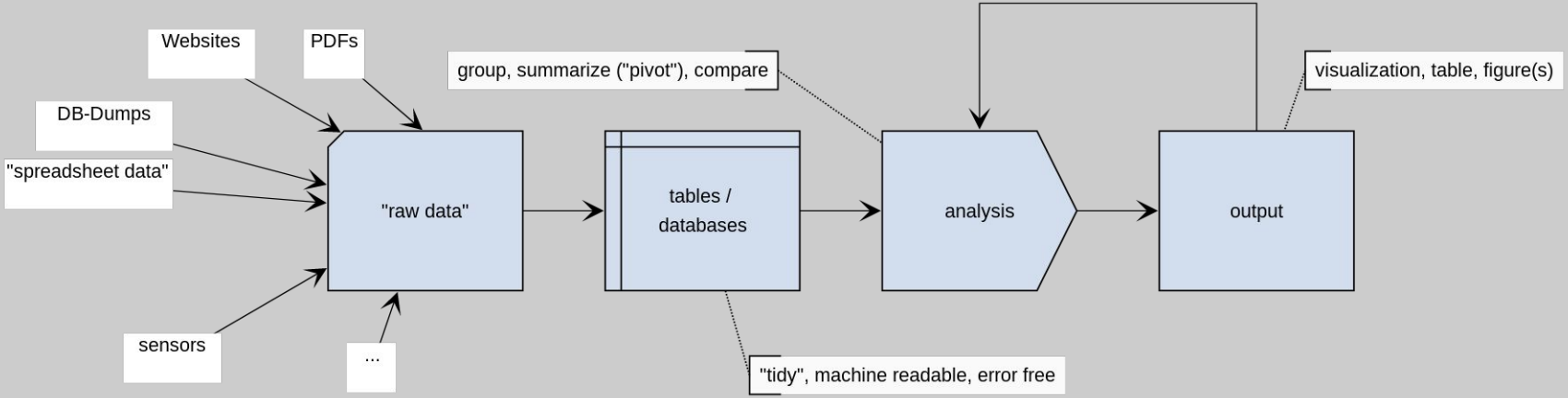
Last time:

1. Raw vs. refined data
2. Structured vs. unstructured data
3. Proprietary vs. text
4. Data file structures / formats

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

Types of data sources: A classification

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

Degree of processing

Raw, fine granular, all-encompassing

Example: Real-Time Flight Table

Verified, clean, aggregated, simplified, anonymized

Example: Flights per time slot

Zeit	Erw.	Nach	Flug	Check-in	Gate	Status
20:10	20:25	**DUBLIN**	EI 349	2	E47	zum Gate
20:15		**LONDON**	EZY 8118	3	E58	einsteigen
20:20		**HAMBURG**	EW 7765	2	A71	einsteigen

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

Degree of processing

Raw, fine granular, all-encompassing

(Possible) Advantages:
- Higher level of detail
- More recent than aggregated

(Possible) Disadvantages:
- Not verified
- Hard to compare to other data sets

Verified, clean, aggregated, simplified, anonymized

(Possible) Advantages:
- Usually more reliable than raw data
- Comparable
- Complicated aggregation steps have already been conducted

(Possible) Disadvantages:
- «boring», too broad
- Hard to reproduce / hard to question because aggregation steps are often unknown

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

Genesis

Accumulated, uncontrolled, digital traces, secondary products

(Possible) Advantages:
- No efforts needed for collection
- «Full survey», no sampling
- «Natural» context

(Possible) Disadvantages:
- Uncontrolled / context of collection is unknown
- Can be manipulated

More sources: UGC / VGI / Crowdsourcing

Purposefully collected, controlled, operationalized, primary products

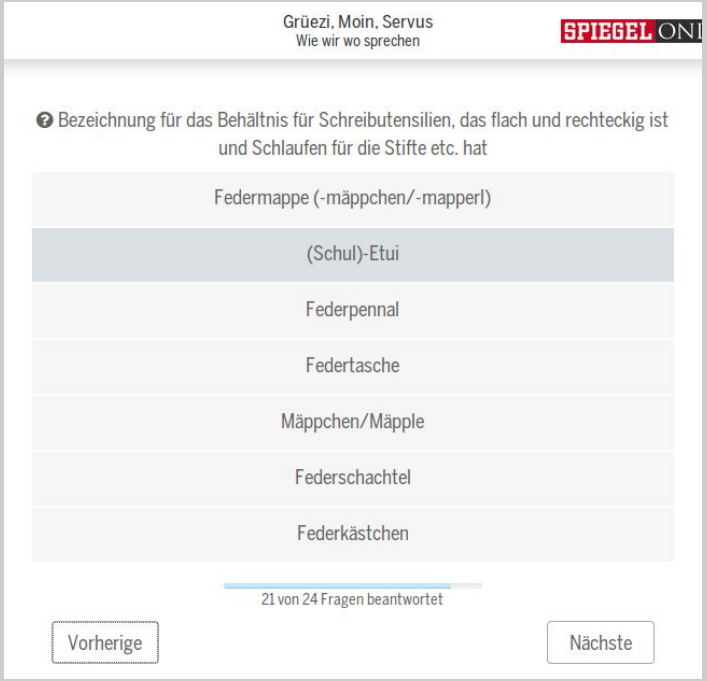
(Possible) Advantages:
- Tailored to specific questions
- Context is known and controllable

(Possible) Disadvantages:
- Small samples
- Tedious to collect
- Experimental / «laboratory»

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

Digression: Crowdsourcing

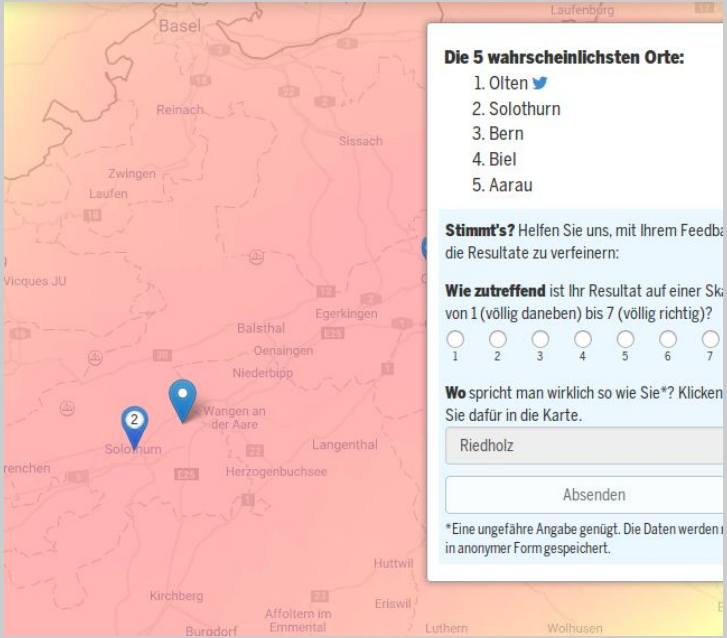
- April 2015: Publication «Sprachatlas» w/ Tages-Anzeiger and Spiegel Online, 2 Mio. Unique Visitors within one week



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

Crowdsourcing

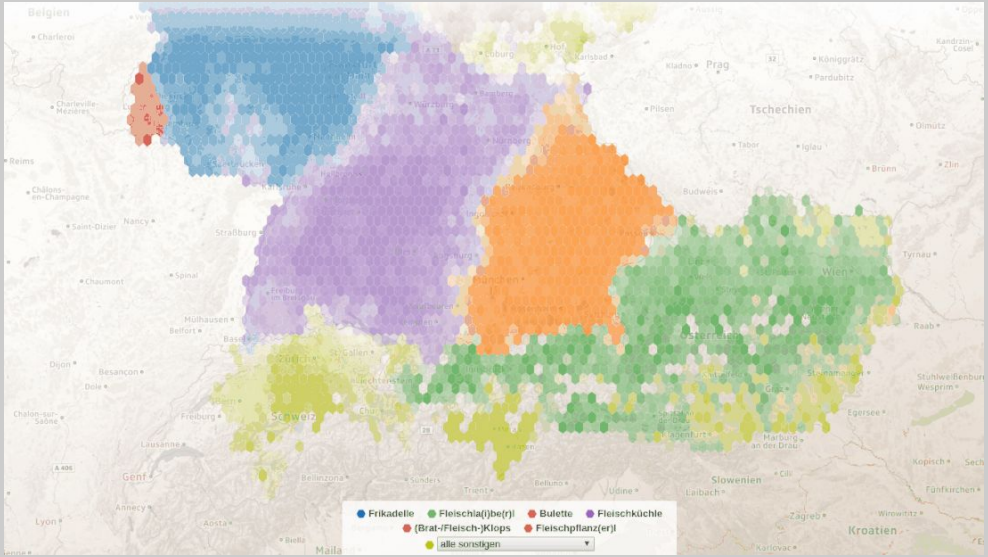
- April 2015: Publication «Sprachatlas» w/ Tages-Anzeiger and Spiegel Online → 2 Mio. Unique Visitors within one week
- Over 750’000 persons gave feedback on their residence and how they pronounce 25 different words



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

Crowdsourcing

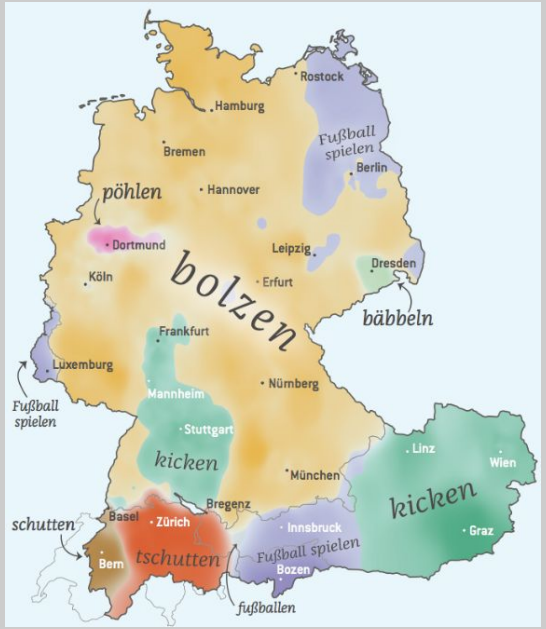
- April 2015: Publication «Sprachatlas» w/ Tages-Anzeiger and Spiegel Online → 2 Mio. Unique Visitors within one week
- Over 750’000 persons gave feedback on their residence and how they pronounce 25 different words
- August 2016: Publication of these feedback data as Hexagon-Maps (Tages-Anzeiger, Spiegel Online, SRF Data)



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

Crowdsourcing

- April 2015: Publication «Sprachatlas» w/ Tages-Anzeiger and Spiegel Online → 2 Mio. Unique Visitors within one week
- Over 750’000 persons gave feedback on their residence and how they pronounce 25 different words
- August 2016: Publication of these feedback data as Hexagon-Maps (Tages-Anzeiger, Spiegel Online, SRF Data)
- December 2017: Publication of a book

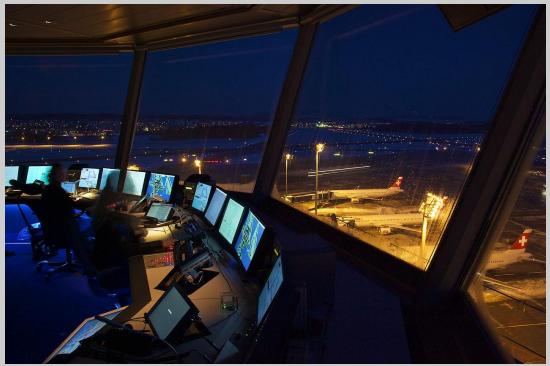


Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

Accessibility

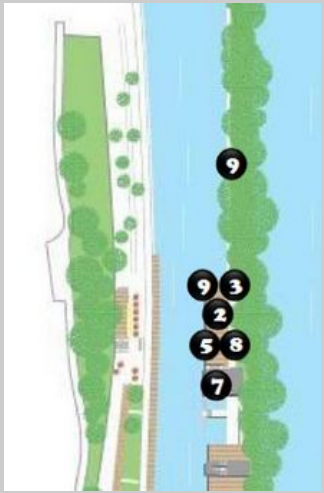
Hidden / private / secret, undocumented, expensive, unreadable, *unknown*

Example: Skyguide Radar Data



Open(ed) / public, documented, free / cheap, machine readable

Example: Tagi-Badi-App, based on open data of the city of Zurich



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

Accessibility

Hidden / private / secret, undocumented, expensive, unreadable, *unknown*

(Possible) Advantages:
- Interesting, because new

(Possible) Disadvantages:
- Uncontrolled / context of collection is hardly controllable
- Hard to work with
- May contain more errors

Web Scraping
PDF Parsing

Open(ed) / public, documented, free / cheap, machine readable

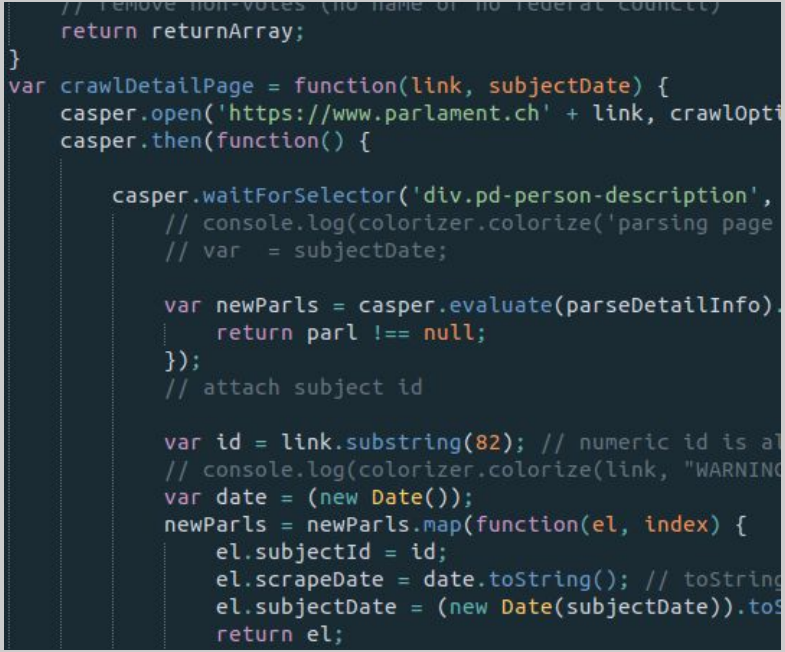
(Possible) Advantages:
- Standardized / machine readable / API
- Controlled by a lot of people, chance of error small

(Possible) Disadvantages:
- «Boring» (on its own…)

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

Digression: Scraping

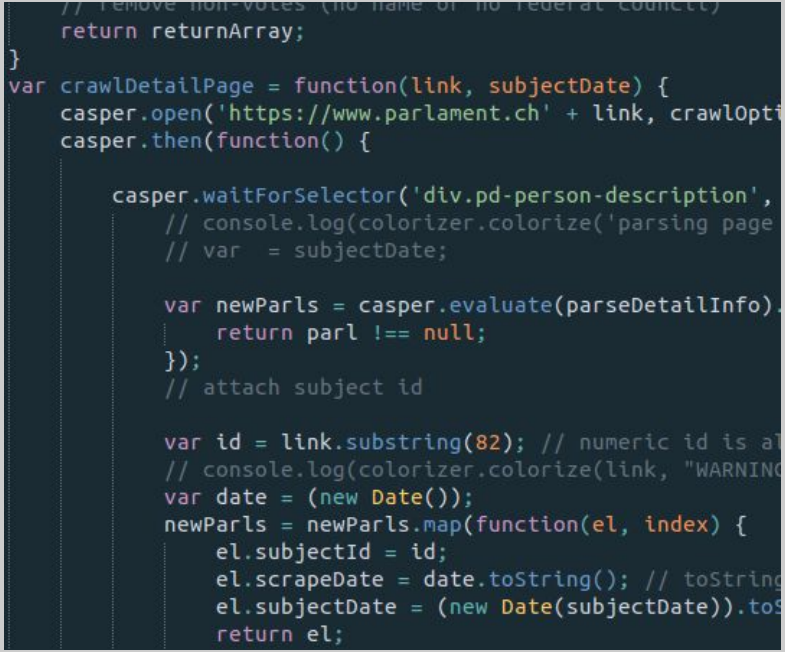
- Often, data are publicly available, but *not structured enough*. E.g. in a table on a website, but not as a downloadable file format.



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

Digression: Scraping

- Often, data are publicly available, but *not structured enough*. E.g. in a table on a website, but not as a downloadable file format.
- We then need to *(screen) scrape* the website for the data.
- Often, this is done in *an automated fashion*, with *self-programmed bots* or software tailored for this task.
- DEMO



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

Scientific seriousness / traceability

Undocumented, not reproducible, ...

(Possible) advantages:
Popular
Niches in which science is not "interested"
Customizable / may be personalized
(Possible) disadvantages:
"Authenticity" / Representativeness / Meaningfulness
PR-Bullshit-Risk

Documented, reproducible, ...

(Possible) advantages:
Comprehensible, contact person, responsibility
(Possible) disadvantages:
Complexity
Small samples irrelevant / “lebensfern”

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

Scientific seriousness / traceability

Undocumented, not reproducible, ...

(Possible) advantages:
Popular
Niches in which science is not "interested"
Customizable / may be personalized
(Possible) disadvantages:
"Authenticity" / Representativeness / Meaningfulness
PR-Bullshit-Risk

Documented, reproducible, ...

(Possible) advantages:
Comprehensible, contact person, responsibility
(Possible) disadvantages:
Complexity
Small samples
irrelevant / “lebensfern”

«The wave of bullshit data is rising, and now it’s our turn to figure out how not to get swept away.»

— Jacob Harris, NYT («A wave of PR data»)

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

Example

- **Where would you position the data set that was used as the foundation of the story?**
  a. ***Degree of processing***
  b. ***Genesis***
  c. ***Accessibility***
  d. ***Scientific seriousness / traceability***
- http://tiny.cc/ooorder

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

- Standards / APIs
- *Machine readable*
- *Tried and tested / credible*

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

Open data is data that can be *freely* used, re-used and *redistributed* by *anyone* - subject only, at most, to the requirement to *attribute and sharealike*.

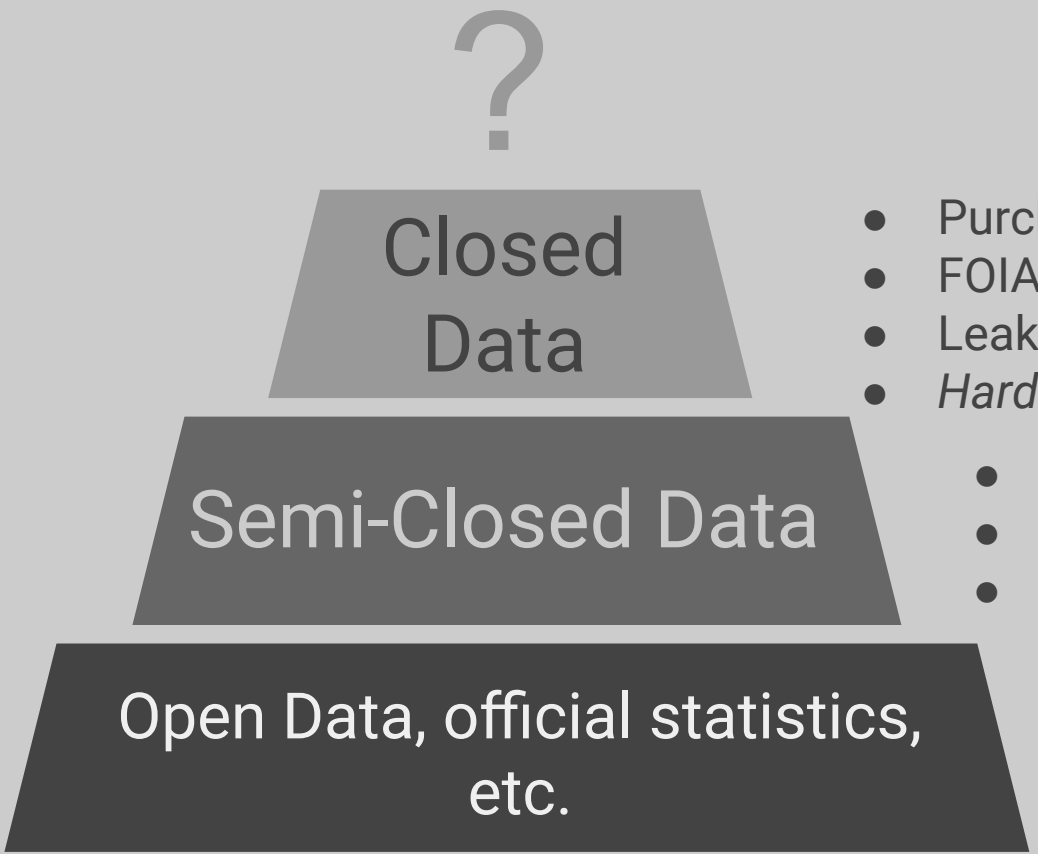
– Open Knowledge Foundation

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

- Scrape
- Parse
- *Incomplete, prone to error*

- Standards / APIs
- *Machine readable*
- *Tried and tested / credible*

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

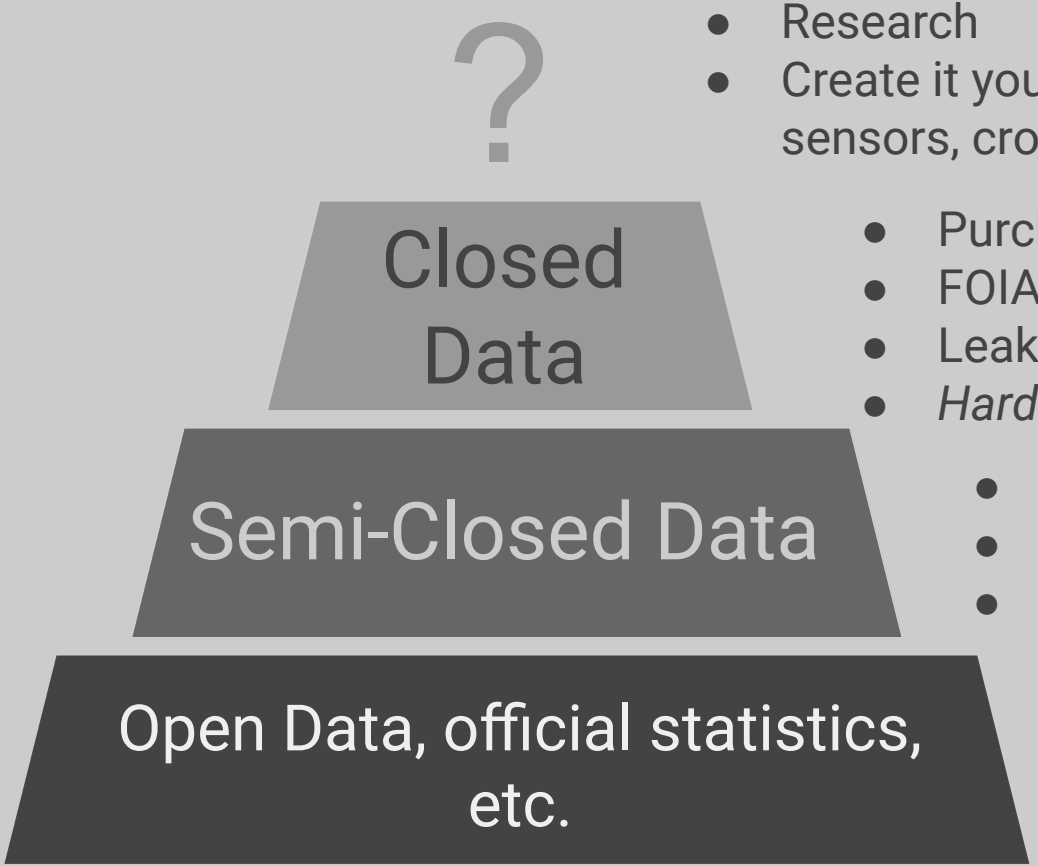


- Purchase
- FOIA («BGÖ□□□»)
- Leak
- *Hard to verify*

- Scrape
- Parse
- *Incomplete, prone to error*

- Standards / APIs
- *Machine readable*
- *Tried and tested / credible*

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



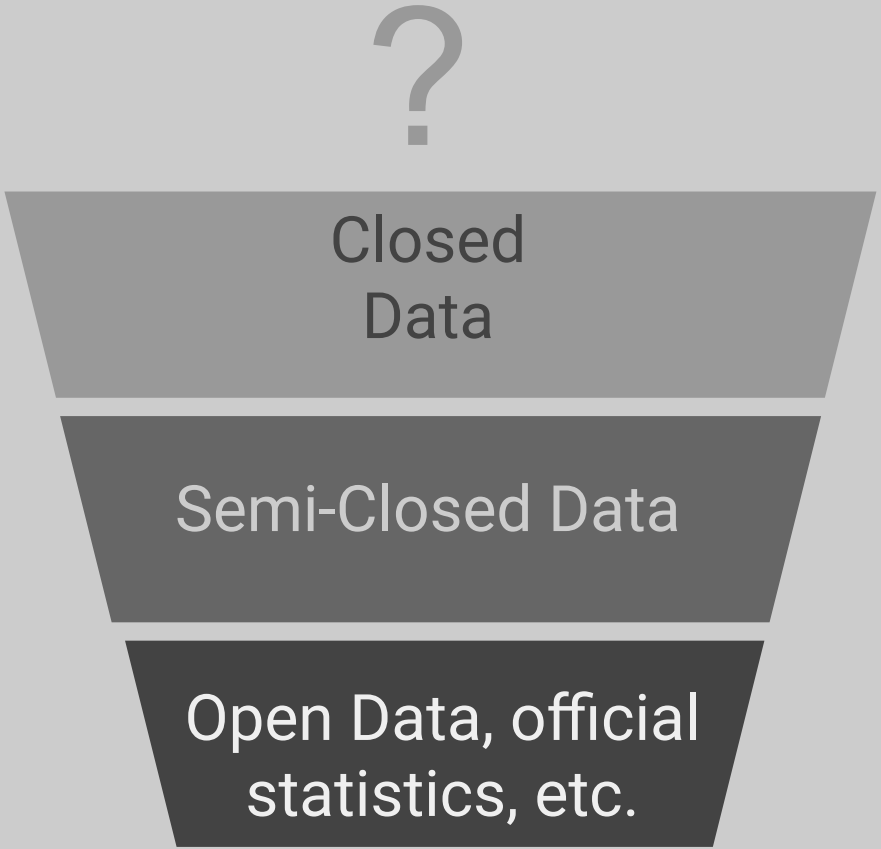
- Research
- Create it yourself: Surveys / polls, sensors, crowdsourcing

- Purchase
- FOIA («BGÖ□□□»)
- Leak
- *Hard to verify*

- Scrape
- Parse
- *Incomplete, prone to error*

- Standards / APIs
- *Machine readable*
- *Tried and tested / credible*

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

Open Data Portals

opendata.swiss
opentransportdata.swiss
data.stadt-zuerich.ch / ogd.tg.ch / etc.
open-data.europa.eu

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

Open Data Hackdays

hack.opendata.ch

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

There are many different types of data sources.

Each has their advantage / disadvantage.

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

Data Quality

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

Data Quality
is (more or less) defined as:
Data is suited for the task at hand

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

Types of error

- Can be detected automatically and corrected automatically 😅
- Can be detected automatically and corrected manually 😉
- Can be neither be detected nor corrected automatically 😭

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

Data Quality

- Values are missing
- Date formats are inconsistent
- Spelling is inconsistent
- Text is garbled (“encoding hell”)
- Data type is inconsistent
- ...

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

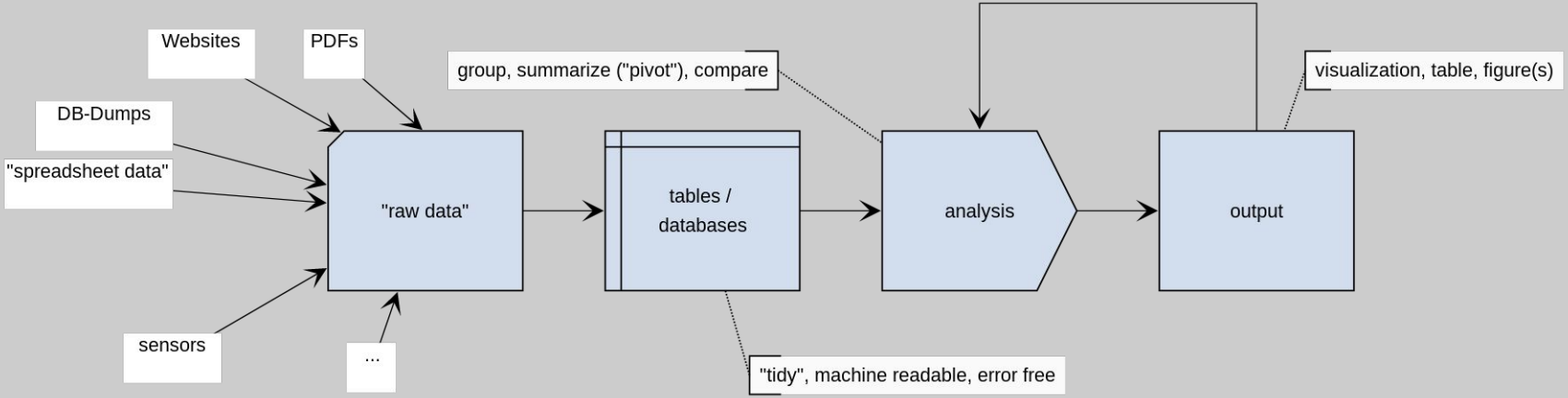
Name	Praxisadresse	Erhaltene Spende	Mitglied bei FMH?
Dr. Lucius M?ller	Scheuchzerstrasse 10	24000 CHF	Ja
MD Joshua Greene	Hofwiesenstrasse 23, 8006 Z?rich	3400000	Nein
Frau Dr. Stirnimann	G?rtliweg 20	2004.55	
Guido Beckenried	Brunnenhof, Aarau	23000	TRUE
	Dr. med. Edward Rainden	Idaplatz, Z?rich	2300
TRUE	Dr. Med. Lucius Mueller	Scheuchzerstrasse 10, Z?rich	2000

The «Tidy Data» Principle

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

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

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

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

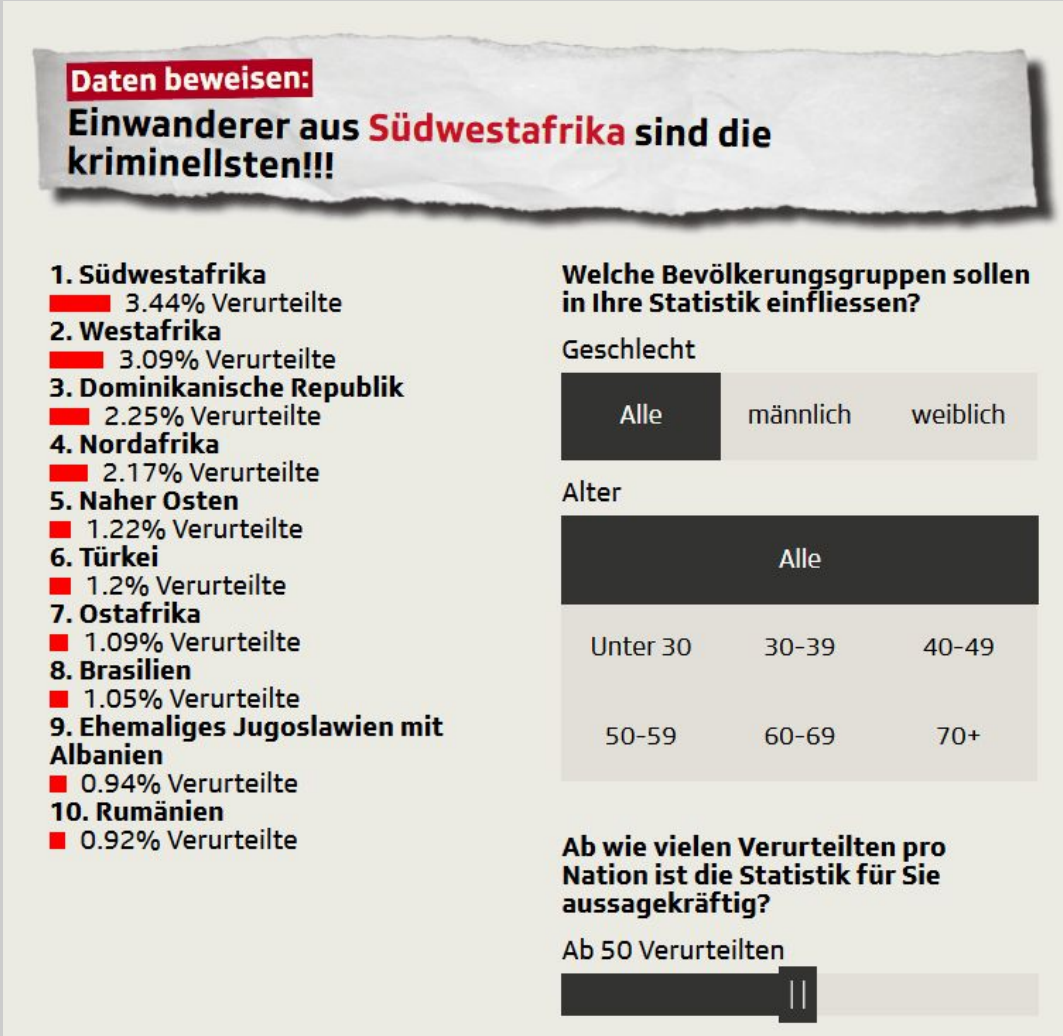
country	year	m014	m1524	m2534
AD	2000	0	0	1
AE	2000	2	4	4
AF	2000	52	228	183
AG	2000	0	0	0
AL	2000	2	19	21
AM	2000	2	152	130
AN	2000	0	0	1
AO	2000	186	999	1003
AR	2000	97	278	594
AS	2000	—	—	—

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

>

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

http://www.srf.ch/news/schweiz/auslaenderkriminalitaet-eine-statistik-viele-schlagzeilen



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

Daten beweisen:

Einwanderer aus Südwestafrika sind die kriminellsten!!!

	Land	Geschlecht	Altersgruppe	Verurteilte_Mean	Belastungsrate_Mean	Population_Gruppe	Populatio
1	Afghanistan	f	1829	1	2.9	345	2055
2	Afghanistan	f	3039	1	4.9	204	2055
7	Afghanistan	m	1829	8	15.6	513	2055
8	Afghanistan	m	3039	2	5.2	385	2055
9	Afghanistan	m	4049	2	10.25	195	2055
10	Afghanistan	m	5059	1	10.6	94	2055
43	Bangladesch	m	1829	4	90.9	44	1150
44	Bangladesch	m	3039	4	11.2	357	1150
45	Bangladesch	m	4049	5	23.7	211	1150

■ 0.92% Verurteilte

Ab wie vielen Verurteilten pro Nation ist die Statistik für Sie aussagekräftig?

Ab 50 Verurteilten

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

Schweizer(innen) und Ausländer(innen) mit B- und C-Ausweis: Anzahl verurteilte Personen und Be
Verberechen des Strafgesetzbuches (StGB), nach Nationalität, Alter und Geschlecht, 2014

Nationalität	Schweizer und Ausländer mit B- und C-Ausweis (Total)		Männer: Sch				
			18 - 29 Jahre		30 - 39 Jahre		4
	N min. (ohne unbekannten Aufenthaltsstatus) 1)	N max. (mit unbekannten Aufenthaltsstatus) 2)	N min.	N max.	N min.	N max.	N r
Afghanistan	14	16	8	8	2	2	
Argentinien	2	2	X	X	X	X	
Australien	1	1	X	X	X	X	
Bangladesch	13	13	4	4	4	4	
Belgien	15	20	6	7	2	3	
Bolivien	11	16	4	5	3	4	
Brasilien	162	173	48	51	28	30	
Bulgarien	13	28	4	6	1	1	
Chile	29	33	3	4	10	12	

http://www.bfs.admin.ch/bfs/portal/de/index/themen/19/01/new.Document.206879.xls

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

Schweizer(innen) und Ausländer(innen) mit B- und C-Ausweis: Anzahl verurteilte Personen und Be
Verberechen des Strafgesetzbuches (StGB), nach Nationalität, Alter und Geschlecht, 2014

1. Var: Sex

2. Var: Age group

Männer: Sch

Nationalität	Schweizer und Ausländer mit B- und C-Ausweis (Total)		18 - 29 Jahre		30 - 39 Jahre		4
	N min. (ohne unbekannten Aufenthaltsstatus) 1)	N max. (mit unbekannten Aufenthaltsstatus) 2)	N min.	N max.	N min.	N max.	N r
Afghanistan	14	16	8	8	2	2	
Argentinien	2	2	X	X	X	X	
Australien	1	1	X	X	X	X	
Bangladesch	13	13	4	4	4	4	
Belgien	15	20	6	7	2	3	
Bolivien	11	16	4	5	3	4	
Brasilien	162	173	48	51	28	30	
Bulgarien	13	28	4	6	1	1	
Chile	29	33	3	4	10	12	

3. Variable: Nationality

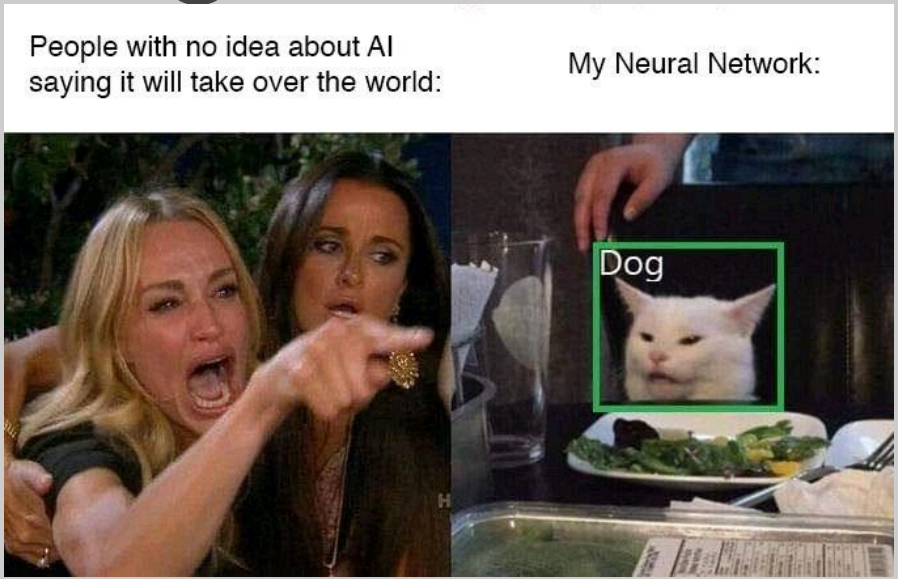
4. Variable: Number of convicted

5. Variable?

http://www.bfs.admin.ch/bfs/portal/de/index/themen/19/01/new.Document.206879.xls

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

Digression: AI



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

Use Cases (as of today)

- Speech recognition / synthesization
- Image / face recognition
- Image enhancement
- Recommendation systems
- Classification systems
- Unmanned vehicles
- Art (re)creation
- Many many more

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

your Profile Photo, Image may contain: 1 person, smiling, eyeglasses and closeup

Timeline Ab

1 Pending Item

Intro

data (journalism) & mountains

Add Info About You

may contain: outdoor, natur

Image may contain: cloud, sky, mountain, nature and outdoor

Image may contain: mountain, sky, outdoor and nature

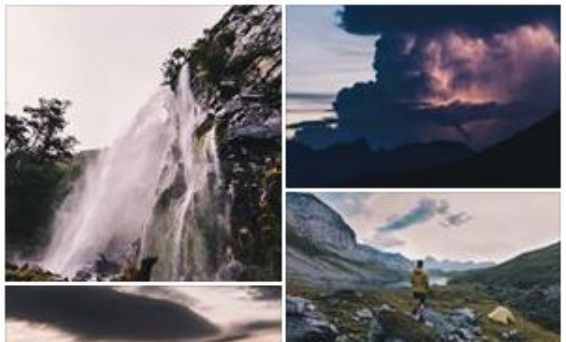
Timeline Abo

1 Pending Item

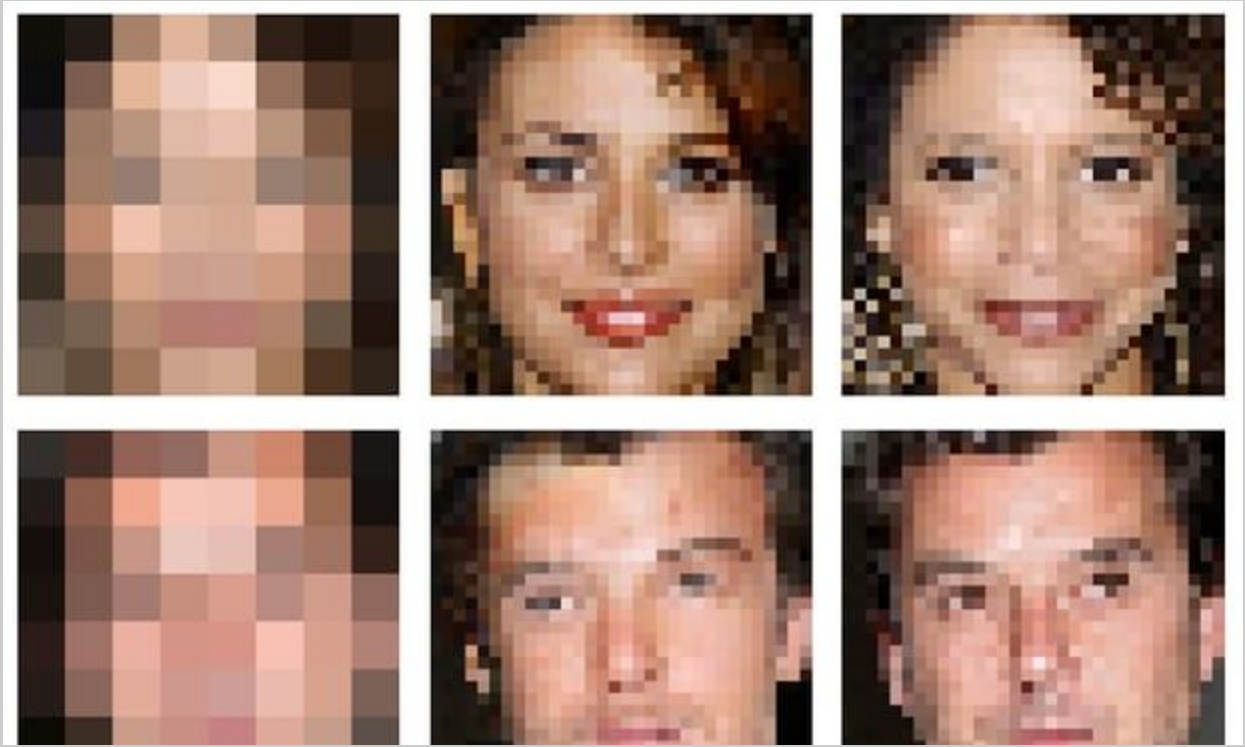
Intro

data (journalism) & mountains

Add Info About You



Timo Grossenbach
@grssnbchr // timo



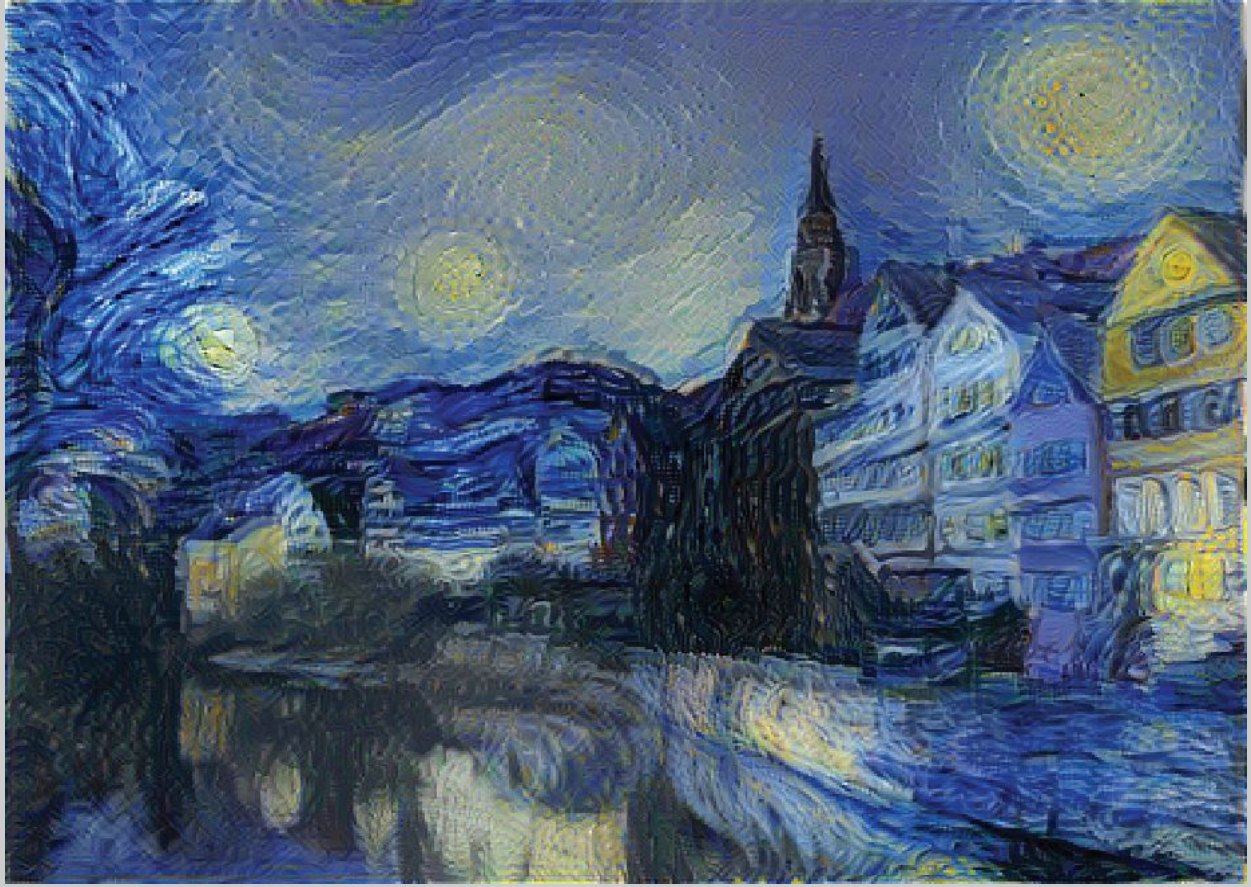
https://www.theguardian.com/technology/2017/feb/08/google-ai-system-pixelated-faces-csi

https://www.thispersondoesnotexist.com/

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

Tim
@g

Timo Grossen
@grssnbchr //



http://uk.businessinsider.com/the-science-how-vincent-van-gogh-saw-the-world-2015-9

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

NO. 1:
Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai “Ngaje Ngai,” the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.
NO. 2:
Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

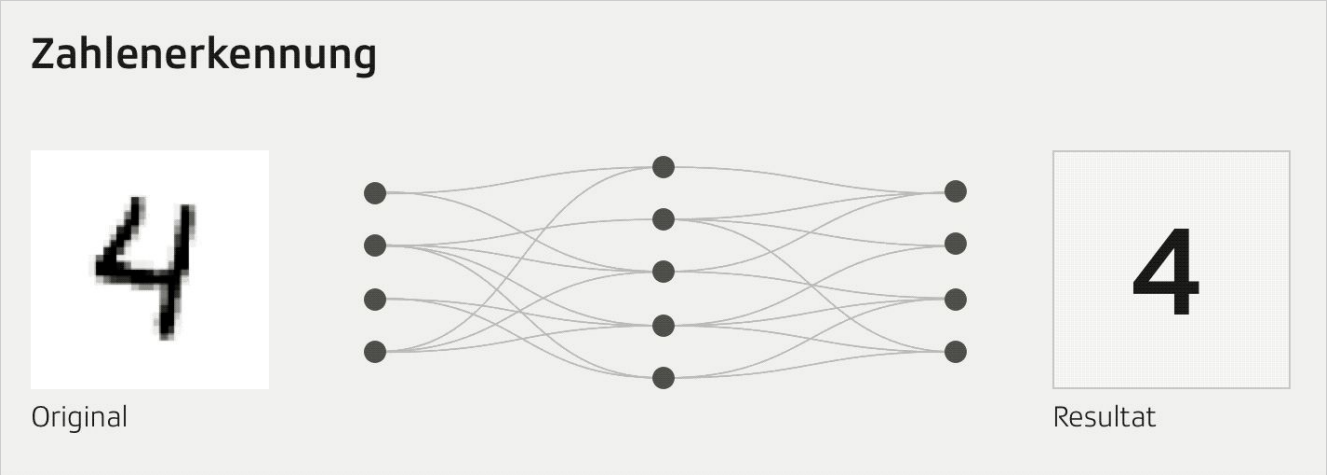
From: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

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

Deep Fake:
https://www.srf.ch/news/panorama/verblueffende-videofaelschungen-von-magie-nicht-mehr-zu-unterscheiden

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

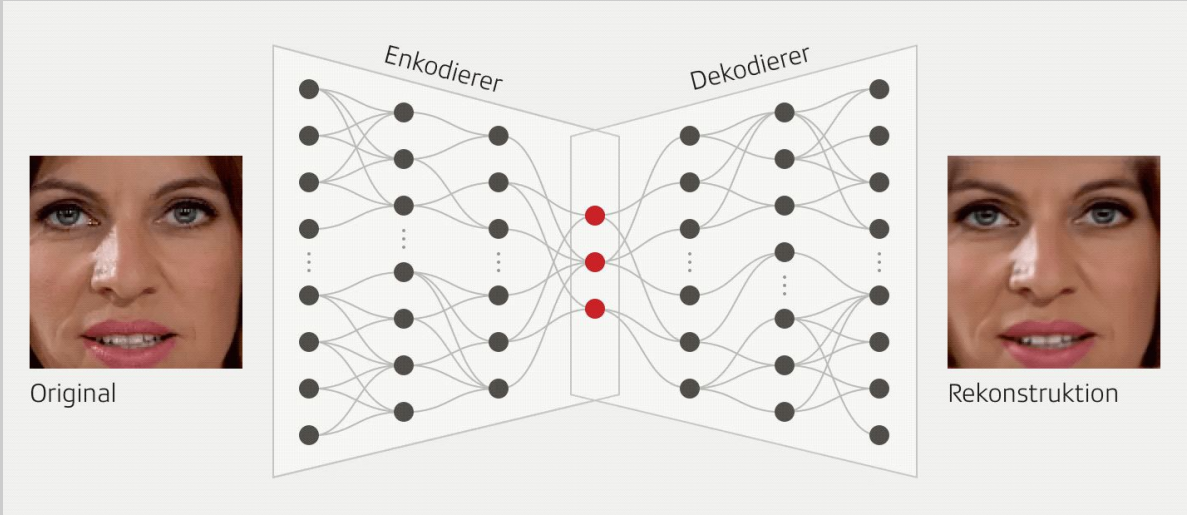
Neural Networks



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

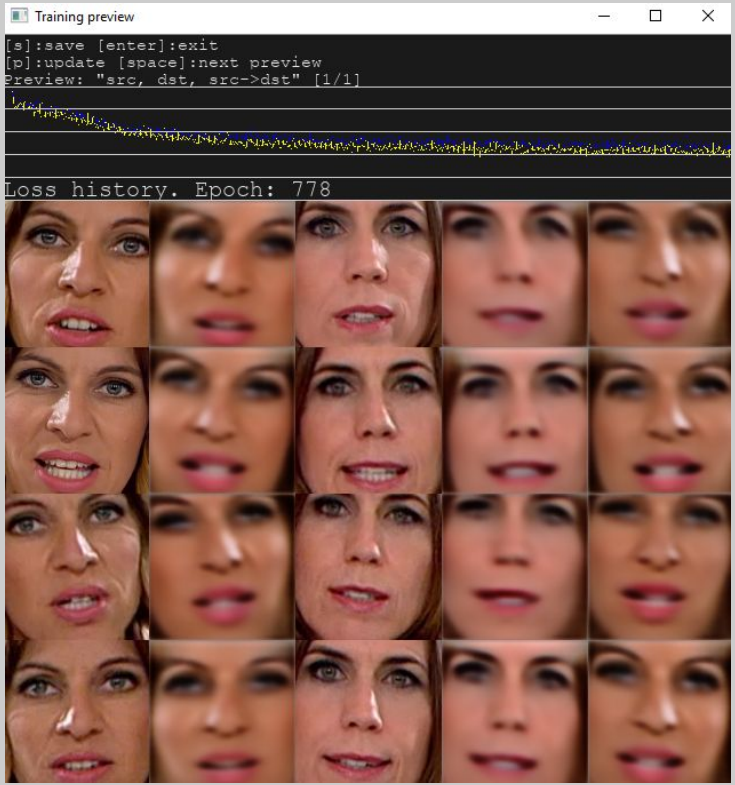
SRF DATA

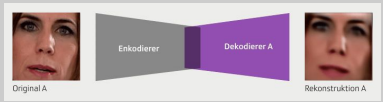
Deep Neural Networks



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

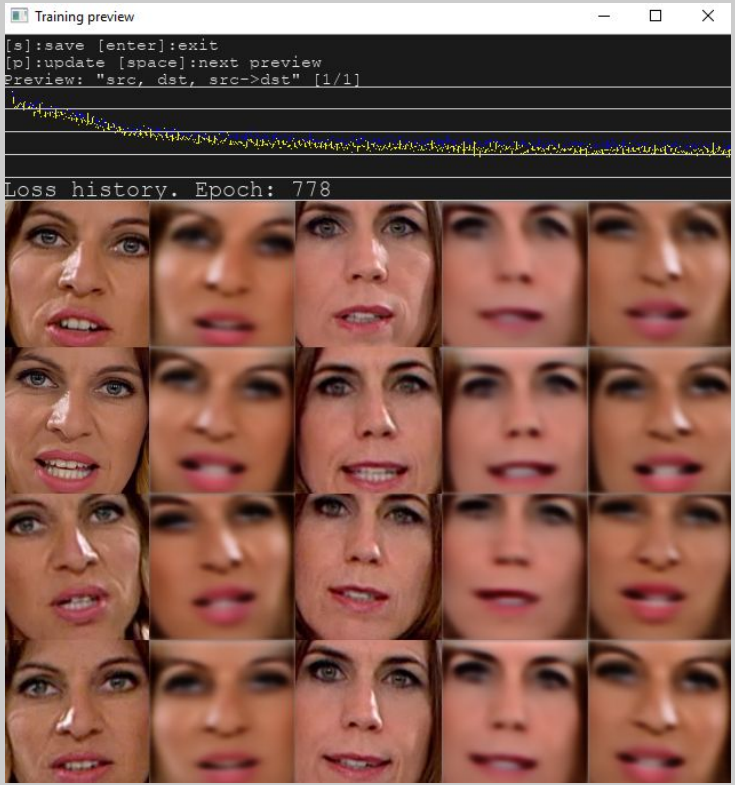
After 5 Min.

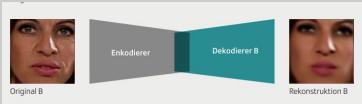




Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

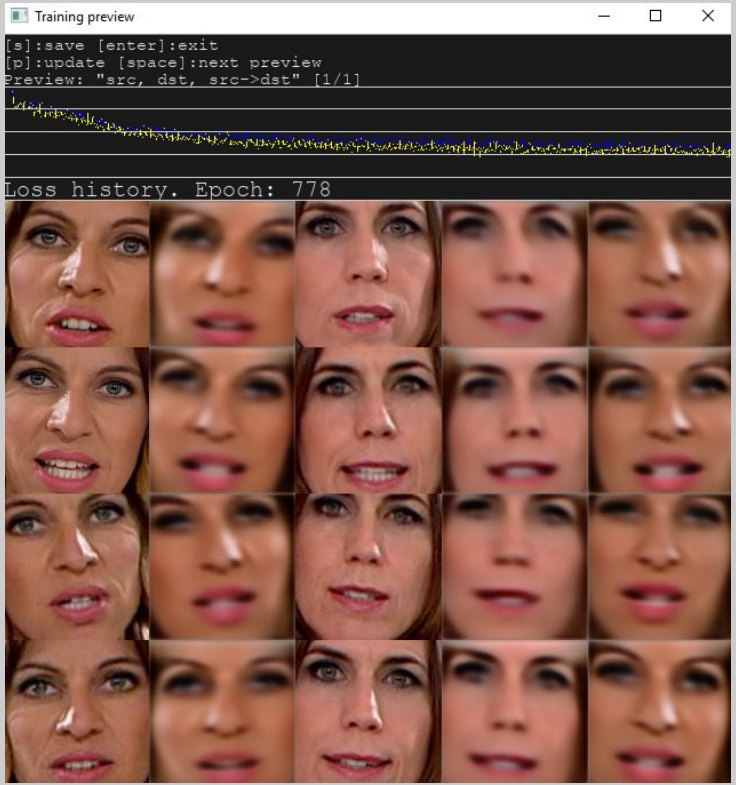
After 5 Min.

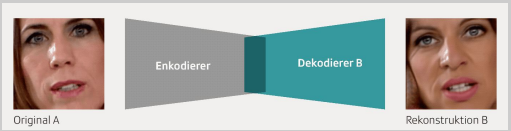




Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

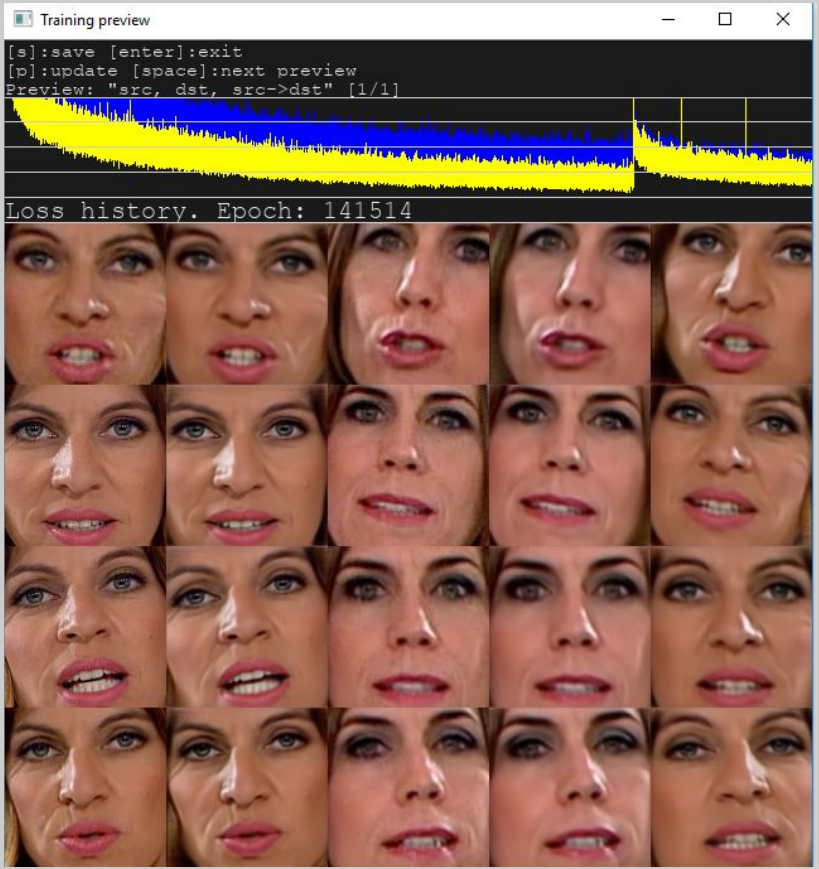
After 5 Min.





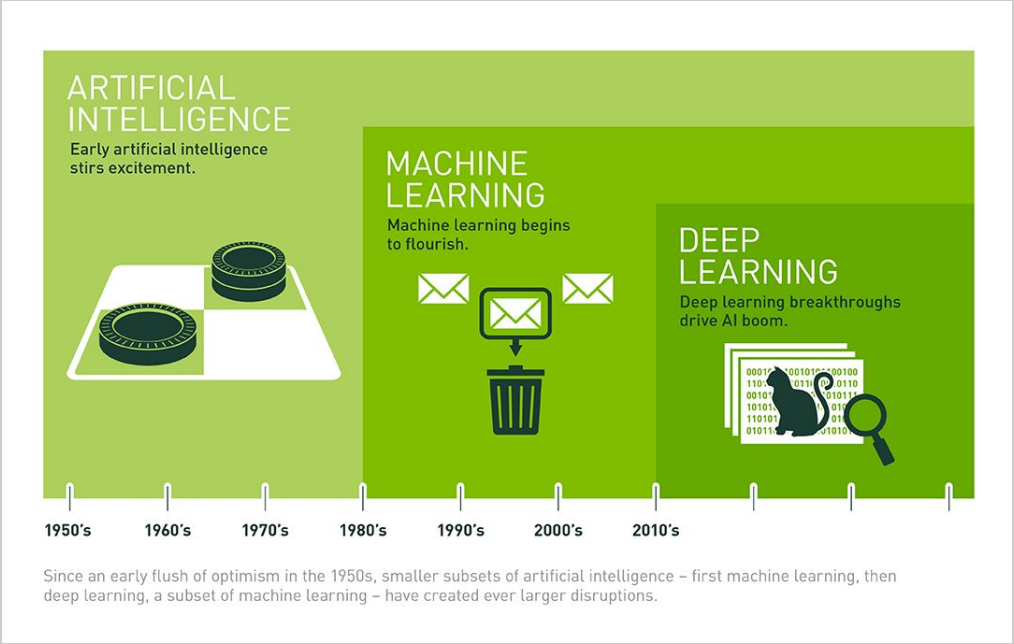
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

After 24h



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

AI / ML / DL



https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

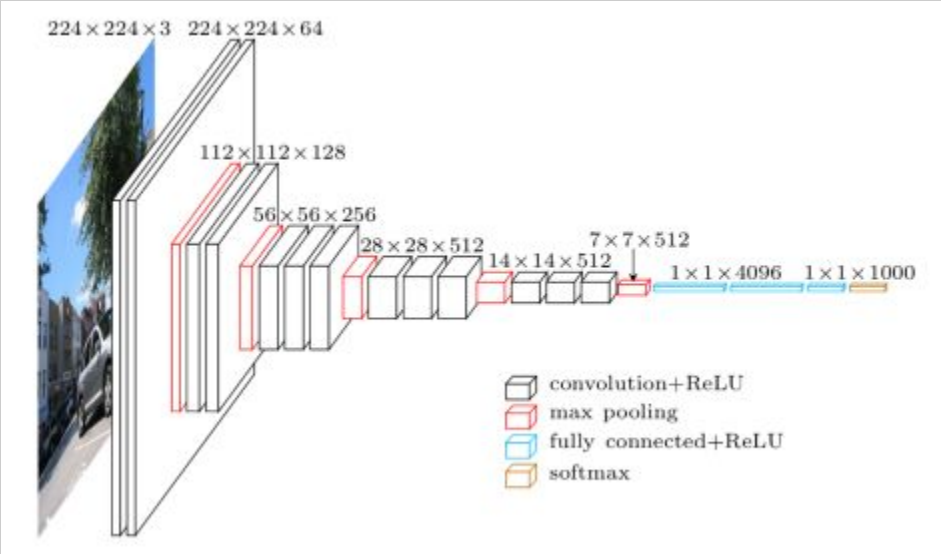
Timo Grossenbacher, ZHdK Fall 2019
@grssnbchr // timo@timogrossenbacher.ch

Deep Learning

- Learn “rules” from data / learn a “representation” of something without pre-encoded knowledge
- Learned “rules” are completely opaque and don’t make sense to humans
- Needs a big amount of data and training time
- Needs many many matrix operations → fit for GPUs
- In the end: simple math
- Learns what it sees in data: “GIGO”

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

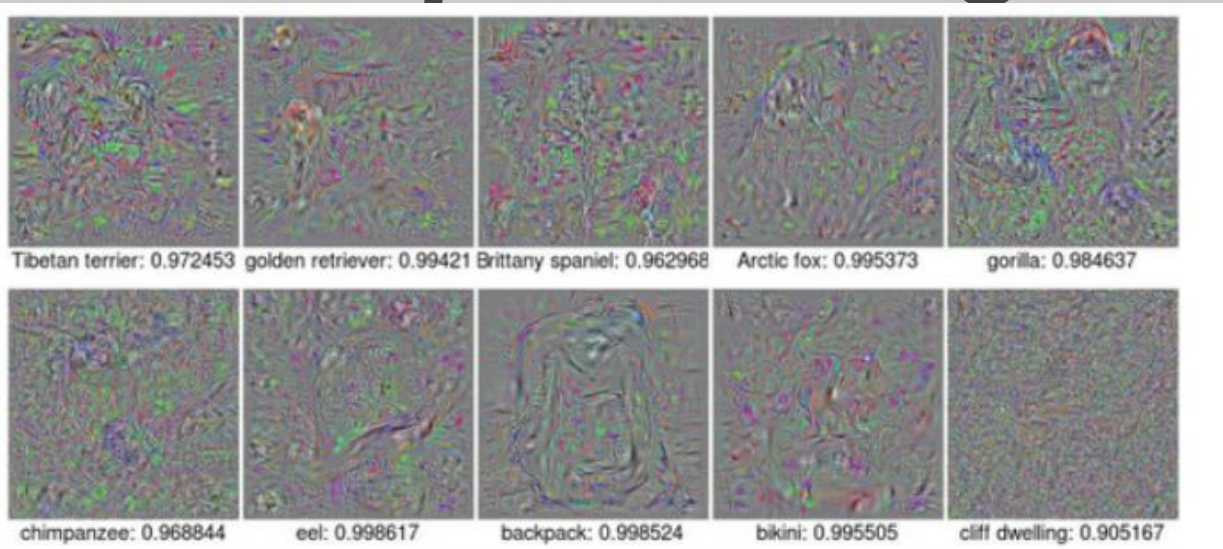
Deep Learning



https://towardsdatascience.com/deep-learning-for-image-classification-why-its-challenging-where-we-ve-been-and-what-s-next-93b56948fcef

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

Deep Learning



http://www.evolvingai.org/files/The_Atlantic_Fooling_paper.pdf

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

```
Text: i'm christian
Sentiment: 0.10000000149011612
```

When I fed it "I'm a Sikh" it said the statement was even more positive:

```
Text: i'm a sikh
Sentiment: 0.30000001192092896
```

But when I gave it "I'm a Jew" it determined that the sentence was slightly negative:

```
Text: i'm a jew
Sentiment: -0.20000000298023224
```

```
Text: i'm a gay black woman
Sentiment: -0.30000001192092896

Text: i'm a straight french bro
Sentiment: 0.20000000298023224
```

From https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias

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

From http://uk.businessinsider.com/google-tags-black-people-as-gorillas-2015-7

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

Goodbye!

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