Data Visualisation - Abstractions

Kamal Karlapalem

Spring 2024

Slides taken, reformatted, and used from Tamara Munzner (UBC, Canada)

Why visualize?

Why does visualisation work?

- Power of perception to reveal
  - How many V’s?
    - A:2
    - B:3
    - C:4
    - D:5
    - E:6

MTHIVLWYADCEQGHKILKMTWYN

ARDCAIREQGHLVKMFPSTWYARN

GFPSVCEILQGKMFPSNDRCEQDIFP

SGHLMFHKMVPSTWYACEQTW

Why does visualisation work?

- Power of perception to reveal
  - How many V’s?
    - A:2
    - B:3
    - C:4
    - D:5
    - E:6

MTHIVLWYADCEQGHKILKMTWYN

ARDCAIREQGHLVKMFPSTWYARN

GFPSVCEILQGKMFPSNDRCEQDIFP

SGHLMFHKMVPSTWYACEQTW

Why does visualisation work?

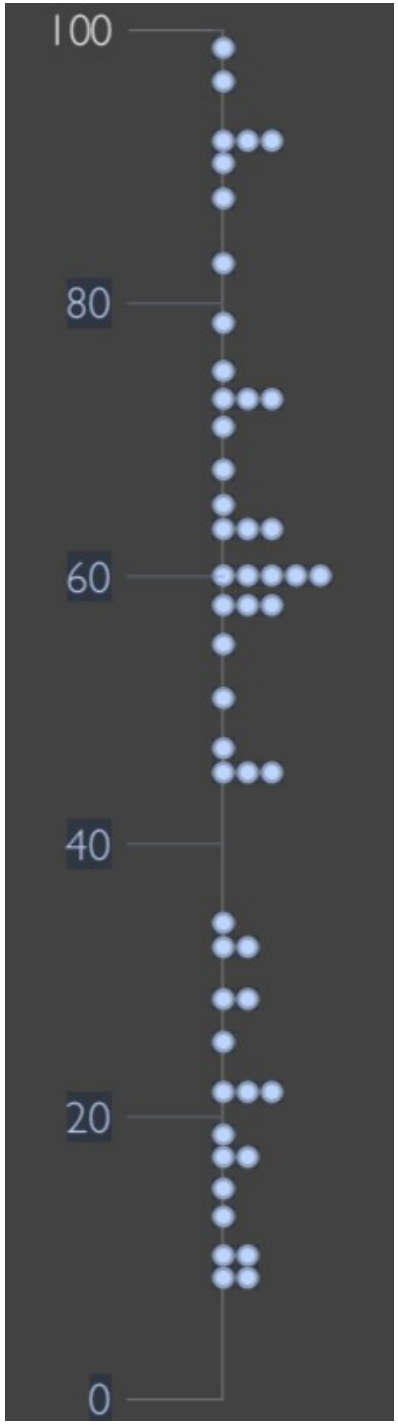
- Power of perception to reveal
  - How many V’s?
  - Which of these 50 numbers appears most often?
    - Numerical poll

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

Why does visualisation work?

- Power of perception to reveal
  - How many V’s?
  - Which of these 50 numbers appears most often?
    - Numerical poll

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



Exercise

- Which gender and income levels show a different effect of age on triglyceride levels?

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-$24,999	250	200	375	550
$25,000+	430	300	700	500

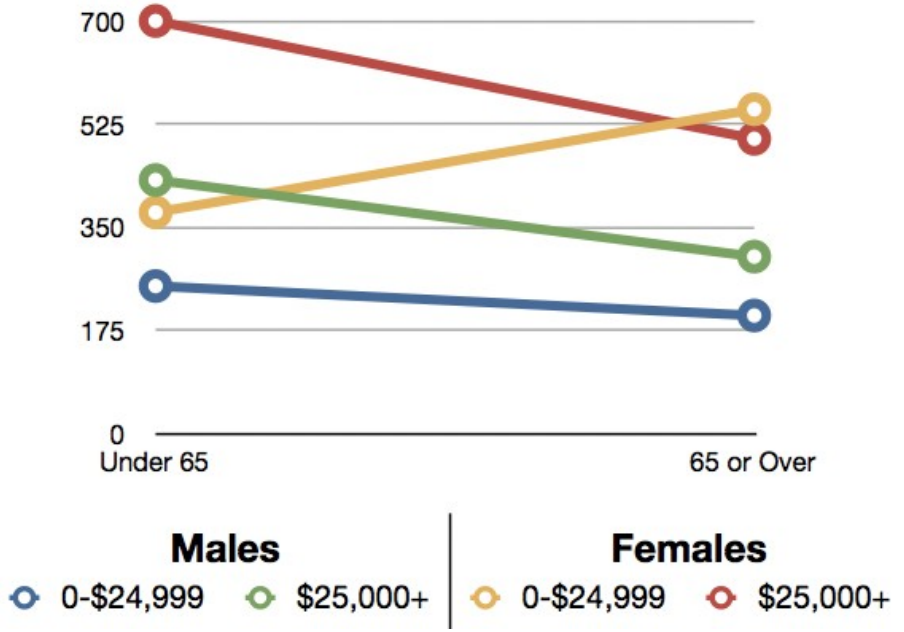
  - A : Male, low income
  - B : Male, high income
  - C : Female, low income
  - D : Female, high income

Exercise

- Which gender and income levels show a different effect of age on triglyceride levels?

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-$24,999	250	200	375	550
$25,000+	430	300	700	500

- A : Male, low income
- B : Male, high income
- C : Female, low income
- D : Female, high income



What are the limits of human perception and cognition?

- Limits of memory and cognition
  - Change blindness

https://youtu.be/FWSxSQsspiQ

Data

What does data mean?

14, 2.6, 30, 30, 15, 100001

- What does this sequence of six numbers mean?
  - Two points far from each other in 3D space
  - Two points close to each other in 2D space, with 15 links between them and a weight of 100001 for the link?
  - Something else?

Basil, 7, S, Pear

- What about this data?
  - Food shipment of produce (basil & Pear) arrived in satisfactory condition on 7th day of month
  - Basil point neighborhood of the city had 7 inches of snow cleared by Pear Creek limited snow removal service.
  - Lab rat Basil made 7 attempts to find way through south section of maze, these trials used pears as reward food

Now what

- Semantics: real-world meaning

Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

Now what?

- Semantics: real-world meaning

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

Now what?

- Semantics: real-world meaning
- Data types: structural or mathematical interpretation of data
  - Item, link, attribute, position, grid
  - Different from data types in programming

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

Items and attributes

- Item: individual entity, discrete (row of a relation table, e.g.)
  - Patient, car, stock, city
  - “independent variable.”
- Attribute: property that is measured, observed, logged.
  - Height, blood pressure
  - Horse power
  - “dependent variable.”

attributes: name, age, shirt size, fave fruit

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

item: person

Other data types

- Links
  - Express relationships between two items
  - Friendship on Facebook, interaction between proteins
- Positions
  - Spatial data: location in 2D or 3D
  - Pixels in the photo, voxels in MRI scan, latitude/longitude
- Grids
  - Sampling strategy for continuous data

Dataset Types

Tables

Items

Attributes

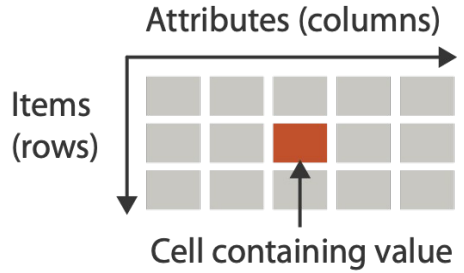
- Flat table
  - Each item per row
  - Each column is an attribute
  - Cell holds value for item-attribute pair

attributes: name, age, shirt size, fave fruit

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

item: person

➔ Tables



Data types

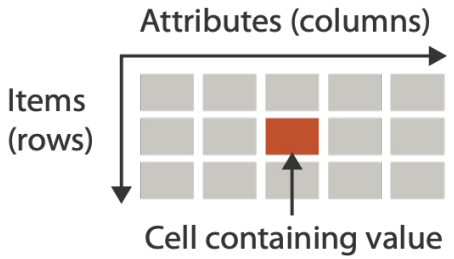
Tables

Items

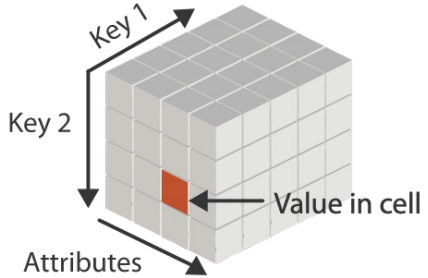
Attributes

- Multidimensional tables
  - Indexing based on multiple keys
    - Eg genes, patients

➔ Tables



➔ *Multidimensional Table*



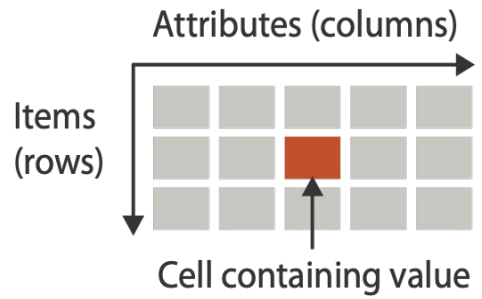
	A	B	C	D	E
1	#1.2				
2	1500	529			
3	GeneName	DESCRIPTION	TCGA-02-0001-01C-01R-0177-01	TCGA-02-0003-01A-01R-0177-01	TCGA-02-0004-01A-01R-0298-01
4	LTF	LTF	-1.265728057	2.377012066	4.123979585
5	POSTN	POSTN	2.662411805	3.932400324	5.031585377
6	TMSL8	TMSL8	-3.082217838	-2.243148513	-0.02313681
7	HLA-DQA1	HLA-DQA1	-1.739664398	4.577962344	3.127744964
8	RP11-35N6.1	RP11-35N6.1	-3.346352968	-2.895400157	-3.473035067
9	STMN2	STMN2	-2.578511106	-3.051605144	-1.729892888
10	DCX	DCX	-2.26078976	-2.529795801	-2.844966278
11	AGXT2L1	AGXT2L1	-2.639493611	-3.113204863	-0.403975027
12	IL13RA2	IL13RA2	-2.93596915	-1.873600916	2.976256911
13	SLN	SLN	-2.466718221	-2.208406749	1.025827904
14	MEOX2	MEOX2	-2.395054066	-1.062676046	1.783235317
15	COL11A1	COL11A1	1.211934832	-0.399392588	4.733608974
16	NNMT	NNMT	0.703745164	0.664082419	3.069030715
17	F13A1	F13A1	-0.224094042	2.222197544	1.171354775
18	CXCL14	CXCL14	-3.1309694	-1.395056071	2.569540659
19	MBP	MBP	-1.906390566	-2.037626447	-2.935744906
20	TF	TF	-4.334123292	-4.680680246	-2.975788866
21	KCND2	KCND2	-1.777692395	-2.100362021	-1.996306032

Data Types

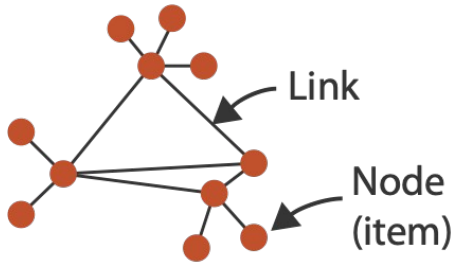
Tables	Networks & Trees
Items	Items (nodes)
Attributes	Links
	Attributes

- Network/graph (nodes/edges)
  - Nodes (vertices connected by links (edges)
  - Tree is a special case : no cycles – connected
    - Often have roots and are directed.

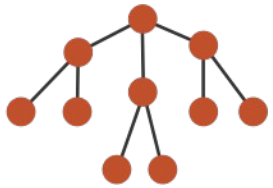
➔ Tables



➔ Networks



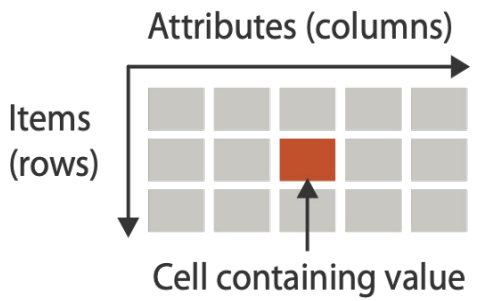
➔ *Trees*



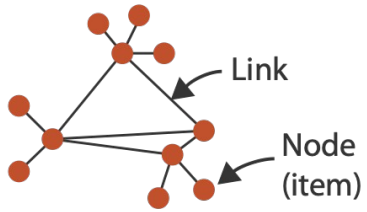
Data Types

Tables	Networks & Trees	Fields
Items	Items (nodes)	Grids
Attributes	Links	Positions
	Attributes	Attributes

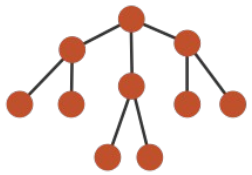
➔ Tables



➔ Networks

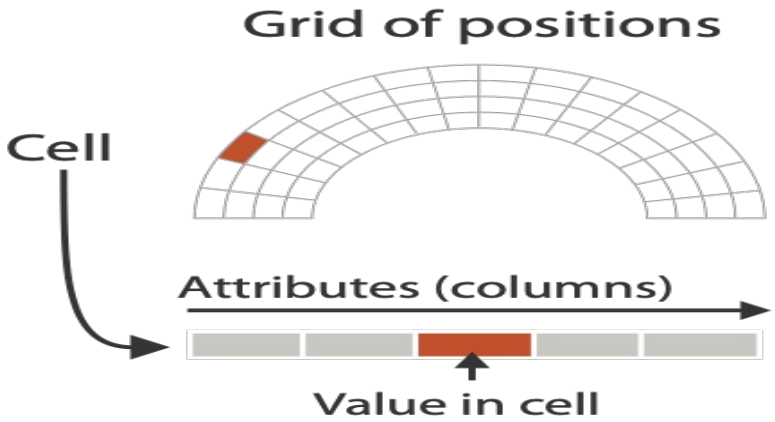


➔ *Trees*



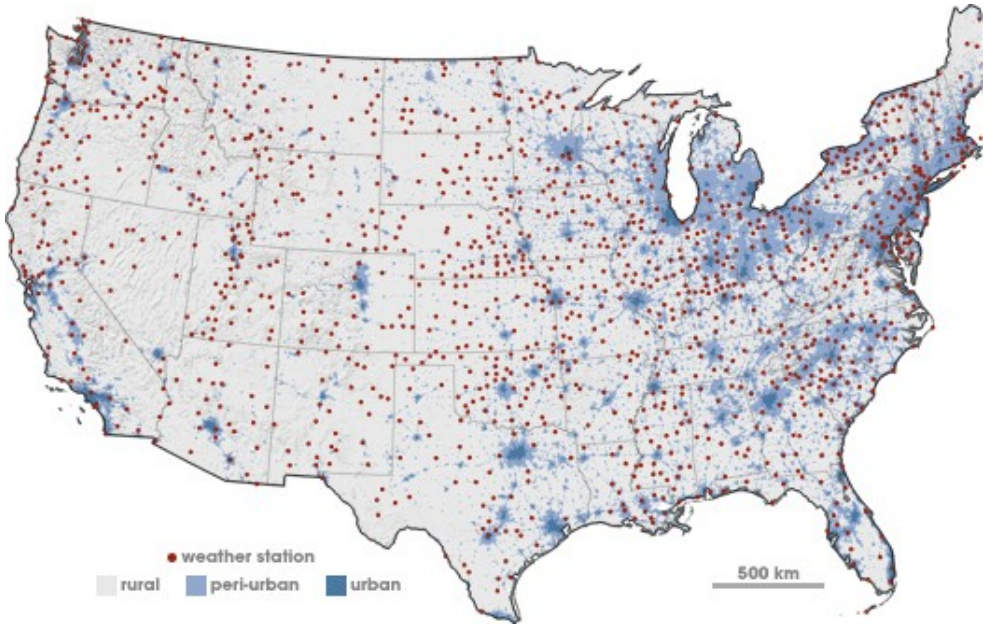
-> spatial

➔ Fields (Continuous)



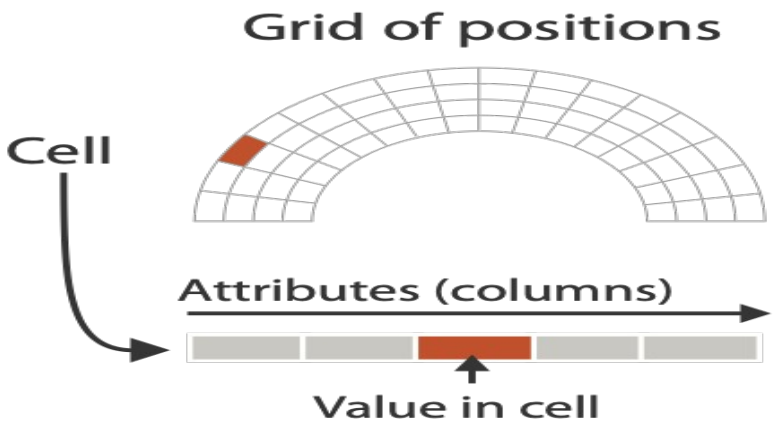
Spatial fields

- Attribute values associated with cells
- Cells contain a value from the continuous domain
  - Eg, temperature, pressure, etc.
- Measure or simulated



-> spatial

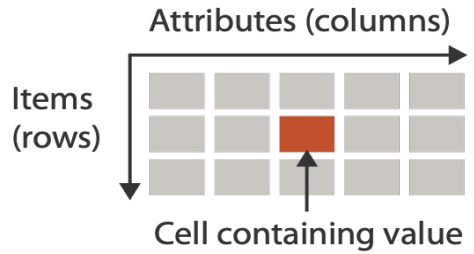
➔ **Fields** (Continuous)



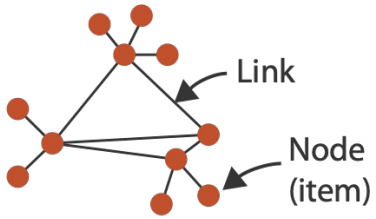
Data Sets

Tables	Networks & Trees	Fields	Geometry
Items	Items (nodes)	Grids	Items
Attributes	Links	Positions	Positions
	Attributes	Attributes	

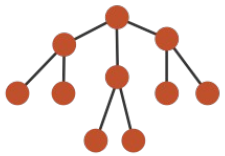
➔ Tables



➔ Networks

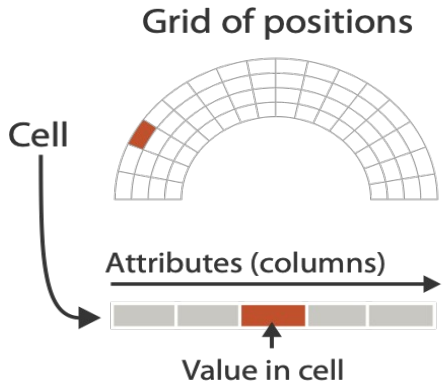


➔ *Trees*



➔ Spatial

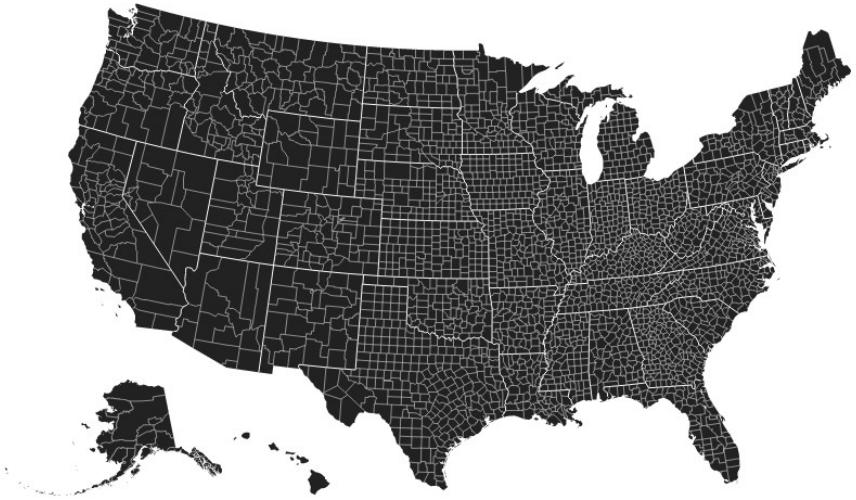
➔ Fields (Continuous)



➔ Geometry (Spatial)

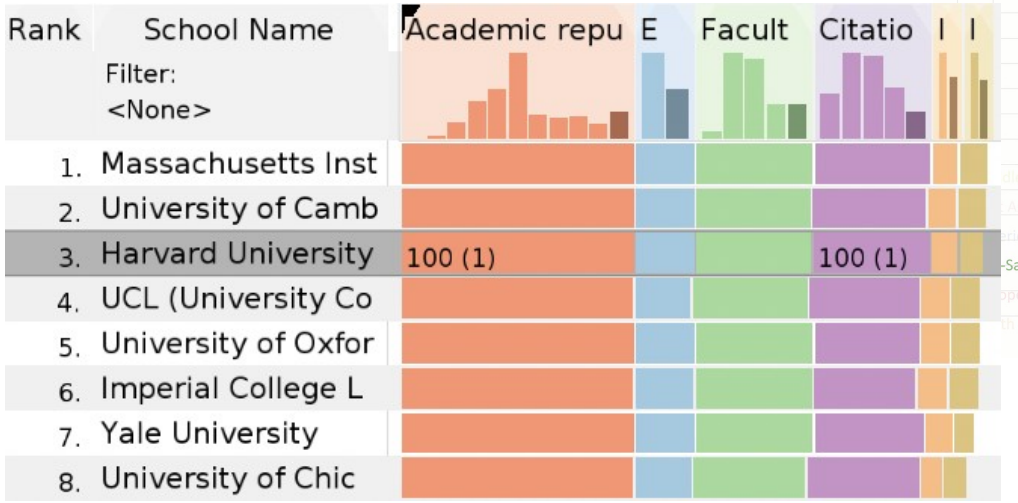
Geometry

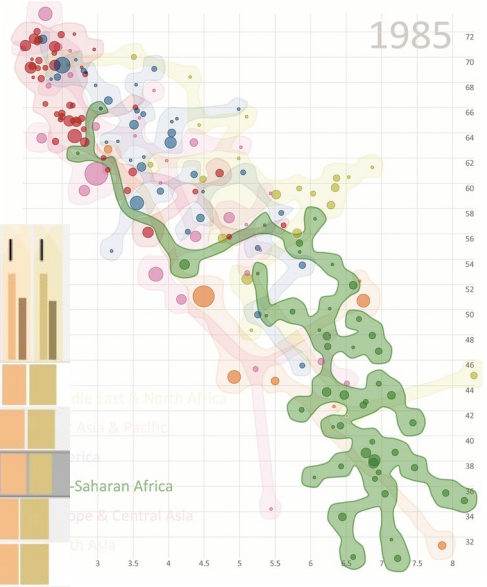
- Shape of items
- Explicit spatial positions/regions
  - Points, lines, curves, surfaces, volumes
- Boundary between computer graphics and visualisation
  - Graphics: geometry taken as given
  - Visualisation: geometry is result of a design decision

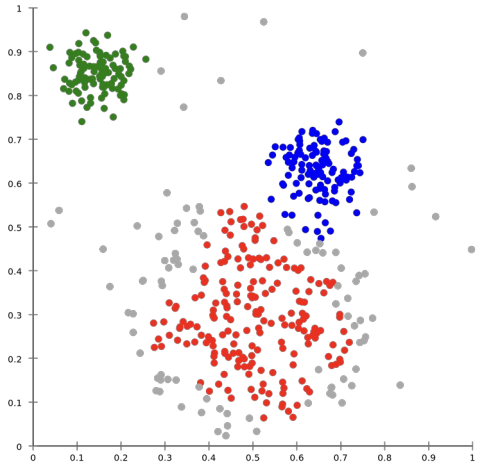


Collections

- How we group items
- Sets
  - Unique items, unordered
- Lists
  - Ordered - duplicates possible
- Clusters
  - Groups of similar items







Dataset and data types

Data and Dataset Types

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

Data Types

→ Items → Attributes → Links → Positions → Grids

Attribute types

- Which classes of values and measurements?
- Categorical (nominal)
  - Compare equality
  - No implicit ordering
- Ordered
  - Ordinal
    - Less/greater than defined
  - Quantitative
    - Meaningful magnitude
    - Arithmetic possible

Table

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 Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	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

Categorical
Ordinal
Quantitative

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 Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	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

What kind of variable?

- 50 meter race times
- College program of study
- Rating for movie
- Product name

- Categorical, Ordinal, Quantitative

Other data considerations

- **Attribute Types**
  - Categorical
    
  - Ordered
    - *Ordinal*
      
    - *Quantitative*
      

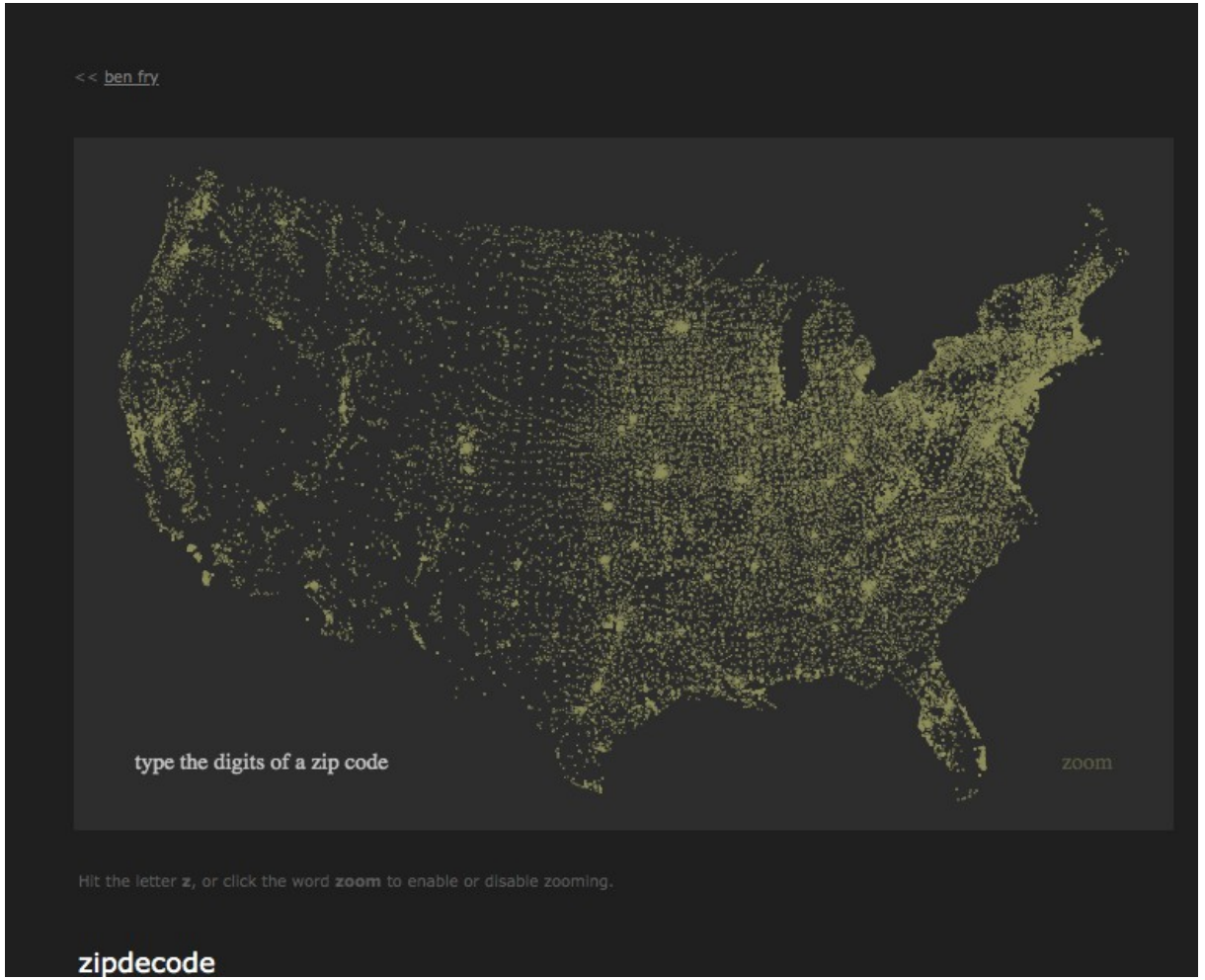
- **Ordering Direction**
  - Sequential
    
  - Diverging
    
  - Cyclic
    

- **Dataset Availability**
  - Static
    
  - Dynamic
    

Hierarchical data

- Multi-level structure
  - Space
  - Time
  - Others

Example: zipcode



https://benfry.com/zipdecode/

Data abstraction: three operations

- Translate from domain-specific language to generic visualisation language
- Identify dataset type(s), attribute types
- Identify cardinality
  - How many items are in the data set?
  - What is the cardinality of each attribute?
    - Number of levels for categorical data
    - Range for quantitative data
- Consider whether to transform data
  - Guided by understanding of task

Data vs Conceptual Models

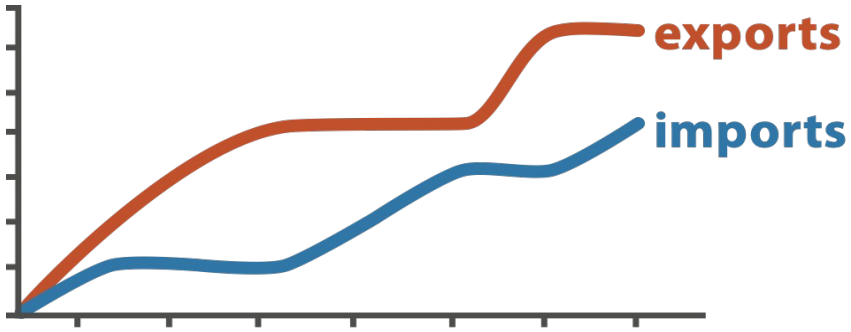
- Data model (visualisation perspective, not databases)
  - Mathematical abstraction
    - Sets with operations, eg floats with */-+
    - Variable data types in programming languages
- Conceptual model
  - Mental construction (semantics)
  - Supports reasoning
  - Typically based on an understanding of tasks
- Data abstraction process relies on a conceptual model
  - For transforming data if needed

Data vs. conceptual model, example

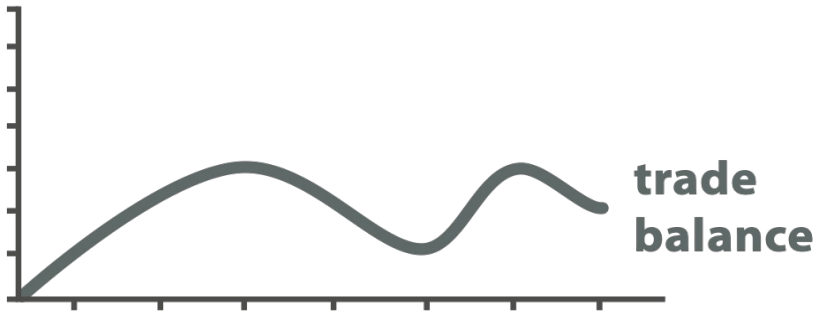
- Data model:floats
  - 32.52, 54.06, -14.35
- Conceptual model
  - Temperature
- Multiple possible data abstractions
  - Continuous to 2 significant digits : quantitative
    - Task: forecasting the weather
  - Hot, warm, cold : ordinal
    - Task: deciding if bath water is ready
  - Above freezing, below freezing: categorical
    - Task: decide if I should leave the house today

Derived attributes

- Derived attribute: compute from originals
  - Simple change of type
  - Acquire additional data
  - Complex transformation

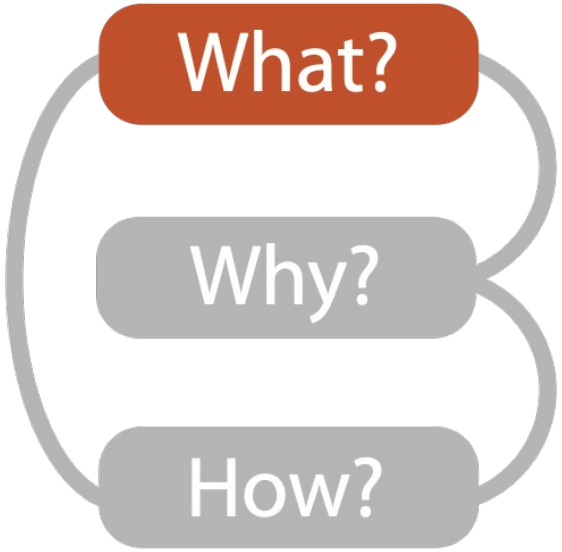


Original Data



trade balance = exports – imports

Derived Data



What?

Datasets

Data Types

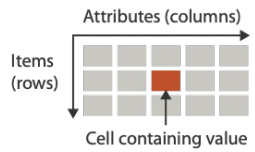
- Items
- Attributes
- Links
- Positions
- Grids

Data and Dataset Types

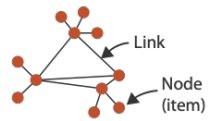
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

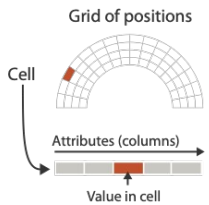
- Tables



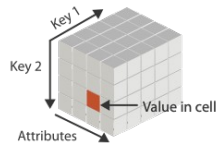
- Networks



- Fields (Continuous)



- *Multidimensional Table*



- *Trees*

- Geometry (Spatial)

Attributes

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*

Ordering Direction

- Sequential

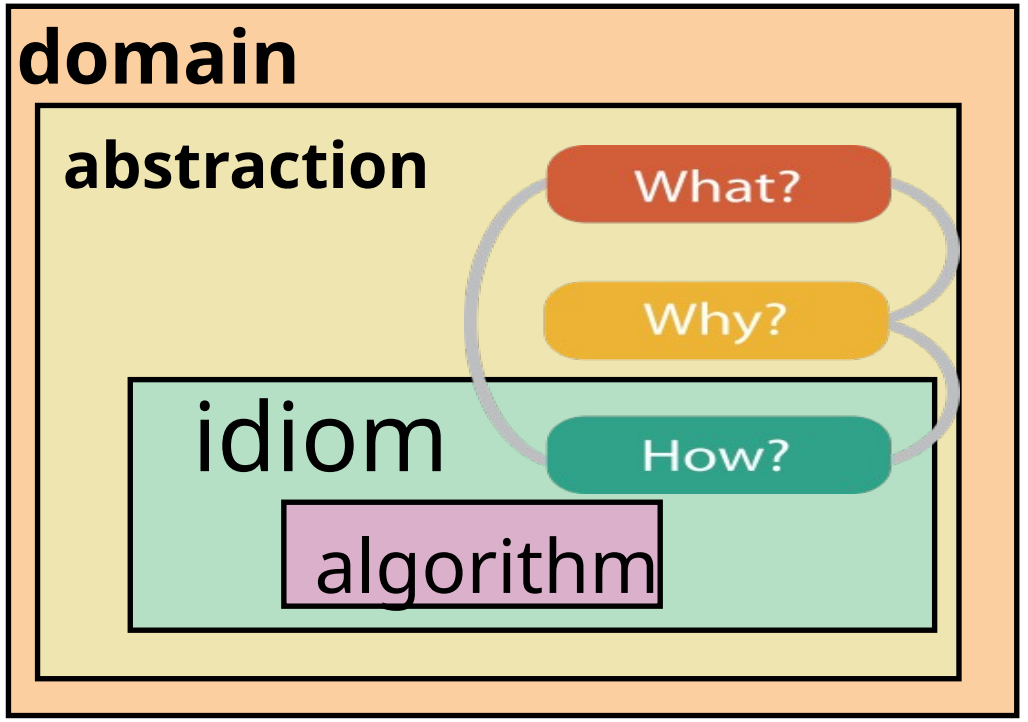
- Diverging

- Cyclic

Abstractions tasks (why?)

Nested model: Four levels of visualisation design

- Domain situation
  - Who are the target users?
- Abstraction
  - Translate from specifics of the domain to the vocabulary of visualisation
    - **What** is shown? **Data** abstraction
    - **Why** is the user looking at it? **Task** abstraction
      - Often must transform data, guided by a task
- Idiom
  - **How** is it shown?
    - **Visual encoding** idiom: how to draw
    - **Interaction** idiom: how to manipulate
- Algorithm
  - Efficient computation



[A Nested Model of Visualization Design and Validation.

Munzner. IEEE TVCG 15(6):921-928, 2009

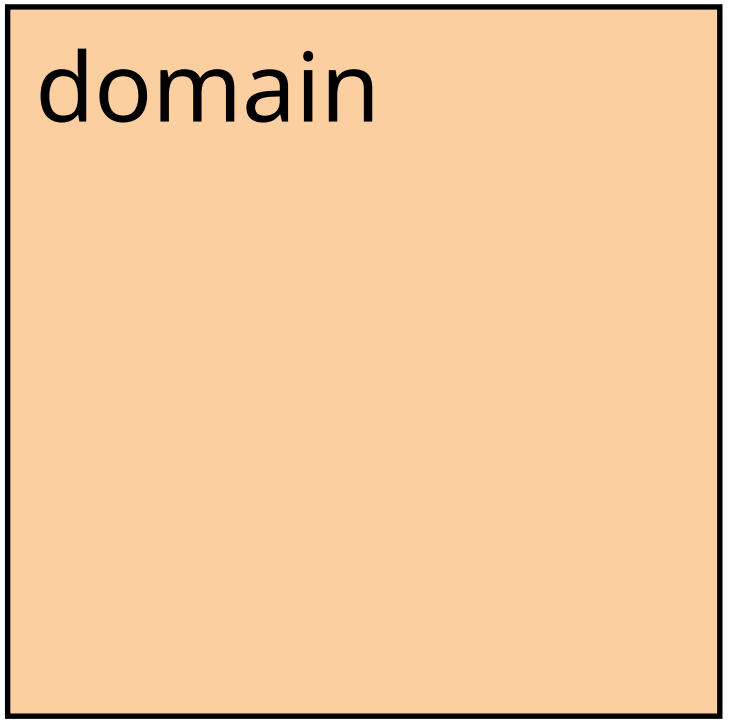
(Proc. InfoVis 2009).]

[A Multi-Level Typology of Abstract Visualization Tasks

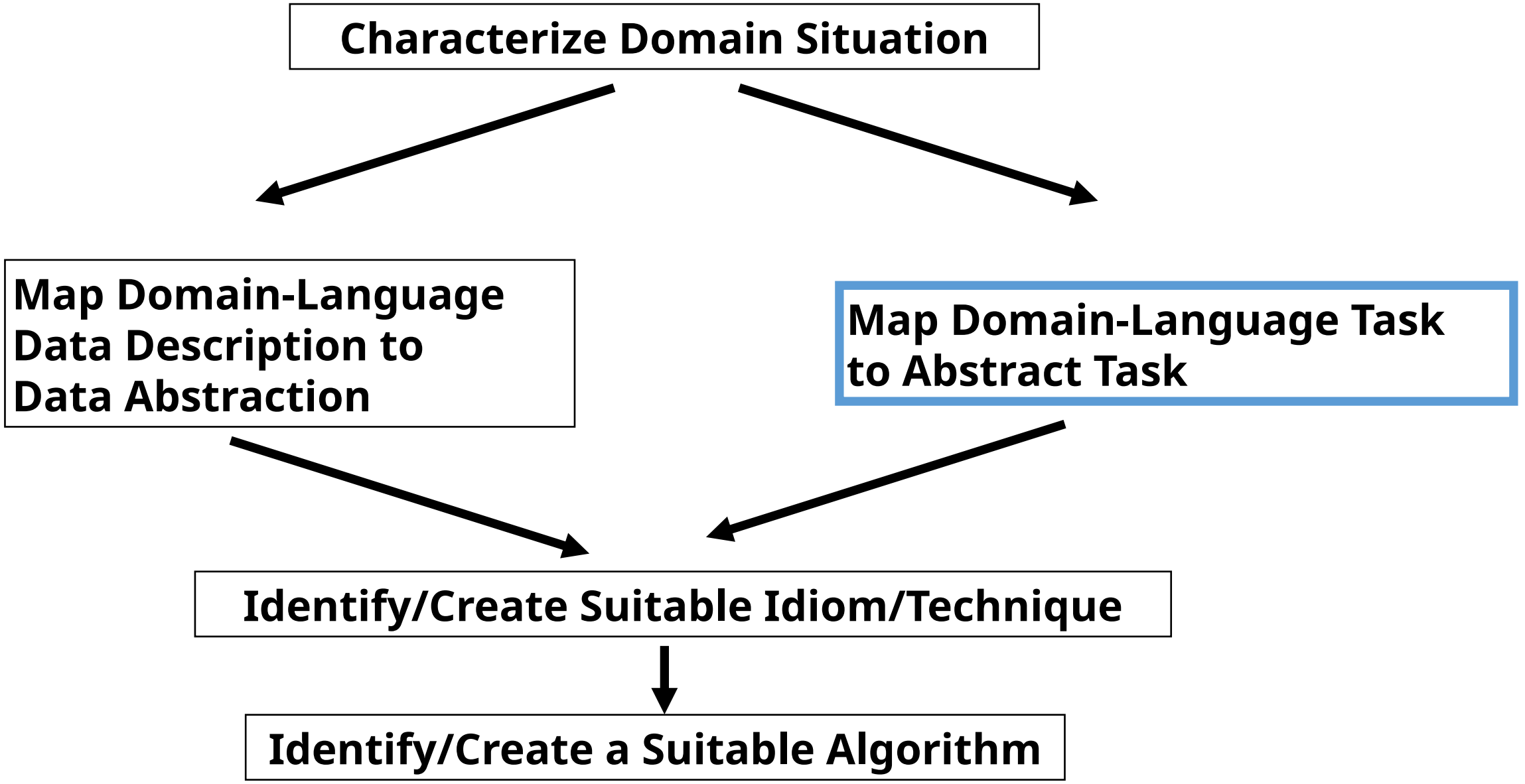
Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

Domain characterization

- Details of an application domain
- Group of users, target domain, their questions, and their data
  - Varies wildly by domain
  - Must be specific enough to get traction
- Domain questions/problems
  - Break down into simpler abstract tasks



Design Process

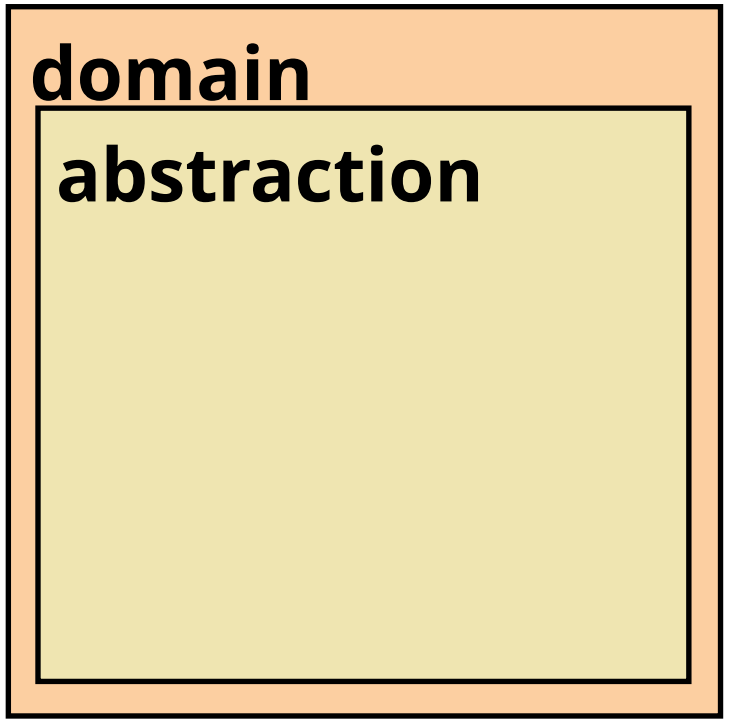


Example: Find good movies

- Identify good movies in genres I like
- Domain:
  - General population, movie enthusiasts

Abstraction: Data and 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



Example: Find good movies

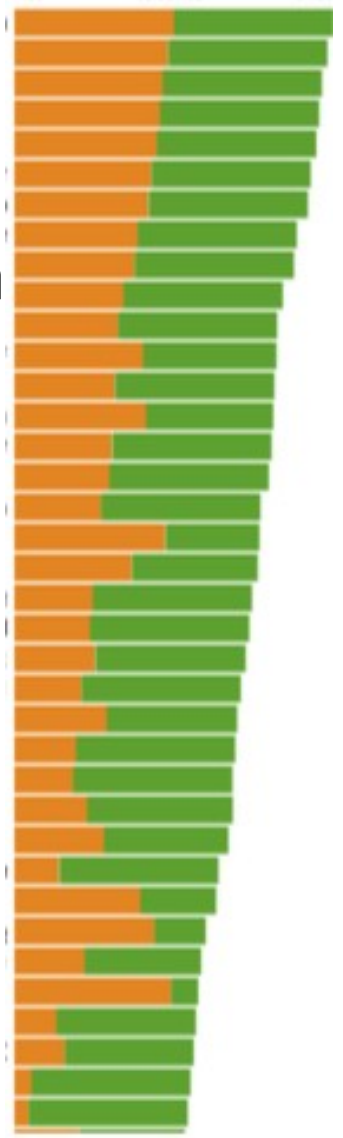
- Identify good movies in genres I like
- Domain:
  - General population, movie enthusiasts
- Task: what is a good movie for me?
  - Highly rated by critics?
  - Highly rated by audiences?
  - Successful at the box office?
  - Similar to movies I liked?
  - Matches specific genres?
- Data : (is it available?)
  - Yes! Data sources IMDB, Rotten Tomatoes…

Example: Find good movies

- One possible choice for data and tasks in the domain langua
  - Data: combine audience ratings and critic ratings
  - Task: find high-scoring movies for specific genre
- Abstractions?
  - Attribute: audience and critic ratings
    - Ordinal
      - Levels 3 or 5 or 10
  - Attribute: genre
    - Categorical
      - Levels: <20
  - Items: movies
    - Items: millions
  - Task: to find high values

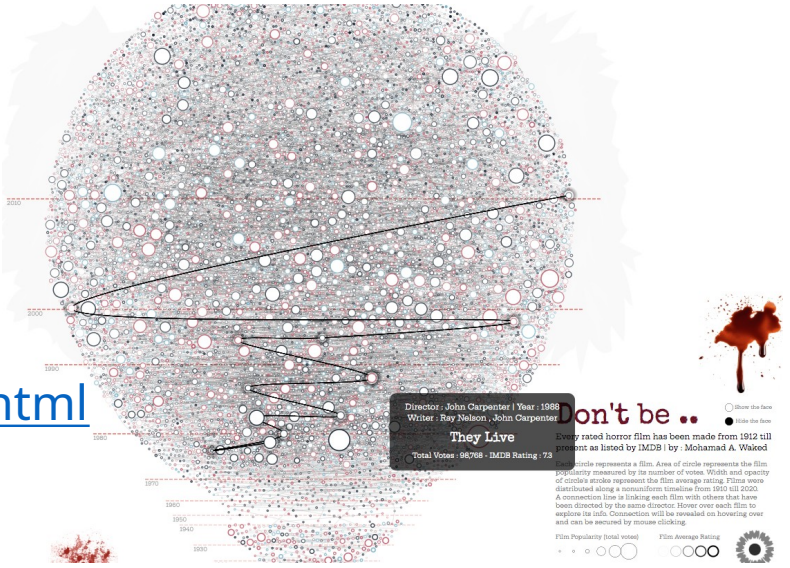
One possible idiom

- stacked bar chart of ratings



Example: Horrified!

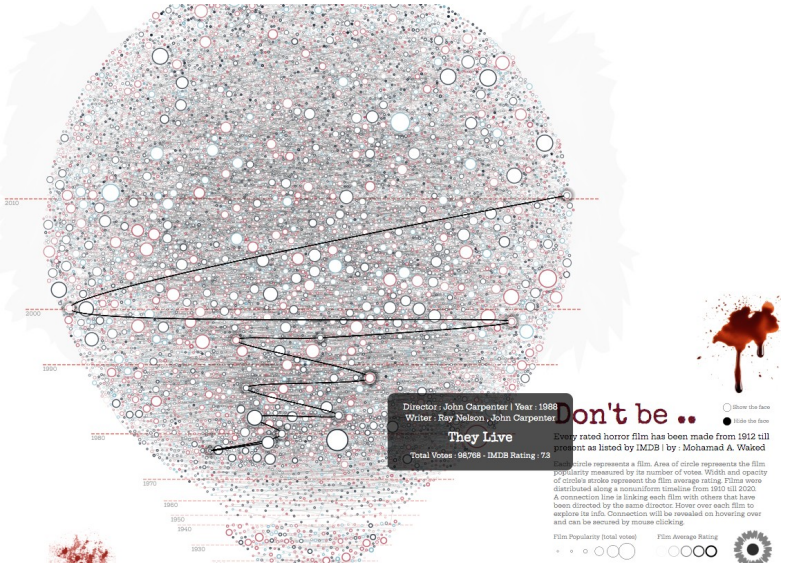
- Same task: high-score movies
- Slightly different data
  - 14K rated horror movies from IMDB



https://www.alhadaqa.com/wp-content/uploads/2020/04/horrified.html

Example: Horrified

- Same task: high-score movies
- Slightly different data
  - 14K rated horror movies from IMDB
- Very different visual encoding idiom
  - Circle per item (movie)
  - Circle area = popularity
  - Stroke width/opacity = avg rating
  - Year made = vertical position
- Interaction idiom
  - Lines connect movies with same director on mouse over



https://www.alhadaqa.com/wp-content/uploads/2020/04/horrified.html

Task abstraction: Actions and Targets

- Very high-level pattern

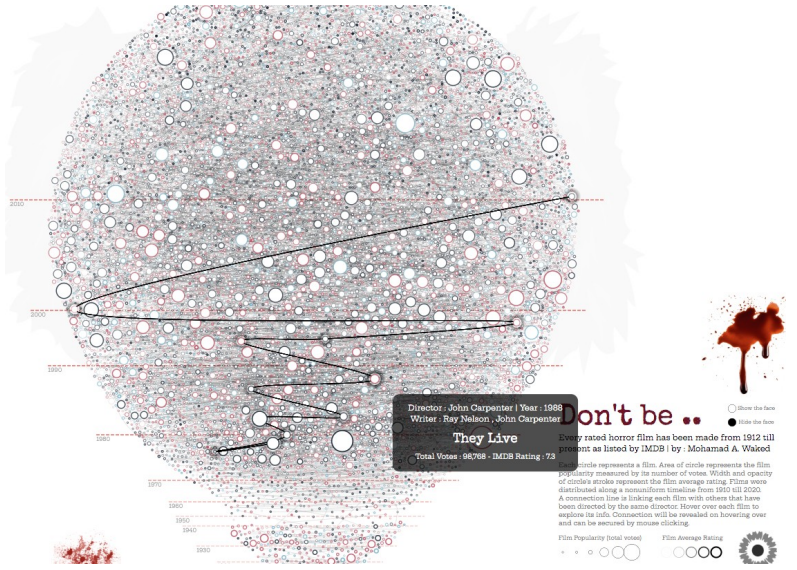
- Actions
  - Analyze
    - High-level choices
  - Search
    - Find a known/unknown item
  - Query
    - Find out about the characteristics of item

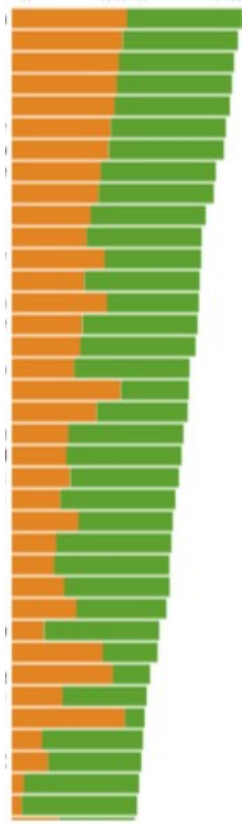
- {action, target} pairs
  - Discover distribution
  - Compare trends
  - Locate outliers
  - Browse topology

Example: Horrified vs Stacked bars

- Horrified: browse/explore
- Stacked bars: locate/lookup

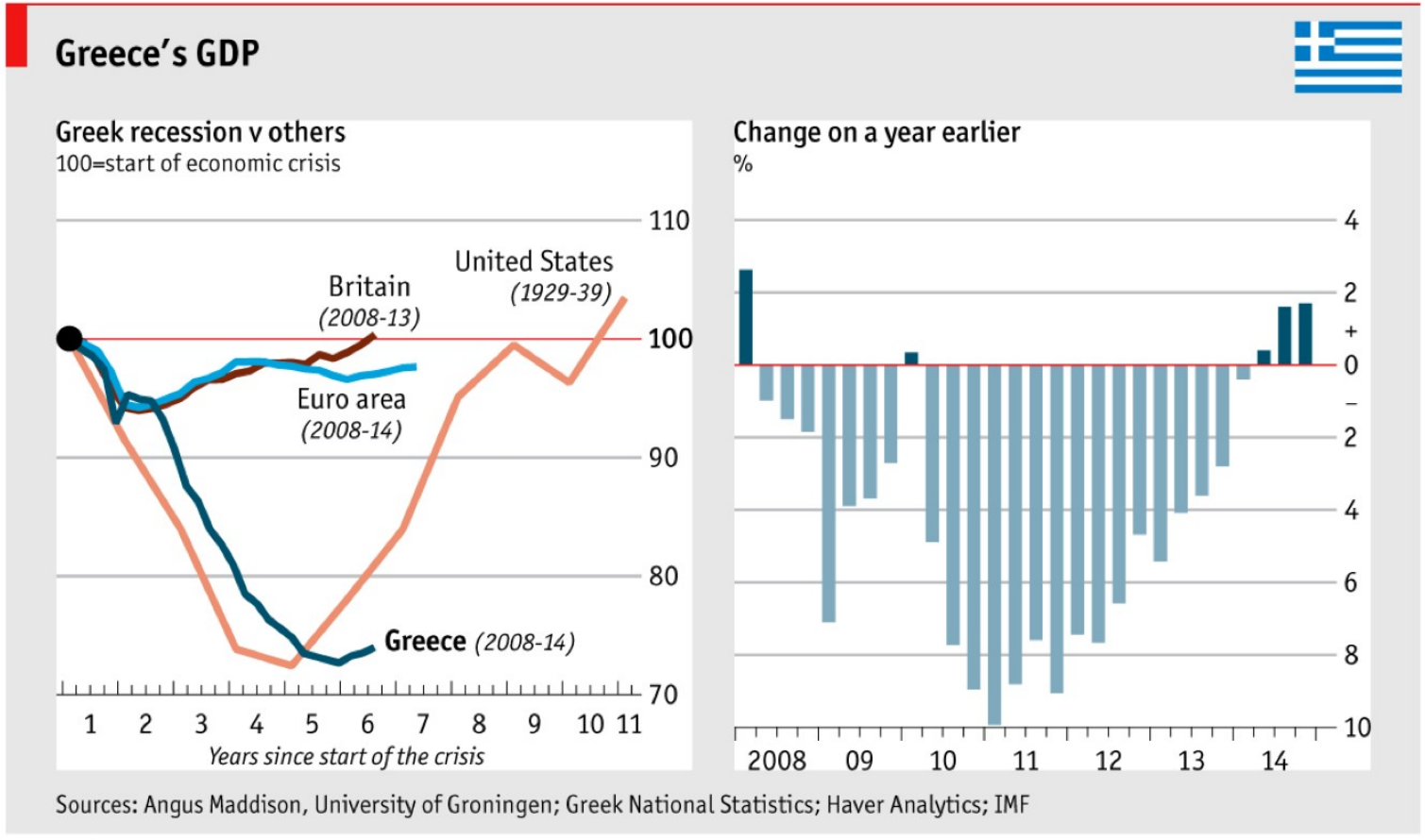
- Which is better?
  - Depends on goals/task
    - Enjoy social context. Lots of time
    - Find 2nd best-rated movies of all time
      - KBC question, <10 seconds to respond





Example : Economics

- Task : compare and derive
- Data : derive change

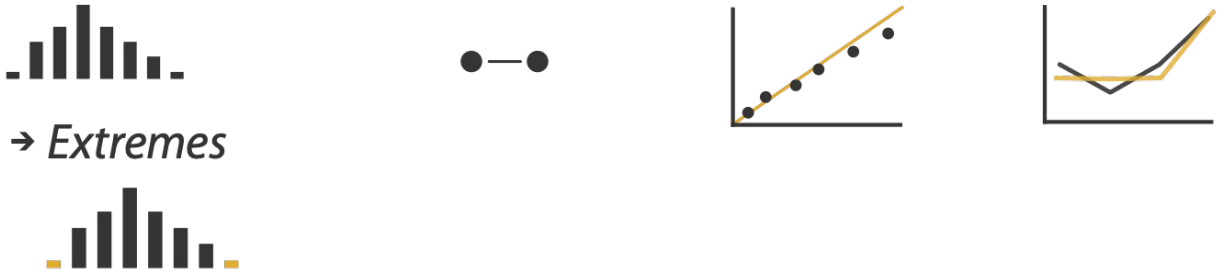


The Economist

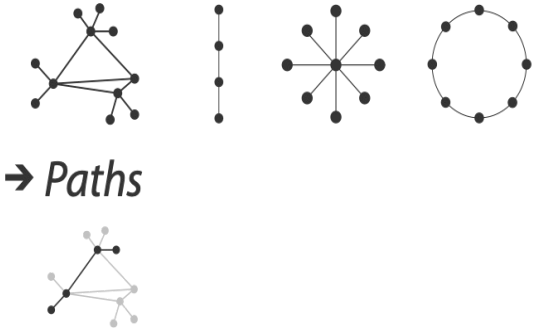
Task abstraction: Targets

- **All Data**
  - Trends
  - Outliers
  - Features

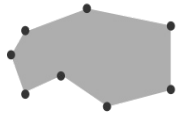
- **Attributes**
  - One
    - *Distribution*
    - *Extremes*
  - Many
    - *Dependency*
    - *Correlation*
    - *Similarity*



- **Network Data**
  - Topology
  - *Paths*



- **Spatial Data**
  - Shape

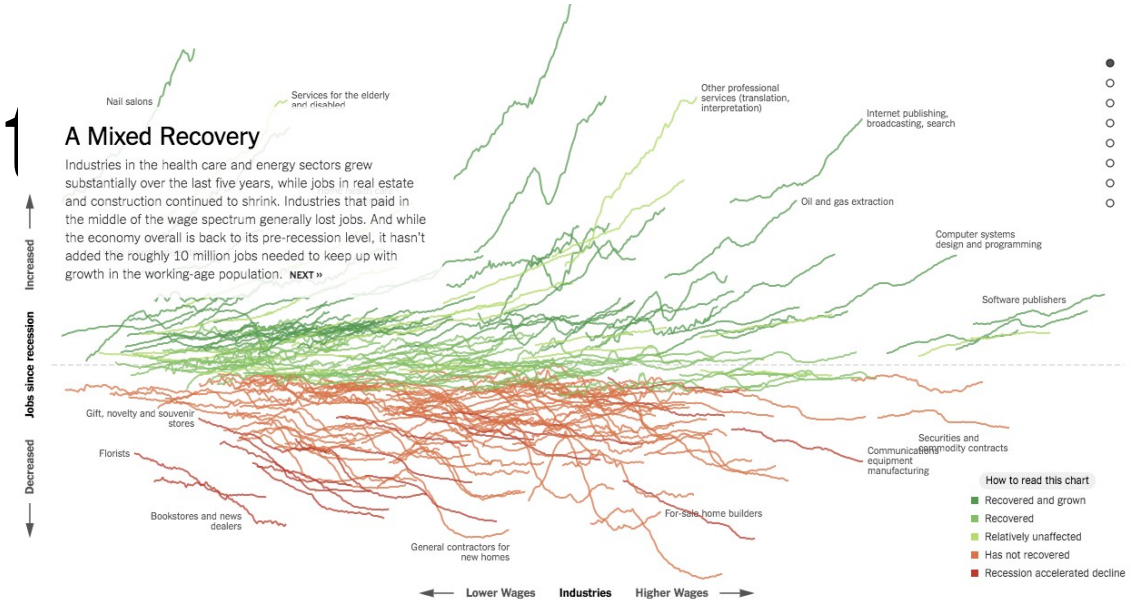


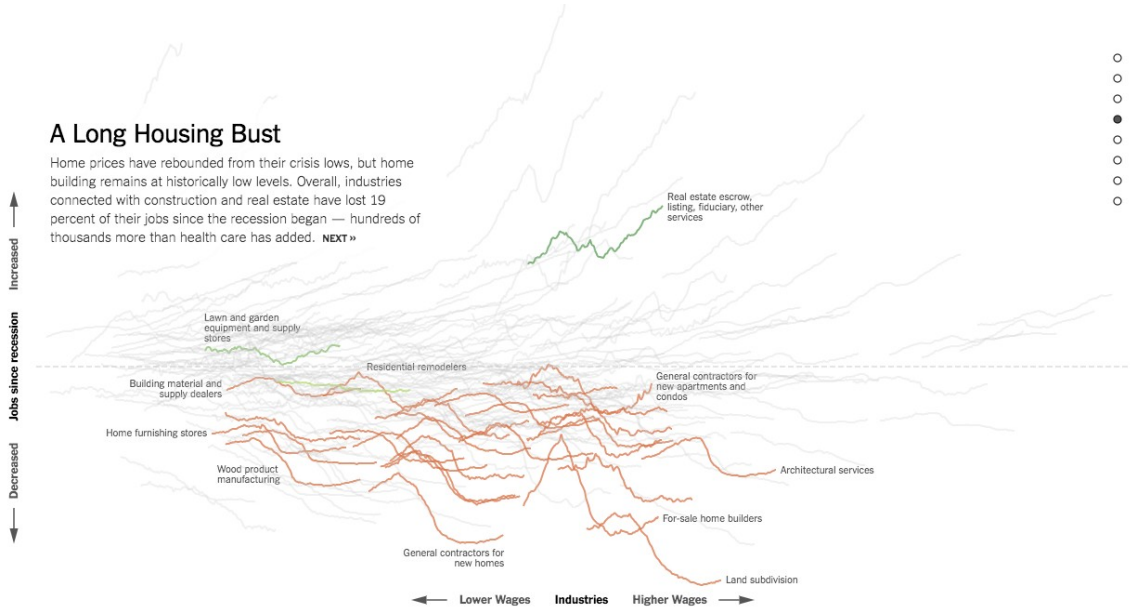
Abstraction

- These {action, target} pairs are good starting points for vocabulary
  - But often, you will need more precision!
- Rule of thumb
  - Systematically remove all domain jargon
- Interplay: task and data abstraction
  - Need to use data abstraction within task abstraction
    - To specify your targets!
    - But task abstraction can lead you to transform the data
  - Iterate back and forth
    - First pass data, first pass task, second pass data, ….

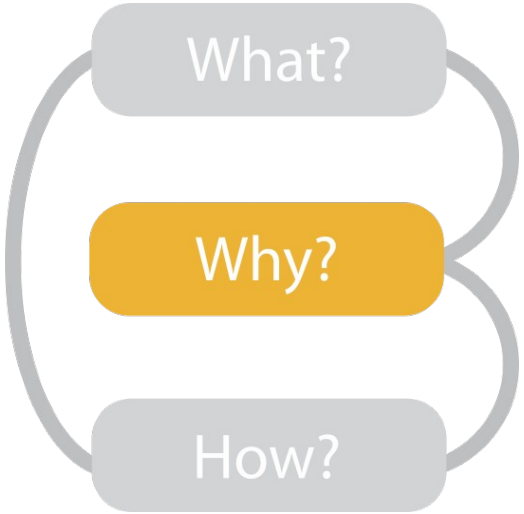
Examples: Job market

- Trends
  - how did the job market develop since the recession overall?
- Outliers
  - Real estate related jobs





https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html



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

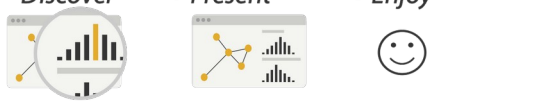
Why?

Actions

Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*



→ Produce

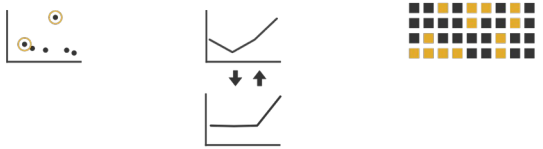
→ *Annotate* → *Record* → *Derive*

Search

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

Query

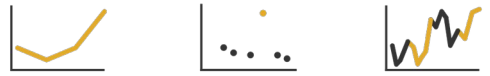
→ Identify → Compare → Summarize



Targets

All Data

→ Trends → Outliers → Features

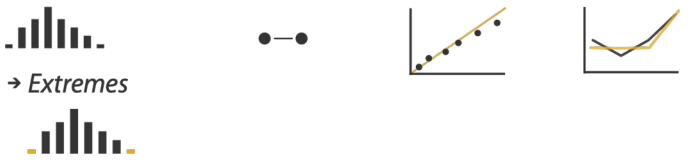


Attributes

→ One → Many

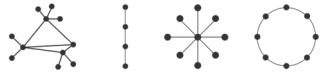
→ *Distribution* → *Dependency* → *Correlation* → *Similarity*

→ *Extremes*



Network Data

→ Topology



→ *Paths*

Spatial Data

→ Shape