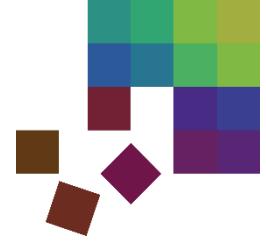


University of Konstanz  
Data Analysis and Visualization Group



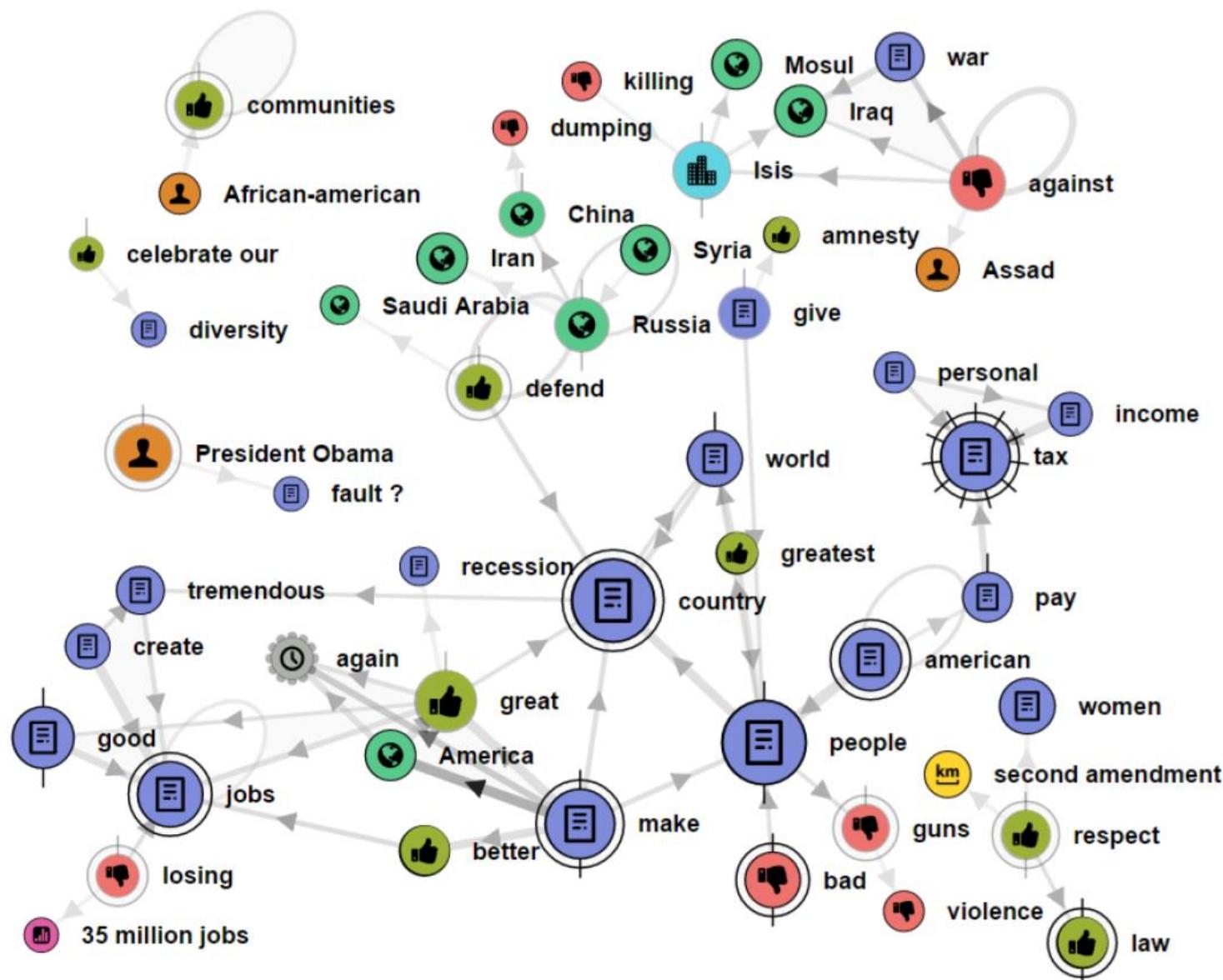
# Visual Analytics for Linguistics

Raphael Buchmüller

Computational Linguistics Fall School 2024

# Presidential Debate Visualization

## Visual Analytics for Linguistics



NEREx (2017)

by Mennatallah El-Assady et al.



Mennatallah El-Assady, Rita Sevastjanova, Bela Gipp, Daniel Keim, and Christopher Collins. NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. Computer Graphics Forum, vol. 36, no. 3, pp. 213-225, 2017.

# Simple Text Visualization Pipeline

Below is a partial transcript of the exchange between Trump and Harris while the debate was ongoing.

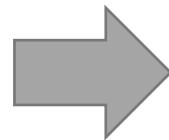
DAVID MUIR: Tonight, the high-stakes showdown here in Philadelphia between Vice President Kamala Harris and former president Donald Trump. Their first face-to-face meeting in this presidential election. Their first face-to-face meeting ever.

LINSEY DAVIS: A historic race for president upended just weeks ago. President Biden withdrawing after his last debate. Donald Trump is now up against a new opponent.

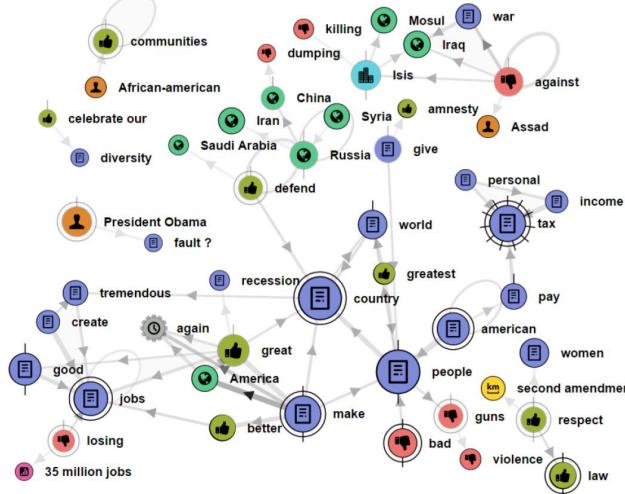
DAVID MUIR: The candidates separated by the smallest of margins. Essentially tied in the polls nationally. And in the key battlegrounds, including right here in Pennsylvania, all still very much in play. The ABC News Presidential Debate starts right now.

DAVID MUIR: Good evening, I'm David Muir. And thank you for joining us for tonight's ABC News Presidential Debate. We want to welcome viewers watching on ABC and around the world tonight. Vice President Kamala Harris and President Donald Trump are just moments away from taking the stage in this unprecedented race for president.

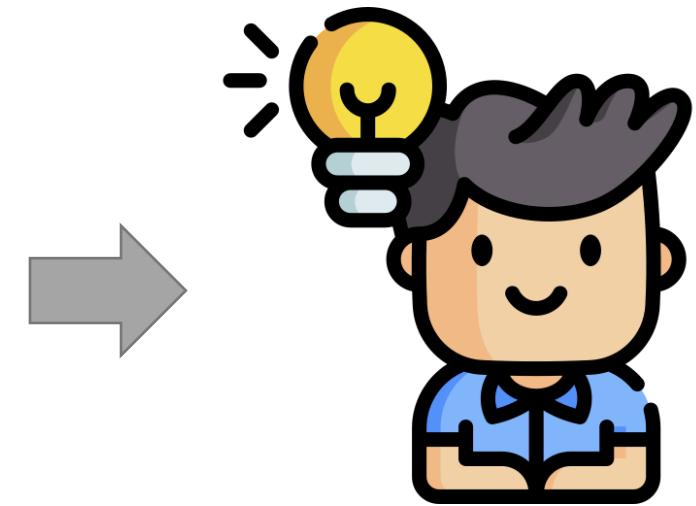
LINSEY DAVIS: And I'm Linsey Davis. Tonight's meeting could be the most consequential event of their campaigns, with Election Day now less than two months away. For Vice President Kamala Harris, this is her first debate since President Biden withdrew from the race on July 21st. Of course, that decision followed his debate against President Donald Trump in June. Since then, this race has taken on an entirely new dynamic.



Text Data



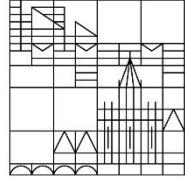
Visualization



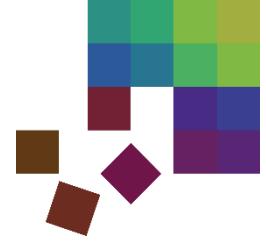
Knowledge

# Whats the Schedule?

	Theory (1:30-3pm)	Practice (3:30-5pm)
Monday	Domain and Design	Design your Approach
Tuesday	Text Processing	Process your Text
Wednesday	Visualization Foundations	Implementation 1
Thursday	Text Visualization	Implementation 2
Friday	Projects, Experiences and Discussion	Present your Approach



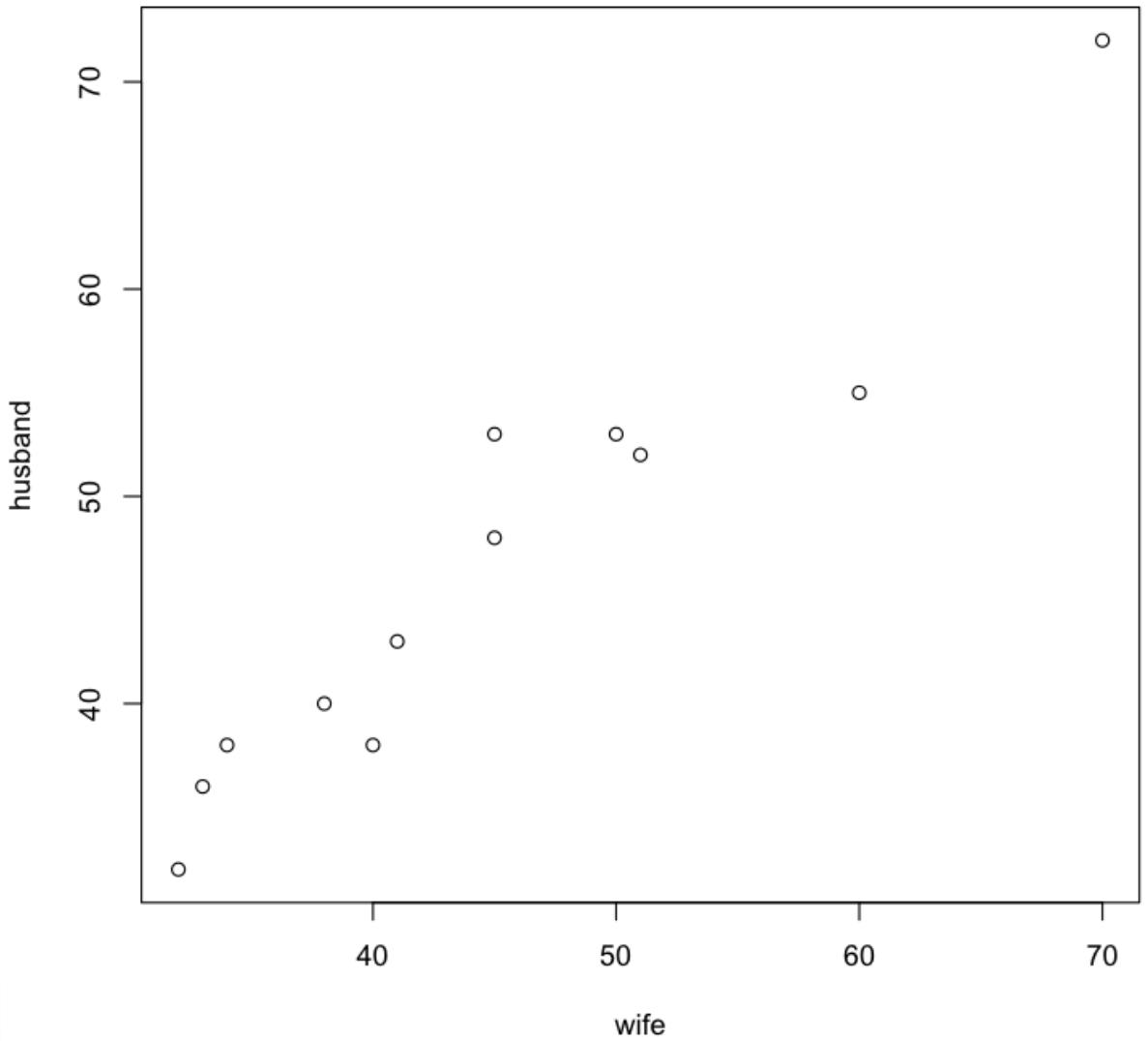
University of Konstanz  
Data Analysis and Visualization Group



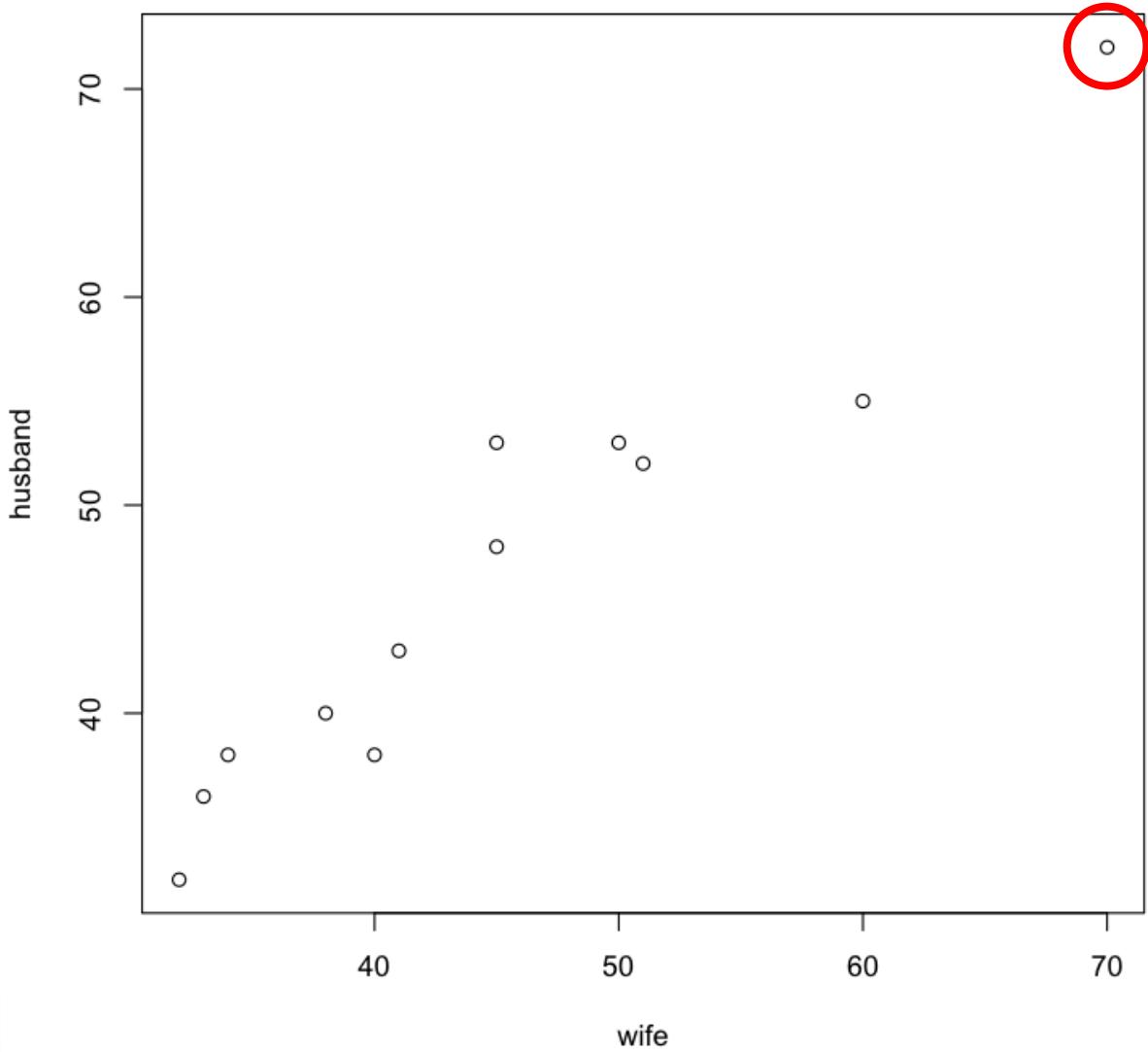
# Why Visualization?

<b>Wife (age)</b>	<b>Husband (age)</b>
45	53
60	55
32	32
33	36
45	48
70	72
50	53
34	38
51	52
40	38
41	43
38	40

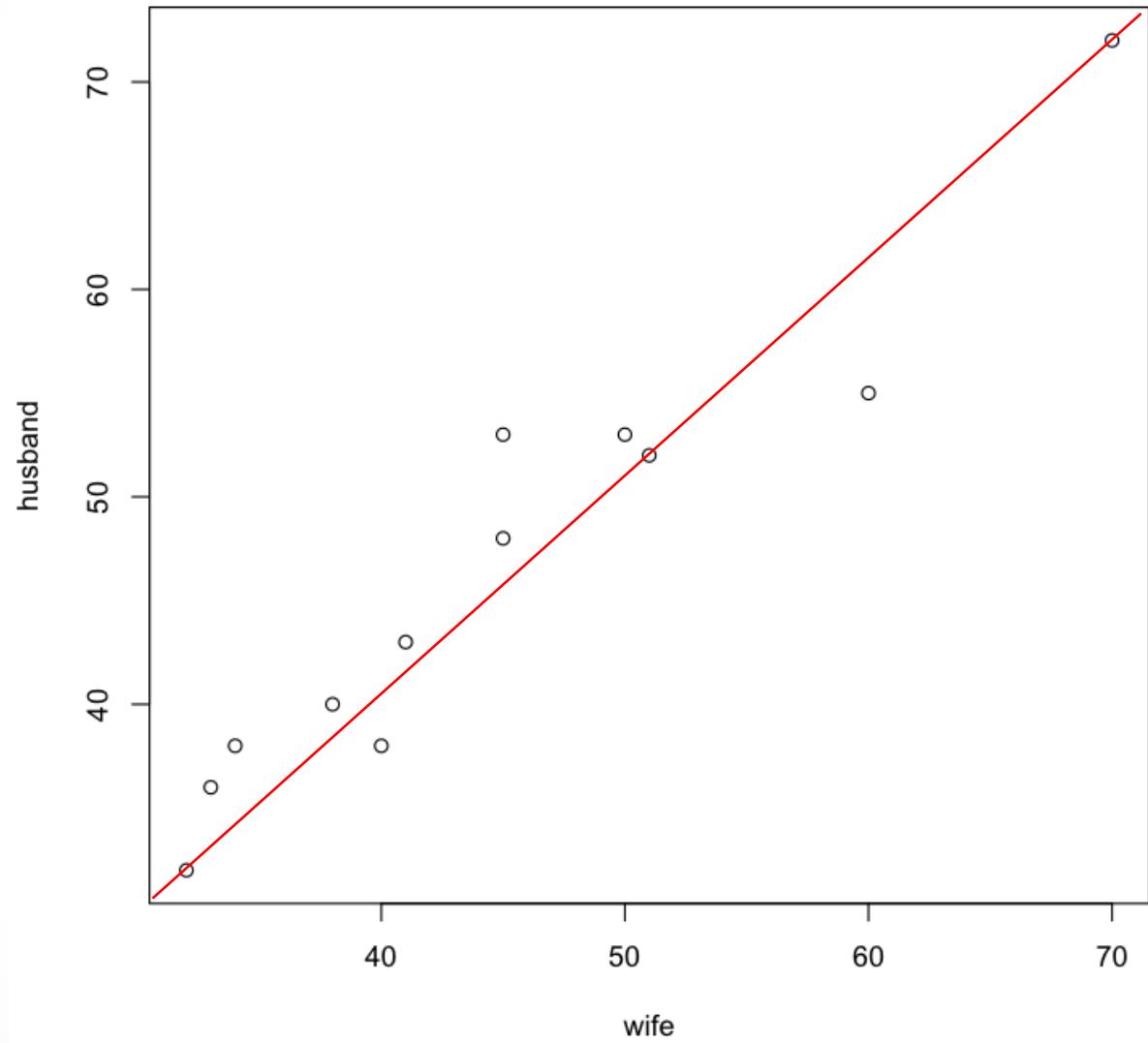
## Presentation of Information



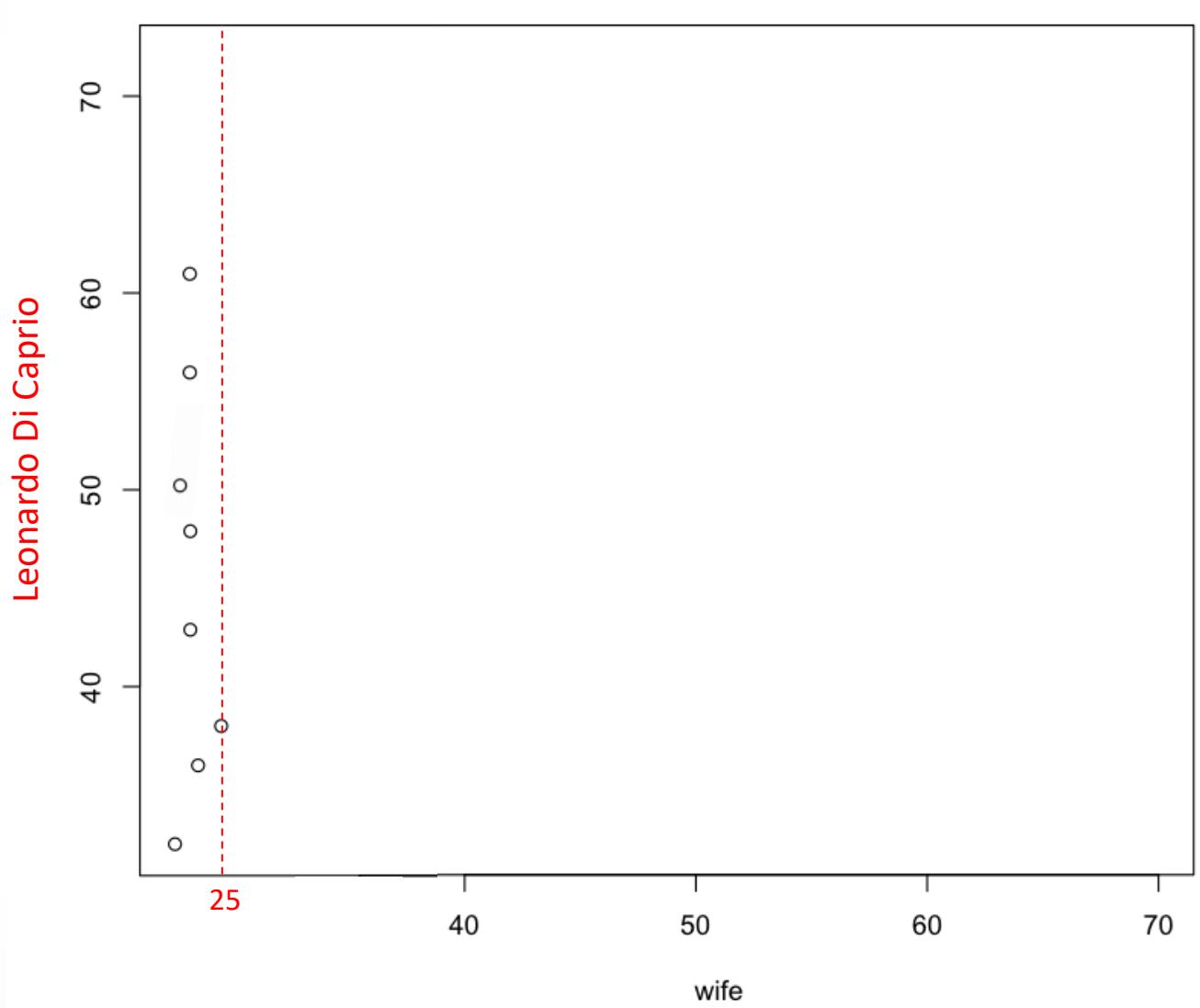
## Presentation of Information



## Presentation of Information



## Presentation of Information



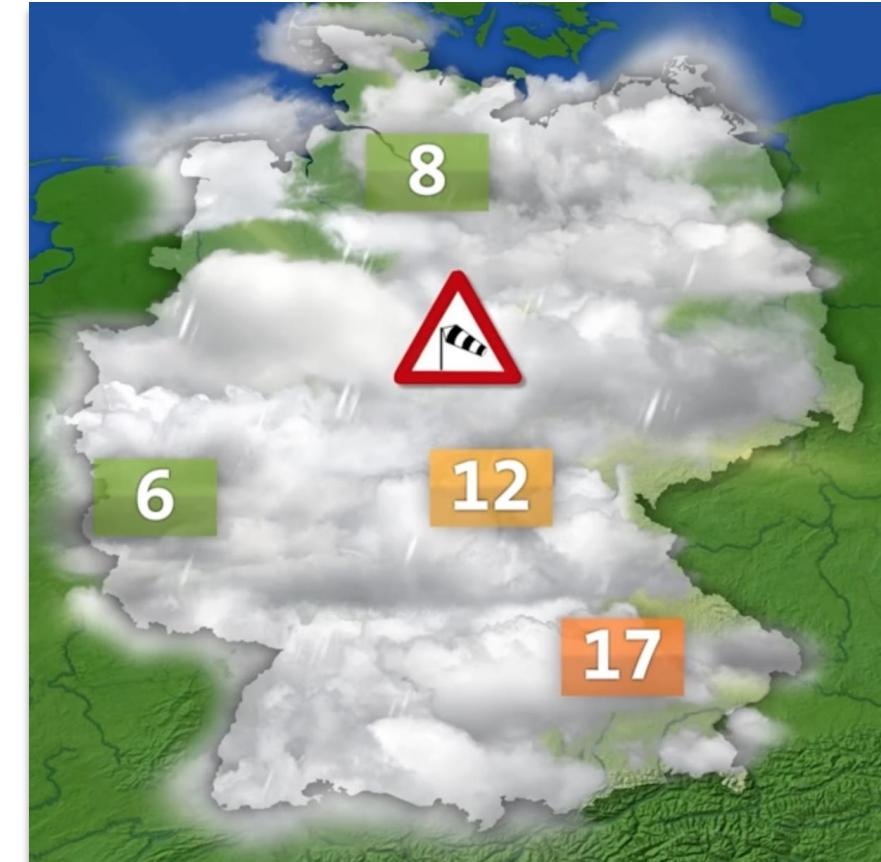
## Presentation of Information



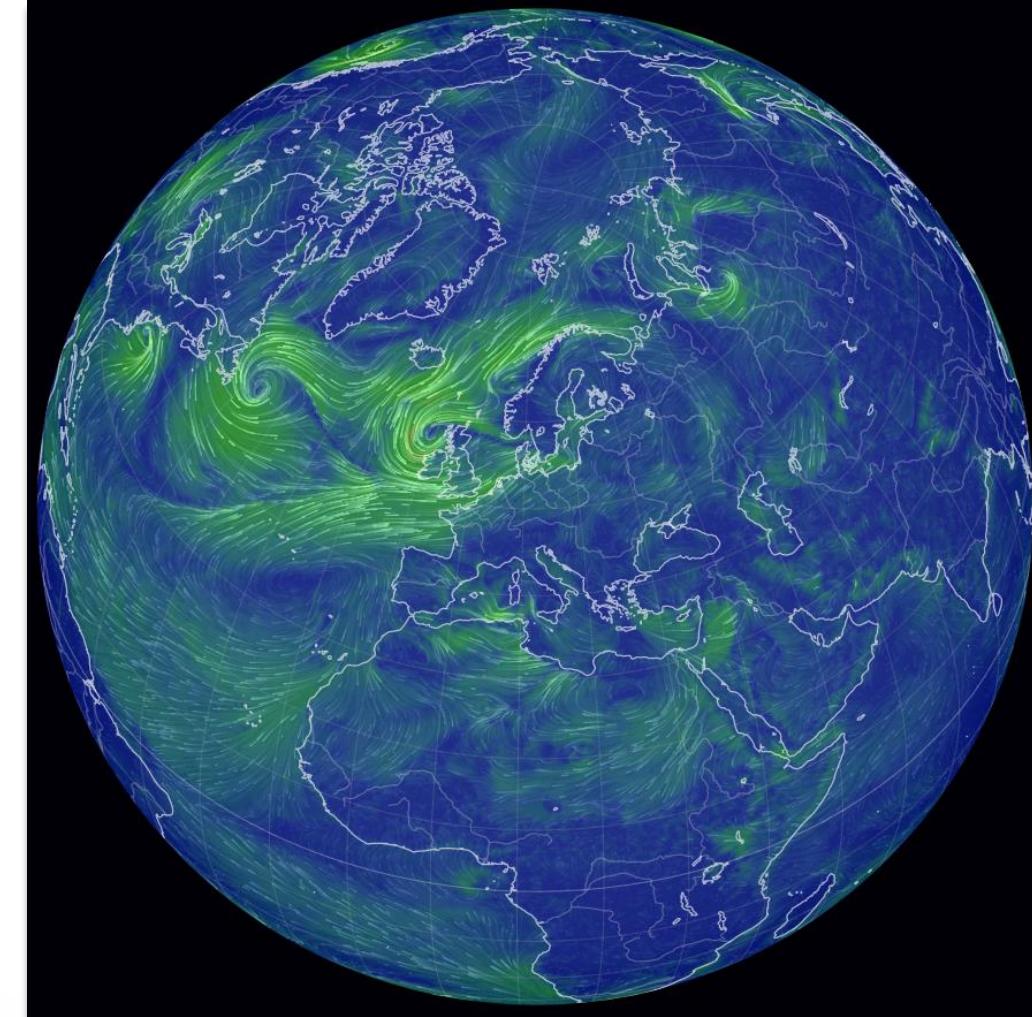
Who is looking at the data?



Weather forecast – temperature and wind information (1990)



Weather forecast – temperature and wind information (2022)



Weather forecast – temperature and wind information (2024)



Who is looking at the data?



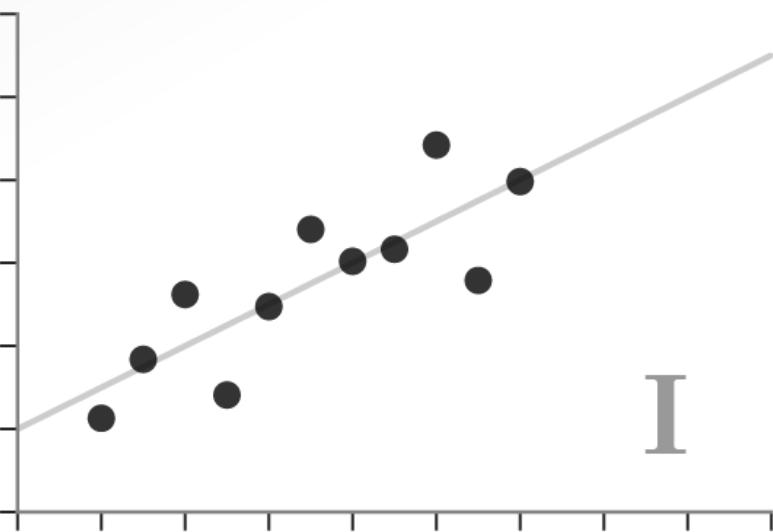


Who is looking at the data?

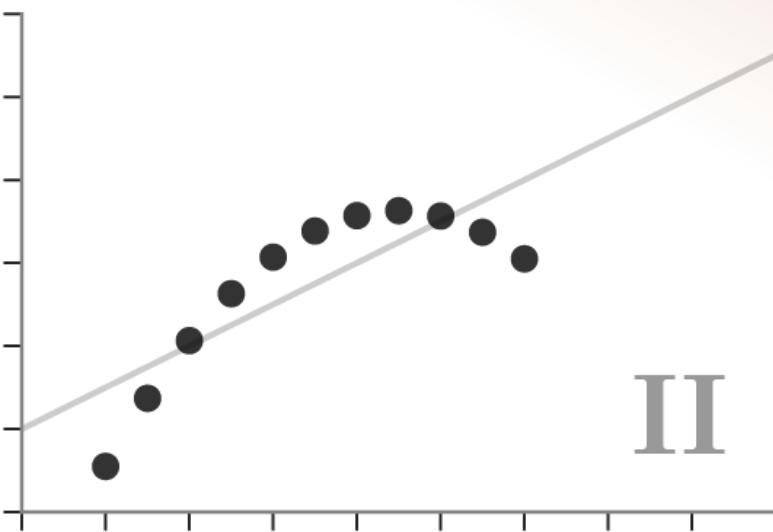


I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	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

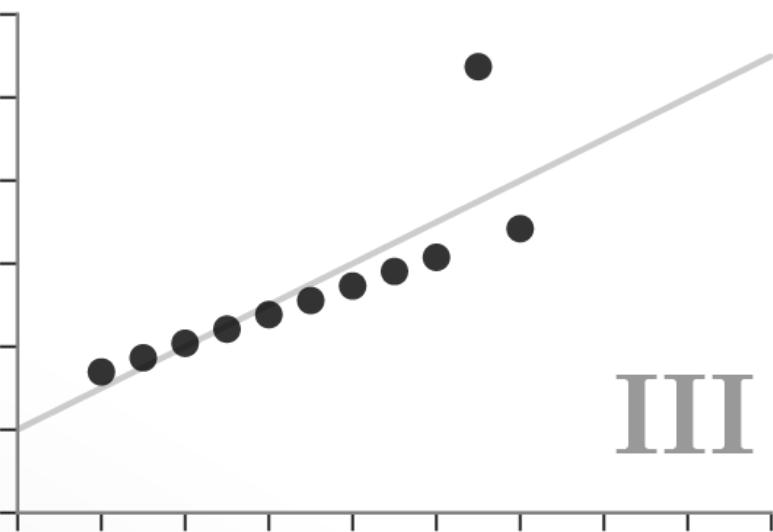
Why should we use visualizations?



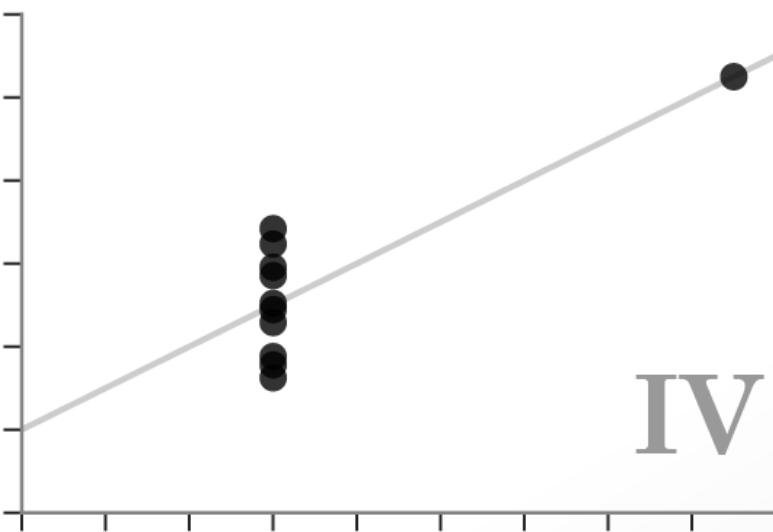
I



II



III



IV

Why visualizations are necessary

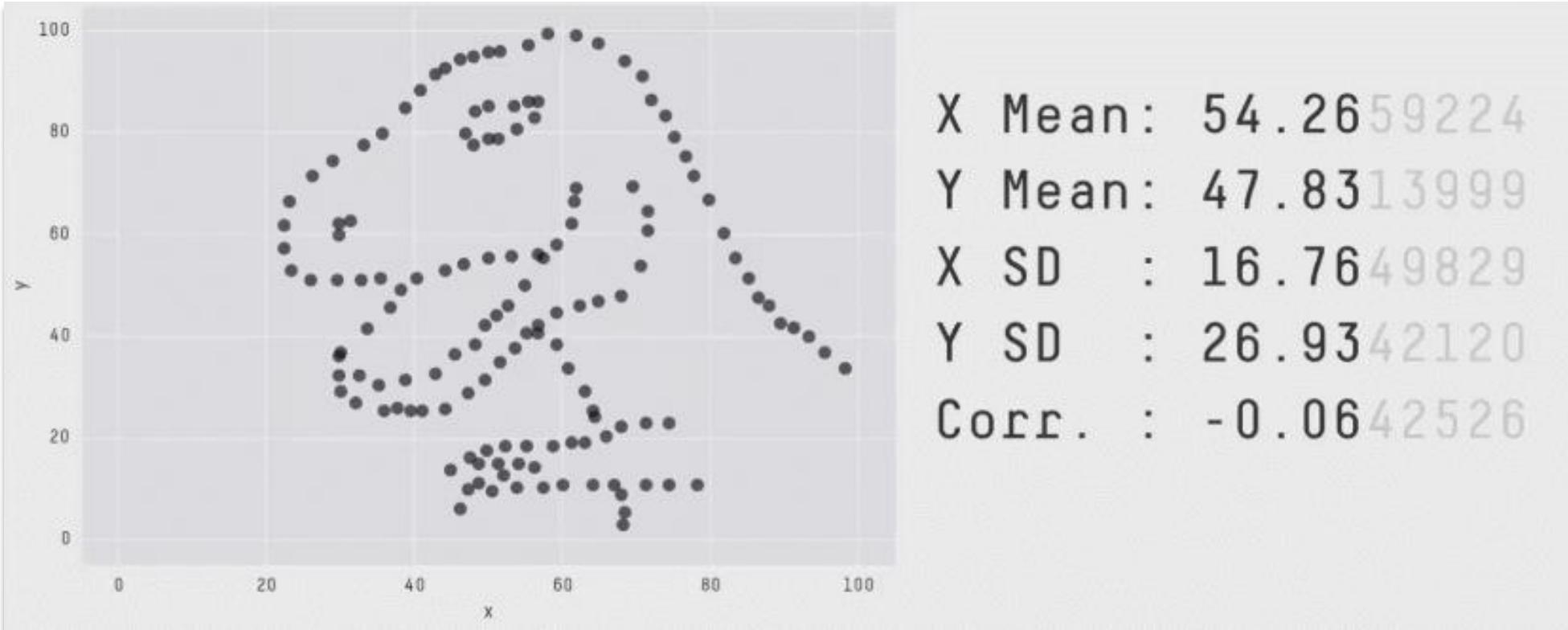


*"...make both calculations and graphs.  
Both sorts of output should be studied;  
each will contribute to understanding."*

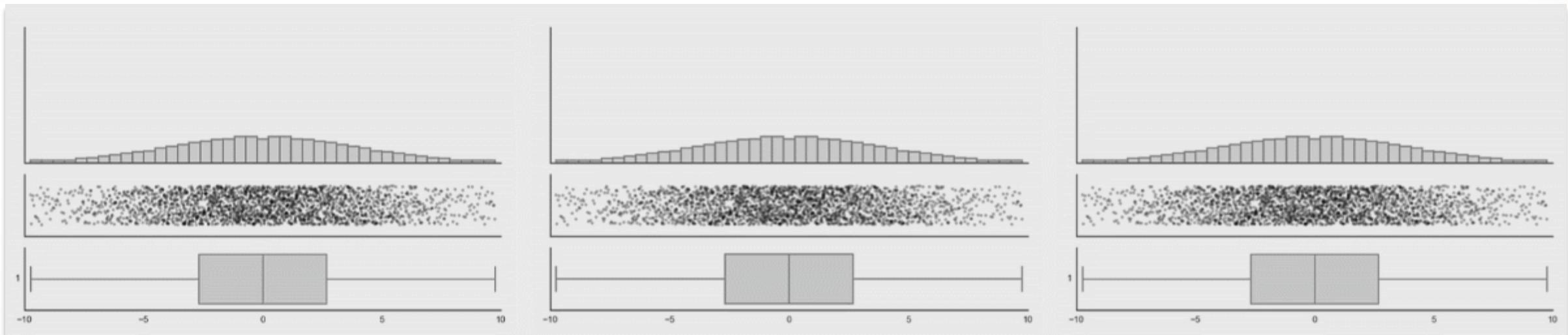
F. J. Anscombe



Identical statistics with varied appearance



Identical statistics with varied appearance



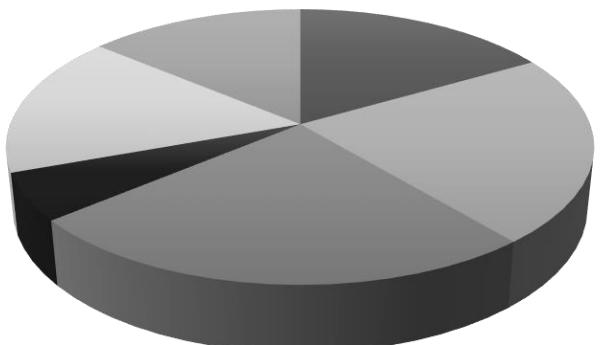
Identical statistics with varied appearance



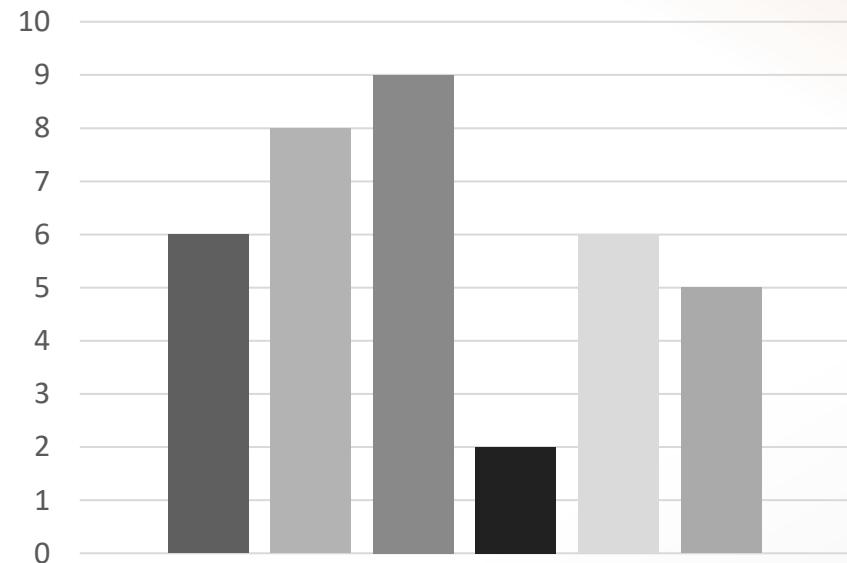
Who is looking at the data?



A

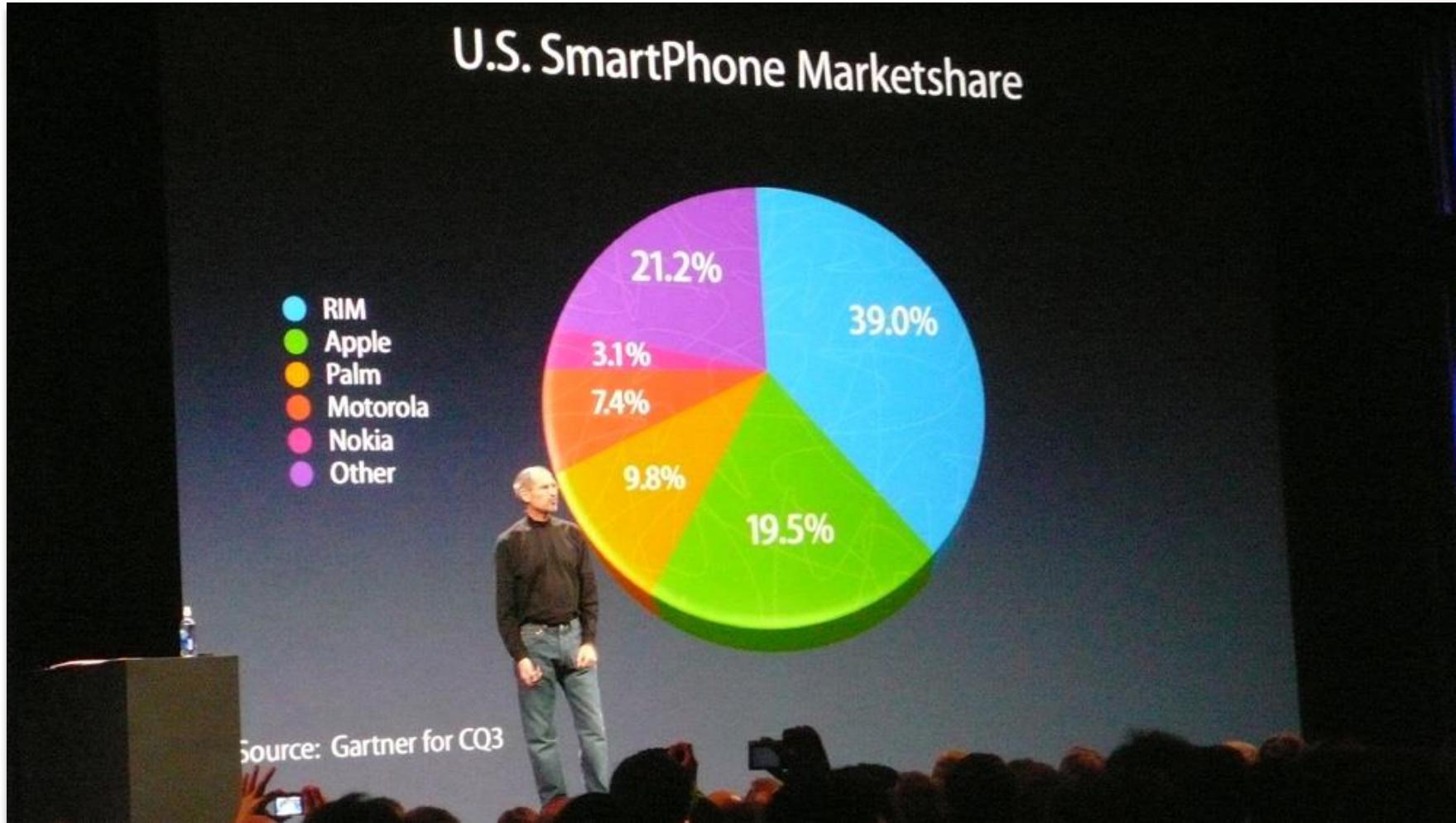


B

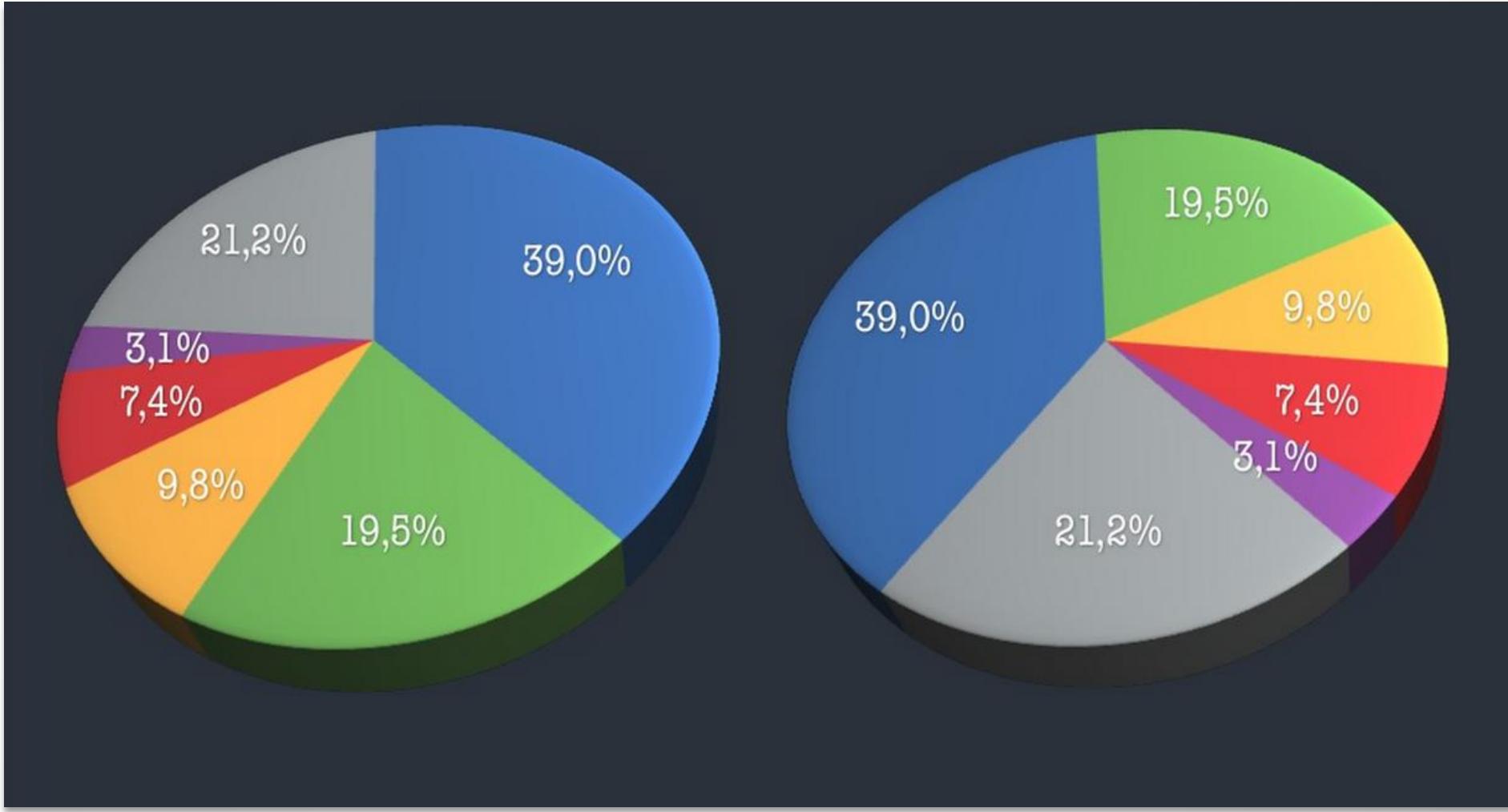


C

Which visualization is the best?



Apple keynote presentation – why 3d is just for marketing purposes



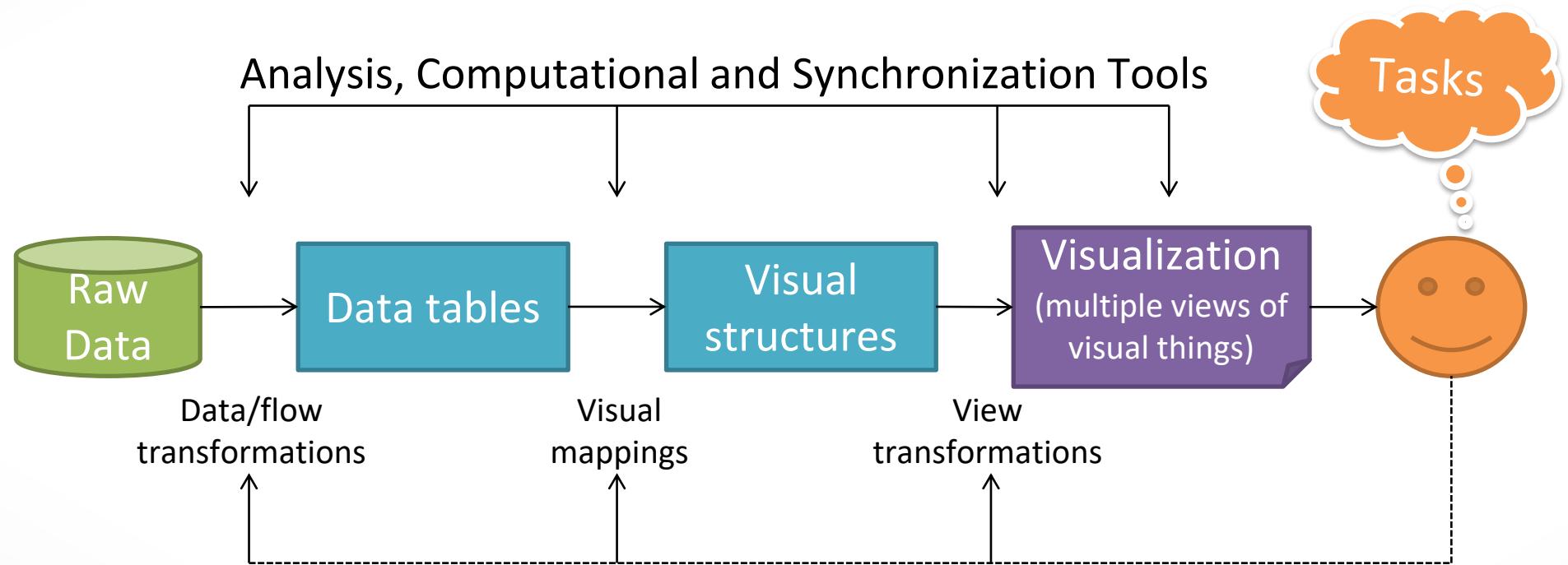
Apple keynote presentation – why 3d is just for marketing purposes

# The Visualization Process

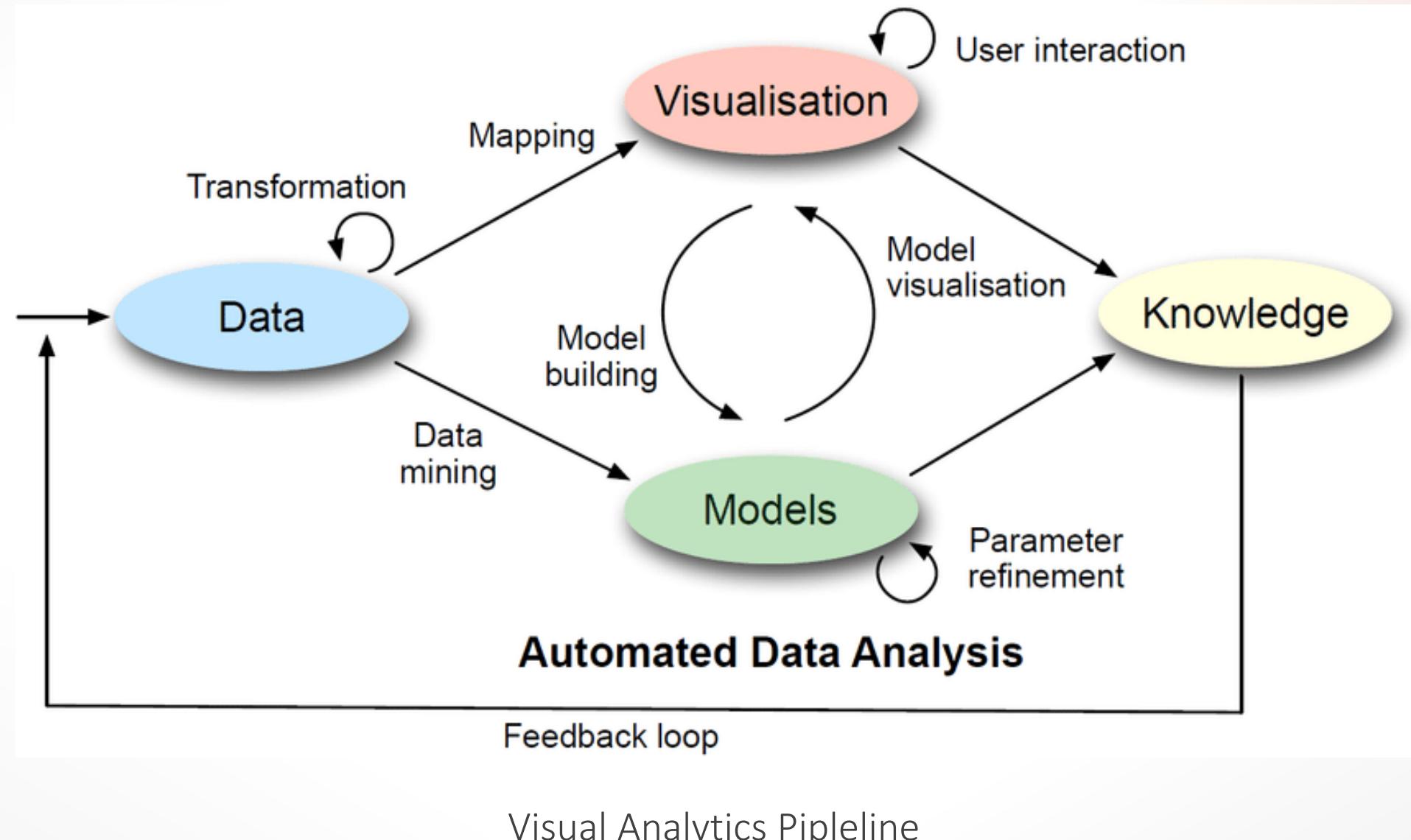
- Analysis of
  - The type of data.
  - The information the viewer hopes to extract.
- Preprocess the data.
- Define a mapping from the data to the display.
- Provide interactive controls if necessary.

# The Visualization Process

- Visualization is often part of a larger process:
  - Explorative data analysis.
  - Knowledge discovery.
  - Visual analytics.
- Visualization and analysis go, thereby, hand in hand.
- The process of starting with data and generating an image, a visualization, or a model is described as a **pipeline**.



Information Visualization Reference Model



# Interdisciplinary Field

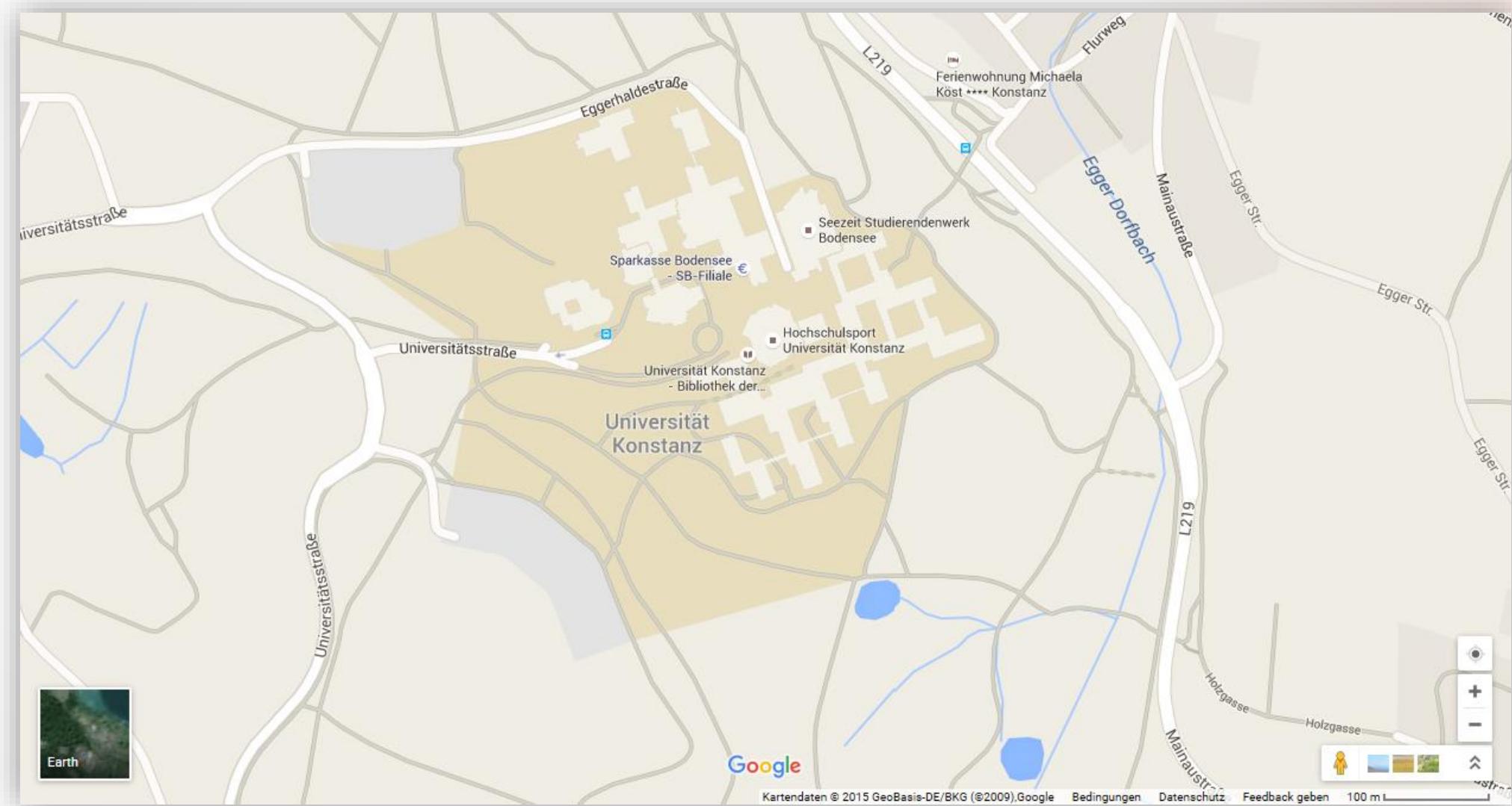
- Human-computer-interaction
- Perceptual psychology
- Databases
- Statistics
- Models
- ...

# Visualization in Everyday Life

- Table in newspaper.
- A train and subway map.
- A map of a region.
- A weather chart.
- A graph of stock market activities.
- A 3D image of your injured knee.
- A highway sign indicating a curve.
- ...



Visualization in Everyday Life



## Visualization in Everyday Life



Why should we use visualizations?

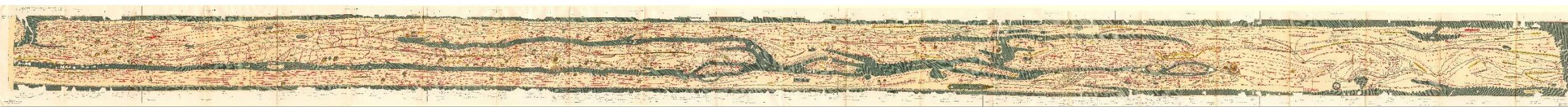
# Early Visualizations

Early visualizations came about out of necessity:

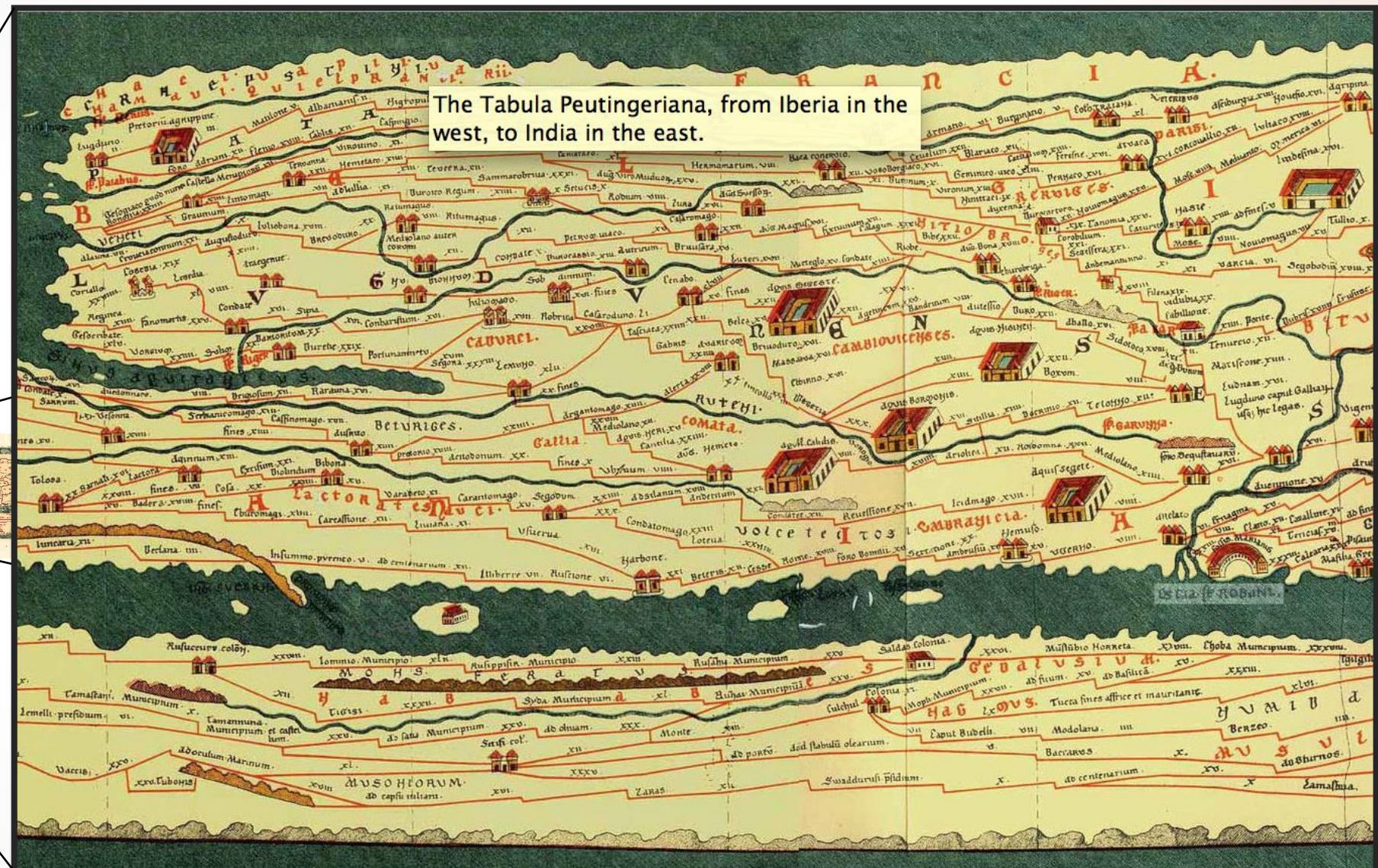
- For travel
- Commerce
- Religion
- Communication
- ...



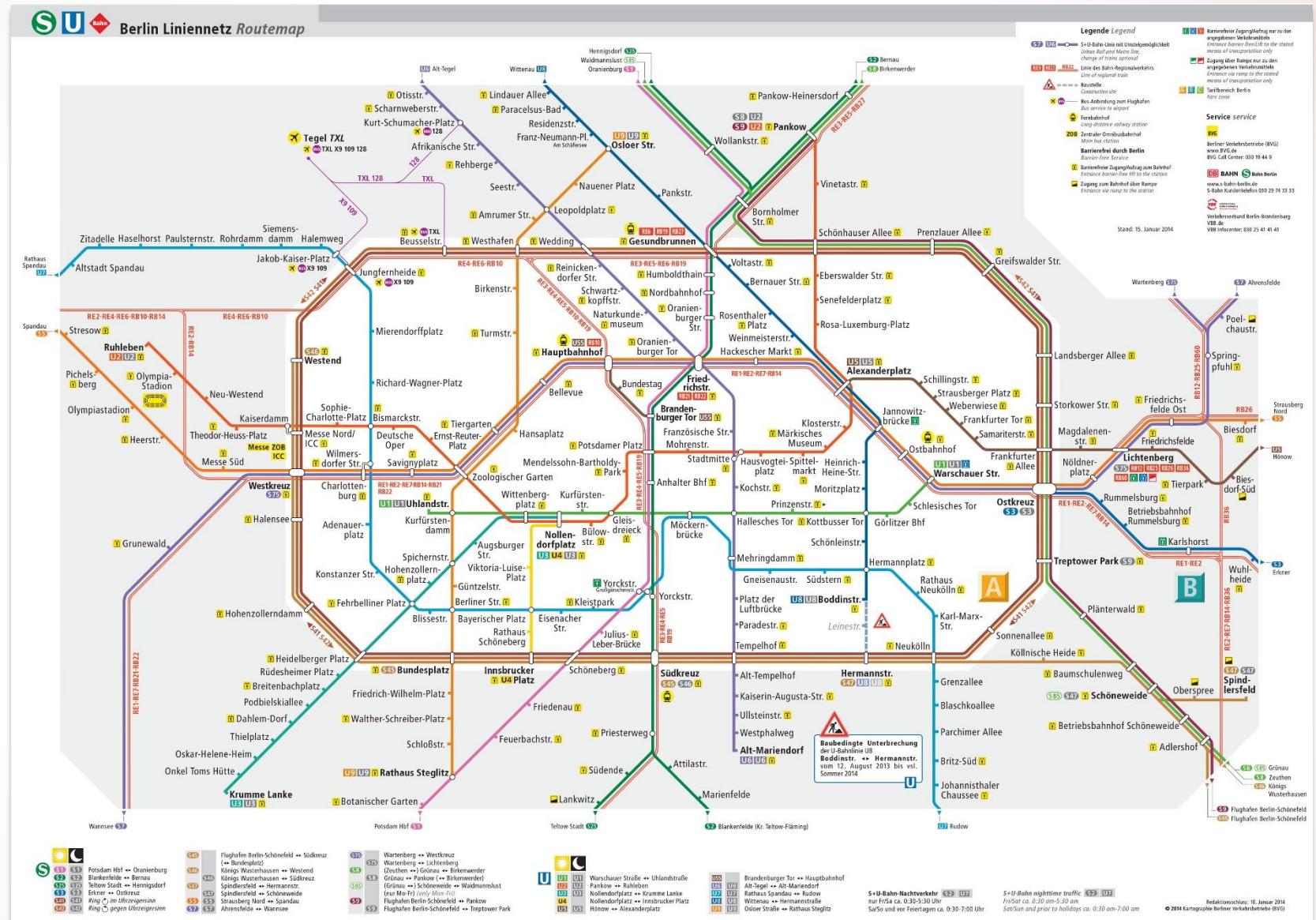
Kish Tablet - early graphical writing (3500 BC)



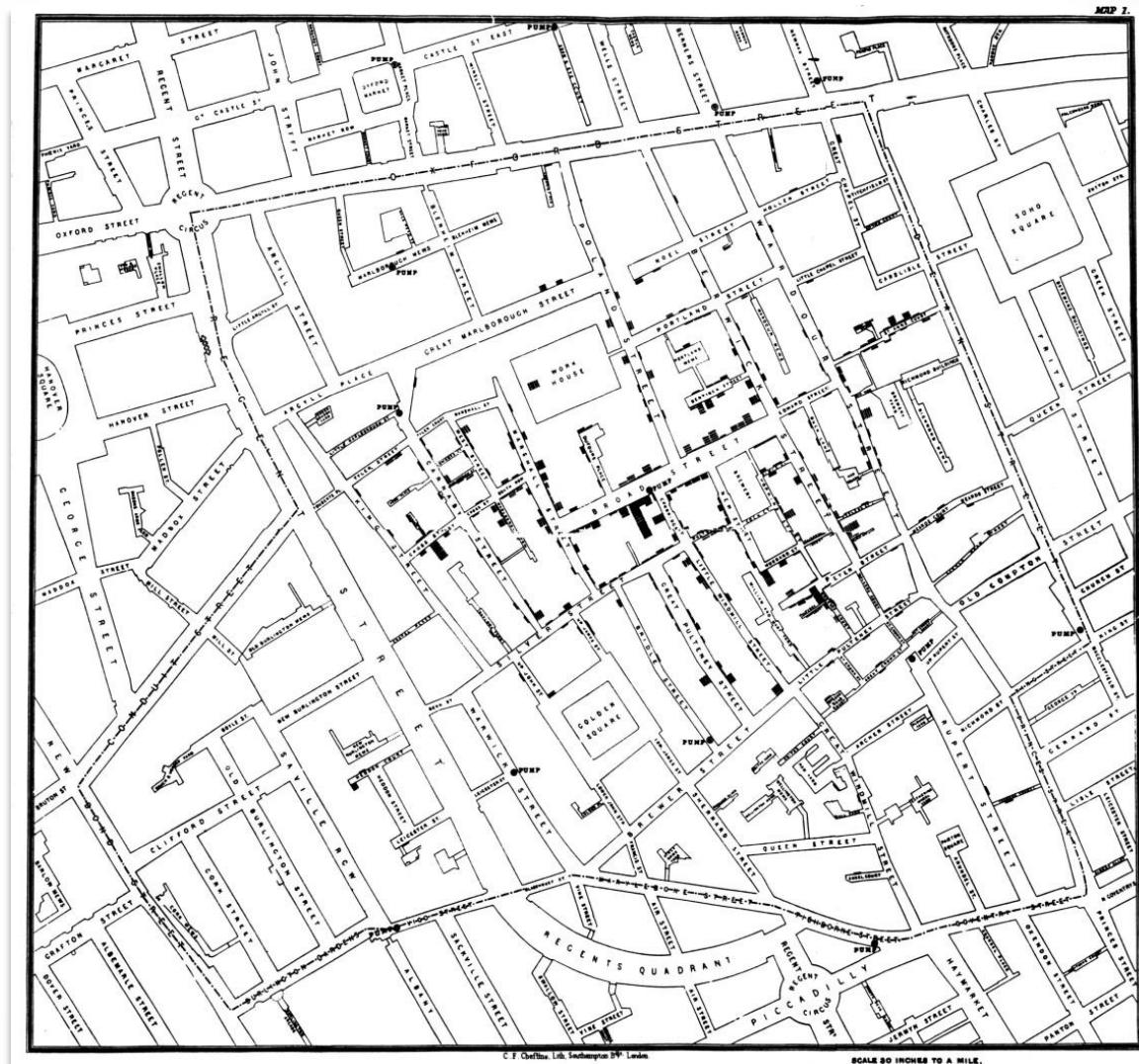
Peutinger Map - scrolls measuring approximately 34 cm high by 6,74 m (1265)



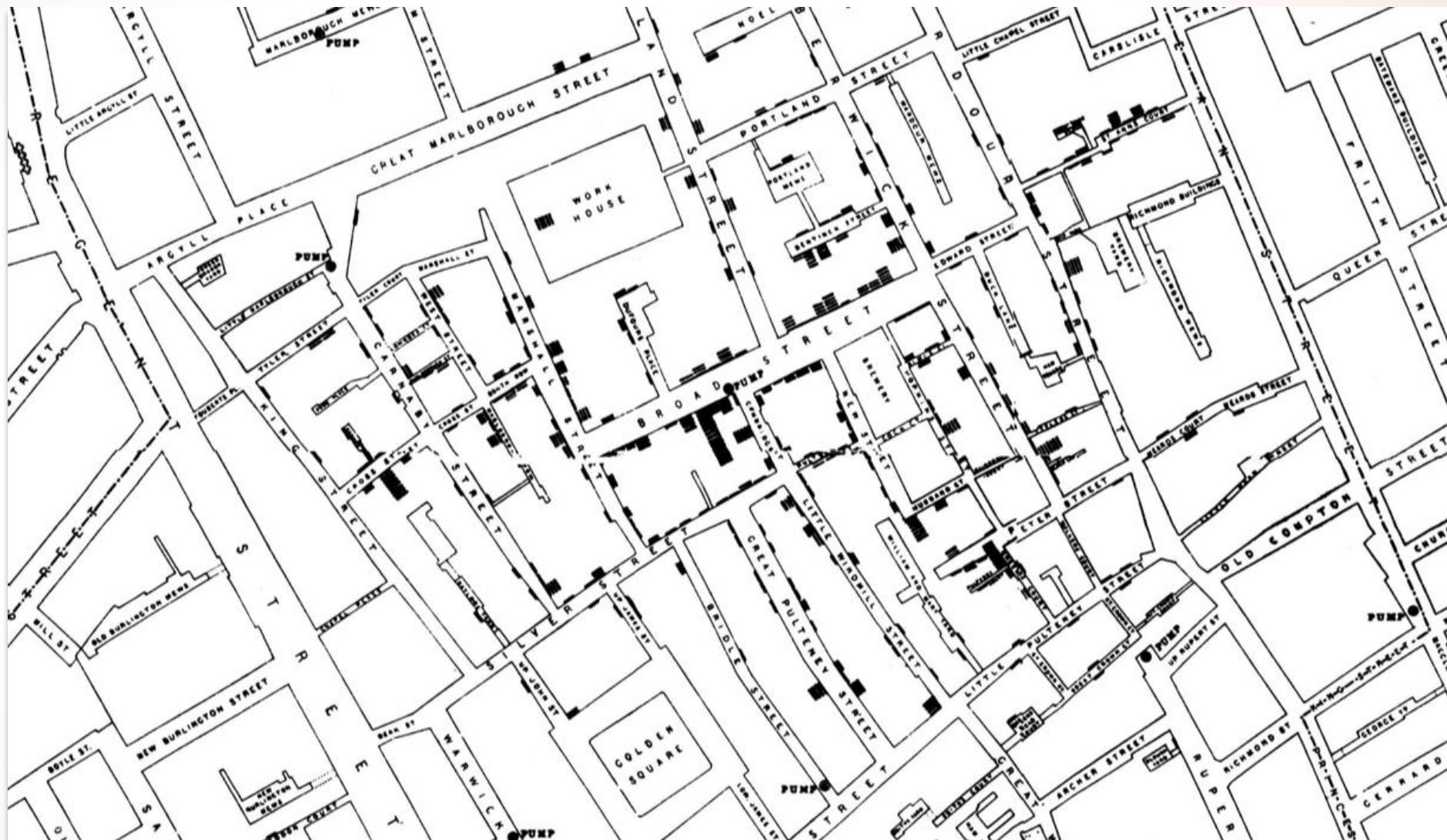
Peutinger Map - scrolls measuring approximately 34 cm high by 6,74 m (1265)



# Berlin Subway (2022)



Deaths from Cholera in London (1854)



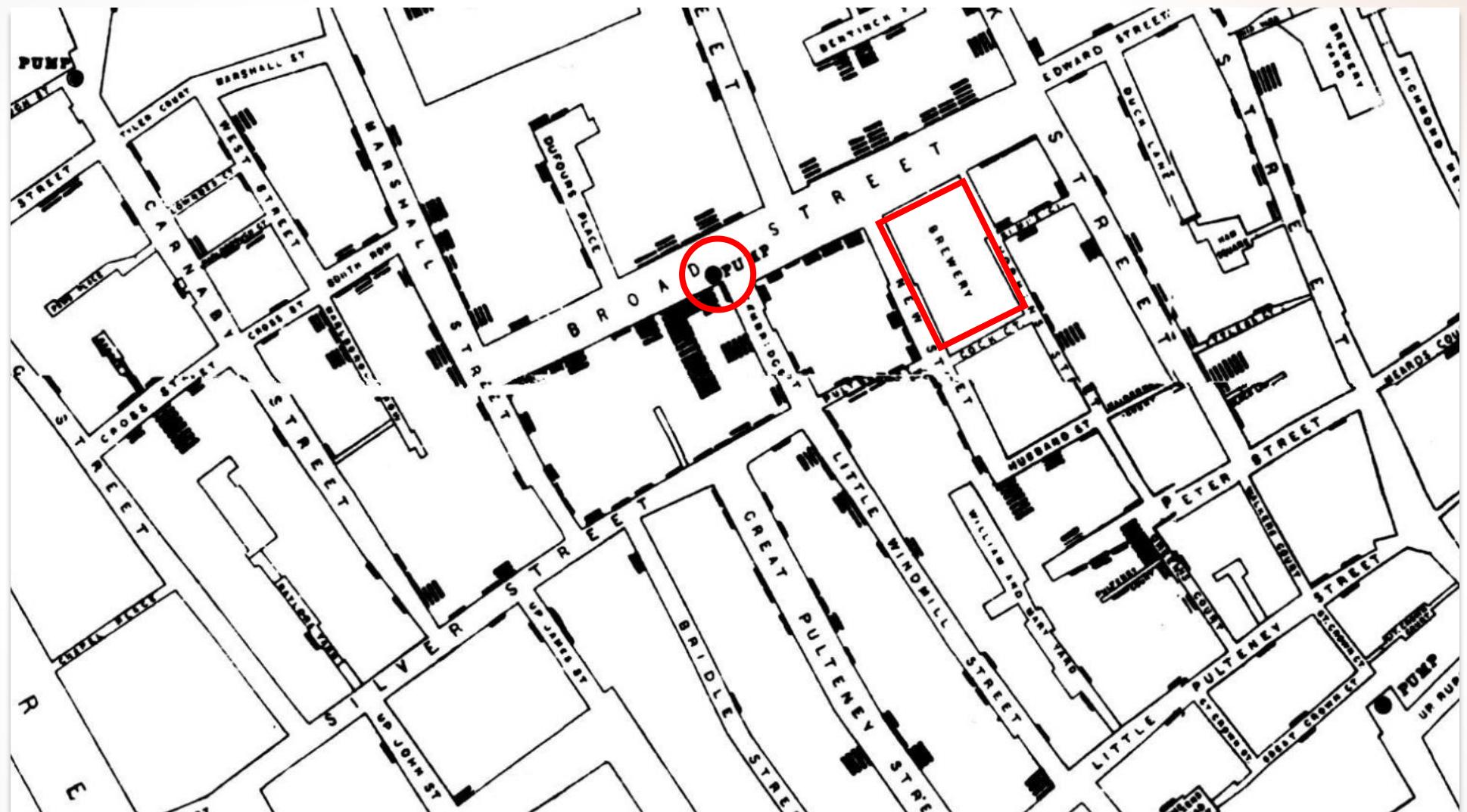
Deaths from Cholera in London (1854)



Deaths from Cholera in London (1854)



Deaths from Cholera in London (1854)



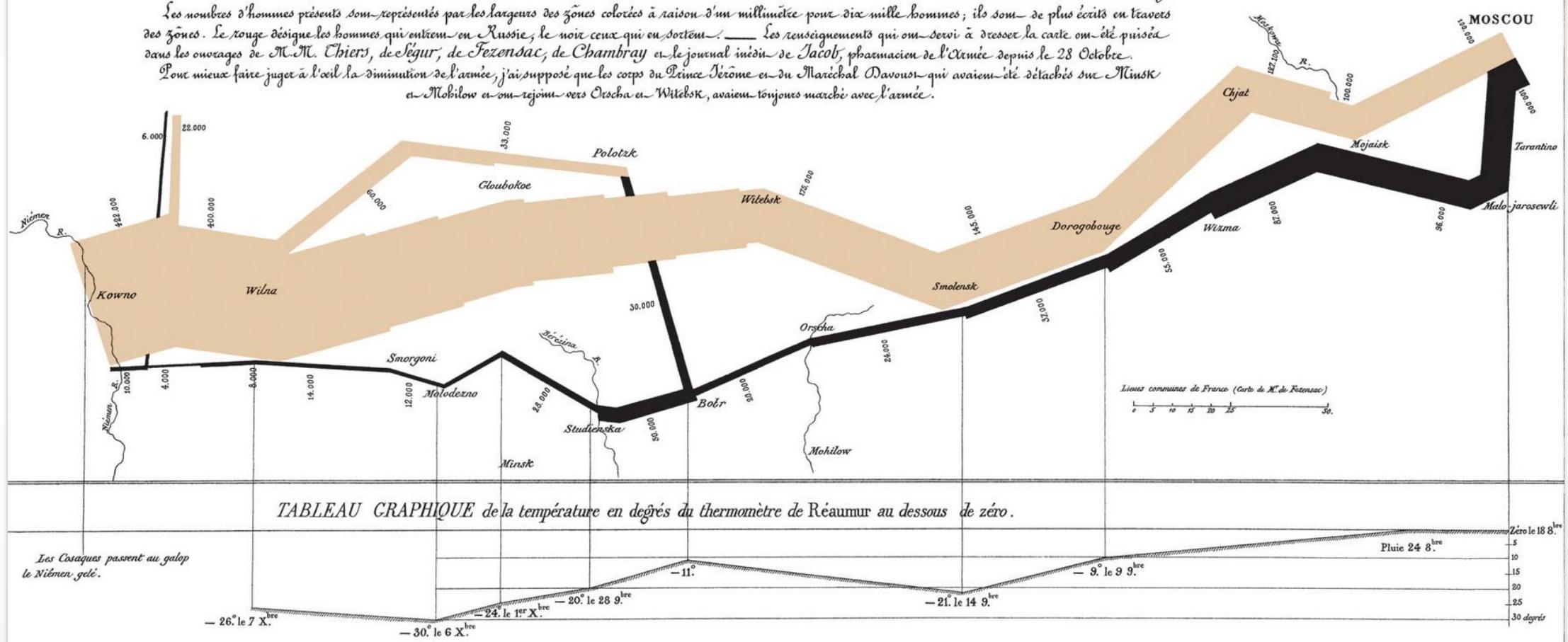
Deaths from Cholera in London (1854)

## Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

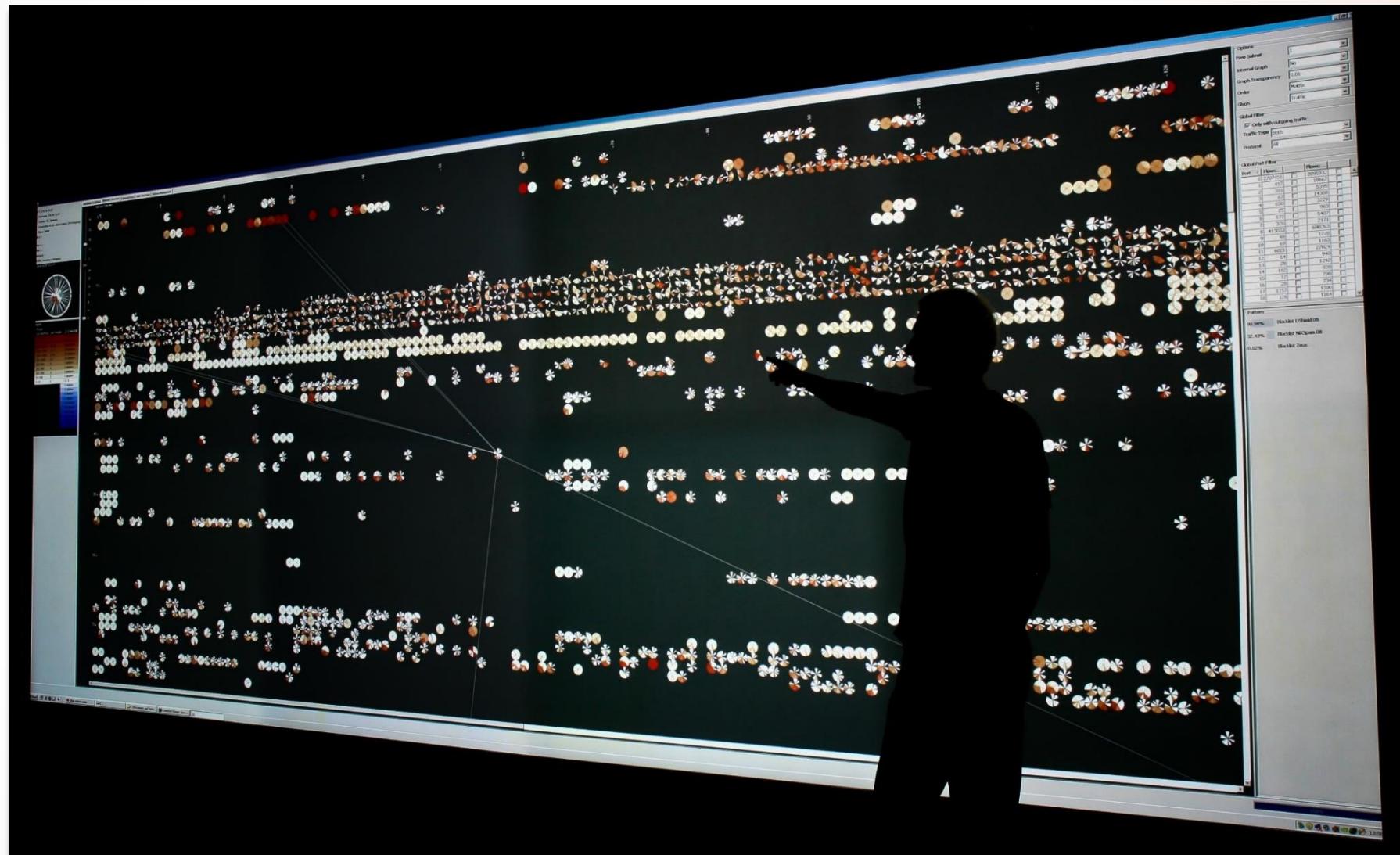
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en lettres des zones. Le rouge désigne les hommes qui entrent en Russie; le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Segur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

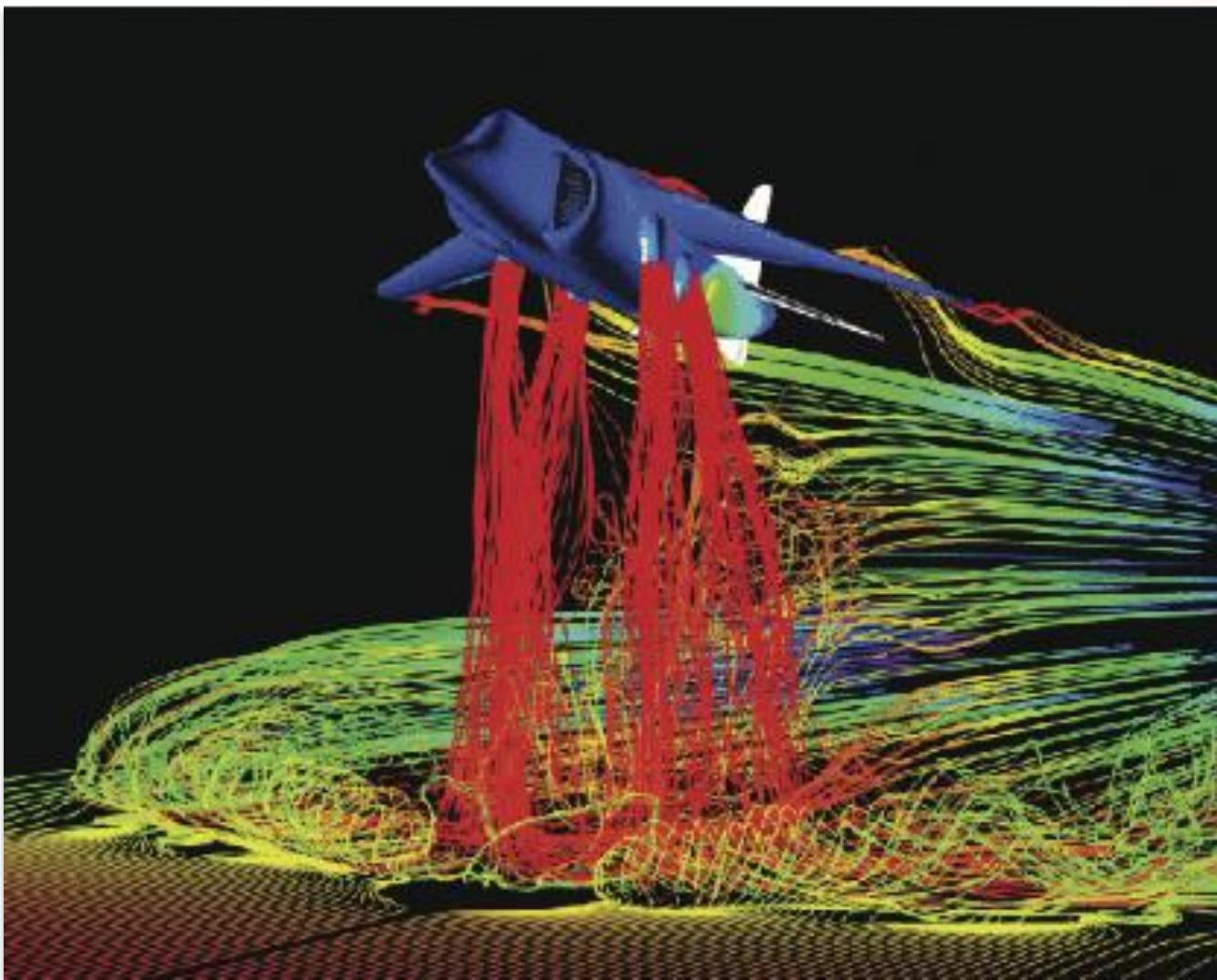
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout, qui avaient été détachés sur Minsk et Mohilow en se rejoignant vers Orsha et Witebsk, avaient toujours marché avec l'armée.



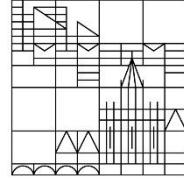
Minard's Map (1869)



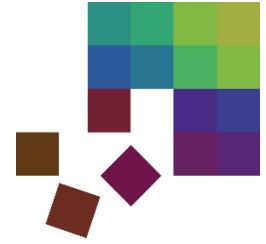
Network traffic – thousands of devices visualized on a Powerwall display (today)



Aircraft flight simulation – harrier hovering over a runway (today)



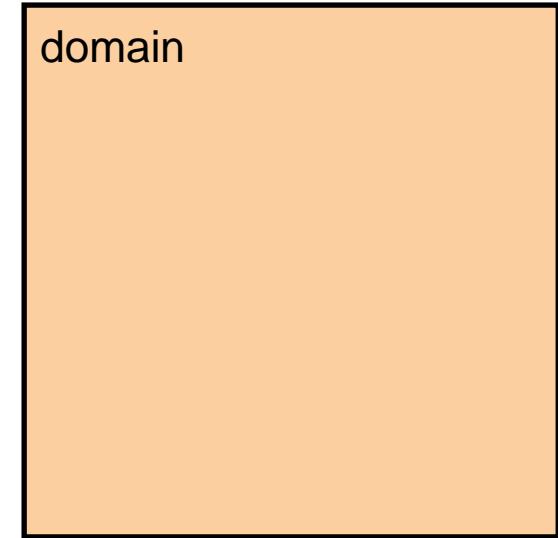
University of Konstanz  
Data Analysis and Visualization Group



# How to design visualizations?

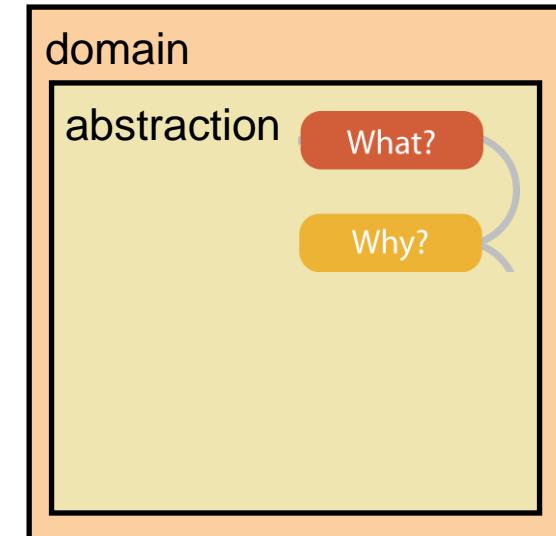
# Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users?



# Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data abstraction**
    - **why** is the user looking at it? **task abstraction**

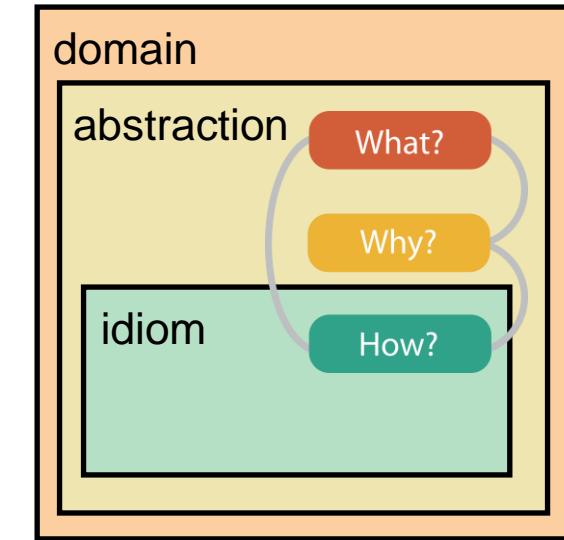


[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]

# Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data abstraction**
    - **why** is the user looking at it? **task abstraction**
- *idiom*
  - how** is it shown?
    - **visual encoding idiom**: how to draw
    - **interaction idiom**: how to manipulate

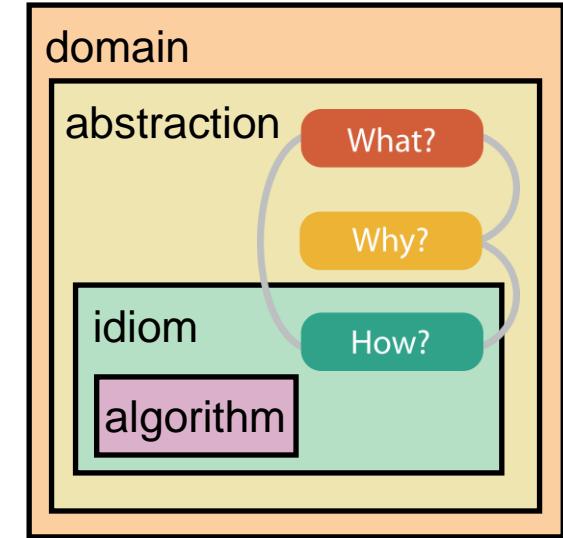


[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]

# Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data abstraction**
    - **why** is the user looking at it? **task abstraction**
- *idiom*
  - **how** is it shown?
    - **visual encoding idiom**: how to draw
    - **interaction idiom**: how to manipulate
- *algorithm*
  - efficient computation

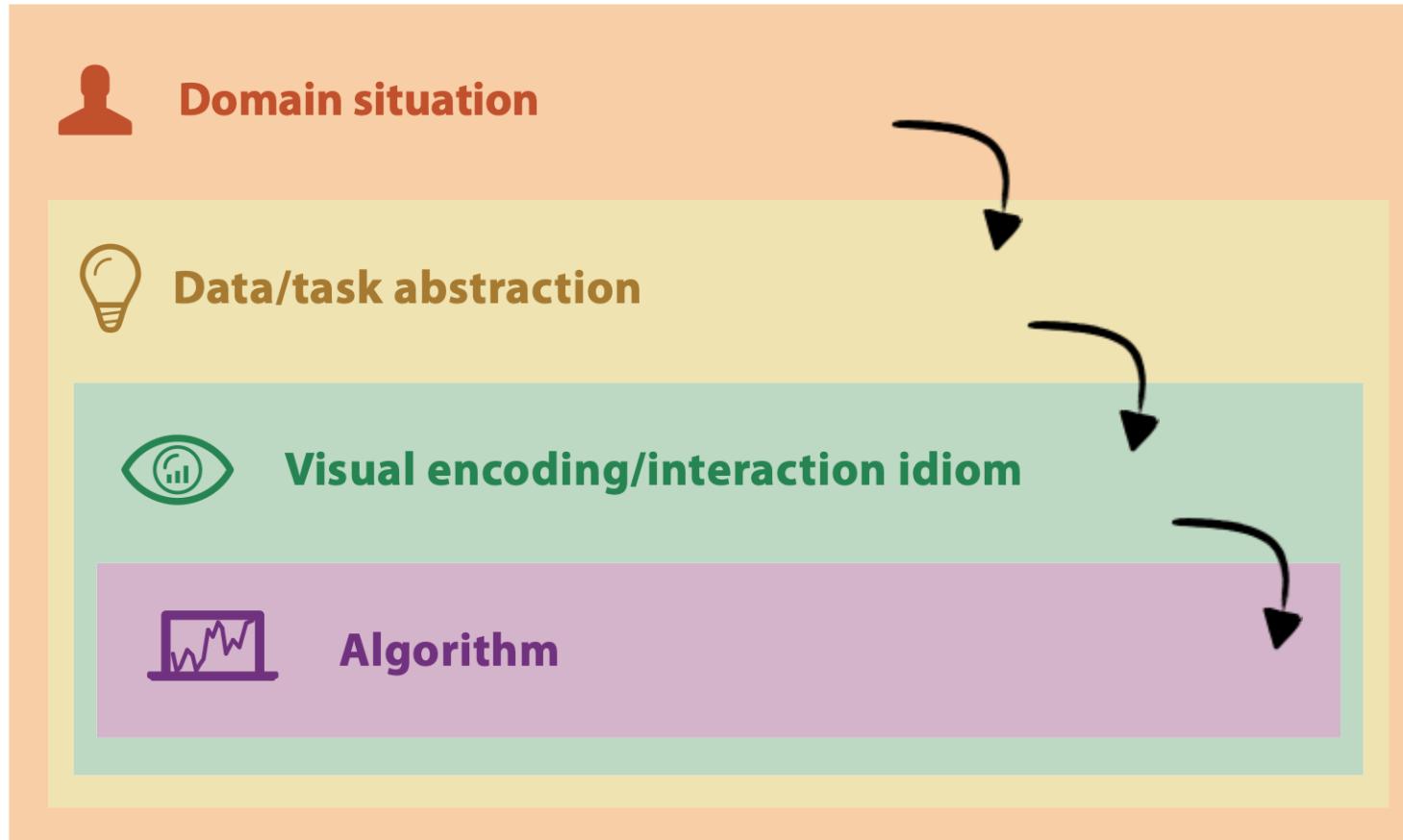


[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]

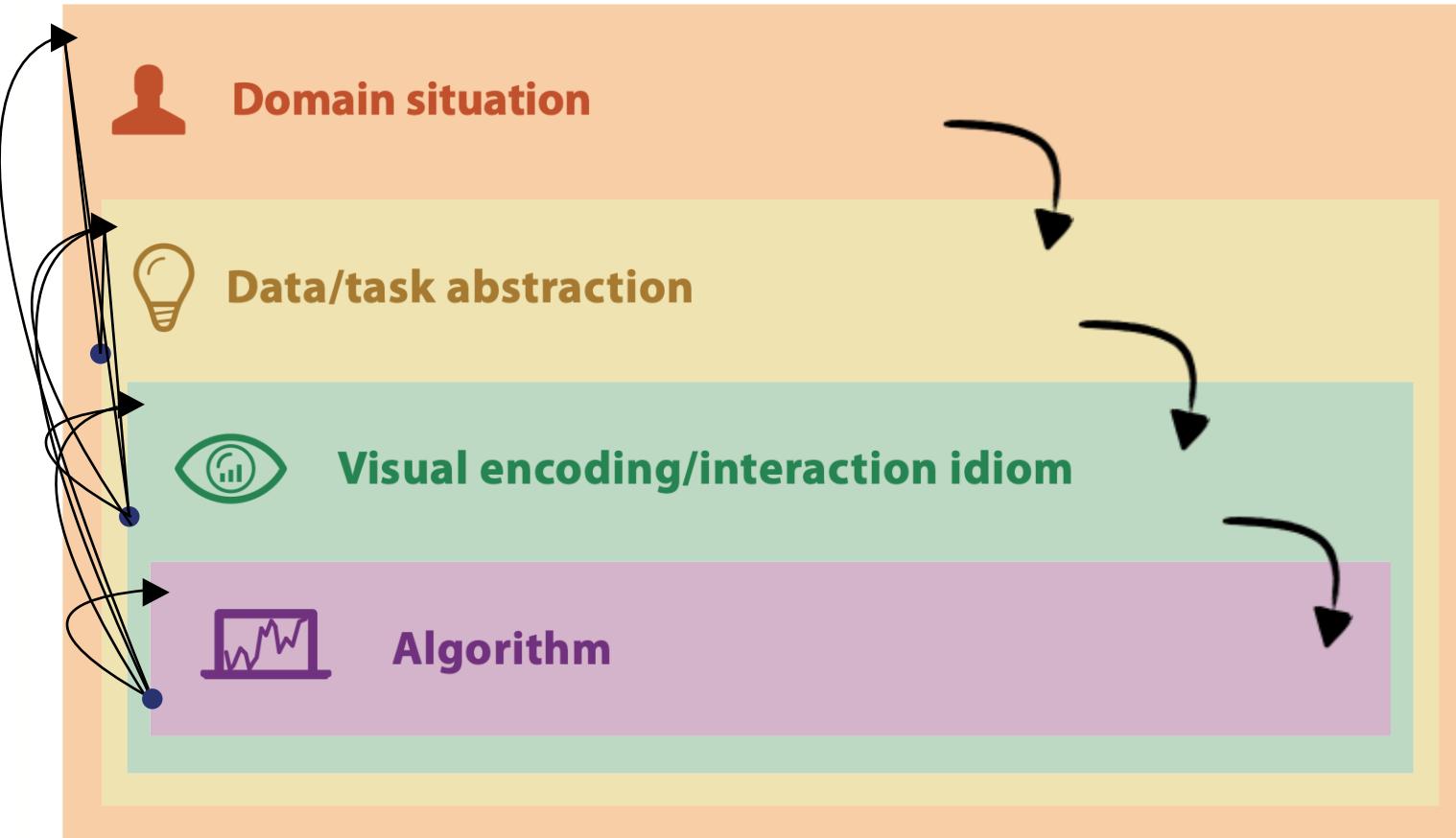
# Nested model

- downstream: cascading effects



# Nested model

- downstream: cascading effects
- upstream: iterative refinement



# Why is validation difficult?

- different ways to get it wrong at each level



## Domain situation

You misunderstood their needs



## Data/task abstraction

You're showing them the wrong thing



## Visual encoding/interaction idiom

The way you show it doesn't work



## Algorithm

Your code is too slow

# Why is validation difficult?

- solution: use methods from different fields at each level

## Algorithm

Measure system time/memory  
Analyze computational complexity

# Why is validation difficult?

- solution: use methods from different fields at each level

computer  
science



## Algorithm

Measure system time/memory  
Analyze computational complexity



technique-  
driven work

# Why is validation difficult?

- solution: use methods from different fields at each level

design

computer  
science

cognitive  
psychology

 <b>Visual encoding/interaction idiom</b> Justify design with respect to alternatives
 <b>Algorithm</b> Measure system time/memory Analyze computational complexity
Analyze results qualitatively Measure human time with lab experiment ( <i>lab study</i> )



technique-  
driven work

# Why is validation difficult?

- solution: use methods from different fields at each level

anthropology/  
ethnography

design

computer  
science

cognitive  
psychology

anthropology/  
ethnography

## 👤 Domain situation

Observe target users using existing tools

## 💡 Data/task abstraction

## 👁️ Visual encoding/interaction idiom

Justify design with respect to alternatives

## 💻 Algorithm

Measure system time/memory

Analyze computational complexity

Analyze results qualitatively

Measure human time with lab experiment (*lab study*)

Observe target users after deployment (*field study*)

Measure adoption

technique-  
driven work

# Why is validation difficult?

- solution: use methods from different fields at each level

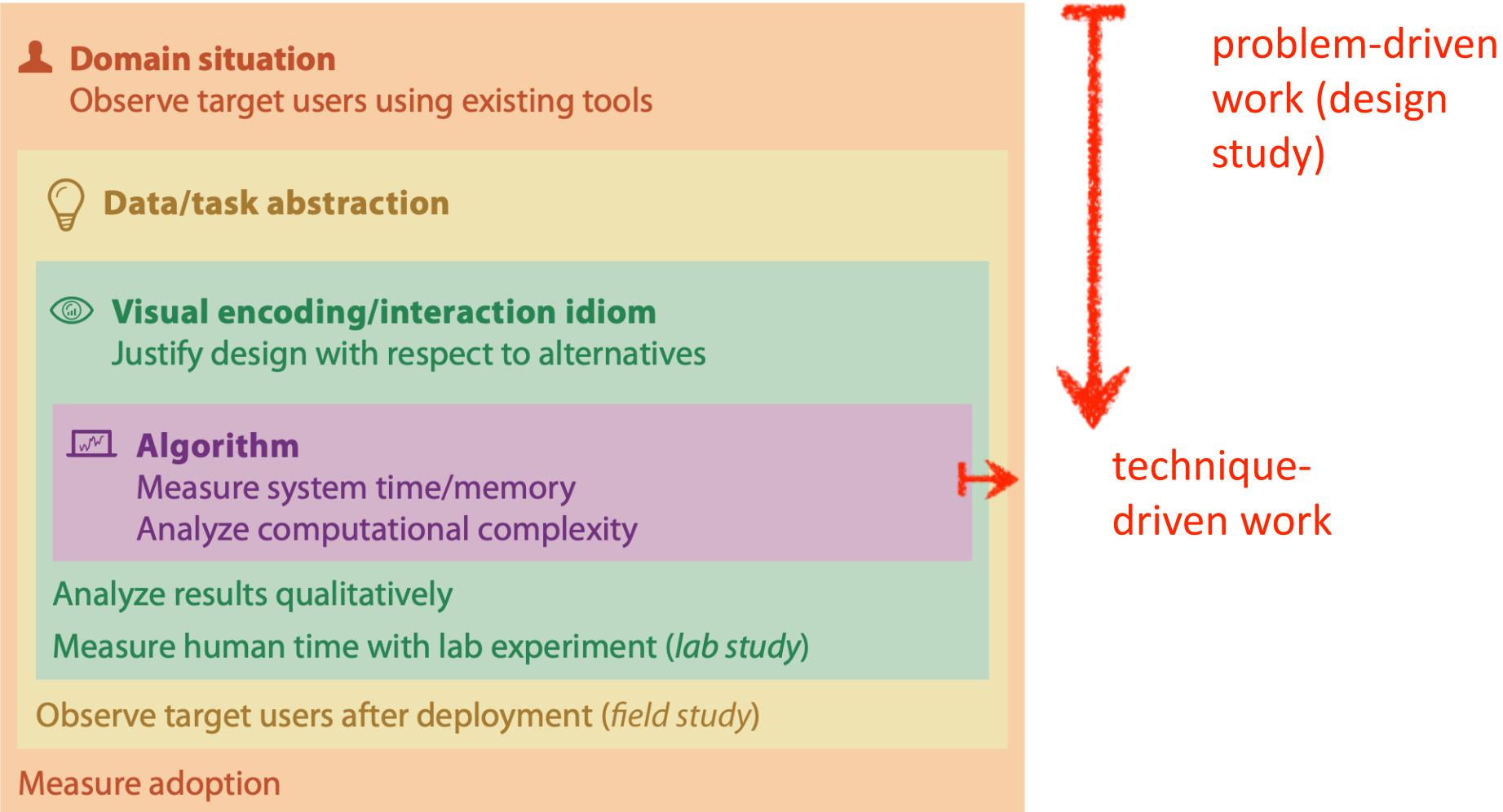
anthropology/  
ethnography

design

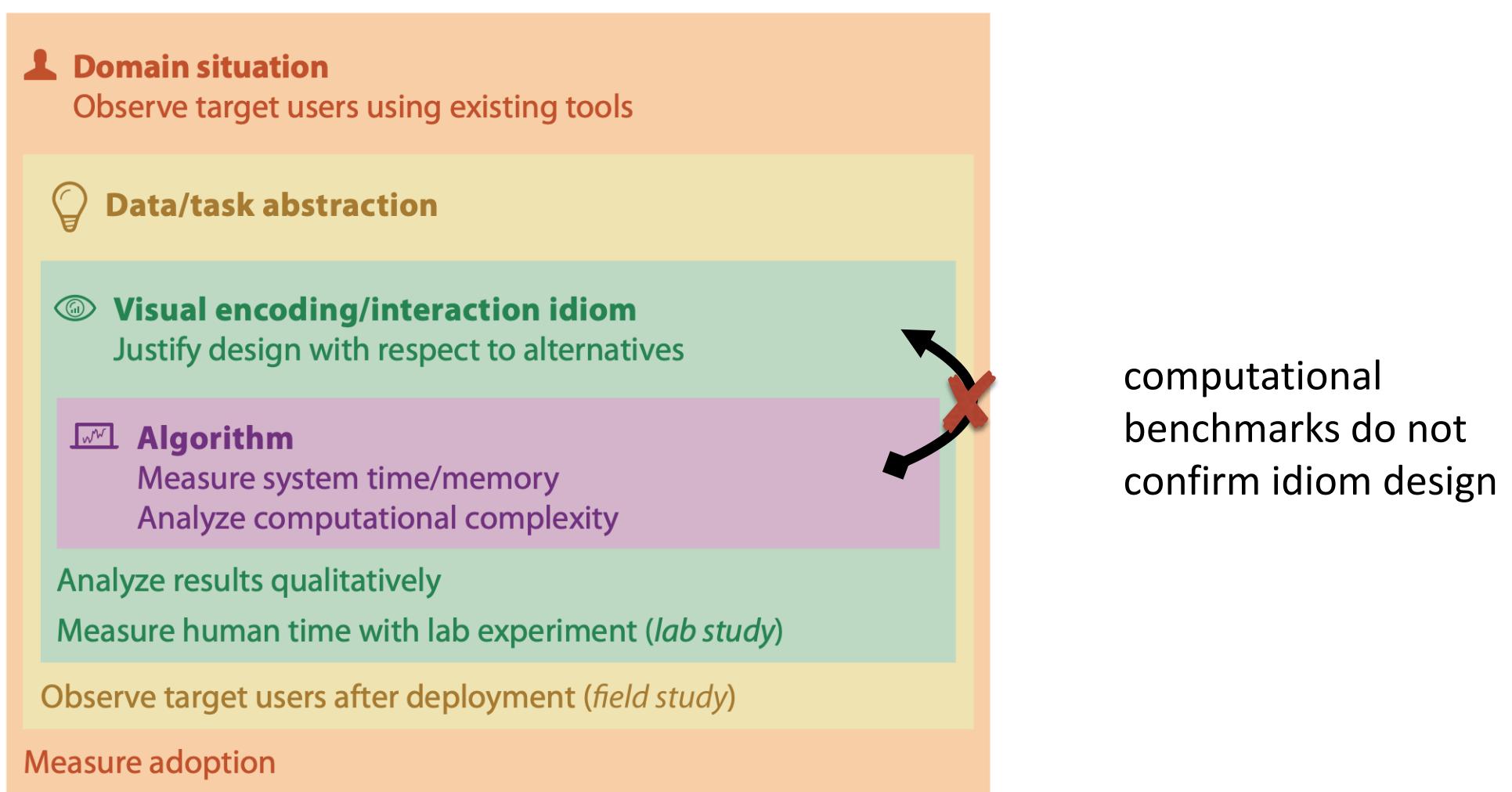
computer  
science

cognitive  
psychology

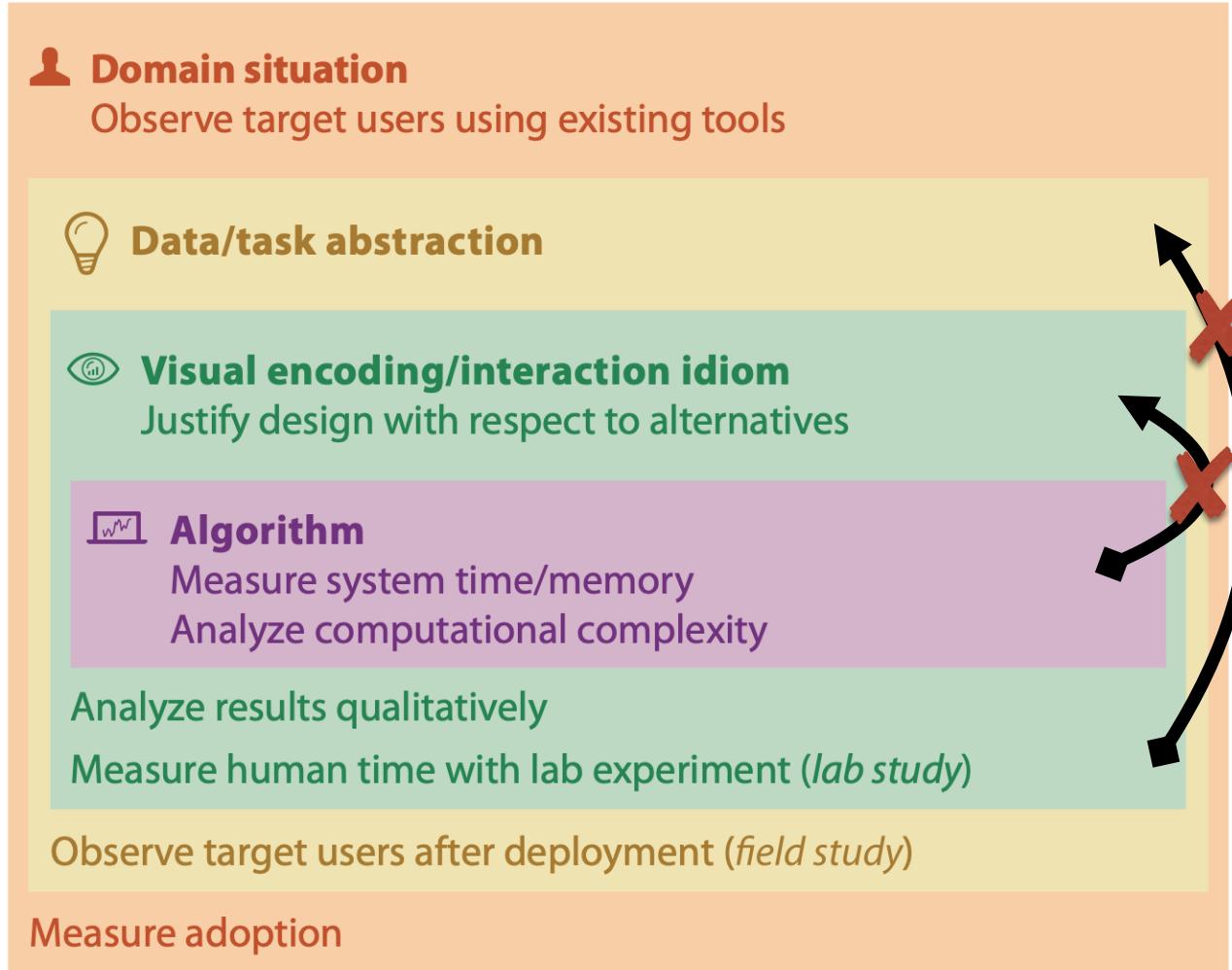
anthropology/  
ethnography



# Avoid mismatches

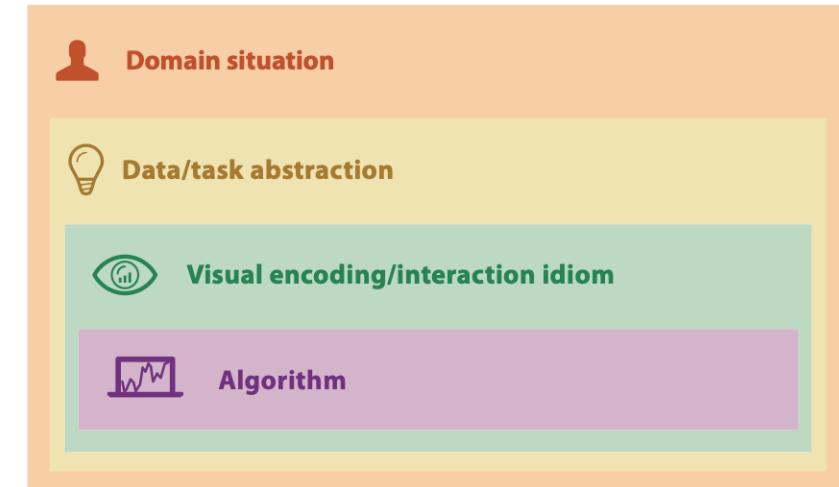


# Avoid mismatches



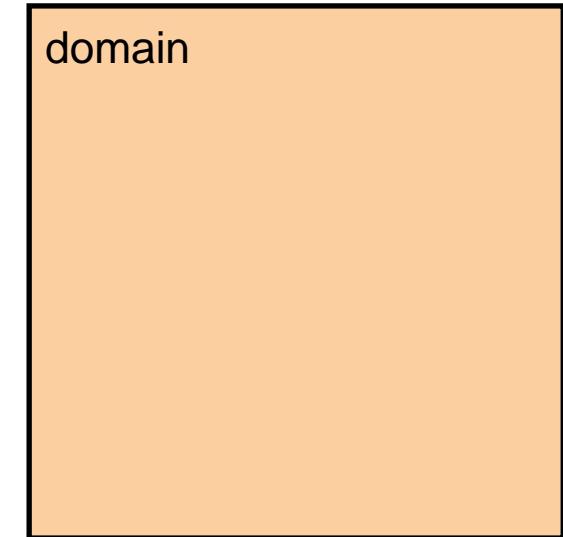
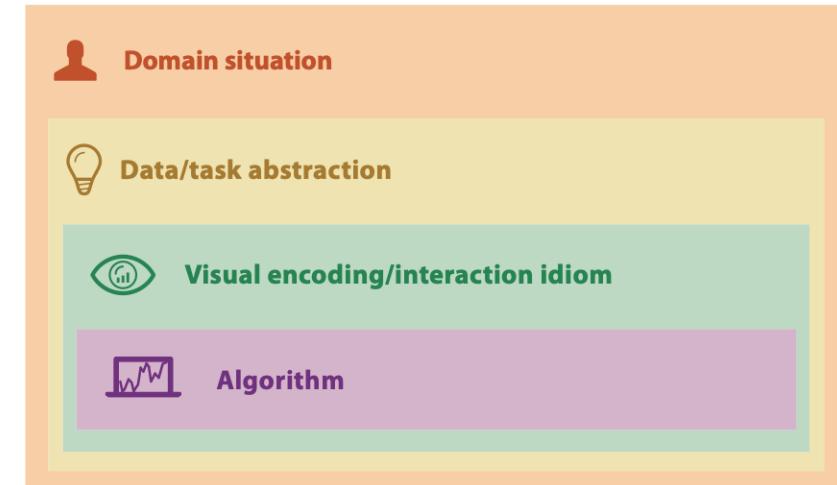
lab studies do not  
confirm task  
abstraction  
computational  
benchmarks do not  
confirm idiom design

# From domain to abstraction



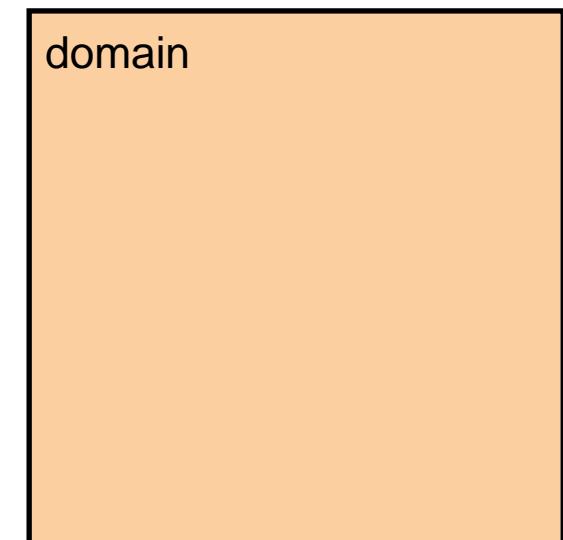
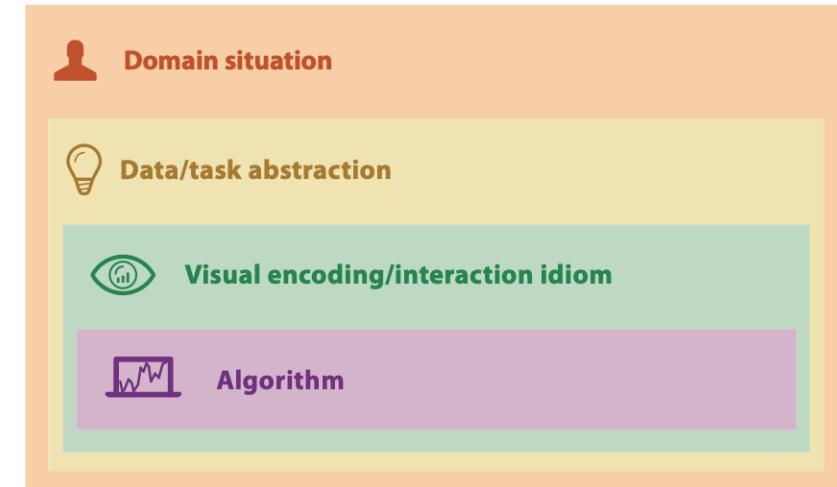
# From domain to abstraction

- domain characterization:  
details of application domain



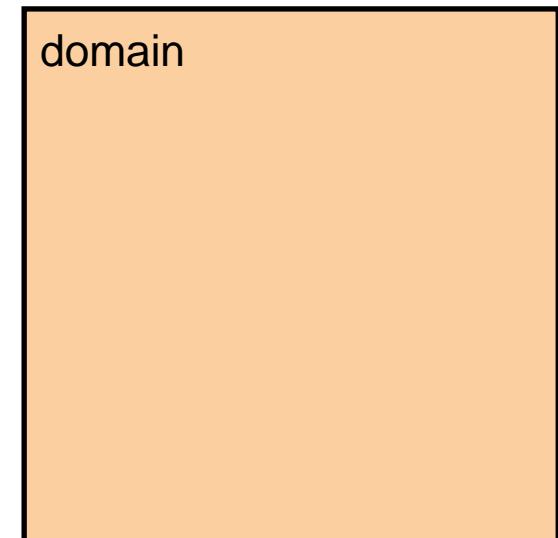
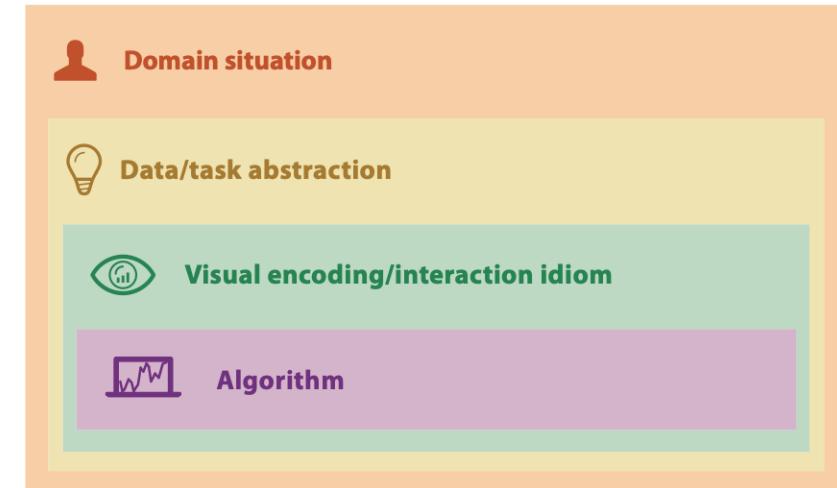
# From domain to abstraction

- domain characterization:  
details of application domain
  - group of users, target domain, their questions & data
    - varies wildly by domain
    - must be specific enough to get traction



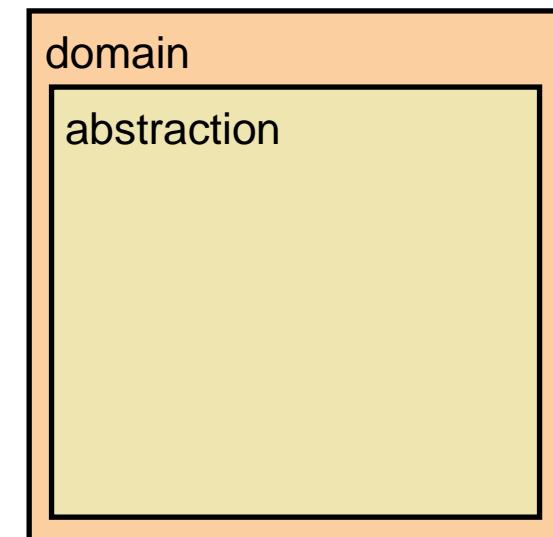
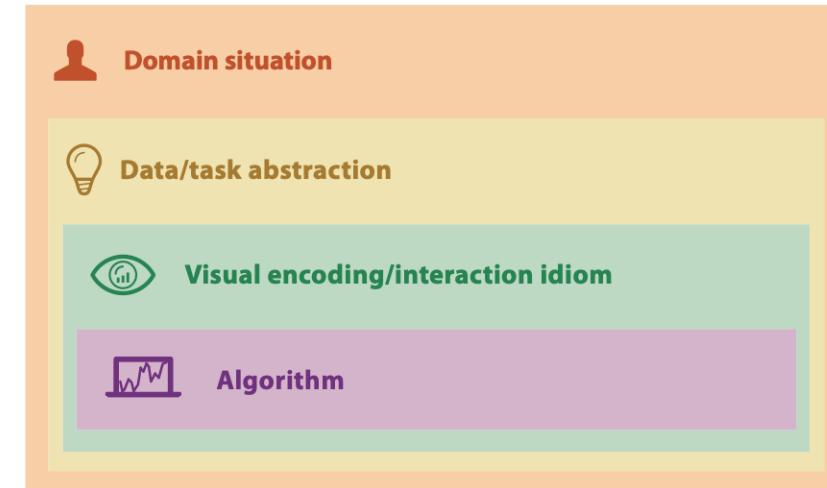
# From domain to abstraction

- domain characterization:  
details of application domain
  - group of users, target domain, their questions & data
    - varies wildly by domain
    - must be specific enough to get traction
  - domain questions/problems
    - break down into simpler abstract tasks



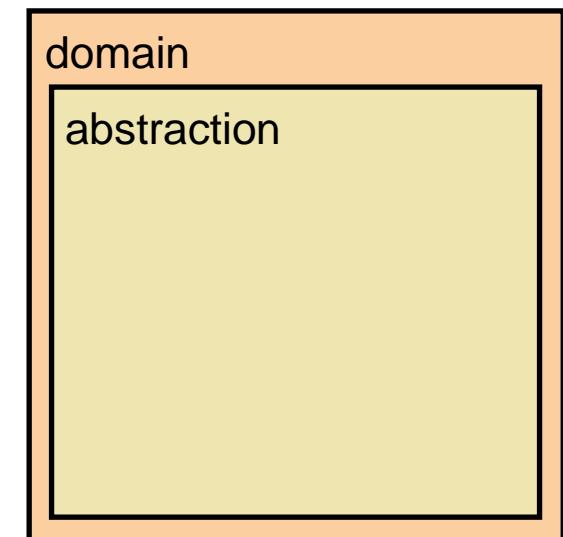
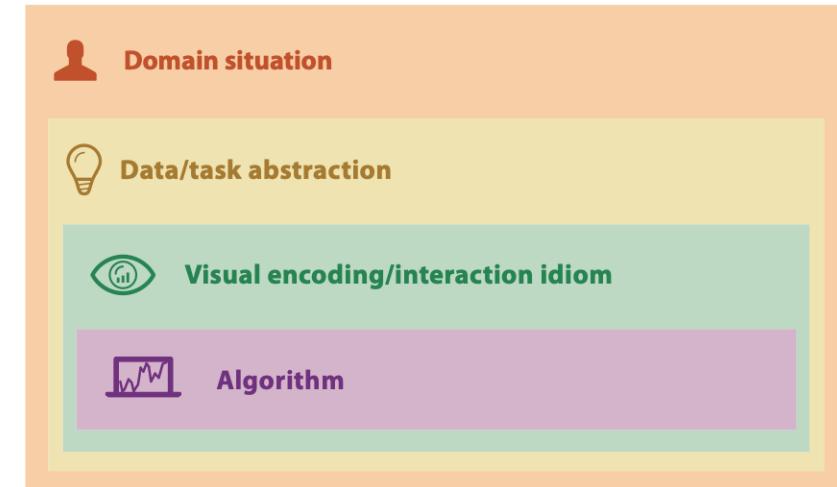
# From domain to abstraction

- domain characterization:  
details of application domain
  - group of users, target domain, their questions & data
    - varies wildly by domain
    - must be specific enough to get traction
  - domain questions/problems
    - break down into simpler abstract tasks
- abstraction: data & task
  - map *what* and *why* into generalized terms



# From domain to abstraction

- domain characterization:  
details of application domain
  - group of users, target domain, their questions & data
    - varies wildly by domain
    - must be specific enough to get traction
  - domain questions/problems
    - break down into simpler abstract tasks
- abstraction: data & task
  - map *what* and *why* into generalized terms
    - identify tasks that users wish to perform, or already do
    - find data types that will support those tasks
      - possibly transform /derive if need be



# Task abstraction: Actions and targets

- very high-level pattern
- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

# Task abstraction: Actions and targets

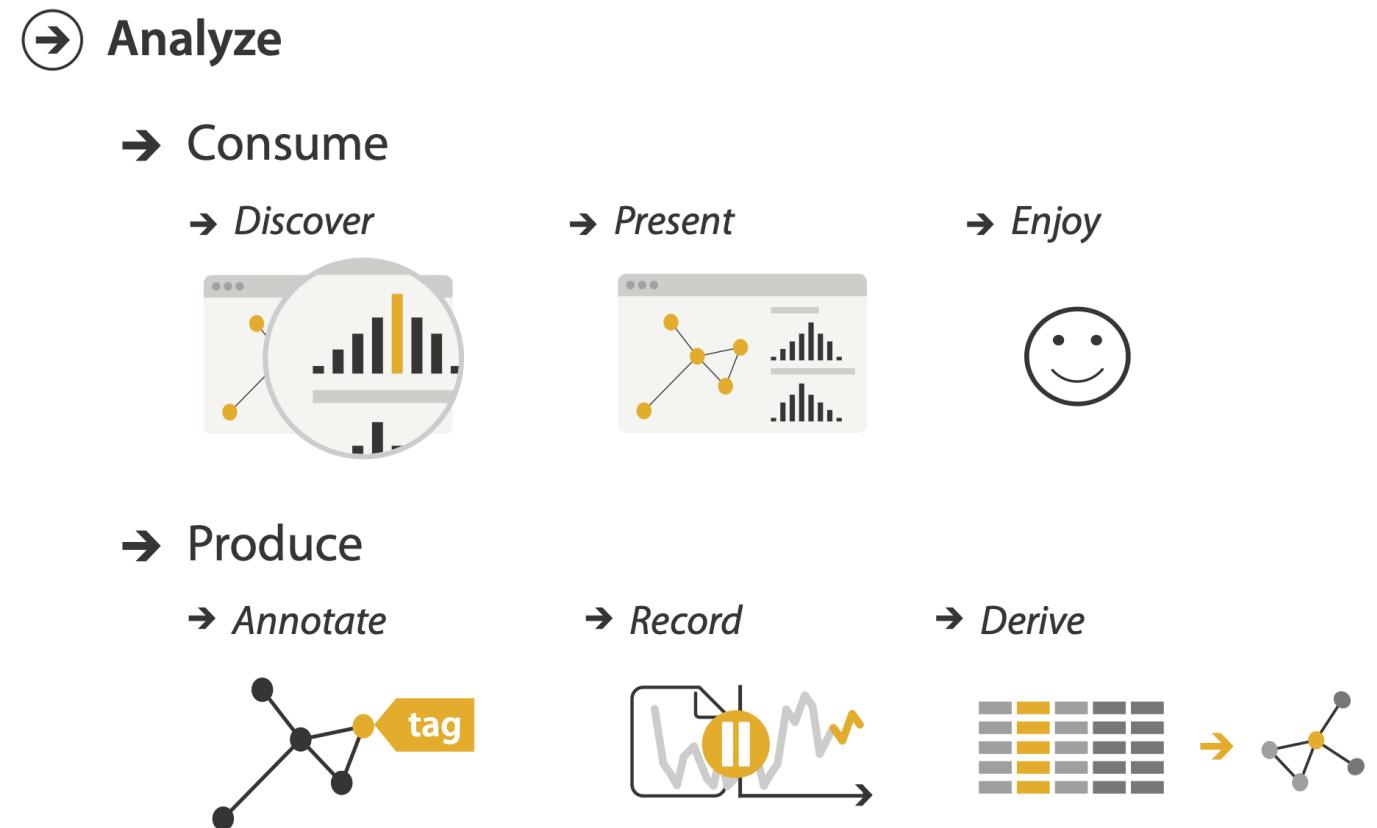
- very high-level pattern
- actions
  - analyze
    - high-level choices
  - search
    - find a known/unknown item
  - query
    - find out about characteristics of item
- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

# Task abstraction: Actions and targets

- very high-level pattern
- actions
  - analyze
    - high-level choices
  - search
    - find a known/unknown item
  - query
    - find out about characteristics of item
- targets
  - what is being acted on
- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

# Actions: Analyze

- consume
  - discover vs present
    - classic split
    - aka explore vs explain
  - enjoy
    - newcomer
    - aka casual, social
- produce
  - annotate, record
  - derive
    - crucial design choice



# Actions: Search

# Actions: Search

- what does user know?  
—target, location

## ➔ Search

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

# Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order

## ➔ Search

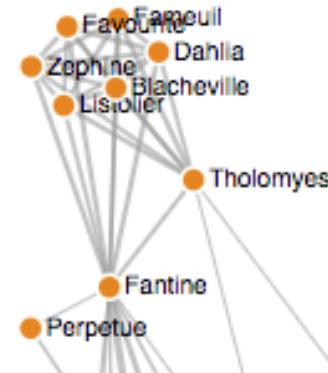
	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>

# Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order
- locate
  - ex: keys in your house
  - ex: node in network

## ➔ Search

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>



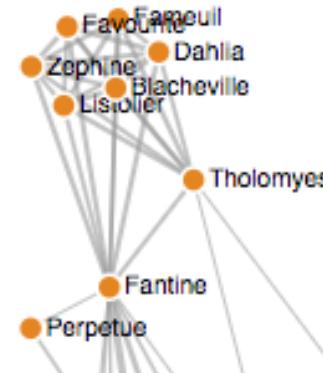
<https://bl.ocks.org/hevignick/3faf257bbbbc7743bb72310d03b86ee8>

# Actions: Search

- what does user know?
  - target, location
- lookup
  - ex: word in dictionary
    - alphabetical order
- locate
  - ex: keys in your house
  - ex: node in network
- browse
  - ex: books in bookstore

## ➔ Search

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>



<https://bl.ocks.org/hevignick/3faf257bbbbc7743bb72310d03b86ee8>

# Actions: Search

- what does user know? → **Search**

- target, location

- lookup

- ex: word in dictionary
    - alphabetical order

- locate

- ex: keys in your house
  - ex: node in network

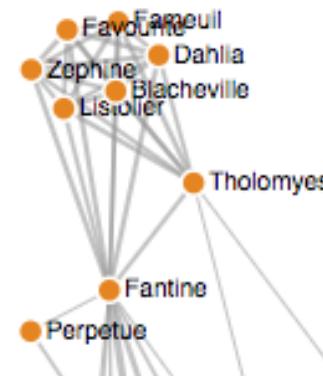
- browse

- ex: books in bookstore

- explore

- ex: find cool neighborhood in new city

	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>



<https://bl.ocks.org/hevignick/3faf257bbbbc7743bb72310d03b86ee8>

# Actions: Query

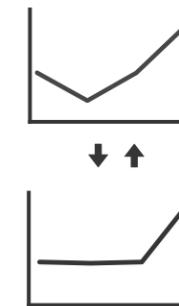
- how much of the data matters?
  - one: identify
  - some: compare
  - all: summarize

➔ **Query**

➔ Identify



➔ Compare



➔ Summarize



# Actions

- independent choices for each of these three levels
  - analyze, search, query
  - mix and match

## Actions

### → Analyze

→ Consume



→ Produce



### → Search

	Target known	Target unknown
Location known	•..• <i>Lookup</i>	•..• <i>Browse</i>
Location unknown	<•○•> <i>Locate</i>	<•○•> <i>Explore</i>

### → Query

→ Identify



→ Compare



→ Summarize



# Task abstraction: Targets

# Task abstraction: Targets

→ All Data

→ Trends



→ Outliers



→ Features



# Task abstraction: Targets

## → All Data

→ Trends      → Outliers      → Features



## → Attributes

→ One      → Many

→ Distribution



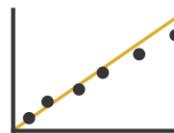
→ Extremes



→ Dependency



→ Correlation



→ Similarity



# Task abstraction: Targets

## → All Data

→ Trends



→ Outliers

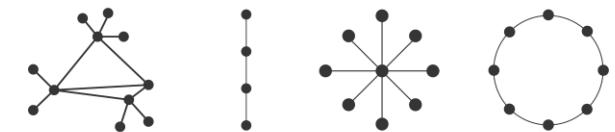


→ Features



## → Network Data

→ Topology



→ Paths



## → Attributes

→ One



→ Many

→ Distribution

→ Dependency

→ Correlation

→ Similarity

→ Extremes



# Task abstraction: Targets

## → All Data

→ Trends



→ Outliers



→ Features



## → Attributes

→ One

→ Distribution



→ Extremes

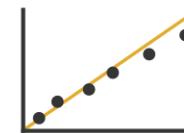


→ Many

→ Dependency



→ Correlation

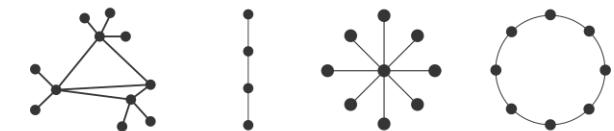


→ Similarity



## → Network Data

→ Topology



→ Paths



## → Spatial Data

→ Shape



# Summary: Goals of Visualization

- Presentation
- Confirmation
- Exploration

# Summary: Goals of Visualization

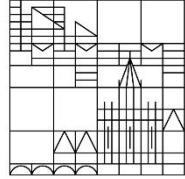
- Presentation
  - Starting point: facts to be presented are fixed a priori.
  - Process: choice of appropriate presentation techniques
  - Result: high-quality visualization of the data to present facts
- Confirmation
- Exploration

# Summary: Goals of Visualization

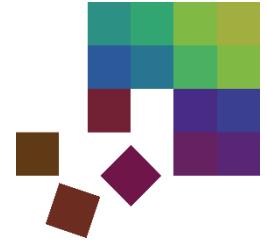
- Presentation
- Confirmation
  - Starting point: hypotheses about the data
  - Process: goal-oriented examination of the hypotheses
  - Result: visualization of data to confirm or reject the hypotheses
- Exploration

# Summary: Goals of Visualization

- Presentation
- Confirmation
- Exploration
  - Starting point: no hypotheses about the data
  - Process: interactive, usually undirected search for structures, trends
  - Result: visualization of data to lead to hypotheses about the data



University of Konstanz  
Data Analysis and Visualization Group



# Break

Thanks for listening

# Practice: Design your own visualization!

Get into teams of three and start the discussion!

## Todos

- Select a text dataset of your choice
  - You could use your own text data set of interest or look into the github repo

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]

# Datasets

- *Script of the Presidential Debate between Trump and Harris from the 11.09.2024 as retrieved from CNN* (<https://www.cnn.com/2024/06/27/politics/read-biden-trump-debate-rush-transcript/index.html>)
- *Reports from the Russian-Ukrainian War from the 17.08 – 17.09.2024* (<https://acleddata.com/about-acled/>)
- *The complete text of Harry Potter by JK Rowling*

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]

# Practice: Design your own visualization!

Get into teams of three and start the discussion!

## Todos

- Select a text dataset of your choice
  - You could use your own text data set of interest or look into the github repo
- Choose your domain situation
  - Who could be the target user of your visualization?
  - What is the persons motivation?
- How could your Visual Encoding look like?
  - Sketch the design! You can use Pen and Paper, Powerpoint or Excalidraw (<https://excalidraw.com/>)
  - Go to <https://www.data-to-viz.com/>, <https://observablehq.com/@d3/gallery> and <https://d3js.org/> and do some research

# A world of possibilities



# Practice: Design your own visualization!

Get into teams of three and start the discussion!

## Todos

- Select a text dataset of your choice
  - You could use your own text data set of interest or look into the github repo
- Choose your domain situation
  - Who could be the target user of your visualization?
  - What is the persons motivation?
- How could your Visual Encoding look like?
  - Sketch the design! You can use Pen and Paper, Powerpoint or Excalidraw (<https://excalidraw.com/>)
  - Go to <https://www.data-to-viz.com/>, <https://observablehq.com/@d3/gallery> and <https://d3js.org/> and do some research
- What are your text processing steps?
  - List your NLP tasks.

We'll have a final discussion on your visualization designs.

[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013). ]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009). ]