Data Models

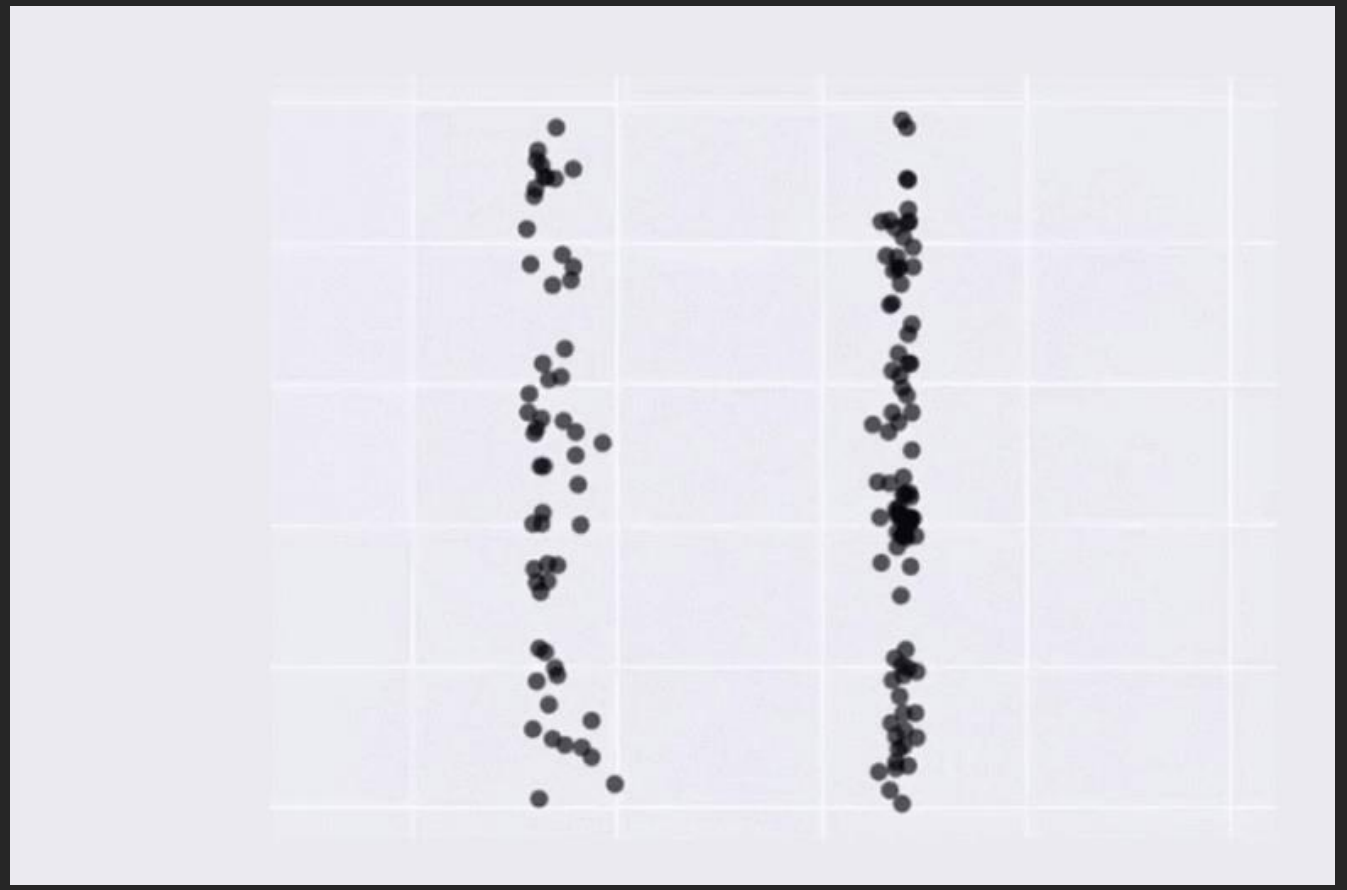
Information Visualization – SI649

Eytan Adar

September 13, 2021

http://www.slido.com

event code **#C674**



Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

https://www.autodesk.com/research/publications/same-stats-different-graphs

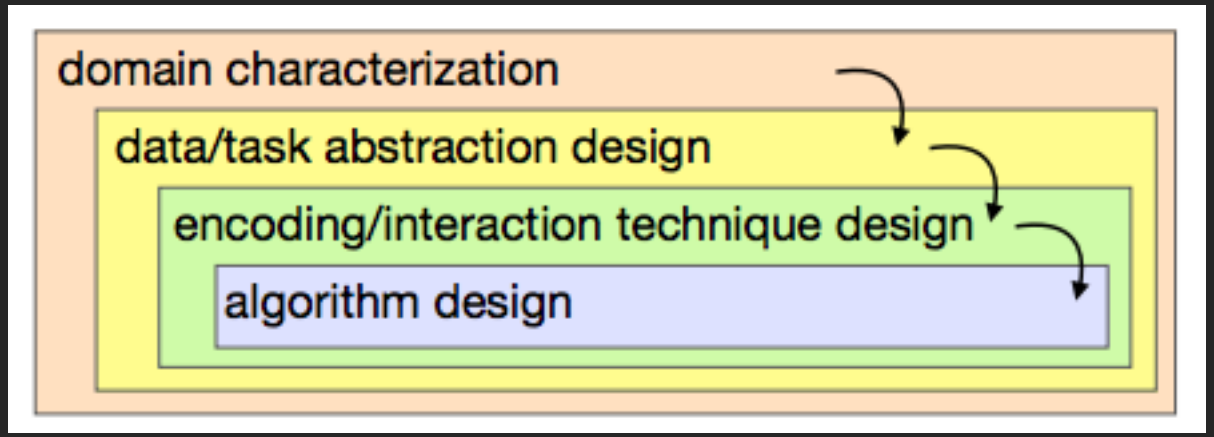
On the menu for today

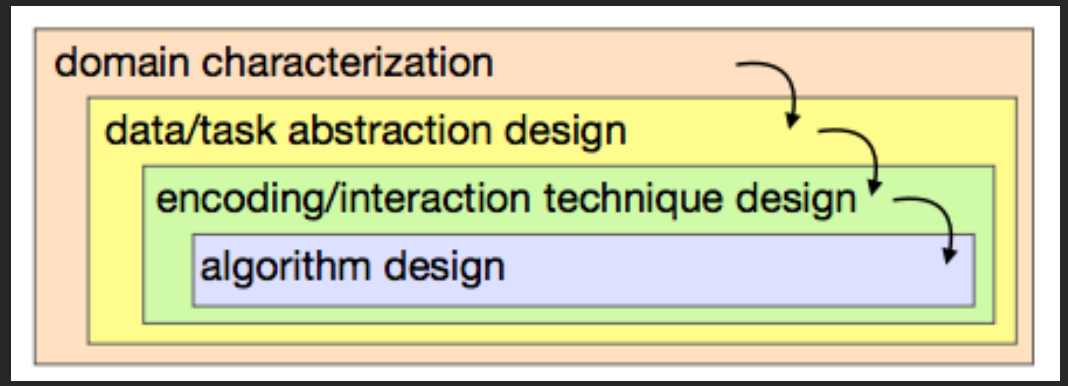
- More on…
  - Expansion on data types + encoding
  - Expressiveness and Effectiveness
- Labs this week:
  - Grammar of Graphics & Altair

Quick review

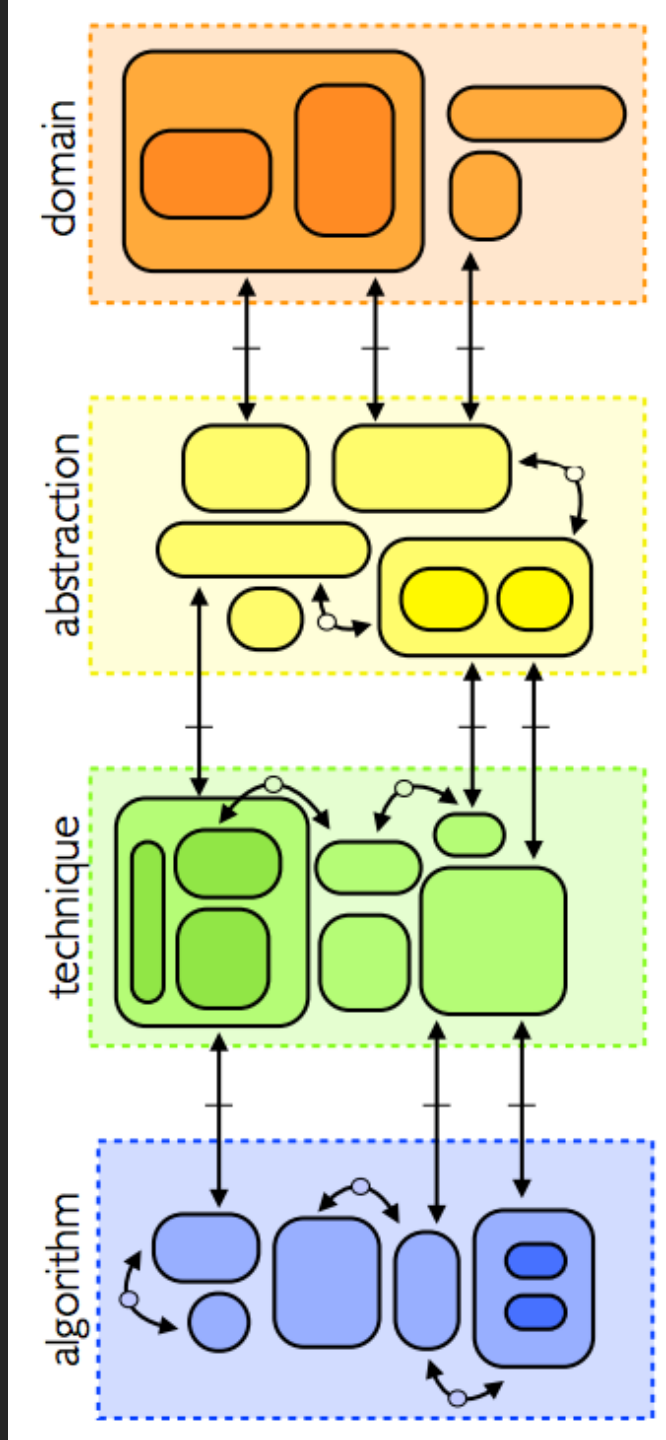
The skill we will develop

1. Identify the job, and job functions
2. Identify the visual comparisons that enable function
3. Identify the best visual encoding/idioms that support those comparisons
4. Identify the best way to implement

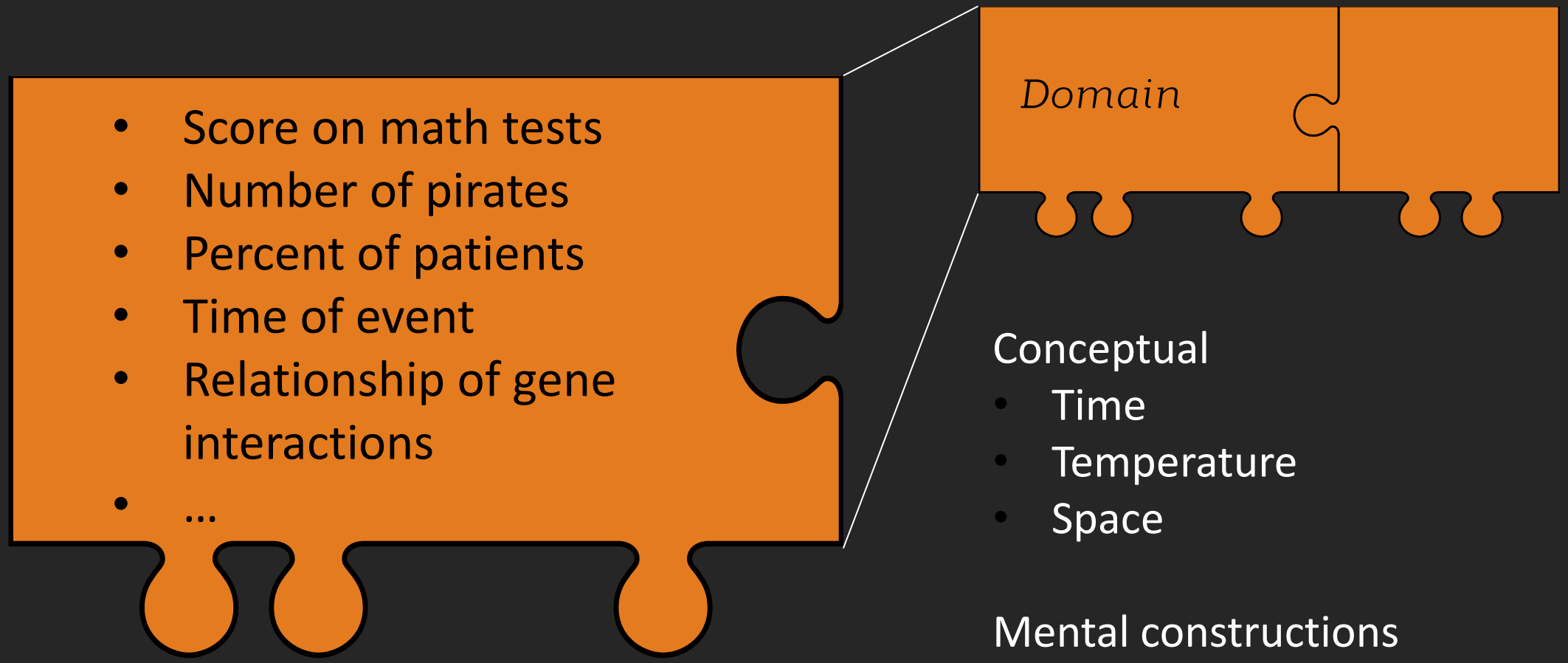




Munzner, 2009



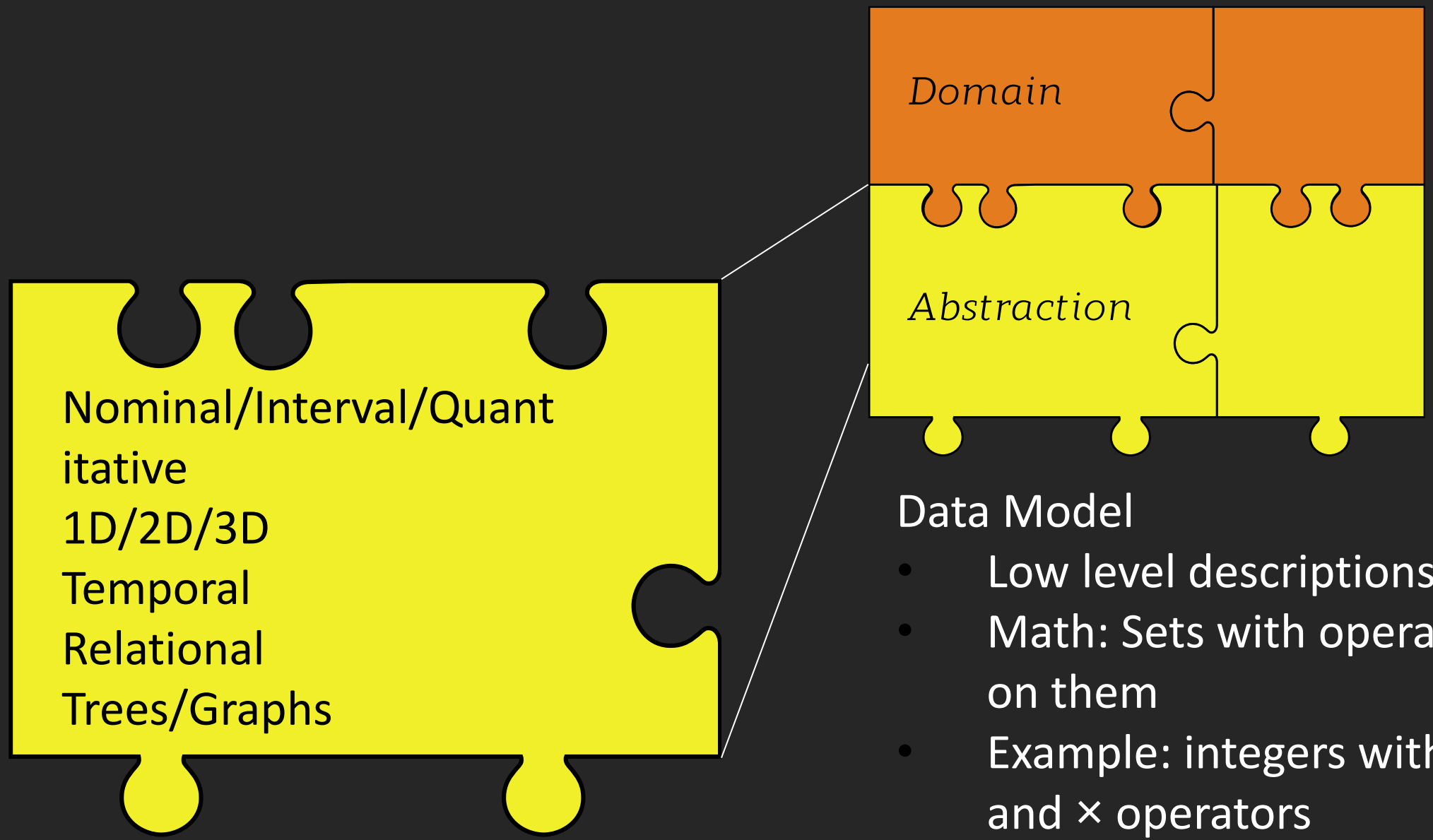
Meyer et al., 2013



Conceptual
- Time
- Temperature
- Space

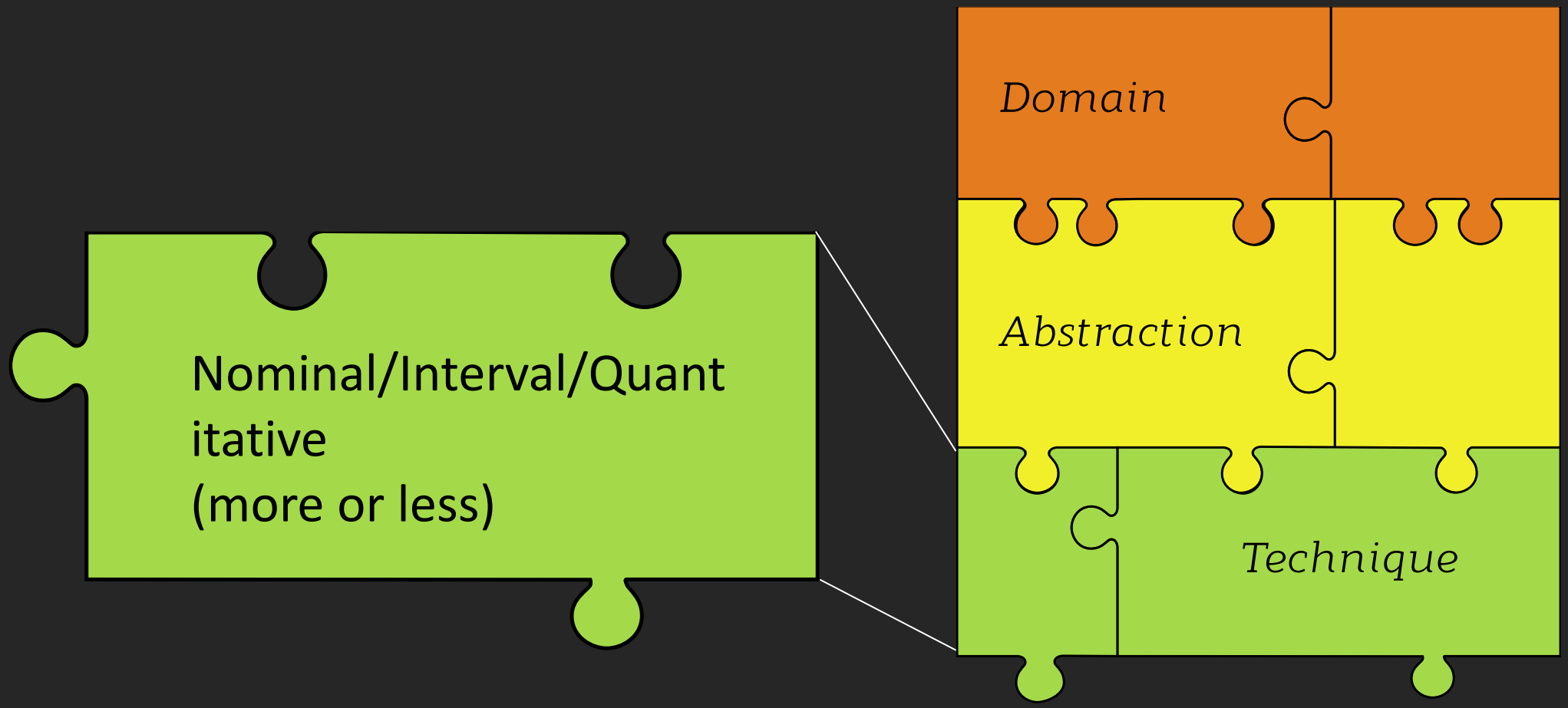
Mental constructions

Include semantics and support reasoning

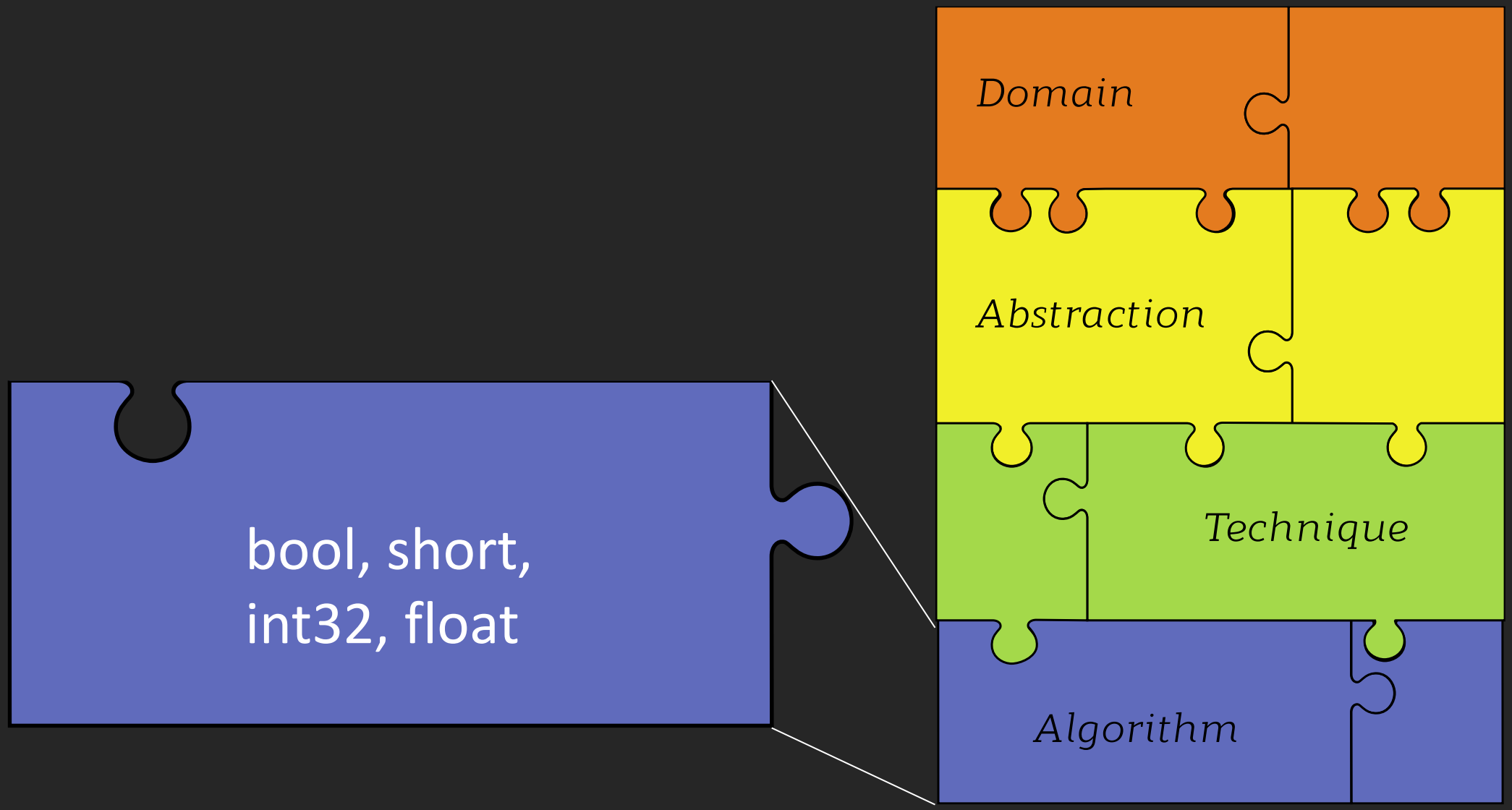


Data Model

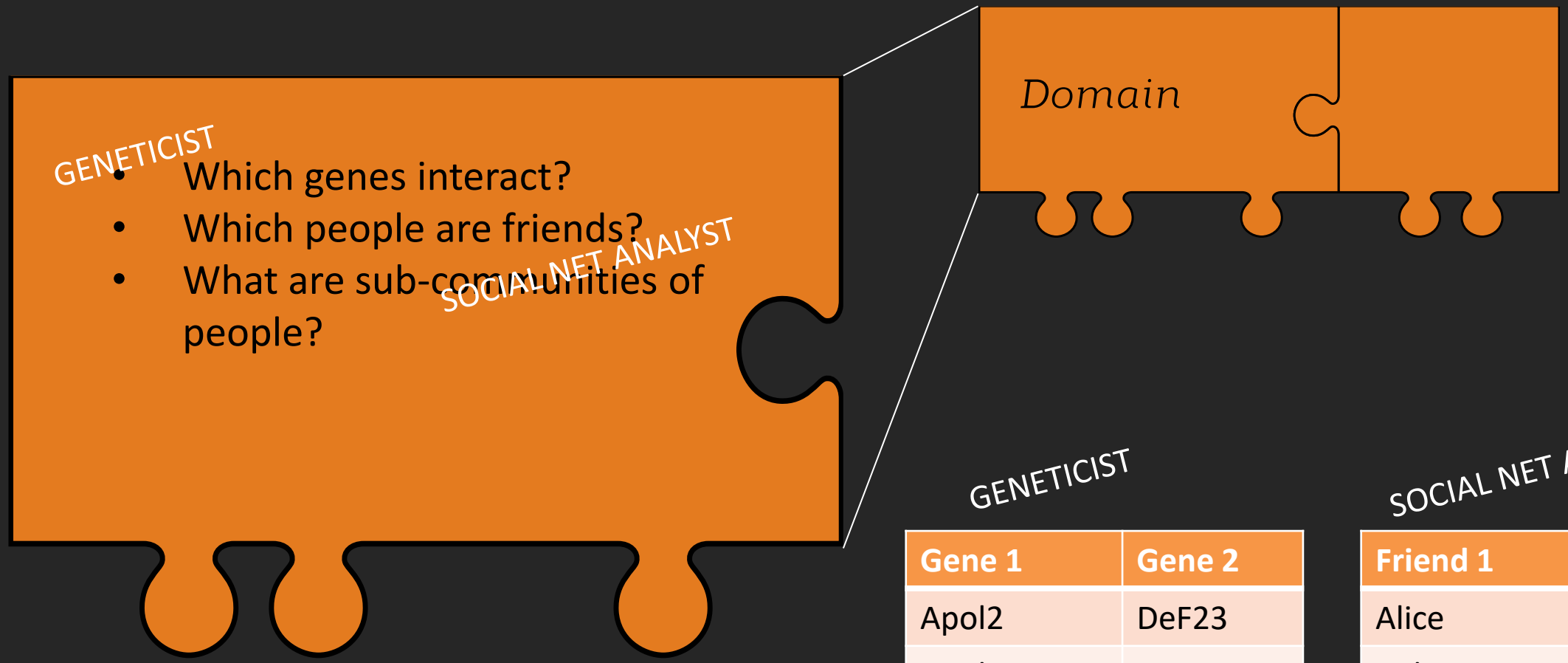
- Low level descriptions
- Math: Sets with operations on them
- Example: integers with + and × operators



Technique = “Idiom” – Bar chart, pie chart, etc.


Domain
Abstraction
Technique
Algorithm
bool, short,
int32, float

Concrete example…



GENETICIST

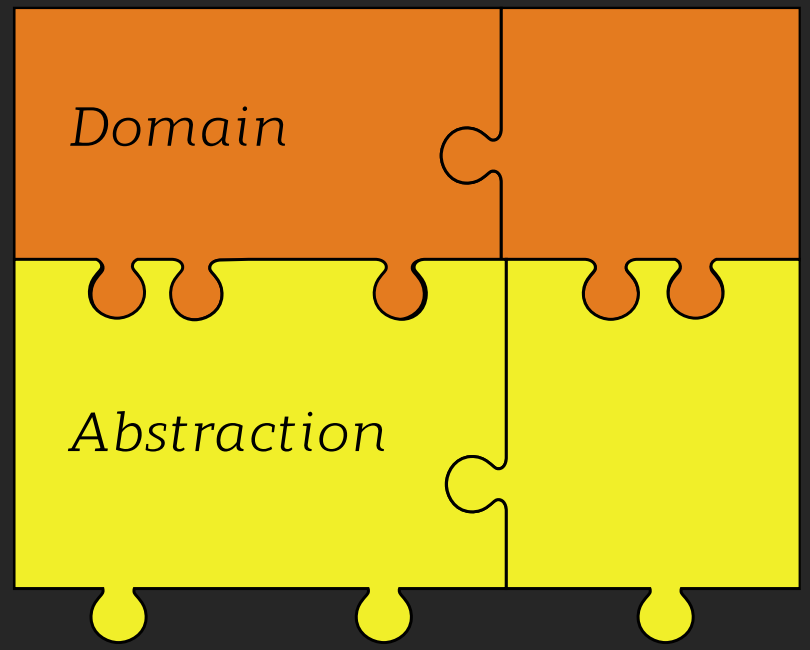
Gene 1	Gene 2
Apol2	DeF23
Apol2	TMK9
TMk9	MnU8
TMk9	MnU10
…	…

SOCIAL NET ANALYST

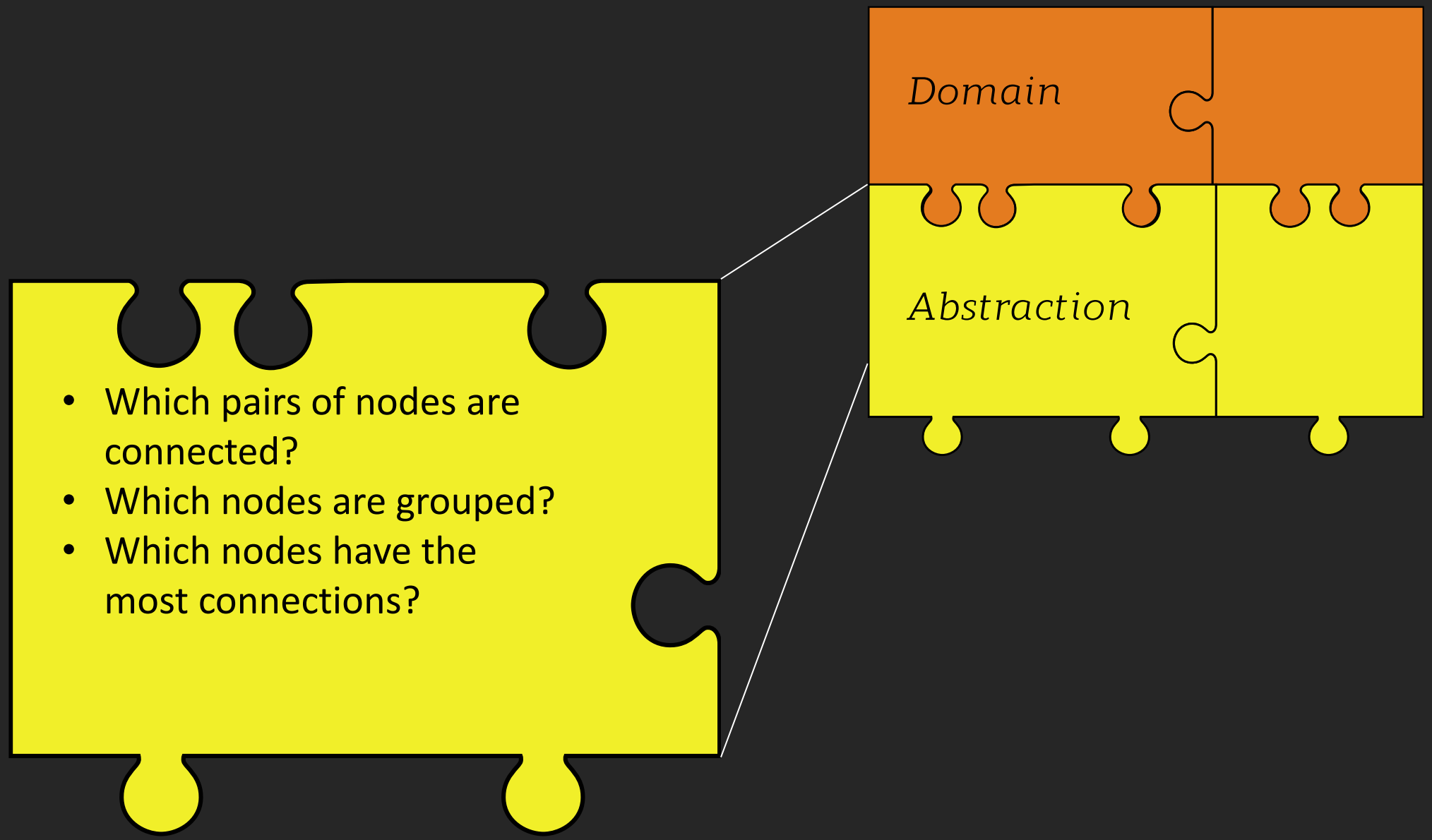
Friend 1	Friend 2
Alice	Bob
Bob	Janice
Bob	Steph
Alice	Rahul
…	…

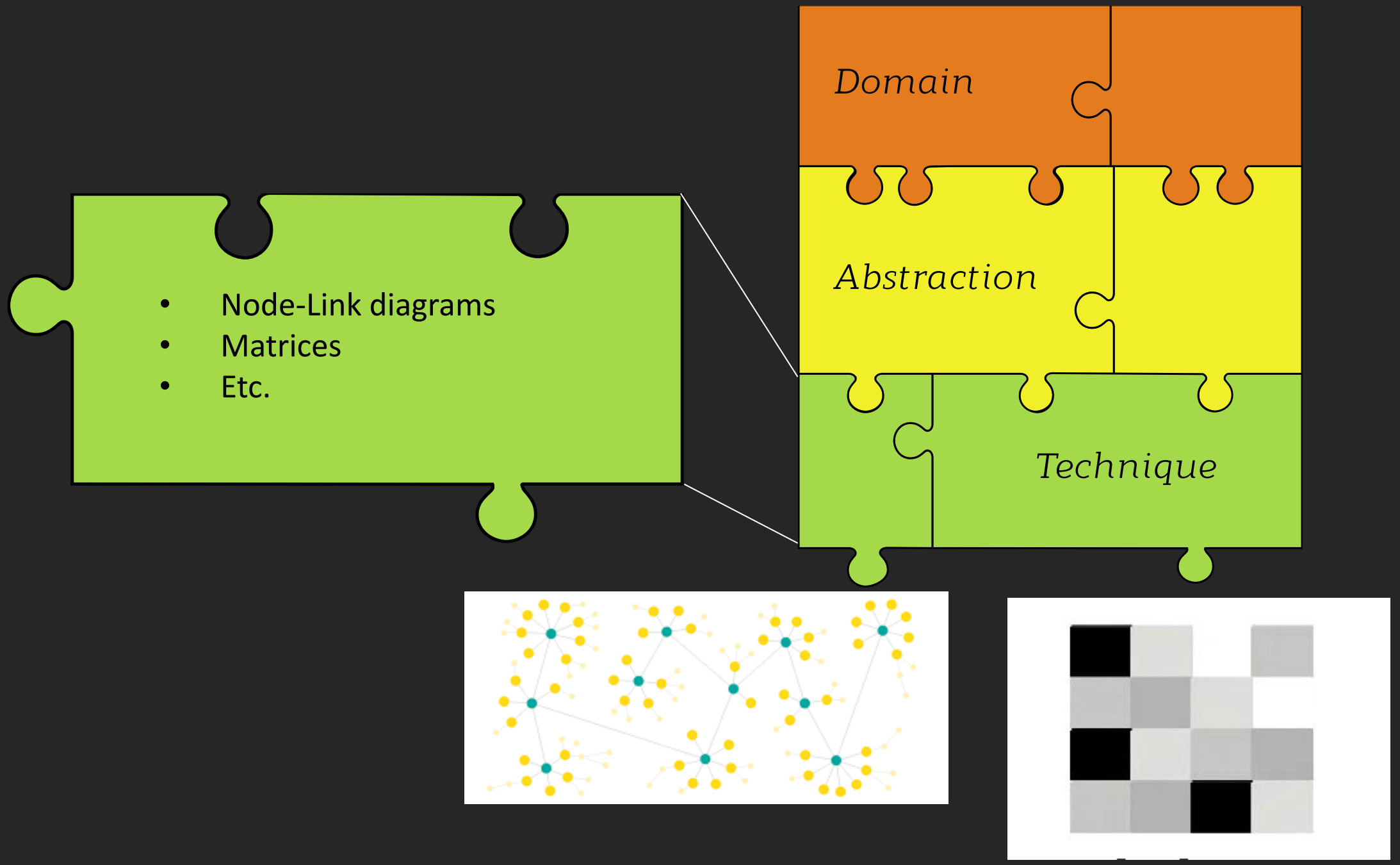
Which g nodes are connected?

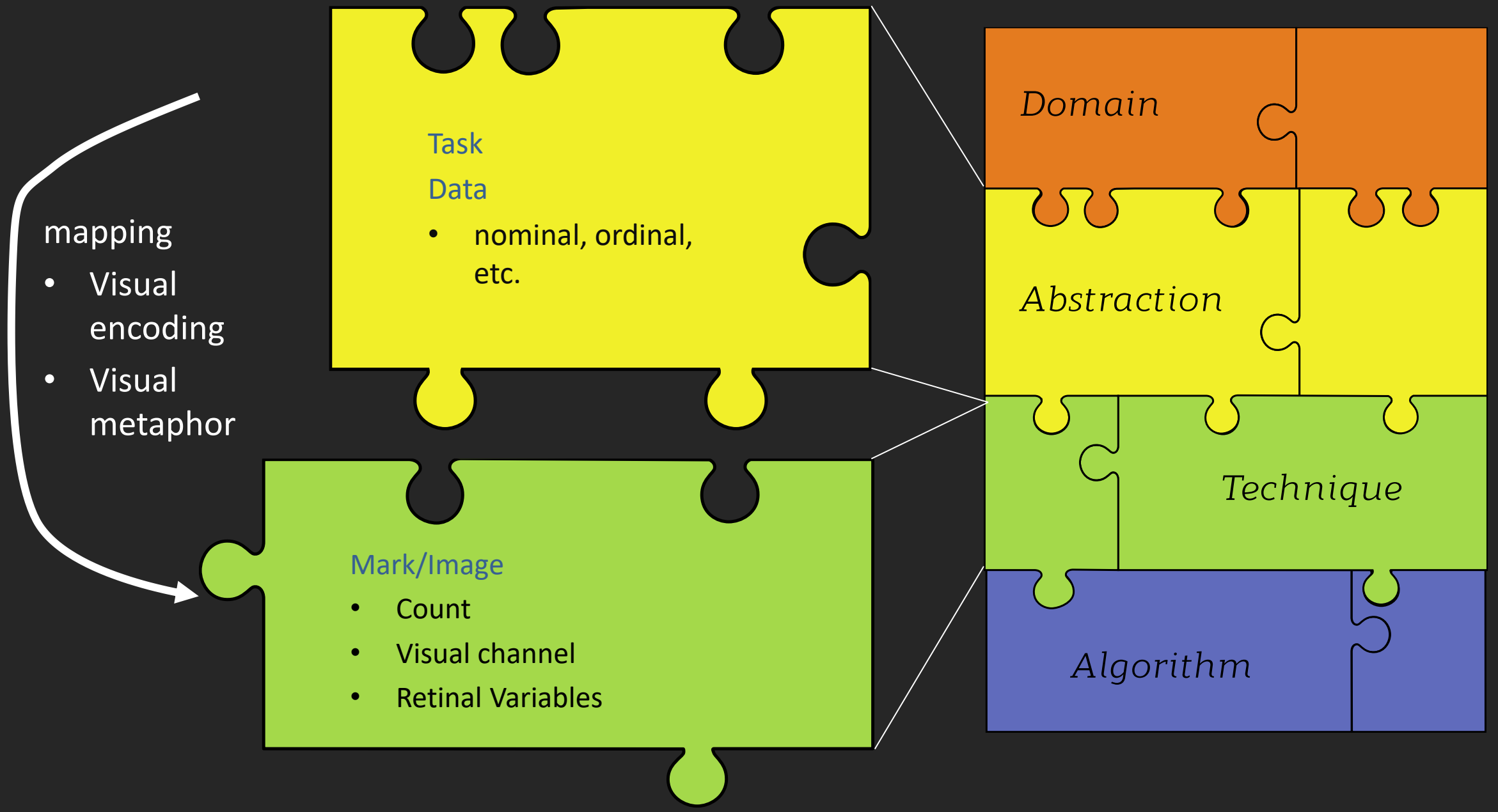
Which pe nodes are connected?



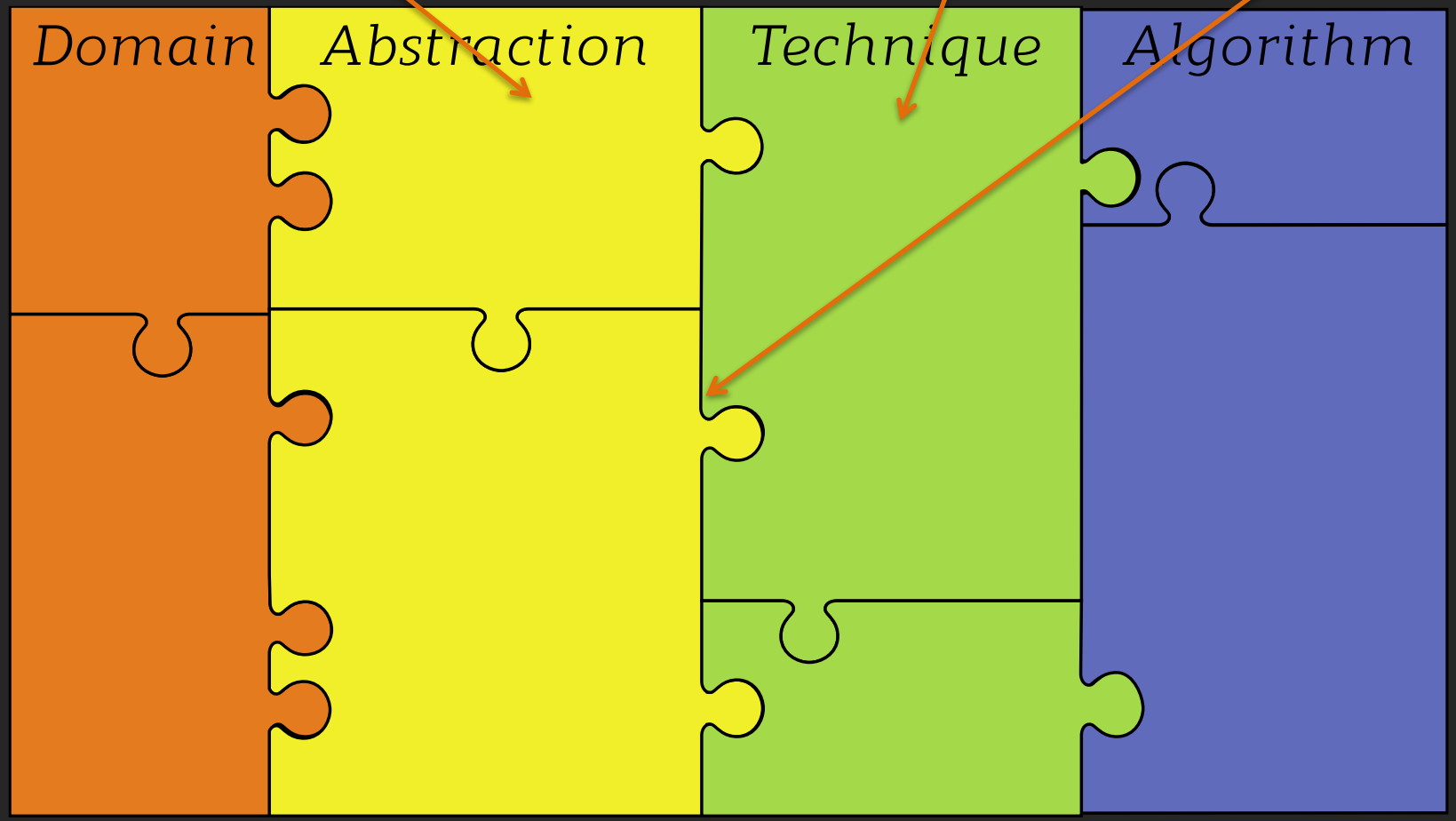
Node 1	Node 2
V1	V2
V2	V3
V2	V4
V9	V12
…	…


Domain
Abstraction
• Which pairs of nodes are connected?
• Which nodes are grouped?
• Which nodes have the most connections?


Domain
Abstraction
Technique
• Node-Link diagrams
• Matrices
• Etc.


mapping
• Visual encoding
• Visual metaphor
Task
Data
• nominal, ordinal, etc.
Mark/Image
• Count
• Visual channel
• Retinal Variables
Domain
Abstraction
Technique
Algorithm

Given the data and tasks, are my choices in representation good?



From the video…

- Data models
  - Relational
  - Statistical
  - Cube
- Taxonomies of Data
- Attributes of Data

http://www.slido.com
event code **#C674**

What is the relationship between the relational, statistical, and cube data “models?”

Relational

Statistical Data

Data Cubes

Relational

Often hold

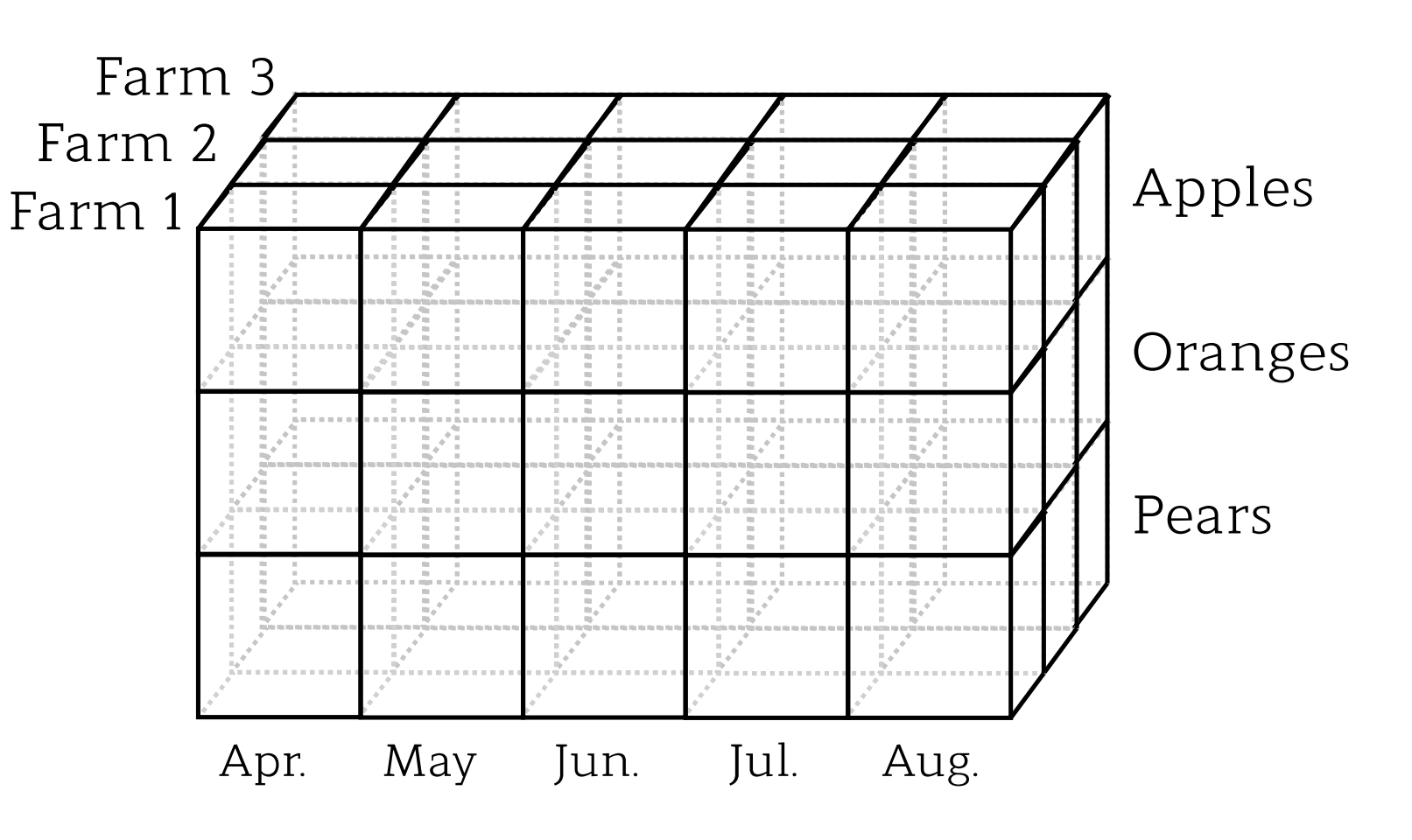
Statistical Data

Trt.	Cor.	$\|\gamma\|_2^2$	ODB	RCT	Wtd.	Spike	Dyn.	Oracle
c	y	3	0.0077	0.0742	0.0072	0.0053	0.0067	0.0059
c	y	6	0.0230	0.0808	0.0212	0.0103	0.0138	0.0134
c	n	3	0.0122	0.1443	0.0116	0.0094	0.0120	0.0104
c	n	6	0.0209	0.1538	0.0196	0.0138	0.0172	0.0158
l	y	3	0.0076	0.0750	0.0072	0.0053	0.0067	0.0060
l	y	6	0.0225	0.0784	0.0209	0.0112	0.0139	0.0134
l	n	3	0.0123	0.1377	0.0116	0.0098	0.0117	0.0104
l	n	6	0.0220	0.1524	0.0204	0.0137	0.0175	0.0162
q	y	3	0.0073	0.0799	0.0069	0.0052	0.0066	0.0058
q	y	6	0.0217	0.0751	0.0201	0.0101	0.0138	0.0132
q	n	3	0.0127	0.1496	0.0120	0.0093	0.0119	0.0107
q	n	6	0.0214	0.1503	0.0201	0.0127	0.0176	0.0160

Two main column “types”: independent & dependent

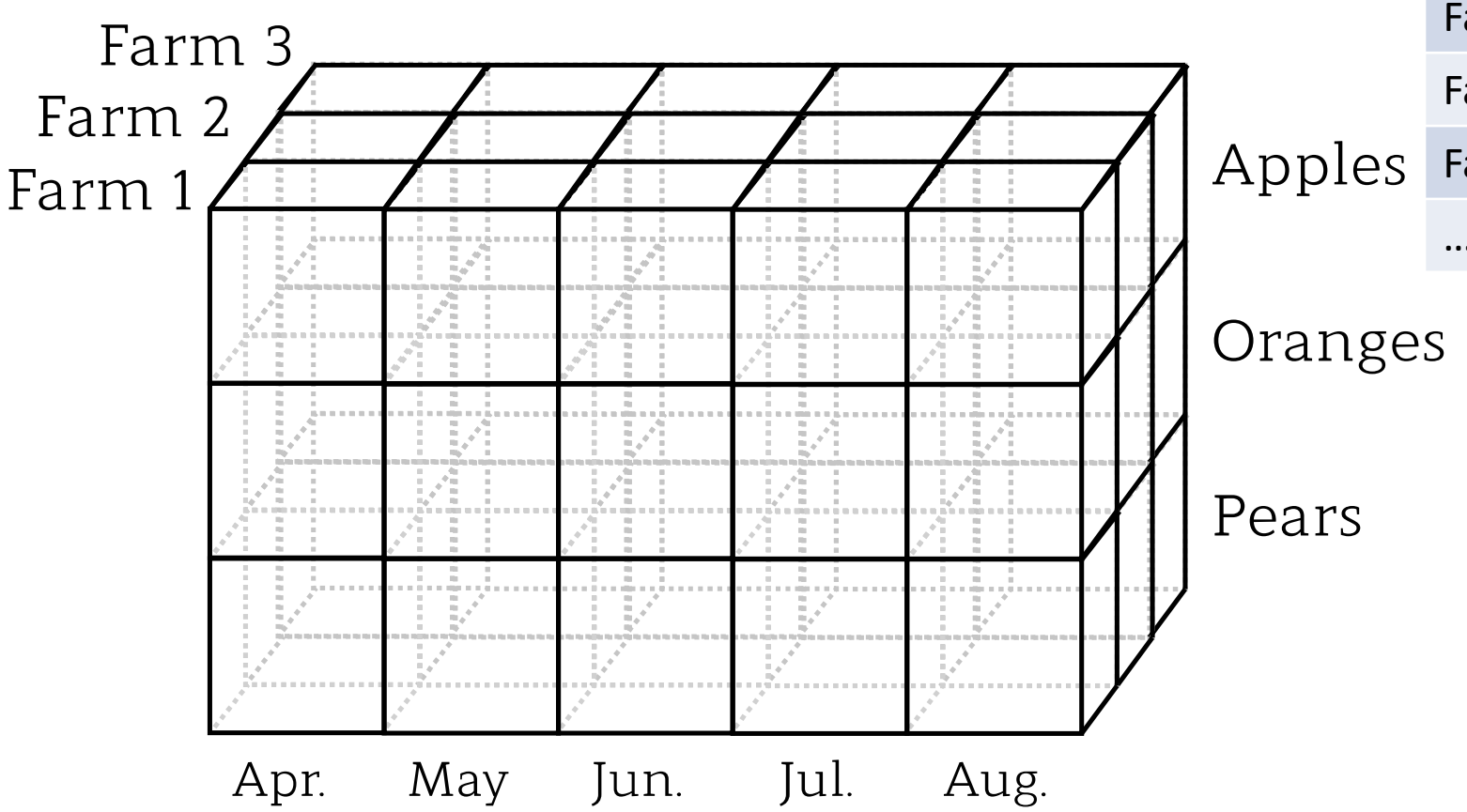
Relational

Conceptualized as…



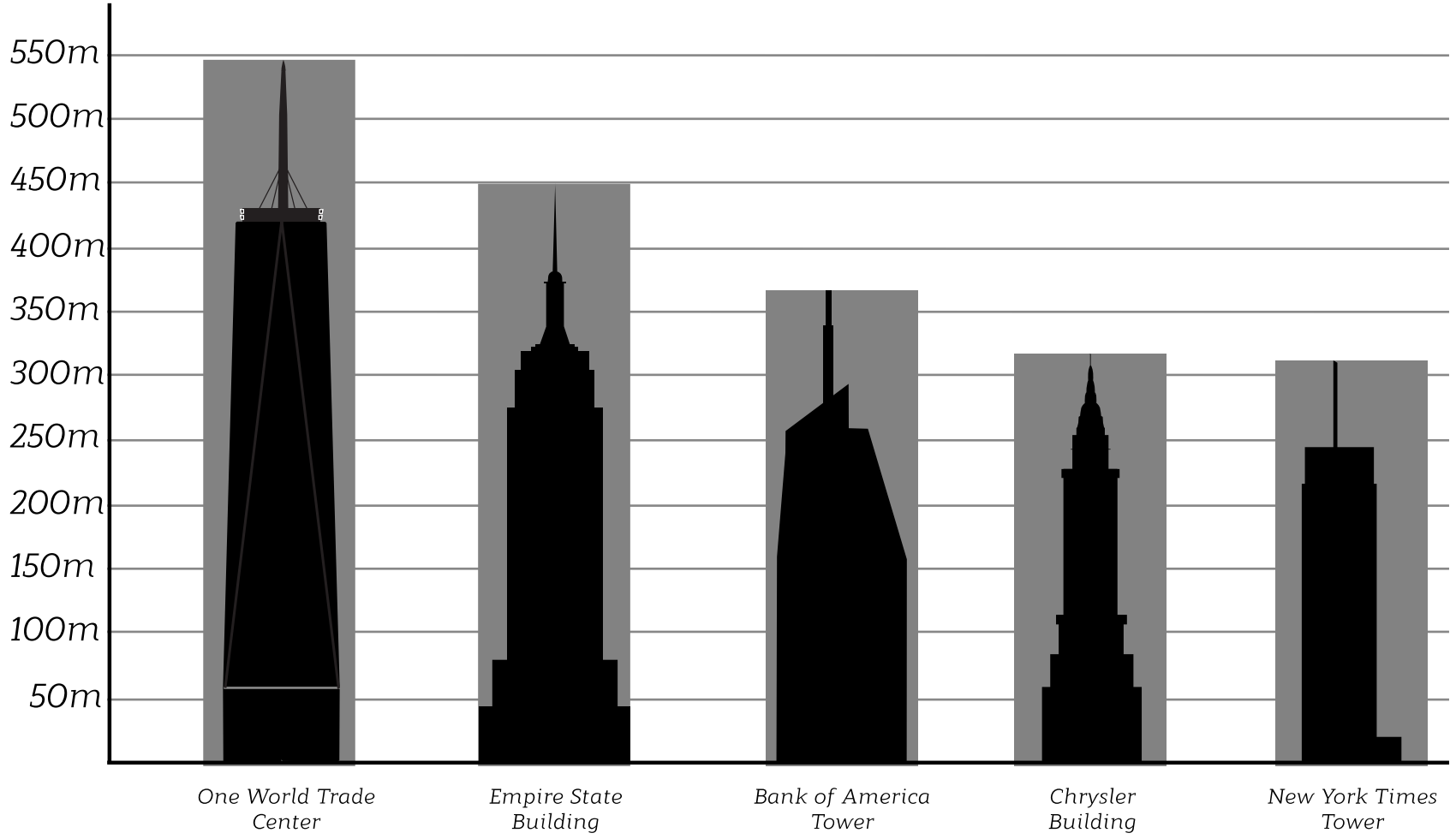
Data Cubes

Same data as…



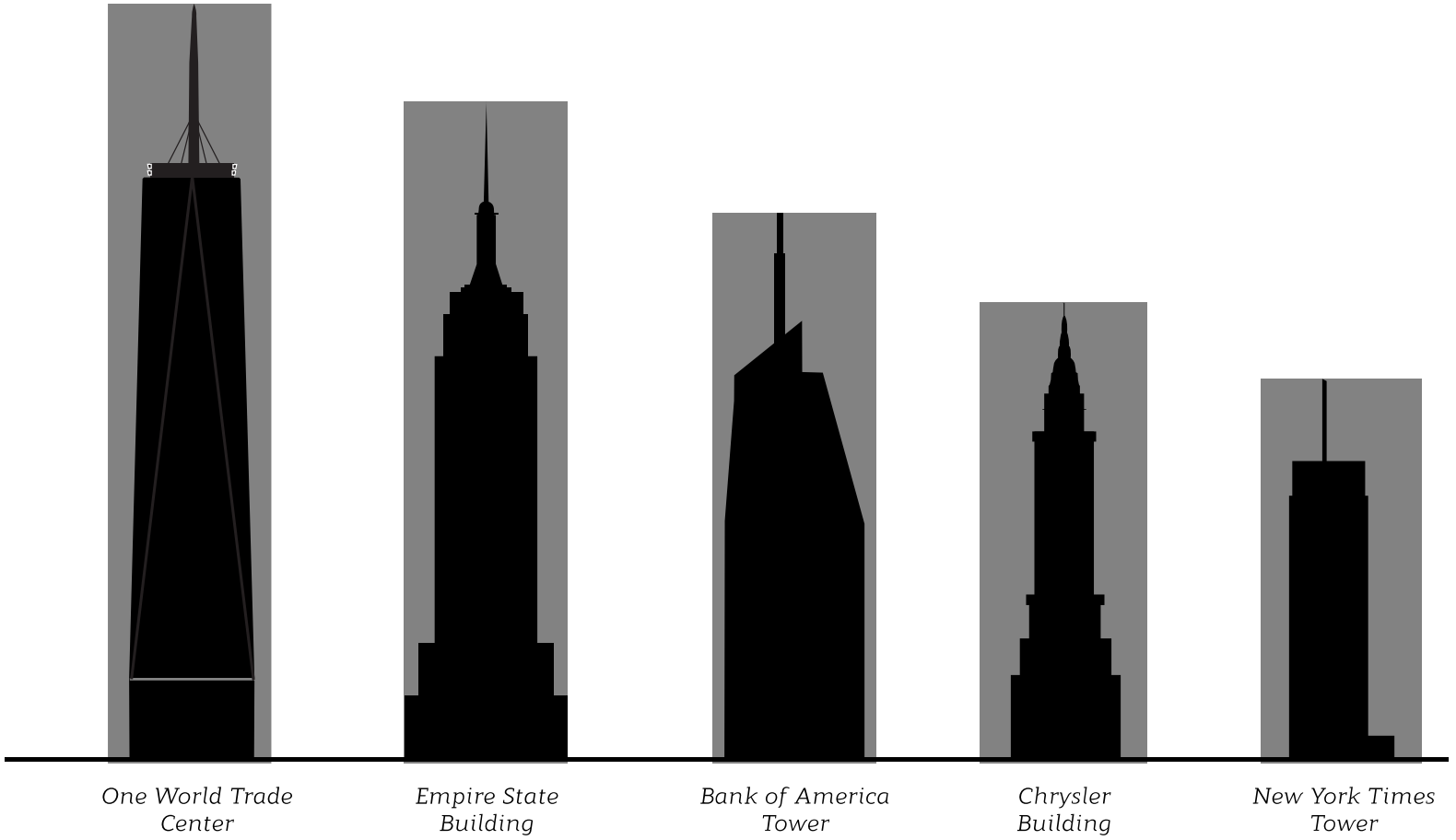
Farm	Month	Fruit	Sales
Farm1	Apr	Pears	400
Farm1	May	Pears	200
Farm1	Apr	Oranges	30
Farm2	Apr	Oranges	100
Farm2	Jul	Apples	100
…	…	…	…

N/O/Q – a taxonomy

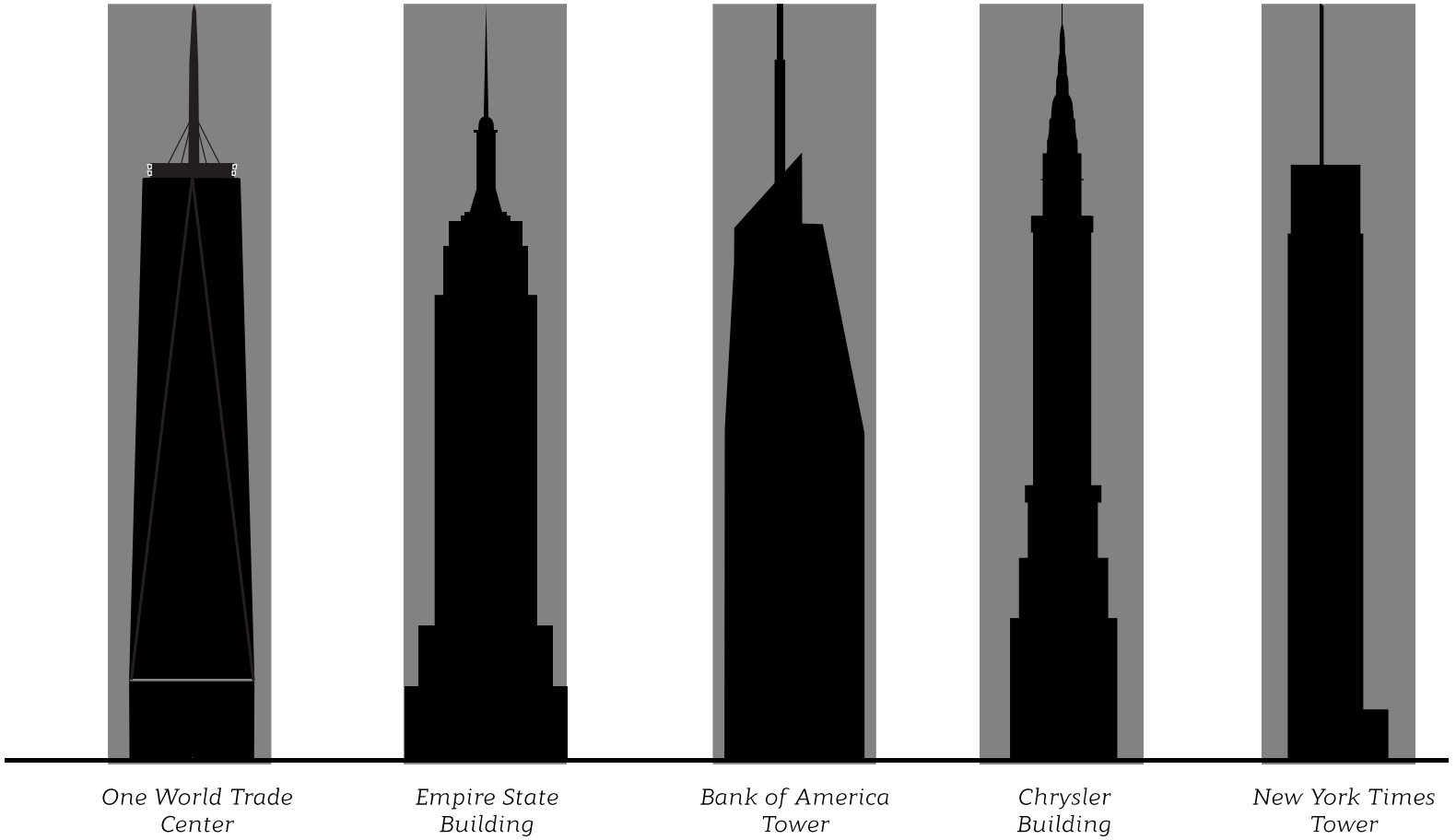


Quantitative
To scale

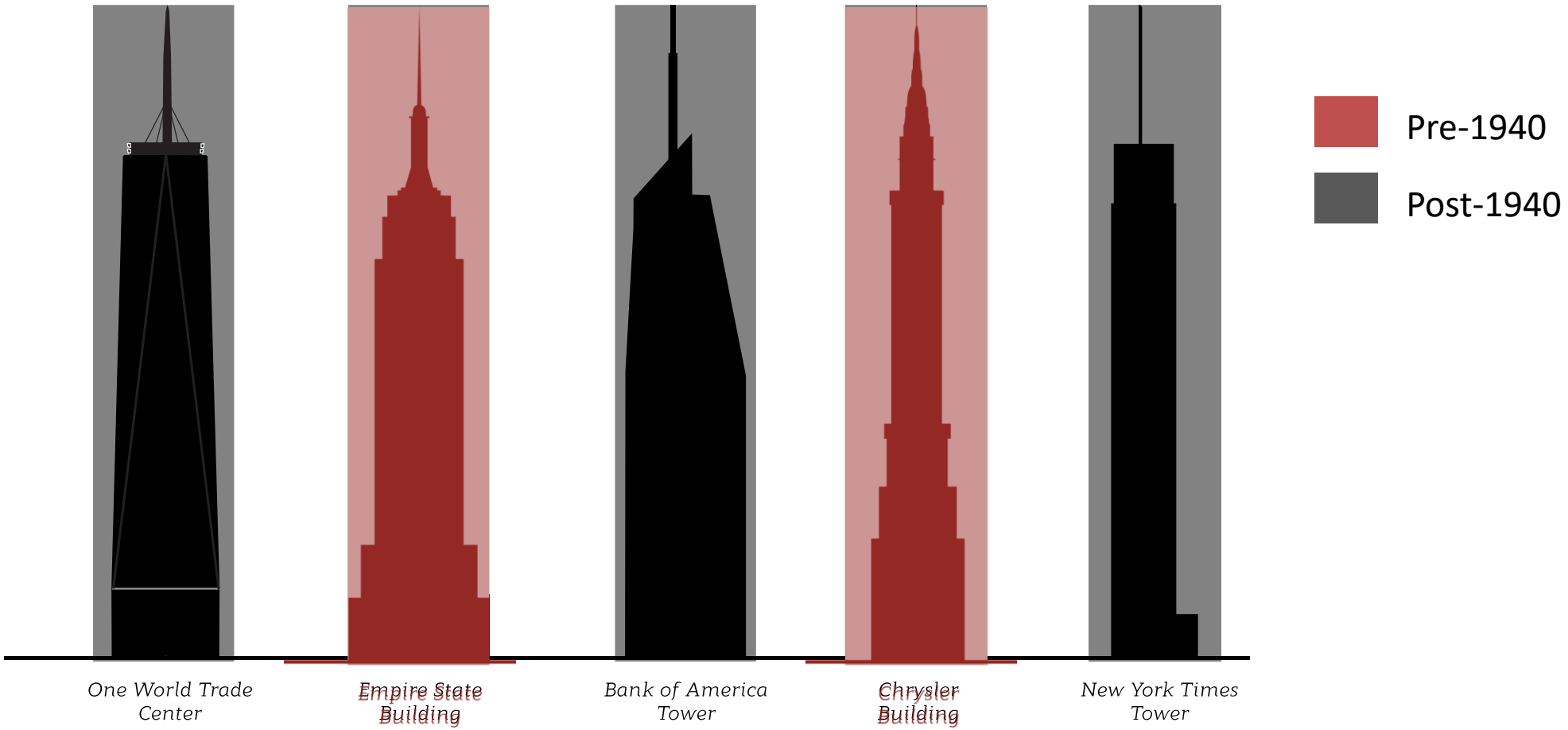
Ordinal
Ranks, Not to scale
(and no scale)

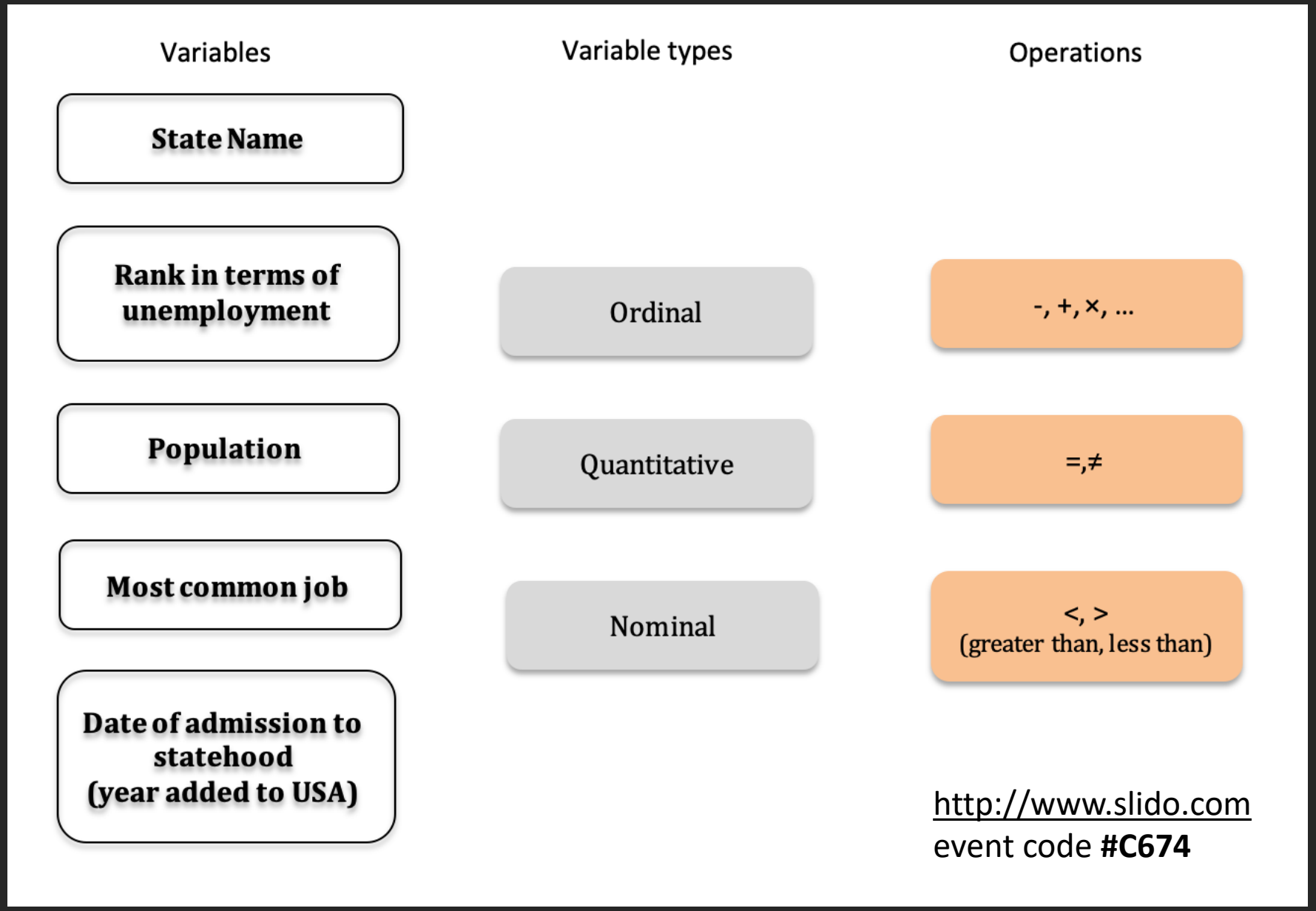


Nominal
Not to scale
(and no scale)



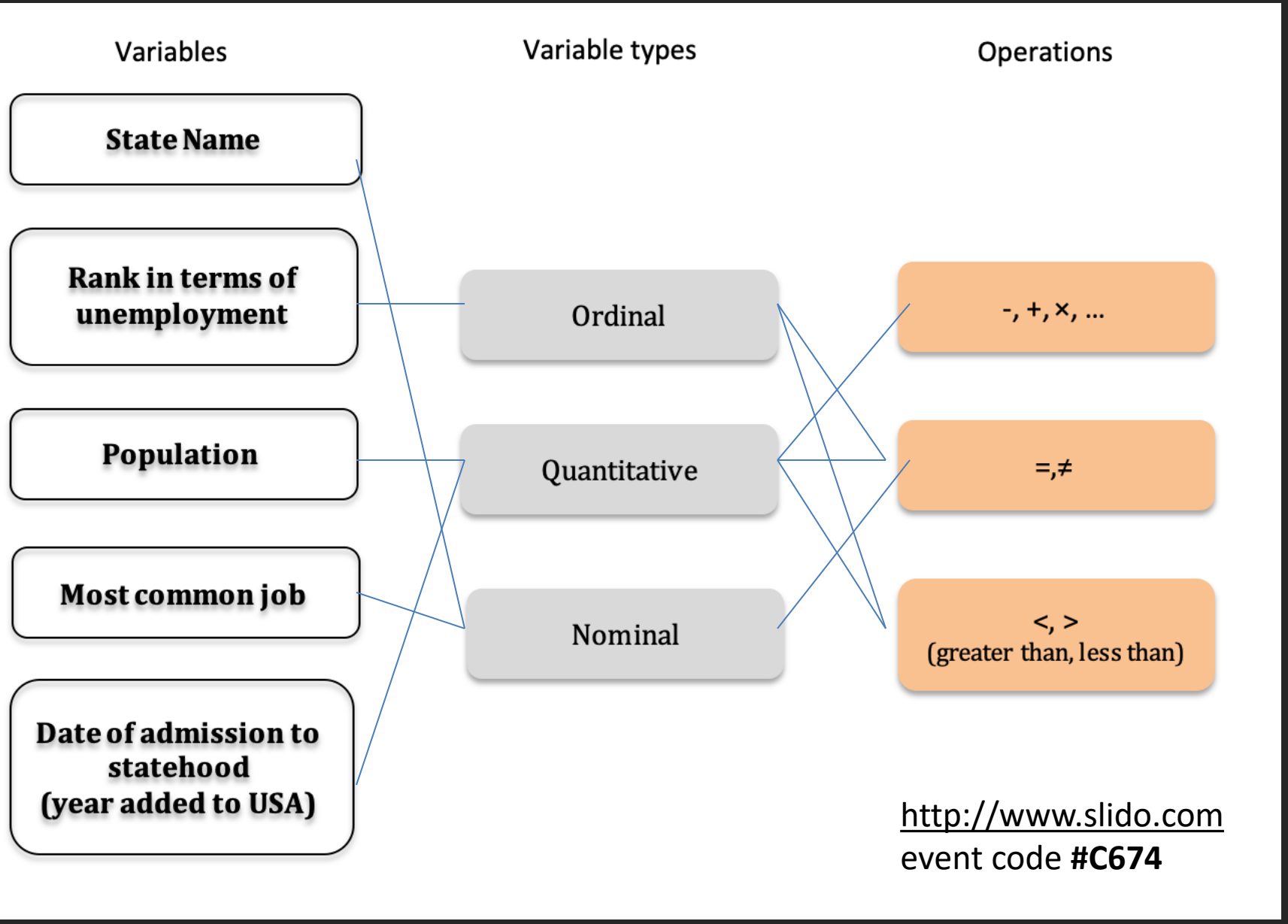
Categorical
Not to scale
(and no scale)





Instructions: (1) Draw one set of lines connecting the ***variables*** to the (best/most likely) ***variable type***. If you feel like there is a second best option, draw a *dashed* line from the variable to the variable type. (2) Draw lines connecting the ***variable types*** to the mathematical ***operations*** that they support.

http://www.slido.com
event code **#C674**



http://www.slido.com
event code **#C674**

Instructions: (1) Draw one set of lines connecting the ***variables*** to the (best/most likely) ***variable type***. If you feel like there is a second best option, draw a *dashed* line from the variable to the variable type. (2) Draw lines connecting the ***variable types*** to the mathematical ***operations*** that they support.

Can we always transform between types?

http://www.slido.com
event code **#C674**

Transformations possible across types, but…

semantics might be weird and transformation might be “lossy”

Roughly…

- You’re constrained a bit by the data you’re given
- But often it’s raw/quant, which means…

- Design Q: What comparison do you need/want to support?

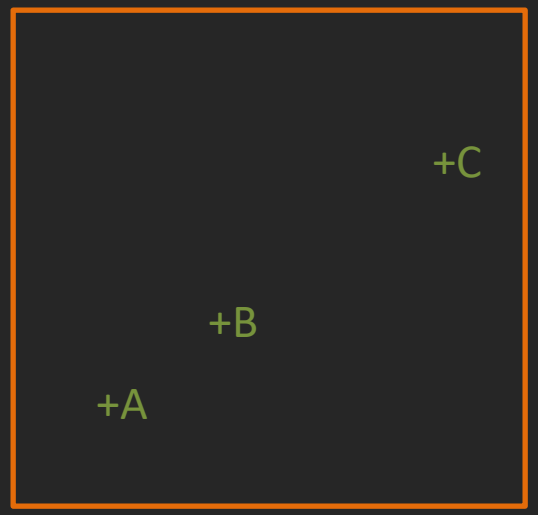
Roughly…

- Should someone be able to tell that two pieces of data are different (alice ≠ bob)?
- Should you be able to tell that one piece of data is larger (alice.height > bob.height)?
- Should you be able to tell how much one piece of data is bigger (average(male height) is 5.5 inches more than average(female height))?

How do we transmit and receive visual information?

Bertin's Semiology

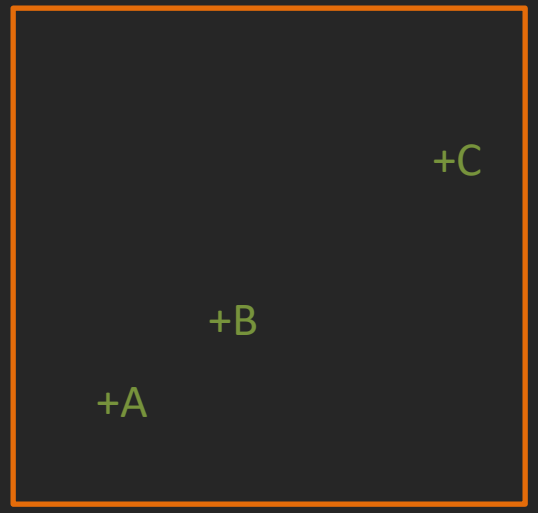
- Images perceived as a set of signs
- Sender encodes information in signs
- Receiver decodes information from signs



Encode quantitative variables

1. A,B, and C and distinguishable
2. B is between A and C
3. BC is twice as long as AB

These are the set of facts this image expresses



Encode quantitative variables

1. A,B, and C and distinguishable
2. B is between A and C
3. BC is twice as long as AB

What is the difference between an analytical task and communicative task?

http://www.slido.com
event code **#C674**

The goal of the visualization

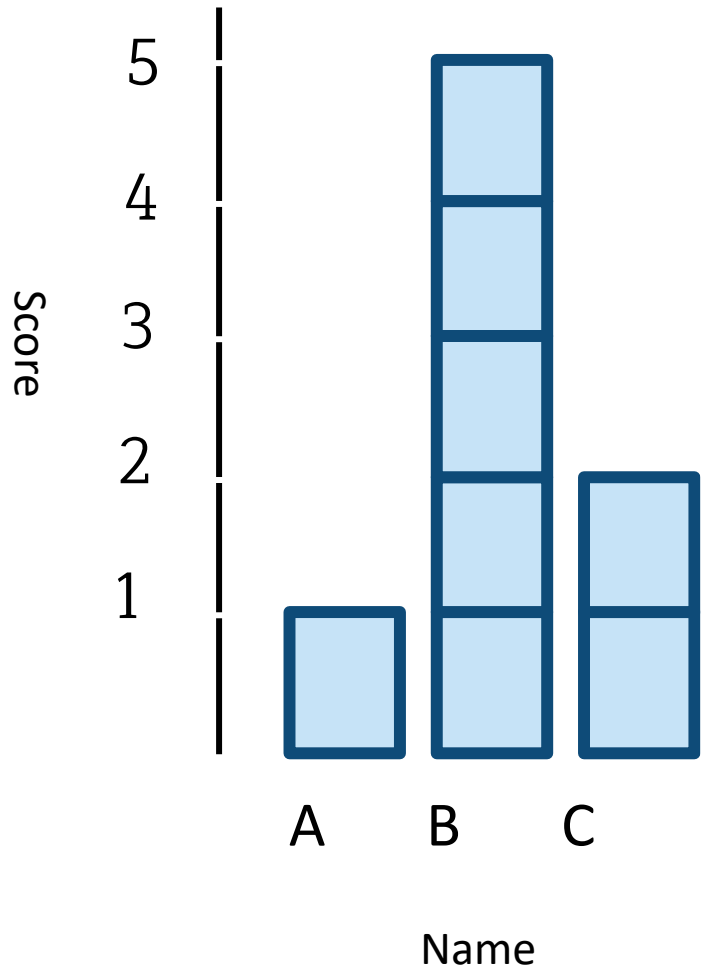
- Analyst role – have data + questions
  - Vis should express the answers
- Communicative role – have data + insights
  - Vis should express the insights
- Everything in between
  - Vis should express answers to questions and insights

Name	Score
A	1
B	5
C	2

Answers

Want to express

1. What is the distribution of values?
2. What is the largest value?
3. What is the smallest?

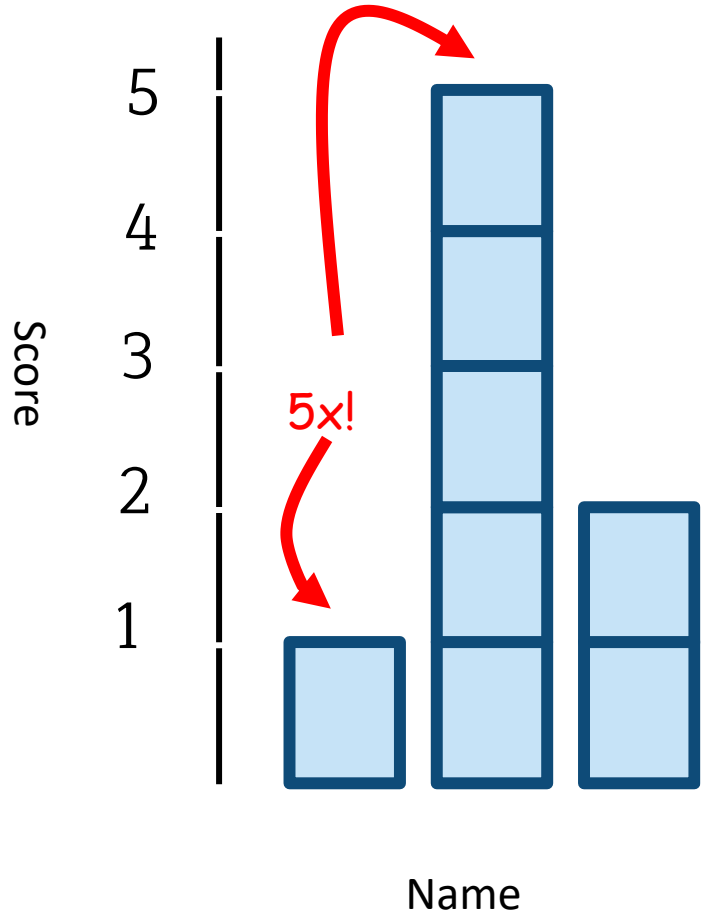


Name	Score
A	1
B	5
C	2

Insights

Want to express

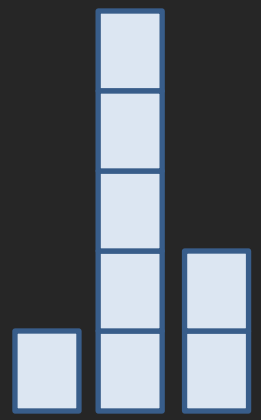
1. A is 1, B is 5, C is 2
2. B is 5 times larger than A
3. C is 2 times larger than A



Name	Score
A	1
B	5
C	2

Sender “Encodes” using Functions/Rules

Draw a bar for every person
Bar’s height should be 1 inch for every point



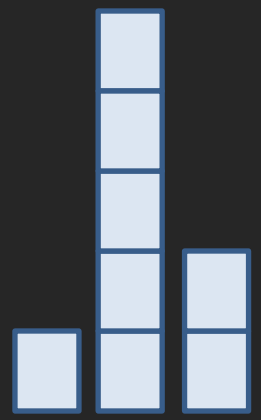
Why is encoding “rule based?”

Name	Score
A	1
B	5
C	2

Sender “Encodes” using Functions/Rules

Name	Score
A	1
B	5
C	2

Receiver “decodes” using Functions/Rules



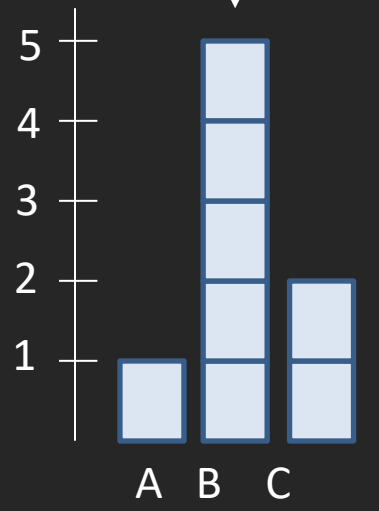
Just seeing the figure, is there enough information to decode?

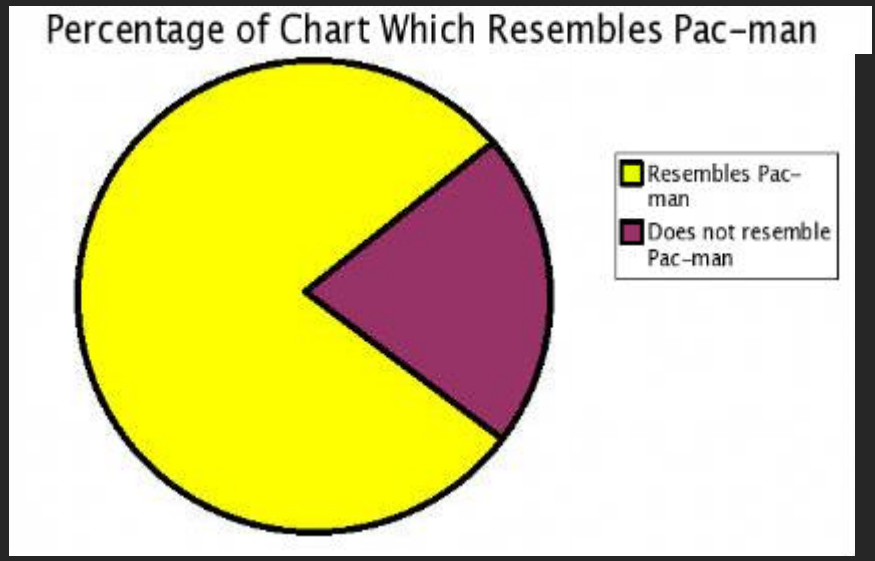
Name	Score
A	1
B	5
C	2

Sender “Encodes” using Functions/Rules

Name	Score
A	1
B	5
C	2

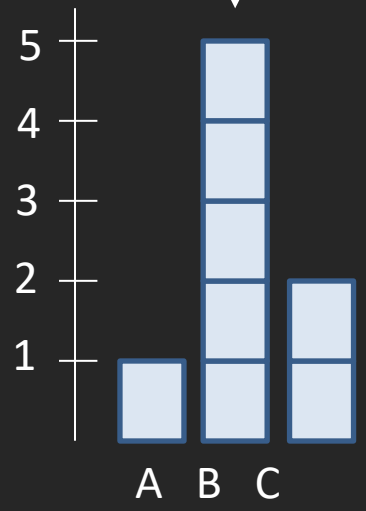
Receiver “decodes” using Functions/Rules





Name	Score
A	1
B	5
C	2

Sender “Encodes” (bar chart)



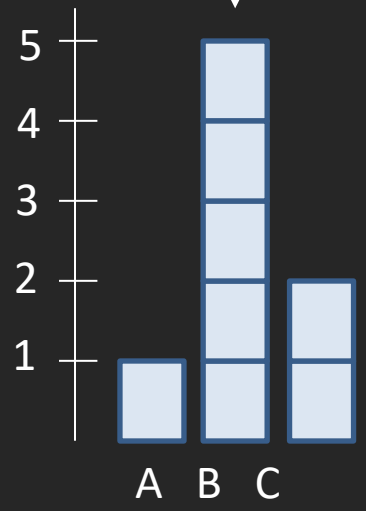
Name	Score
A	1
B	5
C	2

Sender “Encodes” (color dots)

Draw a circle for every person
Set 1 to white, 5 to red, interpolate everything between

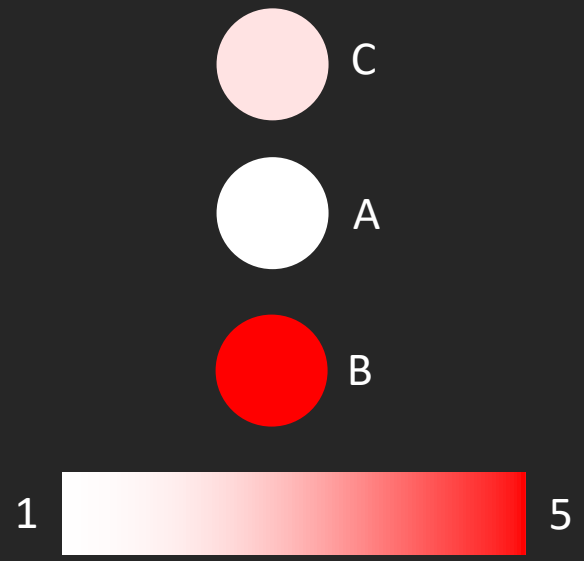
Name	Score
A	1
B	5
C	2

Sender “Encodes” (bar chart)



Name	Score
A	1
B	5
C	2

Sender “Encodes” (color dots)



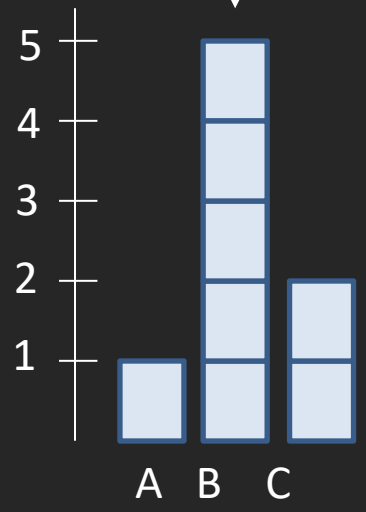
Name	Score
A	1
B	5
C	2

Sender “Encodes” (size stars)

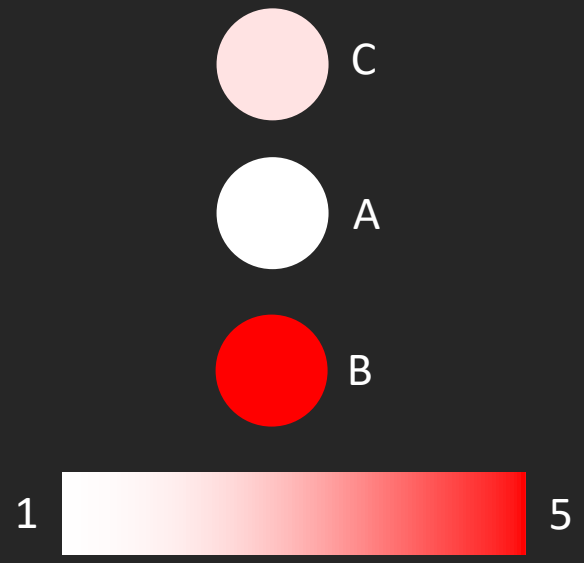
Which is better?

99% of this class is on making the right choice…

Sender “Encodes” (bar chart)

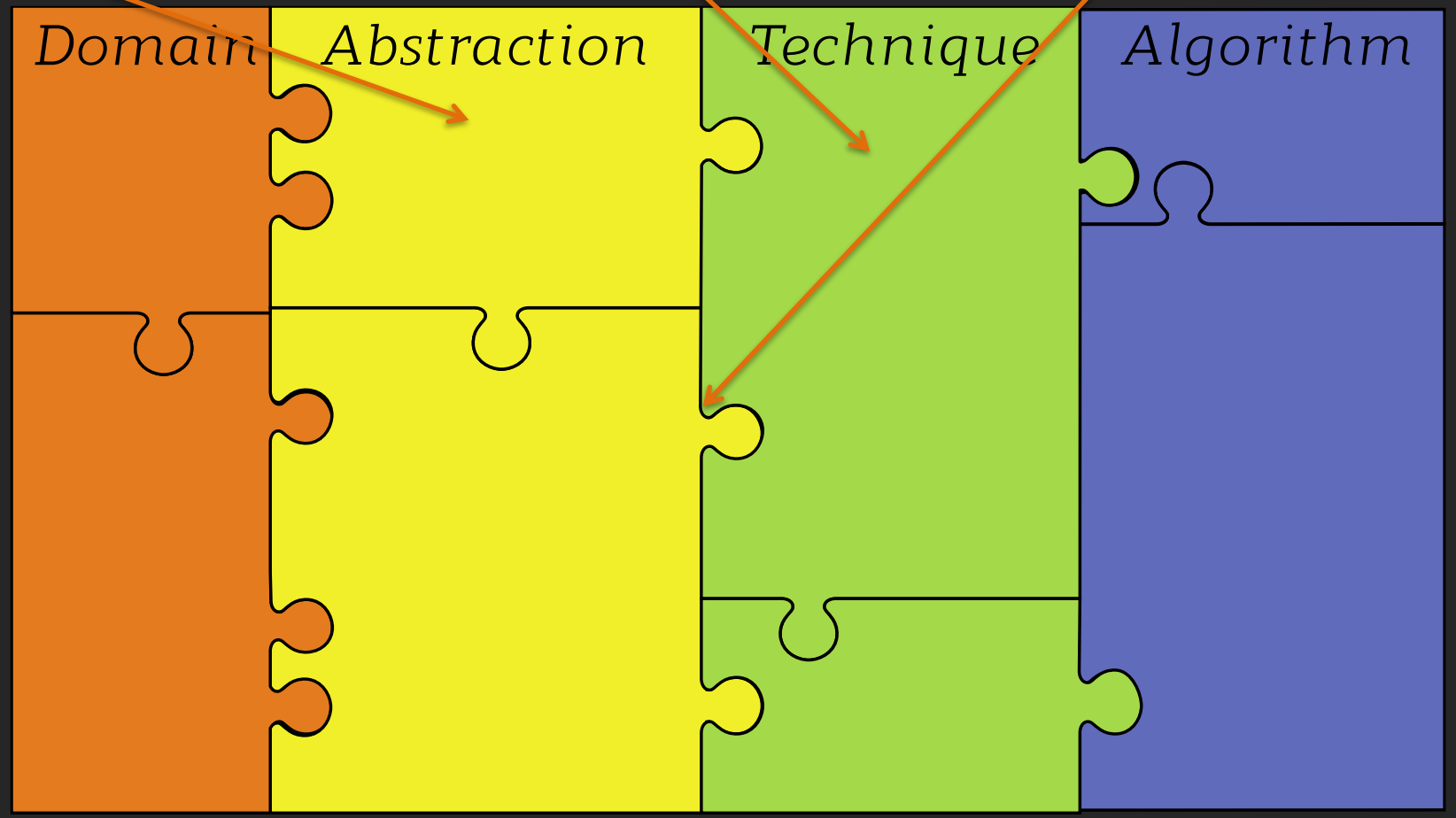


Sender “Encodes” (color dots)



Sender “Encodes” (size stars)

Given the data and tasks, are my choices in representation *good*?

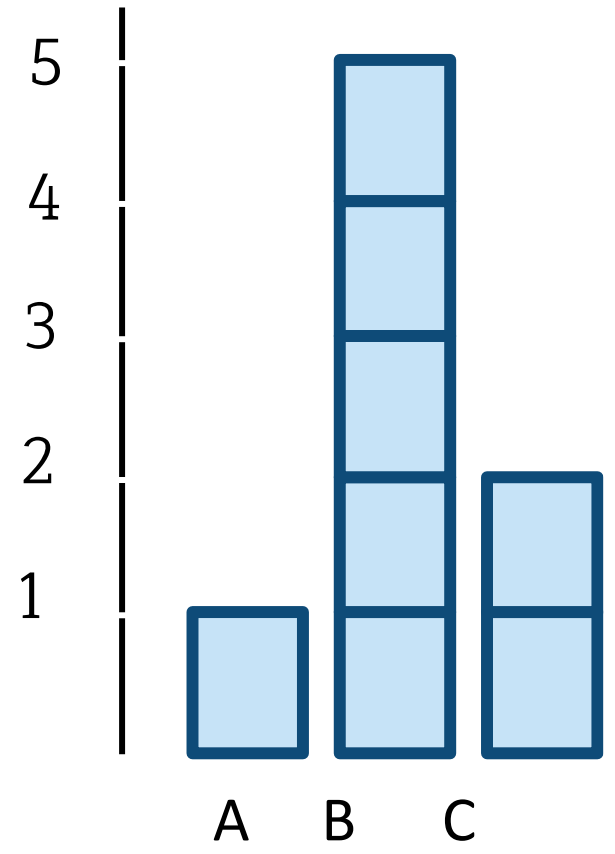


Name	Score
A	1
B	5
C	2

Insights

Want to express

1. A is 1, B is 5, C is 2
2. B is 5 times larger than A
3. C is 2 times larger than A

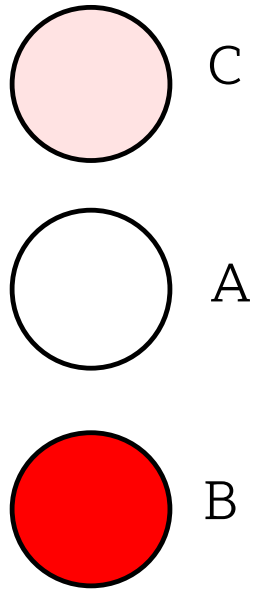


Name	Score
A	1
B	5
C	2

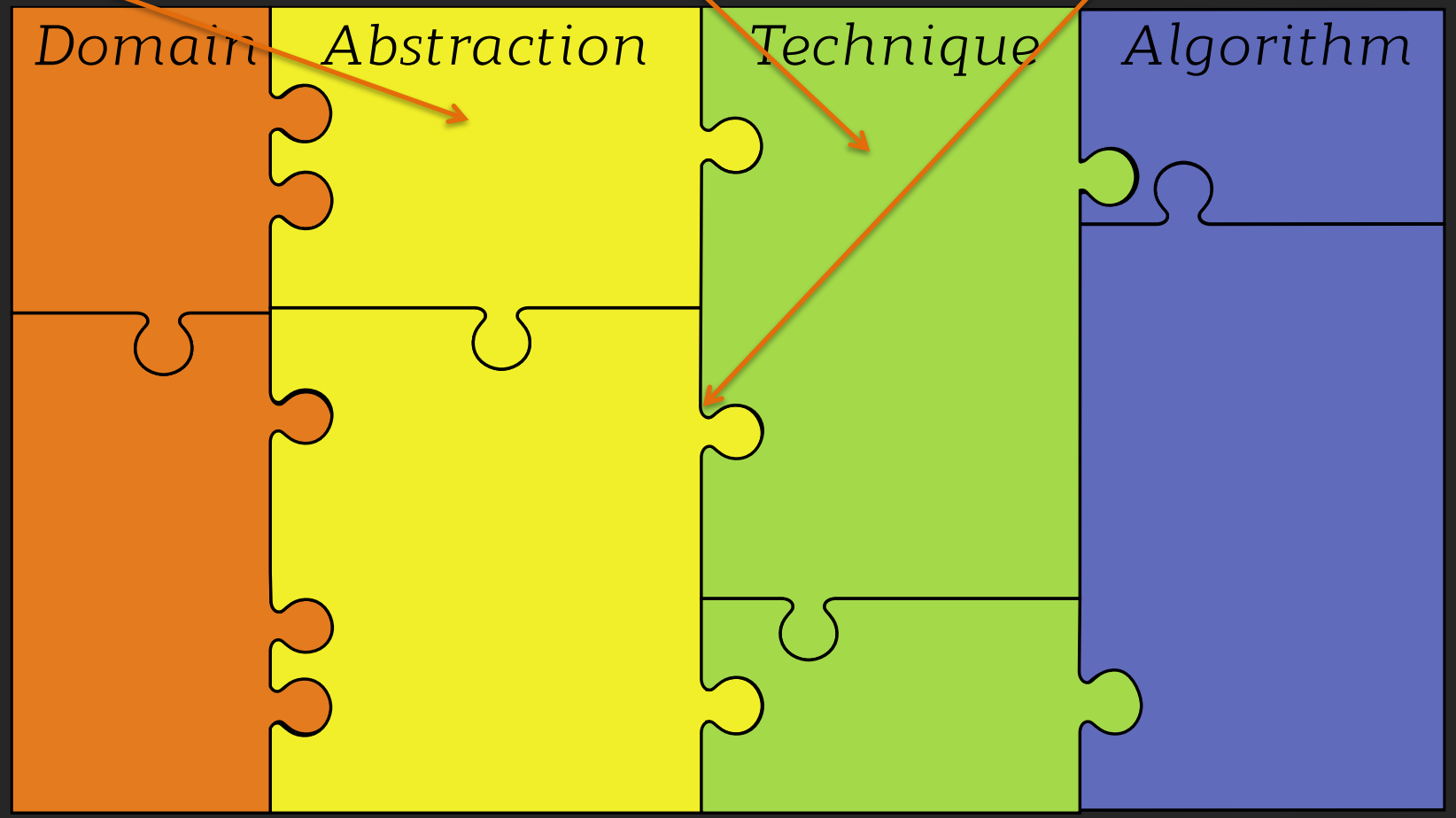
Insights

Want to express

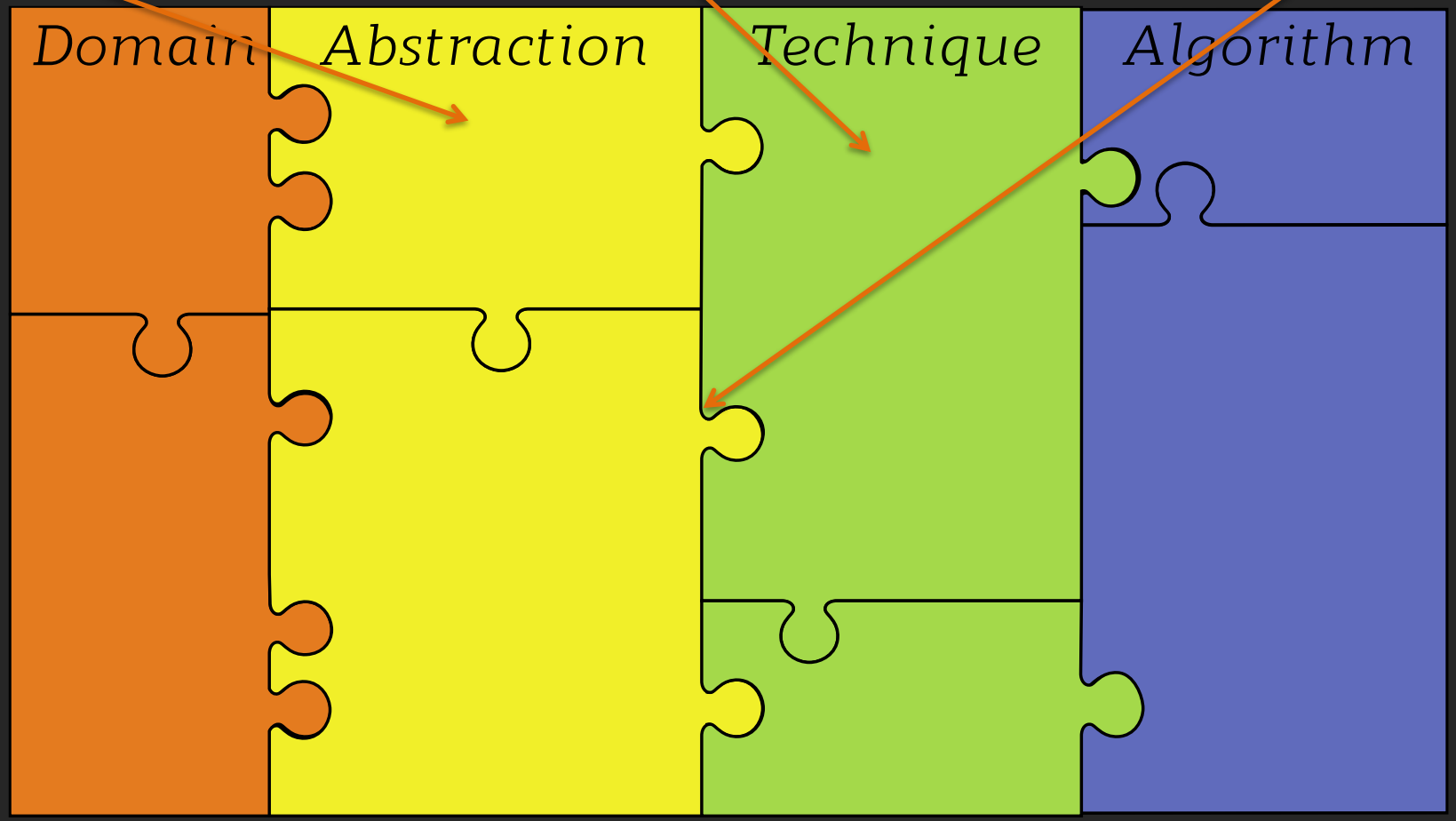
1. A is 1, B is 5, C is 2
2. B is 5 times larger than A
3. C is 2 times larger than A



Given the data and tasks, are my choices in representation *good*?

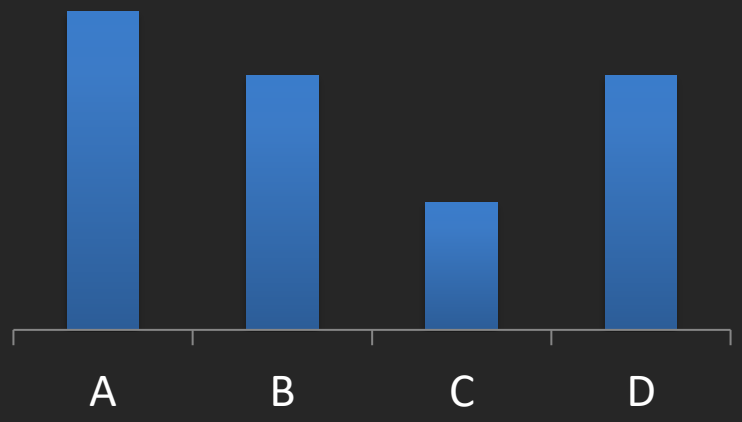


Given the data and tasks, are my choices in representation effective and expressive?

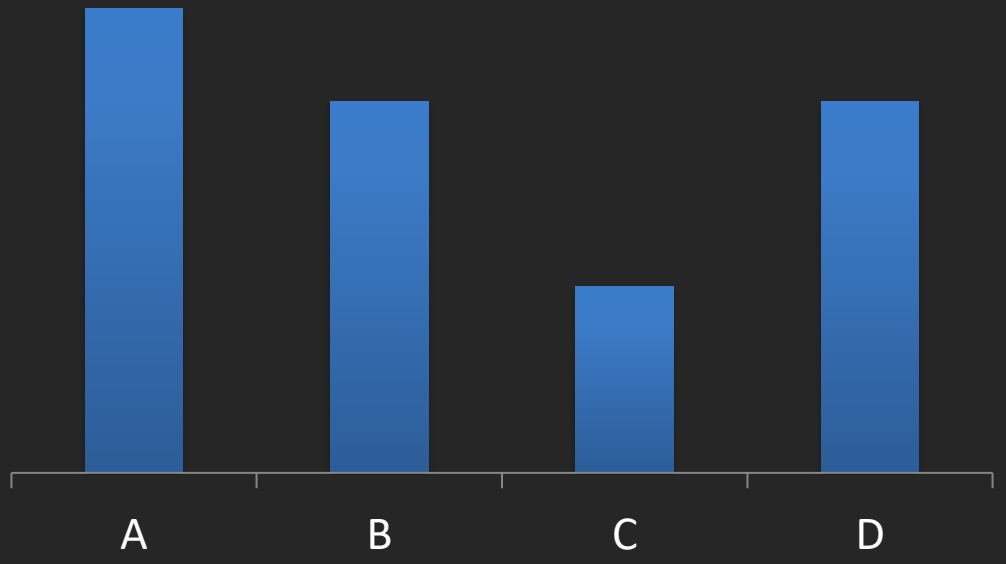


Expressive

- A graph expresses…
  - The statistical “facts” that we can infer from the diagram
- A graph does not express…
  - The statistical “facts” that are *impossible* to infer from the diagram



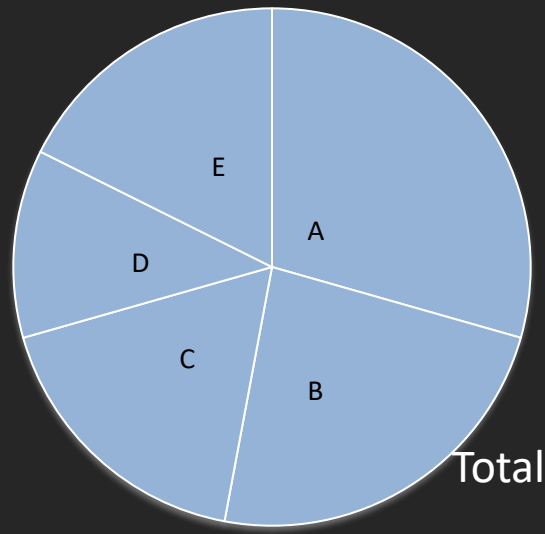
What is expressed here?
What is *not* expressed here?



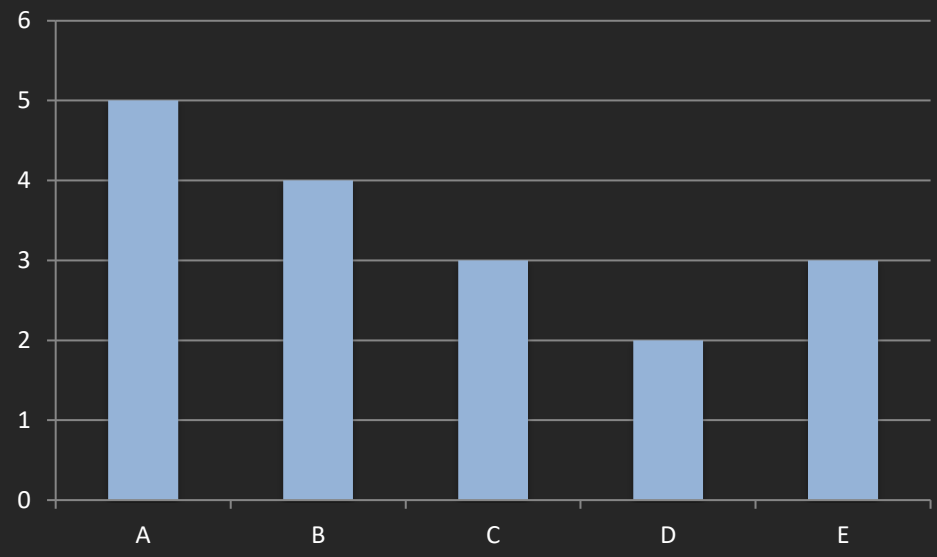
Expressive versus Effective

- Important difference:
  - *(Im)possible* to infer has to do with expressiveness
  - *Easy/hard* to infer has to do with effectiveness

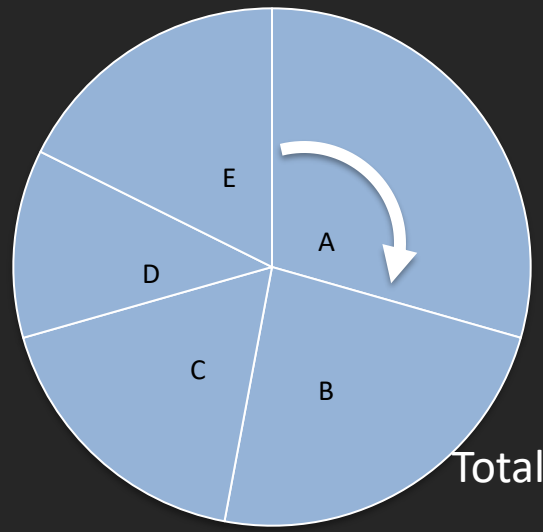
Which is more expressive (expresses more facts)?



Total size: 17

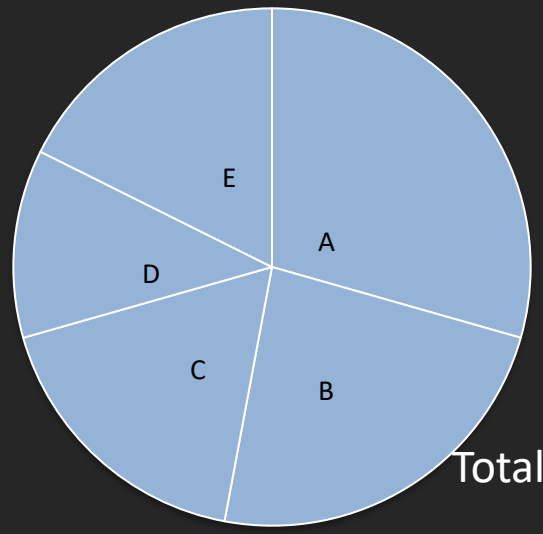


http://www.slido.com
event code **#C674**



Total size: 17

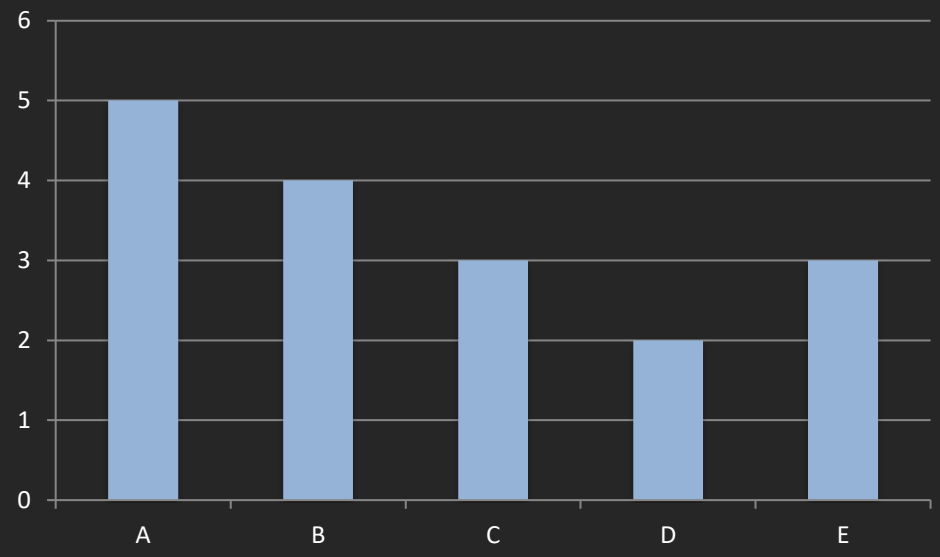
ID	Value
A	5
B	4
C	3
D	2
E	3



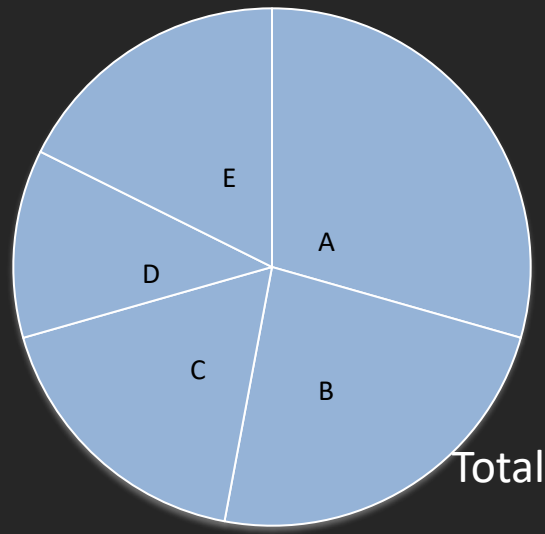
Total size: 17

Both tables are the same, so I can answer the *exact same* set of questions!

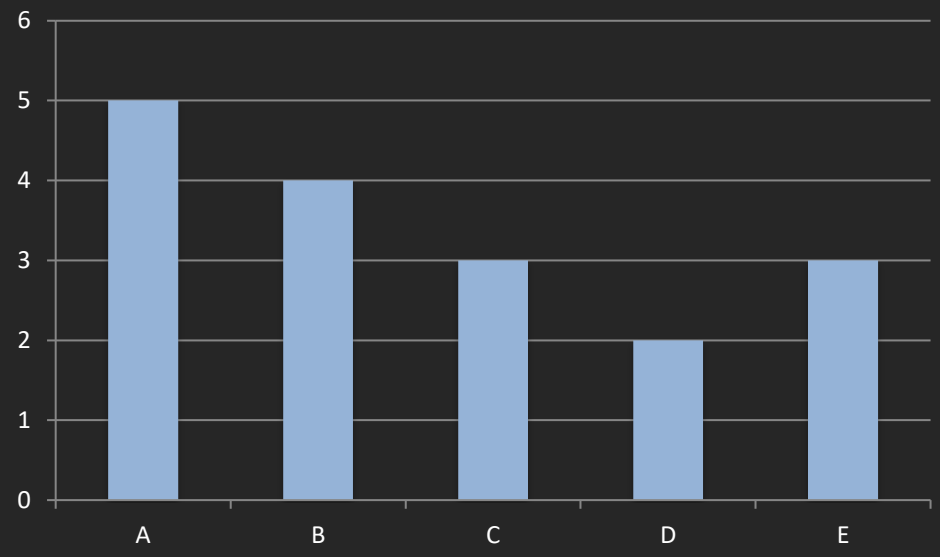
- Is A > B?
- How much bigger is B than C?
- What is the 3rd largest value?
- Are C and E the same?
- And so on…



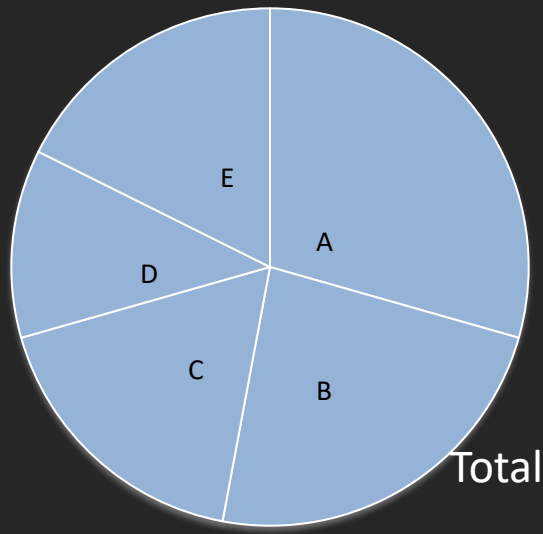
They are equally expressive!



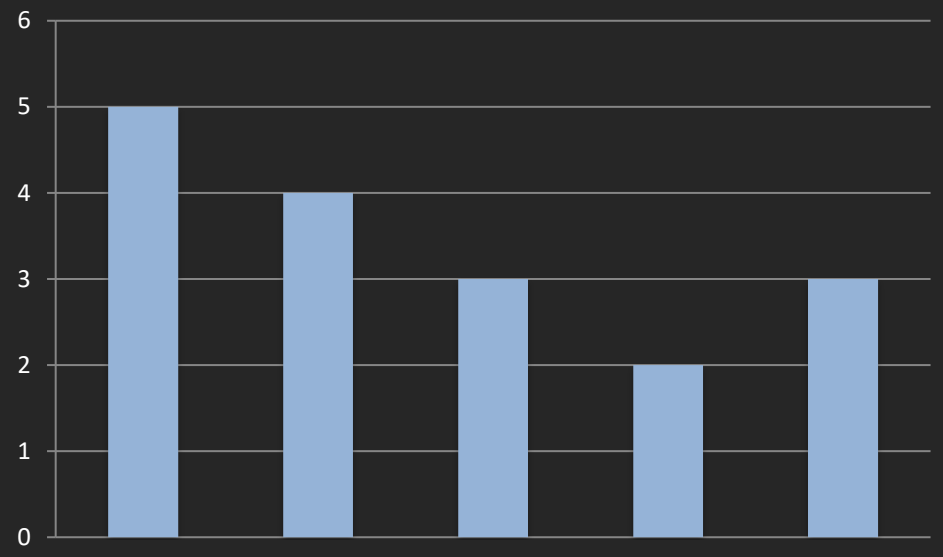
Total size: 17



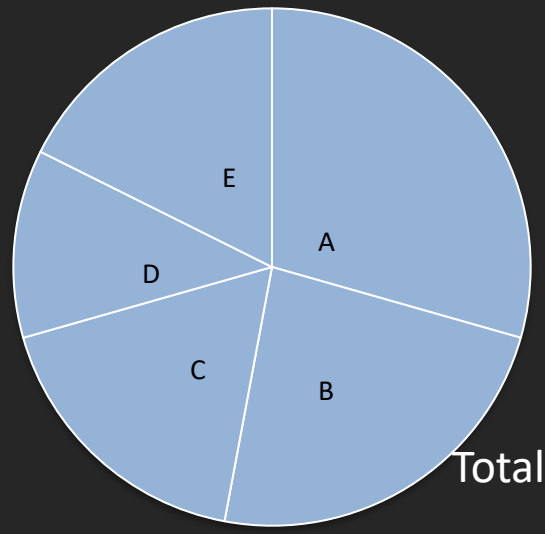
How about now?



Total size: 17

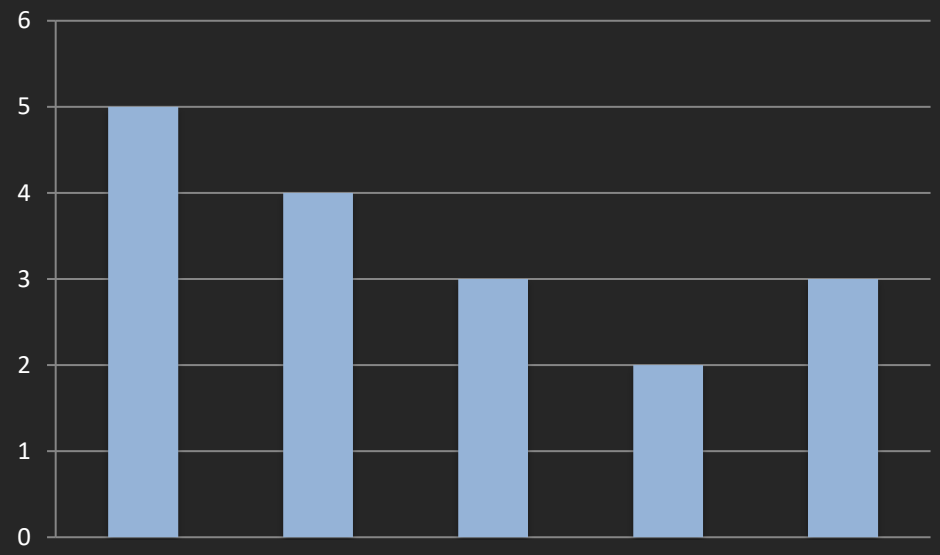


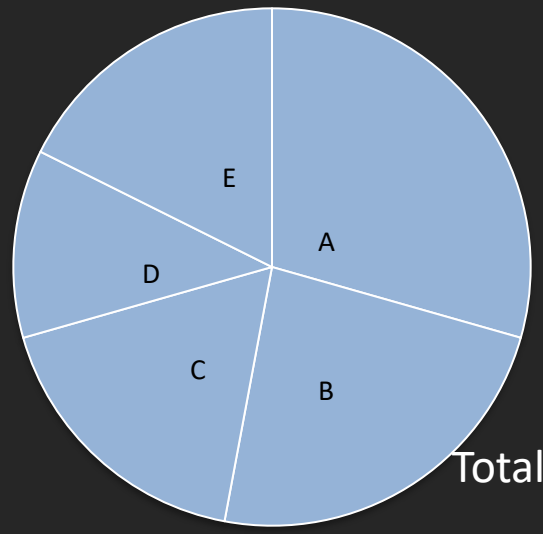
ID	Value
A	5
B	4
C	3
D	2
E	3



Total size: 17

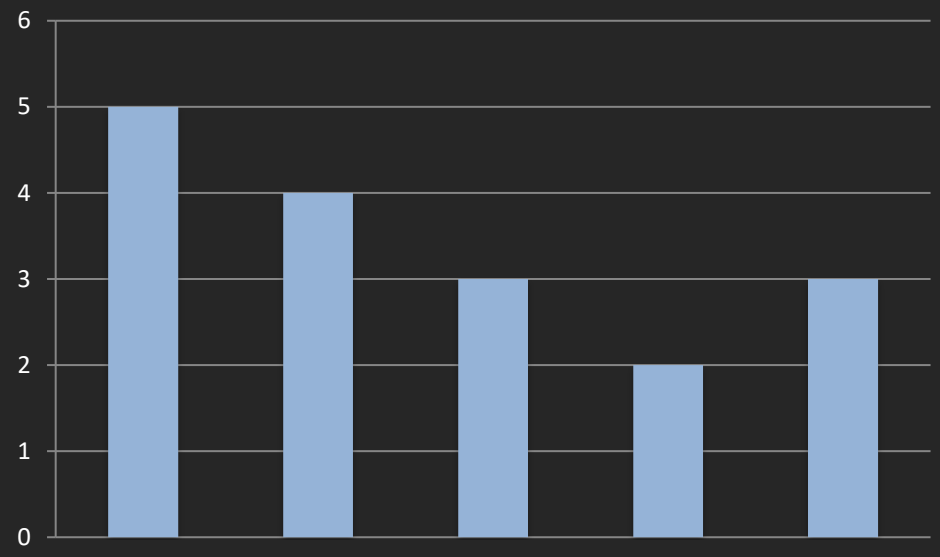
Value
5
4
3
2
3



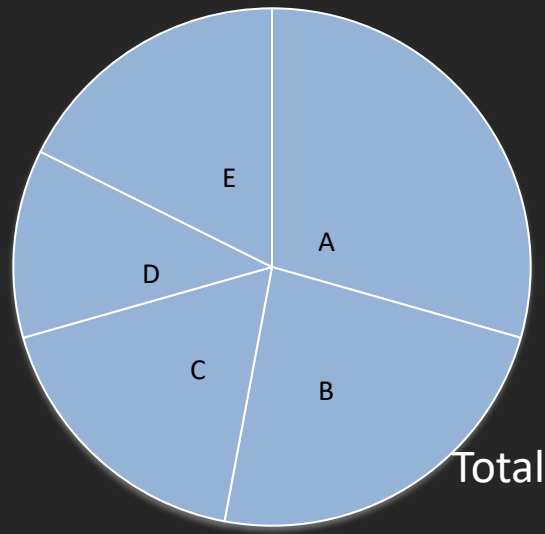


Total size: 17

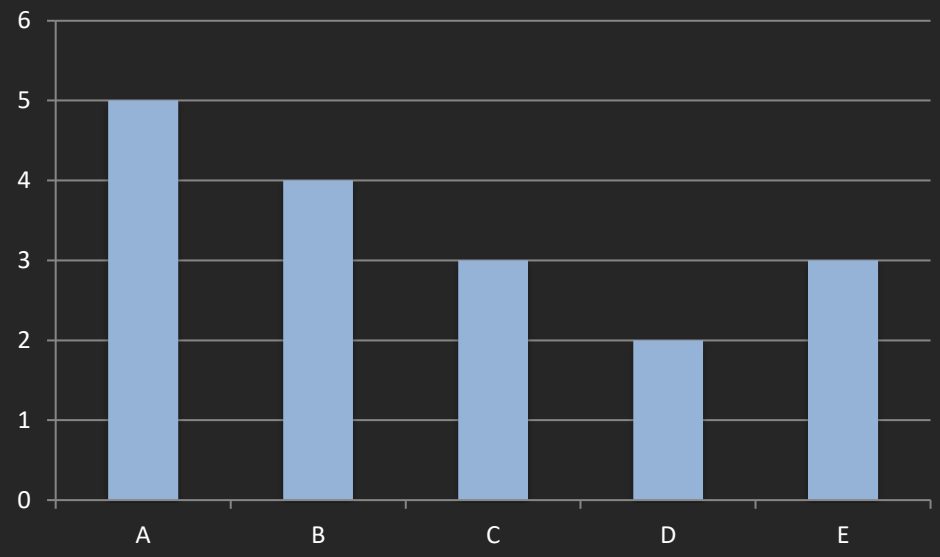
Can no longer answer: Is A>B? or how much bigger is A relative to B?



Which is more effective?



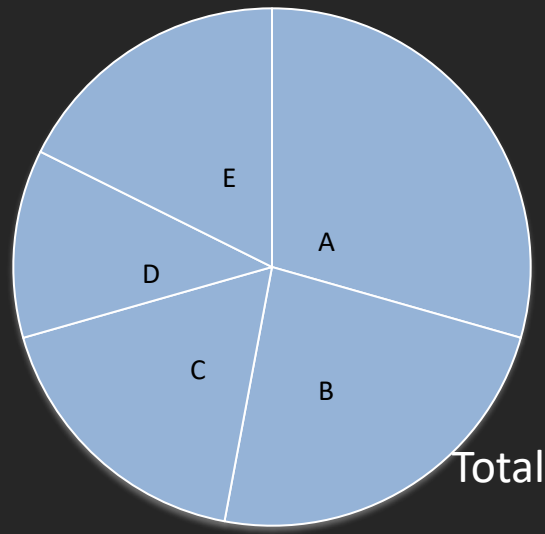
Total size: 17



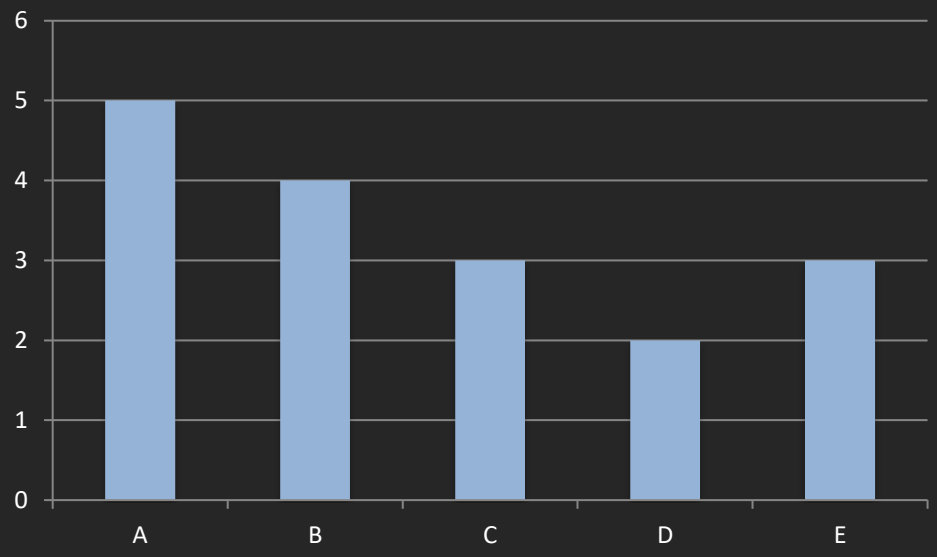
http://www.slido.com
event code **#C674**

Effectiveness?

- You need to ask: “effective for what?”
- What percentage of the whole is A?
- How much bigger is A relative to B?

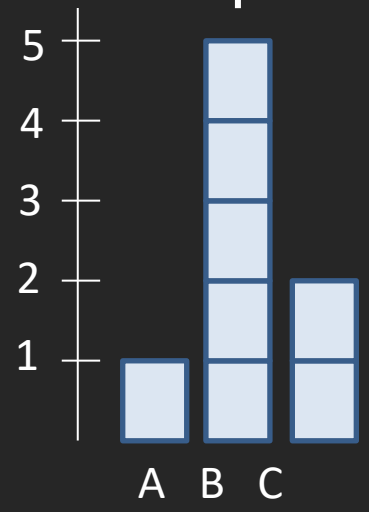


Total size: 17



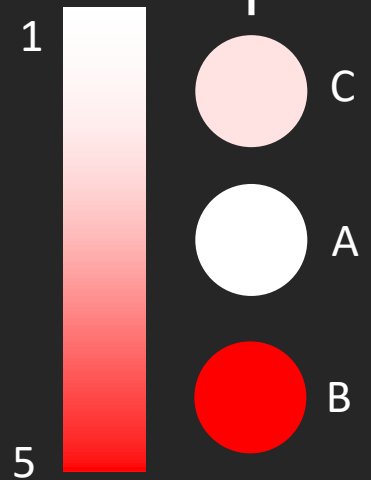
Name	Score
A	1
B	5
C	2

Decode (bar chart)



Name	Score
A	1? 1.1?
B	4.9? 5?
C	1.9? 2? 2.1?

Decode (color dots)



Name	Score
A	1
B	5
C	2? 3?

Decode (size stars)

True answer:

Name	Score
A	1
B	5
C	2

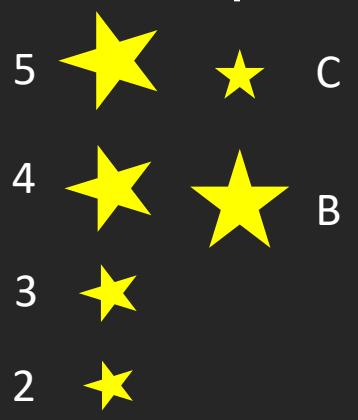
Effectiveness: difference between true answer and what the person says/sees

Name	Score
A	1
B	5
C	2? 3?

Decode (size stars)

Name	Score
B	5
C	2? 3?

Encode/Decode
(size stars++)



Draw a star for every person
Set size proportional to score, throw away scores <= 1

True answer:

Name	Score
A	1
B	5
C	2

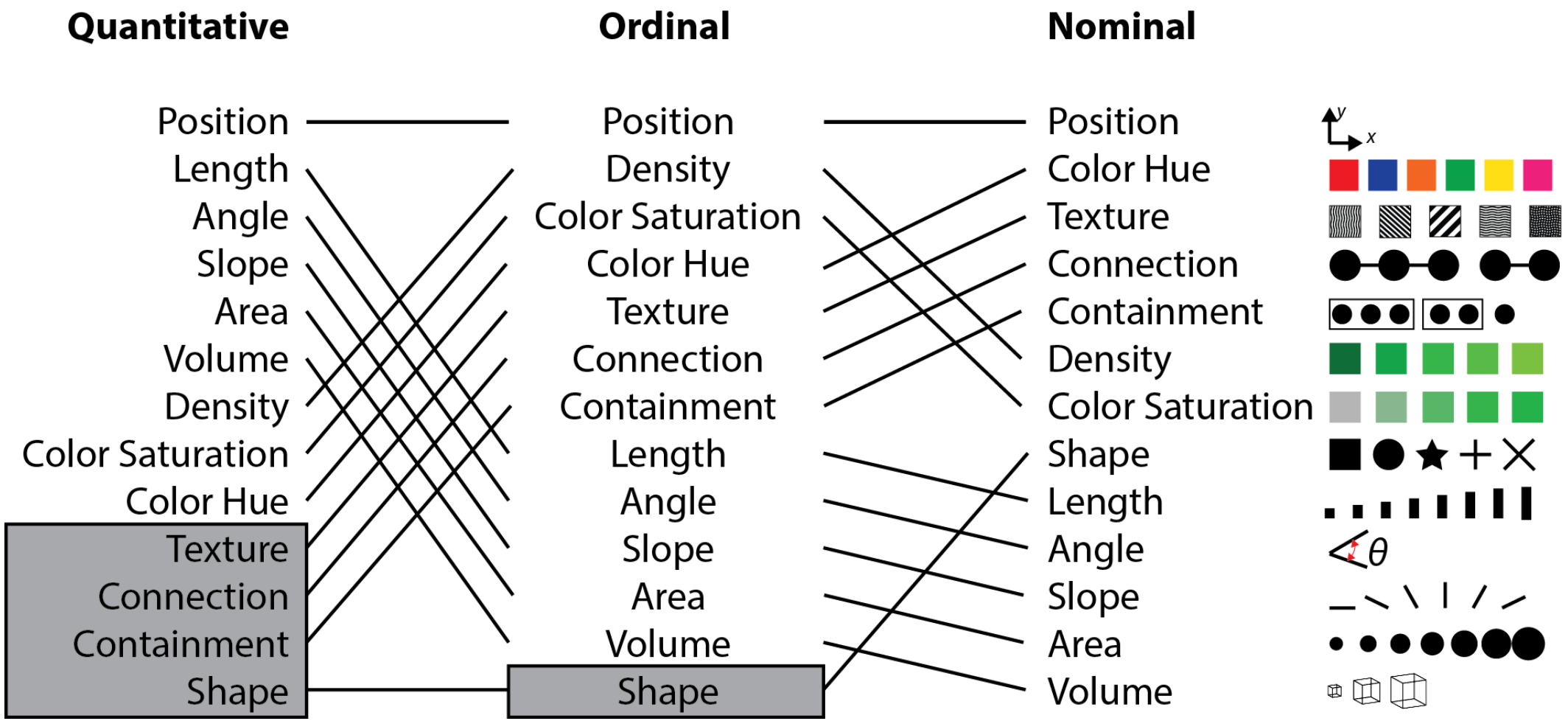
Expressiveness: difference between *amount* of information we can decode

Remember

- To answer expressiveness question:
  - Figure out what data is encoded
  - Distinguish between what is possible/impossible
  - Not important what is easy/hard

Remember

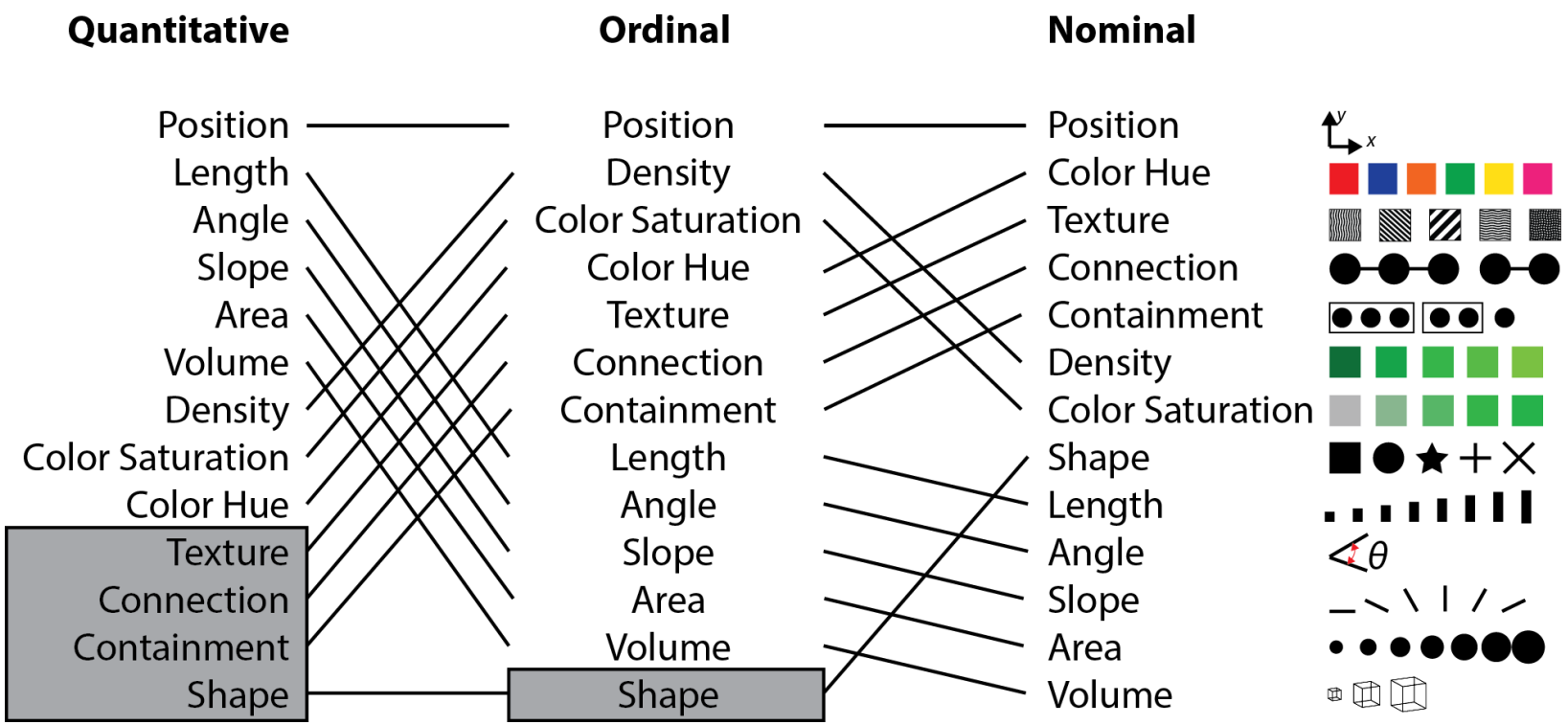
- To answer effectiveness question:
  - First figure out what you need to be effective at…
  - And then you can start applying design philosophy, perceptual rules, and instinct
  - For a human, how much faster/more accurate are they in decoding specific facts?



From Mackinlay

Be greedy:

1. pick best rep. given your most important comparison,
2. cross off list,
3. continue with next most important



From Mackinlay

How do we decide what’s important?

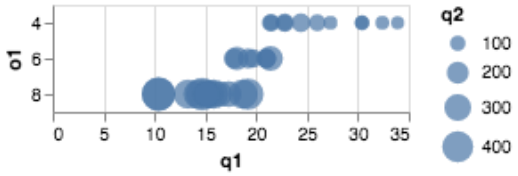
Ranking of Perceptual Tasks

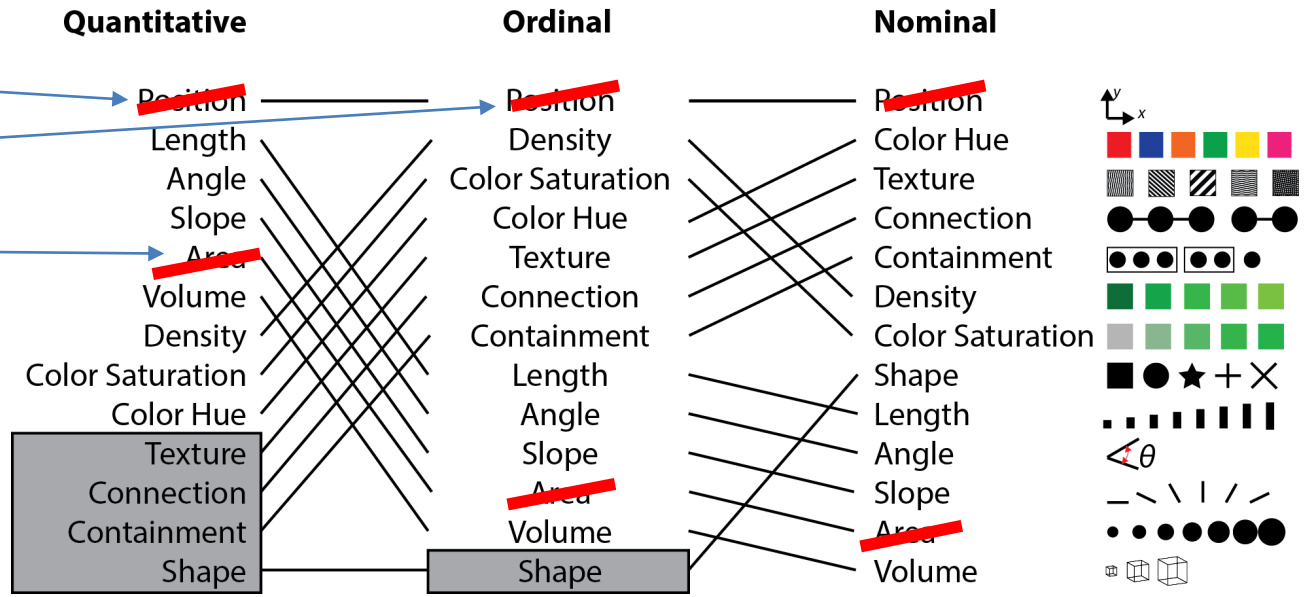
In order of importance

Q1

O1

Q2

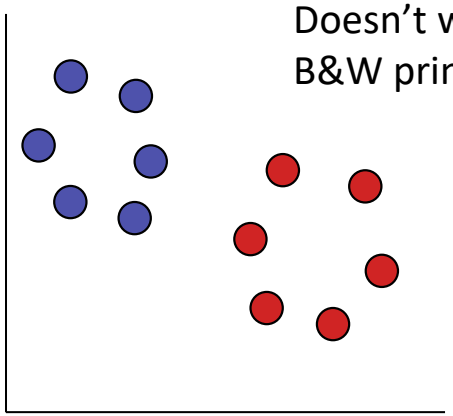




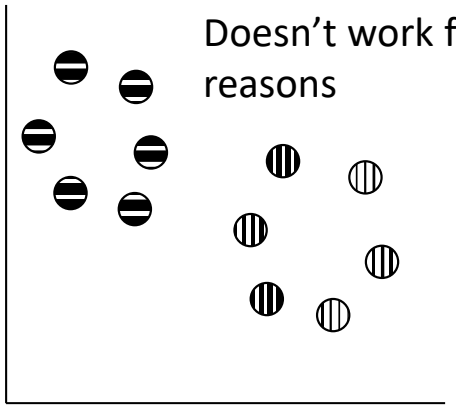
From Mackinlay

Caveats

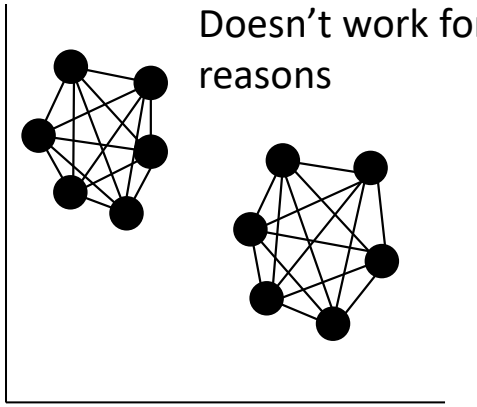
- Sometimes we “skip”



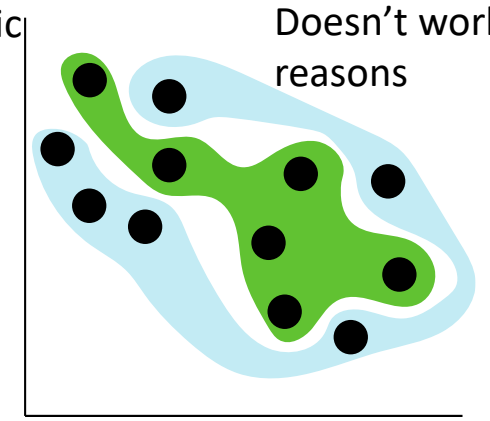
Doesn’t work on B&W printers



Doesn’t work for design reasons



Doesn’t work for semantic reasons



Doesn’t work for data reasons

Nominal

Position

Color Hue

Texture

Connection

Containment

Density

Color Saturation

Shape

Length

Angle

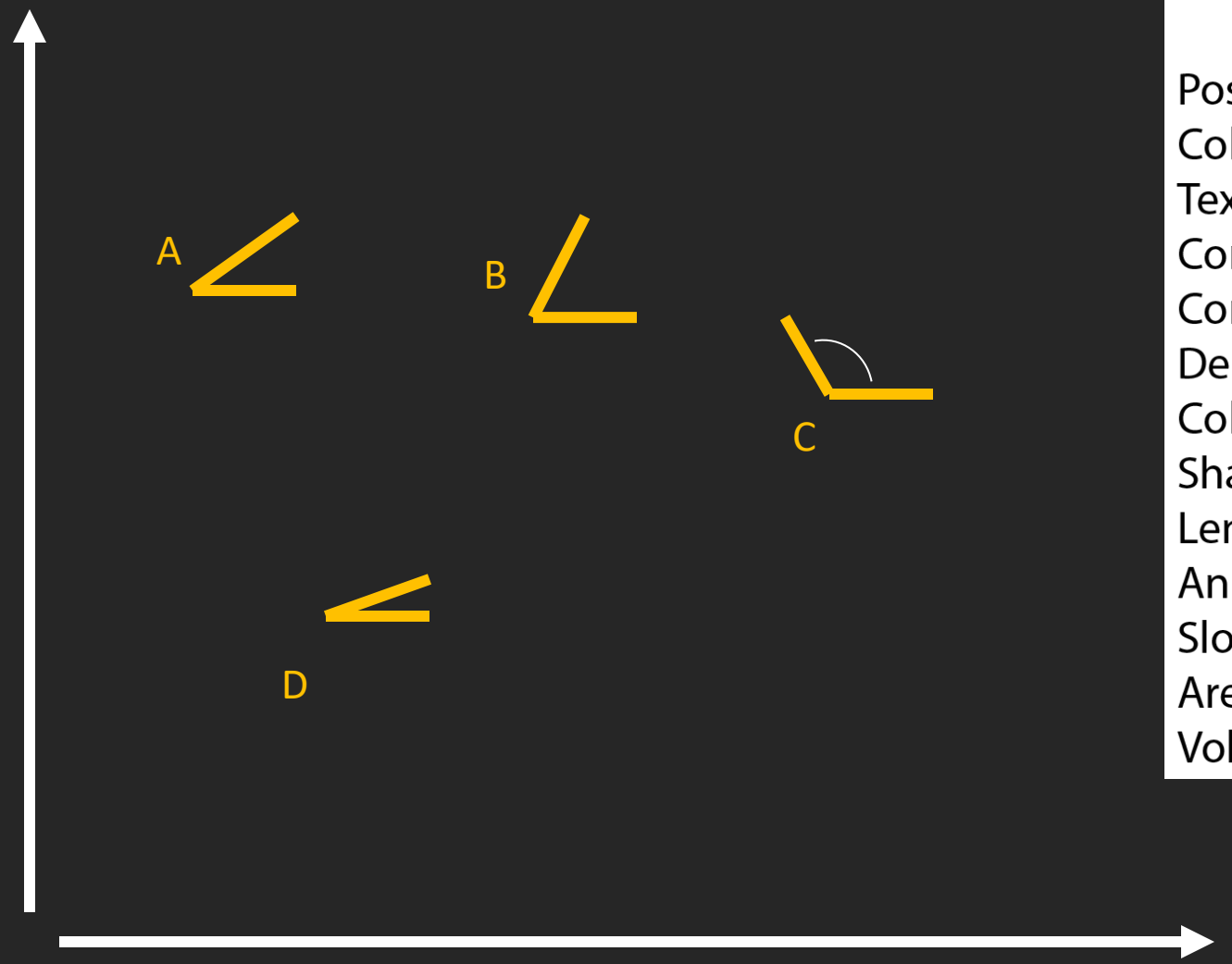
Slope

Area

Volume

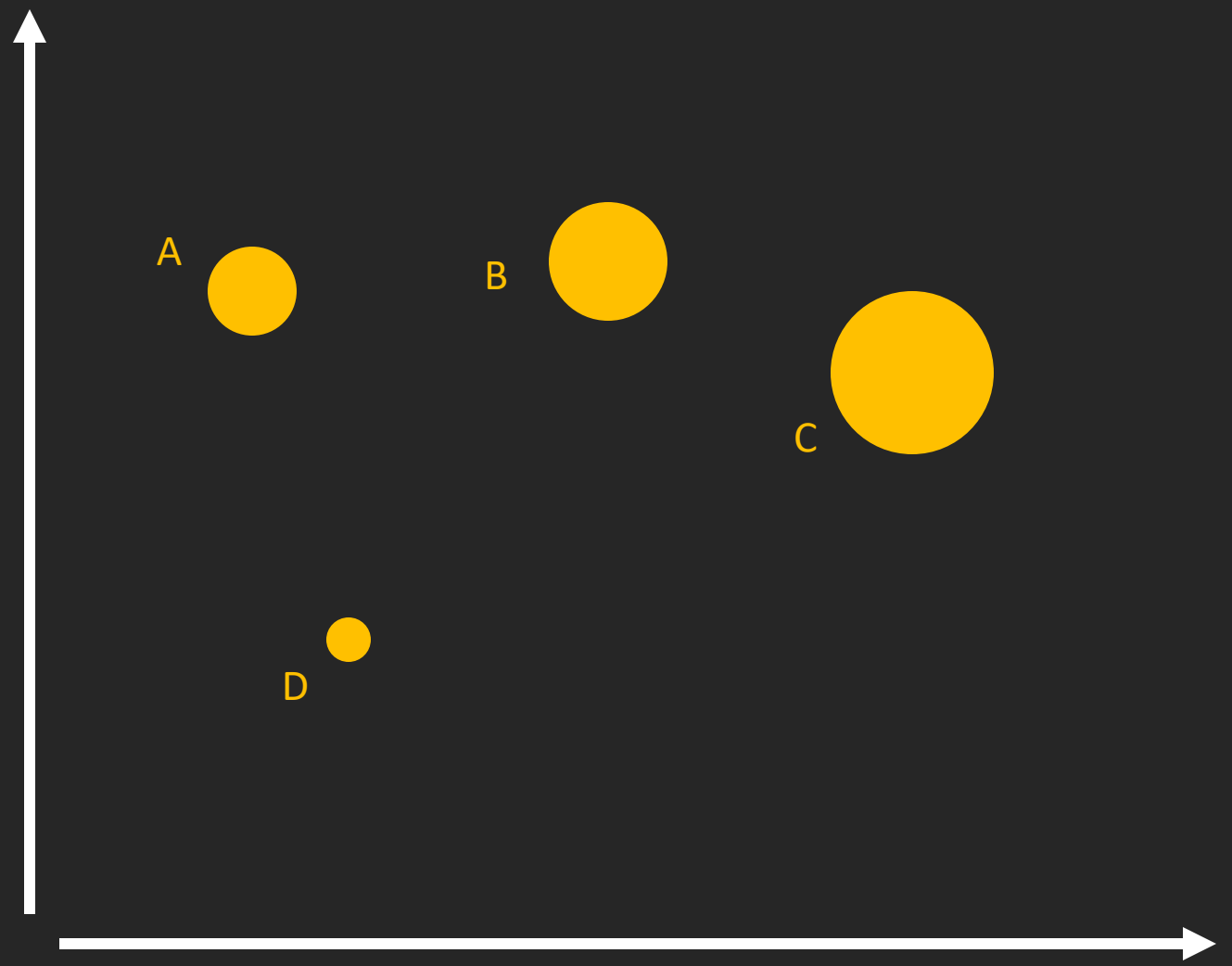
http://www.slido.com
event code **#C674**

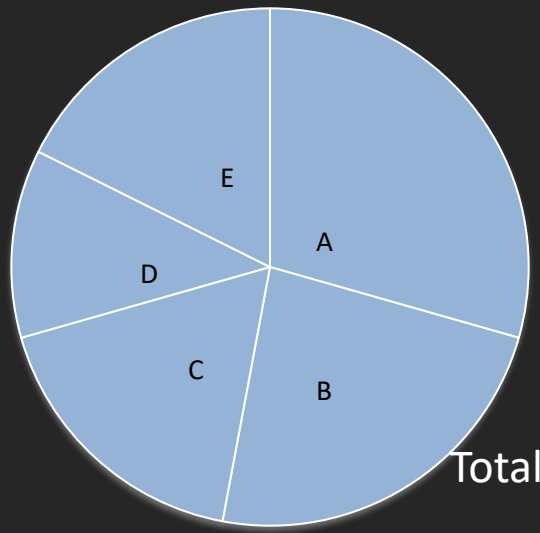
When do we not skip?



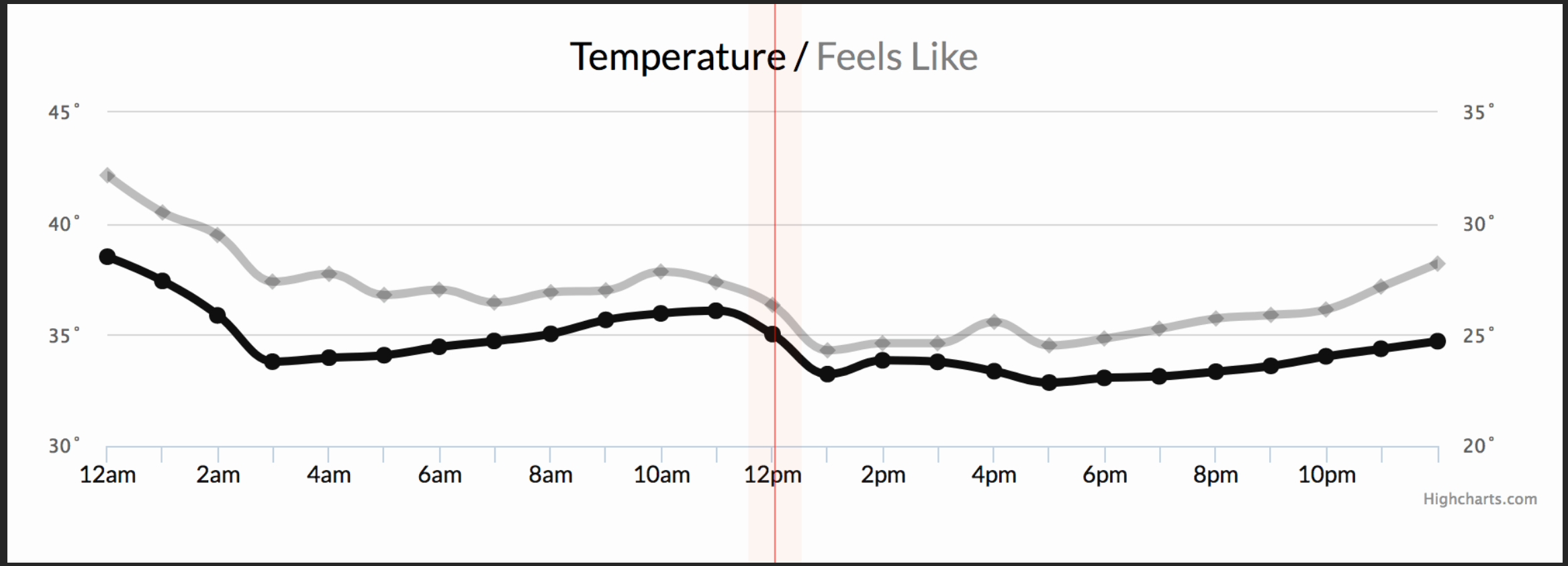
Nominal

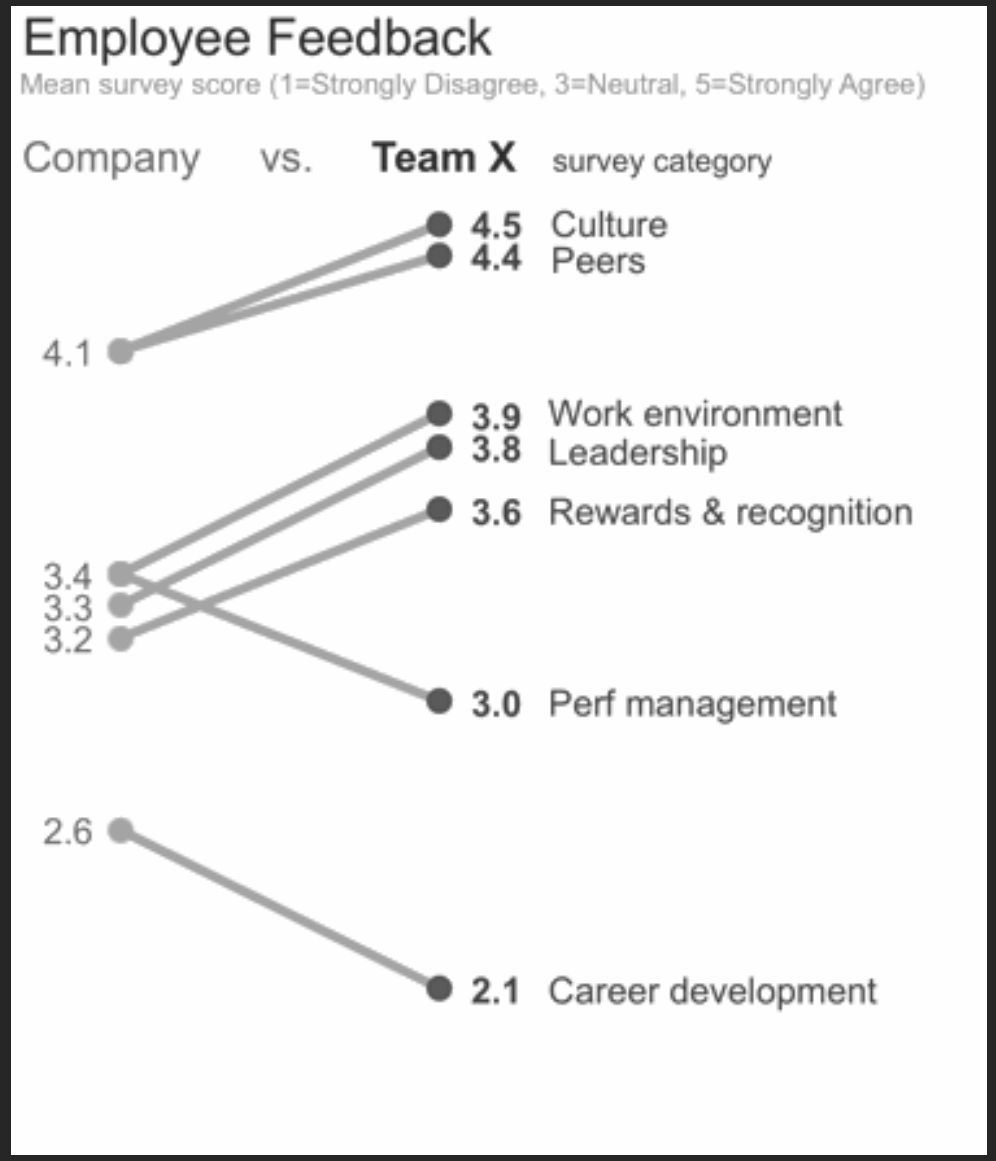
Position
Color Hue
Texture
Connection
Containment
Density
Color Saturation
Shape
Length
Angle
Slope
Area
Volume

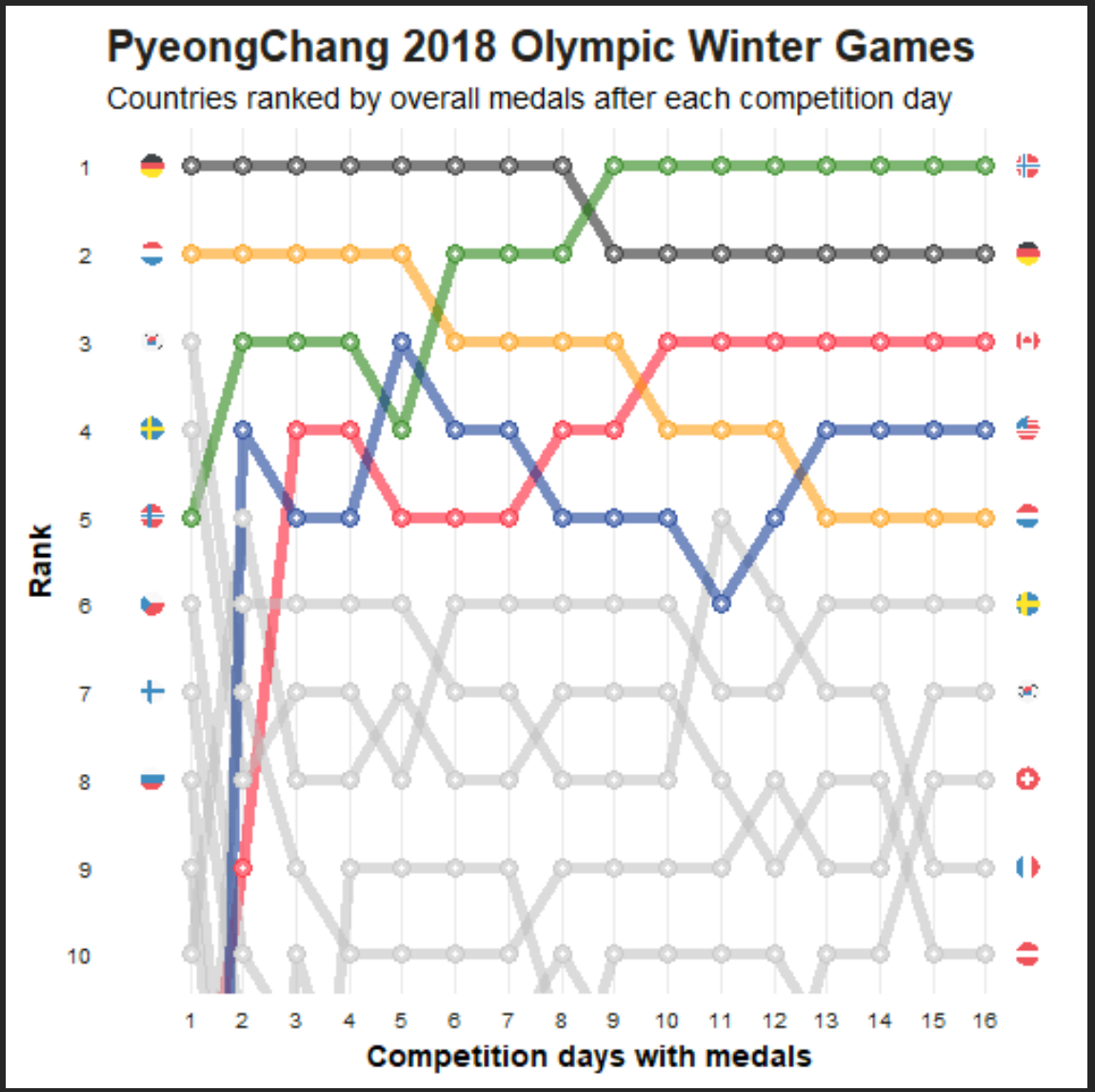

A
B
C
D



Total size: 17


Temperature / Feels Like
45°
40°
35°
30°
35°
30°
25°
20°
12am
2am
4am
6am
8am
10am
12pm
2pm
4pm
6pm
8pm
10pm
Highcharts.com

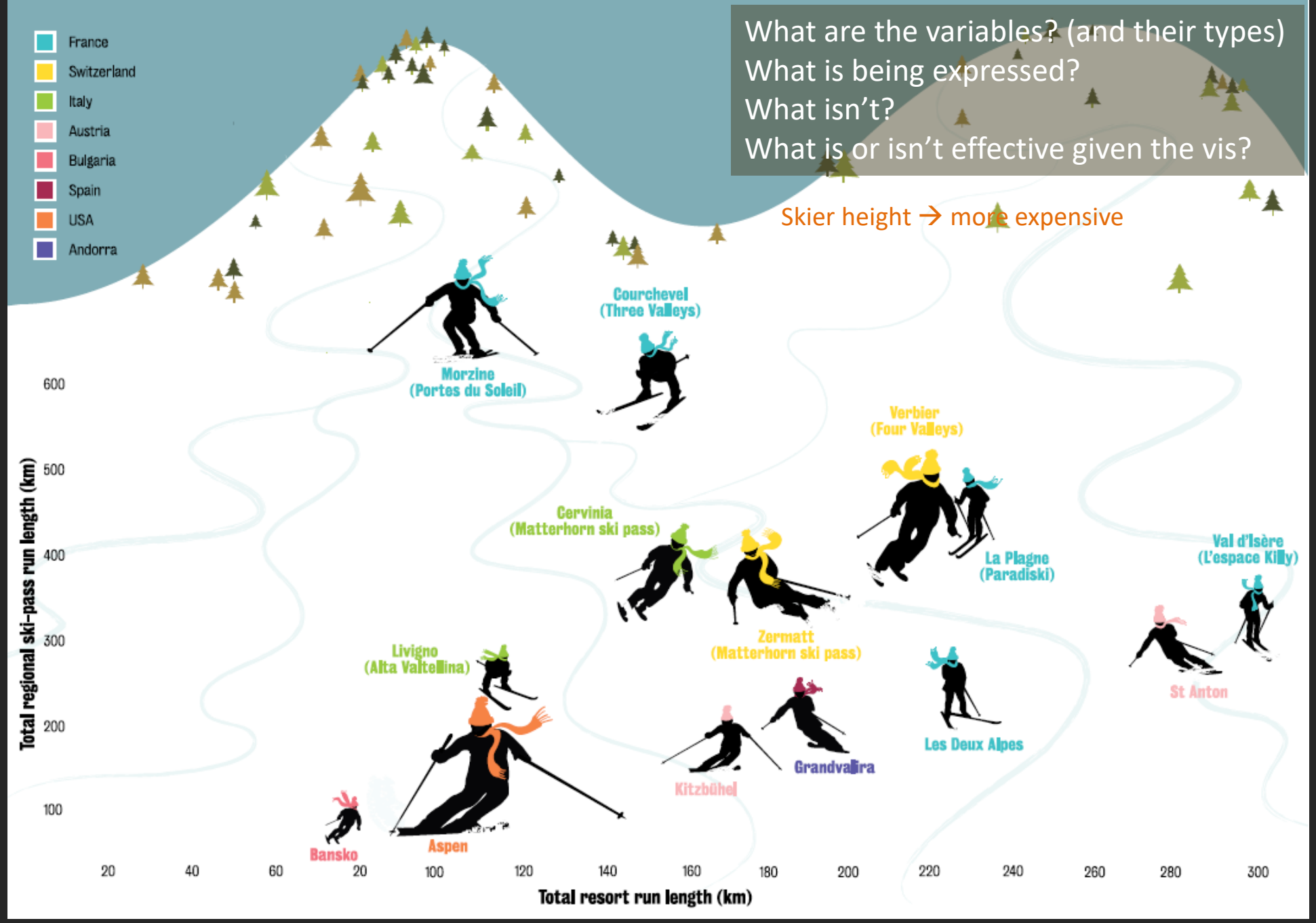

Employee Feedback
Mean survey score (1=Strongly Disagree, 3=Neutral, 5=Strongly Agree)
Company
vs.
Team X
survey category
4.1
4.5 Culture
4.4 Peers
3.4
3.3
3.2
3.9 Work environment
3.8 Leadership
3.6 Rewards & recognition
3.0 Perf management
2.6
2.1 Career development

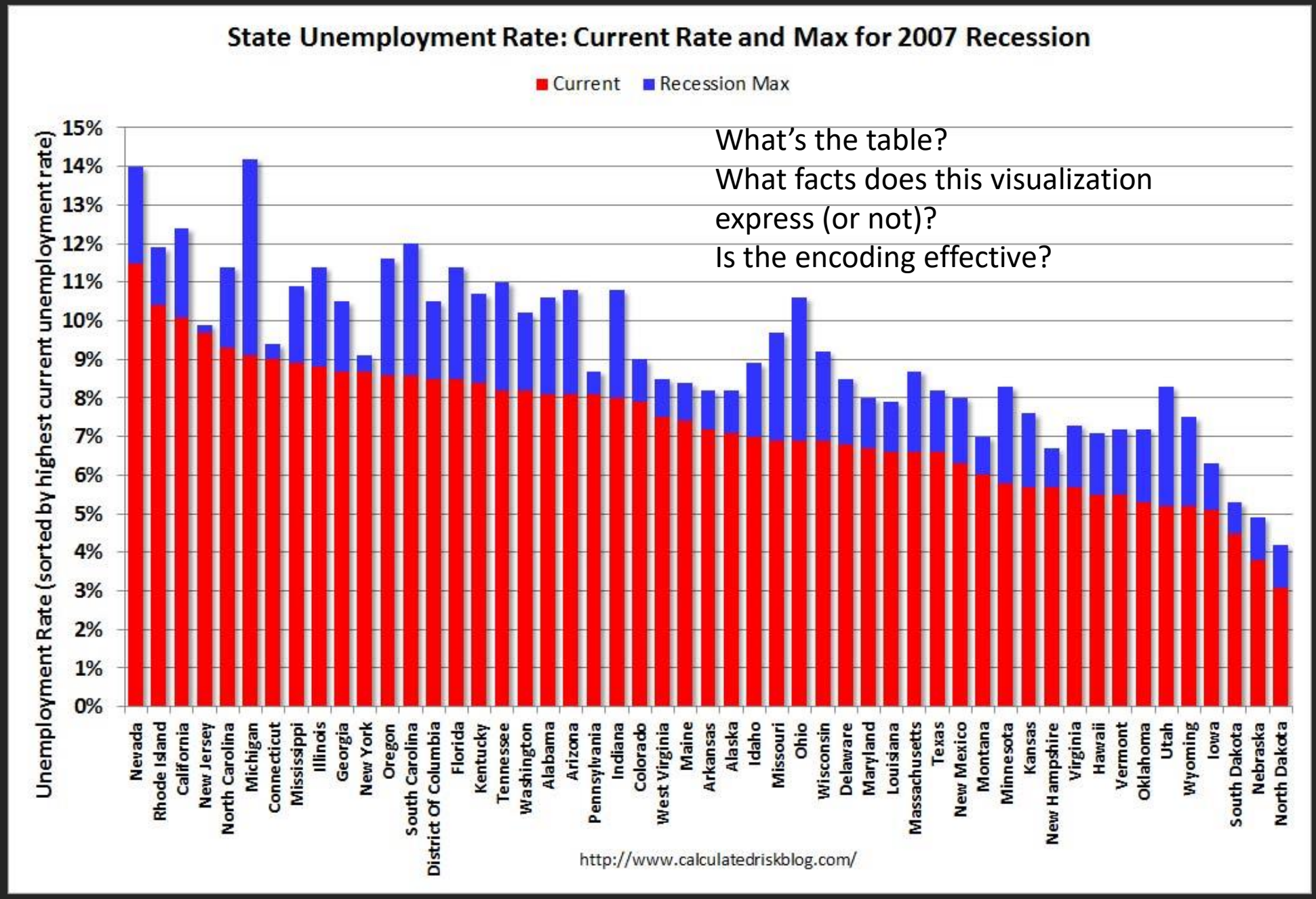

PyeongChang 2018 Olympic Winter Games
Countries ranked by overall medals after each competition day
Rank
1
2
3
4
5
6
7
8
9
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Competition days with medals

Let’s critique some vis

- Look at the visualization
  - Identify the data used (and “type”)
  - Identify what you think they want to express
  - Identify what you think they don’t want to express
  - Determine if the encoding is most effective

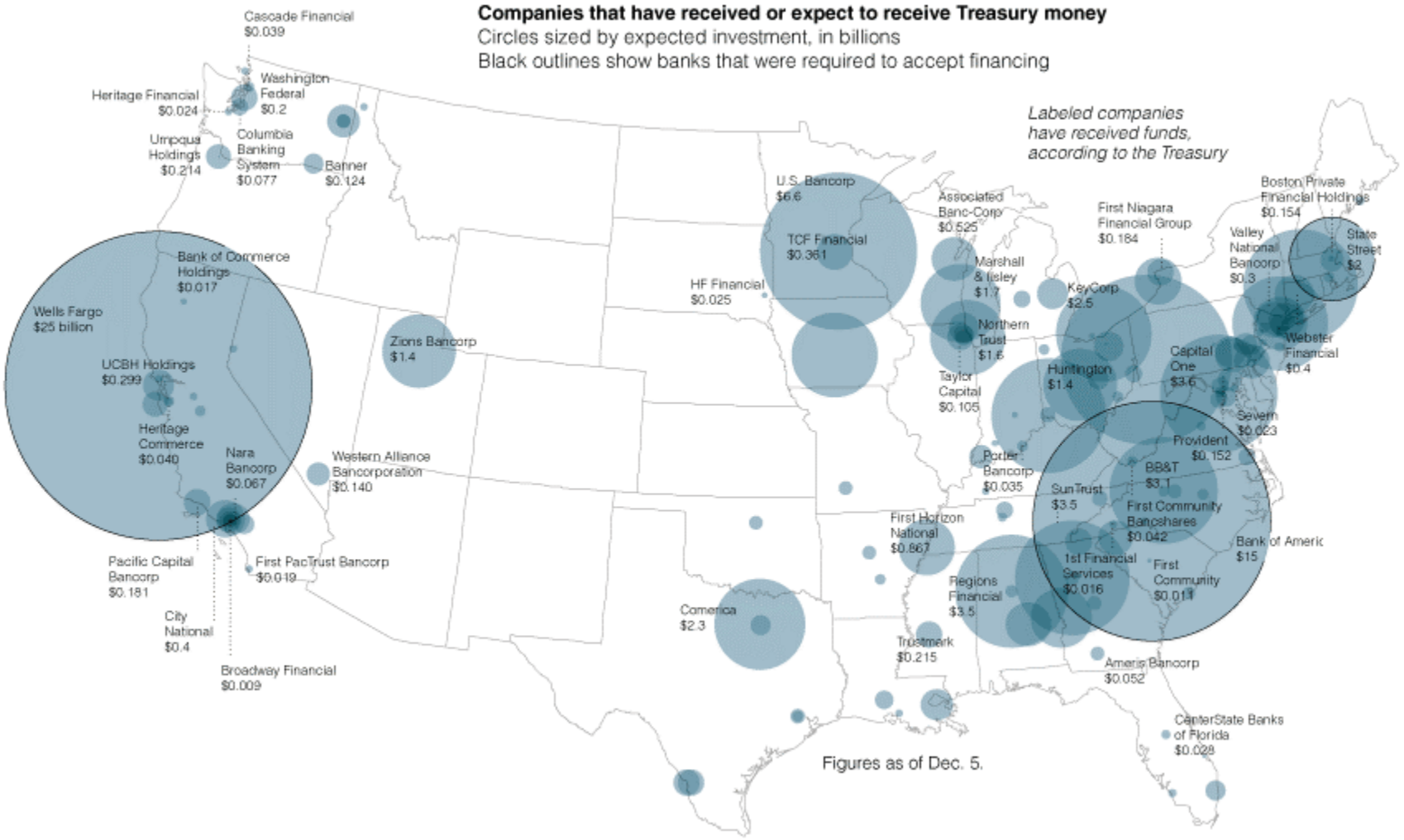
- This is a bit backwards… we’re guessing what they want to express based on their graphical choices (doesn’t mean they made good choices)


What are the variables? (and their types)
What is being expressed?
What isn't?
What is or isn't effective given the vis?
Skier height → more expensive
France
Switzerland
Italy
Austria
Bulgaria
Spain
USA
Andorra
Courchevel (Three Valleys)
Morzine (Portes du Soleil)
Verbier (Four Valleys)
La Plagne (Paradiski)
Cervinia (Matterhorn ski pass)
Zermatt (Matterhorn ski pass)
Val d'Isère (L'espace Killy)
St Anton
Livigno (Alta Valtellina)
Les Deux Alpes
Grandvalira
Kitzbühel
Aspen
Bansko
Total regional ski-pass run length (km)
600
500
400
300
200
100
20
40
60
20
100
120
140
160
180
200
220
240
260
280
300
Total resort run length (km)



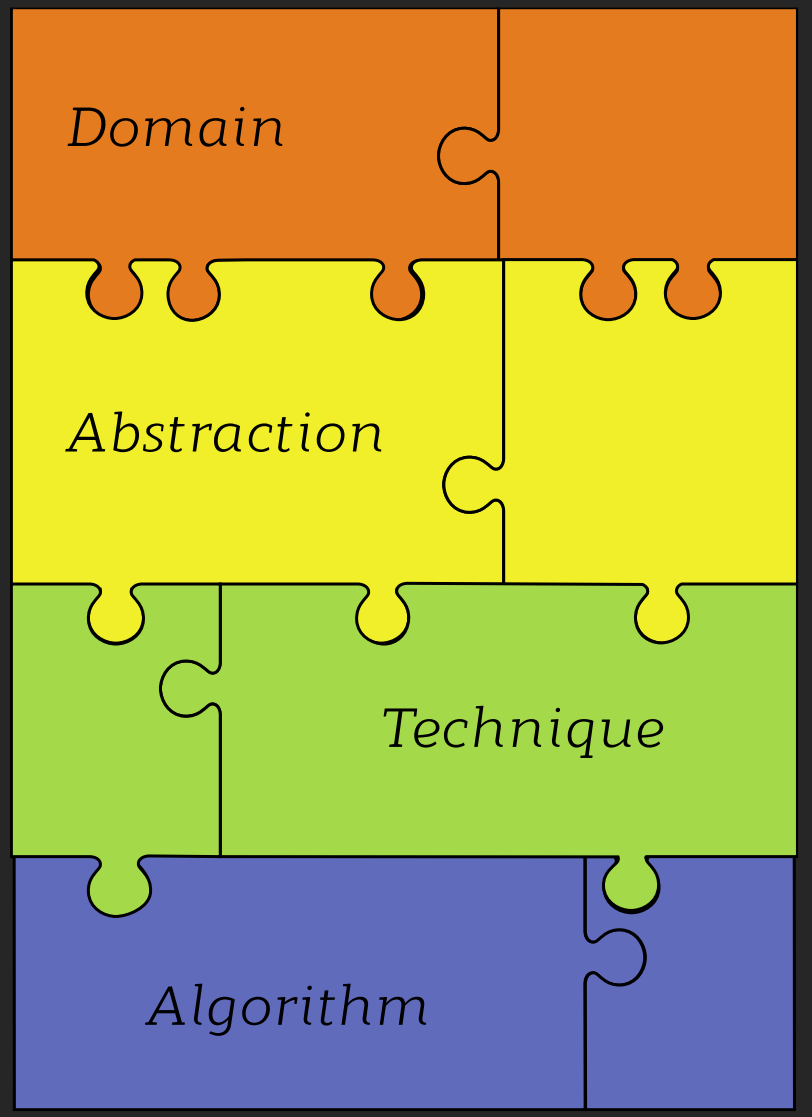
What's the table?
What facts does this visualization express (or not)?
Is the encoding effective?

What's the table?
What facts does this visualization express (or not)?
Is the encoding effective?



You'll get to do this *a lot*

- You'll be repeating this pattern over and over
  - What's the data/use?
  - What do I want to express?
  - What's the most effective way of expressing it?
- But your vocabulary will grow
  - Many more task types
  - Many more encodings/idioms/interaction patterns
  - Many more mechanisms



Meyer et al., 2013

See you soon…