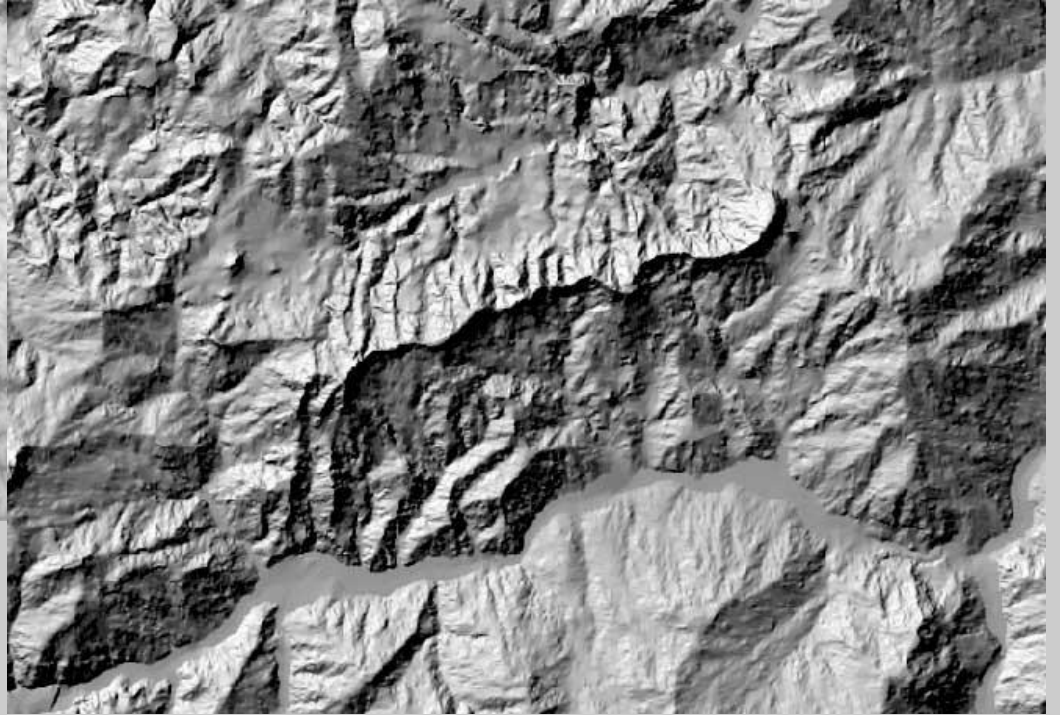
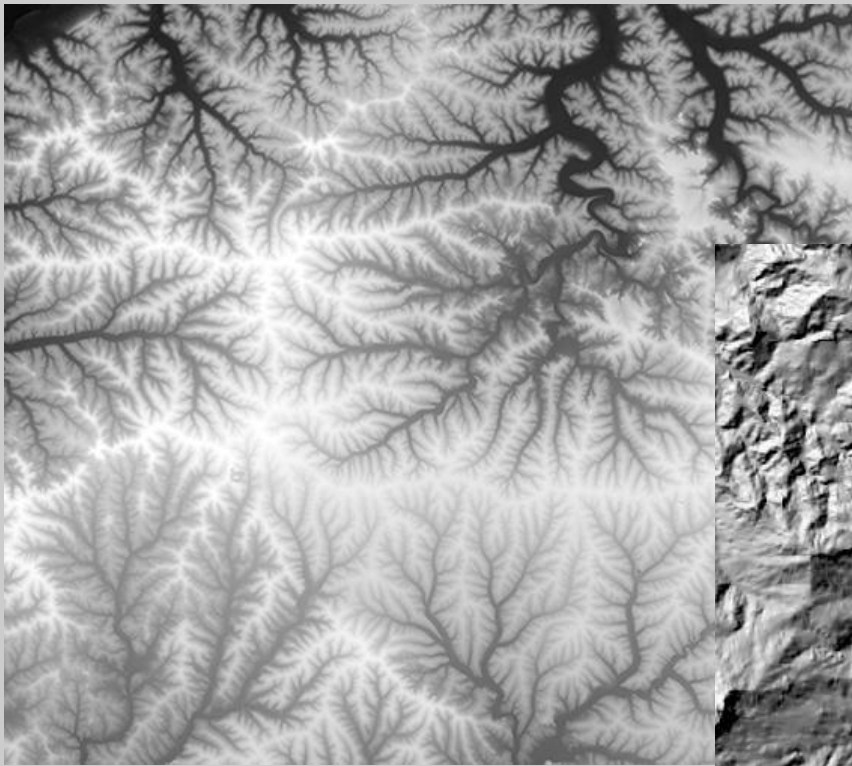


Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

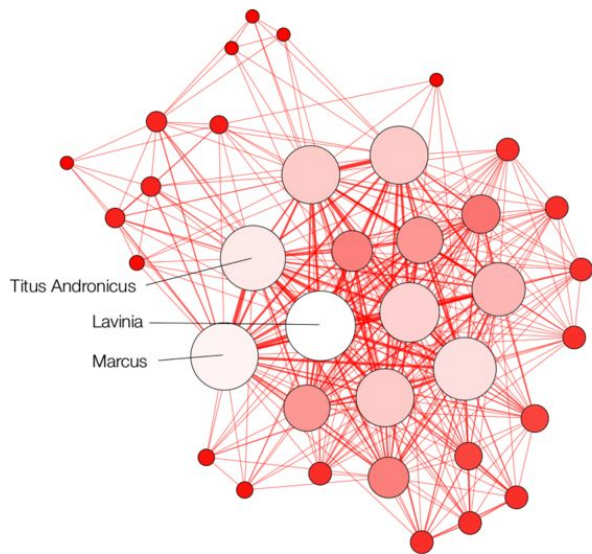


10,000 Cat Pictures (For Science)

Researchers from [Microsoft Research Asia](#) and the [Chinese University of Hong Kong](#) created a database with 10,000 photos containing cat heads to test image recognition algorithms, and now the data set is available freely for research purposes. The photos, which the researchers downloaded mainly from Flickr, are paired with data files that specify the location of each cat's eyes, mouth, and ears. Using these features with their algorithmic method, the researchers were able to distinguish a cat photo from a non-cat photo with better accuracy than previous methods. The creators of the collection, which is affectionately known as the CAT Database, hope their data might be useful to other machine learning researchers working in facial recognition, image comprehension, and other computer vision topics.







TITUS ANDRONICUS

Number of characters **36** | **50%** Network density



ROMEO AND JULIET

Number of characters **41** | **37%** Network density

martingrandjean.ch/network-visualization-shakespeare/

Structured

- Pre-defined data model / annotated
- Suited for a certain processing task
 - Often tables / databases

Unstructured

- Not organized in a pre-defined manner
- Difficult to understand using “traditional” software
 - Not suited for the processing task at hand
 - Often text-heavy / images / music

**Whether data are “structured”
or “unstructured” depends on
the task at hand, imho.**

**There are many different types
of data formats.**

Proprietary vs. text

Proprietary / binary

- Stored according to a particular scheme
 - Designed to be secret
- Can only be read / written with specific software
- Garbled when opened with other software
 - Might be more efficient

Proprietary (some ex.)

- PSD
 - Spotify Tracks
 - WMA
 - PDF (formerly)
- DOC/PPT/XLS (formerly)

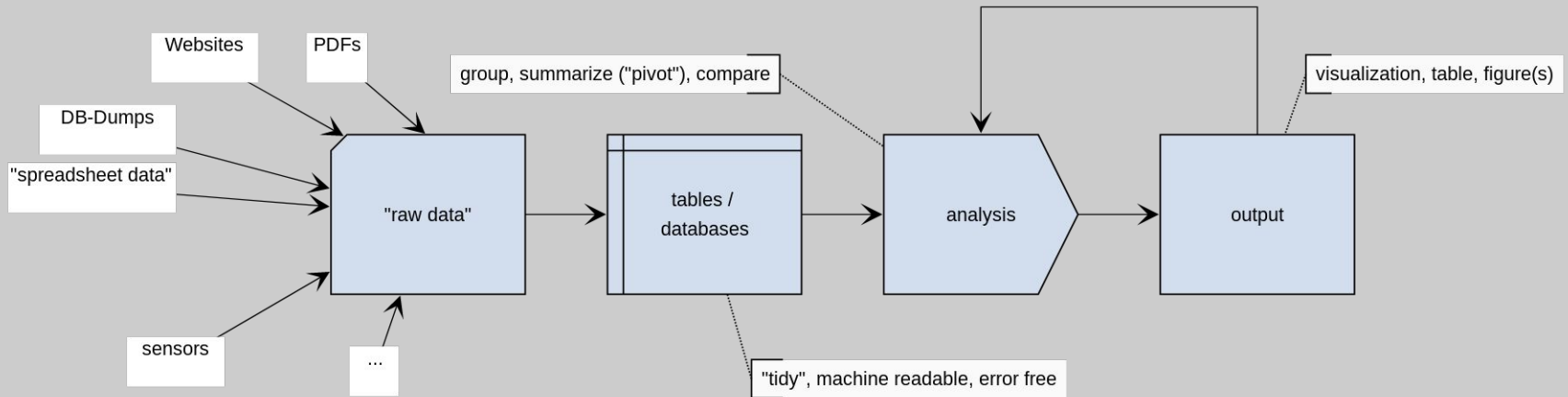
Text (some ex.)

- JSON
- HTML/XML
 - CSV
 - ...

The most common formats:

- XLSX / CSV
 - JSON
 - HTML
 - PDF
- Geodata: SHP, GeoJSON, TIFF

The data processing pipeline



Derived from

<https://www.davidbauer.ch/2013/05/18/datenjournalismus-workflow-simon-rogers-daten-geschichten/>

Timo Grossenbacher, ZHdK Fall 2019

@grssnbchr // timo@timogrossenbacher.ch

The data processing pipeline

1. Collect / create data
2. Pre-process & clean data → table
3. Analyze table data
4. Visualize / transform table data
5. (Make decisions)

Derived from

<https://www.davidbauer.ch/2013/05/18/datenjournalismus-workflow-simon-rogers-daten-geschichten/>

Timo Grossenbacher, ZHdK Fall 2019

@grssnbchr // timo@timogrossenbacher.ch

Break



Exercise:

Find data for a given task

Exercise:

In groups of 4 or 5, try to **find data** that *can / could answer the question / task* you receive. After this, you will present and *live demo* us where you found the data (*Bonus*: how would you process the data to get your answer?)

Also: Write down the website, the data format and the tool(s) you used / would use to look at the data – and get your answer!

Key take aways:

- Sometimes, data is accessible via API
- The preferred data format of APIs is JSON
 - JSON can be converted into CSV
- The preferred way of talking to an API is with code

Key take aways:

- Sometimes there is more than one way to reach a goal
- Excel / *Office have some good filters, get to know them
 - Reduce data volume early on through filters

Key take aways:

- Many interesting data are buried in PDFs
- Use proprietary software or Tabula to extract the data

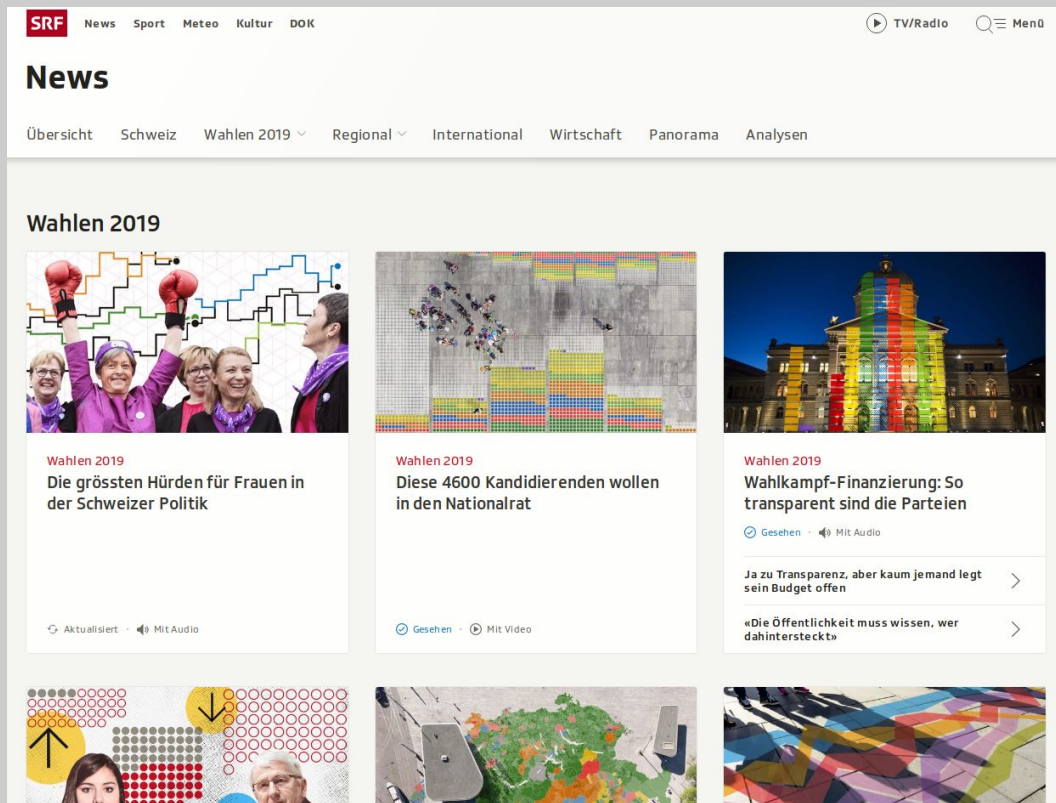
Key take aways:

- Excel can quickly give insights through Pivot data analysis
- Open data portals have funnier data than you might think

My job: Data Journalist

My job: Data Journalist (your future job?)

Who we are and what we do: [srf.ch/data](https://www.srf.ch/data)



The screenshot shows the SRF News website interface. At the top, there is a navigation bar with the SRF logo and links for News, Sport, Meteo, Kultur, and DOK. A TV/Radio icon and a search menu icon are also present. Below the navigation bar, the word 'News' is prominently displayed. A secondary navigation bar includes 'Übersicht', 'Schweiz', 'Wahlen 2019', 'Regional', 'International', 'Wirtschaft', 'Panorama', and 'Analysen'. The main content area is titled 'Wahlen 2019' and features a grid of news articles. Each article includes a thumbnail image, a title, and a 'Gesehen' (Viewed) button with a play icon. The articles are:

- Article 1:** 'Wahlen 2019: Die grössten Hürden für Frauen in der Schweizer Politik'. Thumbnail shows four women in purple jackets. Text: 'Aktualisiert · Mit Audio'.
- Article 2:** 'Wahlen 2019: Diese 4600 Kandidierenden wollen in den Nationalrat'. Thumbnail shows a large grid of colorful dots. Text: 'Gesehen · Mit Video'.
- Article 3:** 'Wahlen 2019: Wahlkampf-Finanzierung: So transparent sind die Parteien'. Thumbnail shows a building facade with colorful lights. Text: 'Gesehen · Mit Audio'. Below the title, there are two sub-headers: 'Ja zu Transparenz, aber kaum jemand legt sein Budget offen' and '«Die Öffentlichkeit muss wissen, wer dahintersteckt»', both with right-pointing arrows.

At the bottom of the grid, there are three more thumbnails: one with a woman's face and a grid of dots, one with a map of Switzerland, and one with a colorful geometric pattern.



We are not alone...

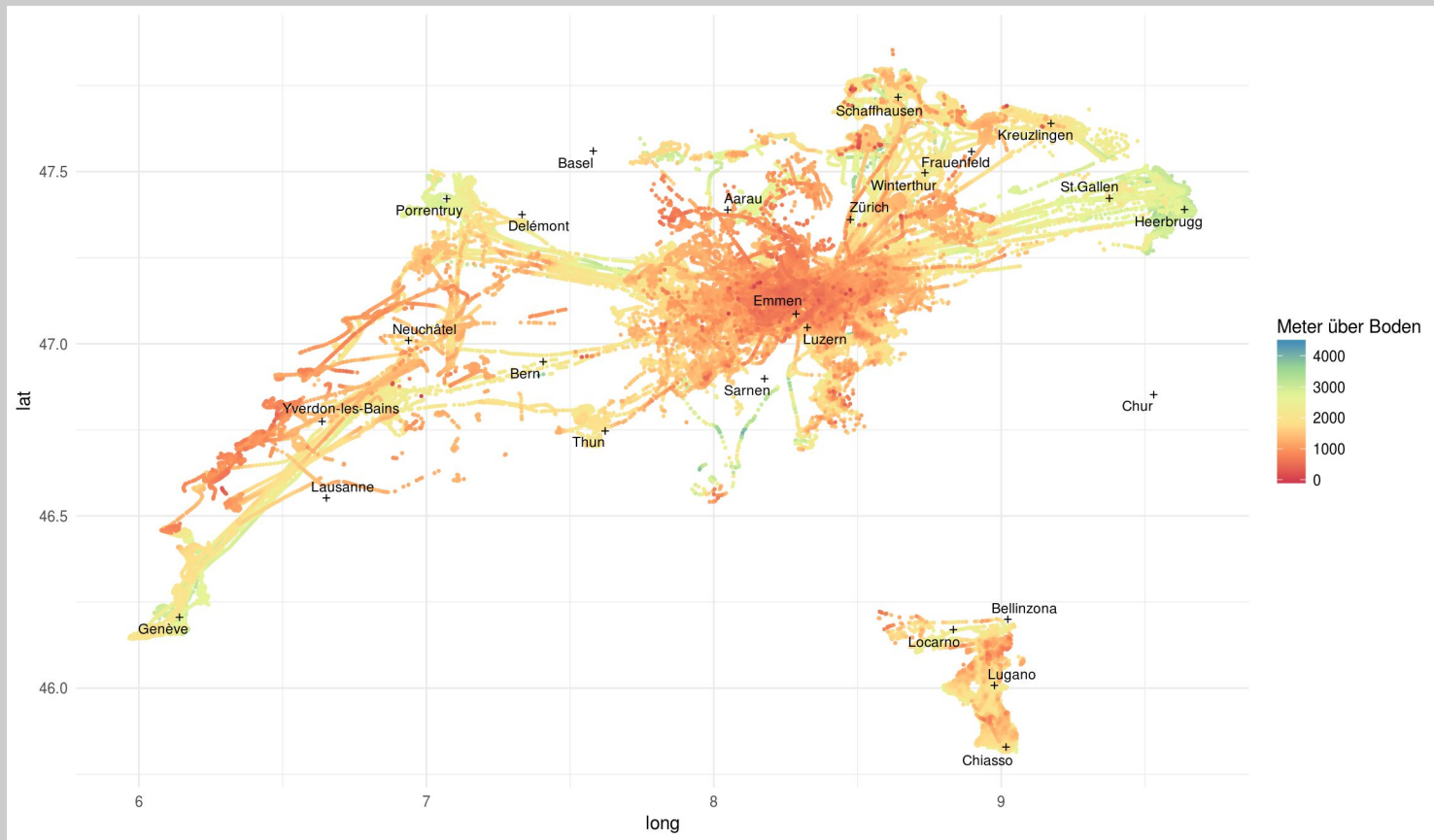
NZZ Visuals (previously NZZ Storytelling)

TA Interaktiv / Datenblog / Tamedia

Data Le Temps / RTS, etc.

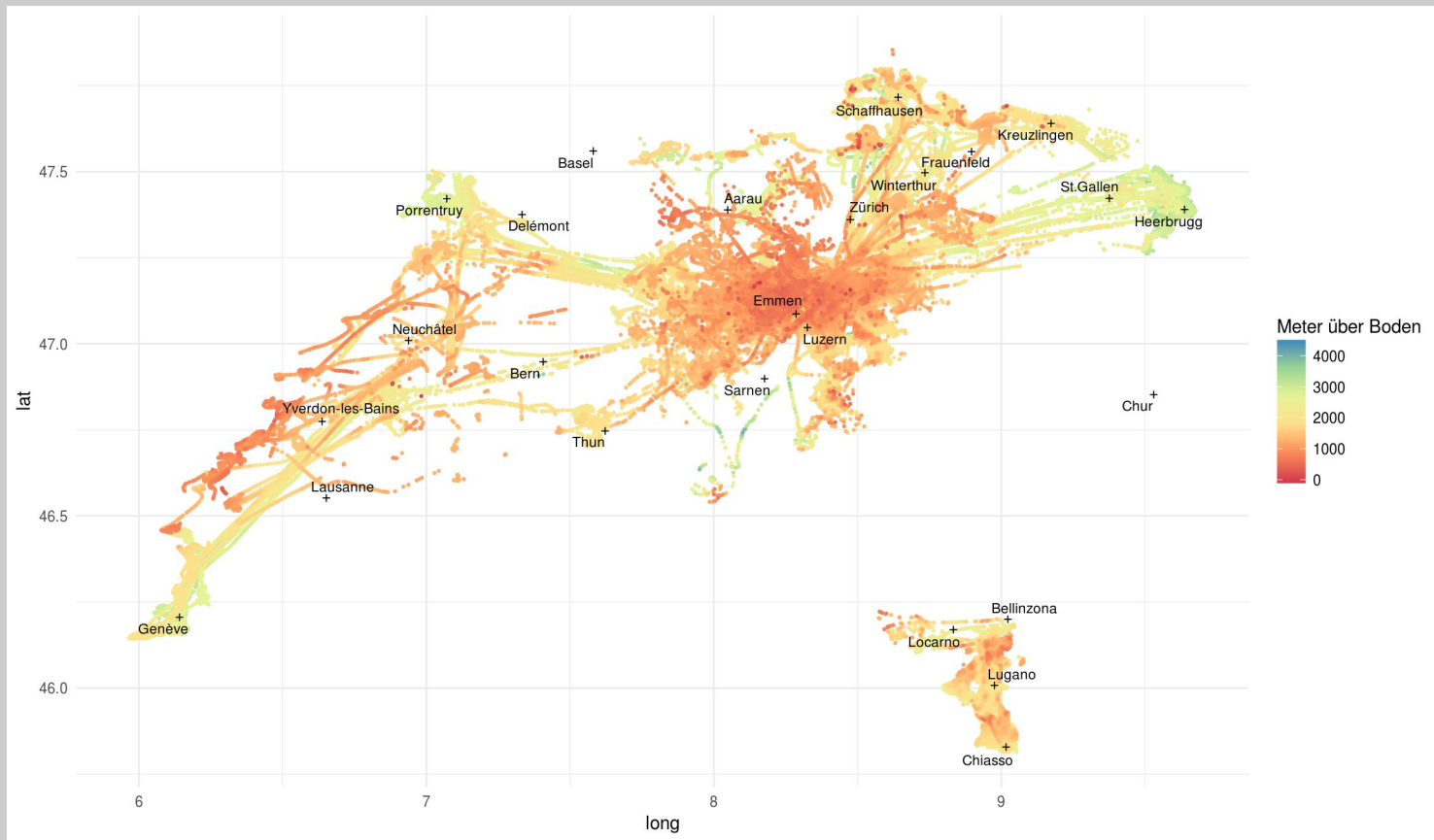
swissinfo.ch

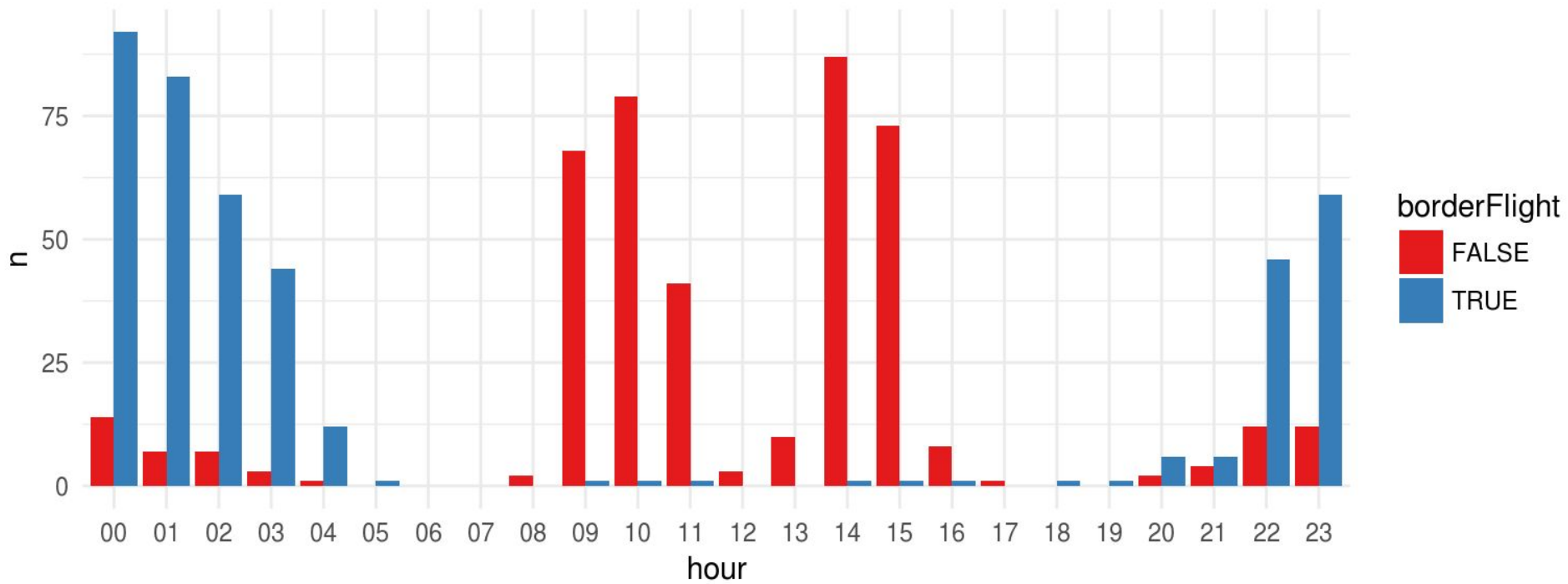
Republik





Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch





Grenzwächter der Läfte

Die Schweizer Armee setzt Drohnen zur Grenzwahe ein. Offiziell für den Kampf gegen Kriminaltourismus. Doch jetzt zeigen Daten: Im Visier sind auch Flüchtlinge.

Reise starten

Nächtlicher Einsatz an der Grenze

Fahndungserfolg

127 Grenzflüge seit April 2015

Klare Ballungspunkte

Offene Balkanroute

Geschlossene Balkanroute



Another example: Identifying Fake Followers



goo.gl/muEUT6

srfdata.github.io