«Data is the new oil»

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 SRF DATA
- **Lecturer** @UZH / @ZHdK / @SfGZ / @DataCamp
- Twitter: **@grssnbchr**
- Website: **timogrossenbacher.ch**

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

Now, what about you?

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

Why this course?

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

Data Literacy: Course contents

	This morning (830-1030): Introduction to Data, Information and Knowledge
Program	• Introduction • What exactly *are* data? • Digital vs. analog data
Readings (links in Wiki)	Economist, T. (2017). The world's most valuable resource is no longer oil, but data. *The Economist: New York, NY, USA*.

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

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	• "Raw" vs. "refined" data • Structured vs. unstructured data • Digression: data journalism • The data processing pipeline • ...	• 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

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

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)

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

Questions?

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

Data: *what is given*

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

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

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

Data = Information?

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

- 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.

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

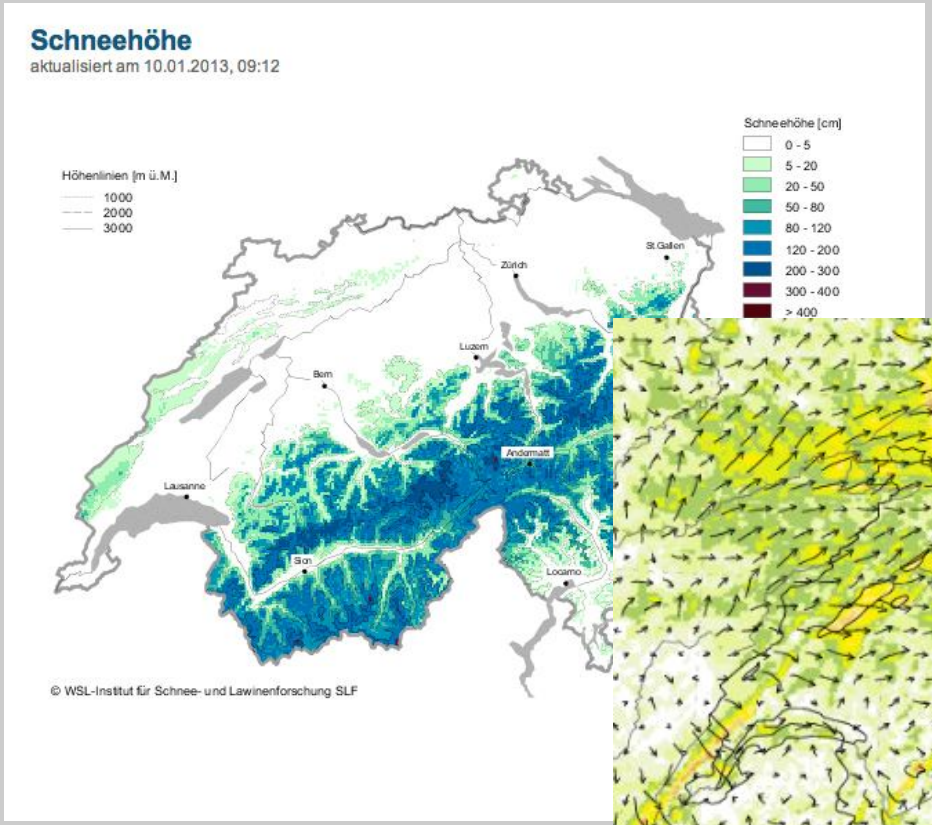
- Once data are used within any **context** (e.g. read by someone, used in a calculation, etc.), we speak of **information**
- Once information is **linked** and **intellectually embedded**, we speak of **knowledge**.

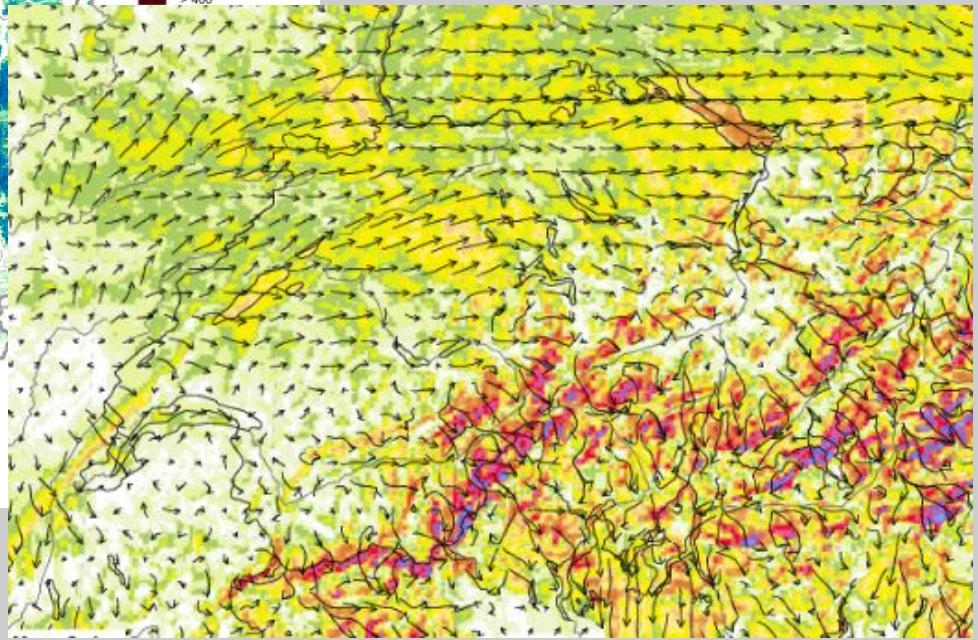
(Characters) > Data > Information > Knowledge > (Wisdom)

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

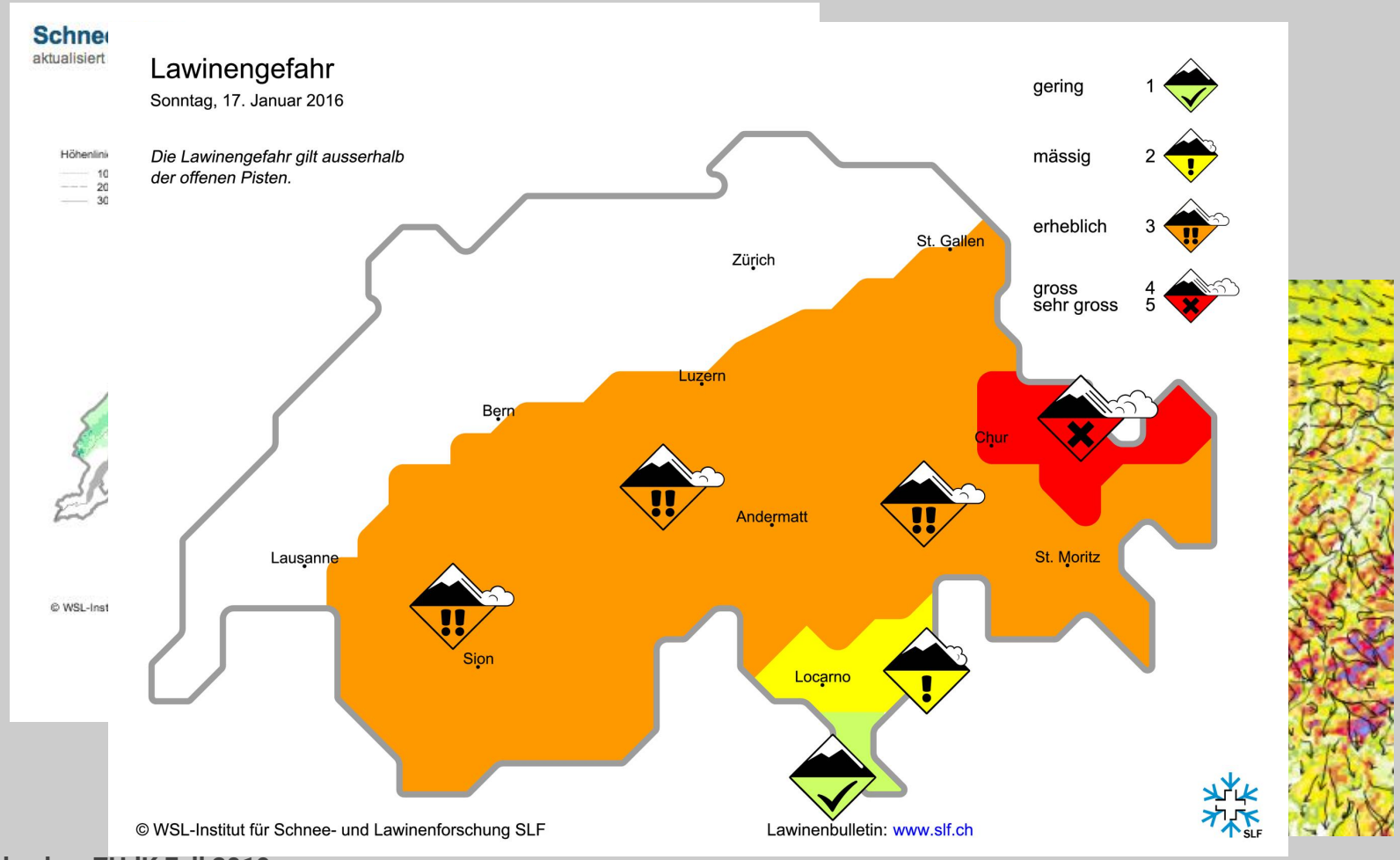
⚠️ BUT: THERE EXIST NO PRECISE DEFINITIONS OF THESE CONCEPTS!!! ⚠️

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

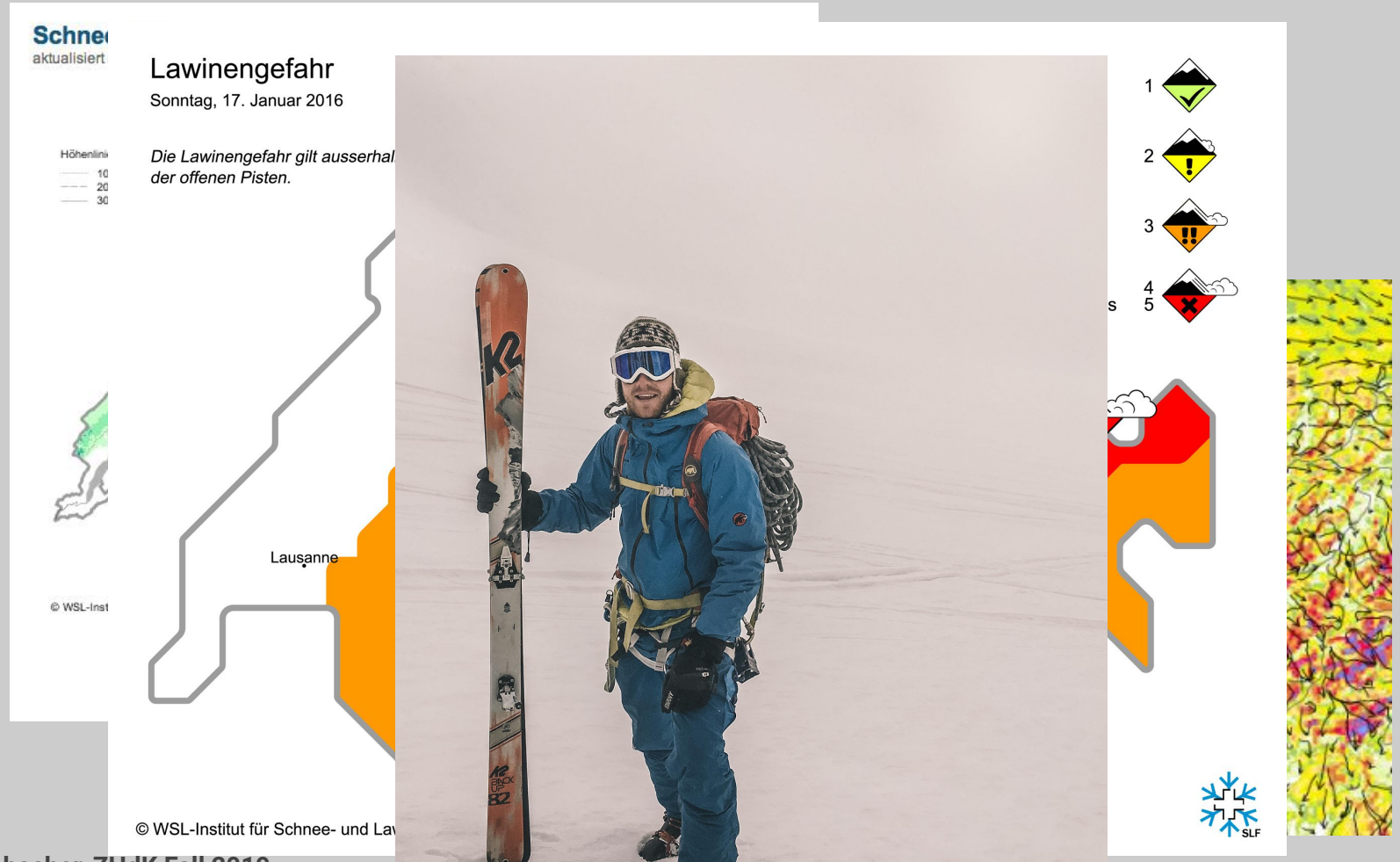




Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch



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



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

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

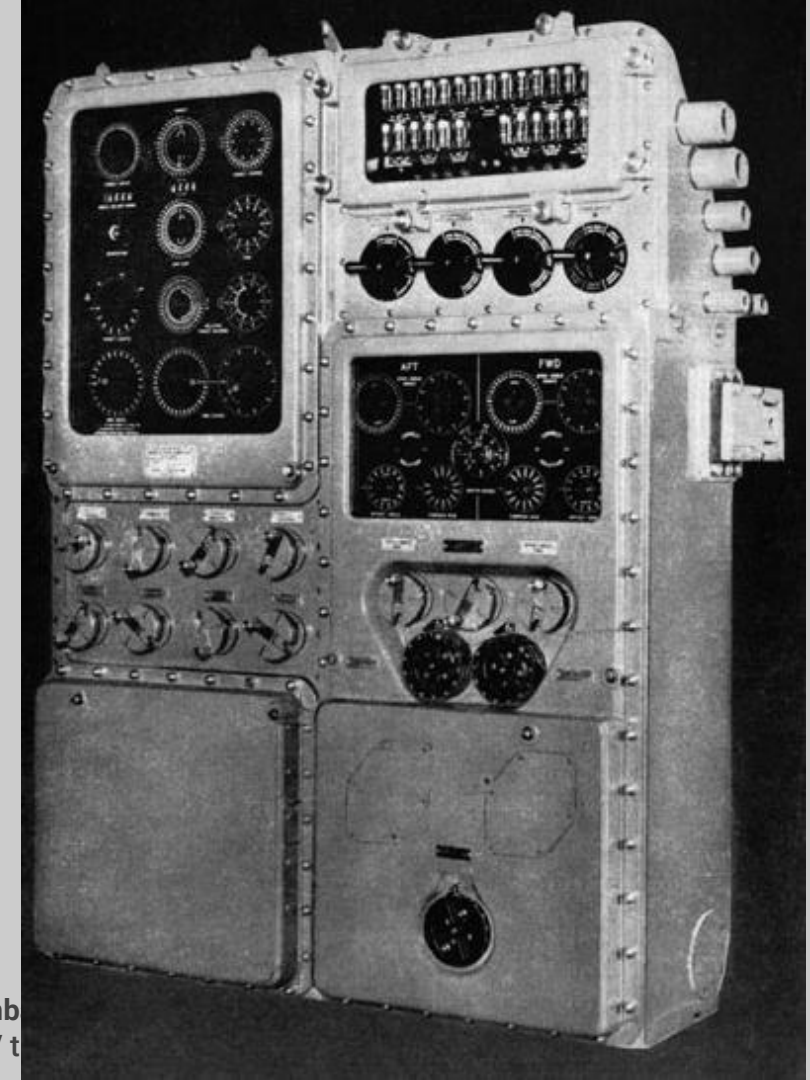
Do you think this example makes sense?

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

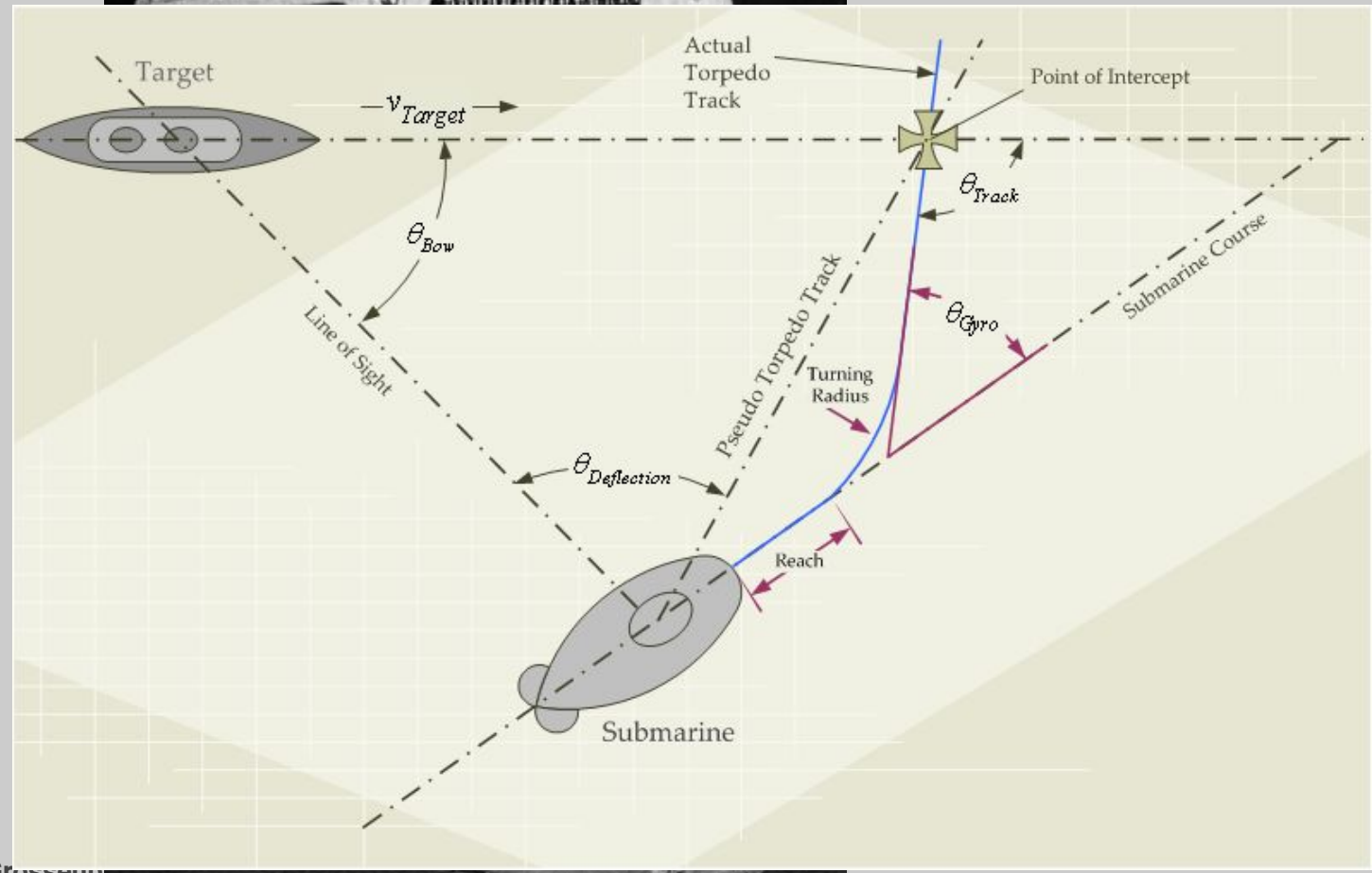
Data storage:
Analog vs. digital

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

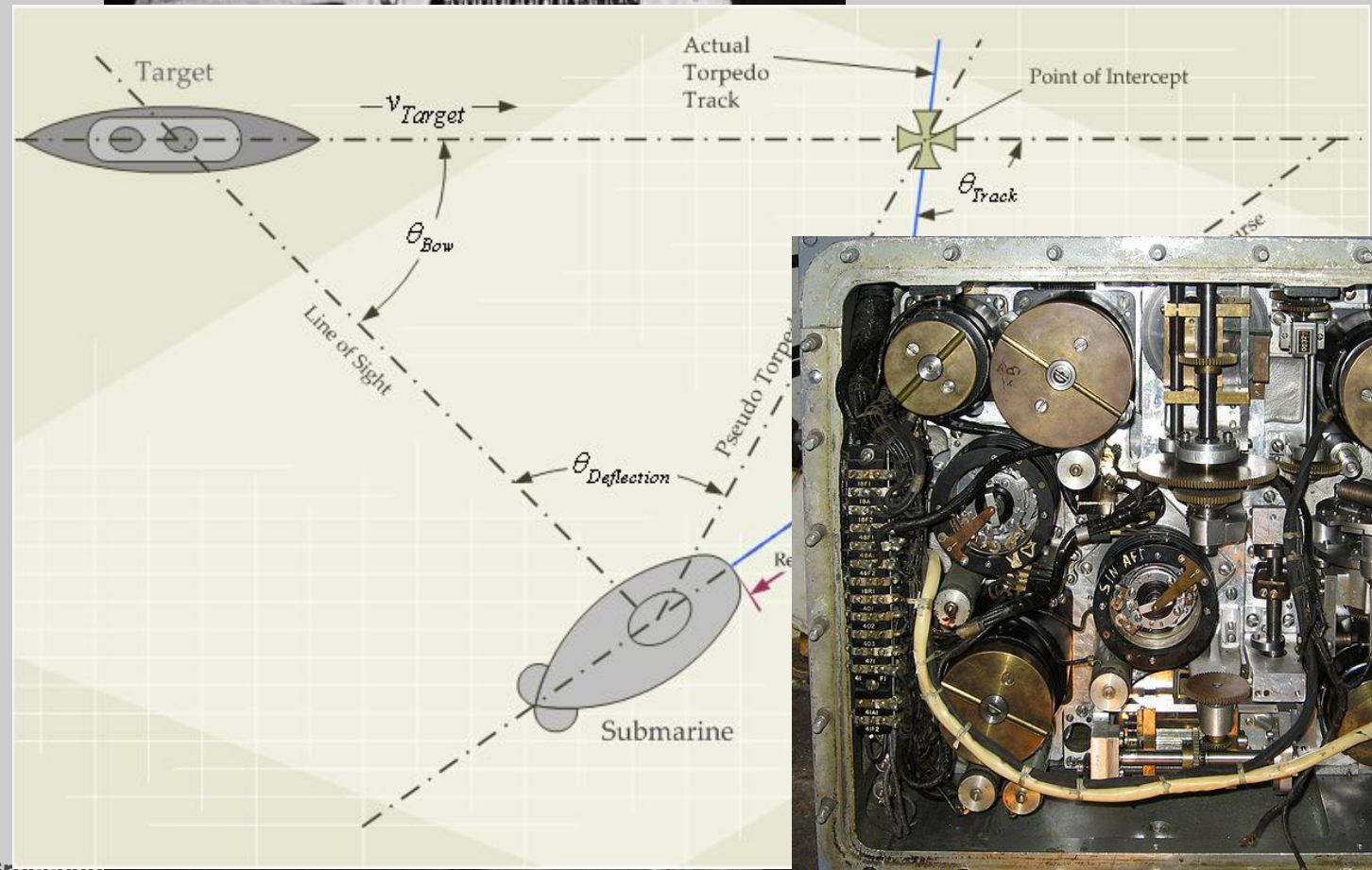
Timo Grossenbacher, ZHd
@grssnbchr // timo@timo



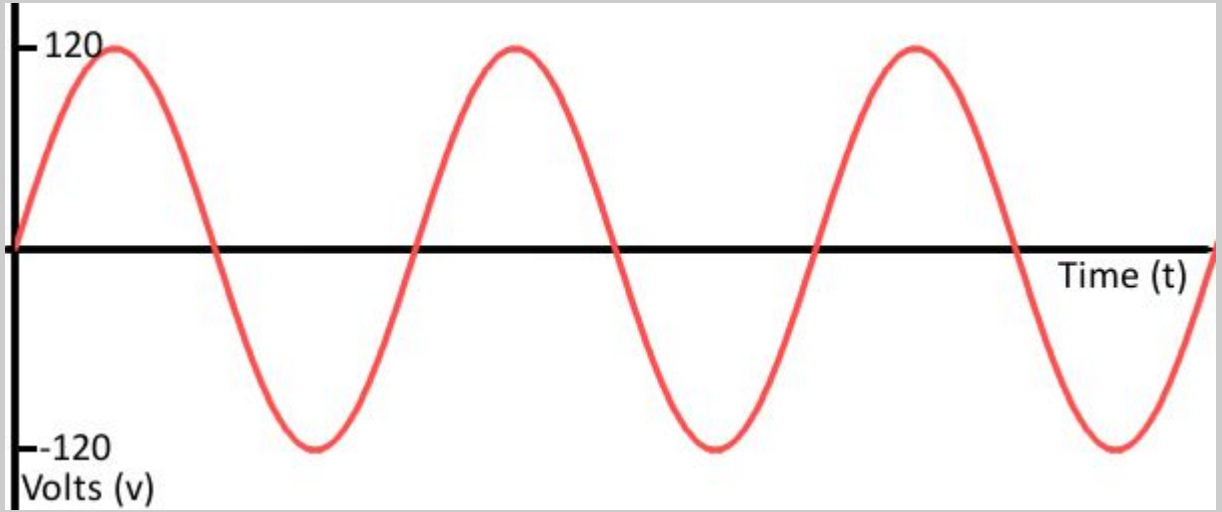
Timo Grossenb
@grssnbchr // t

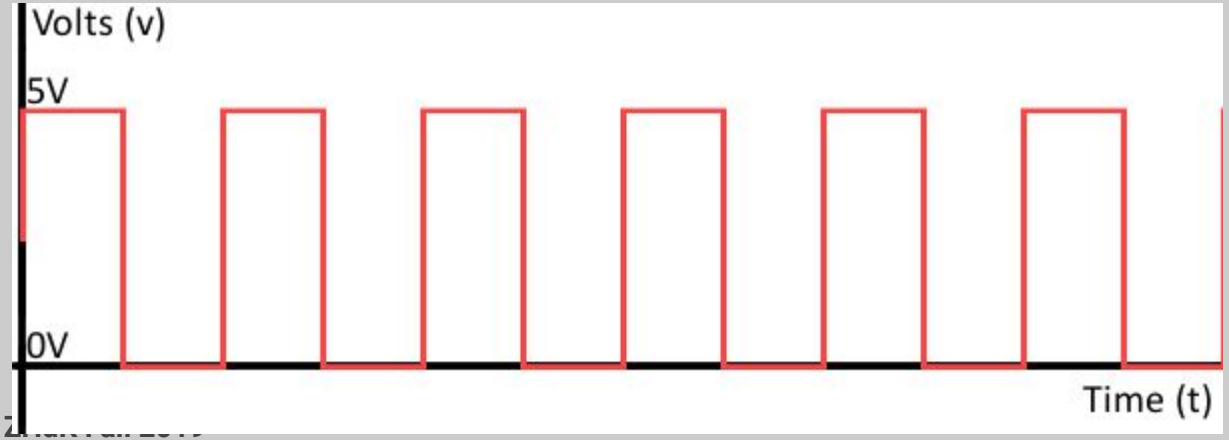


Timo Grossenb
@grssnbchr // t



Timo Grossenb
@grssnbchr // t





Timo Grossenbacher,
@grssnbchr // timo@timogrossenbacher.ch

binarytranslator.com

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

One character = 8 Bit = 1 Byte

D = 01000100 = 68

a = 01100001 = 97

...

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

Monochrome (1-bit)

2-bit Grayscale

4-bit Grayscale

8-bit Grayscale

Timo Grossenbacher, ZH
@grssnbchr // timo@timogrossenbacher.ch

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

Timo Grossenbacher, ZH
@grssnbchr // timo@timogrossenbacher.ch

Monochrome (1-bit)

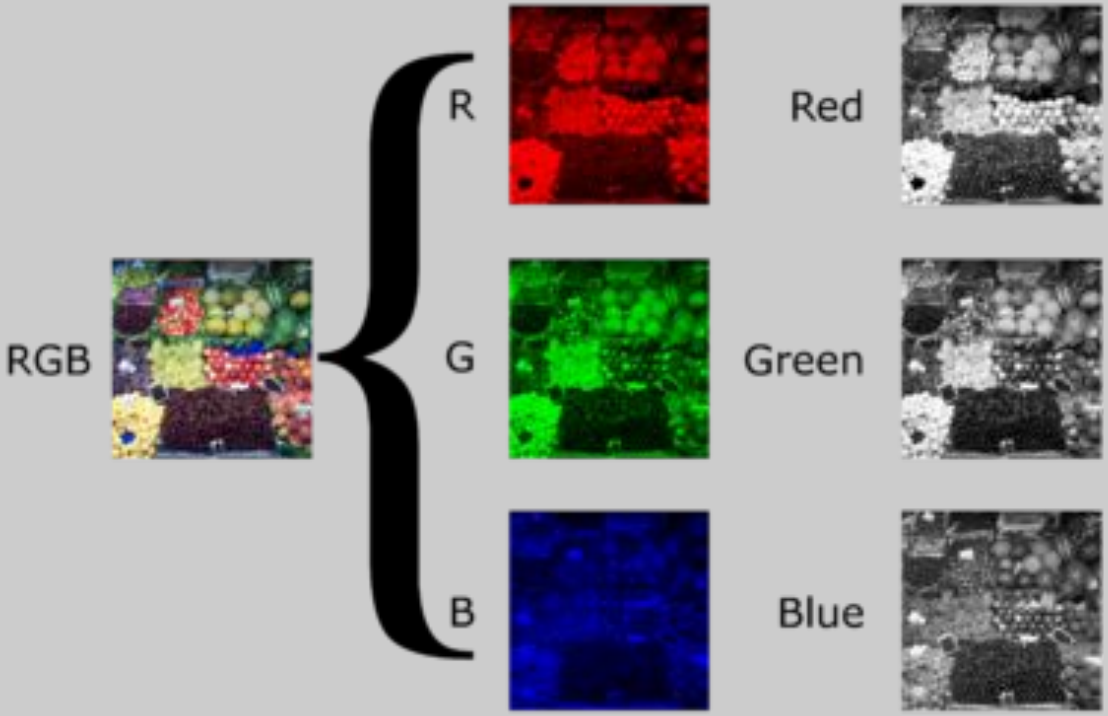
2-bit Grayscale

2^8 = 256 Shades of Grey!

4-bit Grayscale

8-bit Grayscale

Timo Grossenbacher, ZH
@grssnbchr // timo@timogrossenbacher.ch

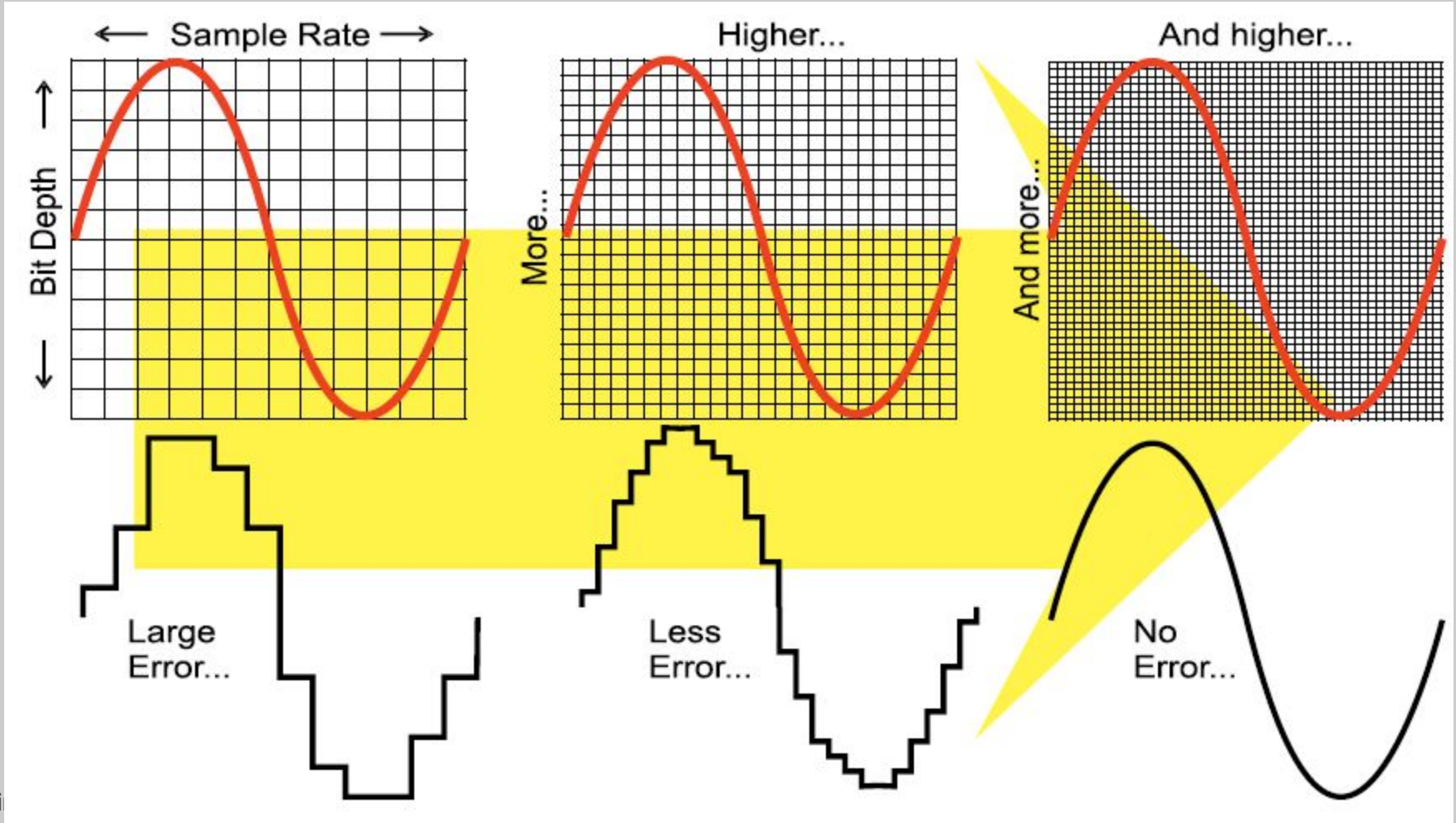


Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

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

Audio:
Sample rate, bit depth, bit rate

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



@grssnbchr // timo@timogrossenbacher.ch

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

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

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

Digital data is all about:
discretization / quantization

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

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)

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

Raw vs. refined data

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

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

Illustrative example

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

Key point 0

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

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

Key point 1

“Collection is also the result of human decisions”

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

Key point 2

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

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

Key point 3

“‘DIKW turned upside down, nowadays”

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

Key point 4

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

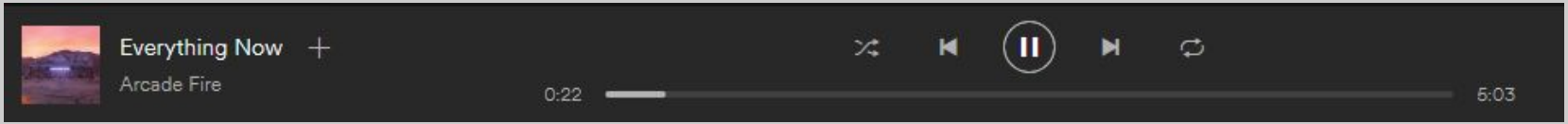
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

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

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

Structured vs. unstructured data

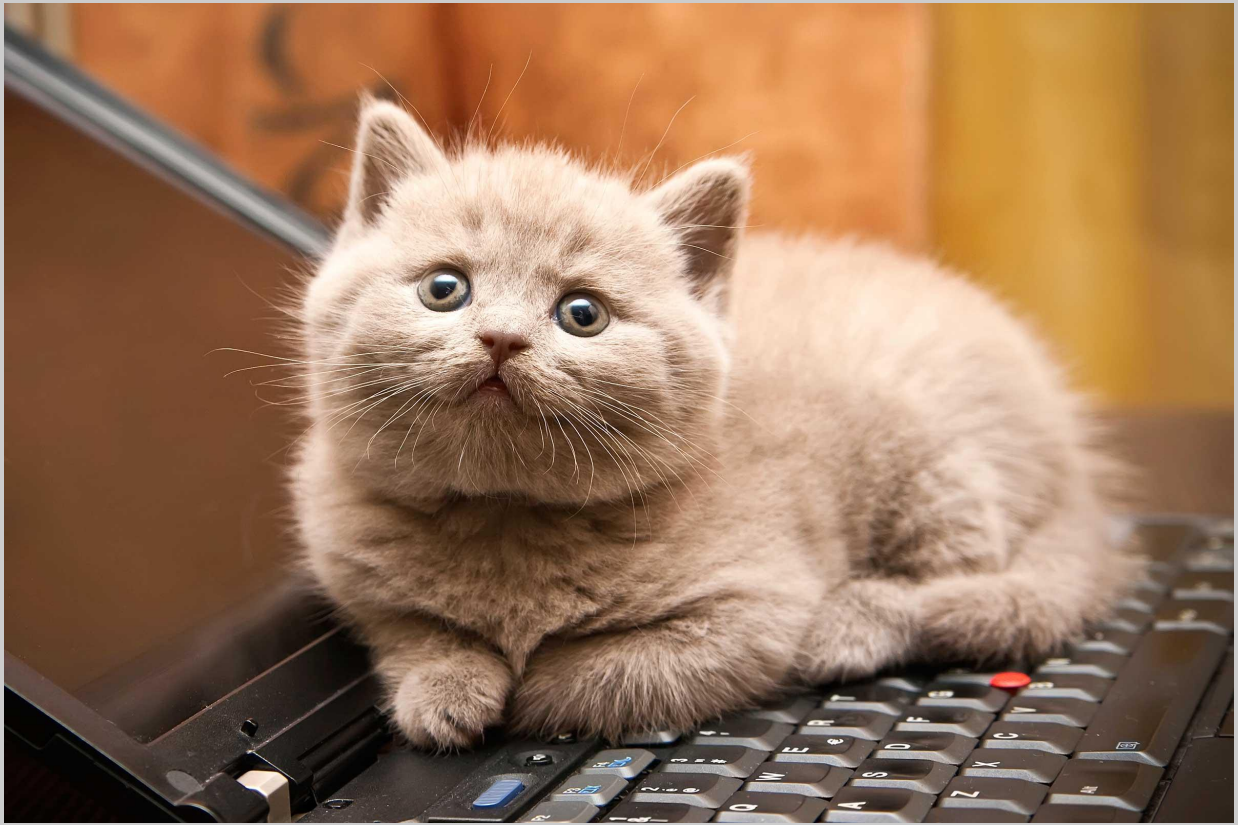
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch



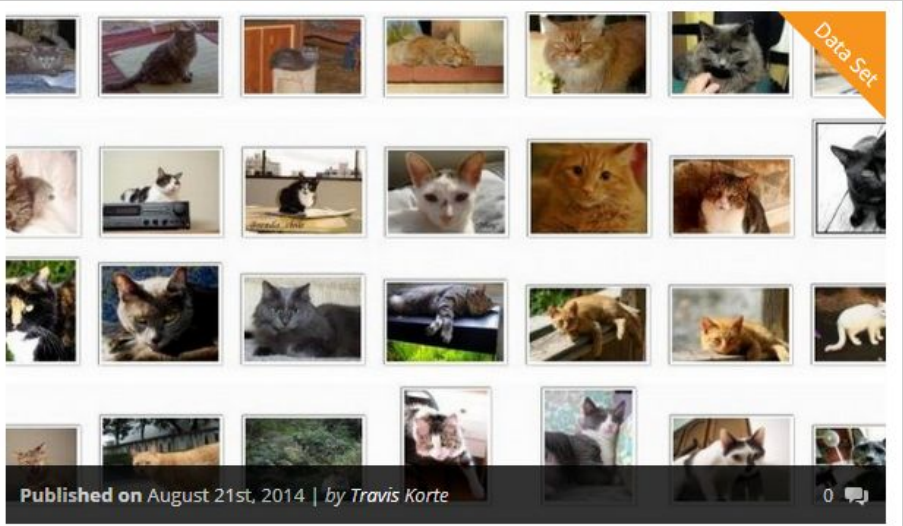
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

```
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"
}
```

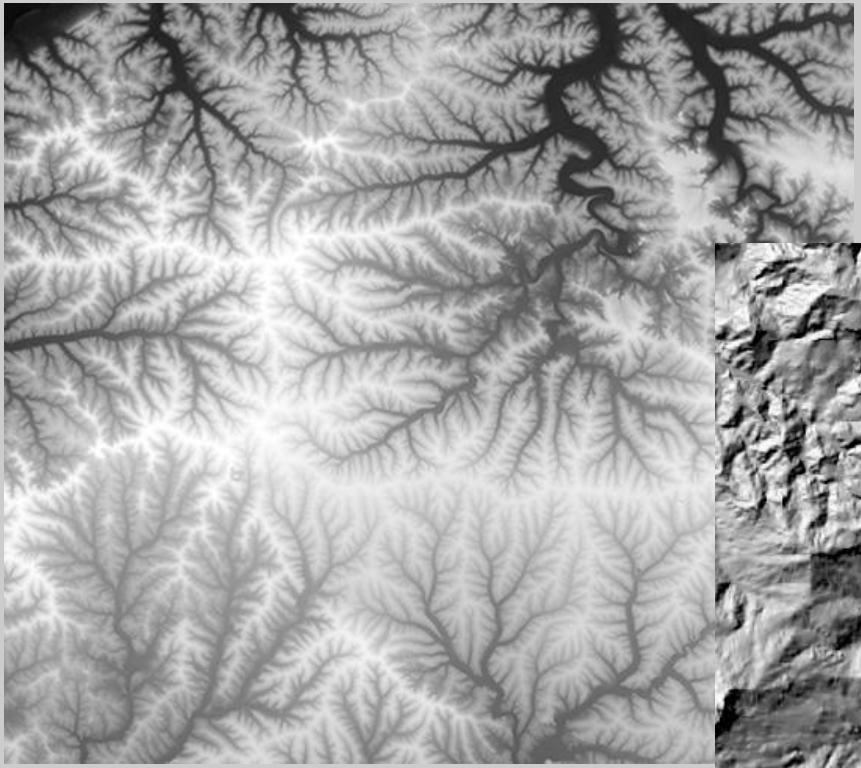


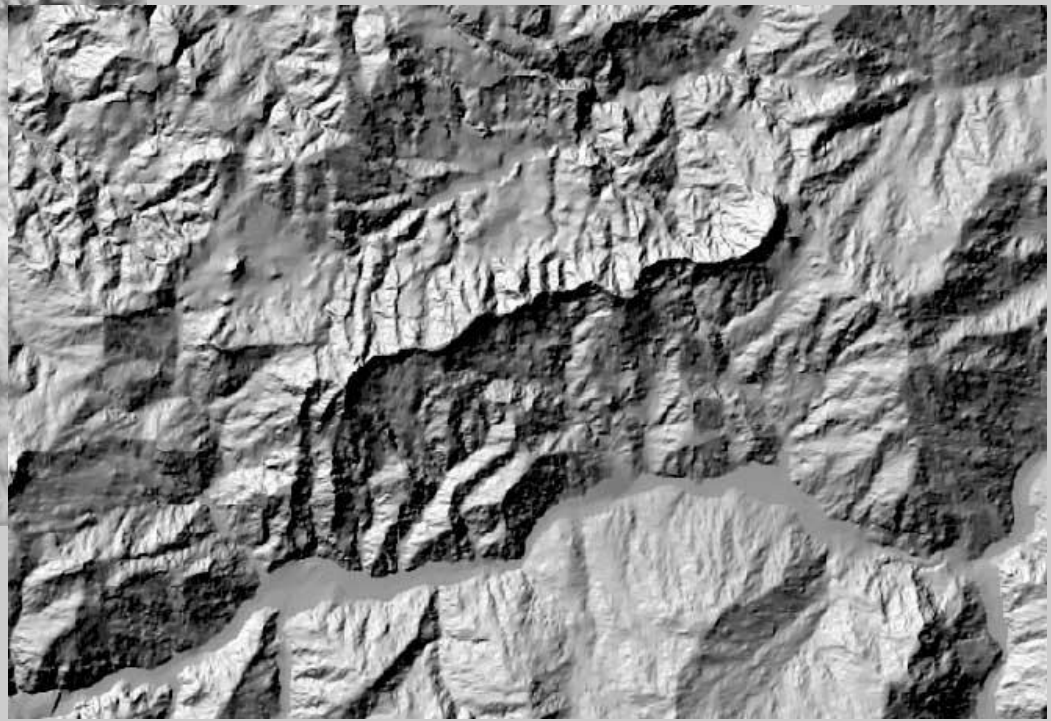
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch



10,000 Cat Pictures (For Science)

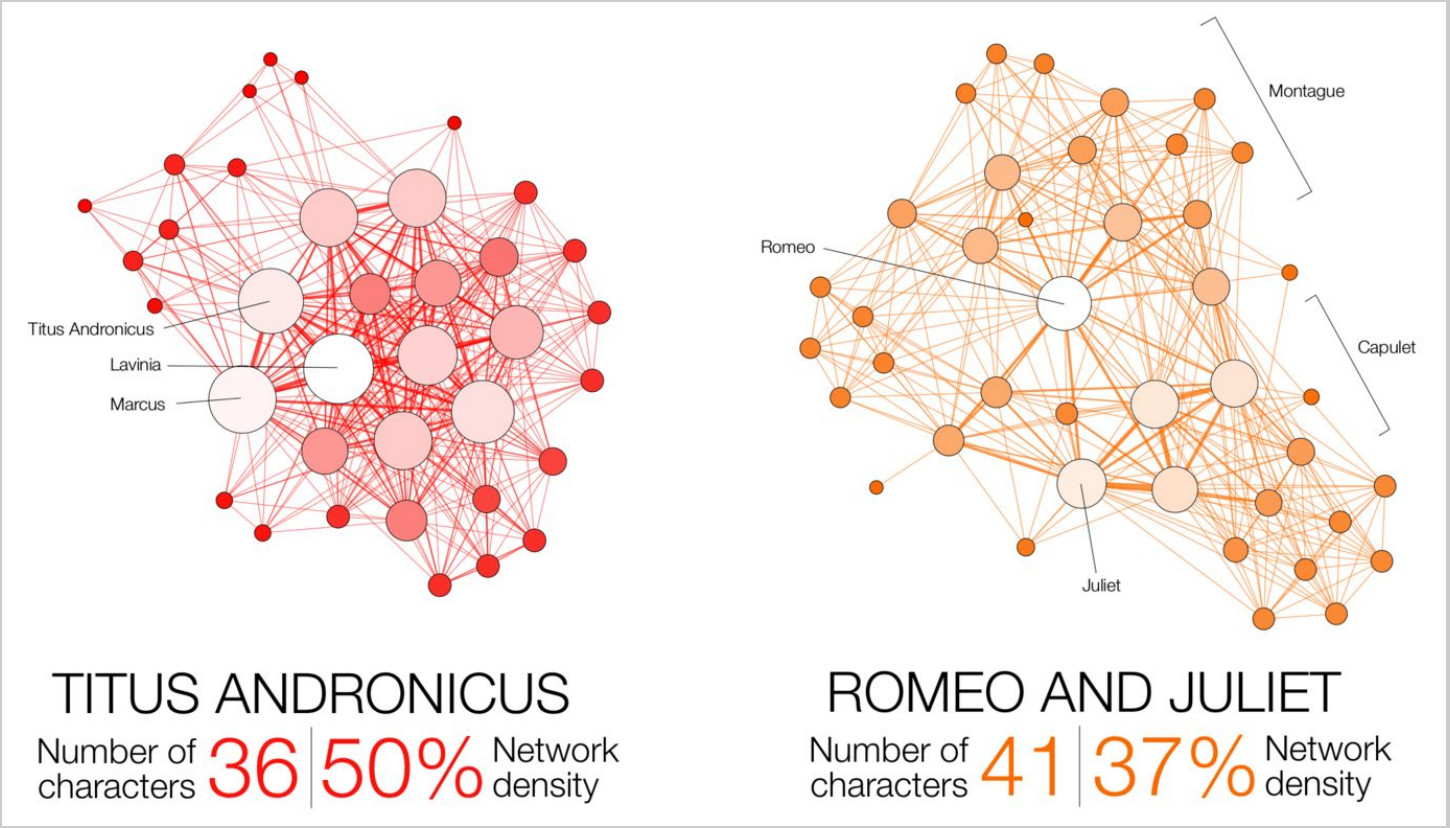
Researchers from **Microsoft Research Asia** and the **Chinese University of Hong Kong** created a database with 10,000 photos containing cat heads to test image recognition algorithms, and now the data set is available freely for research purposes. The photos, which the researchers downloaded mainly 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.





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

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



martingrandjean.ch/network-visualization-shakespeare/

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

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.

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

There are many different types of data formats.

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

Proprietary vs. text

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

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

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

Proprietary (some ex.)

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

Text (some ex.)

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

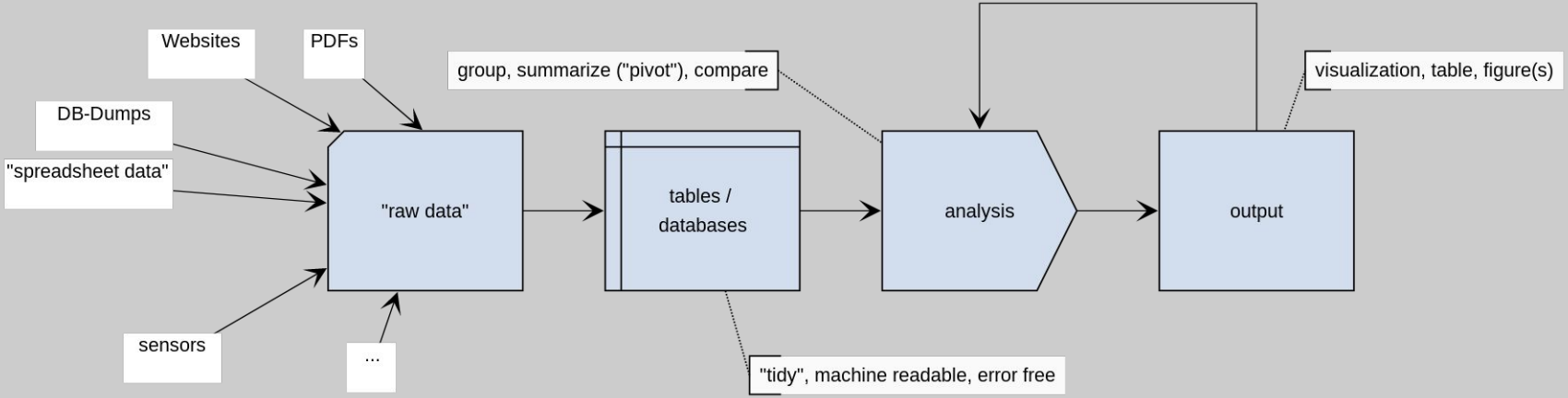
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

The most common formats:

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

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

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

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

Exercise:
Find data for a given task

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

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!

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

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

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

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

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

Key take aways:

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

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

Key take aways:

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

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

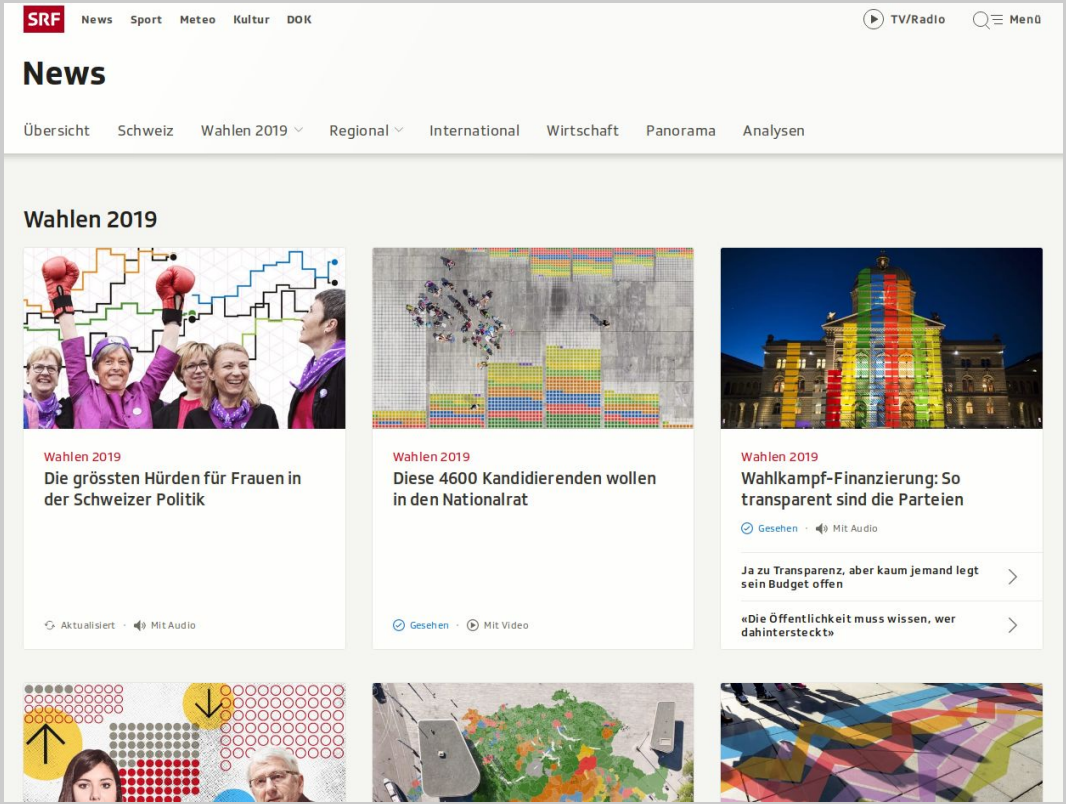
My job: Data Journalist

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

My job: Data Journalist (your future job?)

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

Who we are and what we do: srf.ch/data



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

We are not alone...

NZZ Visuals (previously NZZ Storytelling)

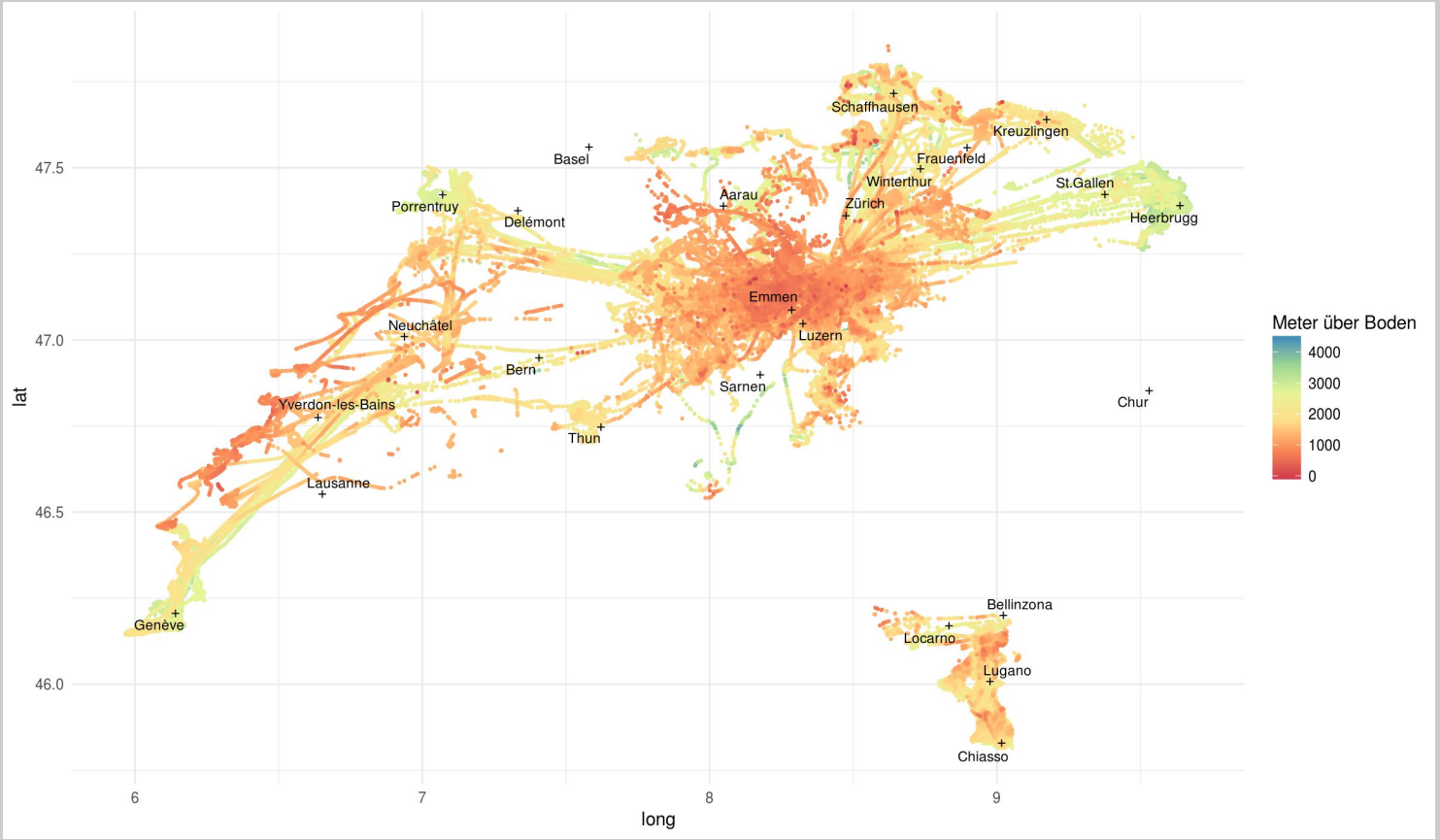
TA Interaktiv / Datenblog / Tamedia

Data Le Temps / RTS, etc.

swissinfo.ch

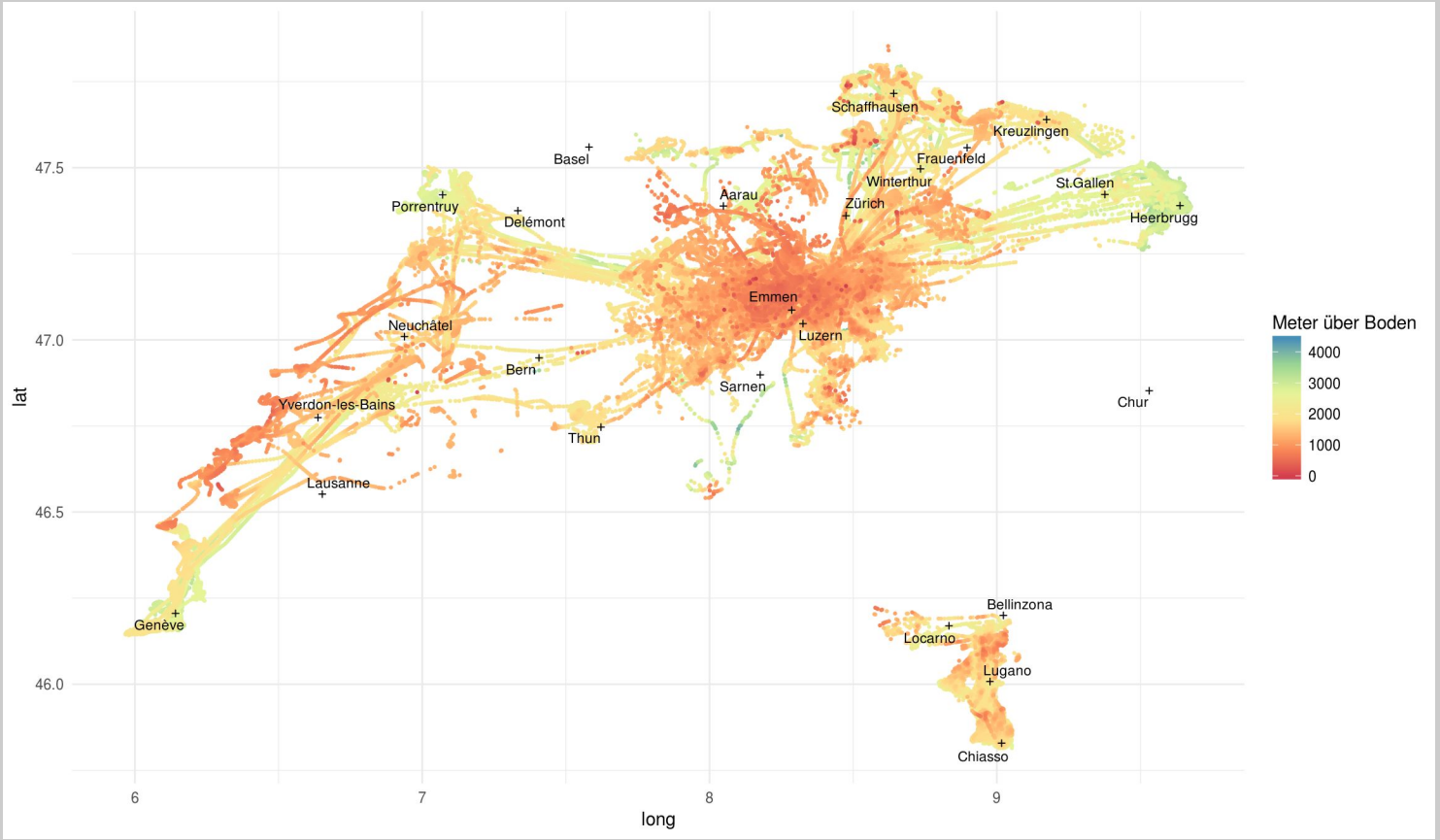
Republik

...

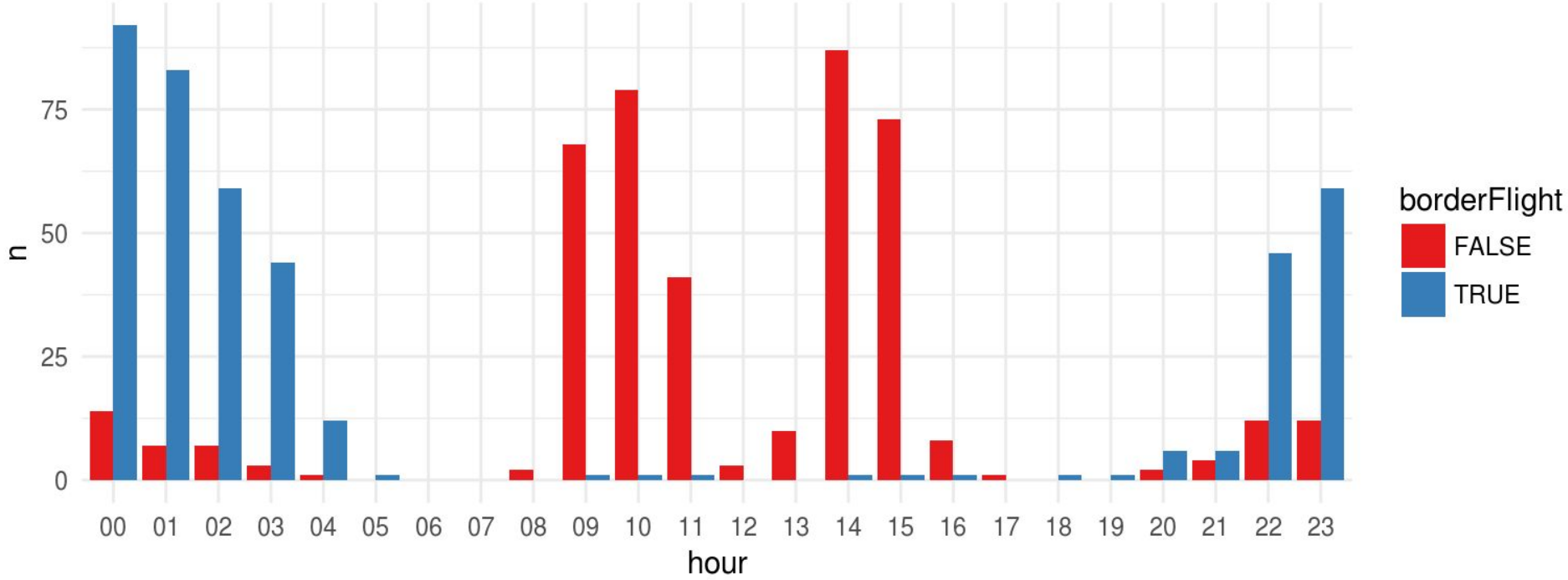


Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

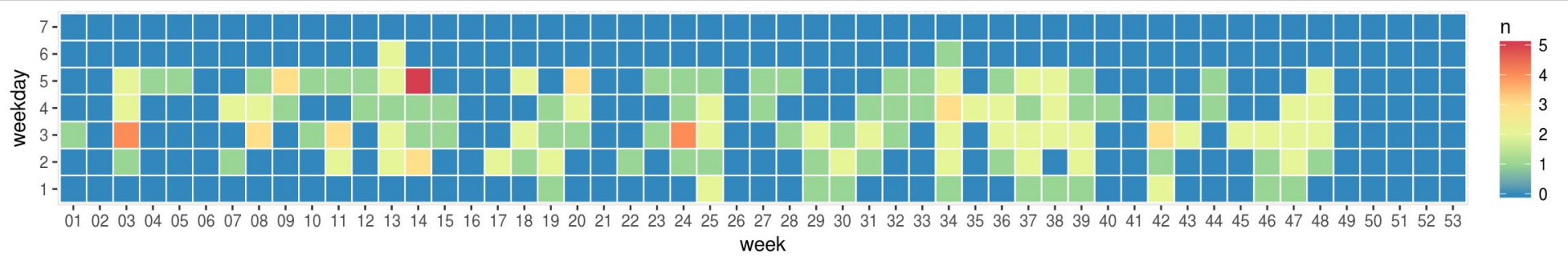
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch



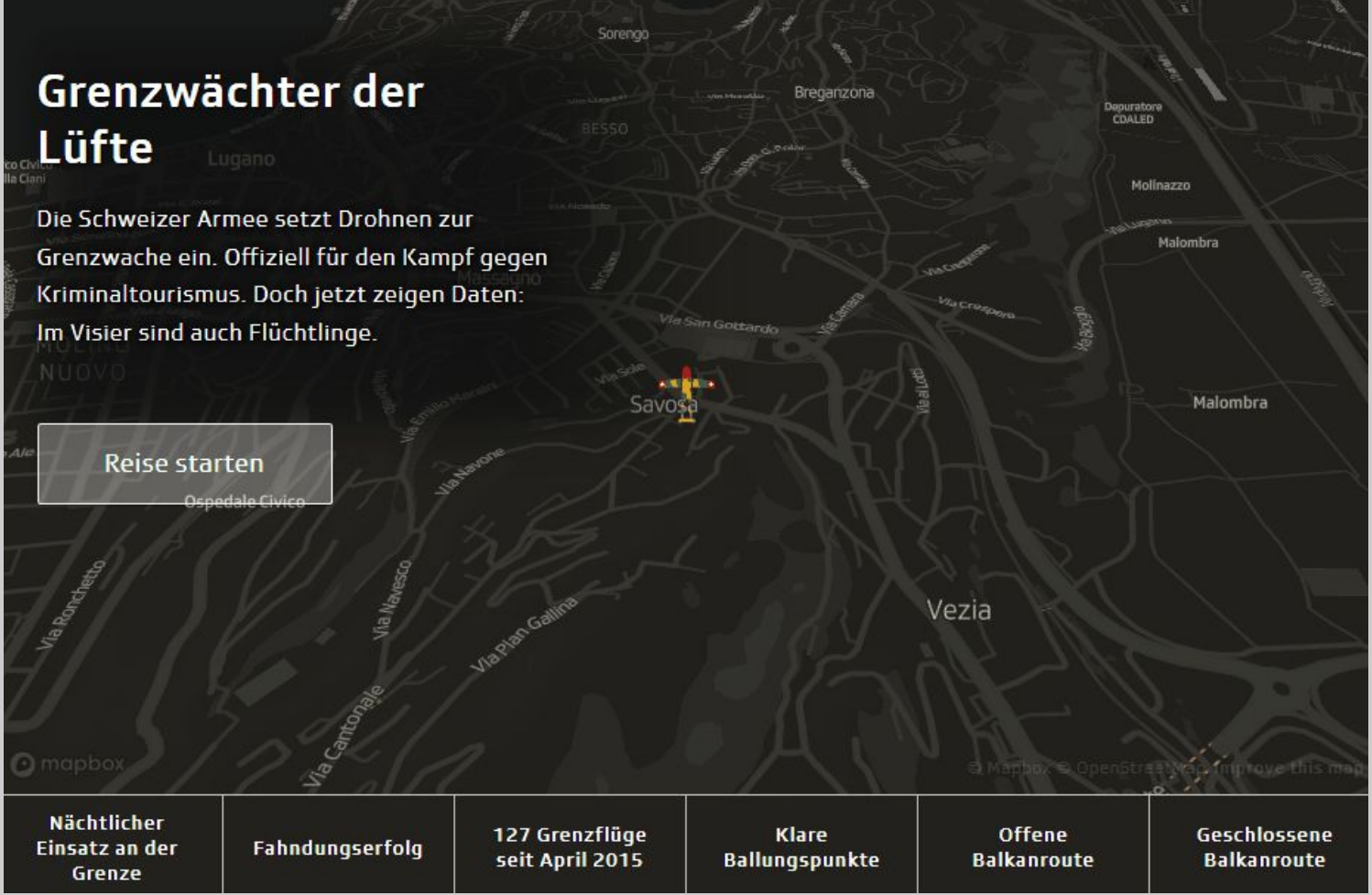
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch



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



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



Nächtlicher Einsatz an der Grenze	Fahndungserfolg	127 Grenzflüge seit April 2015	Klare Ballungspunkte	Offene Balkanroute	Geschlossene Balkanroute

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

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

Another example: Identifying Fake Followers

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

goo.gl/muEUT6

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

srfdata.github.io

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