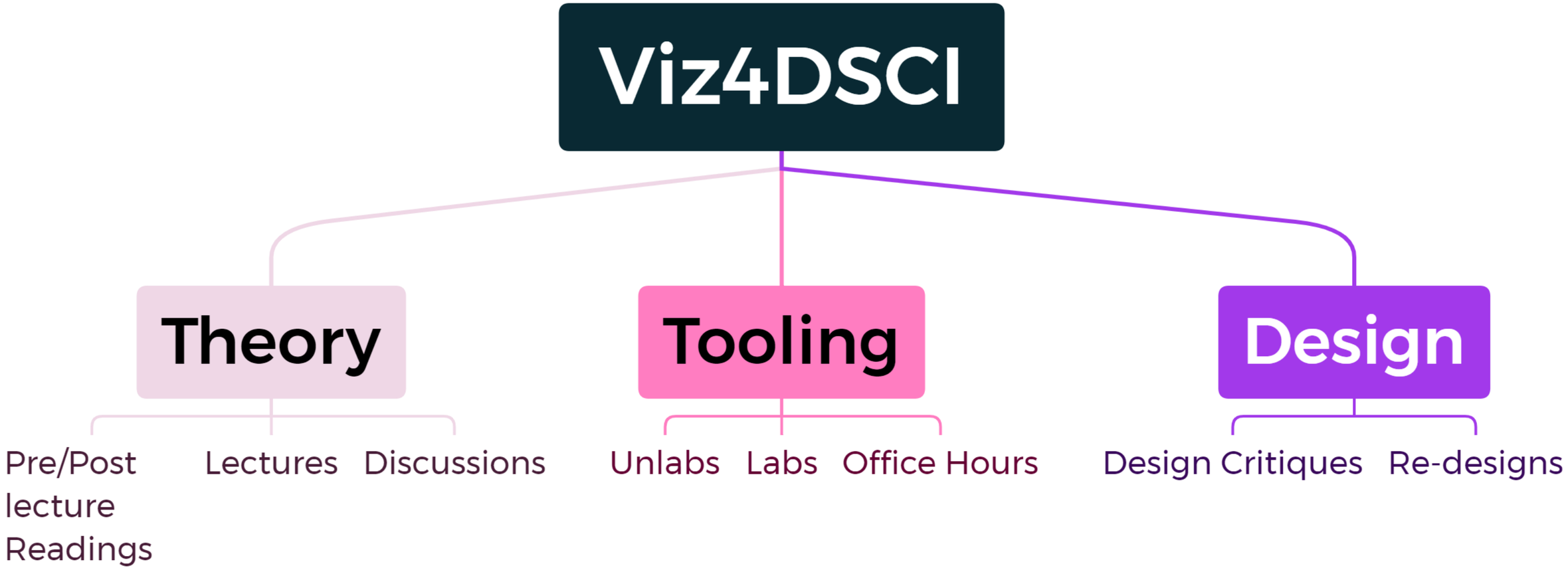
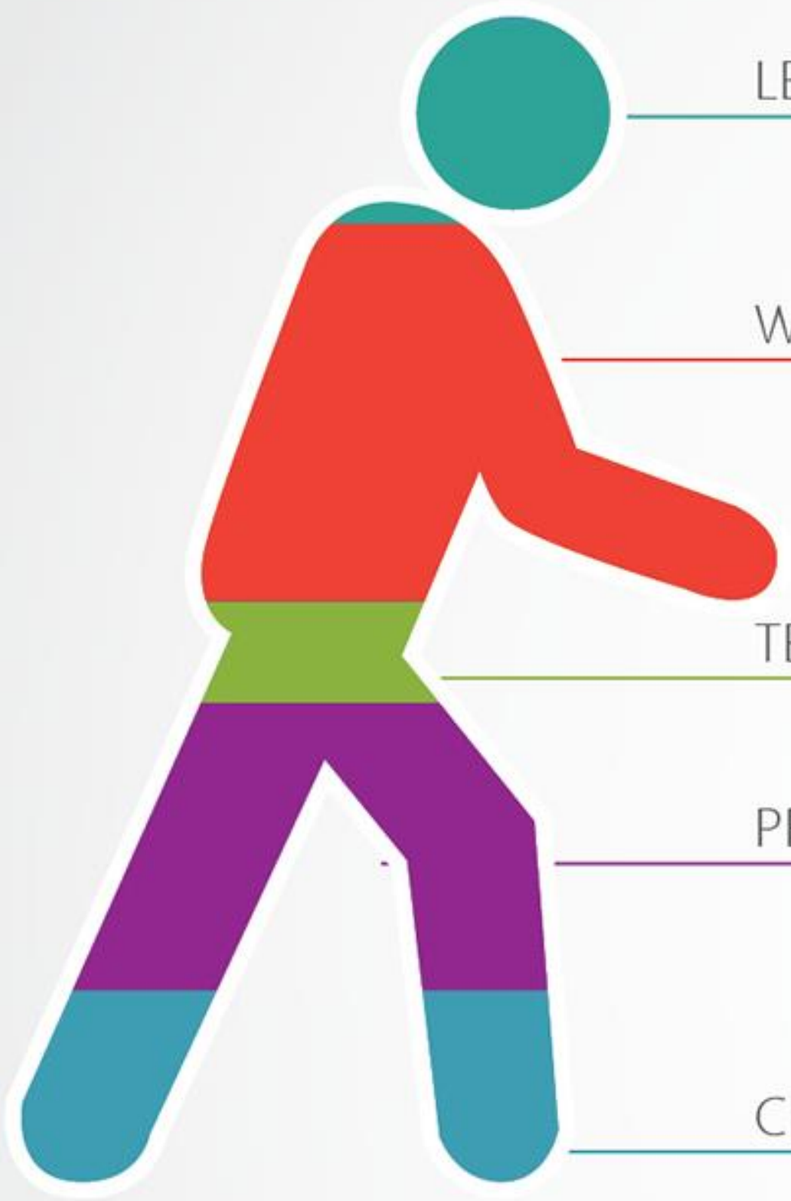
Visualization for Data Science
Data Abstraction


Viz4DSCI
Theory
Tooling
Design
Pre/Post lecture Readings
Lectures
Discussions
Unlabs
Labs
Office Hours
Design Critiques
Re-designs

HOW **BABY BOOMERS** DESCRIBE THEMSELVES



LEADERS **40%**

WILLING TO LEARN **61%**

TECH-SAVVY **22%**

I HAVE A SMART PHONE, DON'T DISCOUNT MY TECHNICAL ABILITIES.

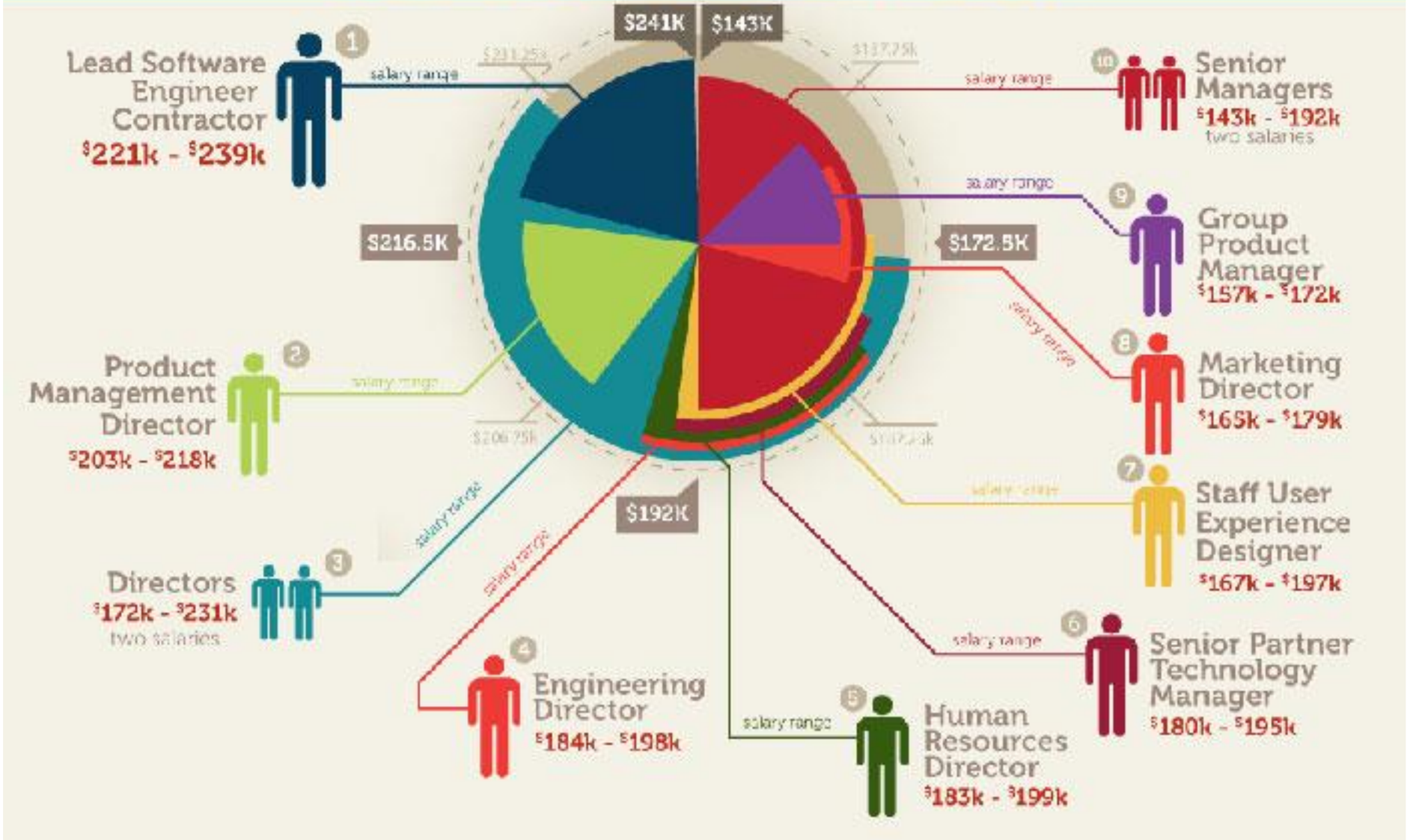
PEOPLE-SAVVY **78%**

CREATIVE **42%**

MY WORK EXPERIENCE HAS MADE ME RESOURCEFUL.

top 10 salaries at Google

RANGE FROM $143,000 TO $241,000 PER YEAR.



Learning Goals

- Describe the difference between how the phrases “dataset types”, “data types” are used in vis. Literature as opposed to programming
- Describe the characteristics of data
- Differentiate between the different types of data and dataset types

Data Types are the fundamental units in which observed **phenomena** are **represented**. The structural or mathematical interpretation of data

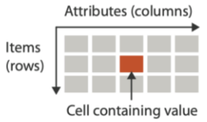
Data Types

➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

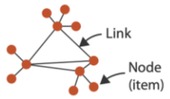
Dataset is any collection of information that is the target of analysis

Dataset Types

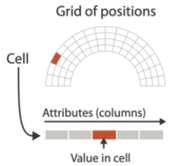
➔ Tables



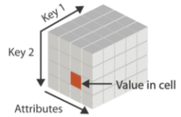
➔ Networks



➔ Fields (Continuous)



➔ *Multidimensional Table*



➔ *Trees*

➔ Geometry (Spatial)

Data and Dataset Types

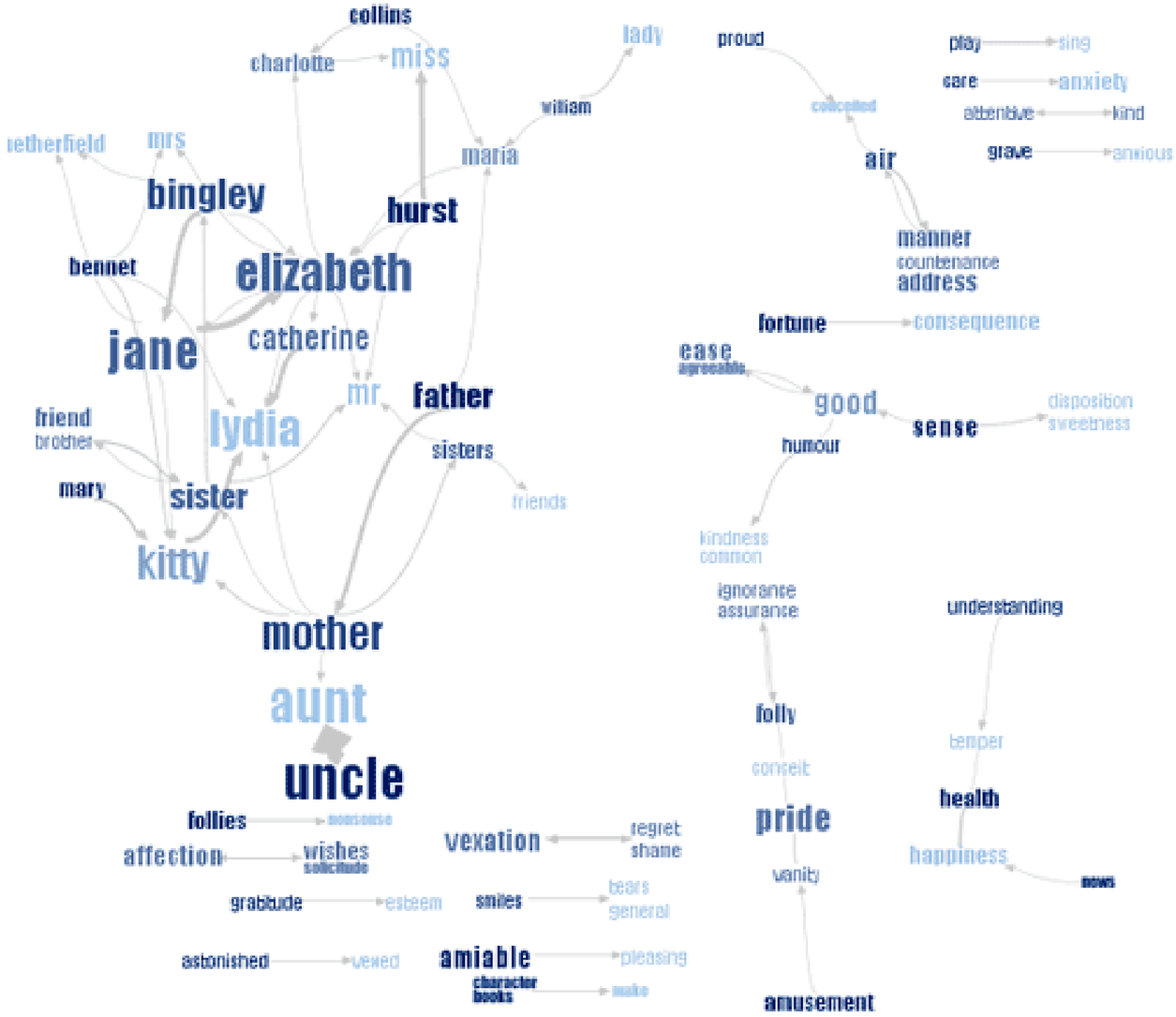
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Characterization

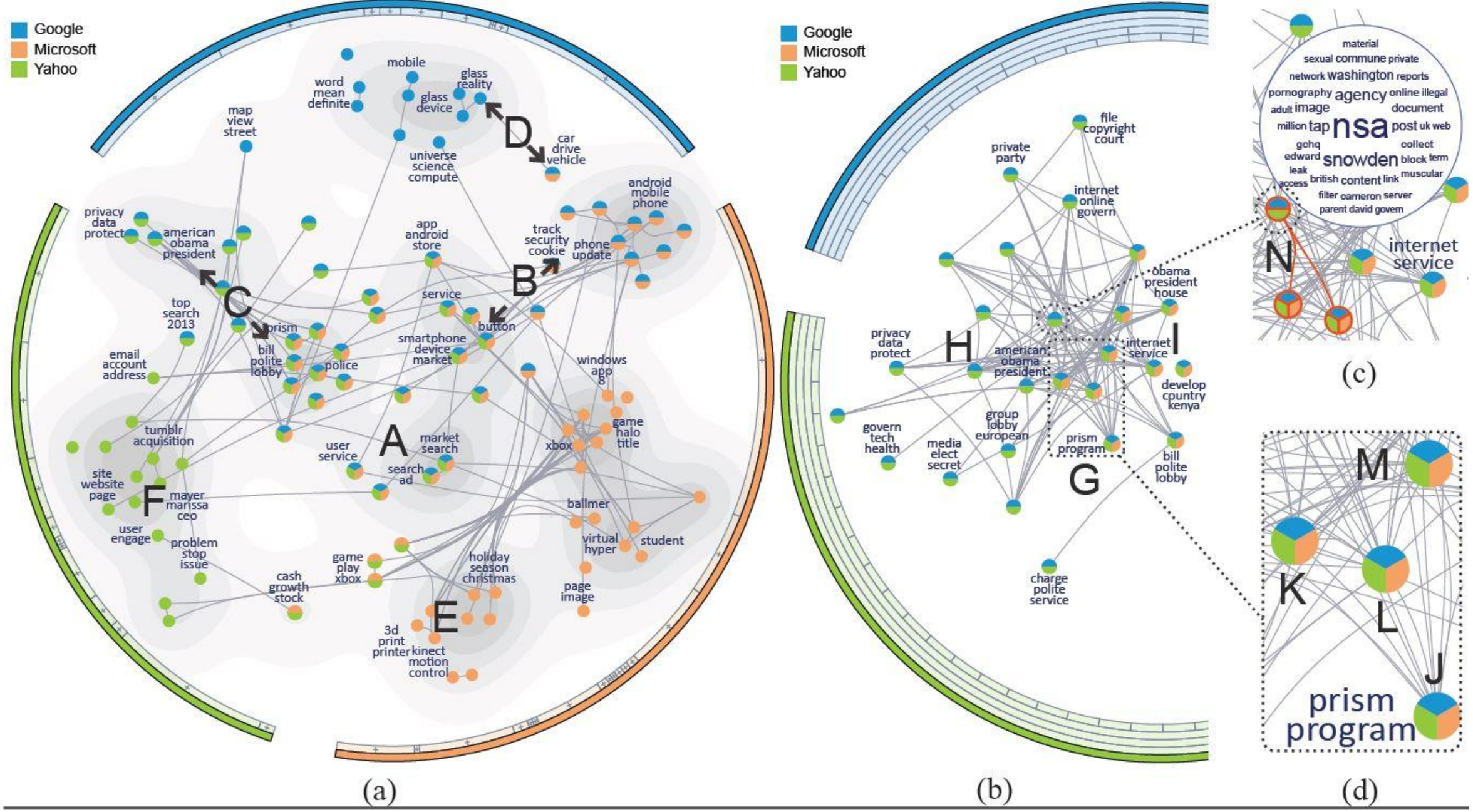
Data is characterized by its

- Size (volume)
- Speed at which it generated (velocity)
- Quality (veracity)
- Structure

Text Example



http://hint.fm/projects/phrasenet/



http://shixialiu.com/publications/TopicPanorama/index.htm

http://shixialiu.com/publications/TopicPanorama/video_eng.mp4

What does data mean?

14, 2.6, 30, 30, 15, 100001

What does this sequence of six numbers mean?

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??

What does data mean?

Basil, 7, S, Pear

What about this data?

What does data mean?

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 city had 7 inches of snow cleared by the Pear Creek Limited snow removal service
- lab rat Basil made 7 attempts to find way through south section of maze, these trials used pear as reward food

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

Semantics

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 & Attributes

- item: individual entity, discrete
  - eg patient, car, stock, city
  - "independent variable"
- attribute: **property that is measured, observed, logged**...
  - eg height, blood pressure for patient
  - eg horsepower, make for car
  - "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

Dataset types

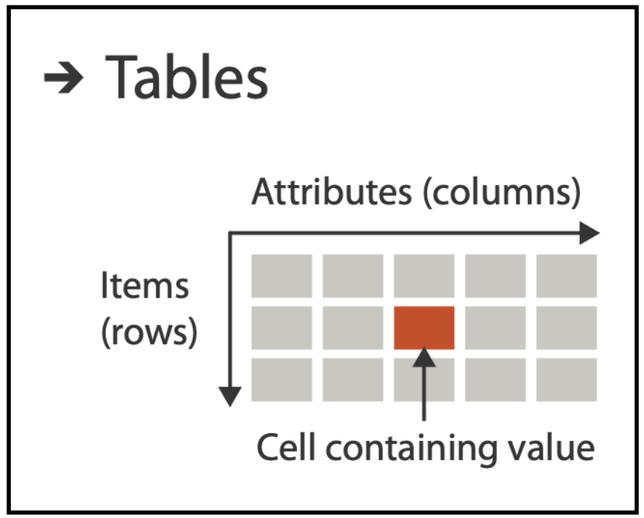
Tables

Items

Attributes

flat table

- one item per row
- each column is 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

Dataset types

Tables

Items

Attributes

flat table

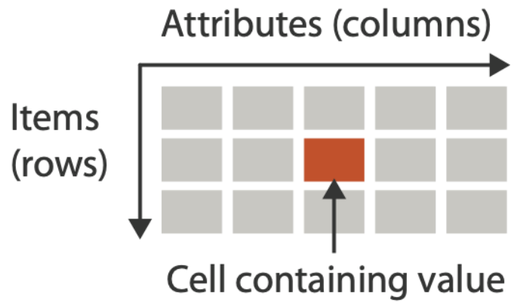
- one item per row
- each column is attribute
- cell holds value for item-attribute pair
- **unique key (could be implicit)**

attributes: name, age, shirt size, fave fruit

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

item: person

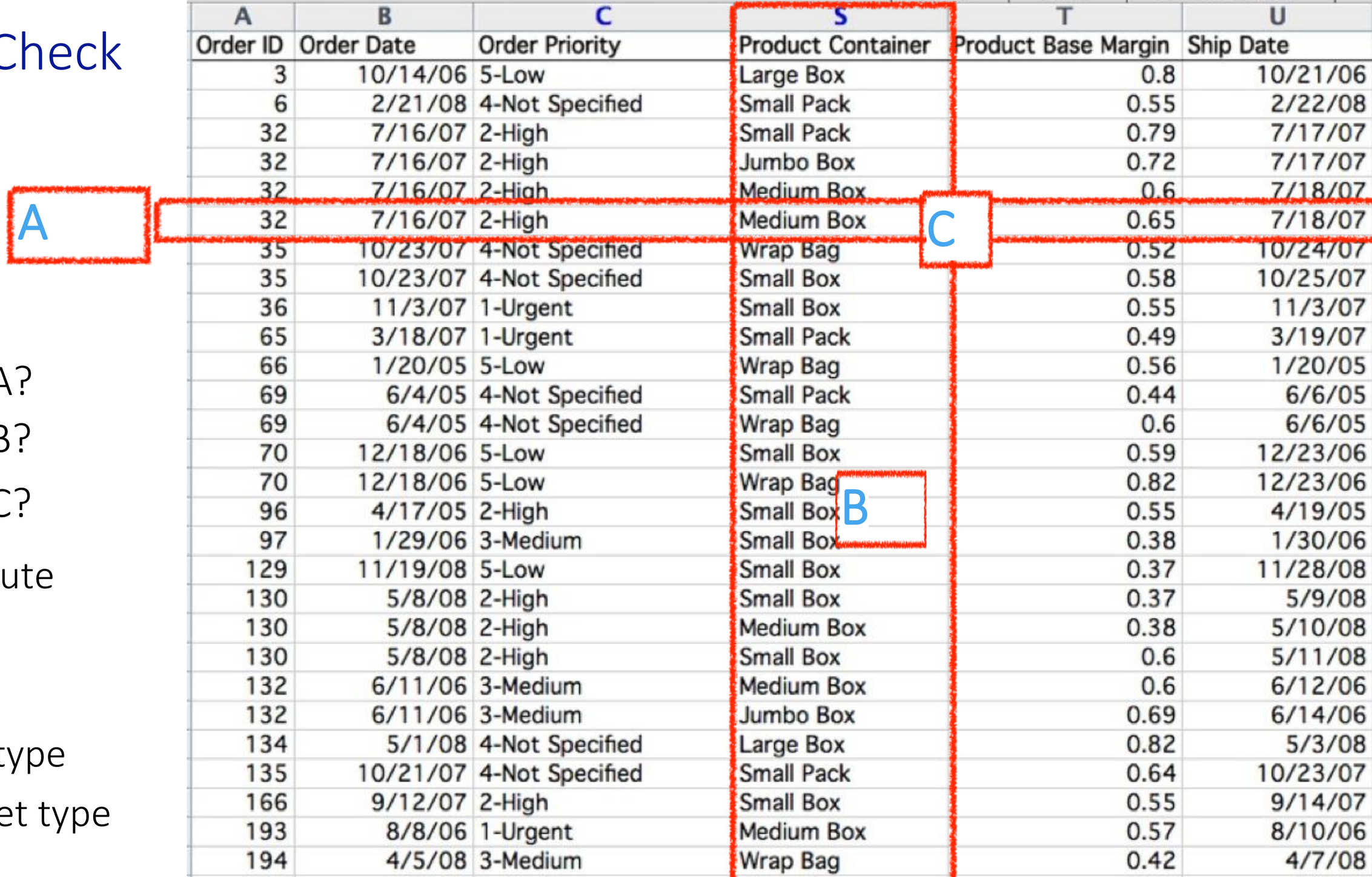
➔ Tables



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

Pulse Check



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 is A?

What is B?

What is C?

A – Attribute

B – Item

C – Cell

D – Data type

E – Dataset type

Dataset types

Tables

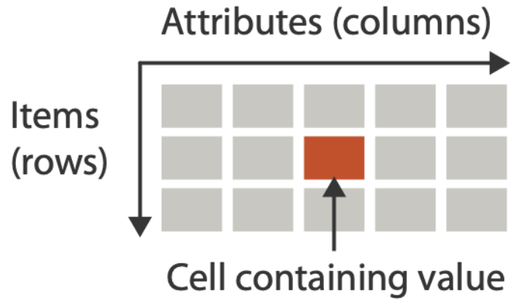
Items

Attributes

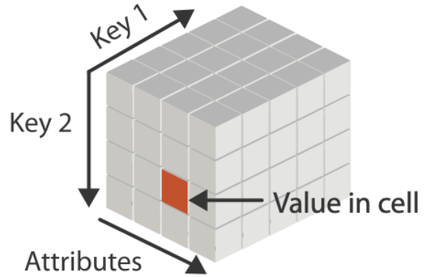
- multidimensional tables
  - –indexing based on multiple keys
    - •eg genes, patients

	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

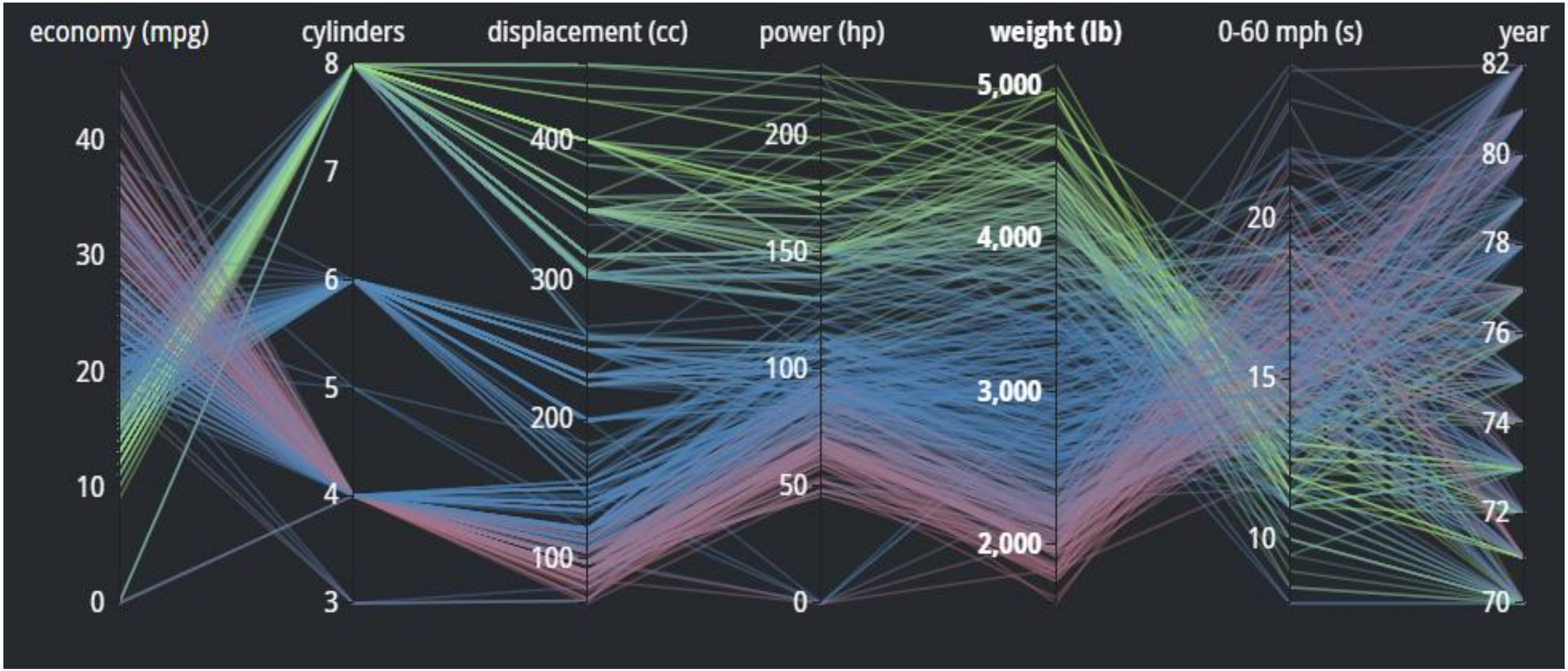
➔ Tables



➔ *Multidimensional Table*



Visualizing Tables: Car Data Parallel Coordinates



http://www.columbia.edu/~cme2126/datavisuals/bigdata_parallelcoordinates.html

Collections

how we group items

sets

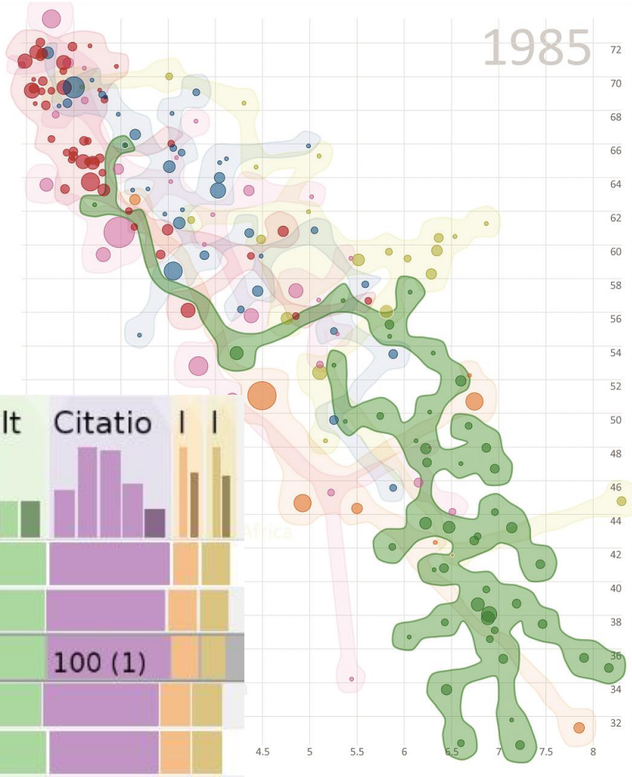
- unique items, unordered

lists

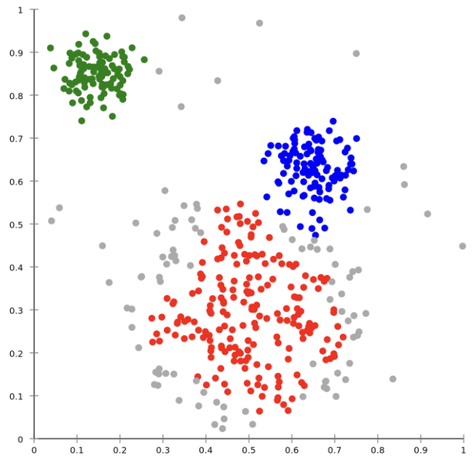
- ordered, duplicates possible

clusters

- groups of similar items



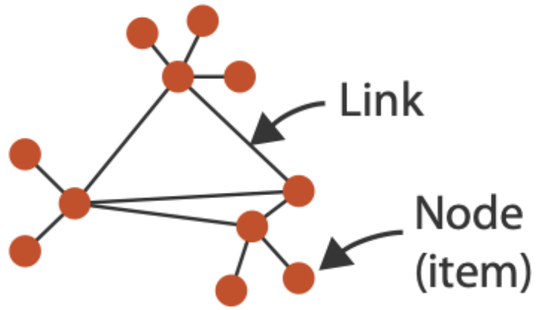
Rank	School Name Filter: <None>	Academic repu	E	Facult	Citatio	I	I
1.	Massachusetts Inst						
2.	University of Camb						
3.	Harvard University	100 (1)			100 (1)		
4.	UCL (University Co						
5.	University of Oxfor						
6.	Imperial College L						
7.	Yale University						
8.	University of Chic						



Other data types

Links

- express relationship between two items
- e.g. friendship on facebook, interaction between proteins

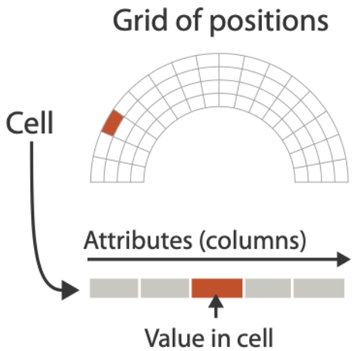


Positions

- spatial data: location in 2D or 3D
- e.g.pixels in photo, voxels in MRI scan, latitude/longitude

Grids

- sampling strategy for continuous data



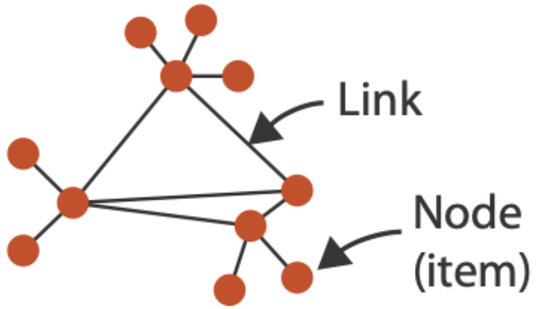
Graphs/Networks

Used to express relationship between two items

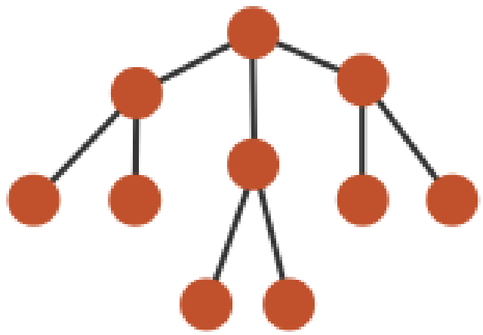
Items (nodes) are connected with links.

Examples

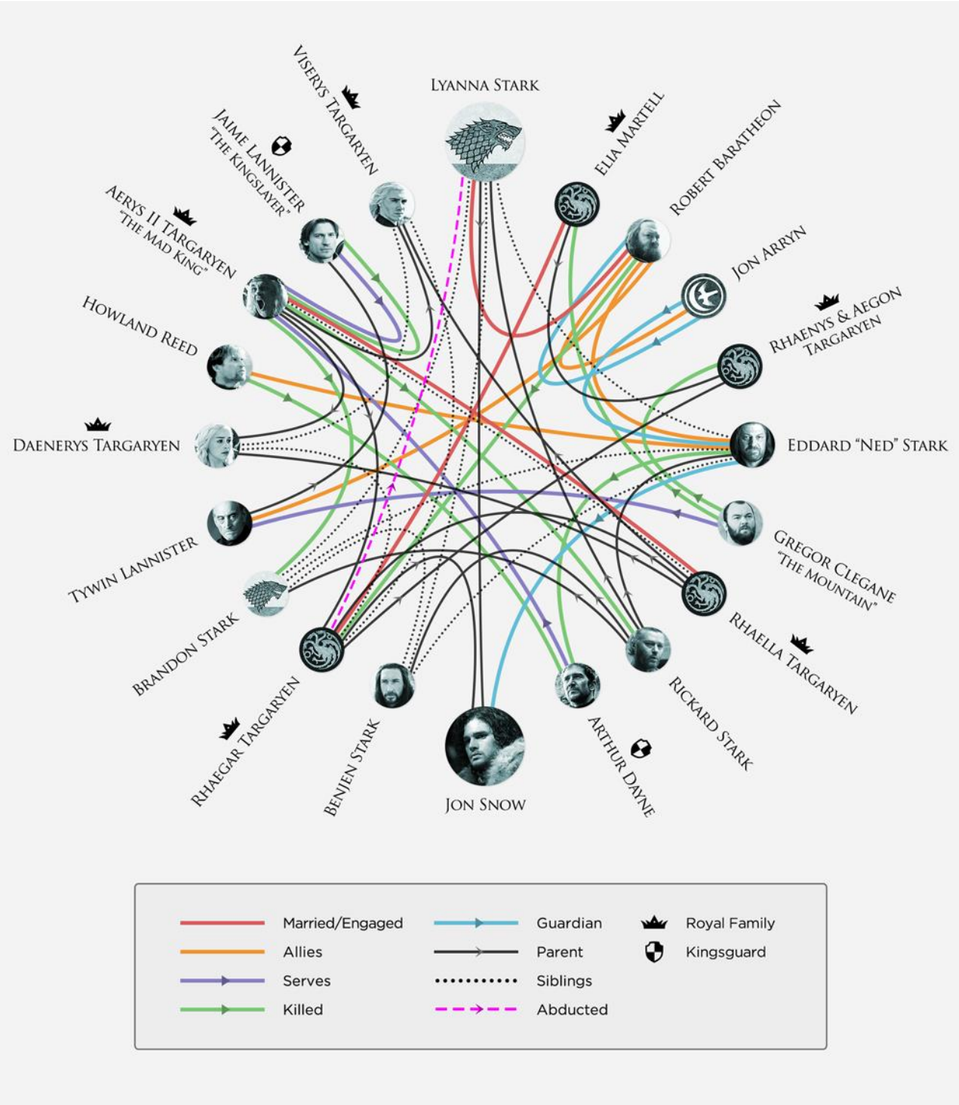
- social networks
- Power grids
- Road networks



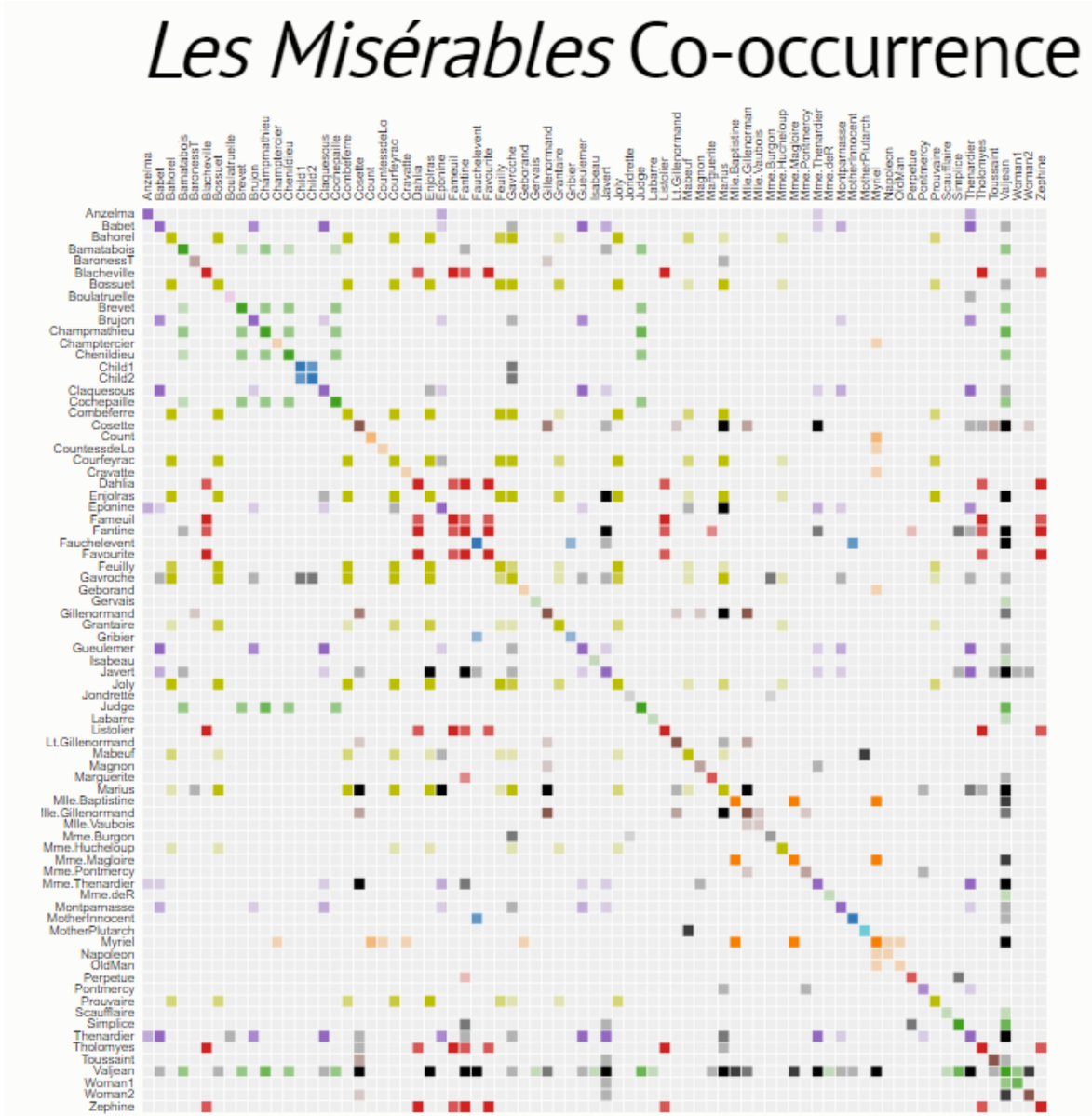
A tree is a subset of graphs
Basically a graph with no cycles.
Trees typically have roots and are directed.



Visualizing Networks/Graphs



https://imgur.com/PG0963W



https://bost.ocks.org/mike/miserables/

Visualizing Trees

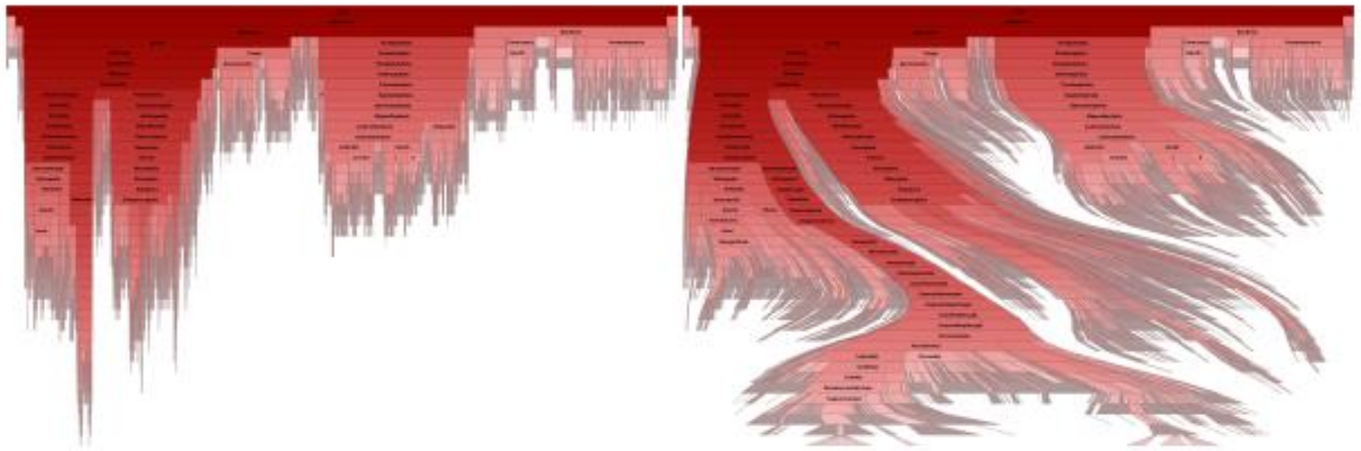
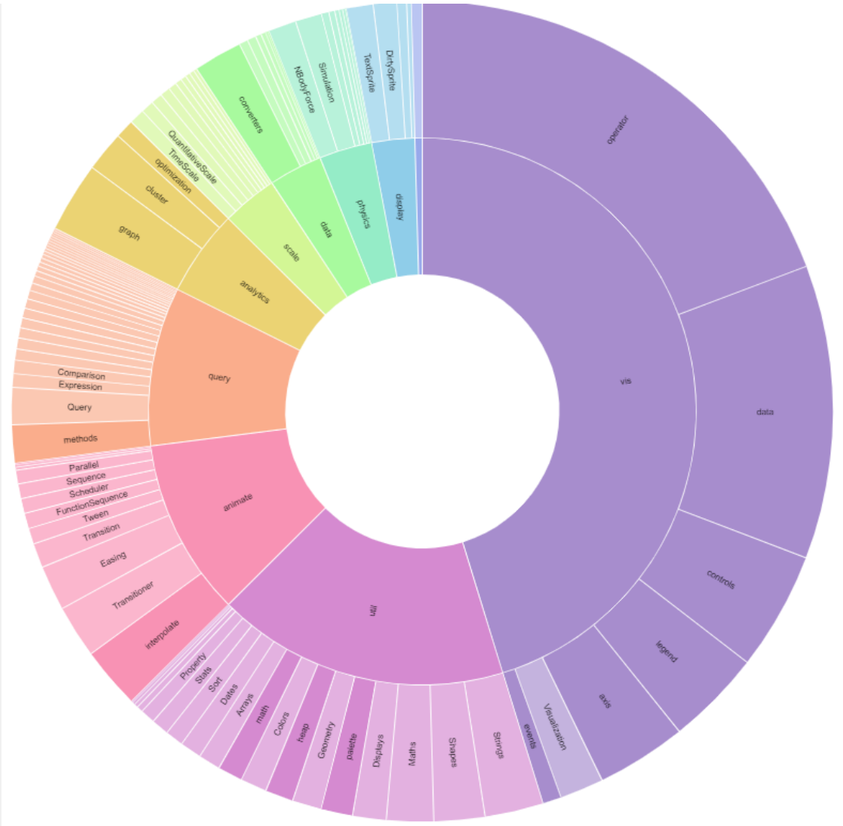


Figure 1: Two visualizations of the NCBI taxonomy dataset [13, 29]: Standard icicle plot (left), space-reclaiming icicle plot (right).



https://ieeexplore.ieee.org/document/9086292

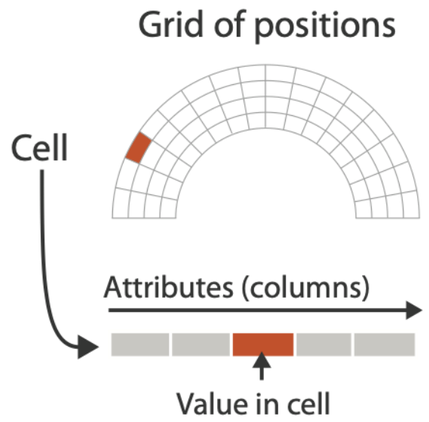
https://observablehq.com/@d3/zoomable-sunburst

Dataset types

Fields	Geometry
Grids	Items
Positions	Positions
Attributes	

➔ Spatial

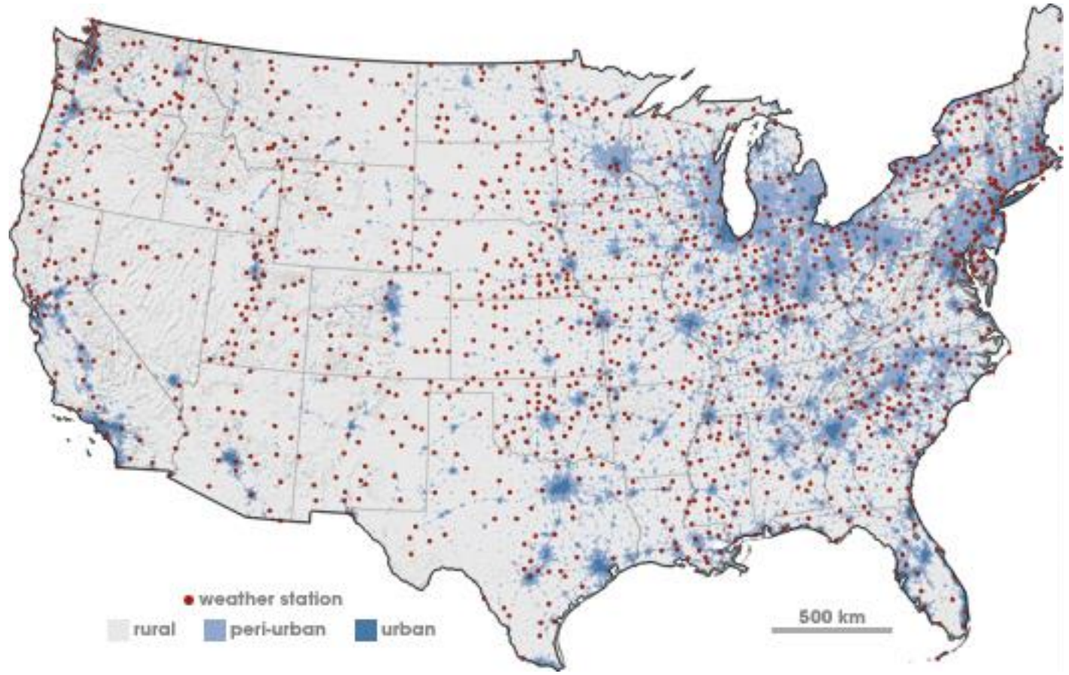
➔ Fields (Continuous)



➔ Geometry (Spatial)

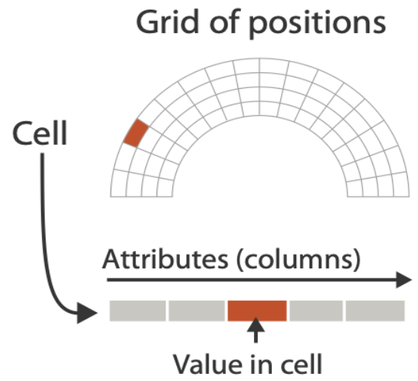
Spatial fields

- attribute values associated w/ cells
- cell contains value from continuous domain
  - eg temperature, pressure, wind velocity
- measured or simulated



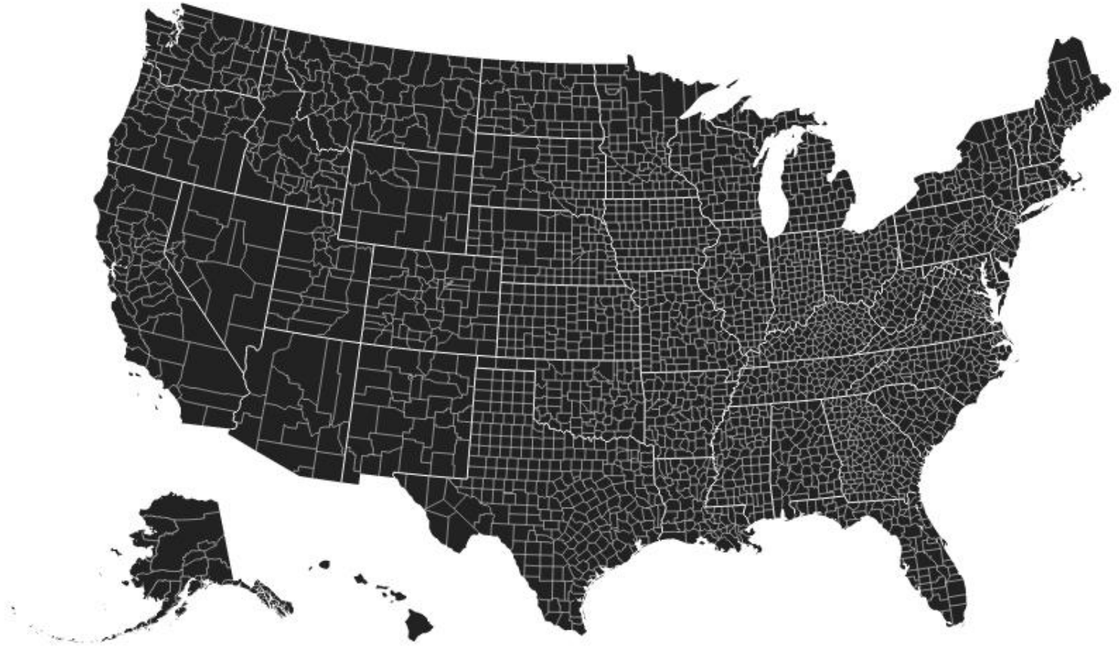
➔ Spatial

➔ Fields (Continuous)



Geometry

- shape of items
- explicit spatial positions / regions
  - points, lines, curves, surfaces, volumes
- boundary between computer graphics and visualization
  - graphics: geometry taken as given
  - vis: geometry is result of a design decision



Attribute Types

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

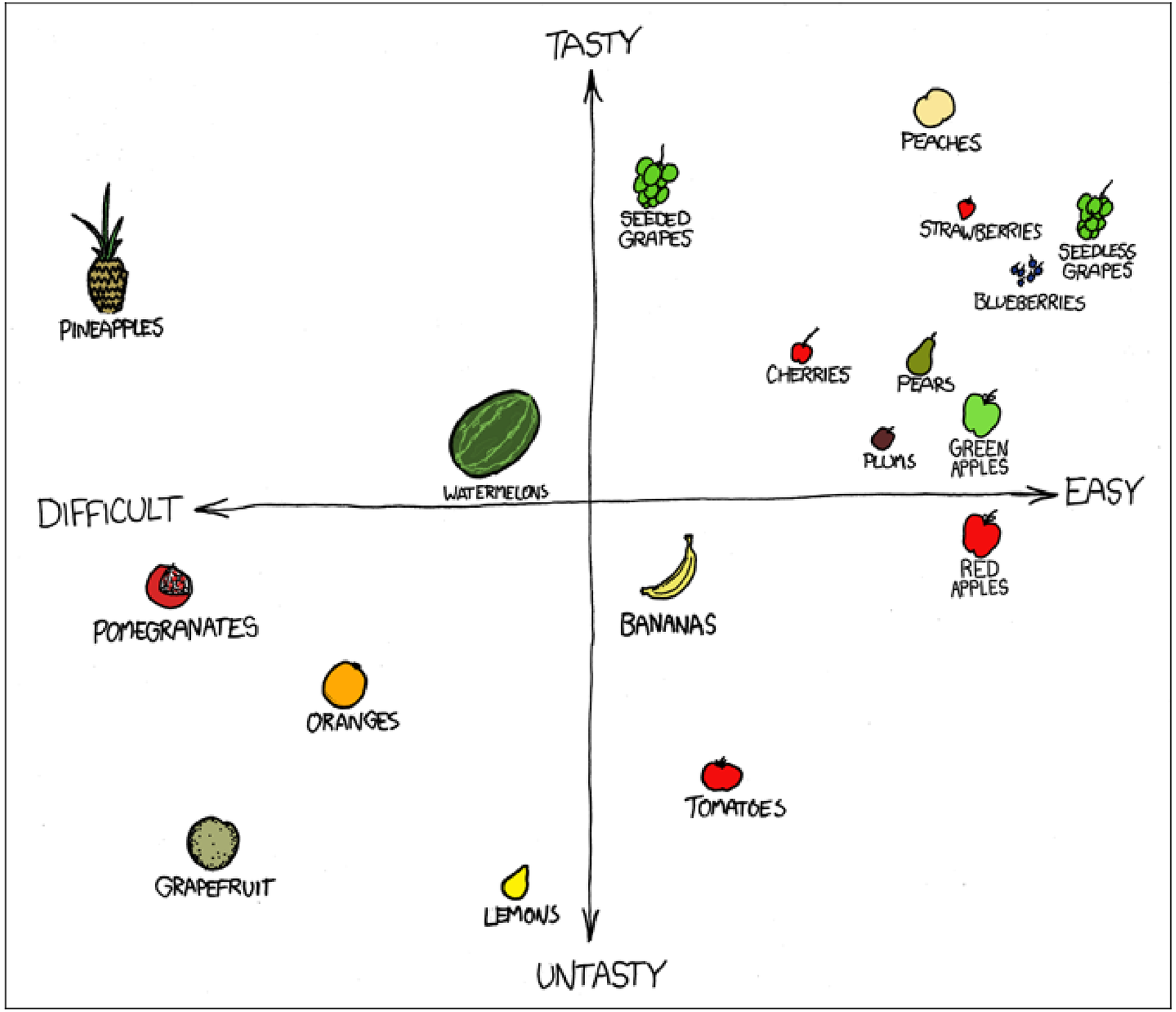
- which classes of values & measurements?
- categorical (nominal)
  - compare equality
  - no implicit ordering
- ordered
  - ordinal
    - less/greater than defined
  - quantitative
    - meaningful magnitude
    - arithmetic possible

➔ Categorical

➔ Ordered

➔ *Ordinal*

➔ *Quantitative*



https://xkcd.com/388/

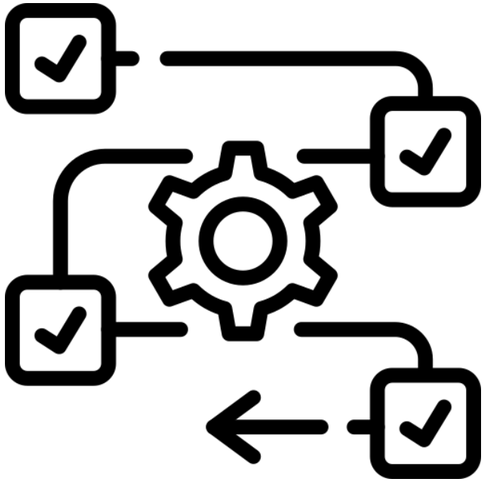
1. What is an item
2. What is an attribute
3. What is a feature
4. What is the semantics
5. What is the keys
6. What is type for Order ID
7. What is the type for Order Date
8. What is the type for Order Priority
9. What is the type for Product Container
10. What is the type for Product Base margin
11. What is the type for Ship Date

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

City	Condition	Temperature
Calgary	Mainly Clear	-4°C
Charlottetown	Light Snowshower	-6°C
Edmonton		-7°C
Fredericton	Clear	-9°C
Halifax	Light Snow	-6°C
Iqaluit	Clear	-28°C
Montréal	Mainly Clear	-9°C
Ottawa (Kanata - Orléans)	Mainly Clear	-10°C
Prince George	Fog	-6°C
Québec	Mainly Clear	-15°C
Regina	Mist	-13°C
Saskatoon	Mist	-11°C
St. John's	Mostly Cloudy	-5°C
Thunder Bay	Light Snow	0°C
Toronto	Cloudy	-1°C
Vancouver	Mainly Clear	7°C
Victoria	Mainly Clear	5°C
Whitehorse	Mostly Cloudy	-12°C
Winnipeg	Mist	-6°C
Yellowknife	Light Snow	-13°C

https://weather.gc.ca/canada_e.html

Data abstraction: Three operations

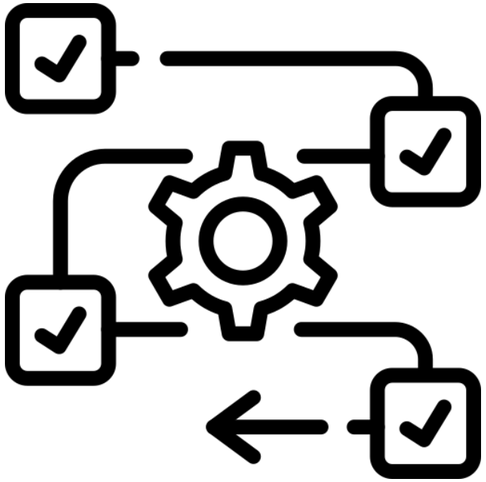


- translate from domain-specific language to generic visualization language

- identify dataset type(s), attribute types
- identify cardinality
  - how many items in the dataset?
  - what is cardinality of each attribute?
    - number of levels for categorical data
    - range for quantitative data

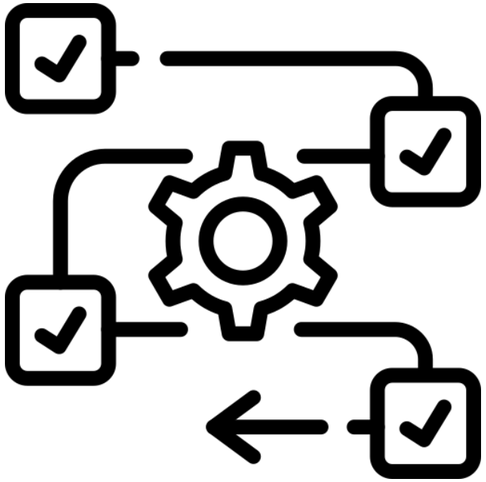
- consider whether to transform data
  - guided by understanding of task

Data vs conceptual models

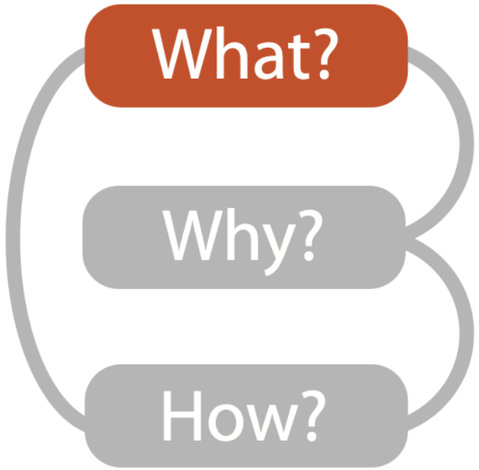


- data model
  - mathematical abstraction
    - sets with operations, eg floats with * / - +
    - variable data types in programming languages
- conceptual model
  - mental construction (semantics)
  - supports reasoning
  - typically based on understanding of tasks [stay tuned!]
- data abstraction process relies on 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 figures: 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



What?

Datasets

Data Types

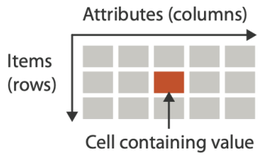
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

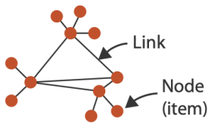
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

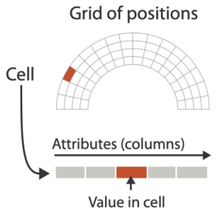
→ Tables



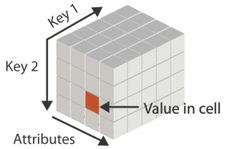
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

→ Geometry (Spatial)

Dataset Availability

→ Static

→ Dynamic

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

→ *Quantitative*

Ordering Direction

→ Sequential

→ Diverging

→ Cyclic

Design Sprint

What kind of questions can we answer?

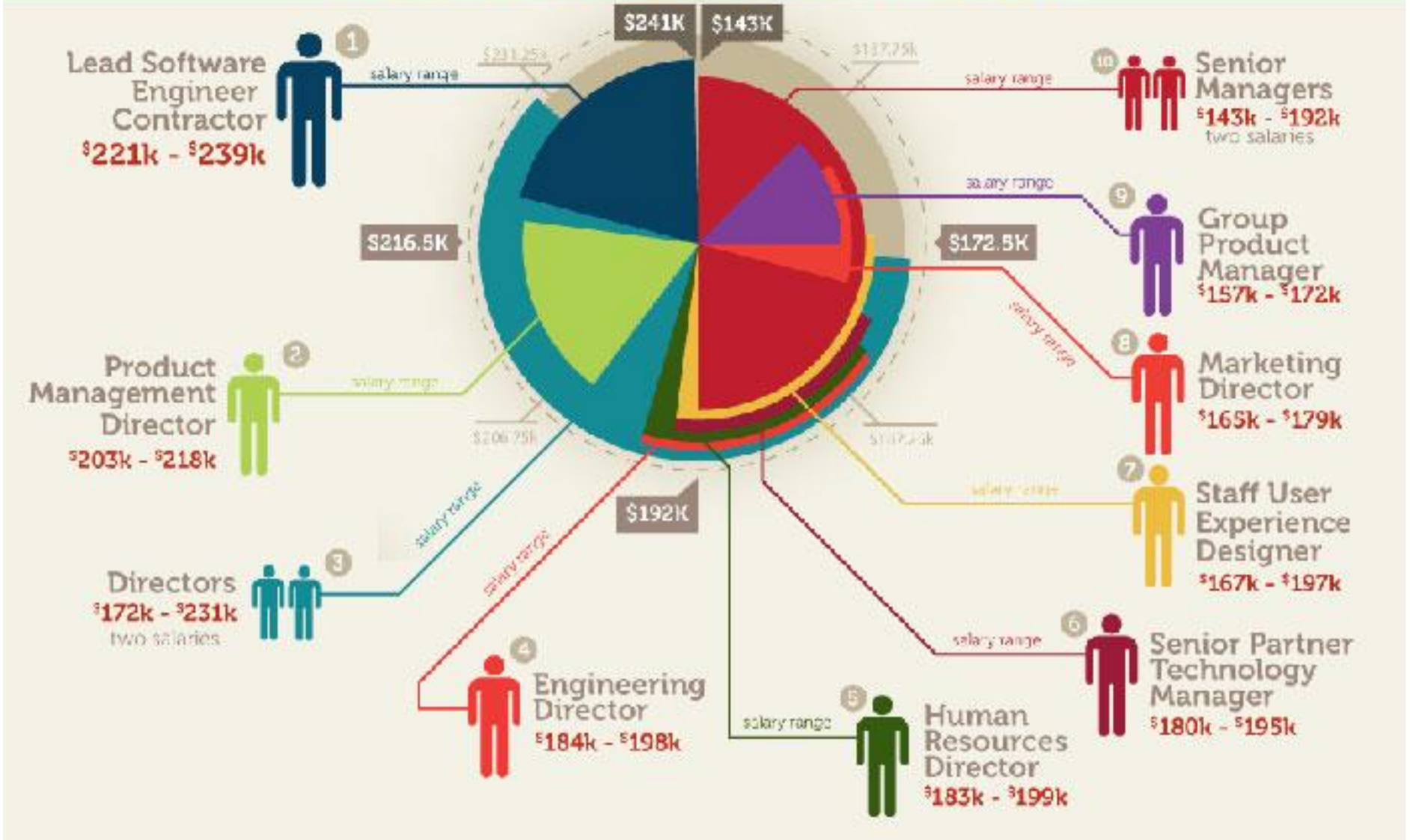
What is the data?

Who might be the audience?

City	Condition	Temperature
Calgary	Mainly Clear	-4°C
Charlottetown	Light Snowshower	-6°C
Edmonton		-7°C
Fredericton	Clear	-9°C
Halifax	Light Snow	-6°C
Iqaluit	Clear	-28°C
Montréal	Mainly Clear	-9°C
Ottawa (Kanata - Orléans)	Mainly Clear	-10°C
Prince George	Fog	-6°C
Québec	Mainly Clear	-15°C
Regina	Mist	-13°C
Saskatoon	Mist	-11°C
St. John's	Mostly Cloudy	-5°C
Thunder Bay	Light Snow	0°C
Toronto	Cloudy	-1°C
Vancouver	Mainly Clear	7°C
Victoria	Mainly Clear	5°C
Whitehorse	Mostly Cloudy	-12°C
Winnipeg	Mist	-6°C
Yellowknife	Light Snow	-13°C

top 10 salaries at Google

RANGE FROM $143,000 TO $241,000 PER YEAR.



Learning Goals

- Describe the difference between how the phrases “dataset types”, “data types” are used in vis. Literature as opposed to programming
- Describe the characteristics of data
- Differentiate between the different types of data and dataset types