Introduction to Advanced Data Visualization Techniques (ADVT)

IBM Predicts Demand For Data Scientists Will Soar 28% By 2020

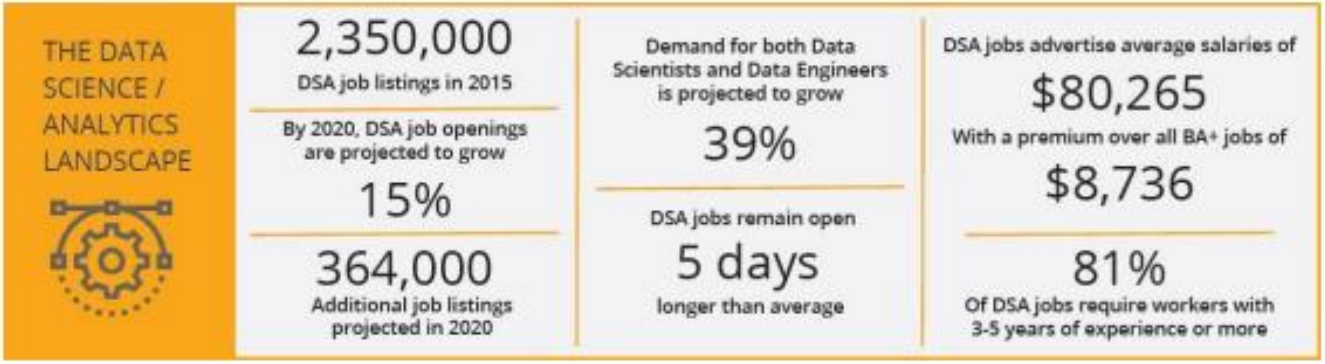
Louis Columbus, CONTRIBUTOR
FULL BIO
Opinions expressed by Forbes Contributors are their own.

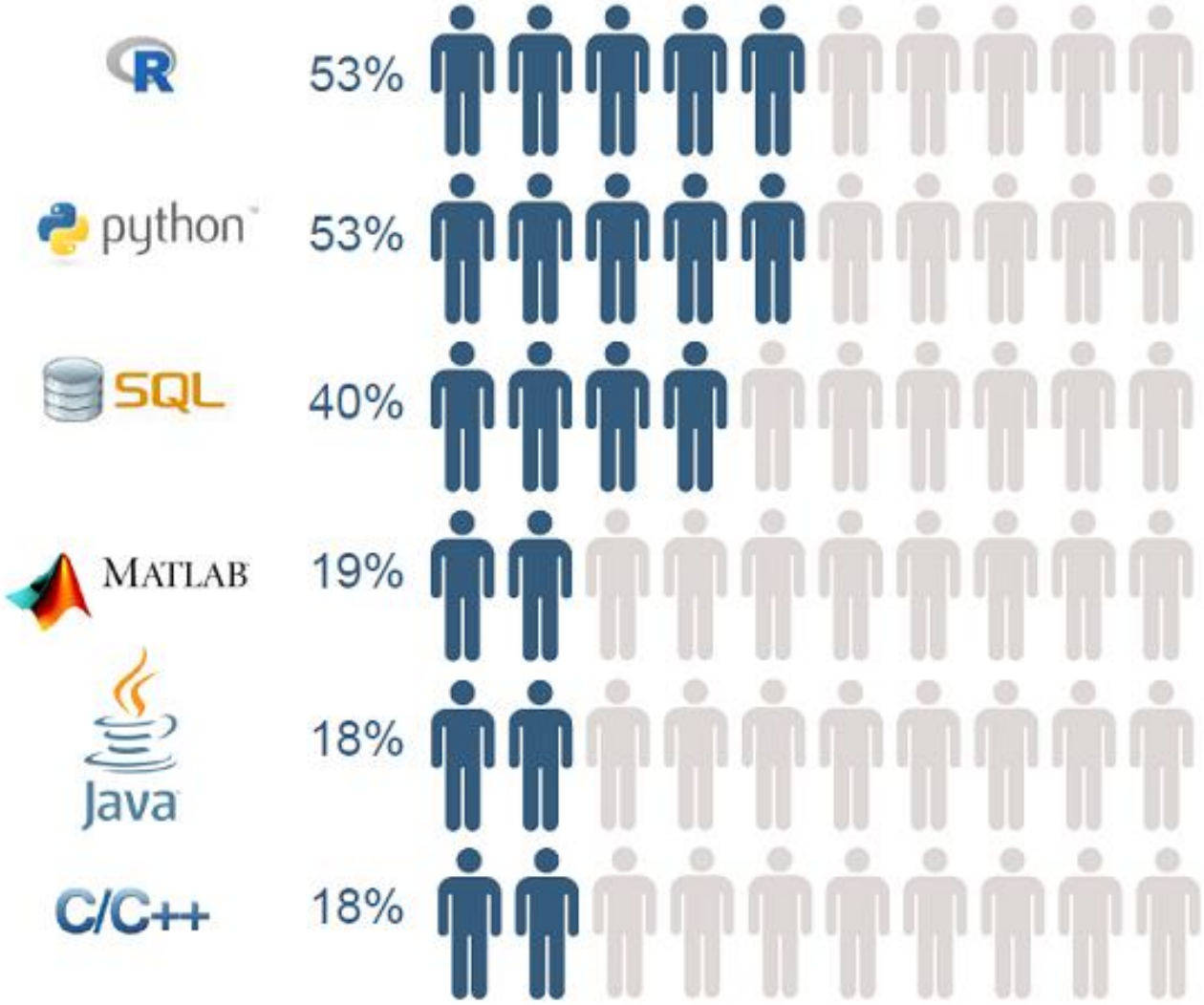
- Jobs requiring machine learning skills are paying an average of $114,000. Advertised data scientist jobs pay an average of $105,000 and advertised data engineering jobs pay an average of $117,000.
- 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.

By 2020, the number of jobs for all US data professionals will increase by 364,000 openings to 2,720,000 according to IBM.

- **59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.** DSA jobs factor most prominently in the Finance and Insurance industry, where they account for 19% of all openings. The Professional Services and IT industries follow with 18% and 17% relative demand for DSA jobs, respectively. The following graphic provides an analysis of DSA job category demand by industry.

- **By 2020 the number of Data Science and Analytics job listings is projected to grow by nearly 364,000 listings to approximately 2,720,000** The following summary graphic from the study highlights how in-demand data science and analytics skill sets are today and are projected to be through 2020.




R
53%
python
53%
SQL
40%
MATLAB
19%
Java
18%
C/C++
18%

Skill Name	Average Salary
MapReduce	$115,907
PIG	$114,474
Machine Learning	$112,732
Apache Hive	$112,242
Apache Hadoop	$110,562
Big Data	$109,895
Data Science	$107,287
NoSQL	$105,053
Predictive Analytics	$103,235
MongoDB	$101,323

Introduction to Data Visualization

- ❑ Overview of data visualization
- ❑ Data Abstraction
- ❑ Task Abstraction
- ❑ Analysis: Four Levels for Validation

Text Book

Tamara Munzer**, Visualization Analysis and Design** -, CRC Press 2014 . **(Chapter 1, 2,3 and 4)**

Visualisation

The broader field of visualisation has three main sub-fields:

- *SciVis*: Scientific Visualisation (SciVis) typically involves concrete (3d) objects, for example a medical scan of part of the body, or a simulation of air flow around an aircraft wing. SciVis visualisations often depict flows, volumes, and surfaces in (3d) space.
- *GeoVis*: Geographic Visualisation (GeoVis) is map-based. The data typically has inherent 2d or 3d spatial coordinates, and is generally shown in relation to a map.
- *InfoVis*: Information Visualisation (InfoVis) deals with *abstract* information structures, such as hierarchies, networks, or multidimensional spaces.

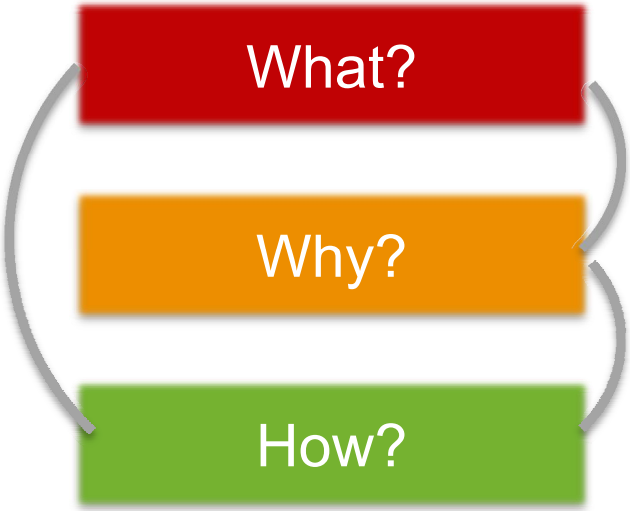
Data Visualisation (DataVis) = InfoVis + GeoVis.

Visual Analytics = DataVis (frontend) + Analytics (backend).

Data & Task Abstraction

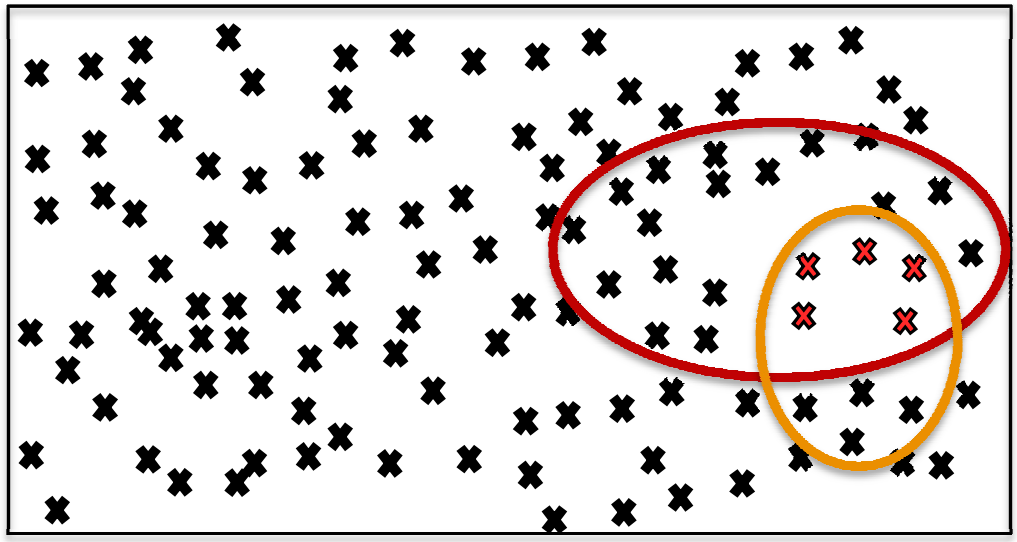
Analysis: what, why, and how

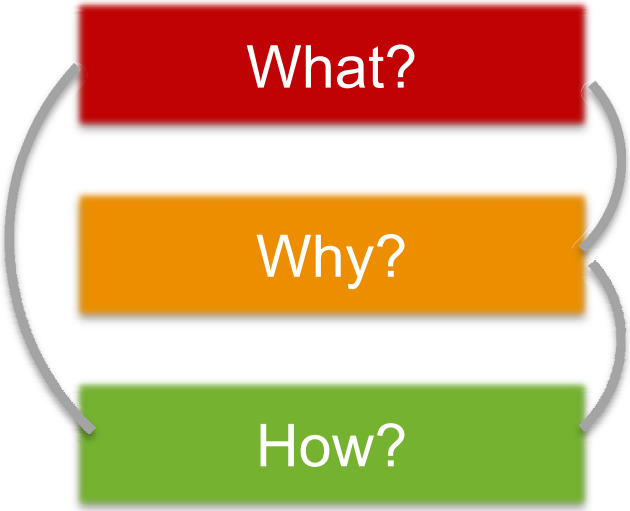
- **What** is shown?
  - Data abstraction
- **Why** is the user looking at it?
  - Task abstraction
- **How** is it shown?
  - Visualization + interaction



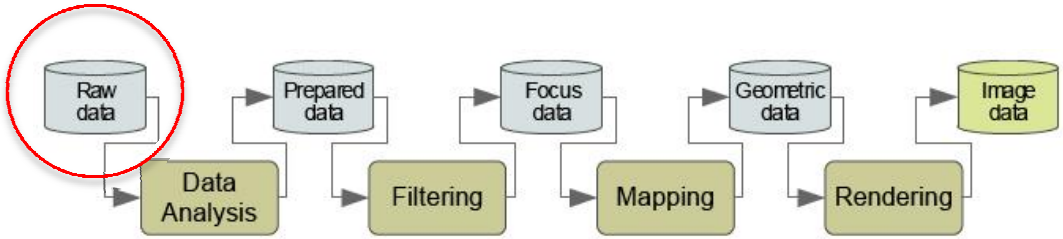
Analysis: what, why, and how

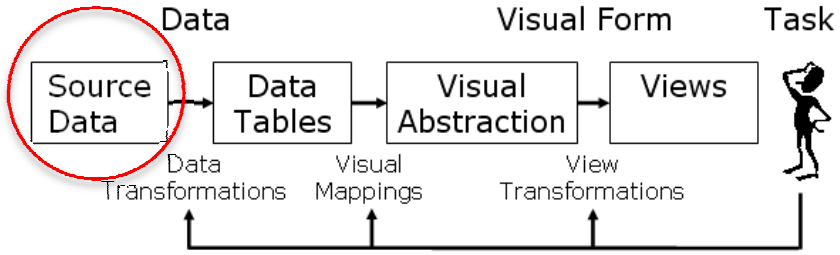
- Answering what & why serve as constraints on design space

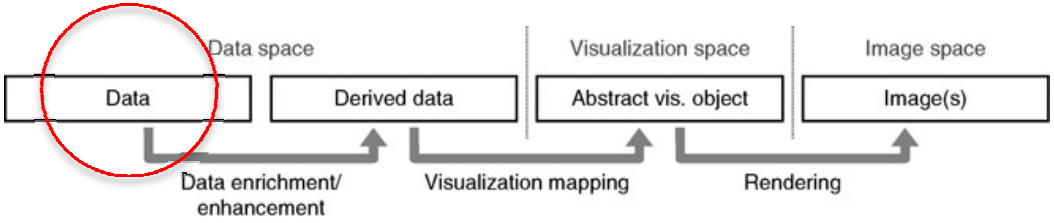


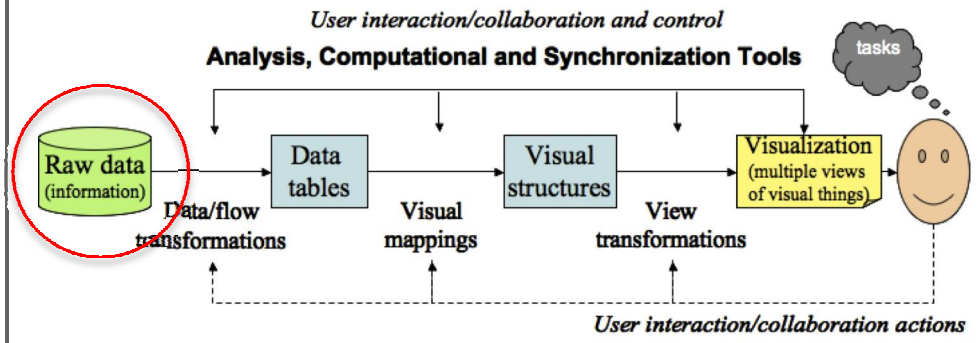


It all starts with data









Data collection and generation

- Big topic! Not addressed in this course.
- Many data collection methods
  - Sensors
  - Logs
  - Experiments
  - Human-generated data
  - Surveys

Data transformation/processing

- Big topic! Not fully addressed in this course.
- Data cab ne transformed in many ways
  - aggregated
  - collated
  - Sub-setted
  - filtered
  - Reshaped
  - Change of scale
  - …

Data abstraction

- Is about understanding your data
- Type
  - Number?
  - Category?
- Organization
  - Table?
  - Network?
- Semantics
  - Meaning of it

Data abstraction

14, 2.6, 30, 30, 15, 100001

(14, 2.6, 30) , (30, 15, 100001)

Point A Point B

(14, 2.6) , (30, 30) , 15 , 100001

Point A Point B Links Weight

Data abstraction

14, 2.6, 30, 30, 15, 100001

Two necessary crosscutting piece of information to move beyond guesswork

- **Semantics**
- **Types**

Data abstraction

- **Semantic** of data is its real-world meaning
  - Link
    - Road?
    - Friendship?
    - Hierarchy?
  - Word
    - Frist name?
    - company name?
    - fruit?
  - Number
    - Day of a month?
    - age?
    - height?

Data abstraction

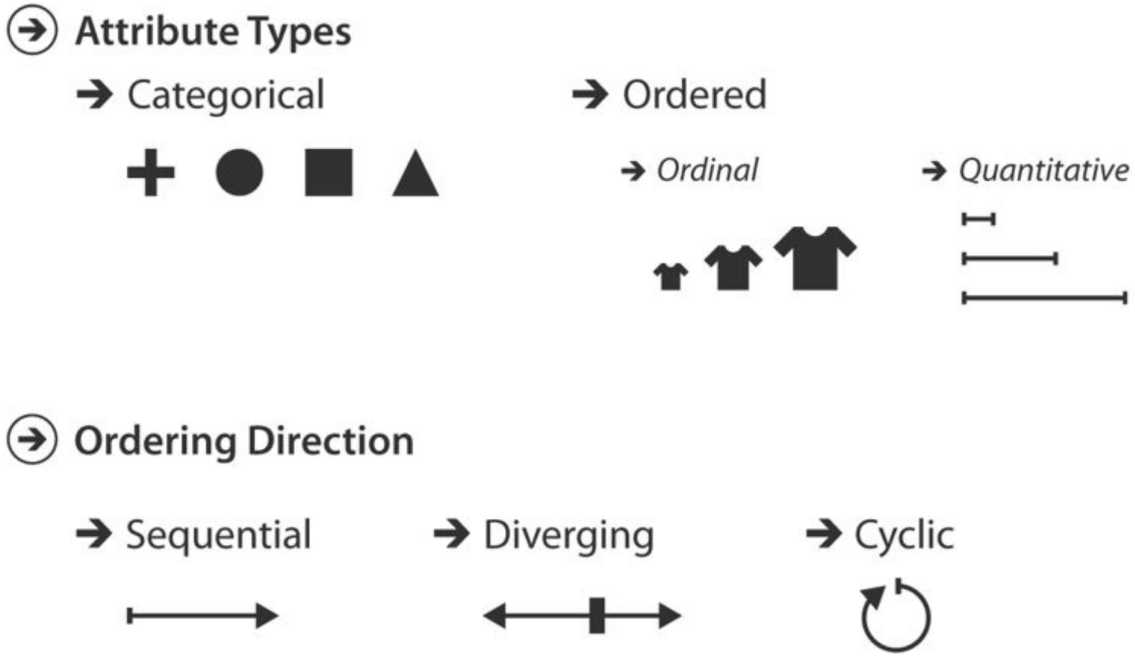
- **Type** of data is its structural or mathematical interpretation
  - **Data type** (i.e. what kind of thing it is?)
    - Item?
    - attribute?
    - …
  - **Dataset type**
    - Table?
    - tree?
    - filed?

Data types

- **Attribute** is something that can be measured, observed, or logged
  - Salary, price, protein expression level,…
- **Item** is an individual entity that is discreet
  - People, stocks, coffee shops, genes,…
- **Link** is a relationship between items
- **Position** a location in (2D) or (3D) space
  - Latitude-longitude
- **Grid** is strategy used for sampling continuous data

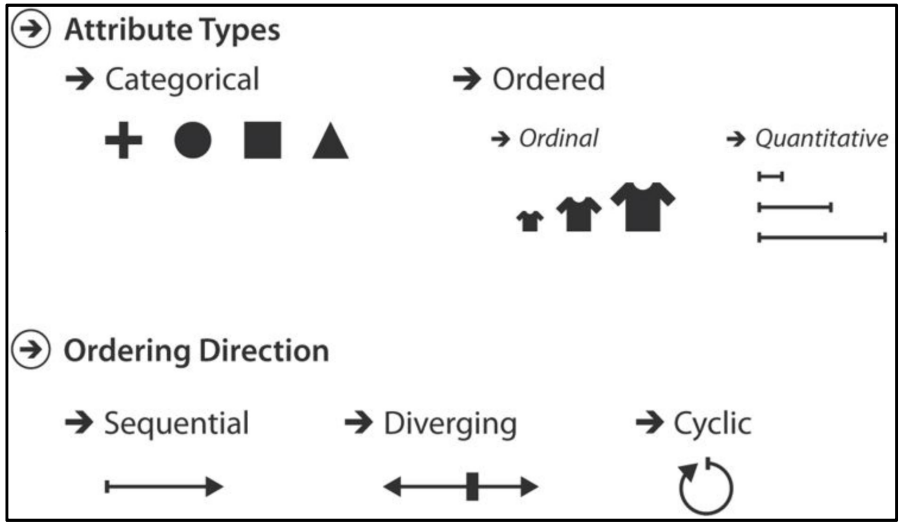
Attribute types

- **Attribute** is something that can be measured, observed, or logged



Attribute types

- **Categorical** (e.g., gender, race, eye color)
- **Ordinal** (e.g., edu level, position in a race)
- **Quantitative** (e.g., age, height, weight)



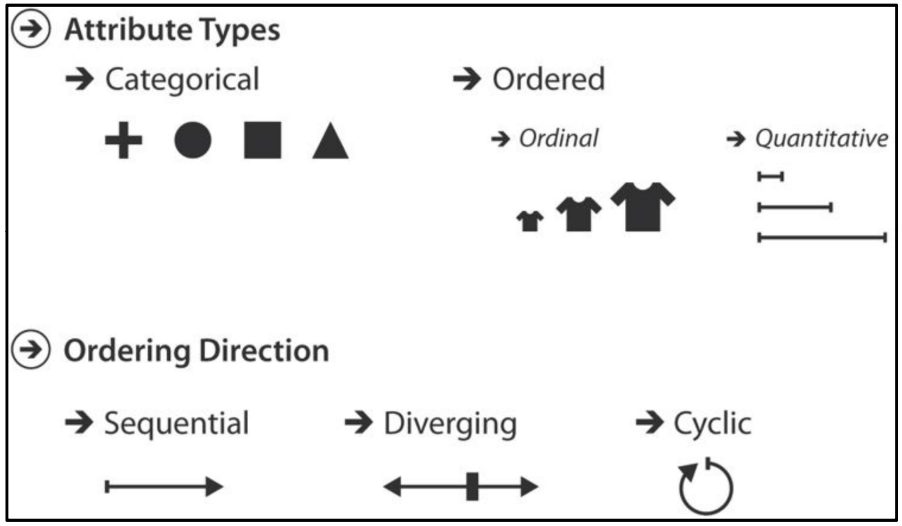
Attribute types

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

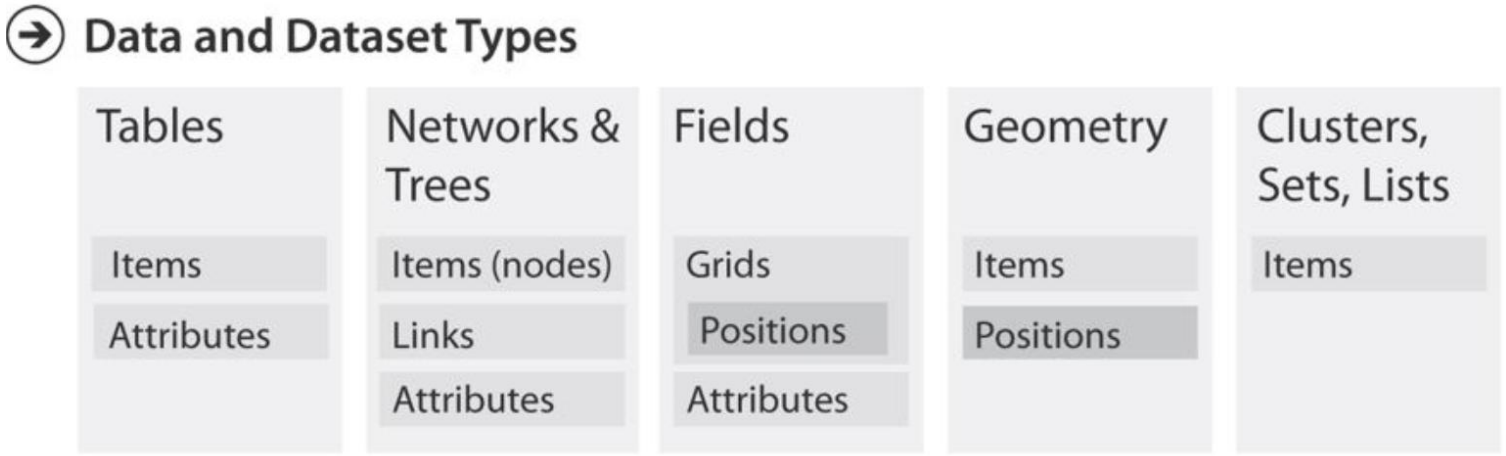
- **Sequential**: e.g., age, height, weight.
- **Diverging**: e.g., temperature, altitude.
- **Cyclic**: e.g., hour, week, month.



Data types

- **Attribute** is something that can be measured, observed, or logged
  - Salary, price, protein expression level,…
- **Item** is an individual entity that is discreet
  - People, stocks, coffee shops, genes,…
- **Link** is a relationship between items
- **Grid** is strategy used for sampling continuous data
- **Position** is spatial data, giving a location in 2D or 3D space
  - Latitude-longitude

Dataset types

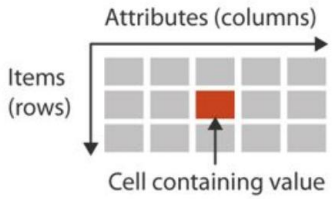


- Four basic database types
- Combination of data types
- In real-word, complex combination of these types are common

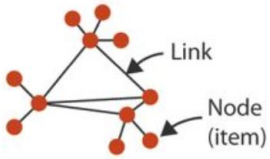
Dataset types

Dataset Types

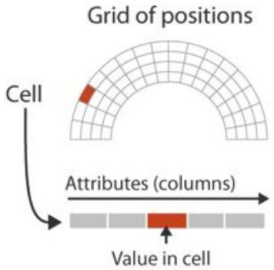
→ Tables



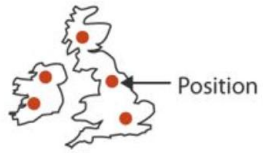
→ Networks



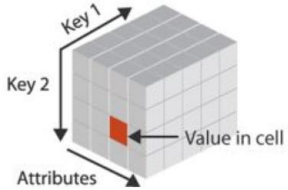
→ Fields (Continuous)



→ Geometry (Spatial)



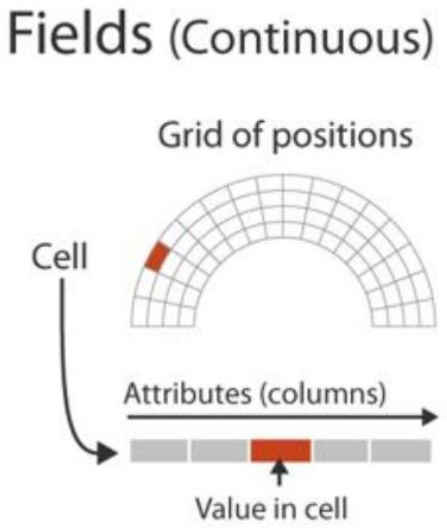
→ *Multidimensional Table*



→ *Trees*

Dataset types

- **Fields**
  - Grid
    - Positions
  - Attributes: values associated with cells
  - Cells: contains measurements or calculations from a continuous domain

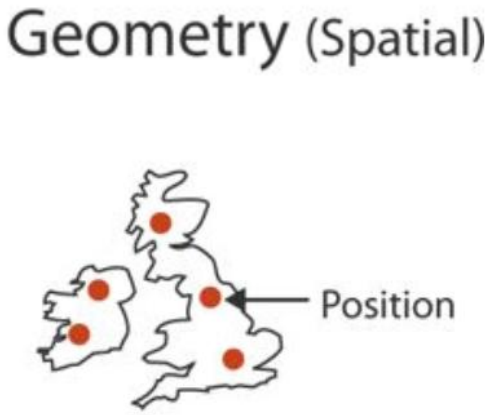


Fields (spatial)

- Medical scan of a human body containing measurements indicating the density of tissue at many sample points
- Animals movements
- Election results in counties
- Simulation of air turbulence

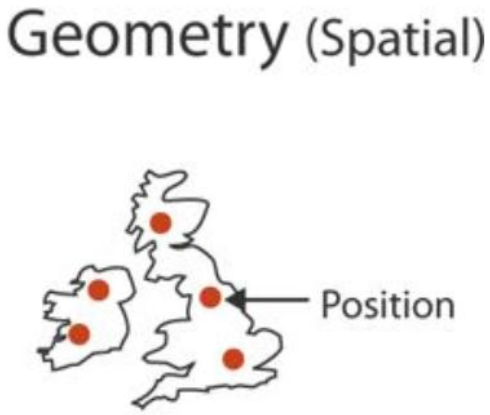
Dataset types

- **Geometry**
  - Items
  - Positions
- Specifies information about the shape of items with explicit spatial positions. The items could be points, or one-dimensional lines or curves, or 2D surfaces or regions, or 3D volumes.



Dataset types

- Geometry datasets are intrinsically spatial
- They typically occur in the context of tasks that require shape understanding
- Geometry datasets do not necessarily have attributes, in contrast to the other three basic dataset types



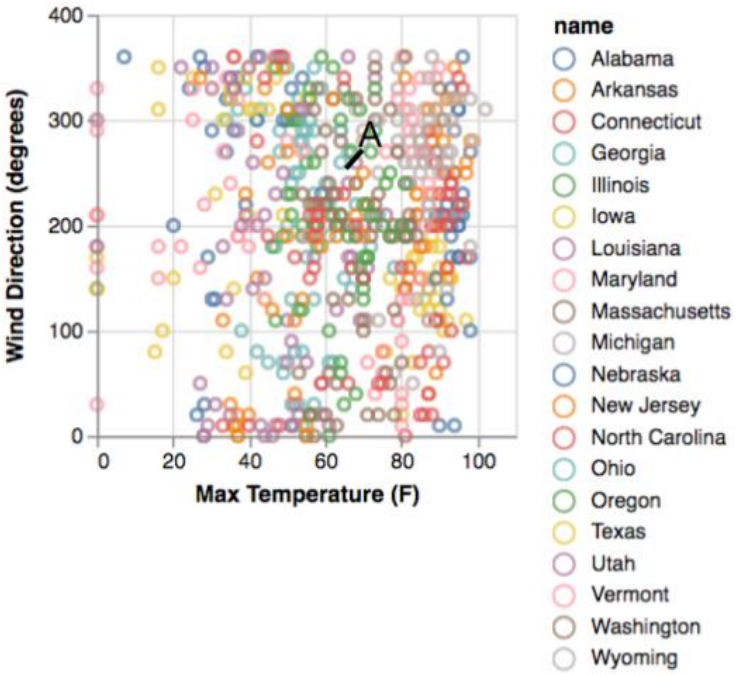
Cardinality

- Effectiveness of data visualization is impacted by:
  - Visual encoding
  - Type of task
  - Distribution of data

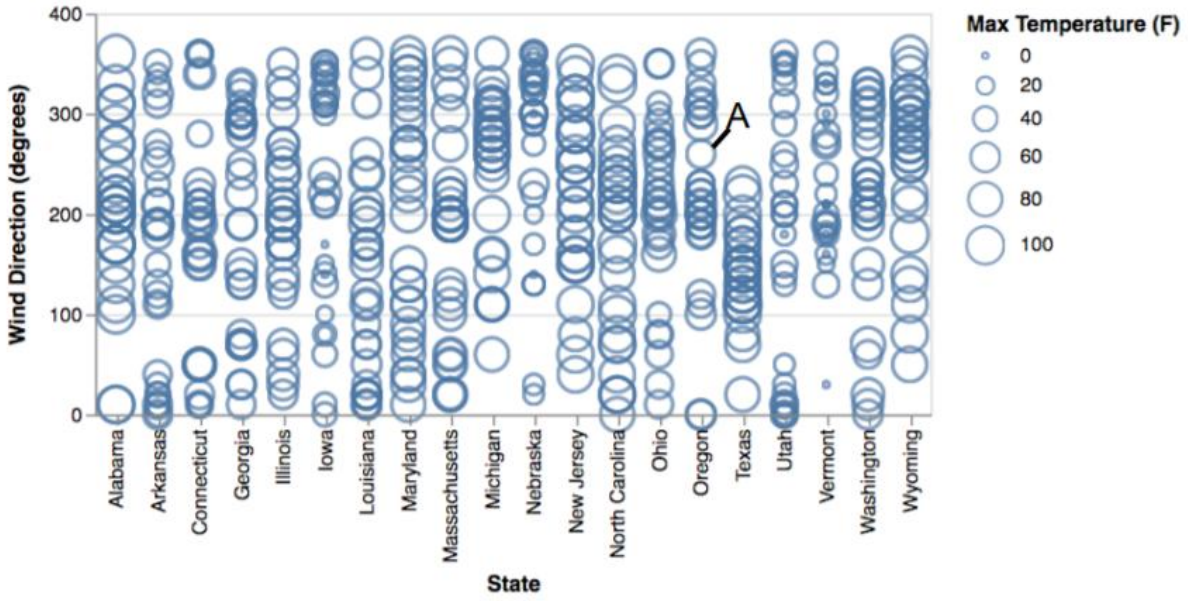
- Various measures describe data distribution
  - Cardinality: number of unique values for an attribute
  - Entropy
  - Clusterdness
  - …

Cardinality

$Q_1 : y, Q_2 : x, N : color$



$Q_1 : size\ Q_2 : y, N : x$



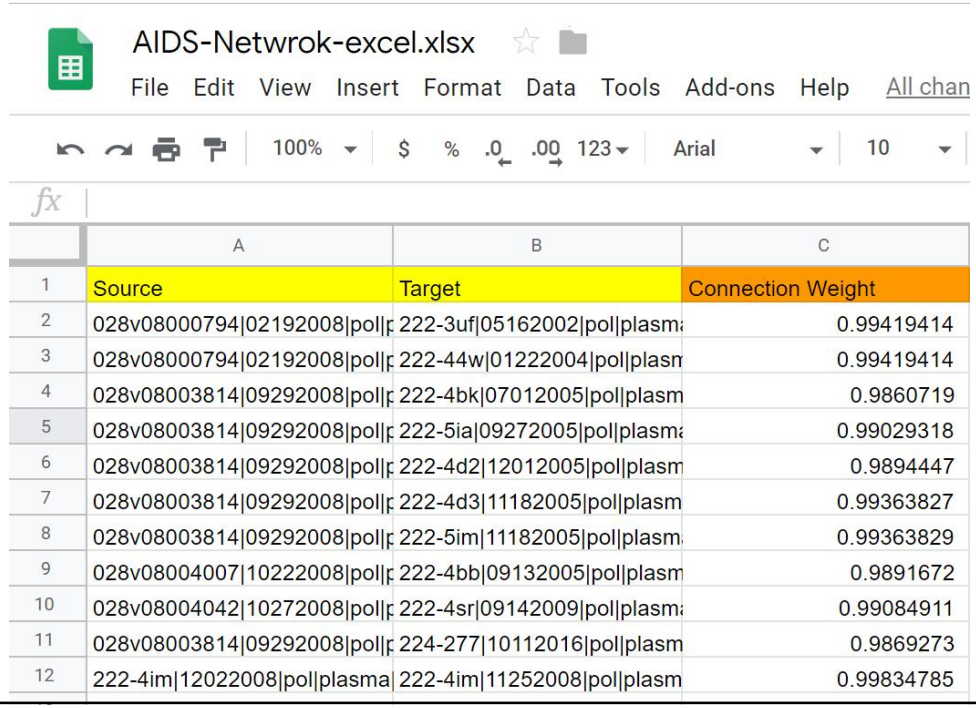
Data abstraction exercise

- Visit Moodle to find exercise!

Questions?

Data abstraction

- Data types & datasets→ templates to help you understand & describe your data


AIDS-Netwrok-excel.xlsx

File Edit View Insert Format Data Tools Add-ons Help All chan

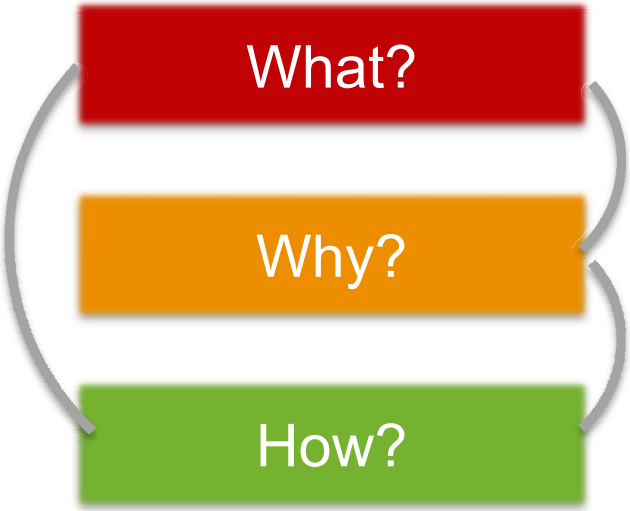
100% $ % .0 .00 123 Arial 10

fx

	A	B	C
1	Source	Target	Connection Weight
2	028v08000794|02192008|pol|p	222-3uf|05162002|pol|plasma	0.99419414
3	028v08000794|02192008|pol|p	222-44w|01222004|pol|plasm	0.99419414
4	028v08003814|09292008|pol|p	222-4bk|07012005|pol|plasm	0.9860719
5	028v08003814|09292008|pol|p	222-5ia|09272005|pol|plasma	0.99029318
6	028v08003814|09292008|pol|p	222-4d2|12012005|pol|plasm	0.9894447
7	028v08003814|09292008|pol|p	222-4d3|11182005|pol|plasm	0.99363827
8	028v08003814|09292008|pol|p	222-5im|11182005|pol|plasma	0.99363829
9	028v08004007|10222008|pol|p	222-4bb|09132005|pol|plasm	0.9891672
10	028v08004042|10272008|pol|p	222-4sr|09142009|pol|plasma	0.99084911
11	028v08003814|09292008|pol|p	224-277|10112016|pol|plasm	0.9869273
12	222-4im|12022008|pol|plasma	222-4im|11252008|pol|plasm	0.99834785

Task abstraction

- **Why** is the user looking at it?
  - Task abstraction
- Goal: transform user task from a domain specific language into a high-level concise representation



Task abstraction

- A biologist studying immune system response might describe her task as:

  "I want to see if the results for the tissue samples treated with LL-37 match up with the ones without the peptide"

Compare values between two groups

Task abstraction

- Business manager:

  “I want to see if new marketing strategy was successfully resulted in selling more products in the home appliances category”

Trends for a group of products

Task abstraction

- You need to collect user's tasks (questions) first
  - Interview
  - Brainstorming
  - Focus groups
  - Exploratory prototypes
  - Observation
  - Surveys
  - …

Task abstraction

Domain questions

"why are there so many failed requests today?"

Abstract tasks

Identify

Analyze

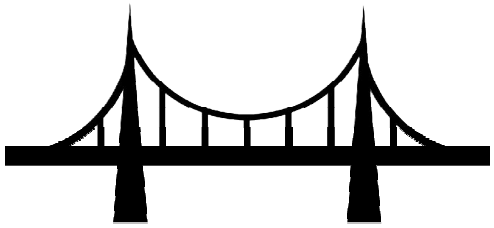
…

Task abstraction

Task classification

Domain questions

“why are there so may failed requests today?”



Abstract tasks

Identify extremums

Analyze outliers

…

Task abstraction

- There are many classifications and taxonomies for visualization tasks
- Low-level
  - Example: Retrieve value, Filter, Order, Find extremums,…
  - Further reading: Low-Level Components of Analytic Activity in Information Visualization (Amar et al., 2005)
- High-level
  - Example: Explore, Describe, Explain,…
  - Further reading: Bridging From Goals to Tasks with Design Study Analysis Reports

Task abstraction

Specificity

Spec\#Pops	Explore	Describe	Explain	Confirm
Single	**Discover Observation** I: Data only O: Obs (Item or Aggregate)	**Describe Observation (Item)** I: Obs (Item) O: Pop Defn (all attributes) **Describe Observation (Aggregate)** I: Obs (Aggregate) O: Pop Defn (all attributes)	**Identify Main Cause (Item)** I: Obs (Item) O: Pop Defn (dominant attribute) **Identify Main Cause (Aggregate)** I: Obs (Aggregate) O: Pop Defn (dominant attribute)	**Collect Evidence** I: Hypothesis O: Confirm / Reject
Multiple		**Compare Entities** I: Pop Defn O: Pop Contrasts (similarities and differences)	**Explain Differences** I: Pop Defn O: Pop Contrasts (differences)	**Evaluate Hypothesis** I: Pop Defn; Hypothesis O: Confirm / Reject

#Population

We will follow task classification by LAM, Tory & Munzner, 2017

Task abstraction

Spec\#Pops	Explore	Describe	Explain	Confirm
Single	**Discover Observation** I: Data only O: Obs (Item or Aggregate)	**Describe Observation (Item)** I: Obs (Item) O: Pop Defn (all attributes) **Describe Observation (Aggregate)** I: Obs (Aggregate) O: Pop Defn (all attributes)	**Identify Main Cause (Item)** I: Obs (Item) O: Pop Defn (dominant attribute) **Identify Main Cause (Aggregate)** I: Obs (Aggregate) O: Pop Defn (dominant attribute)	**Collect Evidence** I: Hypothesis O: Confirm / Reject
Multiple		**Compare Entities** I: Pop Defn O: Pop Contrasts (similarities and differences)	**Explain Differences** I: Pop Defn O: Pop Contrasts (differences)	**Evaluate Hypothesis** I: Pop Defn; Hypothesis O: Confirm / Reject

Specificity is about scope of analysis, breadth VS depth

 # population is about the scope of data selection, single (all data, a single subset), Multiple (two or more subsets of data)

Task abstraction

Spec\#Pops	Explore	Describe	Explain	Confirm
Single	**Discover Observation** I: Data only O: Obs (Item or Aggregate)	**Describe Observation (Item)** I: Obs (Item) O: Pop Defn (all attributes) **Describe Observation (Aggregate)** I: Obs (Aggregate) O: Pop Defn (all attributes)	**Identify Main Cause (Item)** I: Obs (Item) O: Pop Defn (dominant attribute) **Identify Main Cause (Aggregate)** I: Obs (Aggregate) O: Pop Defn (dominant attribute)	**Collect Evidence** I: Hypothesis O: Confirm / Reject
Multiple		**Compare Entities** I: Pop Defn O: Pop Contrasts (similarities and differences)	**Explain Differences** I: Pop Defn O: Pop Contrasts (differences)	**Evaluate Hypothesis** I: Pop Defn; Hypothesis O: Confirm / Reject

Population Definition (Pop Defn) describes how to select (filter) data to get the desired subset

Task abstraction

“I want to know more about how we handled booking requests today”

Discover observation

A subset of data for today’s requests

“Many failed request today!”

“what type of requests were failed the most?”

Describe observation

A subset of data for today’s failed requests

“Class Z and R failed the most! and error code is 100 that means…”

Task abstraction

"Why class Z & R requests failed today?"

Identify main cause

A subset of data for today's failed requests of type Z & R

"100% of failed requests are associated with delayed flight number 4360 in location A"

"We have a problem with maintenance time that causes delays in many flights from location A"

Confirm

A subset of data for today's failed requests, locations, maintenance times …

"Reject"

Task abstraction

- Tree dataset: items (nodes) and links
- “I want to be able to present a path traced between two nodes of interest to a colleague”

  - Find two interesting nodes == Explore, all nodes
    - What defines interesting?
    - # number of childeren

  - Find the path between those two nodes == Describe, observed nodes & their connections

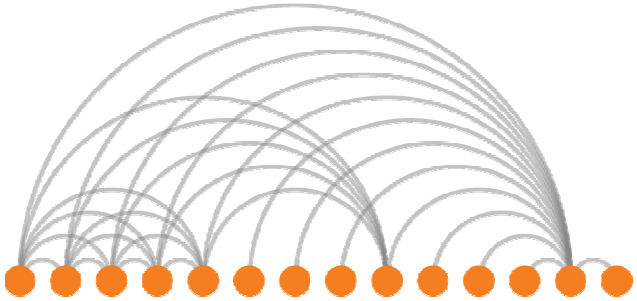
Task abstraction

- Find nodes → find path→ present

Ranked list of nodes

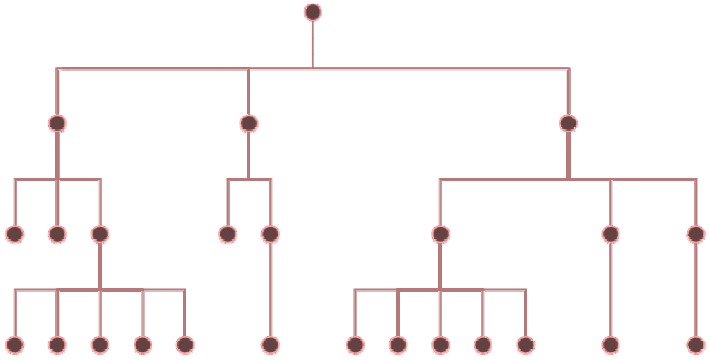
Ordering supports exploration

Arc tree



Ordering supports exploration

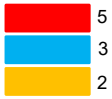
Tree, icicle layout

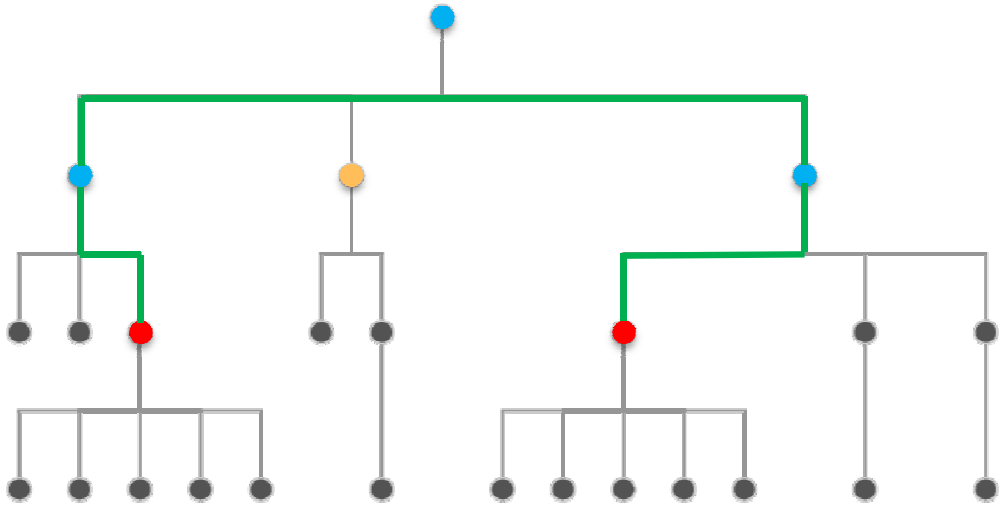


Layout, supports description

Task abstraction

- Find nodes → find path→ present



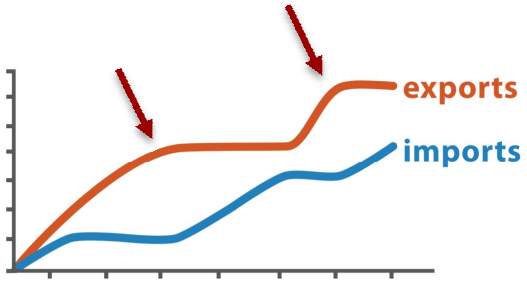


Task abstraction

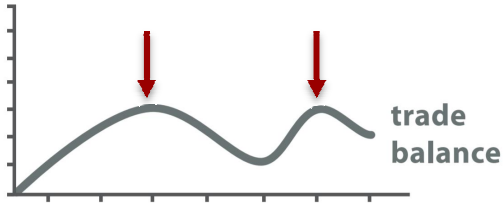
- It is not easy! It takes lots of practice and (failed) attempts to master it
- It is highly iterative!
- People are not very good at describing their tasks
  - Incomplete
  - Incorrect
  - Some times even contradicting

Derive: Crucial design choice

- Don’t just draw what you are given!
  - Decide what the right thing to show is
  - Create it with a series of transformations from original dataset
  - Draw that!
- Deriving is one of our major strategies for handing complexity



Original Data



trade balance = exports – imports

Derived Data