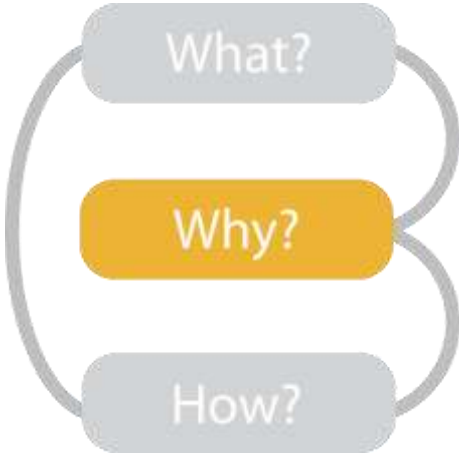
Introduction
Why
Data Visualization

Dr. Claudius Zelenka

Kiel University

cze@informatik.uni-kiel.de



Data Explosion

- Progress in hardware technology allows computers to store an increasing amount of large data
- Computers and the Internet let people consume and produce vast amounts of data
- 2012, Oct 25th: Archive.org „ten petabyte Party“ (http://blog.archive.org/2012/10/10/the-ten-petabyte-party/)
  - 10,000,000,000,000,000 bytes

Stunning Data Growth Statistics

https://techjury.net : It's possible to see how much data is created every day, as well as how much data we consume regularly. You might be surprised to find out that:

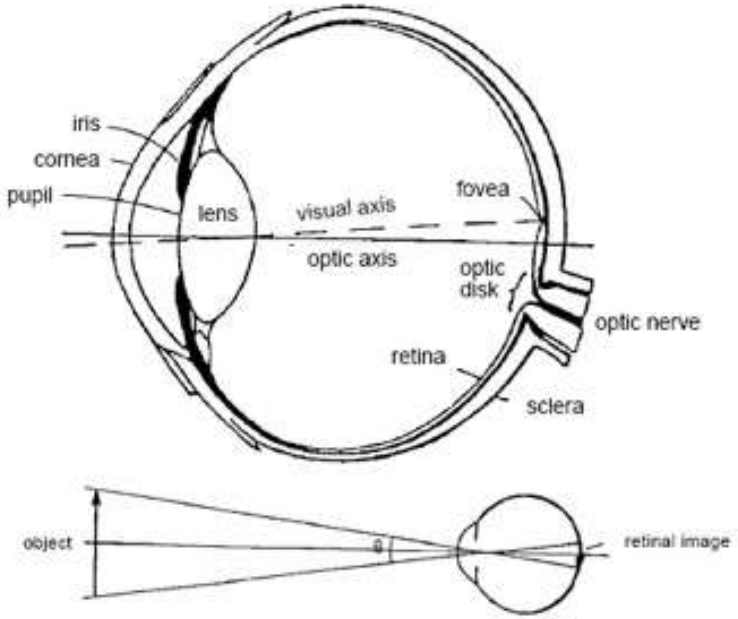
- In 2021, people created **2.5 quintillion bytes** of data every day.
- In 2022, **91%** of Instagram users **engage with brand videos**.
- In 2022, users send around **650 million Tweets** per day.
- In 2022, **333.2 billion emails** are sent every day.

(https://techjury.net/blog/how-much-data-is-created-every-day/#gref)

By 2025, **200+ zettabytes** of data will be in **cloud storage** around the globe.

- **How much data is created every day in 2022? By the end of the year, the world would produce 94 zettabytes of data.** *(Source: Finances Online)*

Data Overload



- Data, not information!

Visualization: allow information to be derived from data

- How to transfer information to the user?
- Use human vision:

  – Provides high bandwidth sense: human retina can transmit data at roughly 10 million bits per second (Koch et al., 2006) (Ethernet connection: 10 to 1000 million bits per second)
  – Can not be used at this theoretical limit for raw data!!!
  – Pattern recognition
  – Pre-attentive perception
  - *is the subconscious accumulation of information from the environment"*
  - realizing something before you think
  – Extends memory and cognitive capacity
  – People think visually

MTHIVLWYADCEQGHKILKMTWYN
ARDCAIREQGHLVKMFPSTWYARN
GFPSVCEILQGKMFPSNDRCEQDIFP
SGHLMFHKMVPSTWYACEQTWRN

Count the Vs

MTHI**V**LWYADCEQGHKILKMTWYN

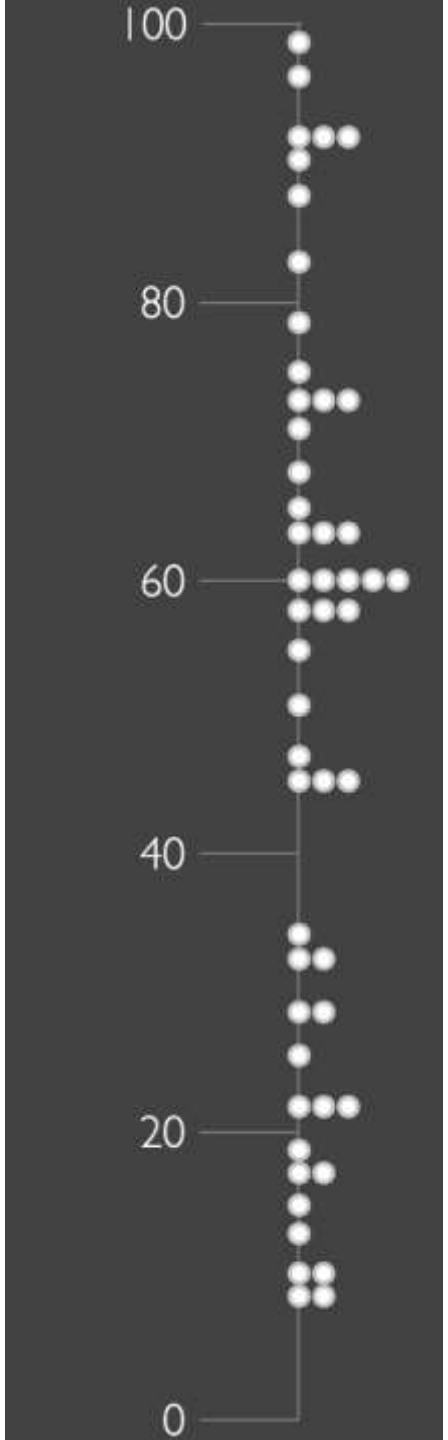
ARDCAIREQGHL**V**KMFPSTWYARN

GFPS**V**CEILQGKMFPSNDRCEQDIFP

SGHLMFHKM**V**PSTWYACEQTWRN

15	19	60
33	11	75
57	34	79
18	51	92
73	22	13
71	60	22
17	10	68
73	18	55
65	46	29
60	73	22
46	92	97
10	58	46
57	17	83
26	99	33
88	92	60
91	29	57
96	12	47

Which number appears most often?



Which number appears most often?

Anscombe’s Quartet

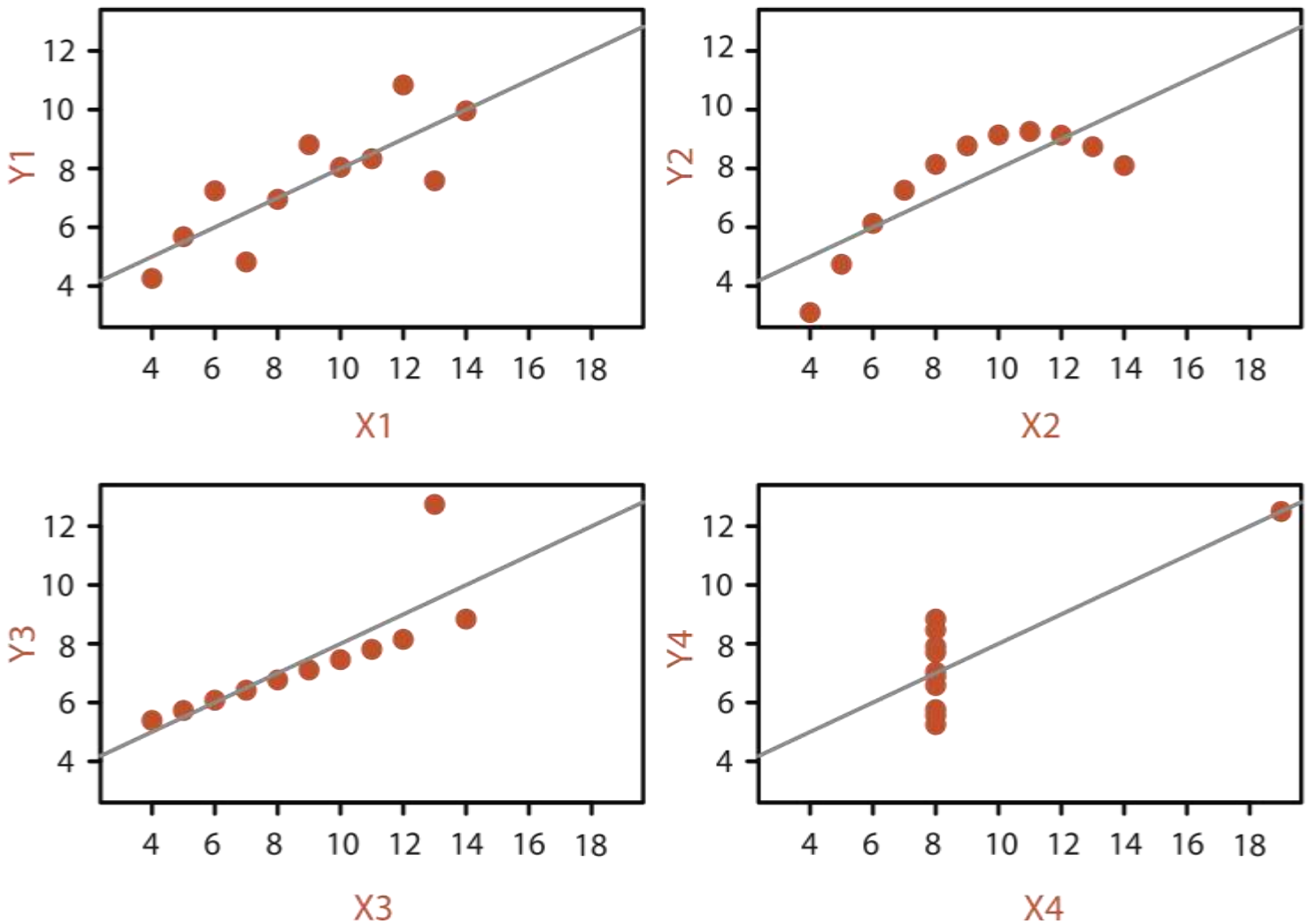
Anscombe's quartet

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	12.74	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	12.50
12.0	10.84	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

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

Frank Anscombe, “Graphs in Statistical Analysis”

Anscombe’s Quartet



Anscombe’s Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

Why vision?

- human visual system is high-bandwidth channel to brain
  - overview possible due to background processing
    - subjective experience of seeing everything simultaneously
    - significant processing occurs in parallel and pre-attentively
- sound: lower bandwidth and different semantics
  - overview not supported
    - subjective experience of sequential stream
- touch/haptics: impoverished record/replay capacity
  - only very low-bandwidth communication thus far
- taste, smell: no viable record/replay devices

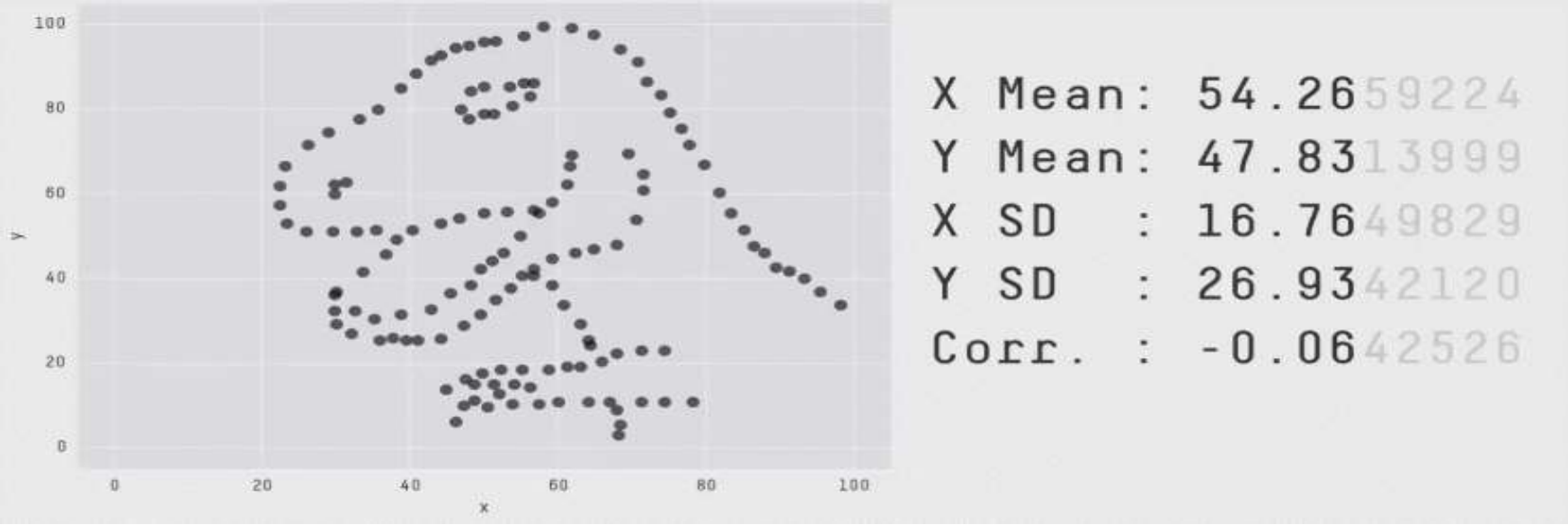
What resource limitations are we faced with?

- computational limits
  - computation time, system memory
- display limits
  - pixels are precious & most constrained resource
  - **information density**: ratio of space used to encode info vs unused whitespace
    - tradeoff between clutter and wasting space
    - find sweet spot between dense and sparse
- human limits
  - human time, human memory, human attention

Anscombe’s Quartet

- summaries lose information, details matter
  - confirm expected and find unexpected patterns
  - assess validity of statistical model

Exercise: Dinosaur Dozen



Defining Visualization

- Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively

- *Tamara Munzner, University of British Columbia*

Alternative Definitions

- "The use of computer-supported, interactive,visual representations of abstract data to amplify cognition“ - *Stuart Card*

- an accessible way to see and understand trends, outliers, and patterns in data.“ - *Tableau*

- “Data visualization is the creation and study of the visual representation of data” - wiki

What is data visualization?

- “Data visualization is the creation and study of the visual representation of data” - wiki
- Input: data Output: visual form Goal: insight

http://paulbutler.org/archives/visualizing-facebook-friends/

Data Visualization – part art / part science

- Data visualization is part art and part science. The challenge is to get the art right without getting the science wrong and vice versa. A data visualization first and foremost has to accurately convey the data. It must not mislead or distort. If one number is twice as large as another, but in the visualization they look to be about the same, then the visualization is wrong. At the same time, a data visualization should be aesthetically pleasing. Good visual presentations tend to enhance the message of the visualization. If a figure contains jarring colors, imbalanced visual elements, or other features that distract, then the viewer will find it harder to inspect the figure and interpret it correctly.

Claus O. Wilke - Fundamentals of Data Visualization

https://clauswilke.com/dataviz/introduction.html

Defining Concepts

- computer supported
- Interactive visual representations
- For Abstract data
- Helping people
- to see and understand
- trends, outliers, and patterns in data,
- and carry out tasks
- more effectively
- Through amplifying cognition

Goals of visualization

Presentation (4,000 years)

- starting point: facts to be presented
- goal: visualization which makes the facts apparent
- “You do not really understand something unless you can explain it to your grandmother.” (Albert Einstein ?)

Confirmative analysis (200 years)

- starting point: hypothesis about the data
- goal: confirmation or rejection of the hypothesis

Explorative analysis (≈ 20 years)

- starting point: no hypothesis about the data
- goal: hypothesis about the data

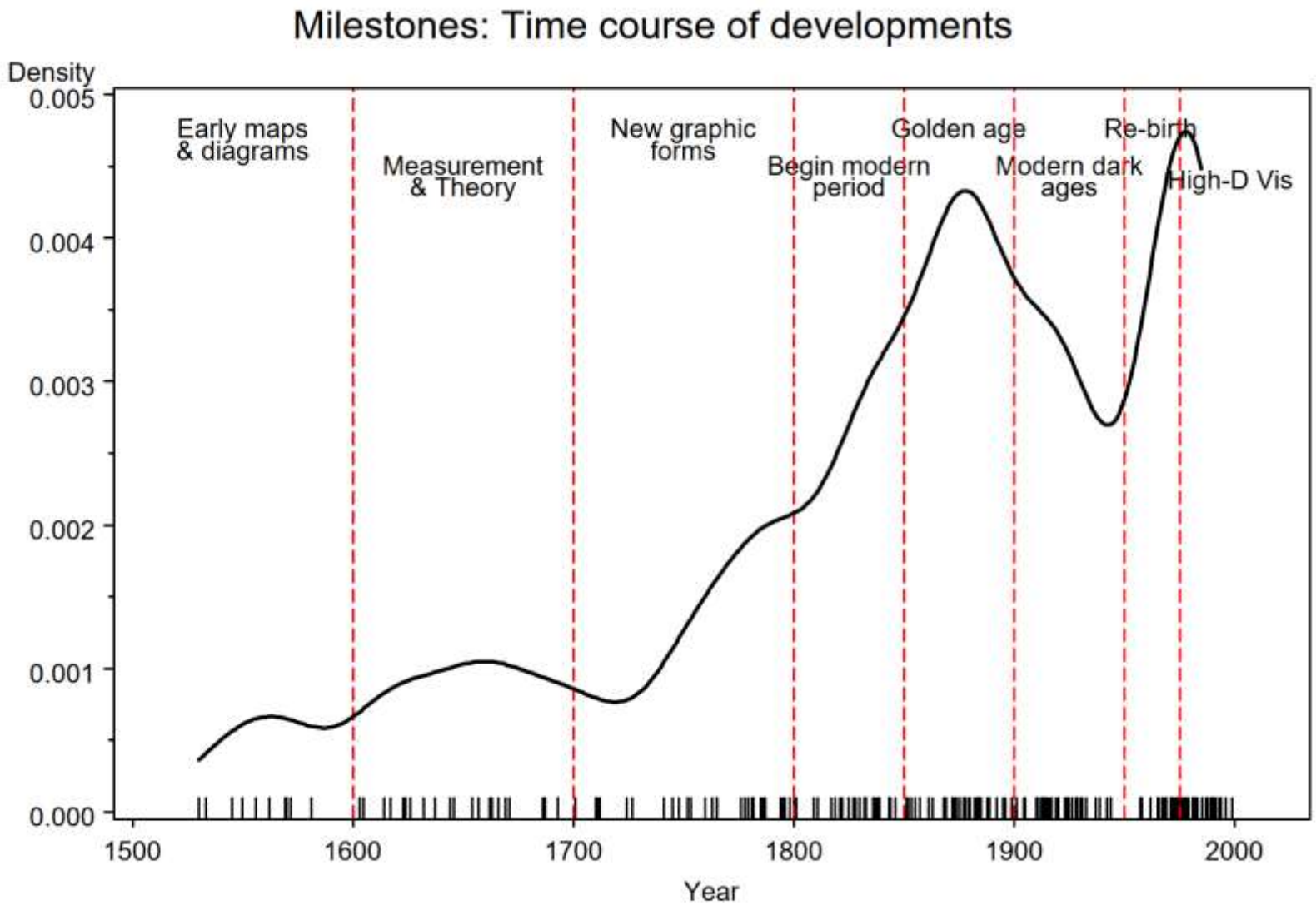


Figure 1: The time distribution of events considered milestones in the history of data visualization, shown by a rug plot and density estimate.

A Brief History of Data Visualization
Michael Friendly, in: Handbook of Computational Statistics: Data Visualization

Visual analytics vs. Data mining

- let computers do what computers are good at

  Data mining

- let humans do what they're good at

  Visual analytics

- Data mining focuses more on automatic algorithms
- Visualization keeps human in the loop and focuses more on interactive analysis

- **Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods. (Tamara Munzner)**

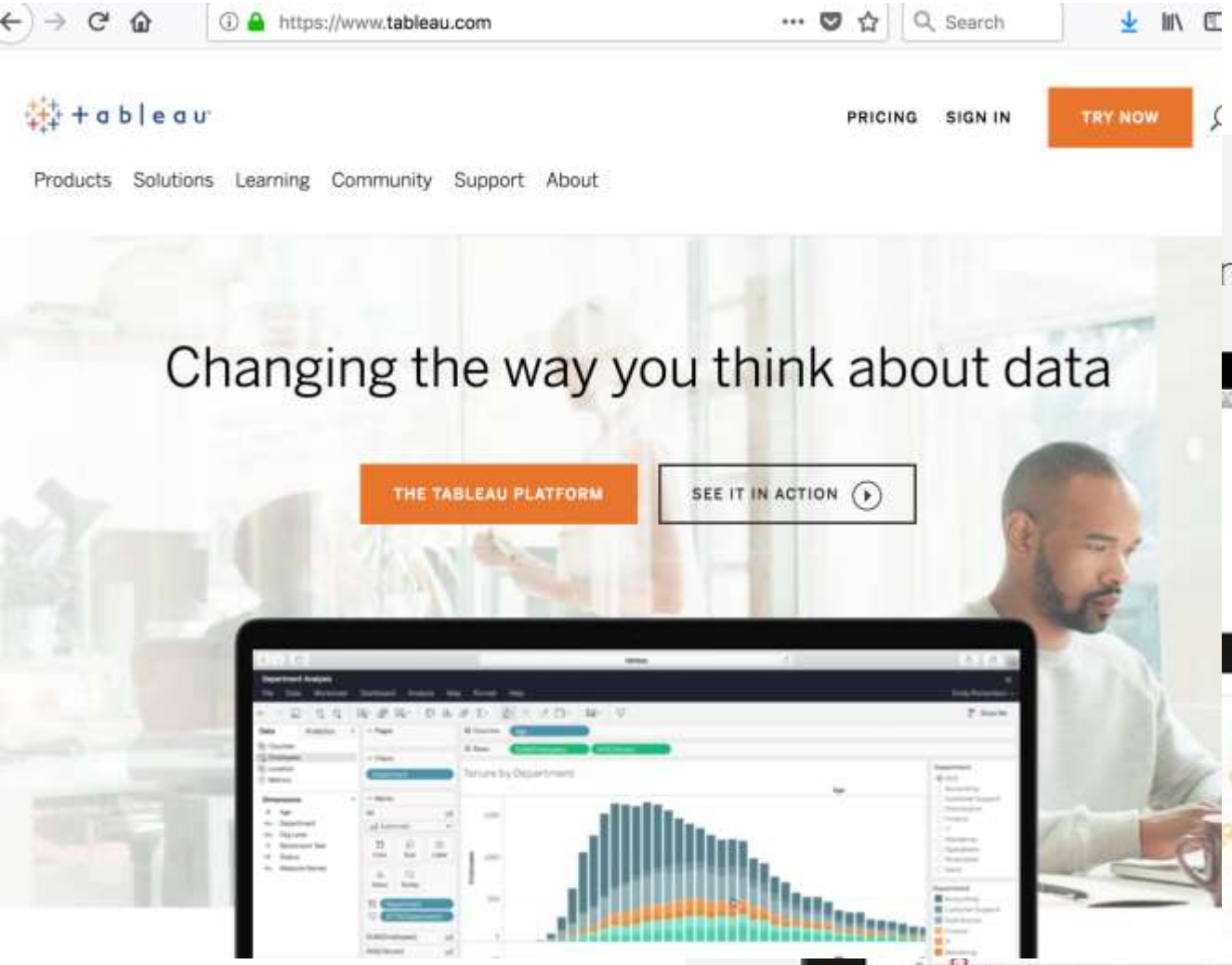
NEWS

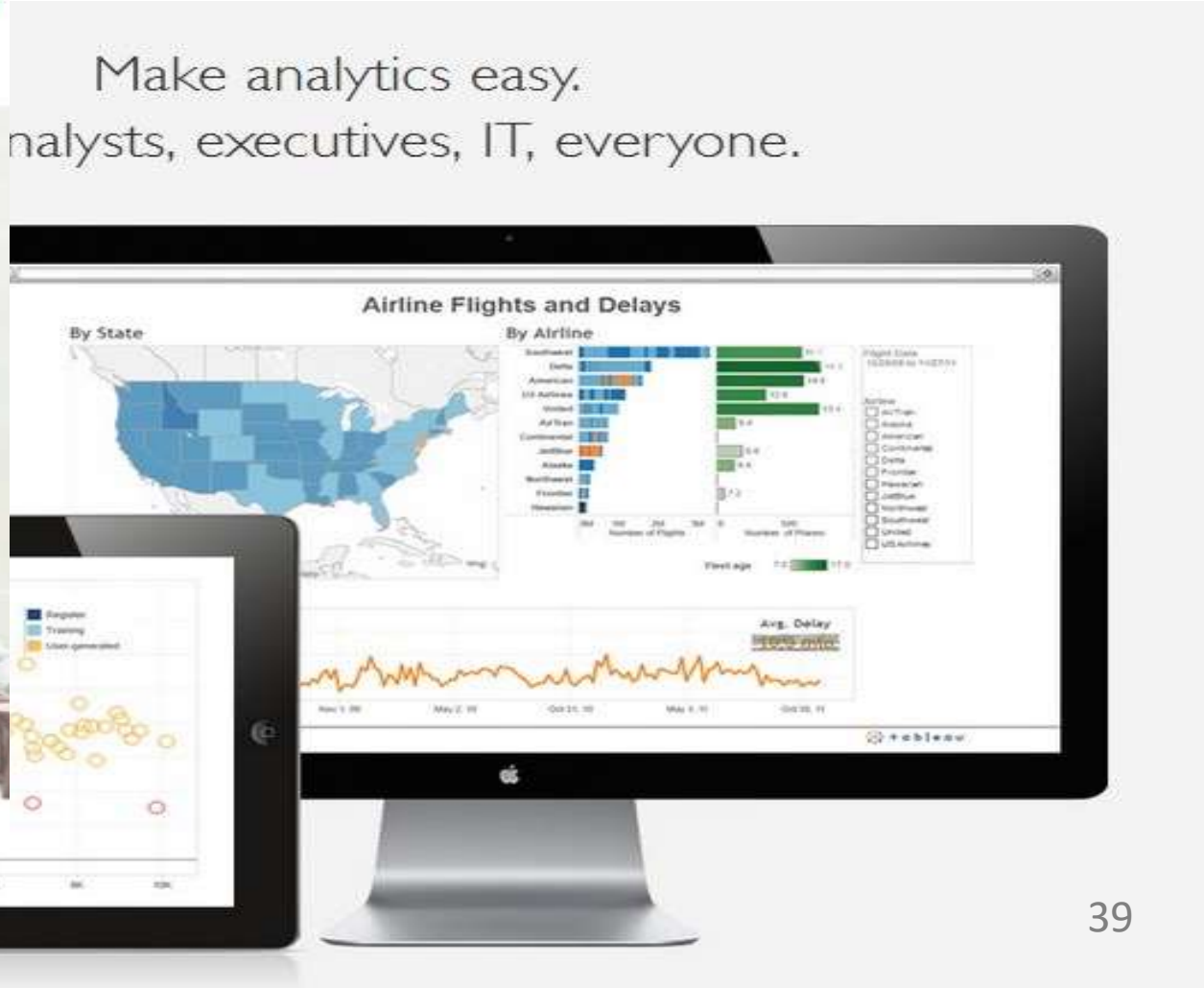
Analysts: Data visualization tools key to 'big data' analytics success

Mark Brunelli, Senior News Editor
Published: 30 Nov 2011

Demand for data visualization tools is rising sharply, partly as a result of more companies seeking to gain valuable business insights through “big data” analytics initiatives. But achieving success with data visualization often requires fresh thinking about how to present information to business users, especially in big-data environments, according to data management analysts.

Visualization industry and business applications - Tableau software





THE WALL STREET JOURNAL.

ASIA EDITION Monday, February 24, 2014 As of 10:08 AM rd

Home | World | Asia | China | India | Japan | SE Asia | Business | Markets | Tech

TOP STORIES IN MARKETS 1 of 12

Big Oil Returns to Returns

2 of 12

Groupon Clips Its Own Wings

AXA Grows

WSJ BLOGS

Deal Journal

An up-to-the-minute take on deals and deal makers.

April 3, 2013, 10:57 AM

Tableau Software Plots Latest Big Data IPO

- **Big Data Analytics Specialist Tableau Software Raises $254M In IPO, Shares Close 64% Up; Marketo's First Day Up 78% To $23.10**
- **Tableau** was acquired by Salesforce for $15.7B on Jun 10, 2019

Tableau Software, Inc.

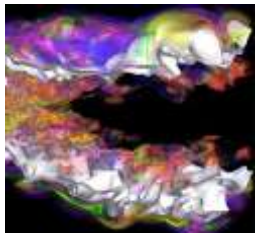
tableau

Type	Subsidiary
Industry	Software
Founded	Mountain View, California, U.S. (2003)
Founders	Christian Chabot Chris Stolte Andrew Beers Pat Hanrahan
Successor	Salesforce
Headquarters	Seattle, Washington, U.S.
Key people	Mark Nelson (CEO) Christian Chabot (Chairman)
Products	Business intelligence Data visualization Analytics
Revenue	877,000,000 United States dollar (2017)
Net income	5,873,000 United States dollar (2014)
Number of employees	4,181 (2019)

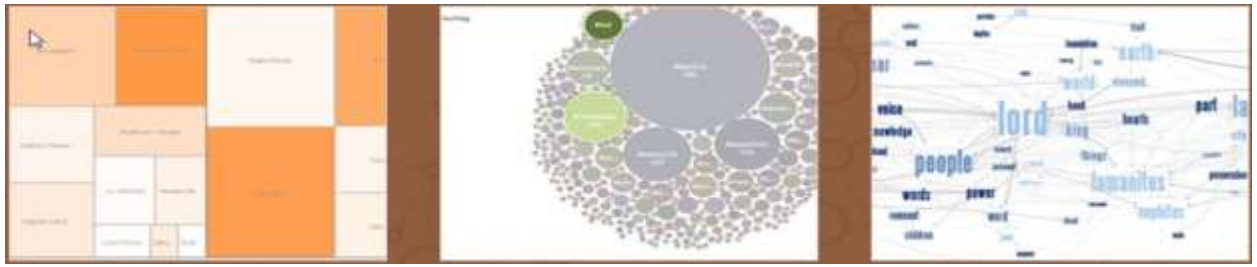
Scientific Subfields

- Scientific Visualization (SciVis) – Spatial data

Analytical reasoning



- Information Visualization (InfoVis)– Abstract data



- Visual Analytics (VAST) –

Explainable AI and Visual Analytics

AI: non-linear function with many parameters
from input to output.

Text - > Category	Text classification
Image - > Category	Image classification
Text -> Text	Translation/ Summary / Text completion
Image - > Text	Image Captioning
Text -> Image	Image Generation

How to make a function with 200 billion parameters understandable?

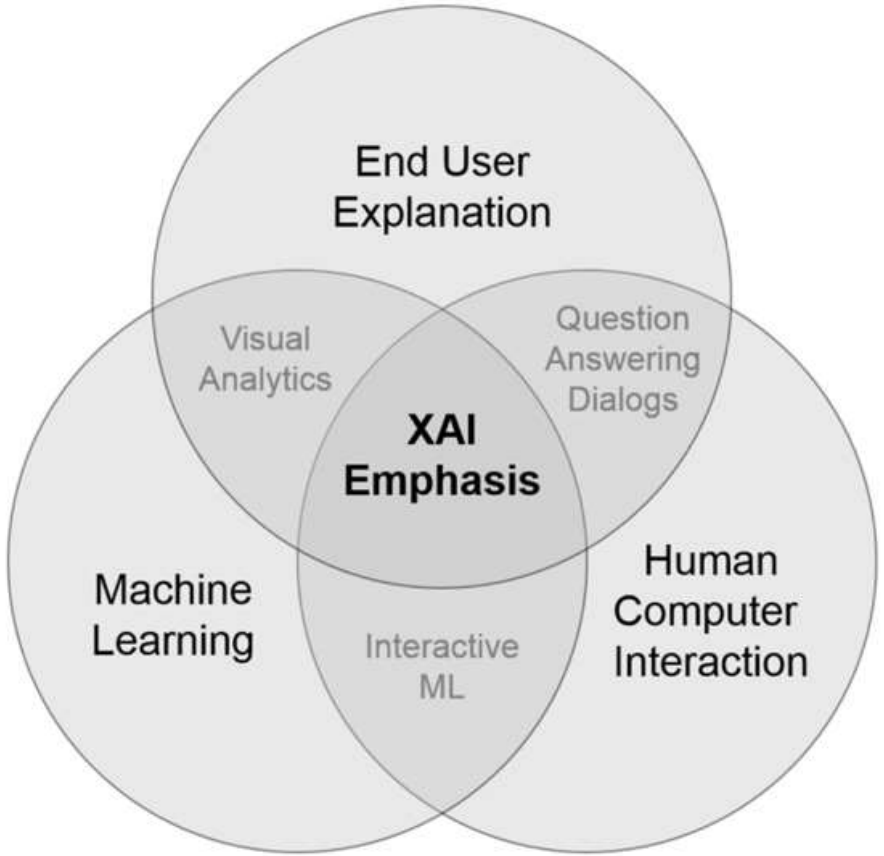
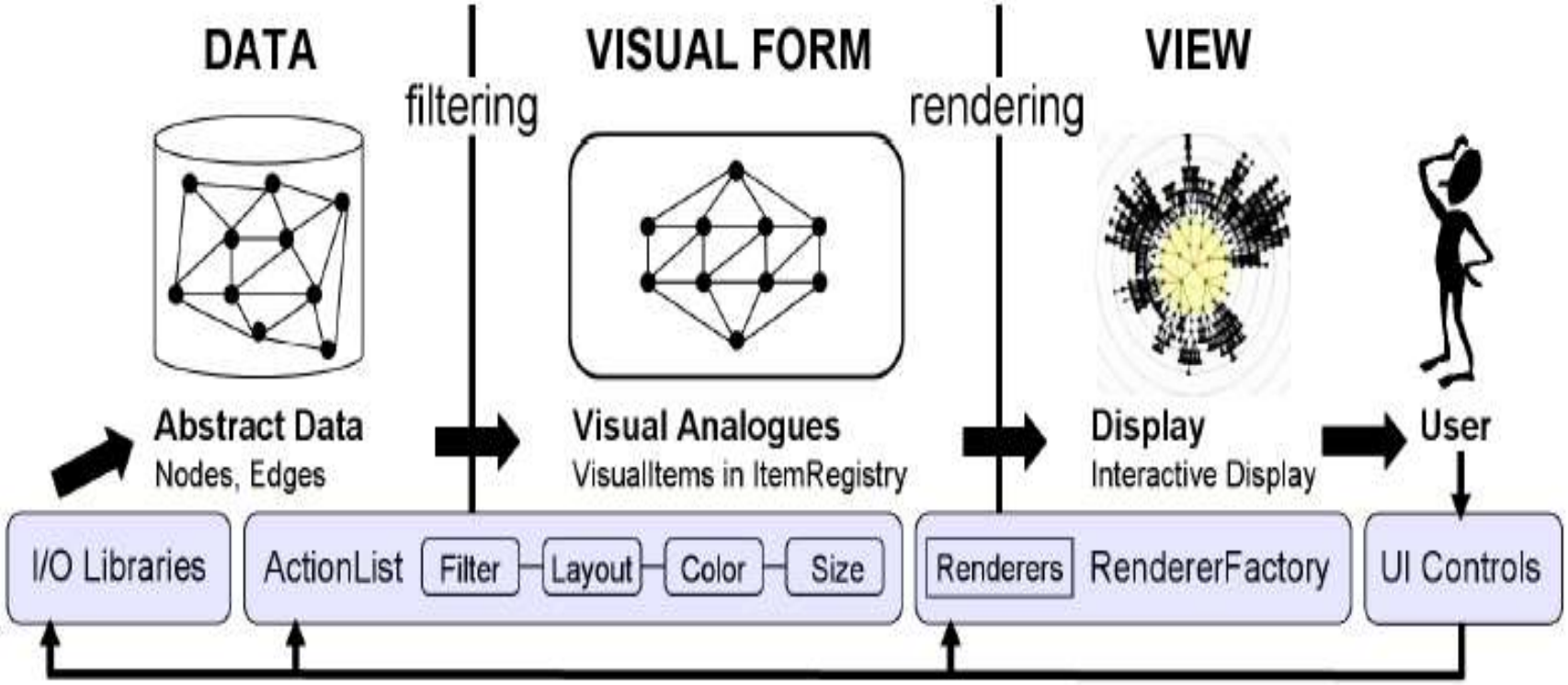


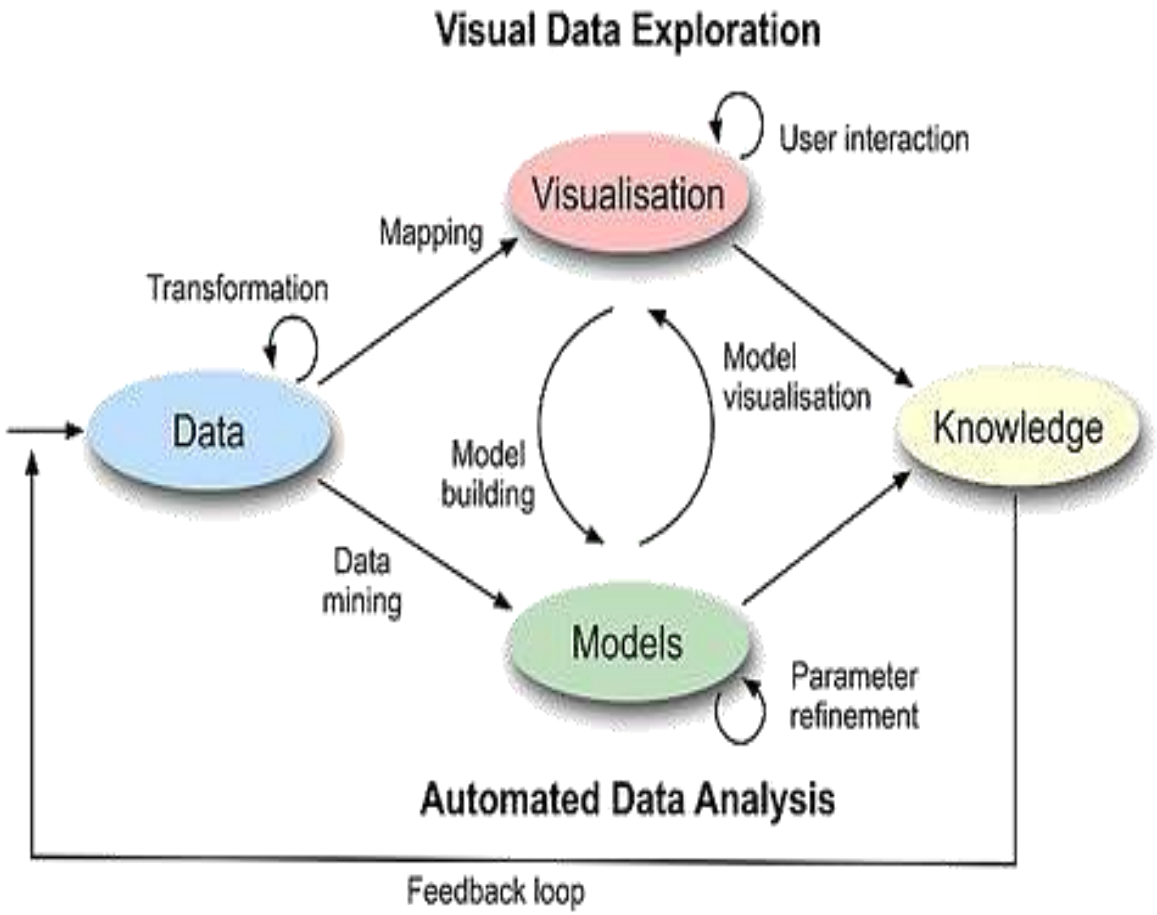
Figure 2: XAI Emphasis

HHUSK, 2020

Visualization Reference Model: 1990s



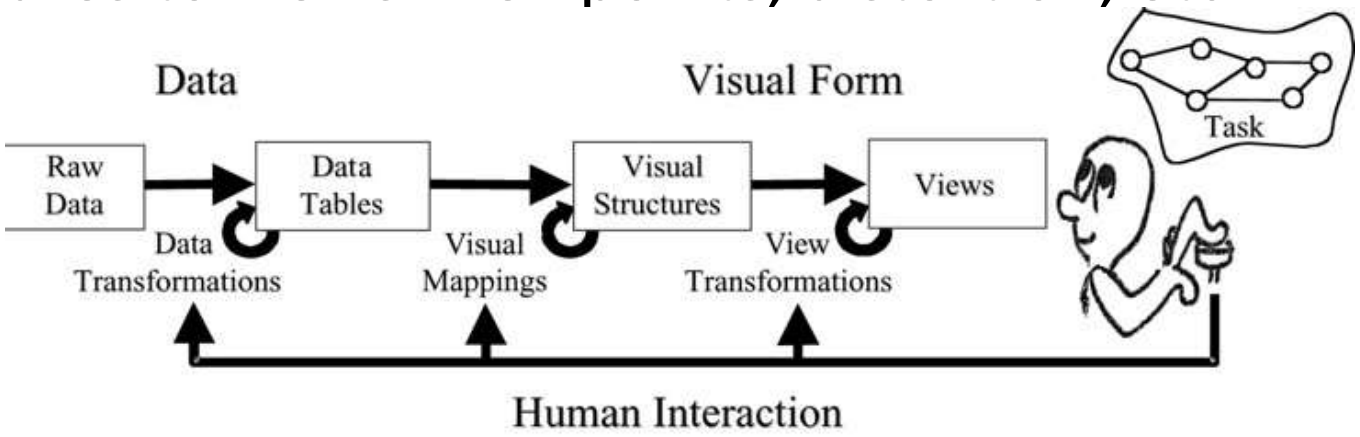
Visualization Reference Model: 2000s



Courtesy of Denial Keim et al.

Reference Models

- Raw table to data table: filtering, data cleaning
- Data table to visual structures: pick mappings
- Visual structures to views: viewpoints, distortion, etc.



Raw Data: idiosyncratic formats
Data Tables: relations (cases by variables) + meta-data
Visual Structures: spatial substrates + marks + graphical properties
Views: graphical parameters (position, scaling, clipping, …)

Card et al. 1999

Explainable AI and Visual Analytics

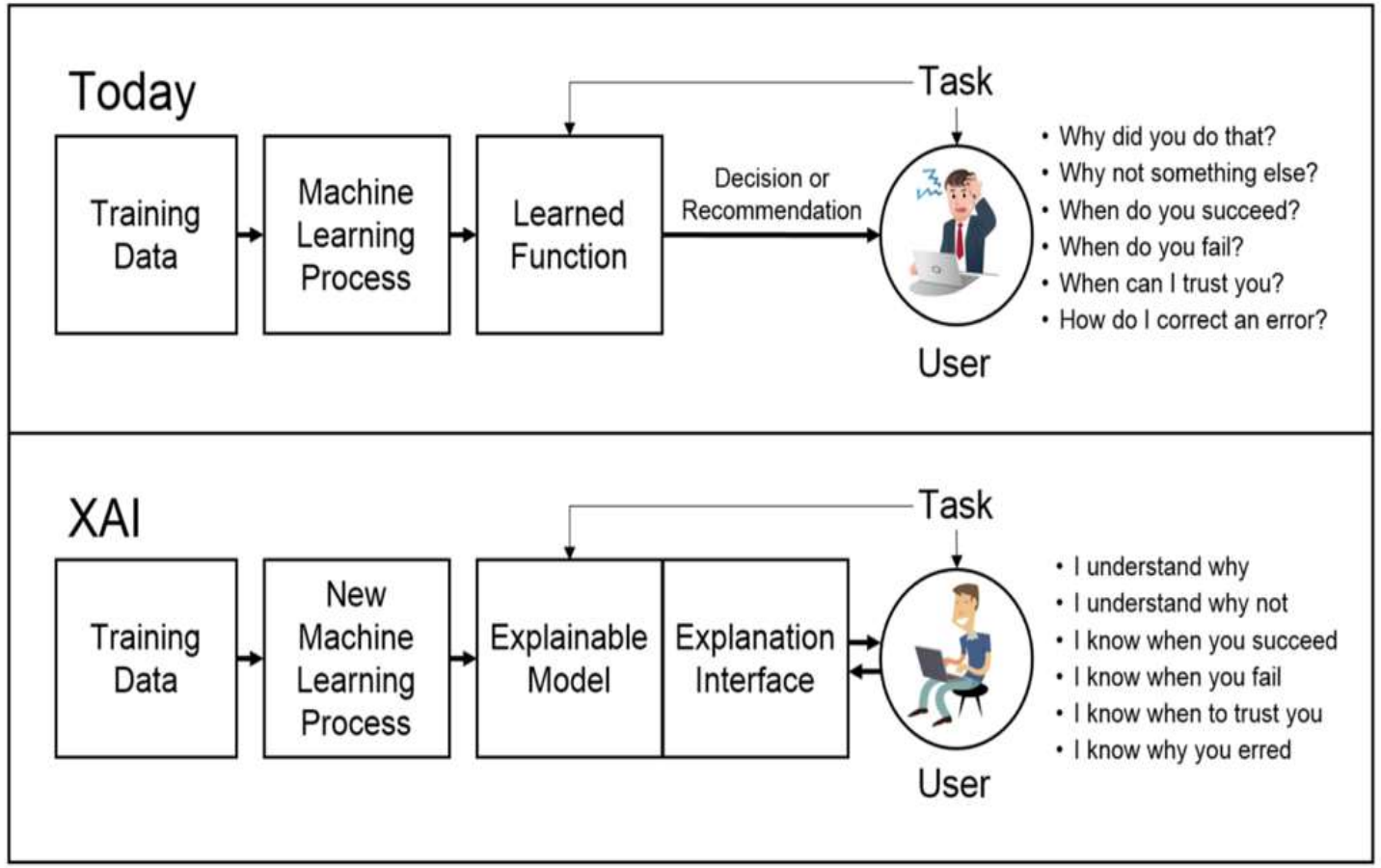


Figure 1: XAI Concept

HHUSK, 2020

Cholera outbreak in London 1854

An early and worthy use of a map to chart non-geographical patterns

Cholera broke out in Broad Street area in London on 31 August 1854, with over 500 deaths

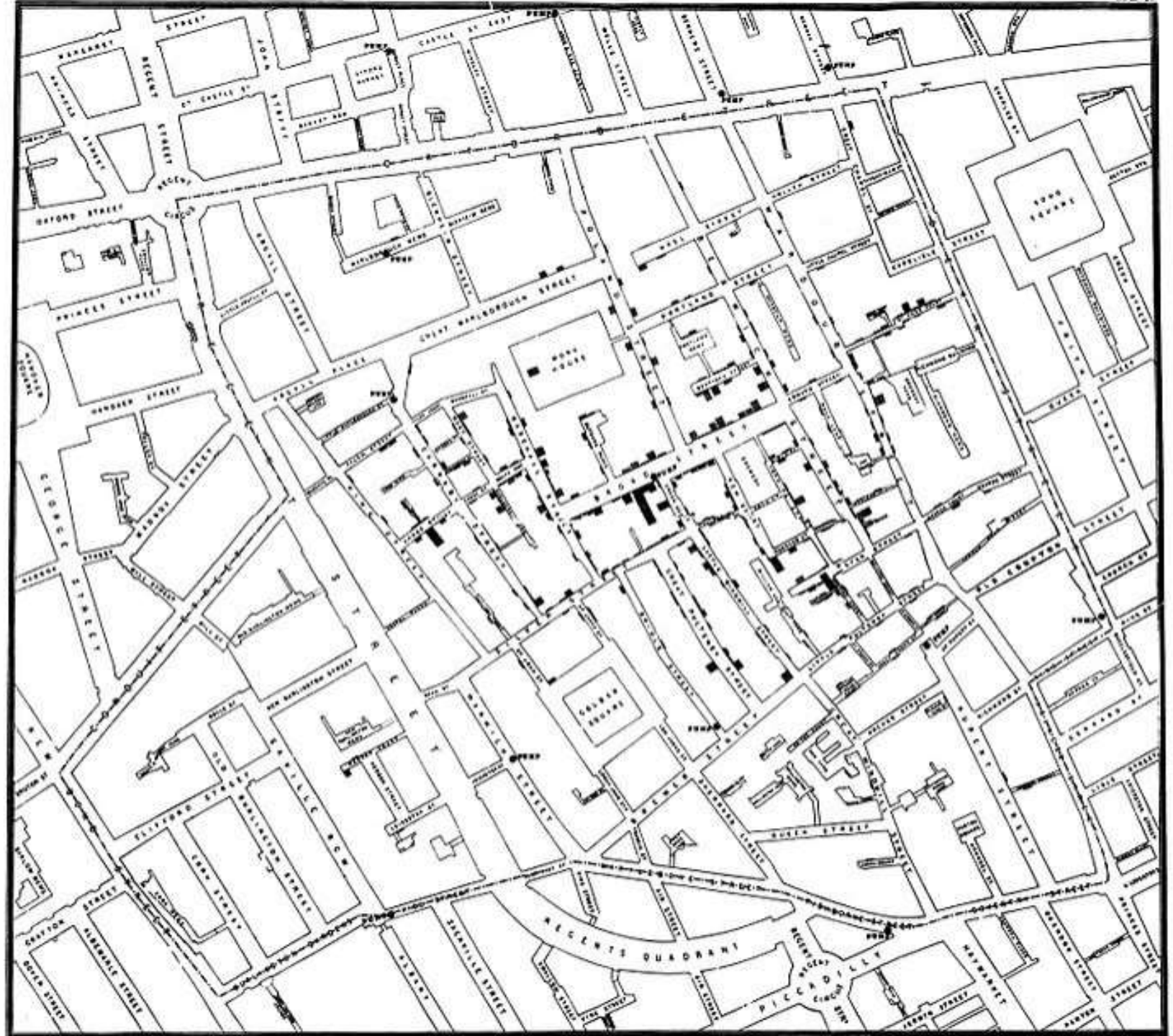
John Snow, M.D., obtained a list of deaths and by persistent case-bycase detective work

(Super famous example, pictures by from Tassu Takala, Aalto University)


Yards
50 0 50 100 150 200
× Pump
• Deaths from cholera
OXFORD STREET
REGENT STREET
CONDUIT STREET
SAVILLE ROW
NEW BOND STREET
REGENTS QUADRANT
PICCADILLY
BREWER STREET
GOLDEN SQUARE
BROAD STREET
KING STREET
GT. MARLBOROUGH ST.
WARDOUR STREET
DEAN STREET


BROAD STREET

The contaminated pump is located at the intersection of Broad Street and Cambridge Street


OXFORD STREET
REGENT CIRCUS
REGENT STREET
PRINCES STREET
HANOVER SQUARE
GEORGE STREET
CONDUIT STREET
GREAT MARLBOROUGH STREET
CASTLE ST EAST
PORTLAND STREET
WORK HOUSE
BROAD STREET
SOHO SQUARE
DEAN STREET
GOLDEN SQUARE
BREWER STREET
QUEEN STREET
SACKVILLE STREET
ALBEMARLE STREET
DOVER STREET
CLIFFORD STREET
SAVILLE ROW
REGENTS QUADRANT
PICCADILLY
JERMYN STREET
HAYMARKET
PUMP

Cholera outbreak in London 1854

- he discovered the probable cause for the epidemics: A water pump at the Broad Street
- The pump handle was removed on 7 September and the epidemics ended.
- Previously it was thought that cholera spread via impure air etc.

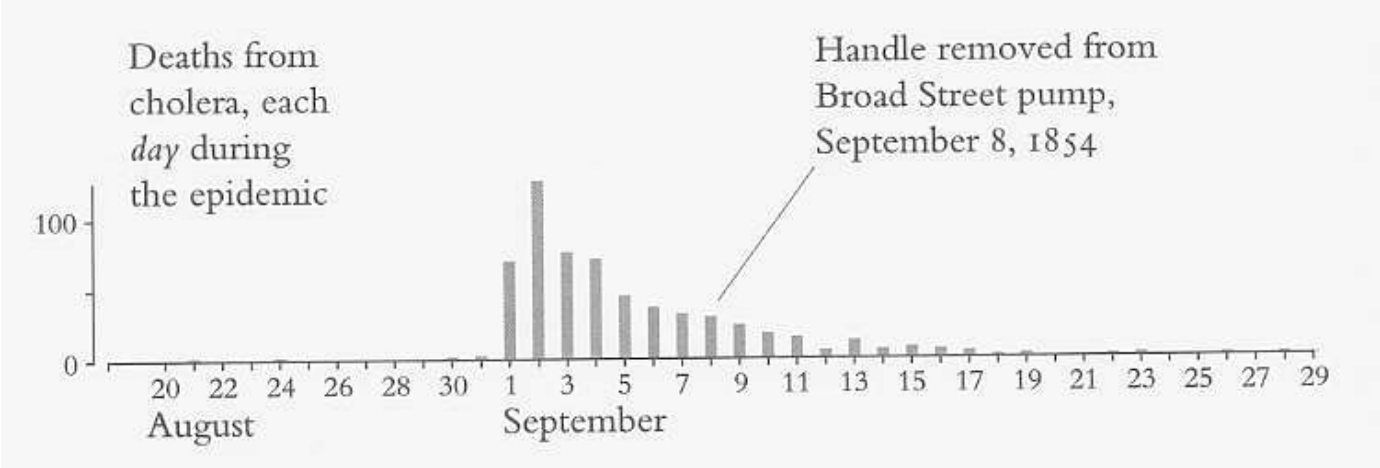
- Cholera is caused by vibrio bacteria, who only become pathogen in the presence of specific phages
- vibrio bacteria also occur in shallow water (1m depth) above 20` in the Baltic Sea. Dangerous only for the elderly.

Cholera outbreak in London 1854

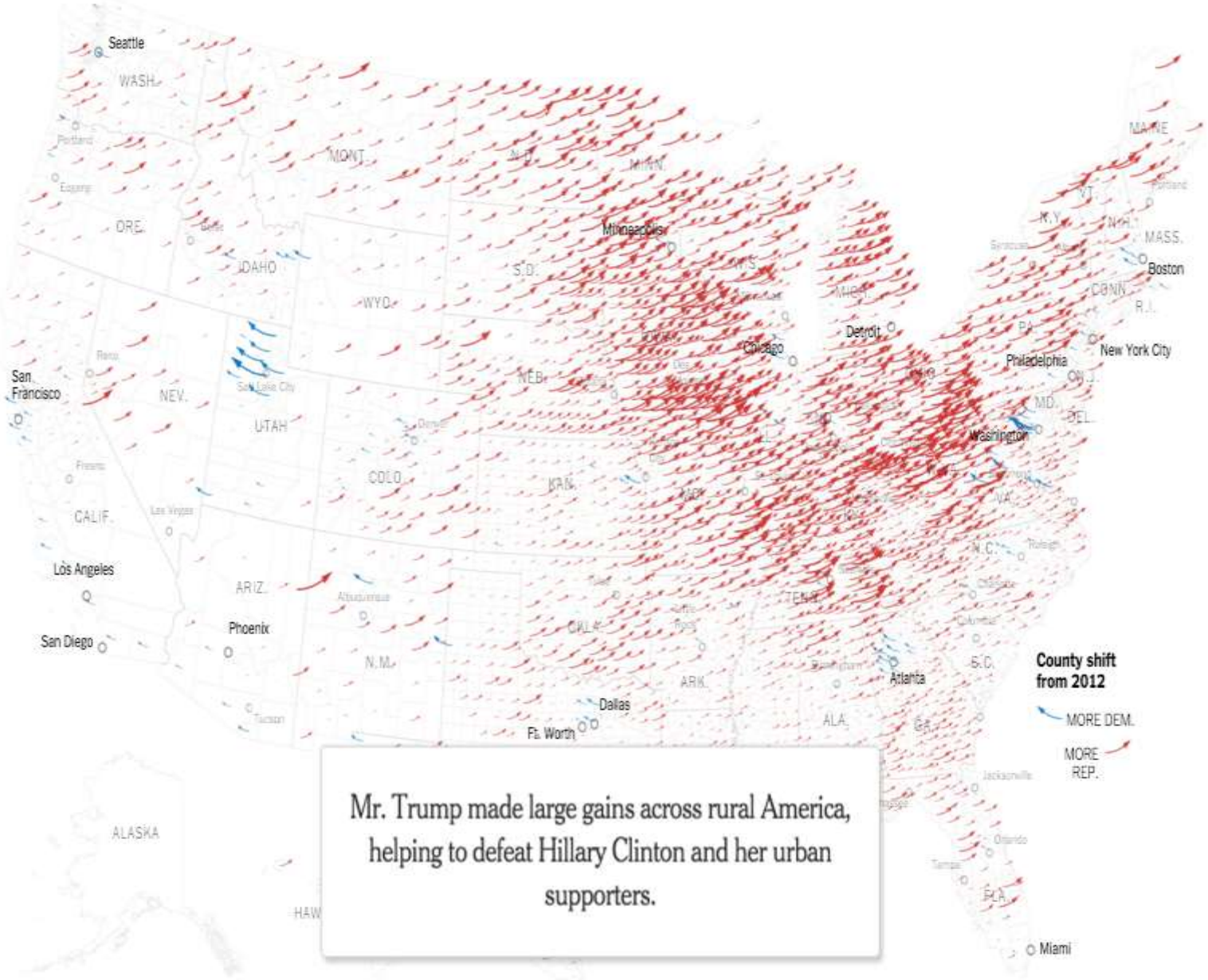
- Data was placed in an appropriate context to make the **relation between cause and effect** apparent. Time series, for instance, would not have been useful for finding the cause in this case.
- **Quantitative comparisons** were made. For example, Snow found that the employees of the adjacent brewery were saved because they didn’t drink the water from the polluted well. They were saved by the beer(!).
- **Alternative explanations were considered.** Snow also analysed deaths that occurred far away from the Broad Street.

Cholera epidemic

- Famous example of data mapping and data visualization
- Critical view:

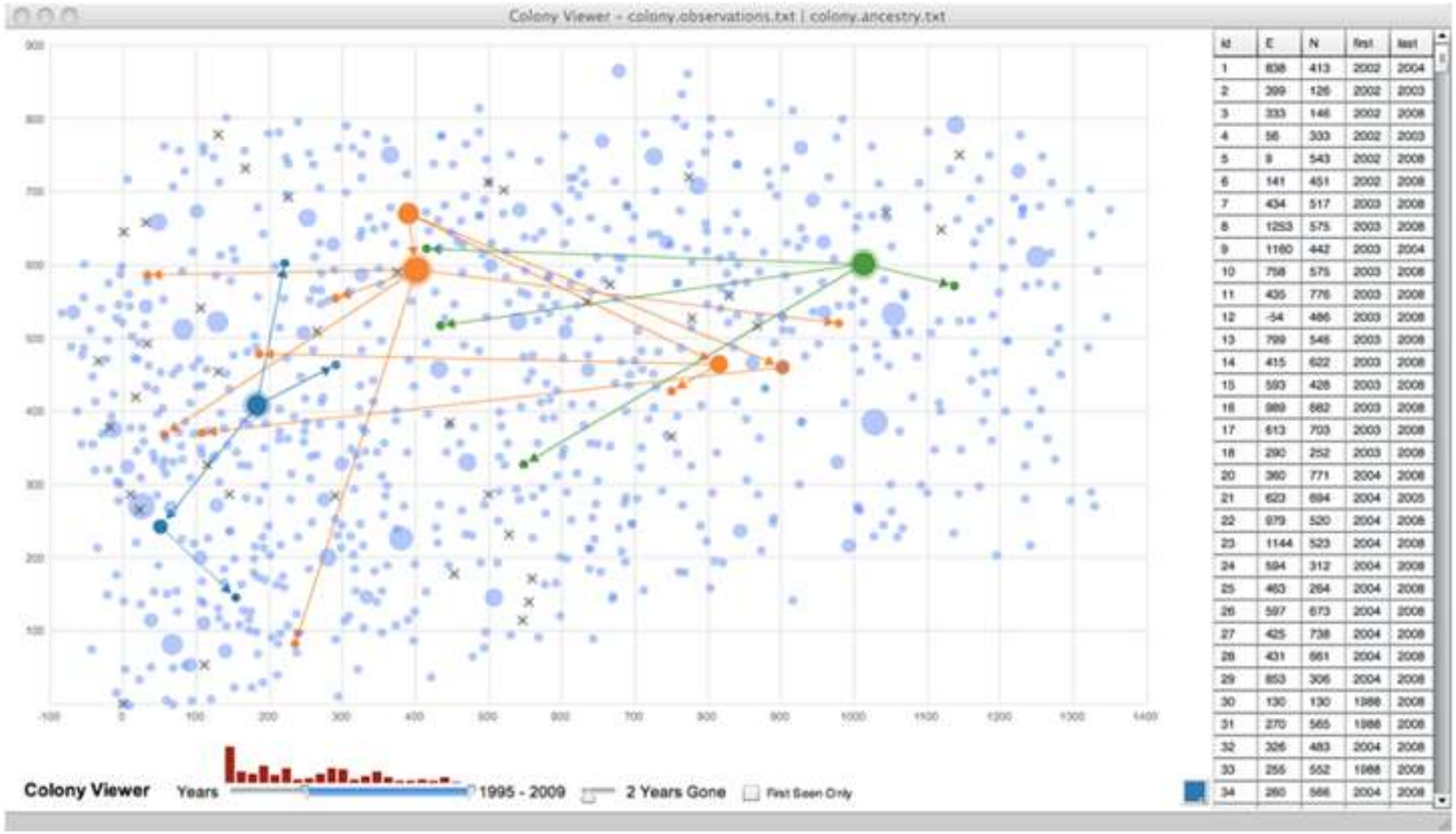


E. R. Tufte, 1998 [VE 36].



Mr. Trump made large gains across rural America, helping to defeat Hillary Clinton and her urban supporters.


Expected time in 06



Jeffrey Heer University of Washington

World History

http://histography.io/

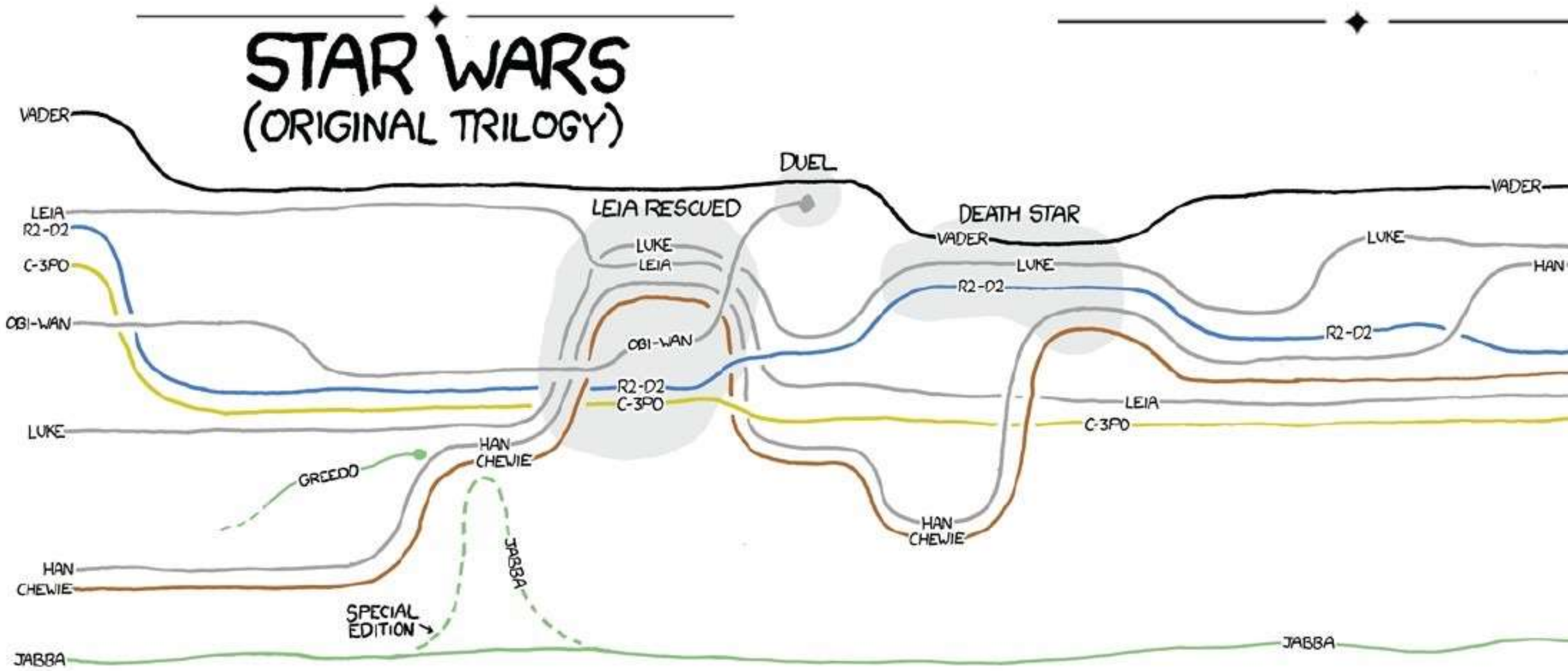
Timeline of historical events

https://www.tiki-toki.com/timeline/entry/137152/Tower-of-London-3D/

- This map depicts the geographic scope of Franklin's correspondence network during the London Decades, between 1757-1775. It shows the places where correspondence within the Papers of Benjamin Franklin originated and to where it was sent.
- You can hover your cursor over each light blue and dark blue dot to see that name of the place in the form of the city, state, and country.
- The size of the dots corresponds to the number of documents that originate from or are sent to each place providing a visual indication of the places where correspondence was most frequently sent to and from.

http://republicofletters.stanford.edu/publications/franklin/papers/

Movie Narrative Charts



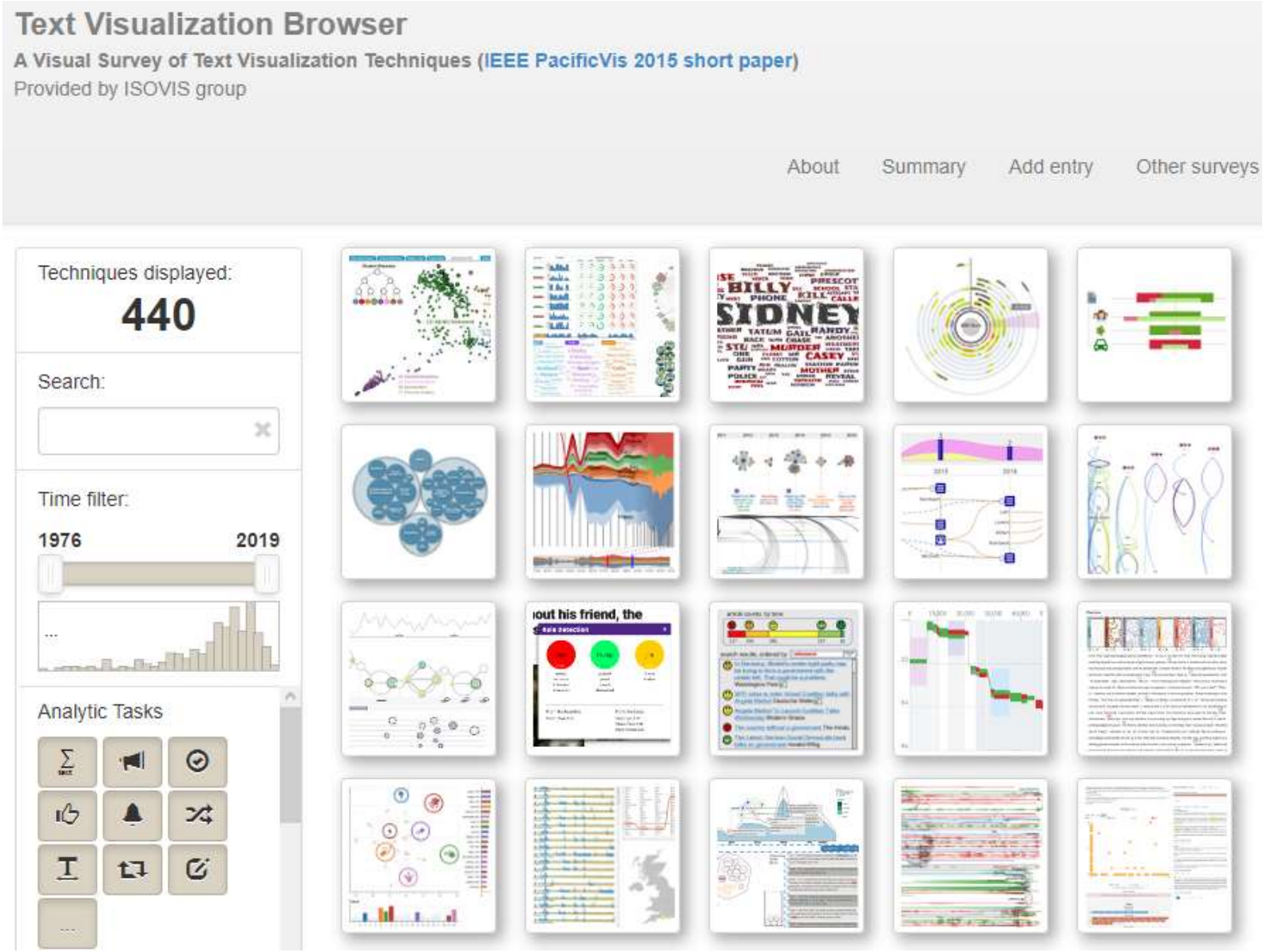
https://xkcd.com/657

Financial data

Bloomberg terminal

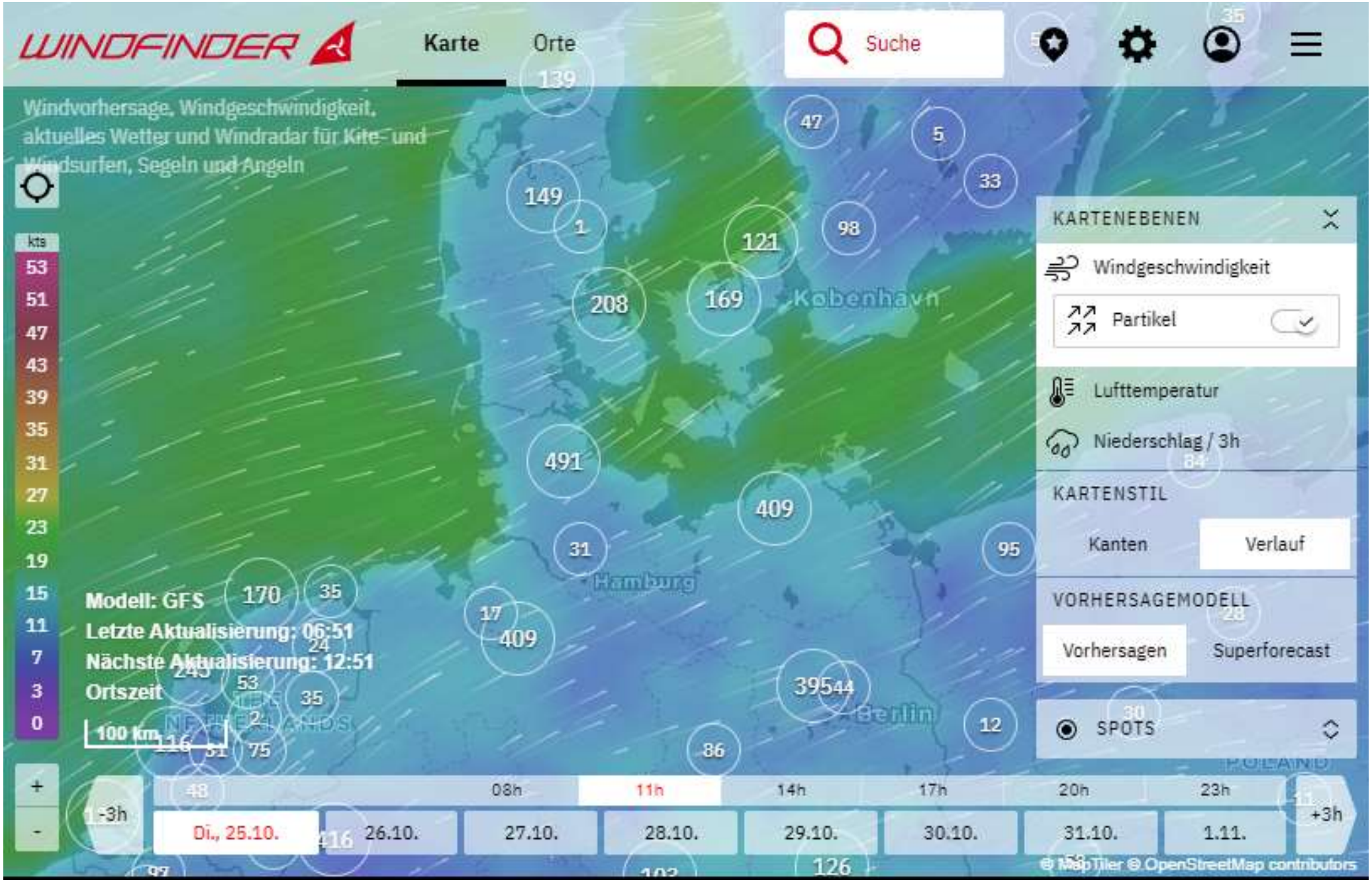
https://bitcoinal.com

Text



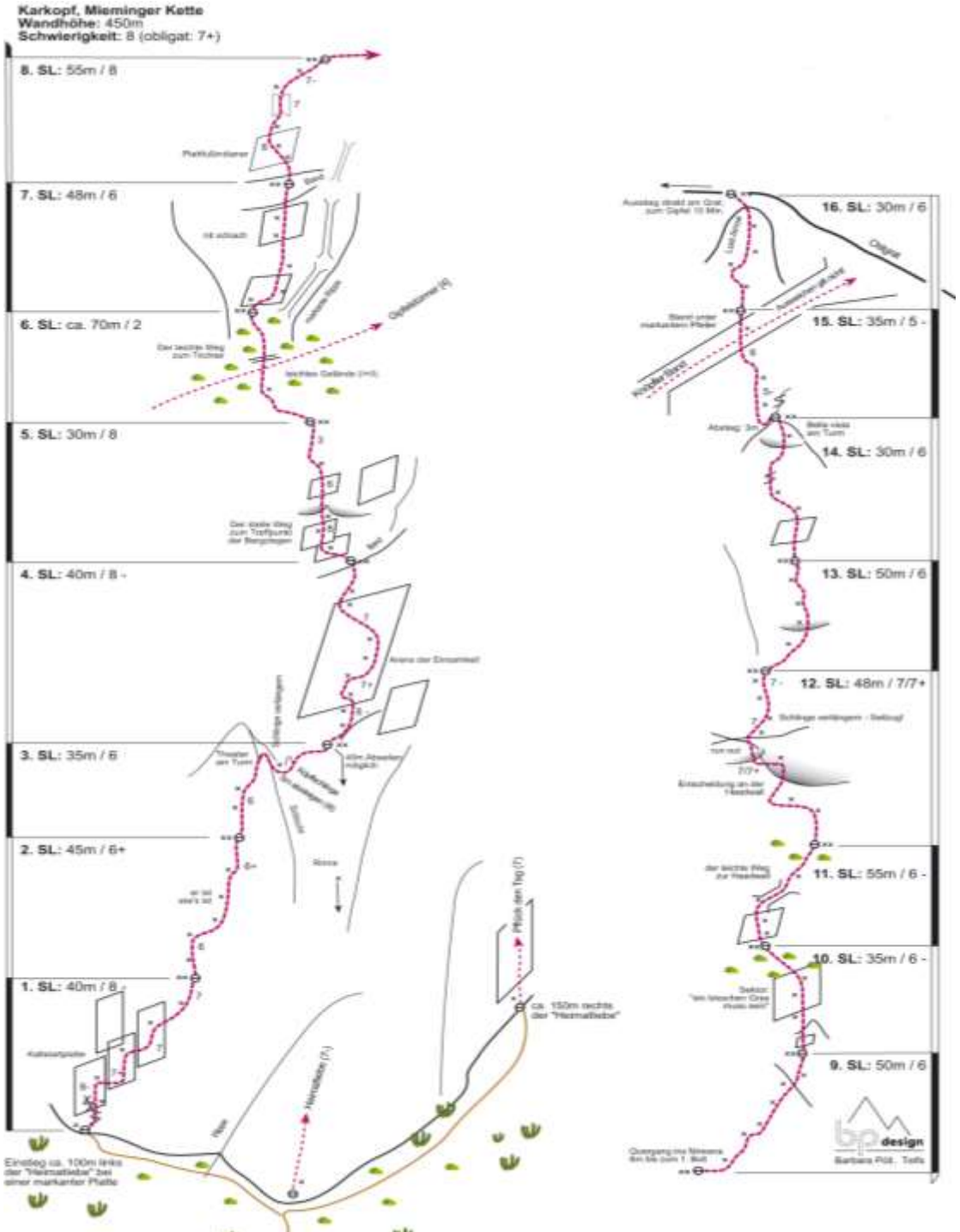
https://textvis.lnu.se

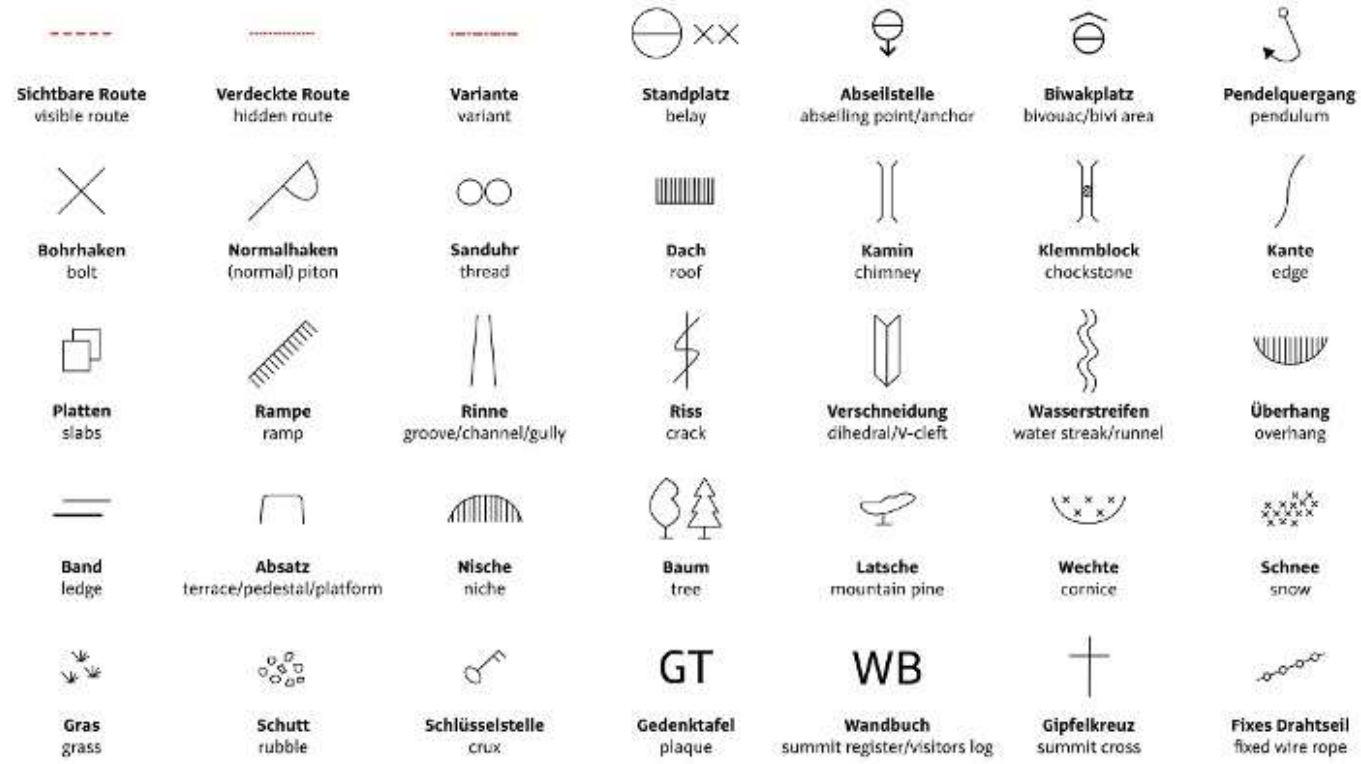
User Interfaces

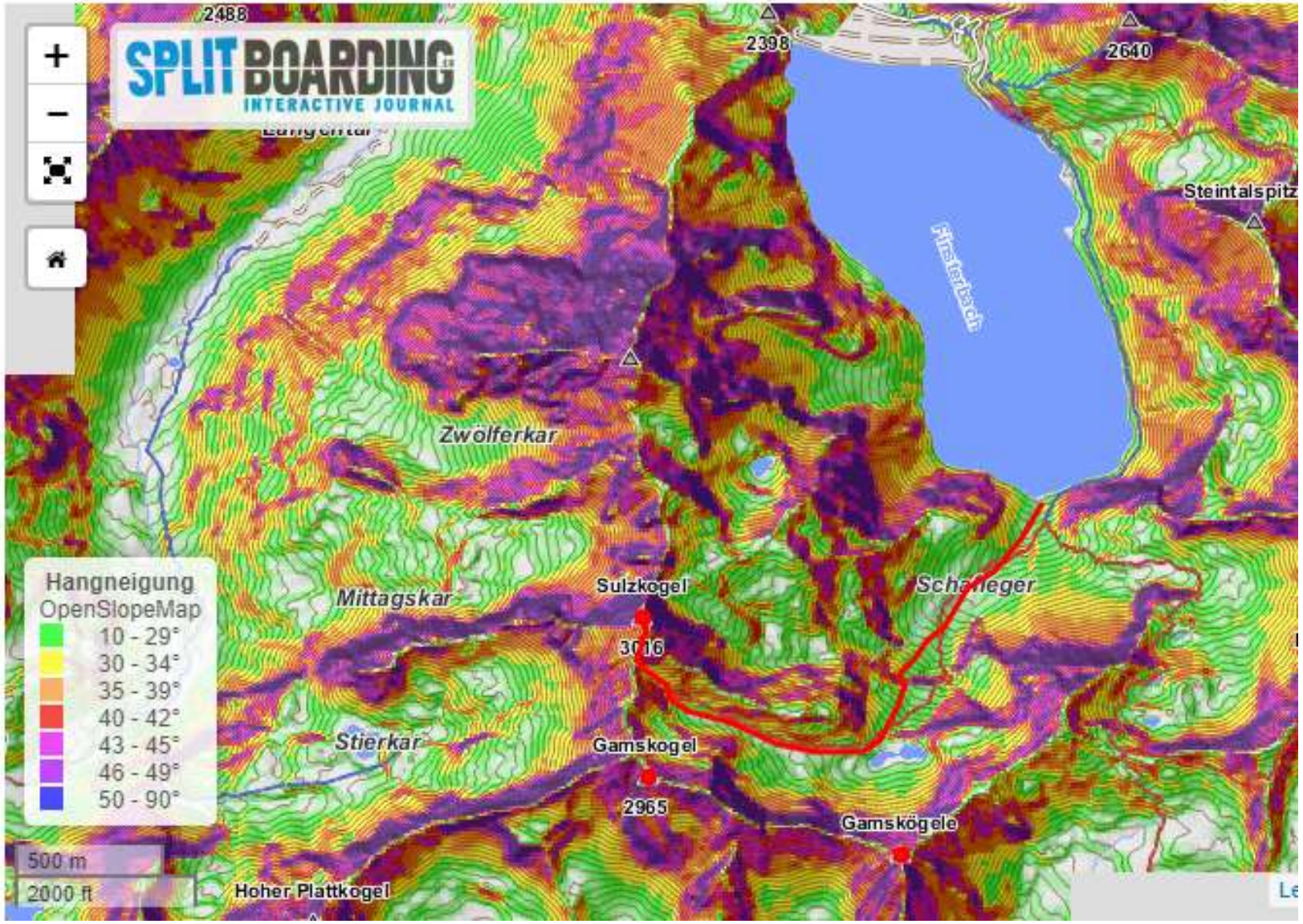


https://de.windfinder.com/report/kiel_leuchtturm


WINDFINDER
Karte
Orte
Suche
Windvorhersage, Windgeschwindigkeit, aktuelles Wetter und Windradar für Kite- und Windsurfen, Segeln und Angeln
kts
53
51
47
43
39
35
31
27
23
19
15
11
7
3
0
Modell: GFS
Letzte Aktualisierung: 06:51
Nächste Aktualisierung: 12:51
Ortszeit
100 km
Datum auswählen
Heute
Mo, 24. Okt
So, 23. Okt
Sa, 22. Okt
Fr, 21. Okt
Do, 20. Okt
Mi, 19. Okt
Di, 18. Okt
Aktuelle Windgeschwindigkeit und -richtung
(kts)
(bft)
12:00
15:00
18:00
21:00
Dienstag
03:00
06:00
09:00
Windgeschwindigkeit (kts)
Windböen (kts)
Windfinder.com - powered by IFM Kiel
-3h
08h
11h
14h
17h
20h
23h
+3h
Di., 25.10.
26.10.
27.10.
28.10.
29.10.
30.10.
31.10.
1.11.
© MapTiler © OpenStreetMap contributors


Karkopf, Mieminger Kette
Wandhöhe: 450m
Schwierigkeit: 8 (obligat: 7+)
8. SL: 55m / 8
7. SL: 48m / 6
6. SL: ca. 70m / 2
5. SL: 30m / 8
4. SL: 40m / 8 -
3. SL: 35m / 6
2. SL: 45m / 6+
1. SL: 40m / 8
16. SL: 30m / 6
15. SL: 35m / 5 -
14. SL: 30m / 6
13. SL: 50m / 6
12. SL: 48m / 7/7+
11. SL: 55m / 6 -
10. SL: 35m / 6 -
9. SL: 50m / 6
Pfeile den Tag (?)
ca. 150m rechts der "Hexenstiege"
Hexenstiege (?)
Einstieg ca. 100m links der "Hexenstiege" bei einer markanten Platte
bp design
Barbara Pöll, Telfs


Sichtbare Route
visible route
Verdeckte Route
hidden route
Variante
variant
Standplatz
belay
Abseilstelle
abseiling point/anchor
Biwakplatz
bivouac/bivi area
Pendelquergang
pendulum
Bohrhaken
bolt
Normalhaken
(normal) piton
Sanduhr
thread
Dach
roof
Kamin
chimney
Klemmblock
chockstone
Kante
edge
Platten
slabs
Rampe
ramp
Rinne
groove/channel/gully
Riss
crack
Verschneidung
dihedral/V-cleft
Wasserstreifen
water streak/runnel
Überhang
overhang
Band
ledge
Absatz
terrace/pedestal/platform
Nische
niche
Baum
tree
Latsche
mountain pine
Wechte
cornice
Schnee
snow
Gras
grass
Schutt
rubble
Schlüsselstelle
crux
GT
Gedenktafel
plaque
WB
Wandbuch
summit register/visitors log
Gipfelkreuz
summit cross
Fixes Drahtseil
fixed wire rope


SPLIT BOARDING
INTERACTIVE JOURNAL
2488
2398
2640
Steintalspitz
Finsterbach
Zwölferkar
Mittagskar
Sulzkogel
3016
Schaneger
Stierkar
Gamskogel
2965
Gamskögele
Hoher Plattkogel
Hangneigung
OpenSlopeMap
10 - 29°
30 - 34°
35 - 39°
40 - 42°
43 - 45°
46 - 49°
50 - 90°
500 m
2000 ft


Stop or Go
Strategische Lawinenkunde für Tourengeher
Entscheidungsstrategie
CHECK 1
Gefahrenstufe und Hangneigung
1 gering
2 mäßig
Verzicht auf 40° und mehr
3 erheblich
Verzicht auf 35° und mehr
4 groß
Verzicht auf 30° und mehr „Spitzkehrengelände"
5 sehr groß
Verzicht auf Touren allgemein
Gefahrenzeichen erkennen
CHECK 2
Wahrnehmen
Beurteilen
Handeln
Neuschnee?
Triebschnee?
Lawinen?
Durchfeuchtung?
Setzungsgeräusche?
GEFÄHRLICH für mich?
YES: STOP
Ausweichen
Abbrechen
NO: GO


LEGENDE
Erste Hilfe Pistenrettung
Defibrillator
Rettungshubschrauber
Hallenbad
Rennstrecke
Funpark
Speedstrecke
Ischgl Skyfly
Eleven 11 km lange Abfahrt
Gratis-Skibus
Information
Museum
Tennishalle
Buckelpiste
Langlaufloipe
Einlaufplatz
Rodelbahn
Nachtrodelbahn ca. 7 km
Winterwanderweg
E-Car Ladestation
bis zu 70% Gefälle (Piste 14a)
Absturzgefahr
Wildruhezone Val Chöglias
Parkplätze
Parkgarage
Tunnel / Fahrtreppe
SKIPISTEN
leicht
mittel
schwer
beleuchtet
Skipisten sind mit den entsprechend der Schwierigkeitsstufe gefärbten Tafeln markiert. Zusätzlich tragen diese Tafeln die jeweilige Abfahrtsnummer. Skipisten und gesicherte Skigebiete sind markiert, nach Möglichkeit präpariert, kontrolliert und vor atypischen alpinen Gefahren, besonders vor Lawinengefahr, gesichert.
SKIROUTEN / VARIANTEN
Skirouten sind in ihrem Verlauf durch auf die Spitze gestellte, quadratische Tafeln gekennzeichnet. Sie sind fallweise teilpräpariert, nur im unmittelbaren Bereich der Markierung vor Lawinengefahr gesichert und nicht kontrolliert.
extrem
Variantes: nicht präpariert, nicht kontrolliert, nicht markiert, nicht gesichert!
ANLAGEN
Länge / Höhendifferenz
A1 Silvrettabahn 24 Personen
A2 Pardatschgratbahn 28 Pers - Sitzheizung
A3 Fimbelbahn 6 Personen - Sitzheizung
B1 Velillbahn
B2 Fünfjochbahn Sitzheizung
B3 Idjochbahn
B4 Greitspitzlift
B5 Sonnenbahn Sitzheizung - Kinderbahn
B6 Velill Übungslift
B7 Viderböden Übungslift
B8 Gratbahn
B9 Übungsareal Förderband
B10 Viderböden Förderband
C1 Höllbodenbahn
C2 Höllkarbahn
C3 Sassgalenbahn
C4 Nachweidebahn
C5 Lange Wandbahn Sitzheizung
D1 Palinkopfbahn Sitzheizung
D2 Zeblasbahn
E1 Paznauner Thayabahn
E2 Möllspitzbahn
E3 Bodenalpbahn
E4 Gamperbahn Sitzheizung
E5 Piz Val Grondabahn 150 Personen
F1 Velilleckbahn
F2 Pandoramabahn 4 Personen
G1 Mathen Übungslift
G2 Schwarzwasserlift
L1 Doppelstöckige Pendelbahn 180 Pers.
L2 Luftseilbahn 80 Personen
M1 Grivaleabahn
M2 Mullerbahn
N1 Fimsattelbahn
N2 Visnitzbahn
N3 Mileilift Übungslift
N4 Viderjochbahn 1
N5 Alp Trider Eclift
N6 Alp Trider Sattelbahn
O1 Greitspitzbahn
O2 Blais Grondalift
O4 Viderjochbahn 2
P1 Musellalift
P2 Musella 1-Förderband
P3
Samnaun
Ischgl
Silvretta Arena
Tirol
www.samnaun.ch

Recap

- What is Data Visualization?
- Why visualize?
- Models of Data Visualization
- Examples

- Next Week: More on **What** and **Why**