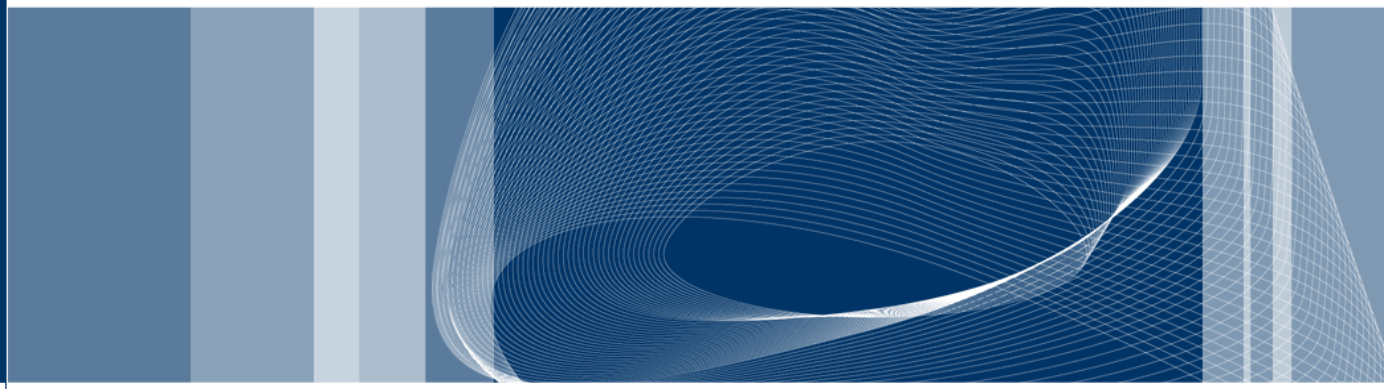
POLITECNICO DI MILANO



Unit 1

Data and Results Visualization

Daniele Loiacono

INFORMATION

Instructor

- **Daniele Loiacono**
  - Contact: daniele.loiacono@polimi.it
  - Office: DEIB, room **150** (over your head ☺)
  - Phone: +39 02 2399 **3615**

- Research interests
  - Data Mining and Machine Learning
  - Games and Immersive Technologies

- Materials
  - https://github.com/DanieleLoiacono/DRViz2018
  - http://www.datacamp.com/  DataCamp

Logistics

- When?
  - 21/5/2018, 9 – 13, Sala Conferenze
  - 23/5/2018, 9 – 13, Sala Conferenze
  - *24/5/2018, 9 – 13, Aula Alpha*
  - 28/5/2018, 9 – 13, Sala Conferenze
  - 29/5/2018, 14 – 18, Sala Conferenze
  - 30/5/2018, 9 – 13, Sala Conferenze
  - *(Opt2) 30/5/2018, 14 – 18, Sala Conferenze*
  - *(Opt3) 31/5/2018, 9 – 13, Sala Conferenze*

Clear?

- How lectures are organized?
  - Start around 9:10 (few minutes for setup and latecomers)
  - Two chunks (1h30m / 1h40m) of lectures
  - A 20/30m break in the middle

Logistics

- When?

Mon	Tue	Wed	Thu	Fri
21/5 AM	22/5	23/5 AM	24/5 AM*	25/5
28/5 AM	29/5 PM	30/5 AM / PM	31/5 AM	1/6
AM(9-13) or PM(14-18) - Sala Conferenze Ed 20 *Aula Alpha Ed. 24				

Confirmed
Alternative 1
Alternative 2
Alternative 3

- How lectures are organized?
  - Start around 9:10 (few minutes for setup and latecomers)
  - Two chunks (1h30m / 1h40m) of lectures
  - A 20/30m break in the middle

How to pass the exam?

- Attending the 70% lectures (~18h)
  - Remember to sign at the beginning and after the break
- Small visualization project
  - Team: 1 or 2 students
  - Subject: either proposed by you or given by instructor
  - Evaluation: report and/or short presentation of results
- Grade: pass or fail

Course topics

- Introduction and data visualization framework
- Tools overview
  - Introduction to Pandas (hands-on)
  - Introduction to Matplotlib (hands-on)
  - Introduction to Seaborn (hands-on)
- Marks and Channels
  - Colormaps in Seaborn / Colorbrewer (hands-on)
- General principles
- Tables visualization(hands-on)
- Spatial data (hands-on)
- Networks and Graphs visualization (hands-on)
- Interactive visualization (hands-on) and dashboard design
- Dealing with complexity
  - Principles
  - Examples (hands-on)
- Validation

Course topics

- Introduction and data visualization framework (21/5)
- Tools overview (21/5)
  - Introduction to Pandas (hands-on)
  - Introduction to Matplotlib (hands-on)
  - Introduction to Seaborn (hands-on)
- Marks and Channels (23/5)
  - Colormaps in Seaborn / Colorbrewer (hands-on)
- General principles (24/5)
- Tables visualization(hands-on) (24/5 – 28/5)
- Spatial data (hands-on) (28/5)
- Networks and Graphs visualization (hands-on) (29/5)
- Interactive visualization (hands-on) and dashboard design (29/5)
- Dealing with complexity (30/5)
  - Principles
  - Examples (hands-on)
- Validation (30/5)

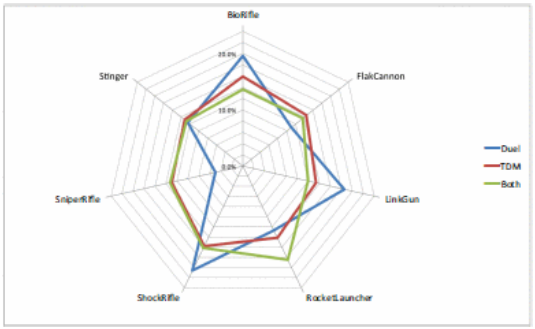
Why this course?

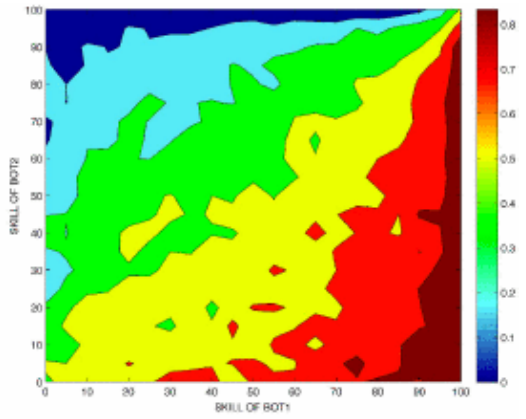
You (will) use data visualization

- Explain data and results…
  - … in lectures
  - … in presentations
  - … in papers
- Analyze data and results…
  - … to explore an unknown problem
  - … to design a new solution of a problem
  - … to assess an existing solution of a problem
- Create a tool to support yours and others decisions.

Disclaimer

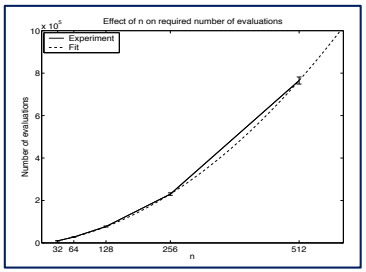
- Data visualization is difficult to evaluate and mistakes are common (lack of time, misjudjment, constraints, etc.).



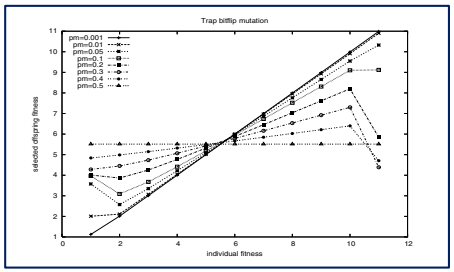


Disclaimer

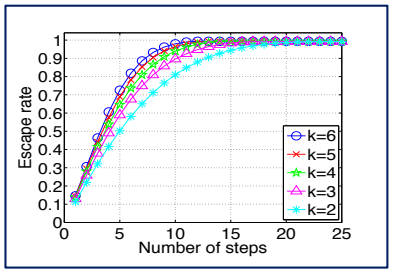
- Data visualization is difficult to evaluate and mistakes are common (lack of time, misjudjment, constraints, etc.).
- Highly problem/domain dependent (also issues keep changing over time).



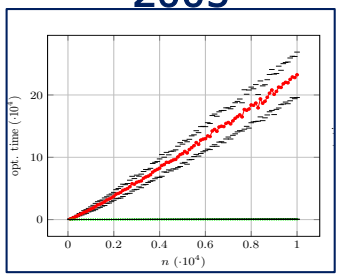
2005



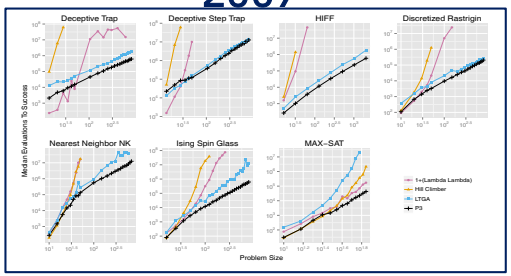
2007



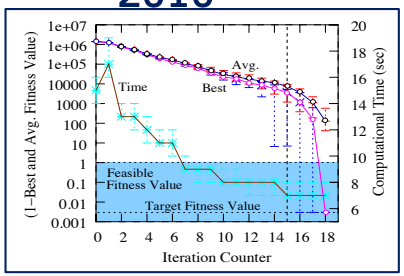
2010



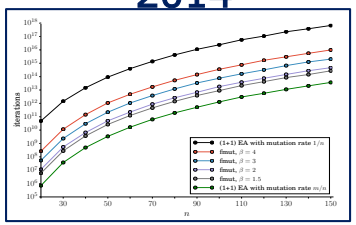
2011



2014



2016

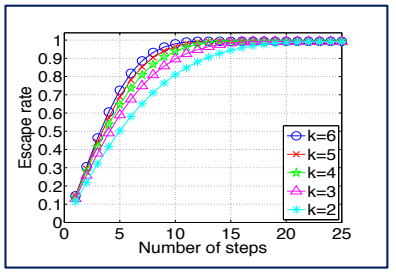


2017

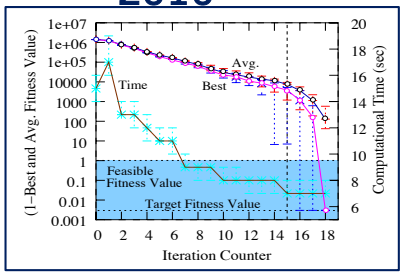
Figures from GECCO best papers in past few years

Disclaimer

- Data visualization is difficult to evaluate and mistakes are common (lack of time, misjudjment, constraints, etc.).
- Highly problem/domain dependent (also issues keep changing over time).



2010



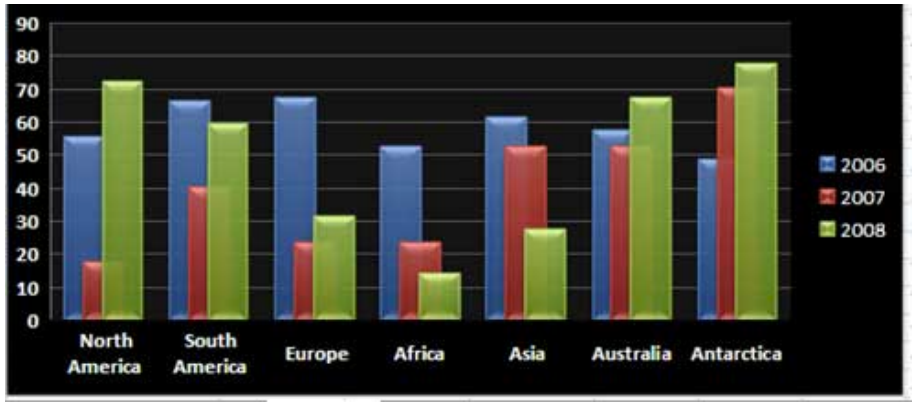
2016

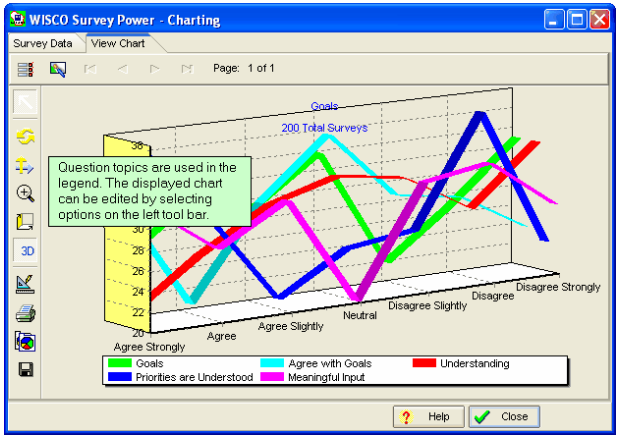
2017

Figures from GECCO best papers in past few years

Disclaimer

- Data visualization is difficult to evaluate and mistakes are common (lack of time, misjudjment, constraints, etc.).
- Highly problem/domain dependent (also issues keep changing over time).
- Probably you won't become a data viz master, but (hopefully) you'll learn to be more effective and (at least) avoid this:

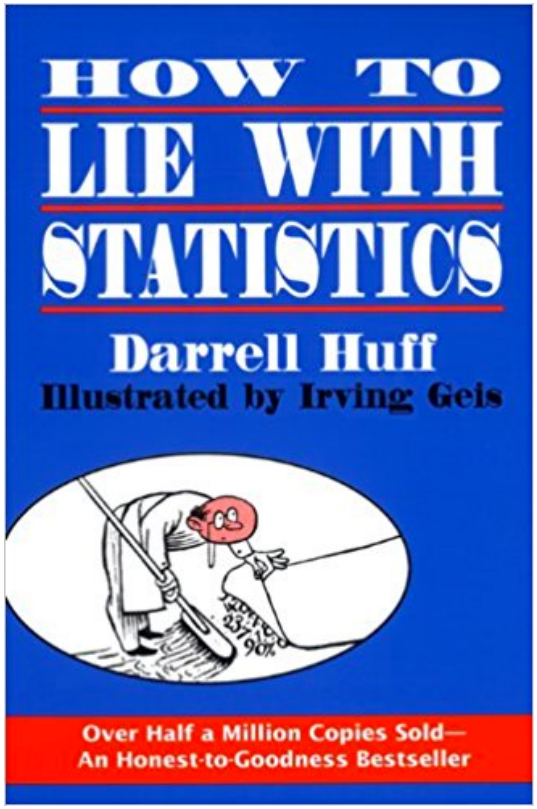


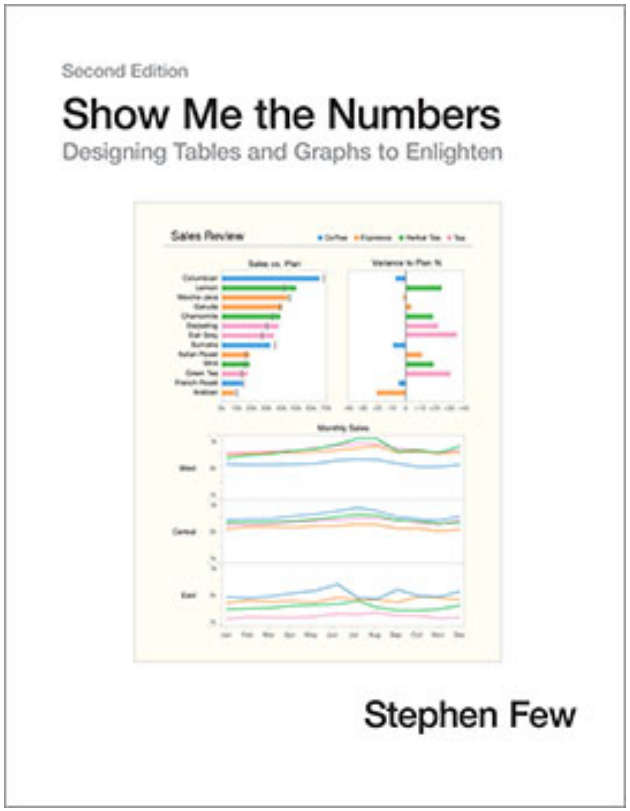


What you will not learn?

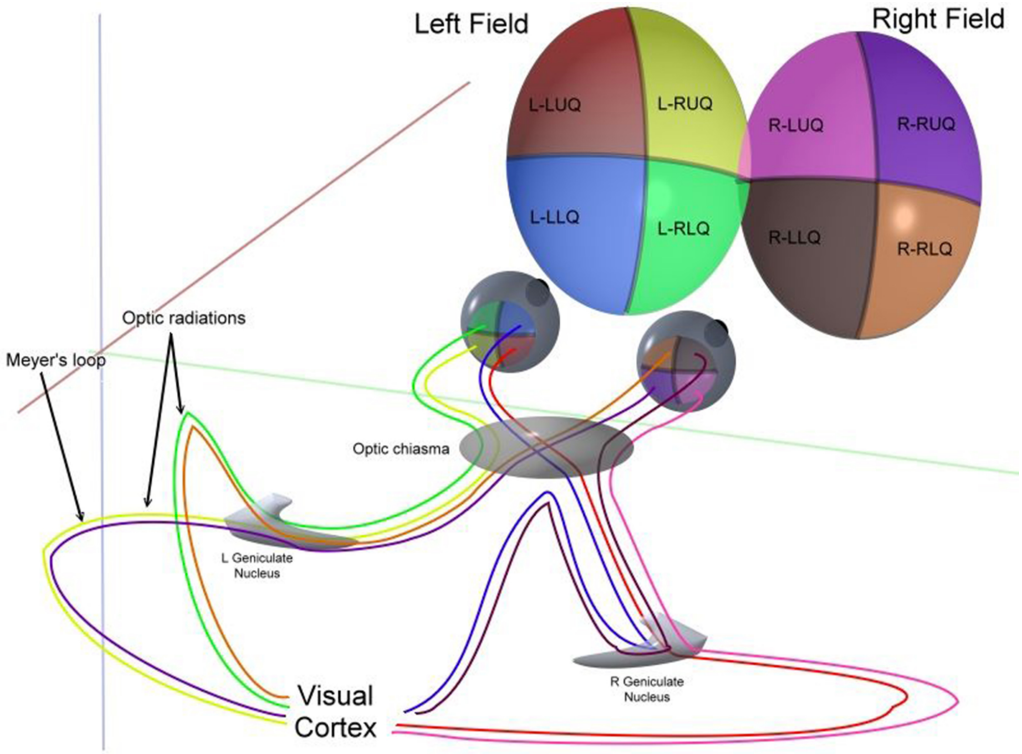
Results presentation

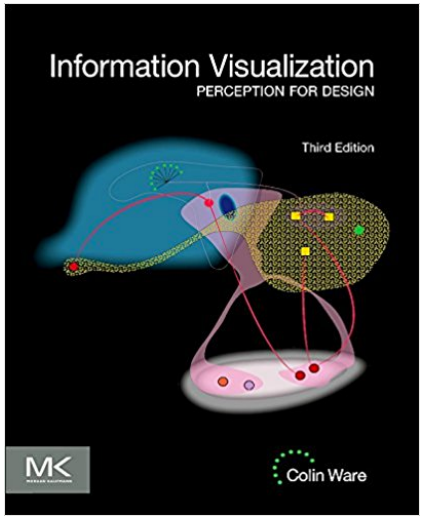
- How to compare and rank results?
- How to design a table?
- What better means? (yes, you do need statistics ☺)



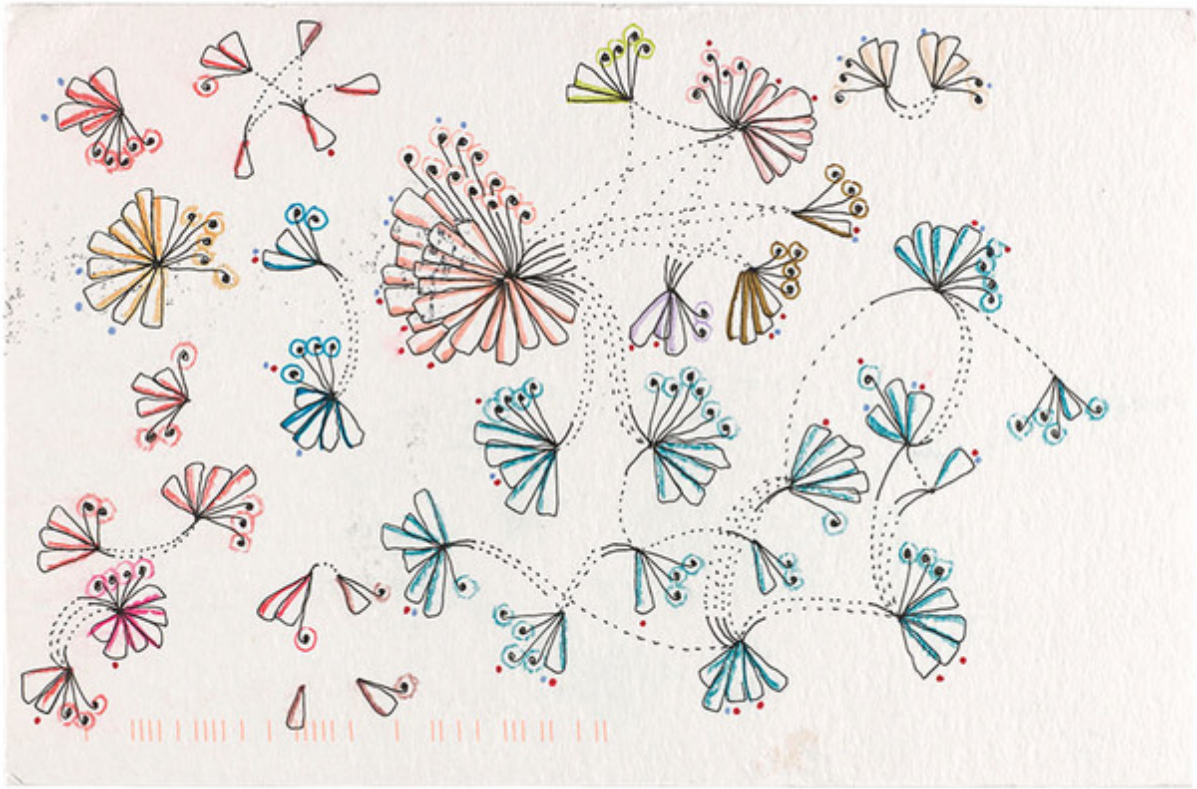


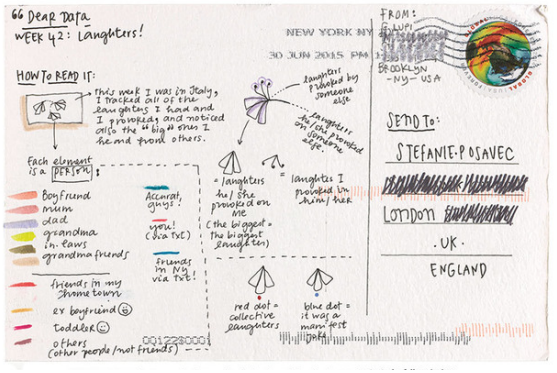
Human visual perception mechanisms





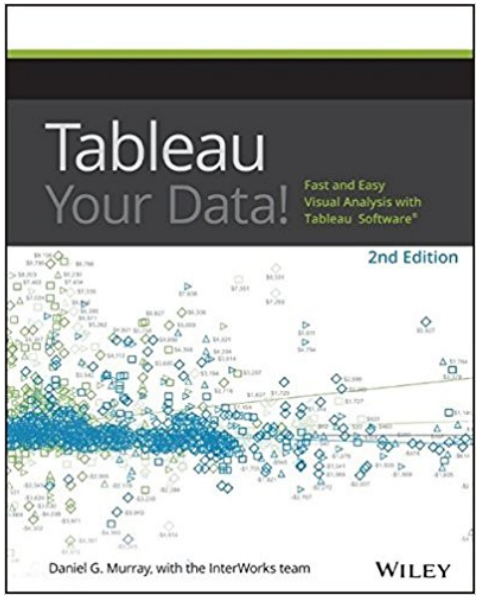
Design beautiful data visualization

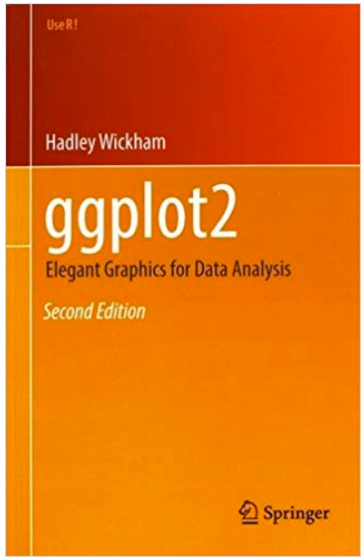


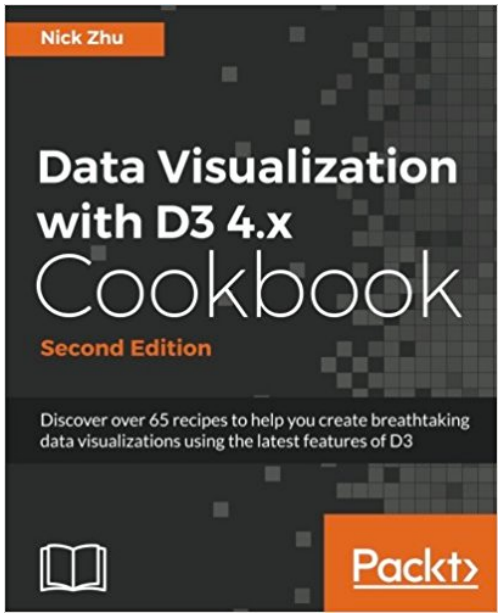


From http://www.dear-data.com/

Master one or more specific tools



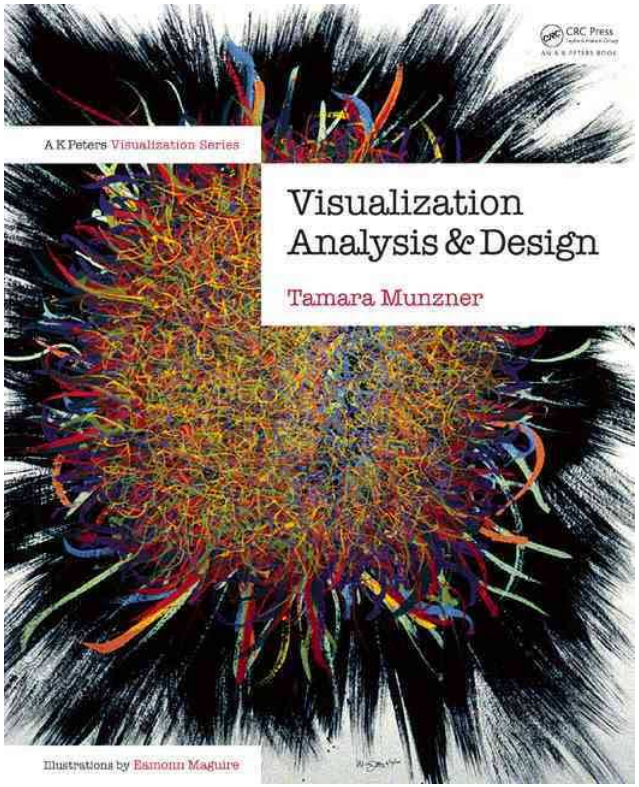


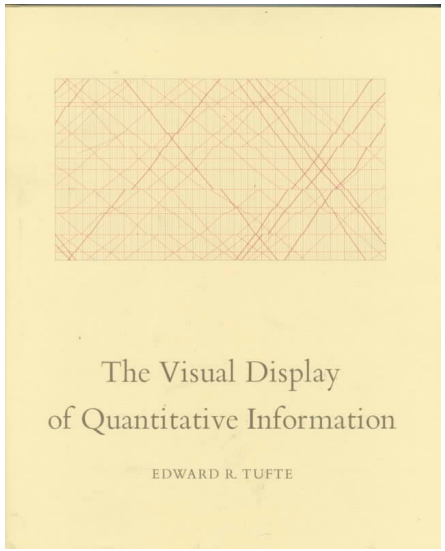


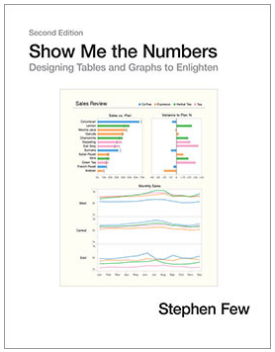
Resources and software

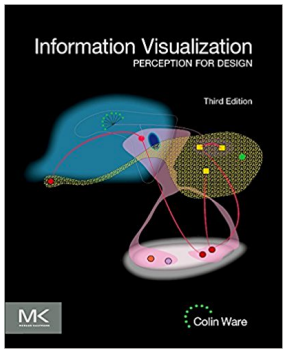
- Web resources
  - https://datavizcatalogue.com/
  - http://junkcharts.typepad.com/
  - https://python-graph-gallery.com/
  - http://www.thefunctionalart.com/
  - http://www.perceptualedge.com/
  - http://chartporn.org/
- Software to install
  - https://www.anaconda.com/ (with Matplotlib, Seaborn, Pandas, Pynum installed)
  - Tableau Public (https://public.tableau.com/)
  - TBA (eventually): more libraries/software

Bibliography









Most of the materials used in this course are based on or taken from Tamara Munzner's book and slides.

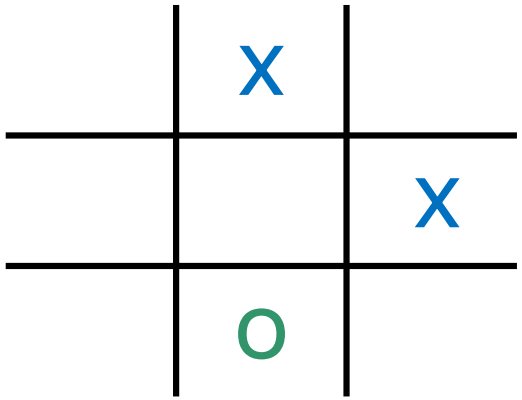
INTRODUCTION

Why data visualization?

Why external representations?

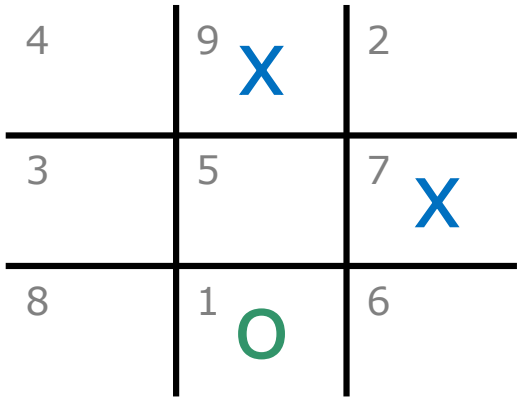
Why vision?

- It is a high-bandwidth channel to our brain!
- An example: the game of 15 (from EuroVis'09 keynote)
  - There are 2 players
  - Each player takes a digit in turn
  - Once a digit is taken, it cannot be used by any of the players again
  - The first player to get three digits that sum to 15 wins
- Familiar?



Why vision?

- It is a high-bandwidth channel to our brain!
- An example: the game of 15 (from EuroVis'09 keynote)
  - There are 2 players
  - Each player takes a digit in turn
  - Once a digit is taken, it cannot be used by any of the players again
  - The first player to get three digits that sum to 15 wins
- Familiar?



Why visualize data?

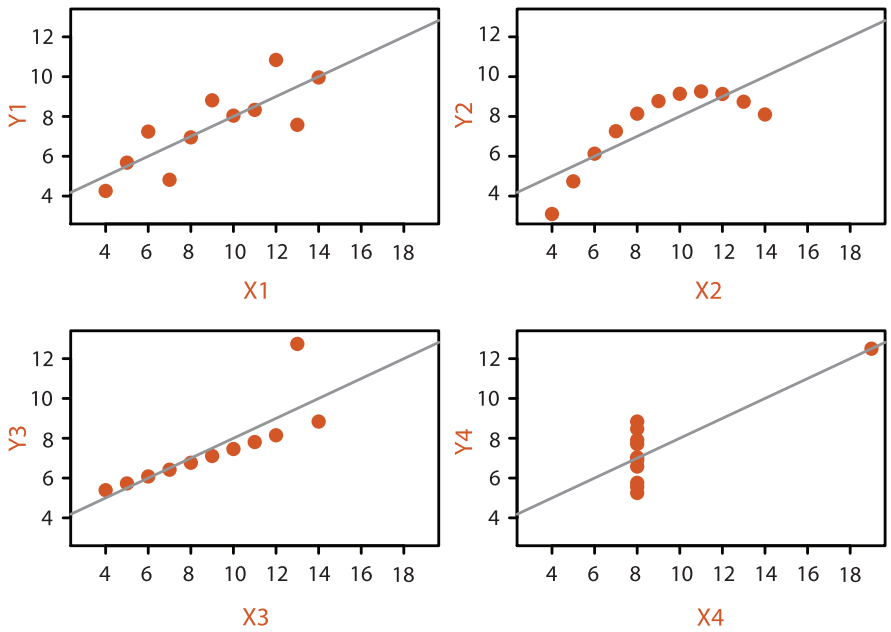
1		2		3		4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	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 visualize data?

	1		2		3		4	
	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
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

Why visualize data?

	1		2		3		4	
	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
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



How to design effective visualization?

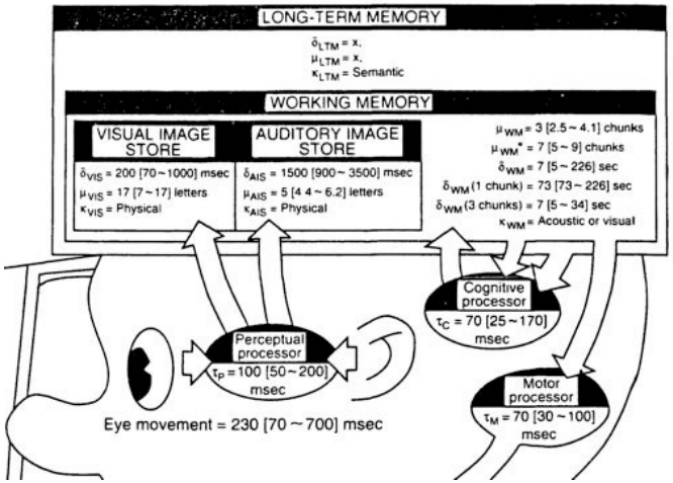
Understand human perception

Define your goal

Explore the design space

Human perception

- How does human perception work?

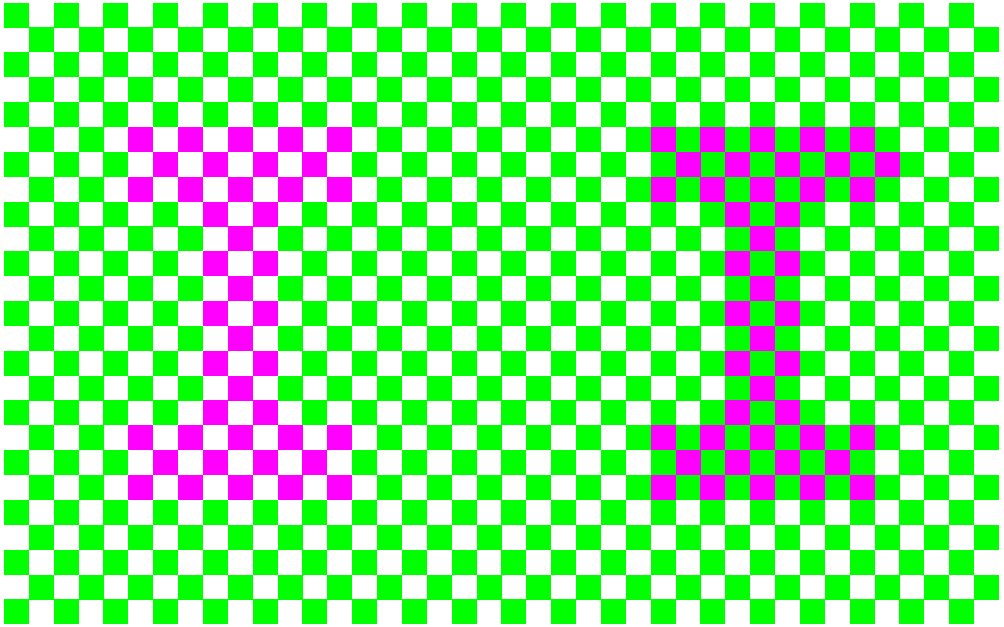


Model Human Processor, Card et al. (1983)

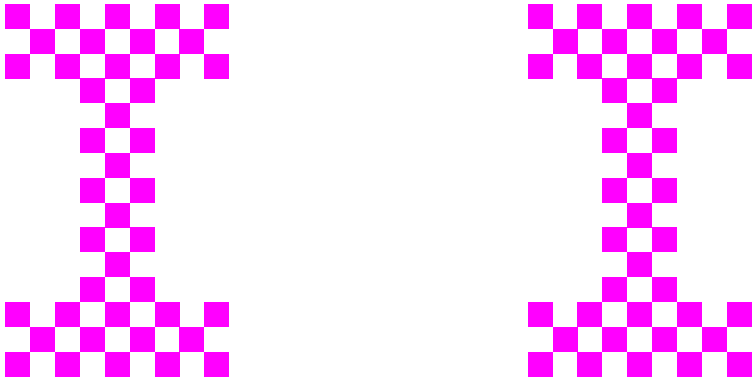
- Major issues
  - Memory and cognitive load
  - Perception and context
  - Change Blindness

Example: memory

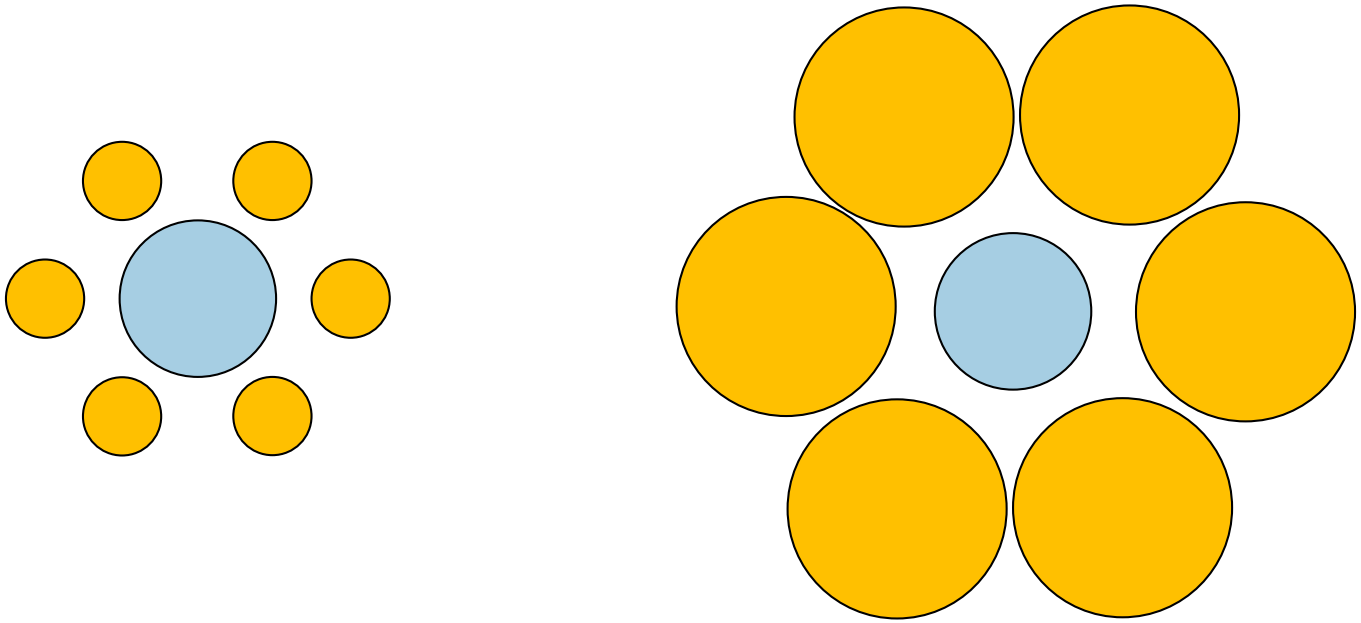
Example: color and context



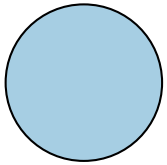
Example: color and context

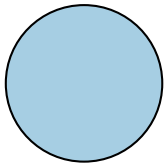


Example: size and context

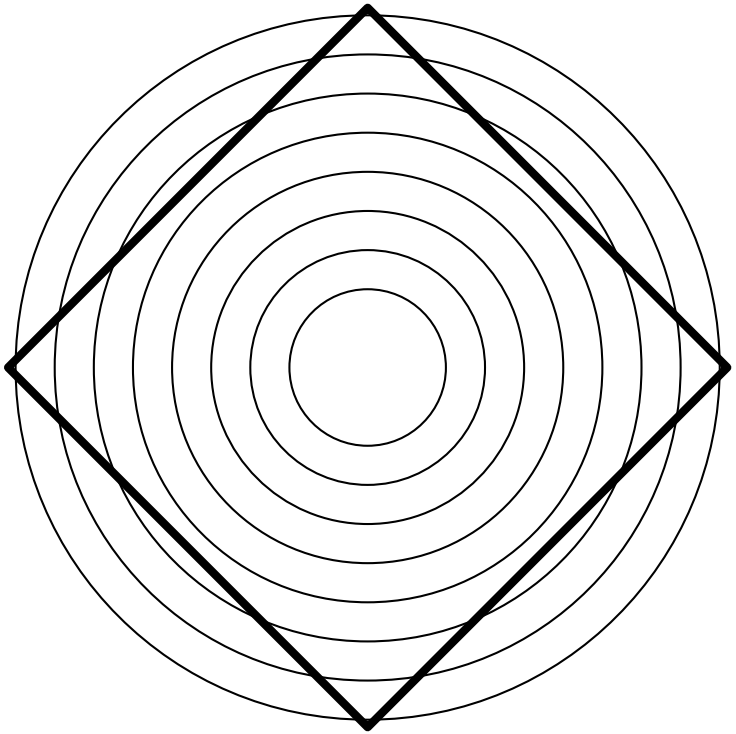


Example: size and context

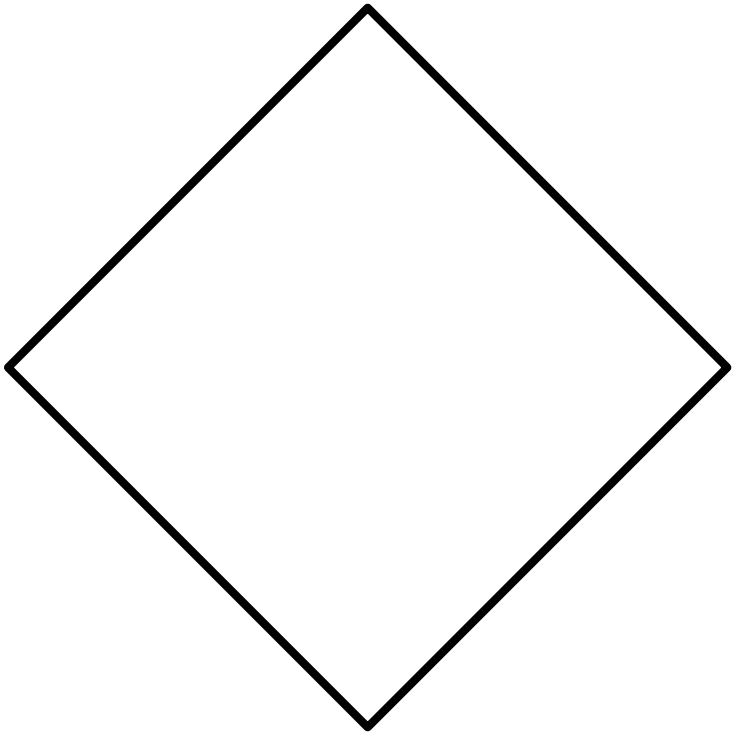




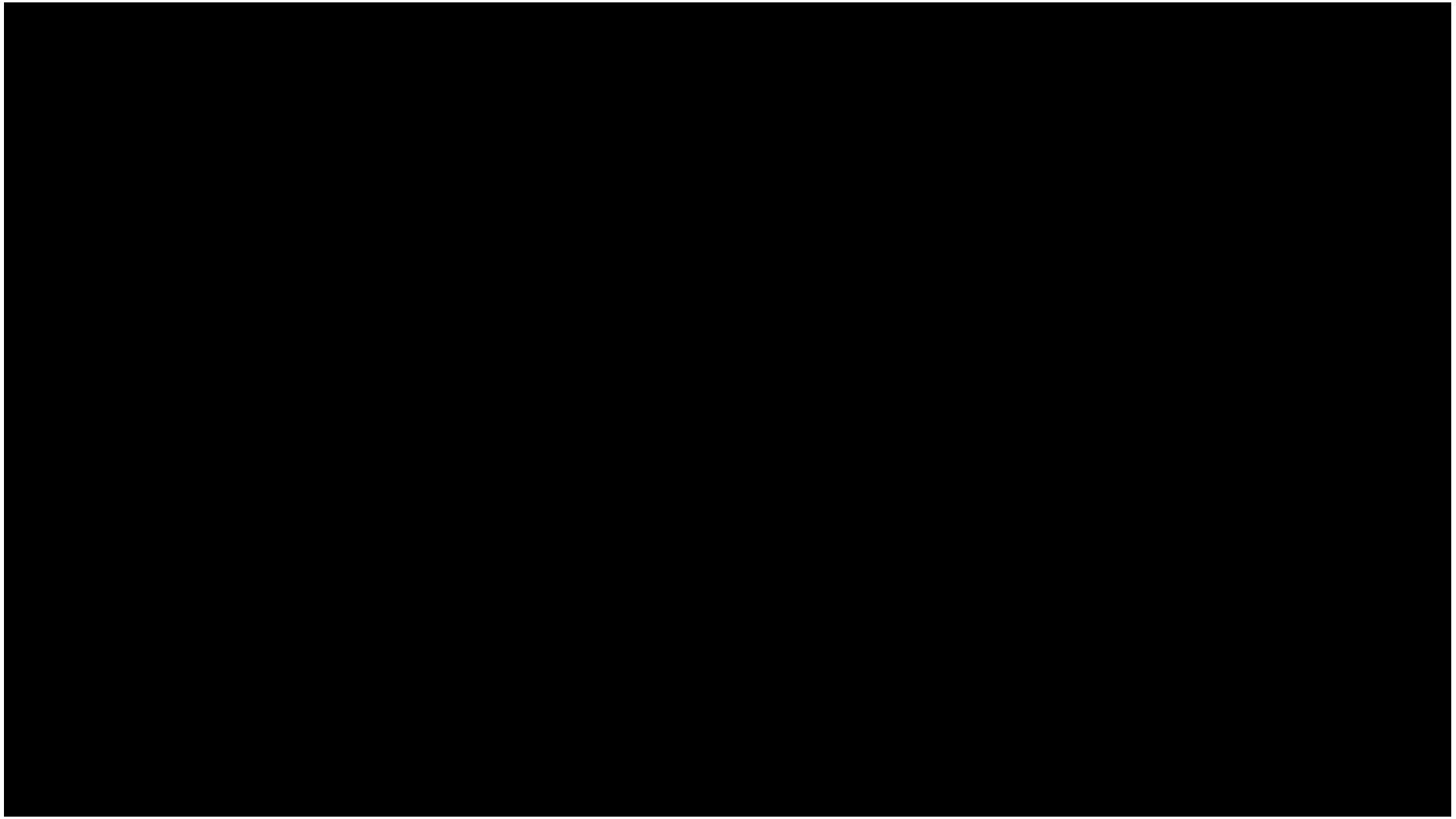
Example: curvature and context



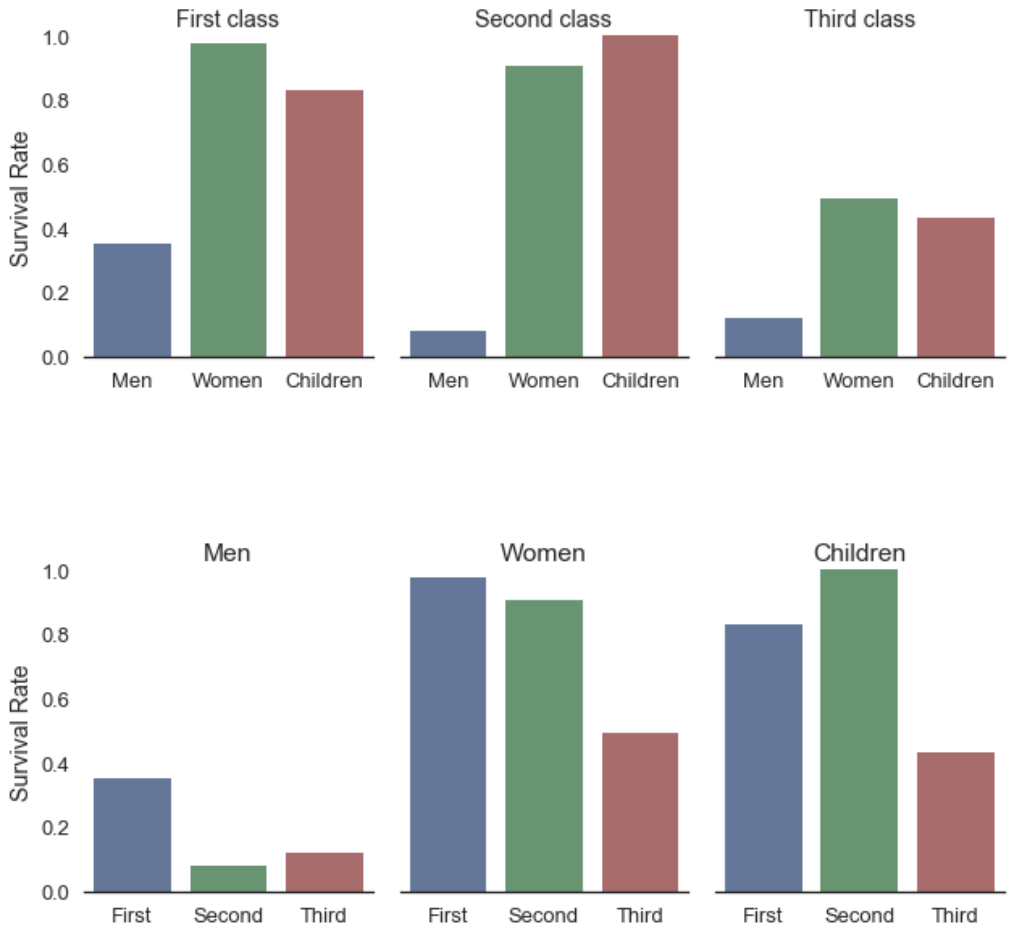
Example: curvature and context



Example: change blindness



Effective?

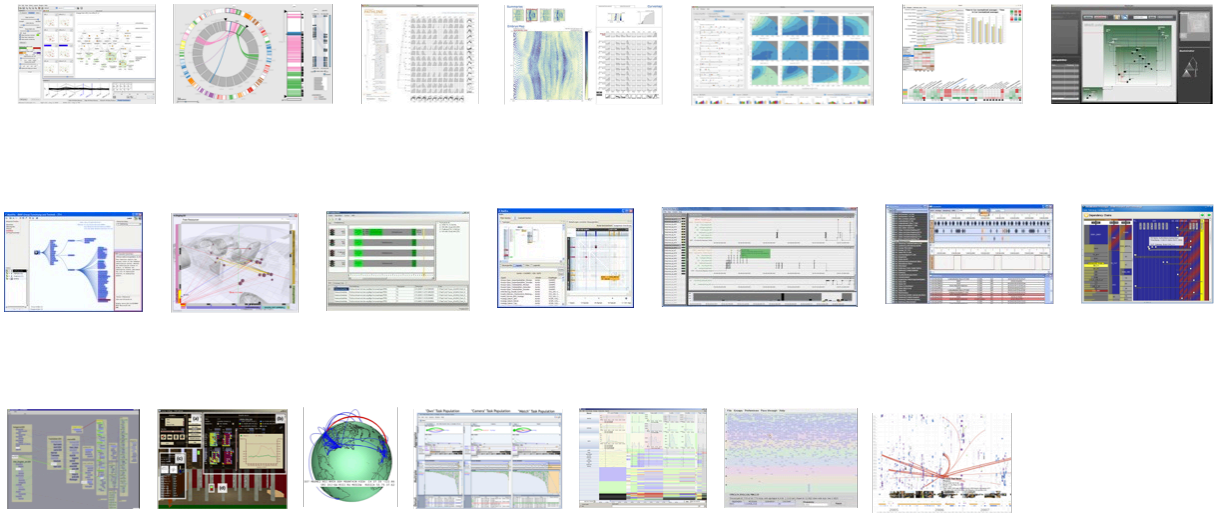

First class
Second class
Third class
Survival Rate
1.0
0.8
0.6
0.4
0.2
0.0
Men
Women
Children
Men
Women
Children
Men
Women
Children
Men
Women
Children
Survival Rate
1.0
0.8
0.6
0.4
0.2
0.0
First
Second
Third
First
Second
Third
First
Second
Third

Effective visualization

- Effectiveness can be measured in terms of:
  - Speed
  - Accuracy
  - Insight
  - Confidence
- Effectiveness can be measured only **in relation** to the **specific** goals and tasks.
- Assessment of a visualization solution is complex.

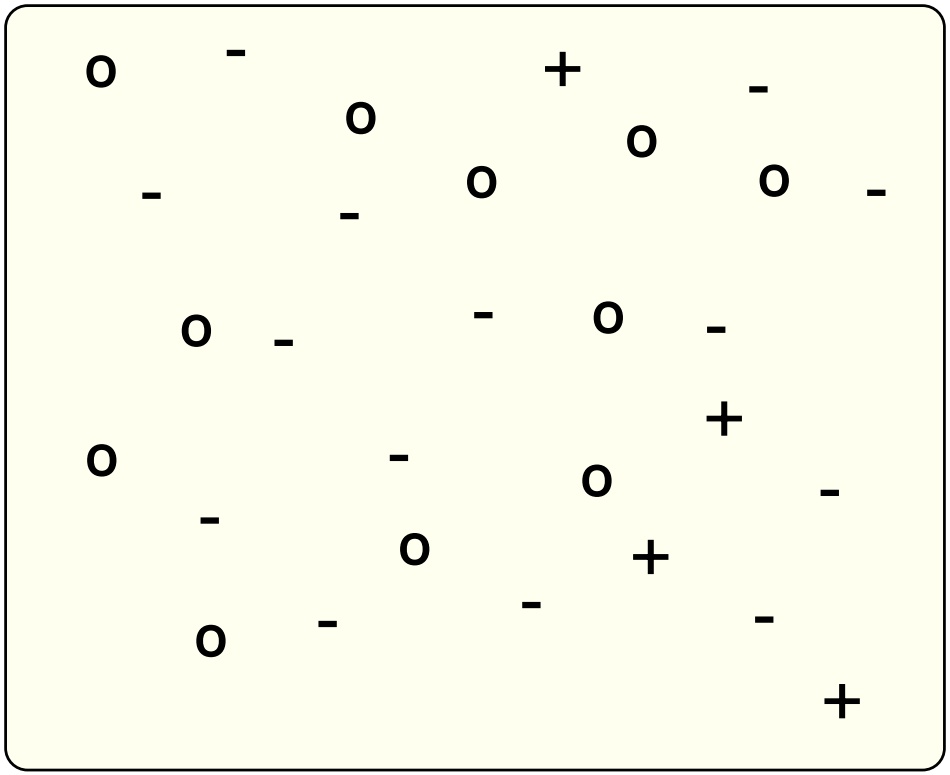
The design space

- The data visualization design space is **huge:**
  - several visual **encodings**
  - several ways to **combine** visualizations
  - several ways to i**nteract** with visualization



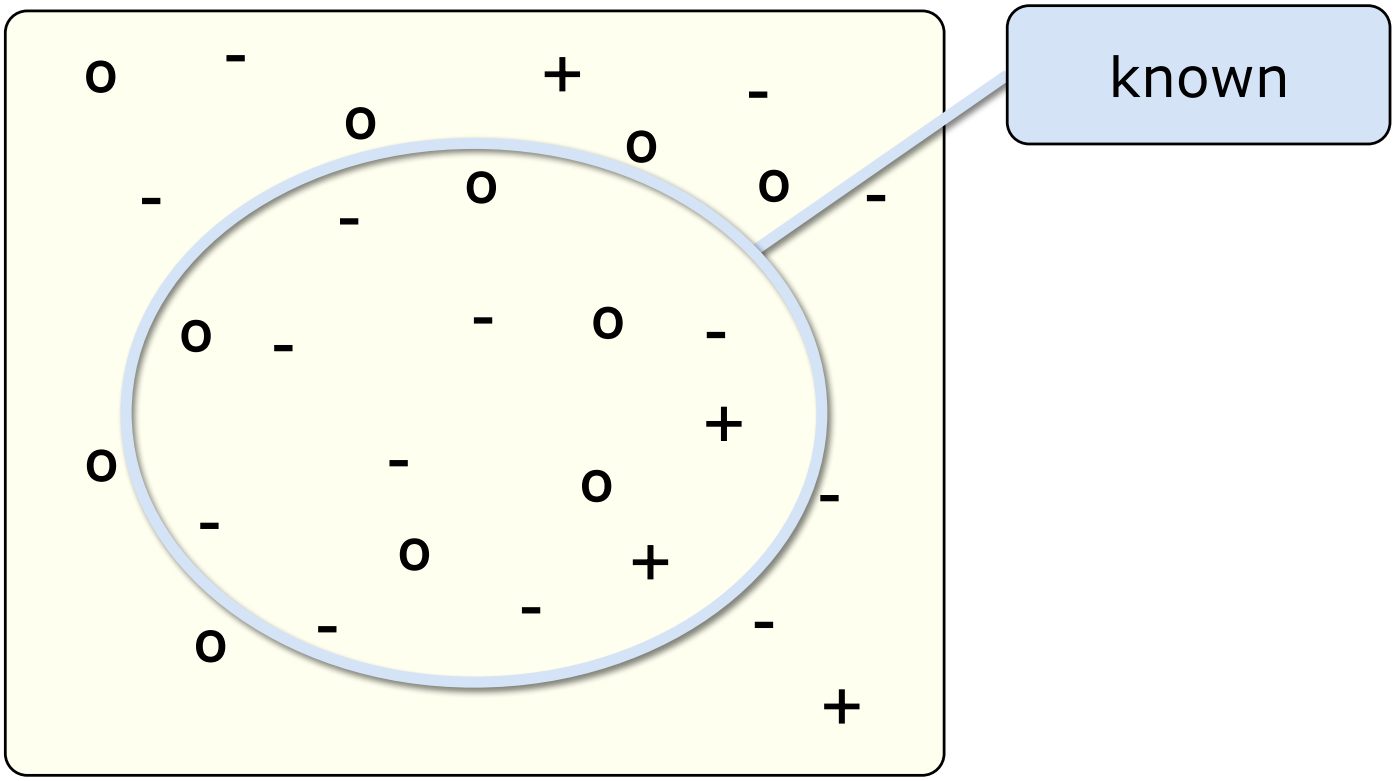
- We call visualization **idiom**, a specific approach to the the visual representation and manipulation of data.

Search the design space

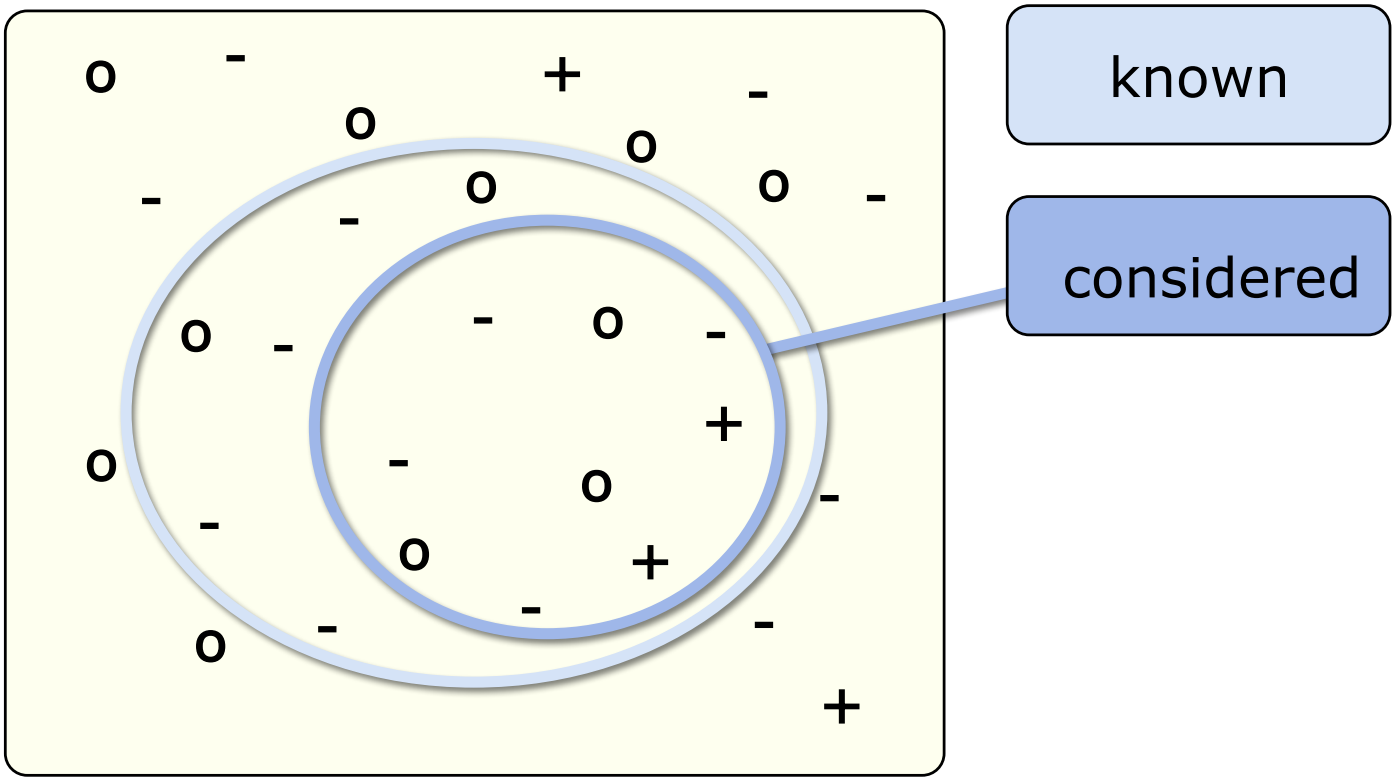


+ good
o okay
- poor

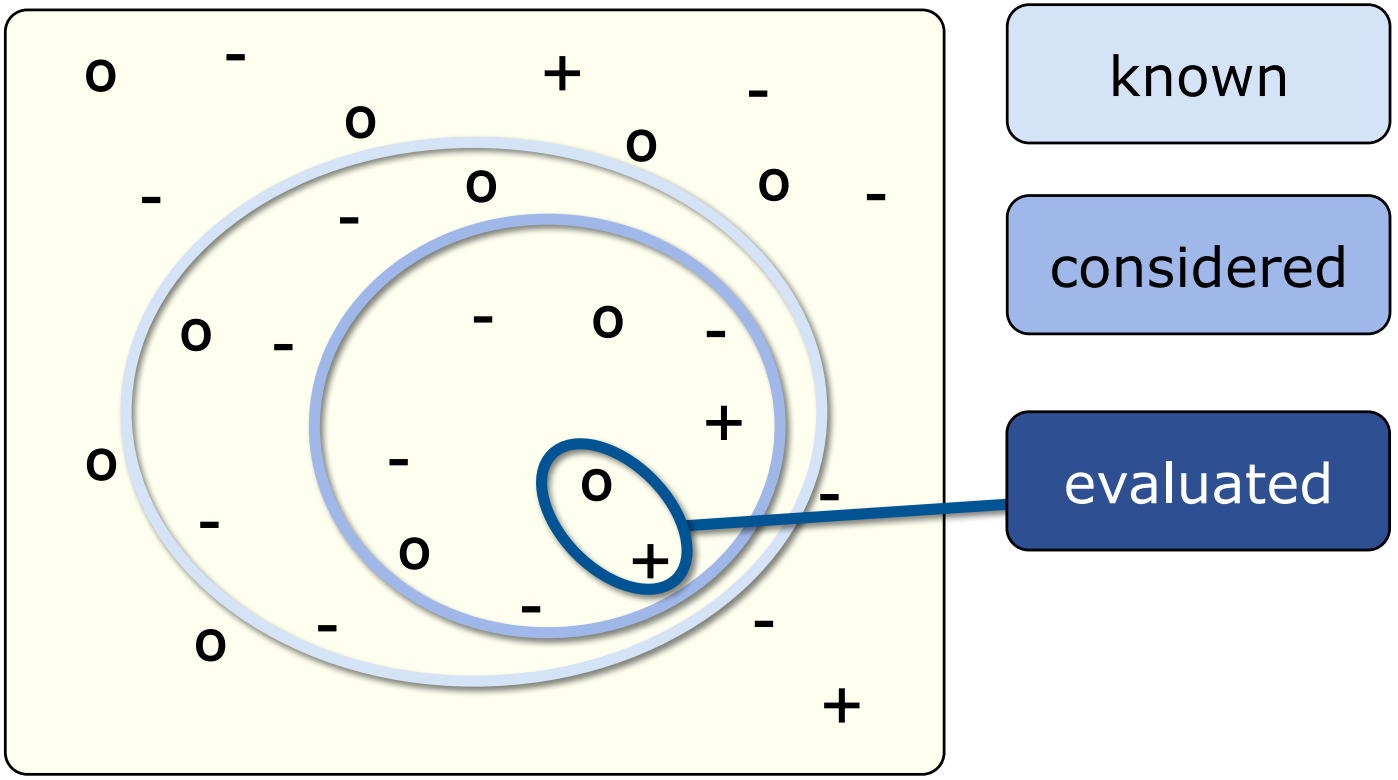
Search the design space



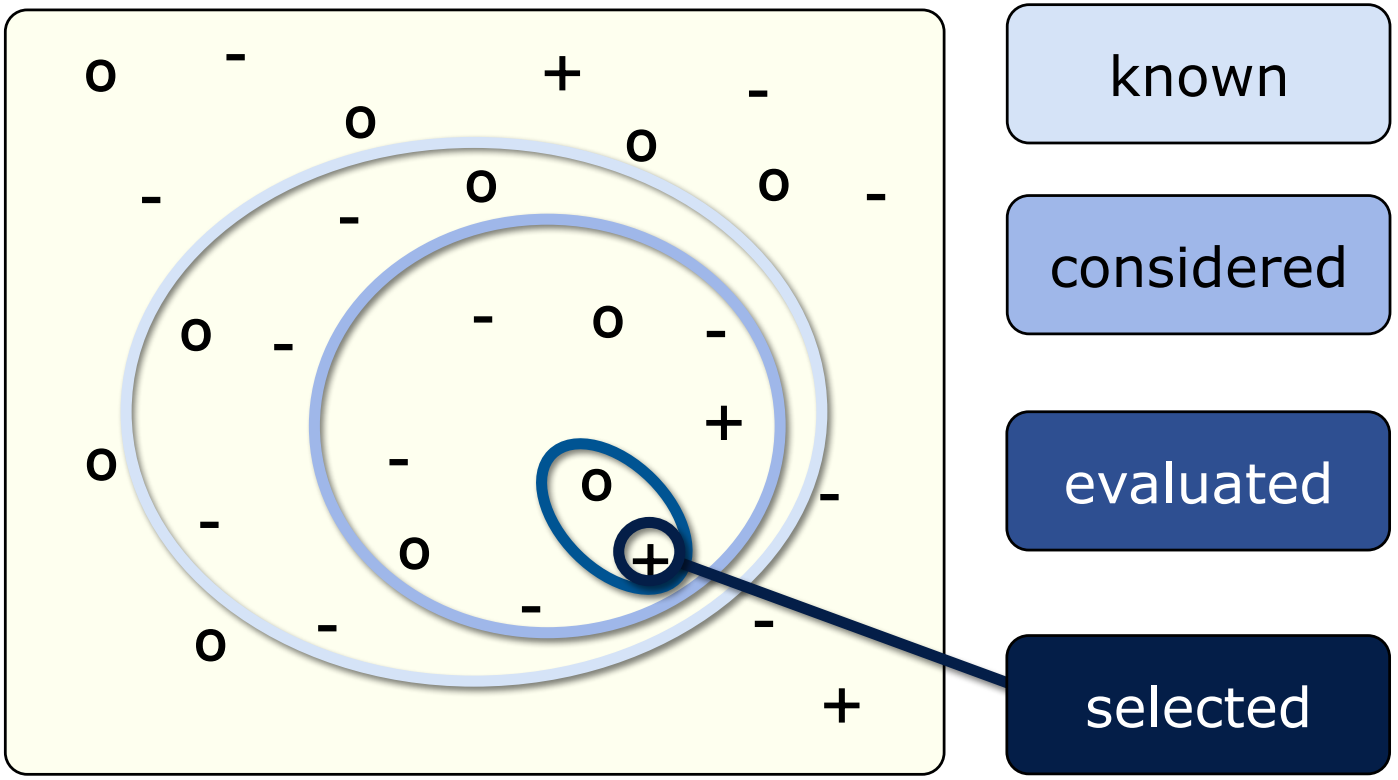
Search the design space



Search the design space

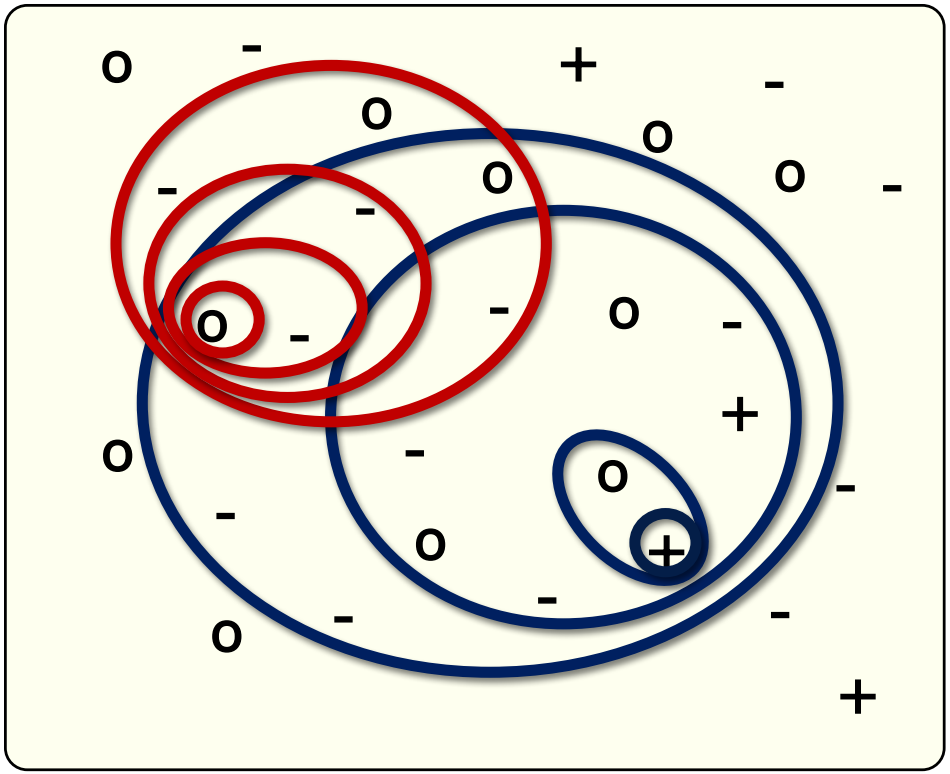


Search the design space



Search the design space

- Think broad!
- Iterate!



good search

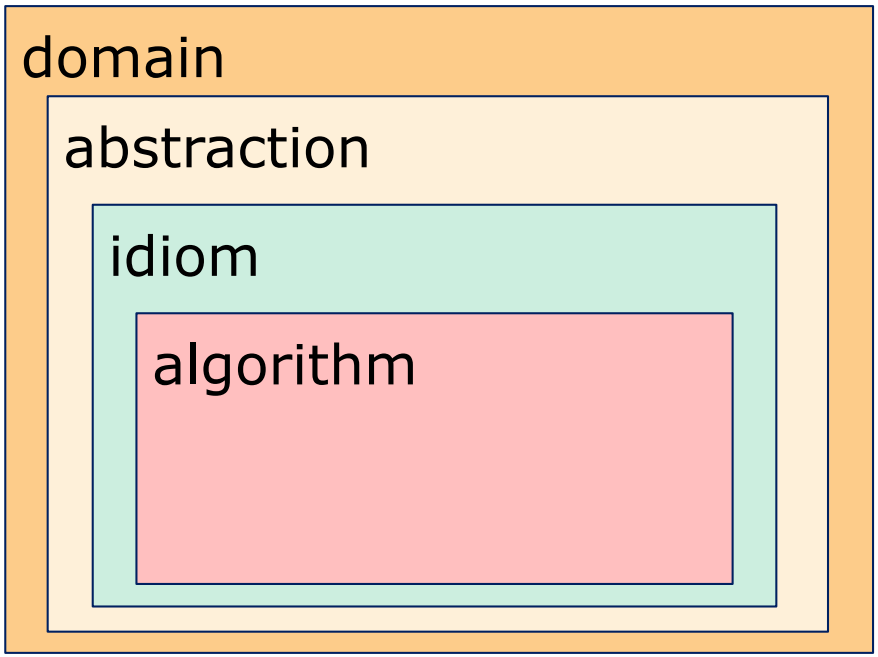
bad search

DESIGN FRAMEWORK

How do we analyze a design?

A four level analysis framework

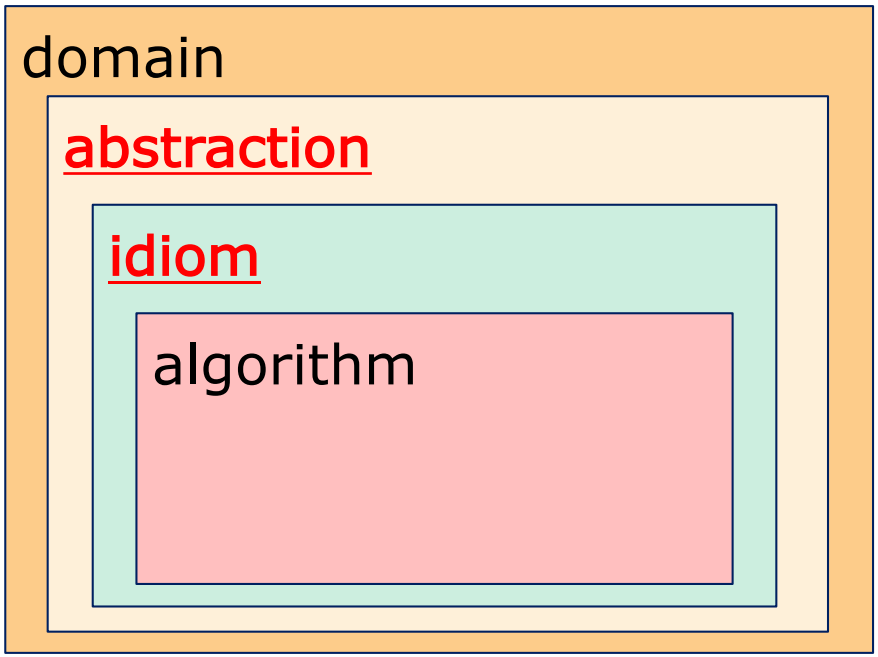
- In the *domain* level we identify **target**, **object, tasks** and **requirements** of the viz
- In the *abstraction* level we **map** domain-specific concepts to general ones
- In the *idiom* level we design **visual encoding** of data and **interaction**
- In the *algorithm* level we design the **computational process** to create visual encoding or to allow visual interaction



Munzner (2009)

A four level analysis framework

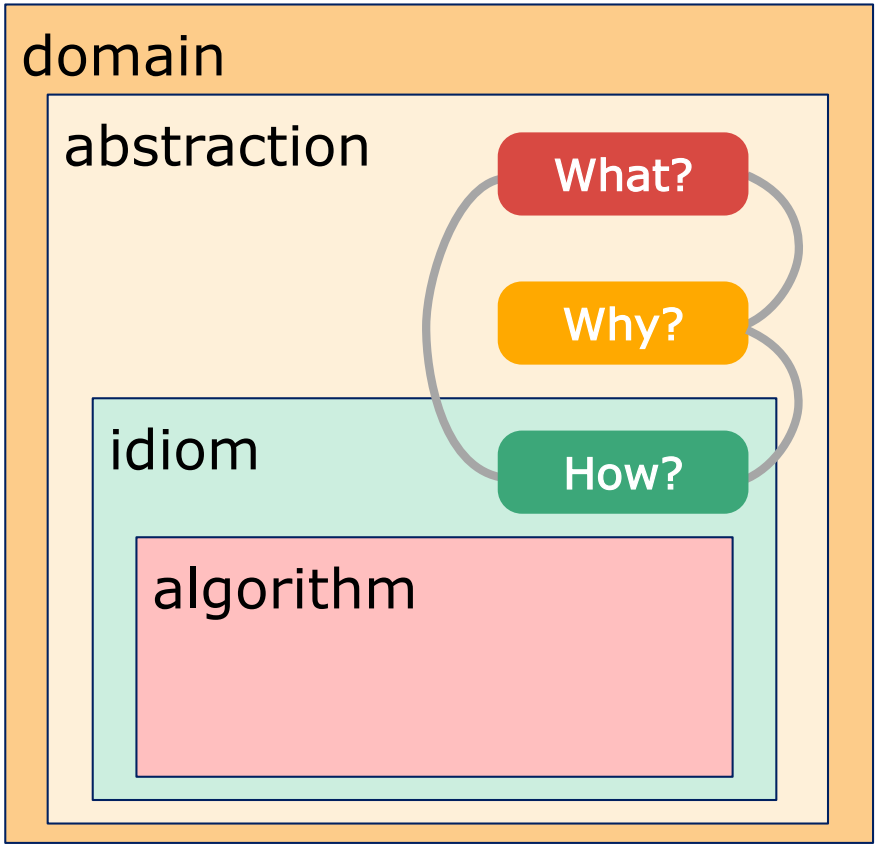
- In the *domain* level we identify **target**, **object, tasks** and **requirements** of the viz
- In the *abstraction* level we **map** domain-specific concepts to general ones
- In the *idiom* level we design **visual encoding** of data and **interaction**
- In the *algorithm* level we design the **computational process** to create visual encoding or to allow visual interaction



Munzner (2009)

Four levels, three questions

- What data is visualized? *(data abstraction)*
  - Type
  - Transformation
- Why data is visualized? *(task abstraction)*
  - Who is the users?
  - Actions
  - Targets
- How data is visualized? *(idiom)*
  - Visual encoding
  - Interaction
- Design process is usually an iterative refinement



Brehmer and Munzner (2013)

What?
Data Abstraction

Data types

- We identify five major abstract types of data:
  - **items** are discrete entities in the data
  - **links** are the relationships between items
  - **grids** are data sampling in a continuous domain (i.e., it is always possible to collect an additional sample between two collected ones)
  - **attributes** are measurable properties of an item, link or sample
  - **positions** are spatial data that locate in space items or samples

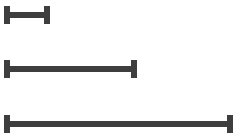
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

Attribute types

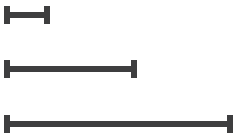
Attribute Types

- Categorical

e.g., gender, race, eye color

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

Attribute types

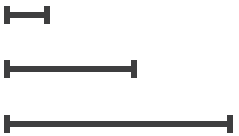
Attribute Types

- Categorical

- Ordered
  - *Ordinal*

e.g., edu level, ranking

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

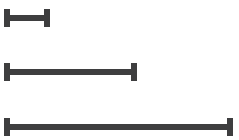
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



e.g., age, height, weight

Ordering Direction

- Sequential

- Diverging

- Cyclic

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

quantitative
ordinal
categorical

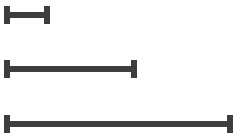
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

e.g., age, height, weight

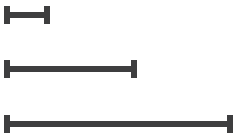
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

e.g., temperature, altitude

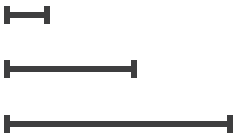
Attribute types

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*



Ordering Direction

- Sequential

- Diverging

- Cyclic

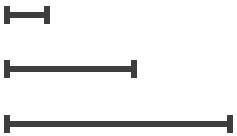
e.g., hour, week, year

Attribute types

- **Attribute Types**
  - Categorical

  - Ordered
    - *Ordinal*

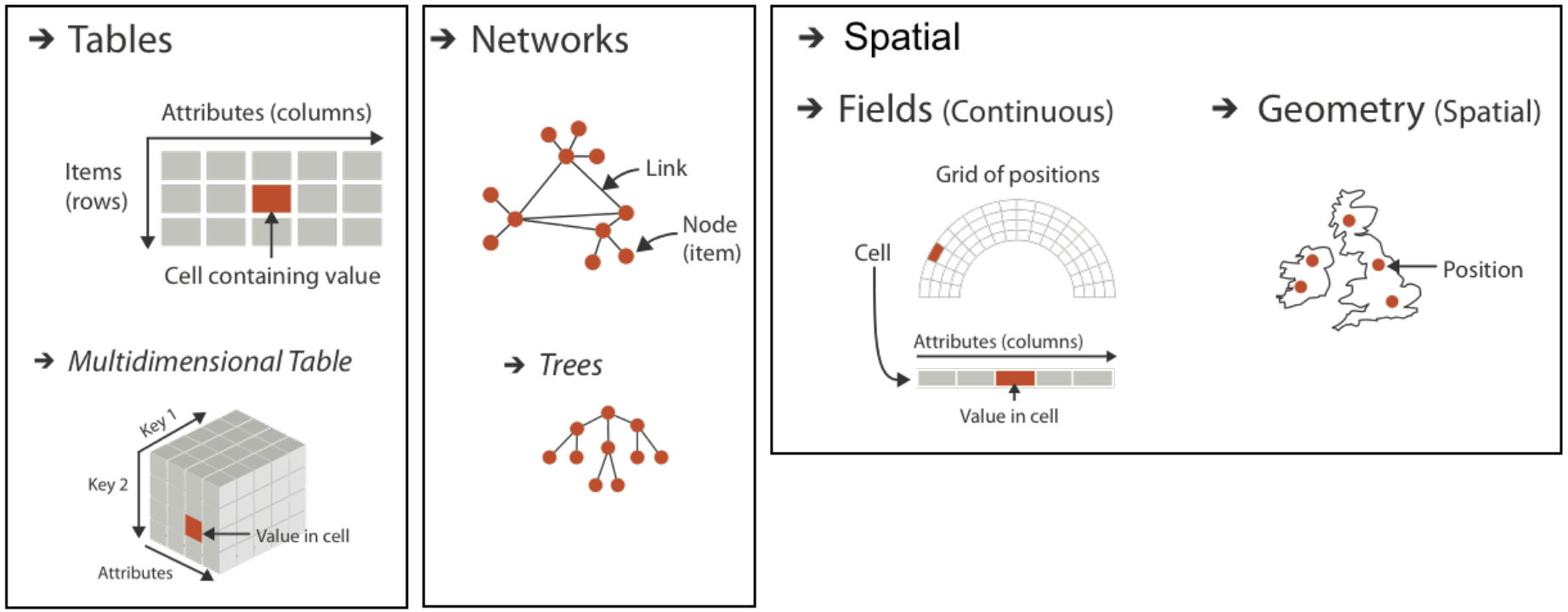
    - *Quantitative*



Some attributes may have an internal hierarchical structure (e.g., dates, spatial regions, taxonomies)

Dataset types

- A type of dataset is how data is arranged/collected
- Major types are:



- Dataset can be either **static** or **dynamic** (that is changing over time)

Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Tables

item

key

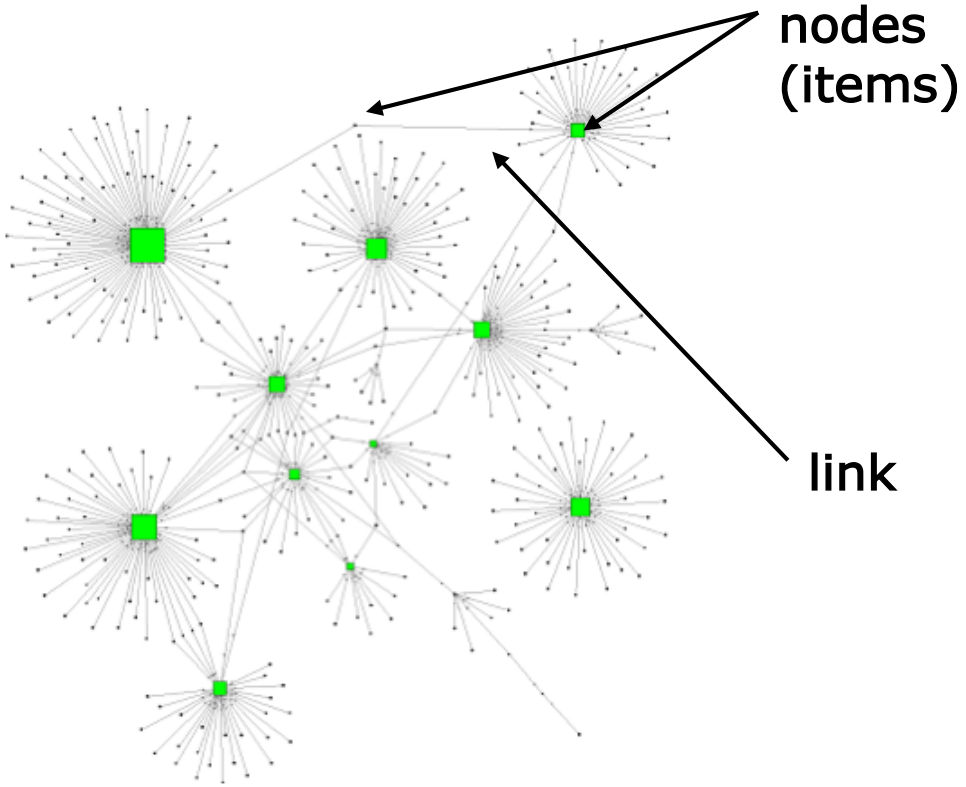
attributes

value

	A	B	C	D	E
1	Last Name	Sales	Country	Quarter	
2	Smith	$16,753.00	UK	Qtr 3	
3	Johnson	$14,808.00	USA	Qtr 4	
4	Williams	$10,644.00	UK	Qtr 2	
5	Jones	$1,390.00	USA	Qtr 3	
6	Brown	$4,865.00	USA	Qtr 4	
7	Williams	$12,438.00	UK	Qtr 1	
8	Johnson	$9,339.00	UK	Qtr 2	
9	Smith	$18,919.00	USA	Qtr 3	
10	Jones	$9,213.00	USA	Qtr 4	
11	Jones	$7,433.00	UK	Qtr 1	
12	Brown	$3,255.00	USA	Qtr 2	
13	Williams	$14,867.00	USA	Qtr 3	
14	Williams	$19,302.00	UK	Qtr 4	
15	Smith	$9,698.00	USA	Qtr 1	
16					

- In multidimensional tables, each item is identified by multiple **keys**

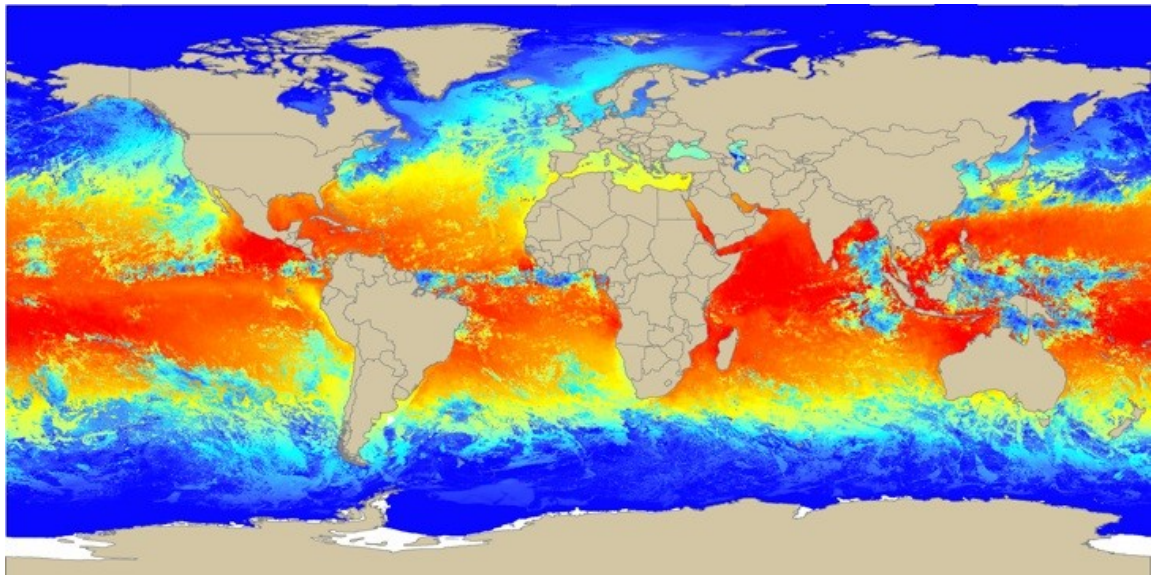
Networks (and trees)

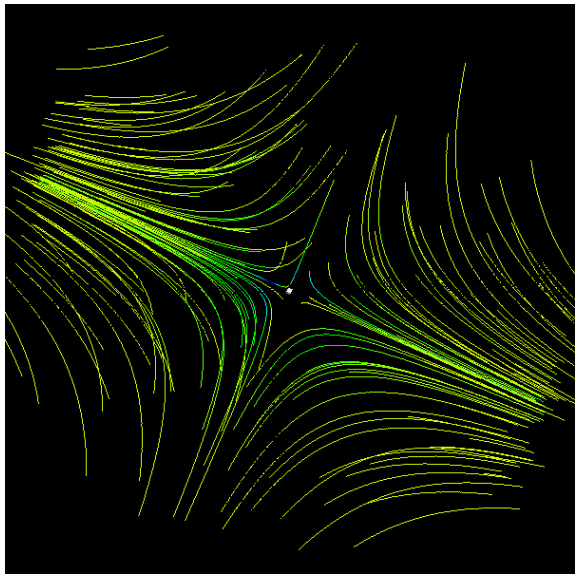


- Trees have a **hierarchical** structure where each node has only one parent.
- Nodes and links can also have **attributes**

Fields

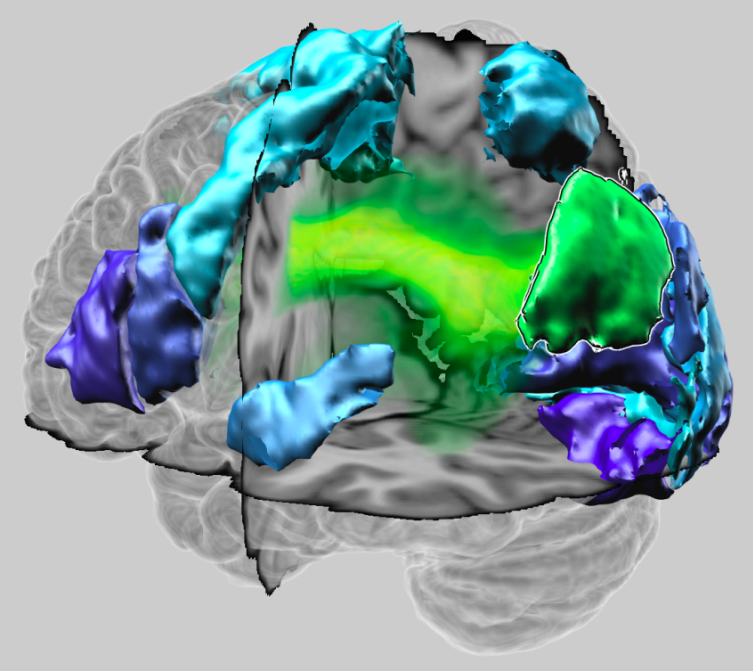
- Each **sample** of data is identified by a **position** and one ore more **attributes (scalar field, vector field, tensor field)**
- Sampling **grid** might not be **uniform** and can have **complex structure.**





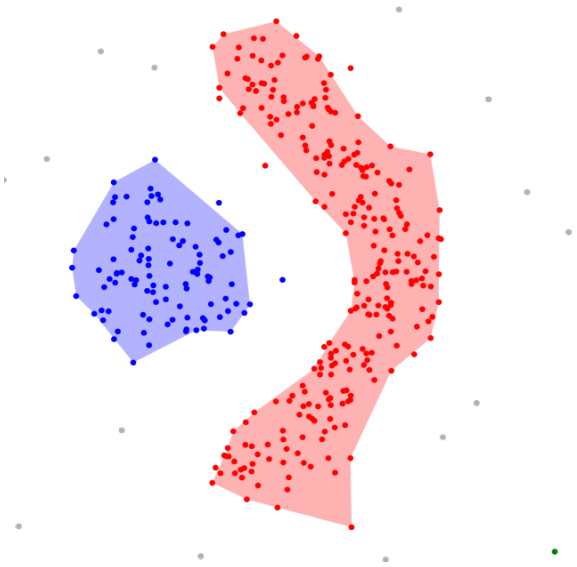
Geometry

- Dataset consist of **items** with **positions** (spatial or geographical)
- Items might have associated attributes



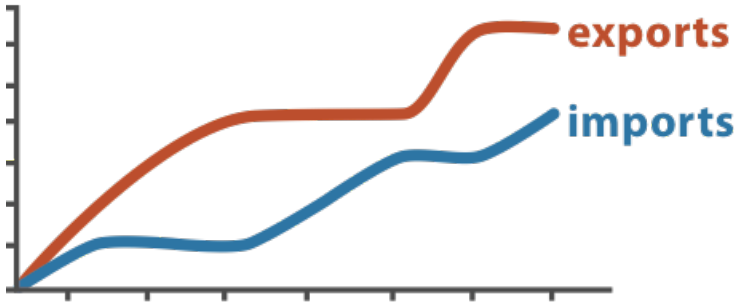
Clusters, sets, list

- Collection of **items** (grouped and/or ordered)

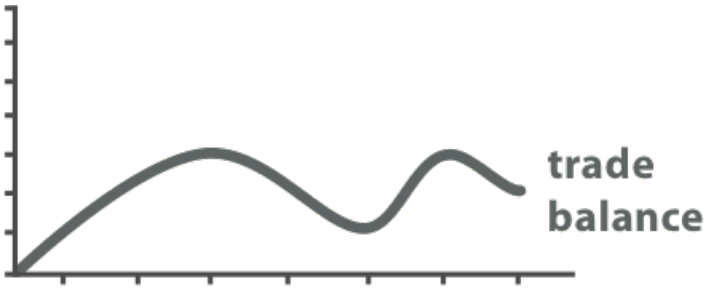


Derived data

- A major strategy to deal with complexity is to transform data
- You don’t have to just draw what you’re given!
  - identify the right data to show
  - derive it transforming the original dataset
  - draw it



Original Data



trade balance = exports – imports

Derived Data

Why?
Task Abstraction

From domain problem to abstract task

- Map specific domain problems to general viz task
- **Who** will perform the task?
  - End-users
    - Not trained
    - Limited interaction (to prevent ineffective viz)
  - Visualization designer
    - Trained
    - Advanced interaction (flexible)
- A task is defined by two elements:
  - Action
  - Target

Actions

- User actions can be described at three different levels:
  - High-level actions: analyze
  - Mid-level actions: search
  - Low-level actions: query
- These three levels represent independent choices
- A specific action can be described by a combination of high-level, mid-level and low-level actions.

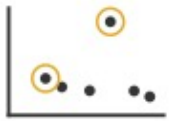
Mid-level and low-level actions

Search

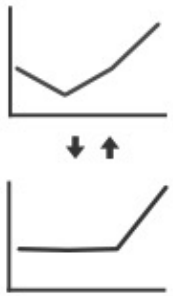
	Target known	Target unknown
Location known	 *Lookup*	 *Browse*
Location unknown	 *Locate*	 *Explore*

Query

- Identify



- Compare



- Summarize

High-level actions

Analyze

Consume

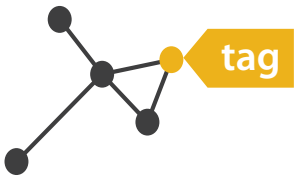
- *Discover*

- *Present*

- *Enjoy*

Produce

- *Annotate*



- *Record*

- *Derive*

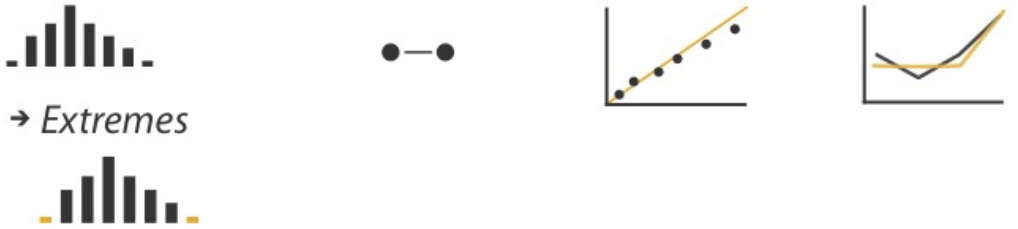
Targets

All Data

- Trends
- Outliers
- Features

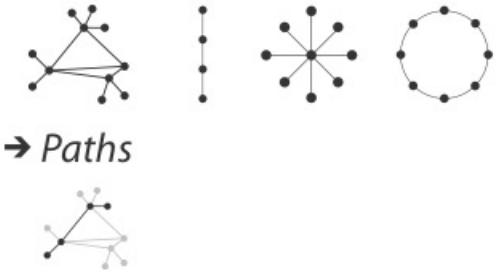
Attributes

- One
  - *Distribution*
  - *Extremes*
- Many
  - *Dependency*
  - *Correlation*
  - *Similarity*



Network Data

- Topology
- *Paths*



Spatial Data

- Shape

How?

Idiom

We will focus on this in the rest of the course!