Visualization Analysis & Design Full-Day Tutorial *Session 1*

Tamara Munzner
Department of Computer Science
University of British Columbia

Sanger Institute / European Bioinformatics Institute
June 2014, Cambridge UK

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

Outline

- Visualization Analysis Framework
  Session 1 *9:30-10:45am*
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels
- Idiom Design Choices
  Session 2 *11:00am-12:15pm*
  - Arrange Tables
  - Arrange Spatial Data
  - Arrange Networks and Trees
  - Map Color
- Idiom Design Choices, Part 2
  Session 3 *1:15pm-2:45pm*
  - Manipulate: Change, Select, Navigate
  - Facet: Juxtapose, Partition, Superimpose
  - Reduce: Filter, Aggregate, Embed
- Guidelines and Examples
  Session 4 *3-4:30pm*
  - Rules of Thumb
  - Validation
  - BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

Outline

- **Visualization Analysis Framework**
  Session 1 *9:30-10:45am*
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels
- **Idiom Design Choices**
  Session 2 *11:00am-12:15pm*
  - Arrange Tables
  - Arrange Spatial Data
  - Arrange Networks and Trees
  - Map Color
- **Idiom Design Choices, Part 2**
  Session 3 *1:15pm-2:45pm*
  - Manipulate: Change, Select, Navigate
  - Facet: Juxtapose, Partition, Superimpose
  - Reduce: Filter, Aggregate, Embed
- **Guidelines and Examples**
  Session 4 *3-4:30pm*
  - Rules of Thumb
  - Validation
  - BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

Defining visualization (vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Why?...

Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

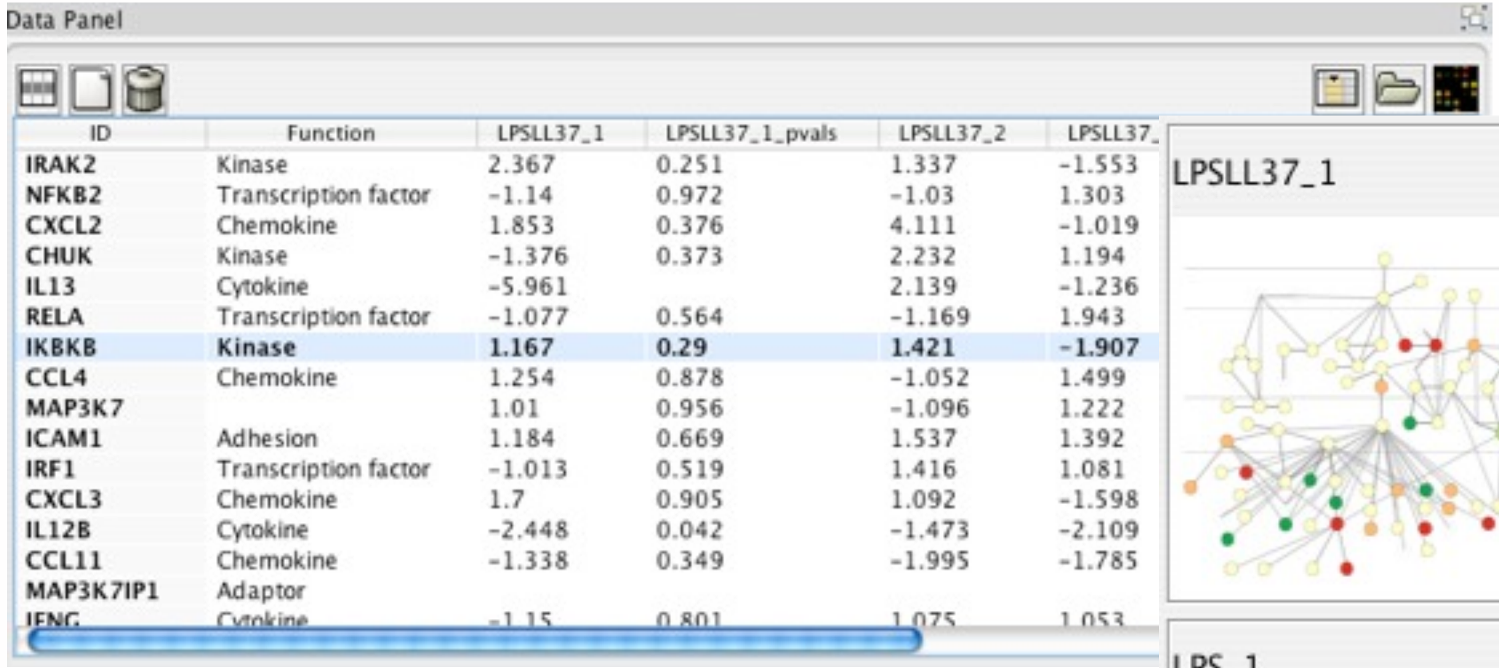
Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

- don't need vis when fully automatic solution exists and is trusted
- many analysis problems ill-specified
  - don't know exactly what questions to ask in advance
- possibilities
  - long-term use for end users (e.g. exploratory analysis of scientific data)
  - presentation of known results
  - stepping stone to better understanding of requirements before developing models
  - help developers of automatic solution refine/debug, determine parameters
  - help end users of automatic solutions verify, build trust

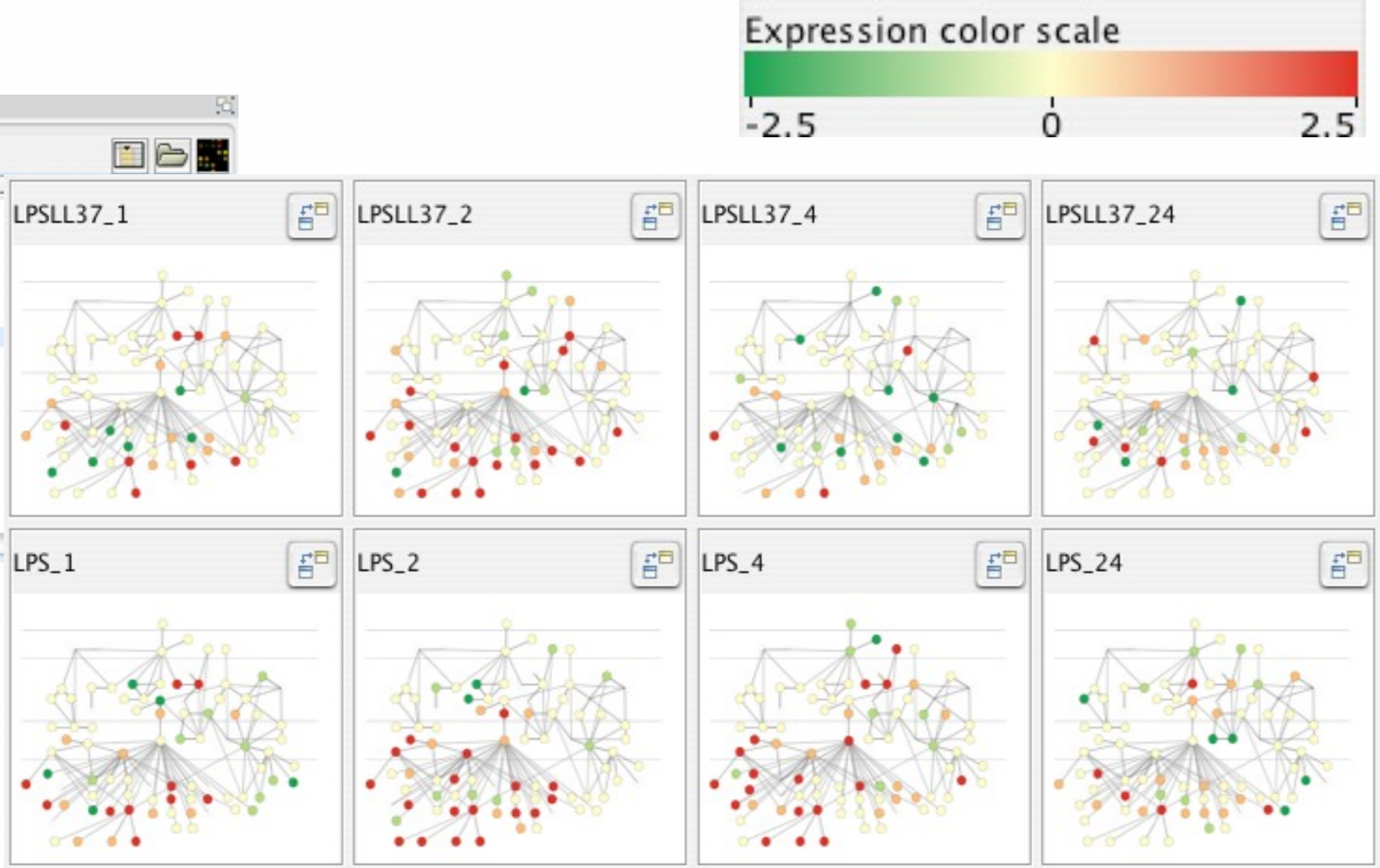
Why use an external representation?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- external representation: replace cognition with perception



ID	Function	LPSLL37_1	LPSLL37_1_pvals	LPSLL37_2	LPSLL37_
IRAK2	Kinase	2.367	0.251	1.337	-1.553
NFKB2	Transcription factor	-1.14	0.972	-1.03	1.303
CXCL2	Chemokine	1.853	0.376	4.111	-1.019
CHUK	Kinase	-1.376	0.373	2.232	1.194
IL13	Cytokine	-5.961		2.139	-1.236
RELA	Transcription factor	-1.077	0.564	-1.169	1.943
IKBKB	Kinase	1.167	0.29	1.421	-1.907
CCL4	Chemokine	1.254	0.878	-1.052	1.499
MAP3K7		1.01	0.956	-1.096	1.222
ICAM1	Adhesion	1.184	0.669	1.537	1.392
IRF1	Transcription factor	-1.013	0.519	1.416	1.081
CXCL3	Chemokine	1.7	0.905	1.092	-1.598
IL12B	Cytokine	-2.448	0.042	-1.473	-2.109
CCL11	Chemokine	-1.338	0.349	-1.995	-1.785
MAP3K7IP1	Adaptor				
IFNG	Cytokine	-1.15	0.801	1.075	1.053

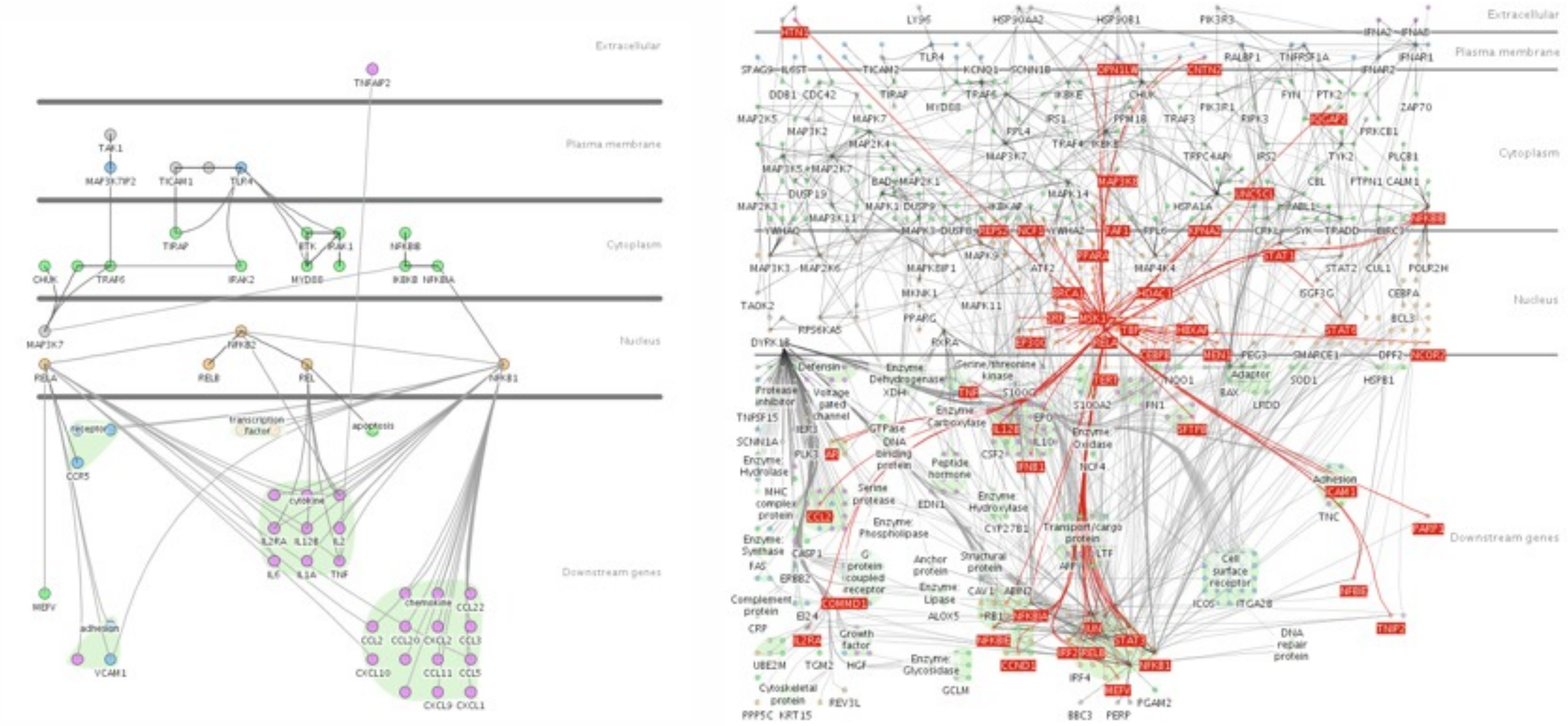


[Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky, Munzner, Gardy, and Kincaid. IEEE TVCG (Proc. InfoVis) 14(6):1253-1260, 2008.]

Why have a computer in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- beyond human patience: scale to large datasets, support interactivity
  - consider: what aspects of hand-drawn diagrams are important?



[Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. Barsky, Gardy, Hancock, and Munzner. Bioinformatics 23(8):1040-1042, 2007.]

Why depend on vision?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

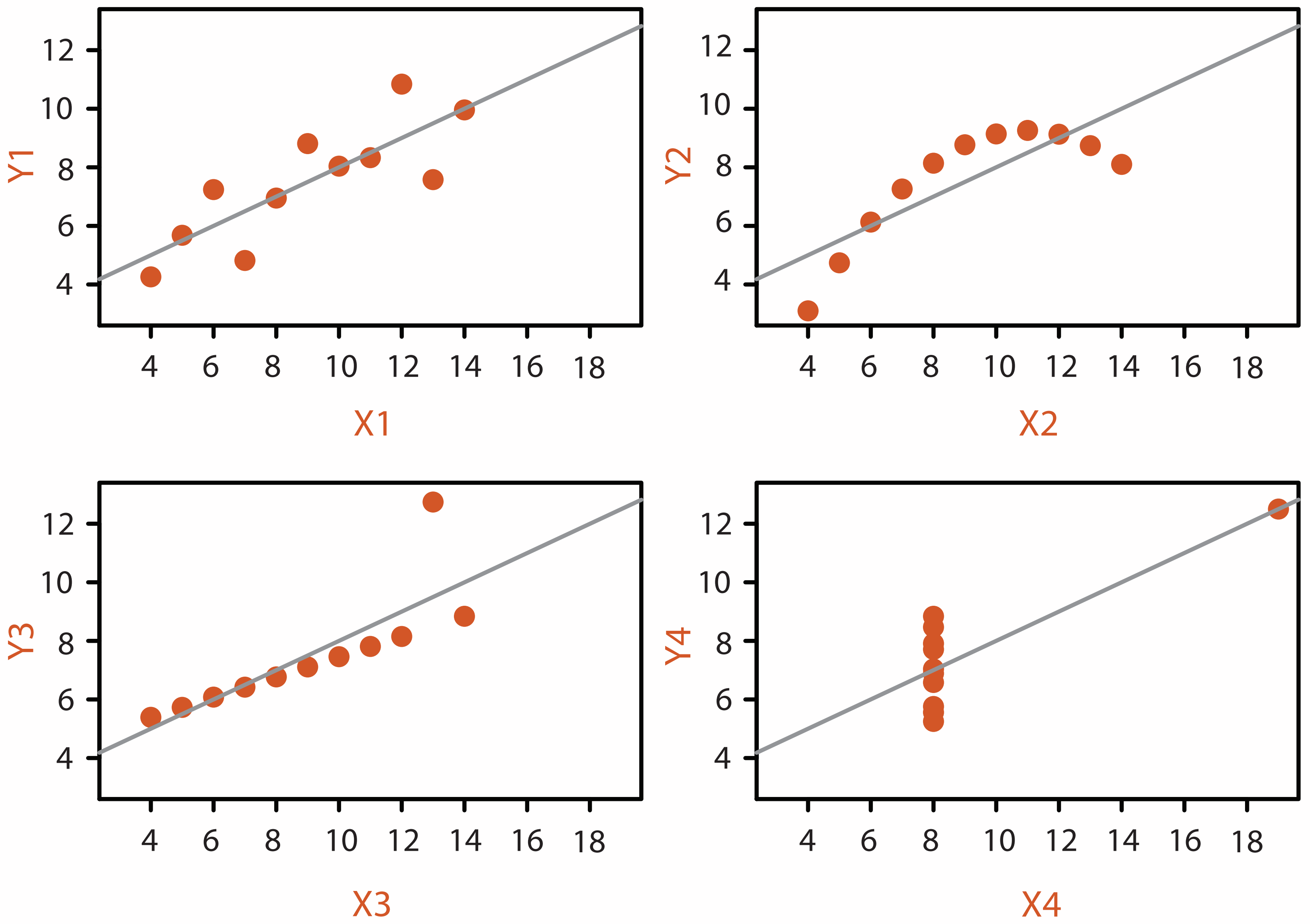
- human visual system is high-bandwidth channel to brain
  - overview possible due to background processing
    - subjective experience of seeing everything simultaneously
    - significant processing occurs in parallel and pre-attentively
- sound: lower bandwidth and different semantics
  - overview not supported
    - subjective experience of sequential stream
- touch/haptics: impoverished record/replay capacity
  - only very low-bandwidth communication thus far
- taste, smell: no viable record/replay devices

Why show the data in detail?

- summaries lose information
  - confirm expected and find unexpected patterns
  - assess validity of statistical model

Anscombe's Quartet

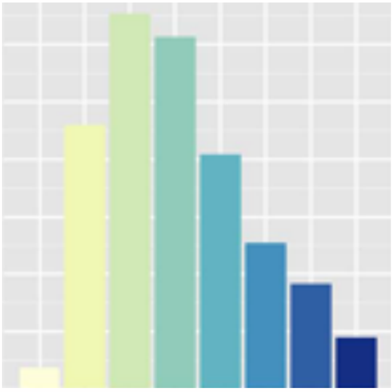
Identical statistics	
x mean	9
x variance	10
y mean	8
y variance	4
x/y correlation	1



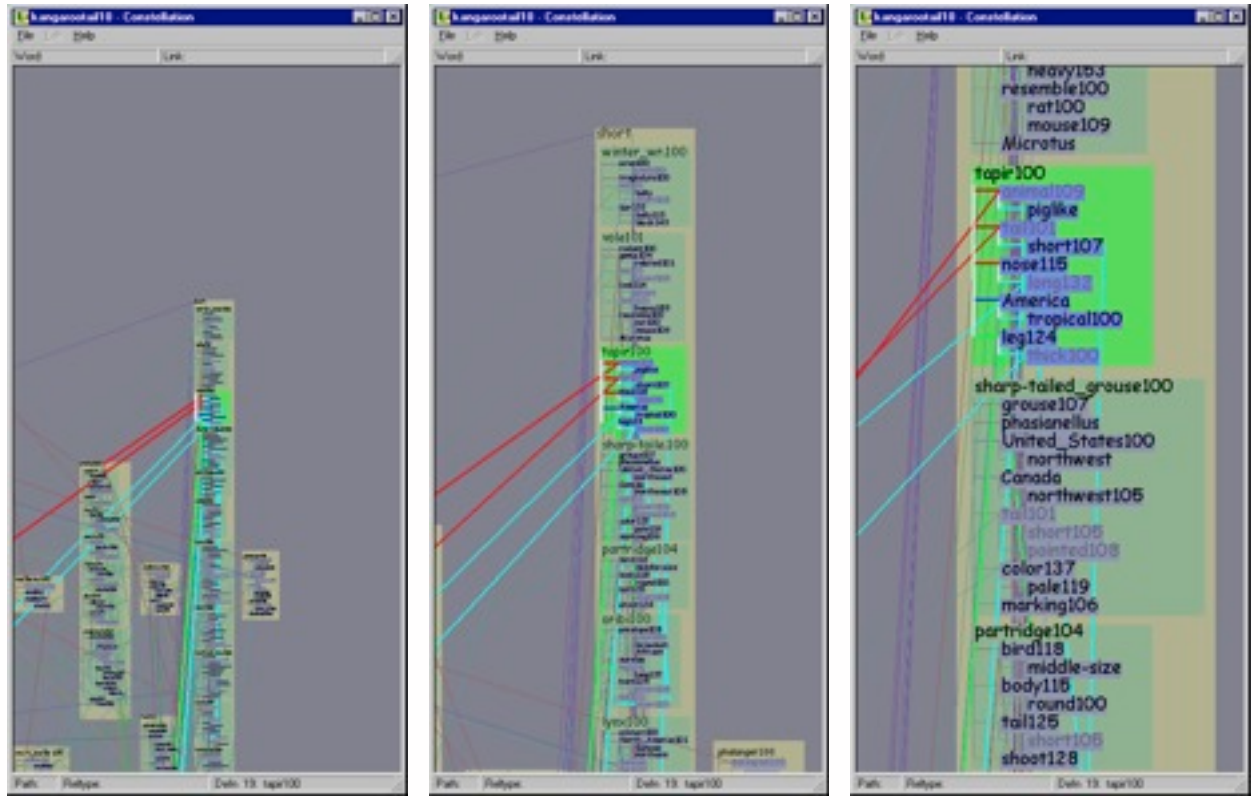
Idiom design space

The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations.

- **idiom**: distinct approach to creating or manipulating visual representation
  - how to draw it: **visual encoding** idiom
    - many possibilities for how to create



  - how to manipulate it: **interaction** idiom
    - even more possibilities
      - make single idiom dynamic
      - link multiple idioms together through interaction



[A layered grammar of graphics. Wickham. Journal of Computational and Graphical Statistics 19:1 (2010), 3–28.]

[Interactive Visualization of Large Graphs and Networks. Munzner. Ph.D. thesis, Stanford University Department of Computer Science, 2000.]

Why focus on tasks and effectiveness?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- tasks serve as constraint on design (as does data)
  - idioms do not serve all tasks equally!
  - challenge: recast tasks from domain-specific vocabulary to abstract forms
- most possibilities ineffective
  - validation is necessary, but tricky
  - increases chance of finding good solutions if you understand full space of possibilities
- what counts as effective?
  - novel: enable entirely new kinds of analysis
  - faster: speed up existing workflows

Resource limitations

Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.

- computational limits
  - processing time
  - system memory
- human limits
  - human attention and memory
- display limits
  - pixels are precious resource, the most constrained resource
  - **information density**: ratio of space used to encode info vs unused whitespace
    - tradeoff between clutter and wasting space, find sweet spot between dense and sparse

Further reading

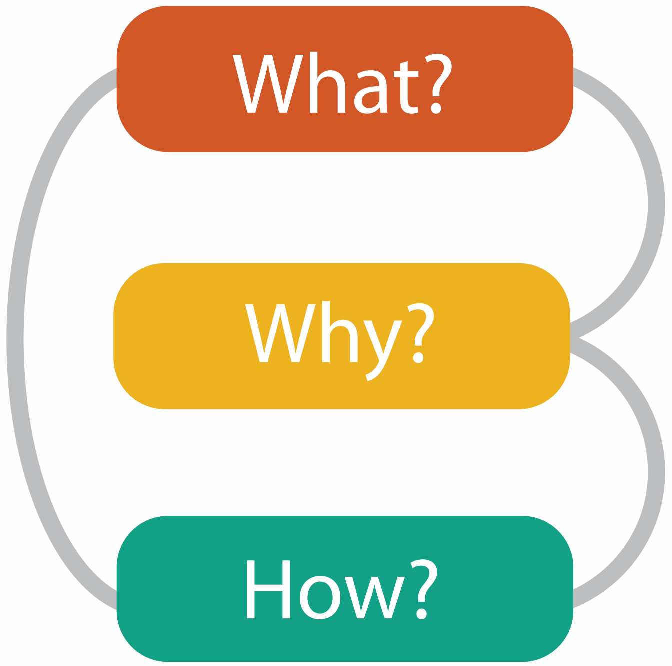
- Visualization Analysis and Design. Munzner. AK Peters / CRC Press, Oct 2014.
  - *Chap 1: What's Vis, and Why Do It?*

Outline

- **Visualization Analysis Framework**
  Session 1 *9:30-10:45am*
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels
- **Idiom Design Choices, Part 2**
  Session 3 *1:15pm-2:45pm*
  - Manipulate: Change, Select, Navigate
  - Facet: Juxtapose, Partition, Superimpose
  - Reduce: Filter, Aggregate, Embed
- **Idiom Design Choices**
  Session 2 *11:00am-12:15pm*
  - Arrange Tables
  - Arrange Spatial Data
  - Arrange Networks and Trees
  - Map Color
- **Guidelines and Examples**
  Session 4 *3-4:30pm*
  - Rules of Thumb
  - Validation
  - BioVis Analysis Example

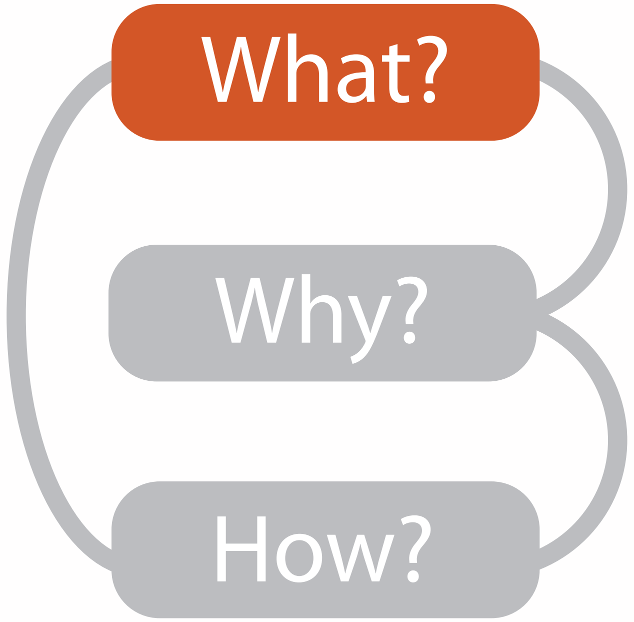
http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

Analysis: What, why, and how



- what is shown?
  - **data** abstraction
- why is the user looking at it?
  - **task** abstraction
- how is it shown?
  - **idiom**: visual encoding and interaction

- abstract vocabulary avoids domain-specific terms
  - translation process iterative, tricky
- what-why-how analysis framework as scaffold to think systematically about design space



What?

Datasets

Data Types

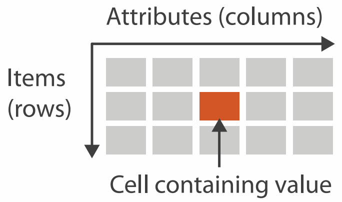
- Items
- Attributes
- Links
- Positions
- Grids

Data and Dataset Types

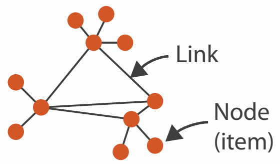
Tables	Networks & Trees	Fields	Geometry	Clusters, sets, lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

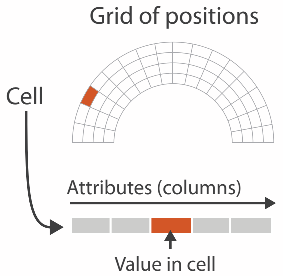
- Tables



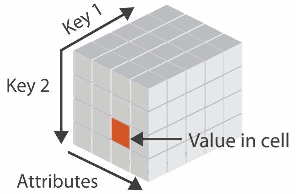
- Networks



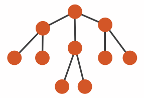
- Fields (Continuous)



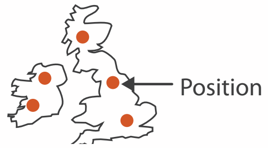
- *Multidimensional Table*



- *Trees*



- Geometry (Spatial)



Dataset Availability

- Static

- Dynamic

Attributes

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

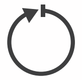
  - *Quantitative*

Ordering Direction

- Sequential

- Diverging

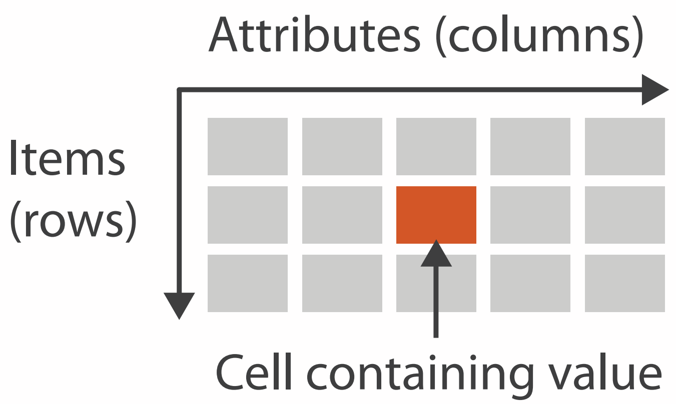
- Cyclic



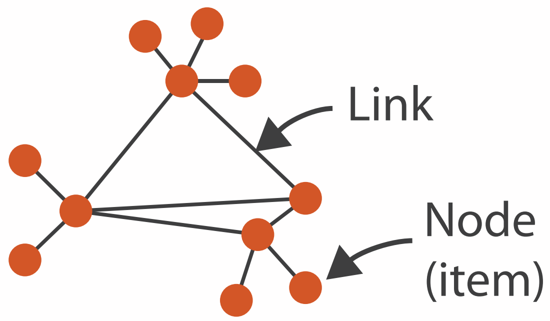
Dataset types

➔ Dataset Types

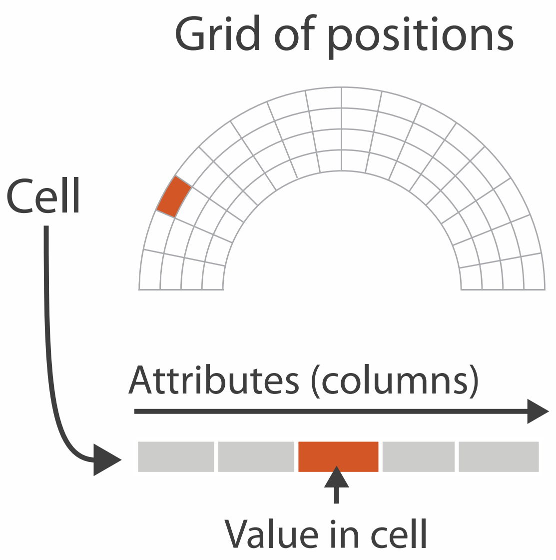
➔ Tables



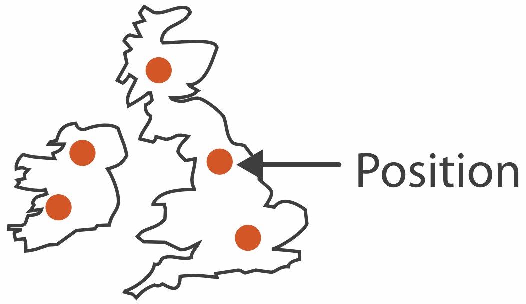
➔ Networks



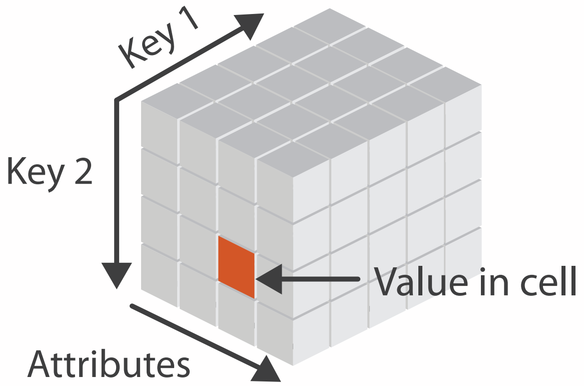
➔ Fields (Continuous)



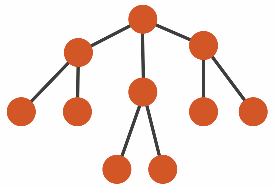
➔ Geometry (Spatial)



➔ *Multidimensional Table*



➔ *Trees*



Dataset and data types

Data and Dataset Types

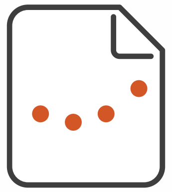
Tables	Networks & Trees	Fields	Geometry	Clusters, sets, lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Data Types

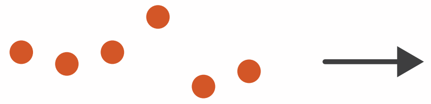
- Items
- Attributes
- Links
- Positions
- Grids

Dataset Availability

- Static



- Dynamic

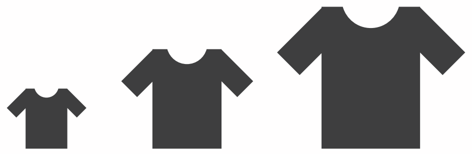


Attribute types

➔ Attribute Types

- ➔ Categorical

- ➔ Ordered
  - ➔ *Ordinal*



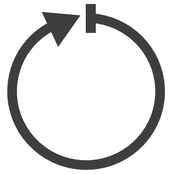
  - ➔ *Quantitative*

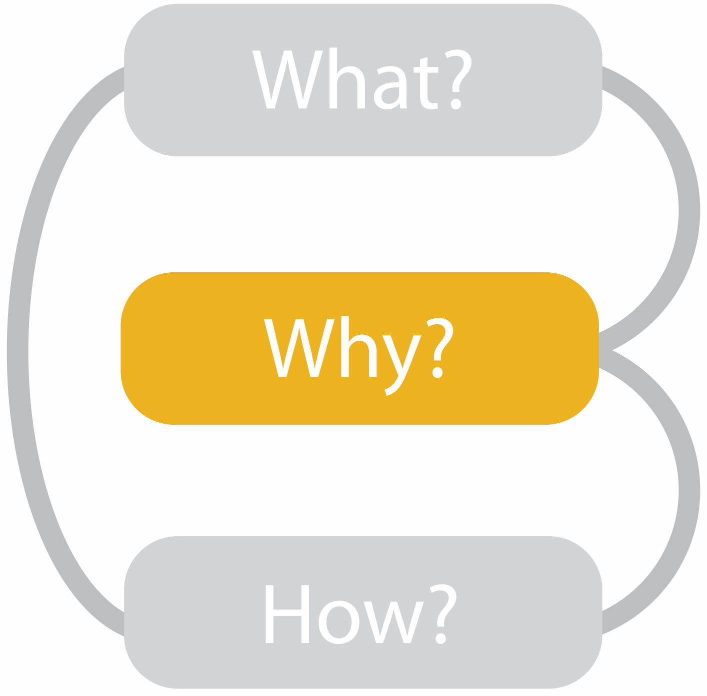
➔ Ordering Direction

- ➔ Sequential

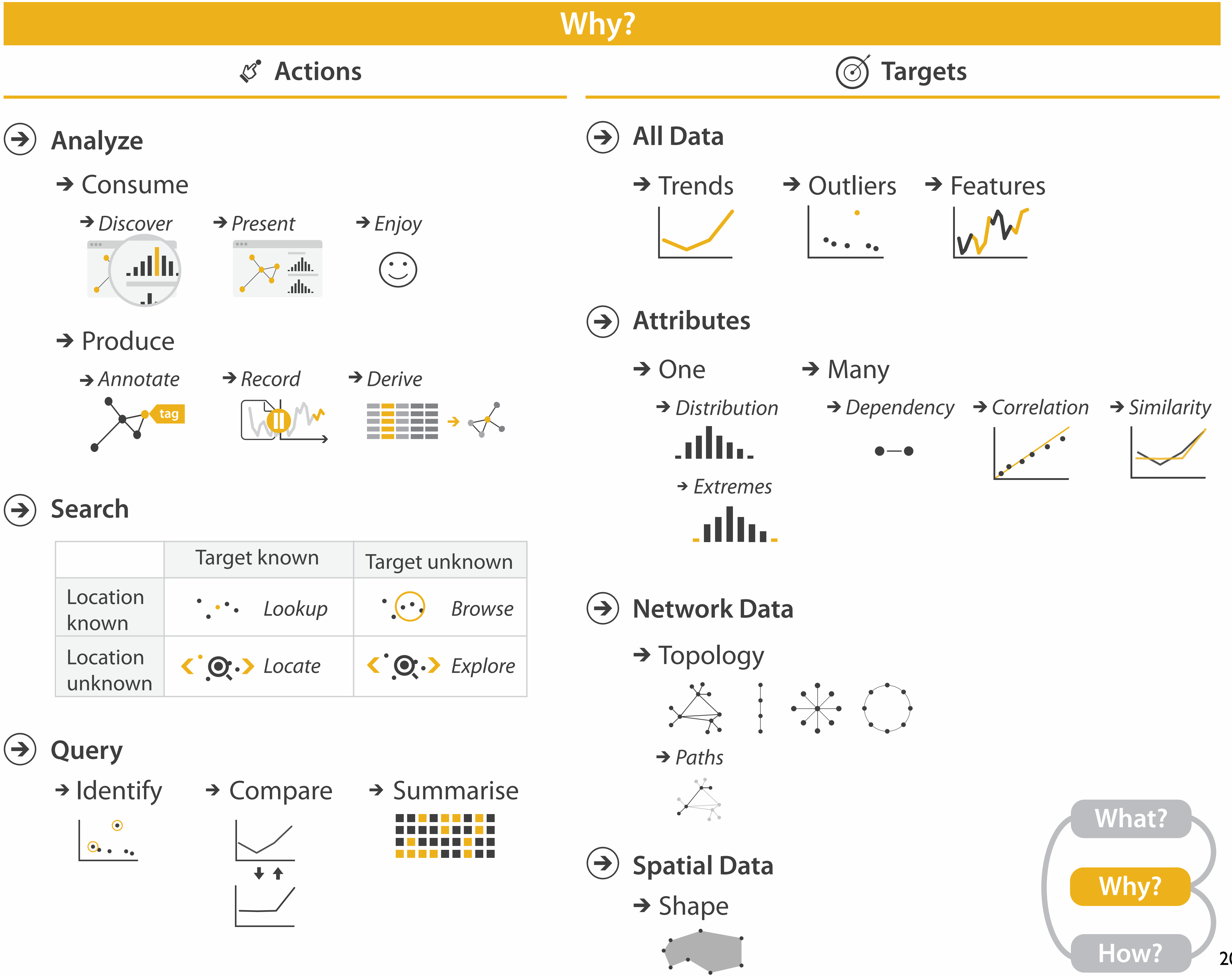
- ➔ Diverging

- ➔ Cyclic



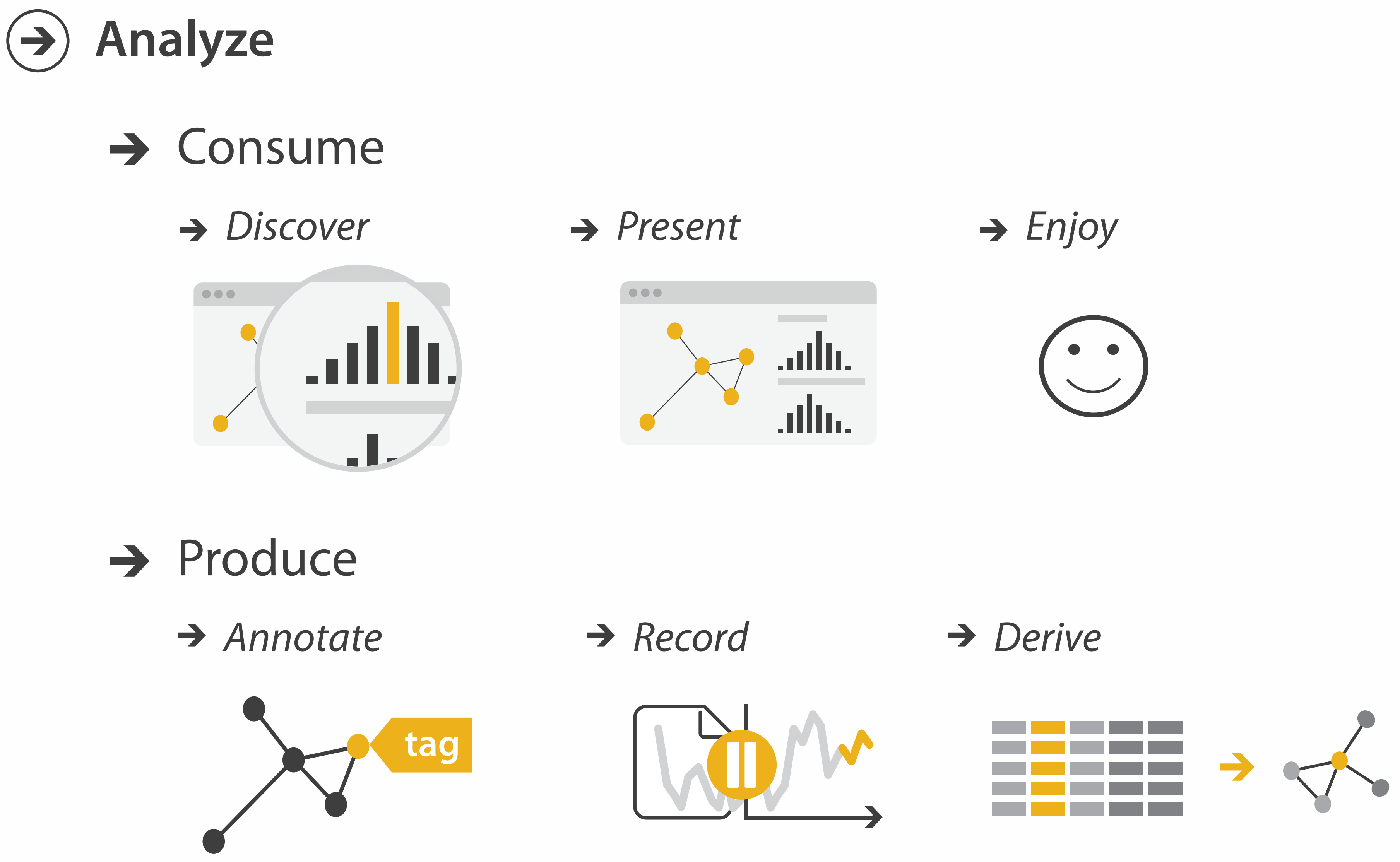


- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*



High-level actions: Analyze

- consume
  - discover vs present
    - classic split
    - aka explore vs explain
  - enjoy
    - newcomer
    - aka casual, social



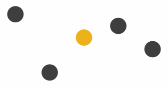
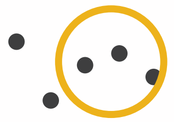
- produce
  - annotate, record
  - derive
    - crucial design choice

Actions: Mid-level search, low-level query

- what does user know?
  - target, location

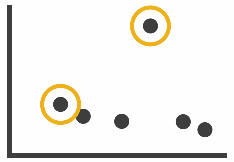
- how much of the data matters?
  - one, some, all

➔ Search

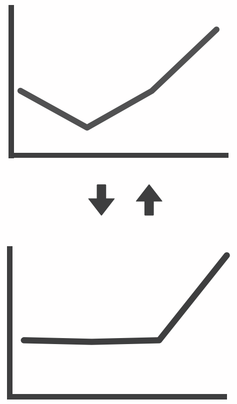
	Target known	Target unknown
Location known	 *Lookup*	 *Browse*
Location unknown	 *Locate*	 *Explore*

➔ Query

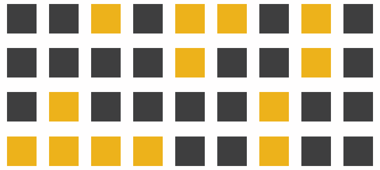
➔ Identify



➔ Compare



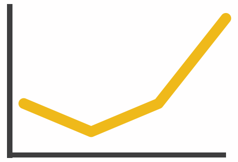
➔ Summarise



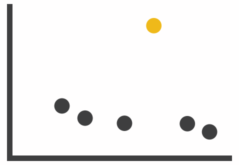
Why: Targets

ALL DATA

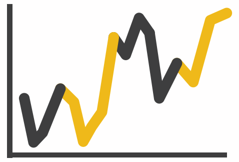
- Trends



- Outliers

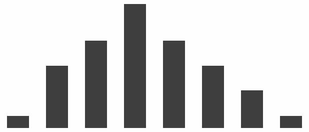


- Features

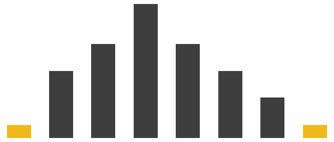


ATTRIBUTES

- One
  - *Distribution*

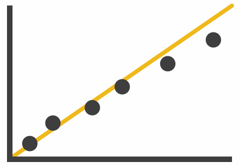


  - *Extremes*

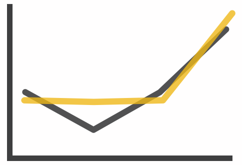


- Many
  - *Dependency*

  - *Correlation*

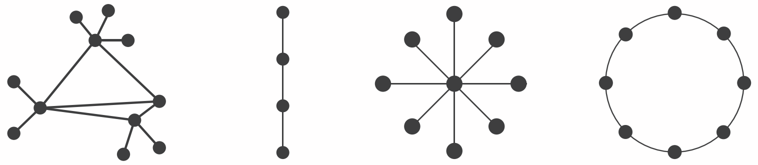


  - *Similarity*

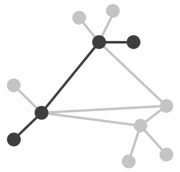


NETWORK DATA

- Topology

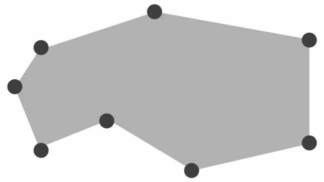


  - *Paths*

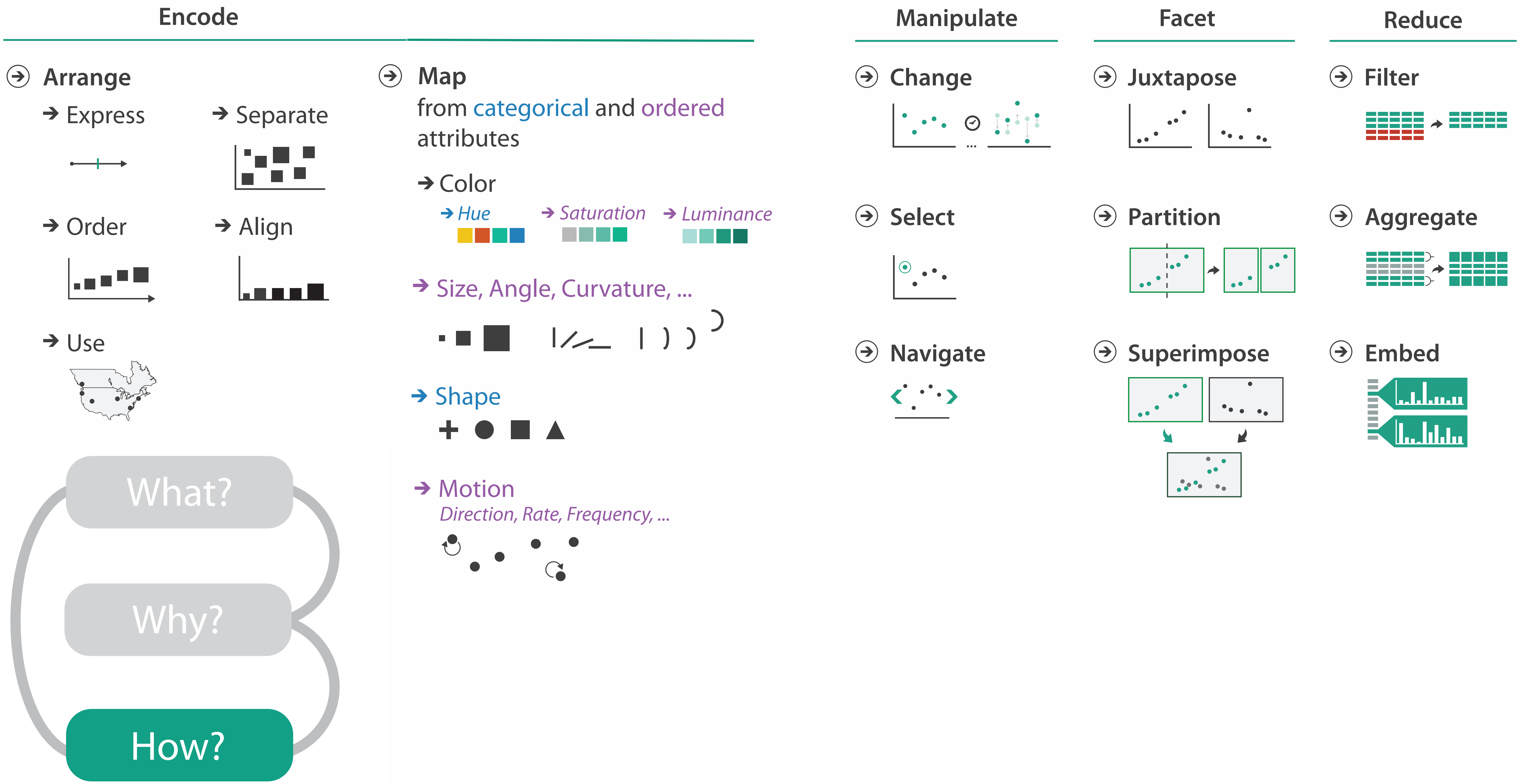


SPATIAL DATA

- Shape

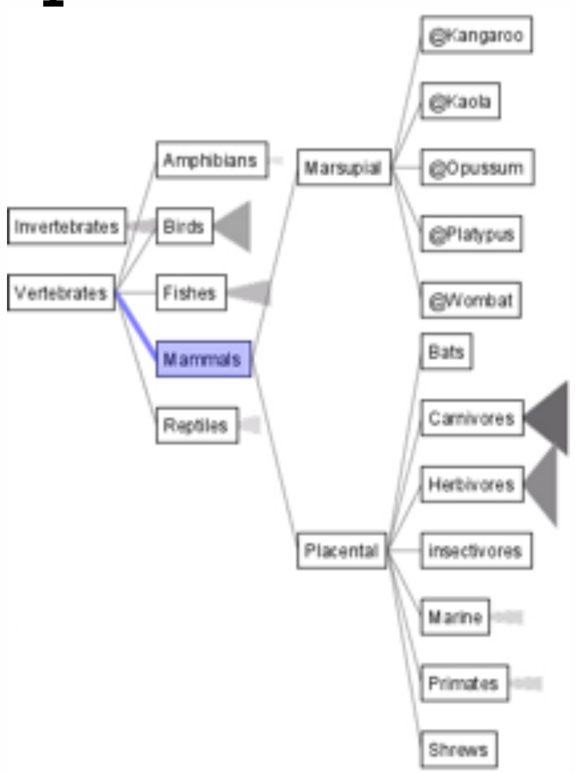


How?

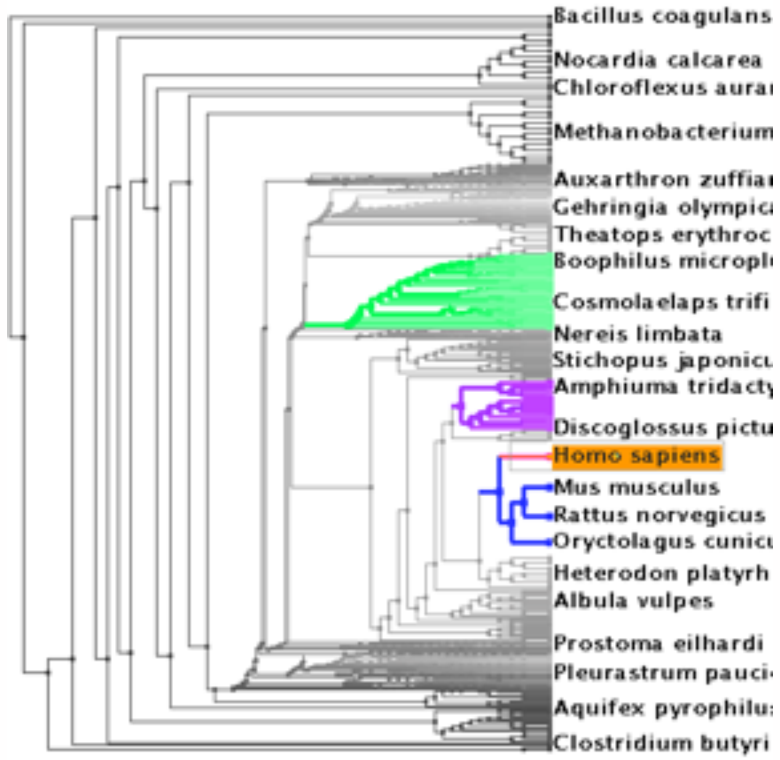

Encode
Manipulate
Facet
Reduce
Arrange
→ Express
→ Separate
→ Order
→ Align
→ Use
Map
from categorical and ordered attributes
→ Color
→ Hue
→ Saturation
→ Luminance
→ Size, Angle, Curvature, ...
→ Shape
→ Motion
Direction, Rate, Frequency, ...
Change
Select
Navigate
Juxtapose
Partition
Superimpose
Filter
Aggregate
Embed
What?
Why?
How?

Analysis example: Compare idioms

SpaceTree

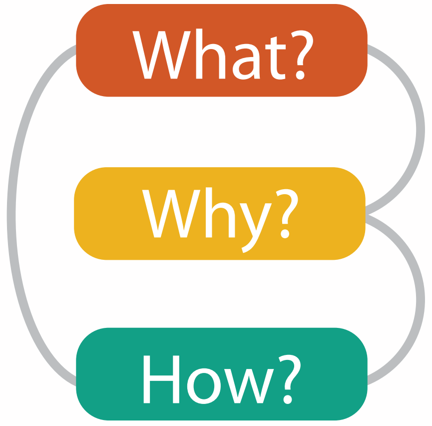


TreeJuxtaposer



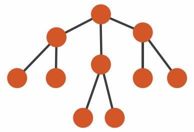
[SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation. Grosjean, Plaisant, and Bederson. Proc. InfoVis 2002, p 57–64.]

[TreeJuxtaposer: Scalable Tree Comparison Using Focus +Context With Guaranteed Visibility. ACM Trans. on Graphics (Proc. SIGGRAPH) 22:453– 462, 2003.]



What?

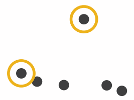
➔ **Tree**



Why?

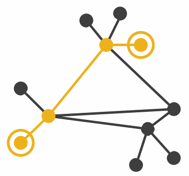
➔ **Actions**

➔ Present ➔ Locate ➔ Identify

➔ **Targets**

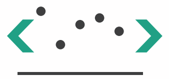
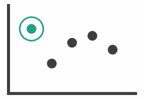
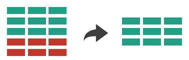
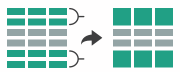
➔ Path between two nodes



How?

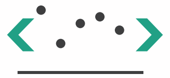
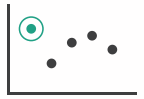
➔ **SpaceTree**

➔ Encode ➔ Navigate ➔ Select ➔ Filter ➔ Aggregate

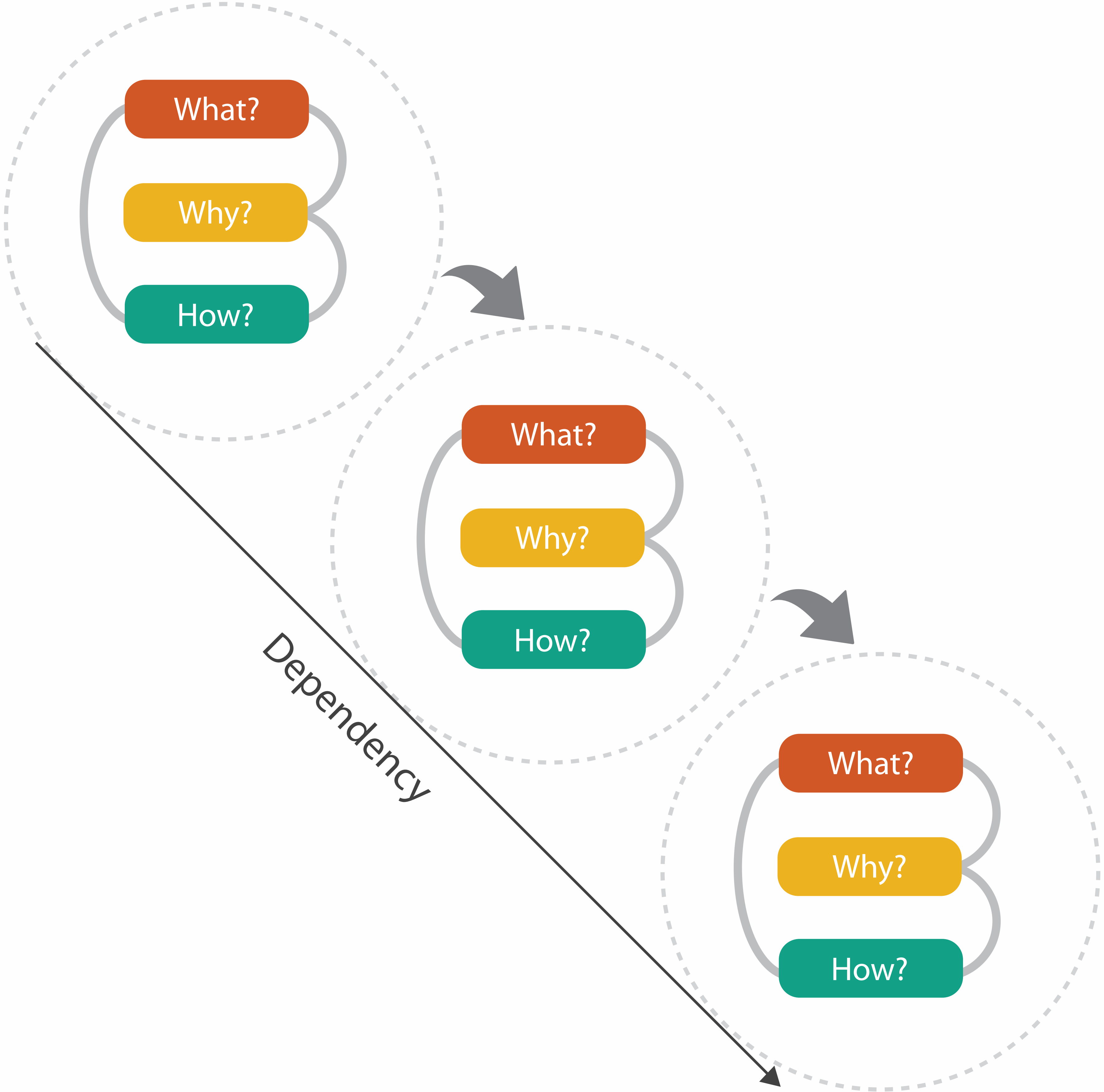
➔ **TreeJuxtaposer**

➔ Encode ➔ Navigate ➔ Select ➔ Arrange

Chained sequences

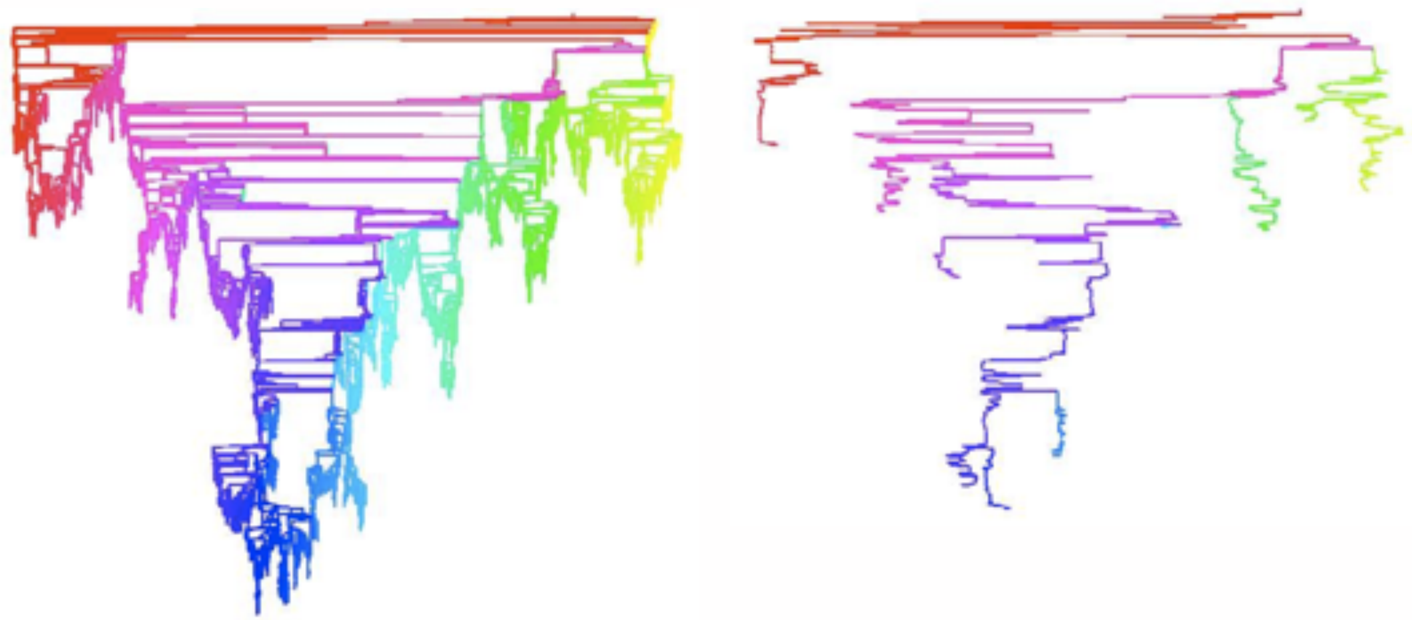
- output of one is input to next
  - express dependencies
  - separate means from ends



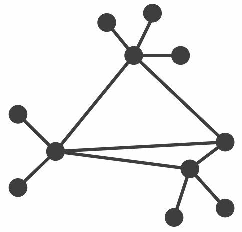
Analysis example: Derive one attribute

- Strahler number
  - centrality metric for trees/networks
  - derived quantitative attribute
  - draw top 5K of 500K for good skeleton

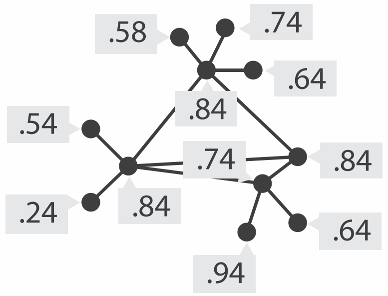
[Using Strahler numbers for real time visual exploration of huge graphs. Auber. Proc. Intl. Conf. Computer Vision and Graphics, pp. 56–69, 2002.]



Task 1



In
Tree

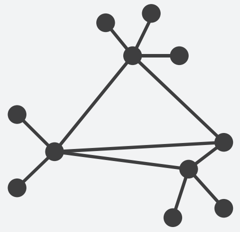


Out
Quantitative attribute on nodes

What?
- In Tree
- Out Quantitative attribute on nodes

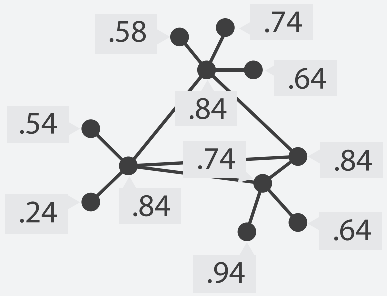
Why?
- Derive

Task 2

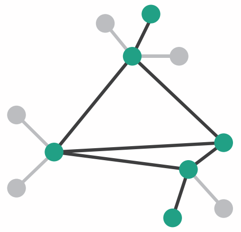


In
Tree

+



In
Quantitative attribute on nodes



Out
Filtered Tree
Removed unimportant parts

What?
- In Tree
- In Quantitative attribute on nodes
- Out Filtered Tree

Why?
- Summarize
- Topology

How?
- Reduce
- Filter

Further reading

- Visualization Analysis and Design. Munzner. AK Peters / CRC Press, Oct 2014.
  - *Chap 2: What: Data Abstraction*
  - *Chap 3: Why: Task Abstraction*
- *A Multi-Level Typology of Abstract Visualization Tasks*. Brehmer and Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 19:12 (2013), 2376–2385.
- *Low-Level Components of Analytic Activity in Information Visualization*. Amar, Eagan, and Stasko. Proc. IEEE InfoVis 2005, p 111–117.
- *A taxonomy of tools that support the fluent and flexible use of visualizations*. Heer and Shneiderman. Communications of the ACM 55:4 (2012), 45–54.
- *Rethinking Visualization: A High-Level Taxonomy*. Tory and Möller. Proc. IEEE InfoVis 2004, p 151–158.
- Visualization of Time-Oriented Data. Aigner, Miksch, Schumann, and Tominski. Springer, 2011.

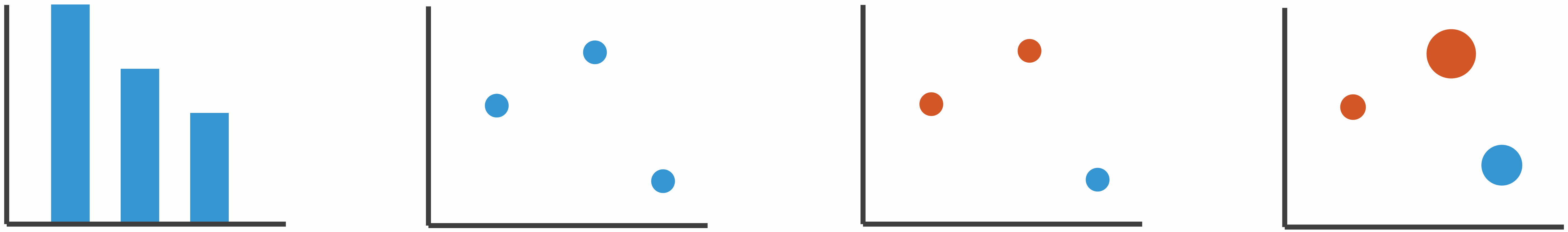
Outline

- **Visualization Analysis Framework**
  Session 1 *9:30-10:45am*
  - Introduction: Definitions
  - Analysis: What, Why, How
  - Marks and Channels
- **Idiom Design Choices, Part 2**
  Session 3 *1:15pm-2:45pm*
  - Manipulate: Change, Select, Navigate
  - Facet: Juxtapose, Partition, Superimpose
  - Reduce: Filter, Aggregate, Embed
- **Idiom Design Choices**
  Session 2 *11:00am-12:15pm*
  - Arrange Tables
  - Arrange Spatial Data
  - Arrange Networks and Trees
  - Map Color
- **Guidelines and Examples**
  Session 4 *3-4:30pm*
  - Rules of Thumb
  - Validation
  - BioVis Analysis Example

http://www.cs.ubc.ca/~tmm/talks.html#minicourse14

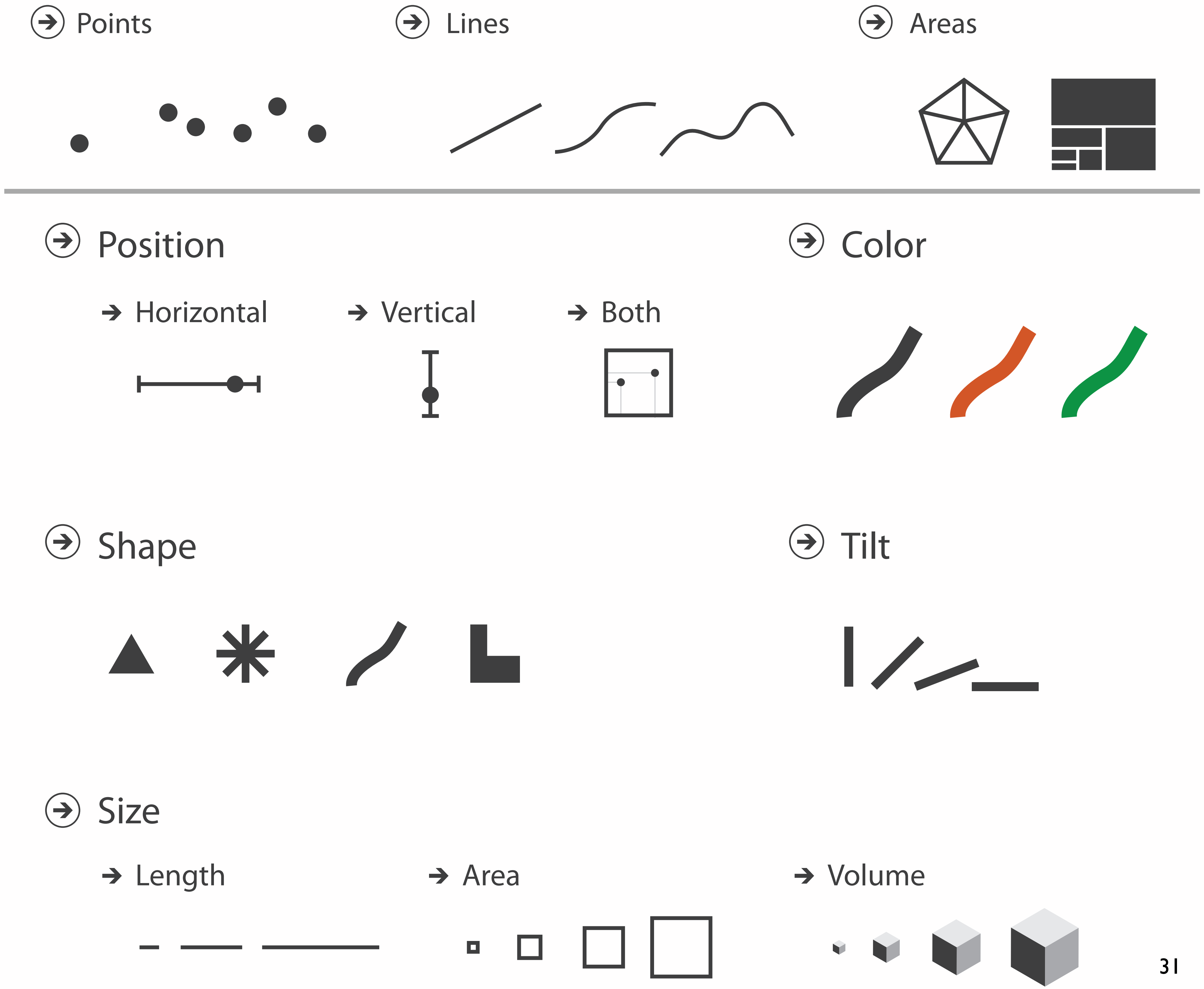
Visual encoding

- analyze idiom structure



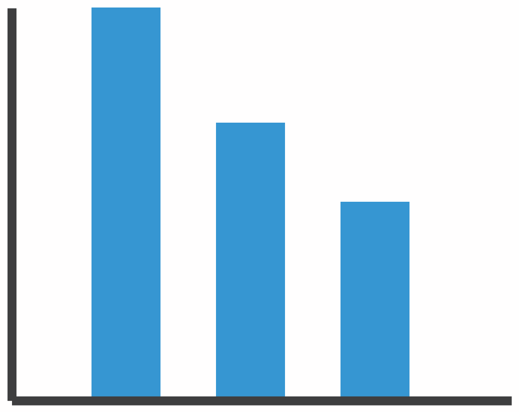
Definitions: Marks and channels

- marks
  - geometric primitives
- channels
  - control appearance of marks
  - can redundantly code with multiple channels
- interactions
  - point marks only convey position; no area constraints
    - can be size and shape coded
  - line marks convey position and length
    - can only be size coded in 1D (width)
  - area marks fully constrained
    - cannot be size or shape coded



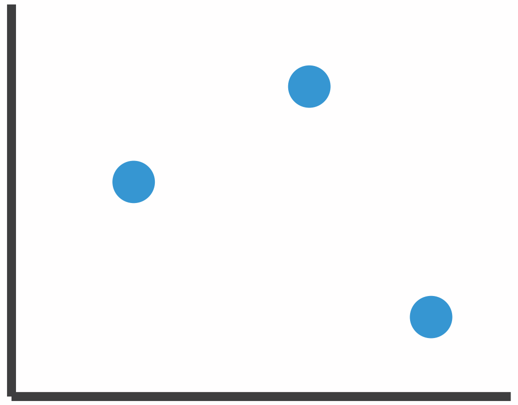
Visual encoding

- analyze idiom structure
  - as combination of marks and channels



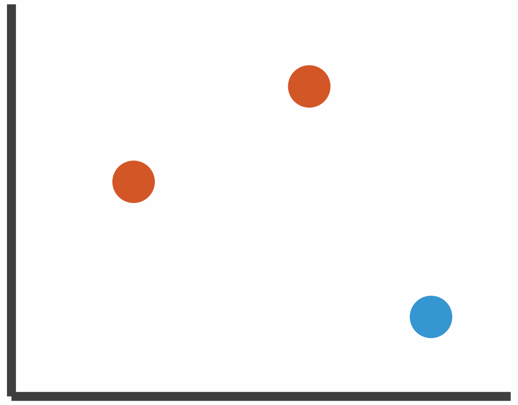
1:
vertical position

mark: line



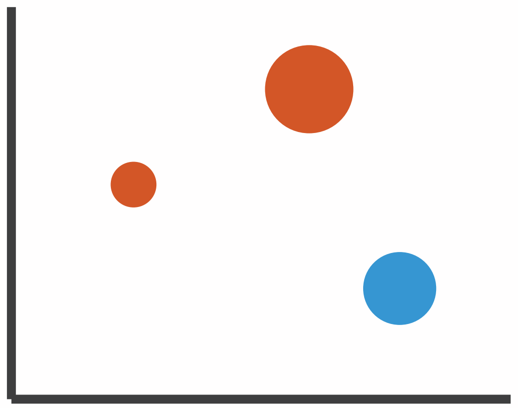
2:
vertical position
horizontal position

mark: point



3:
vertical position
horizontal position
color hue

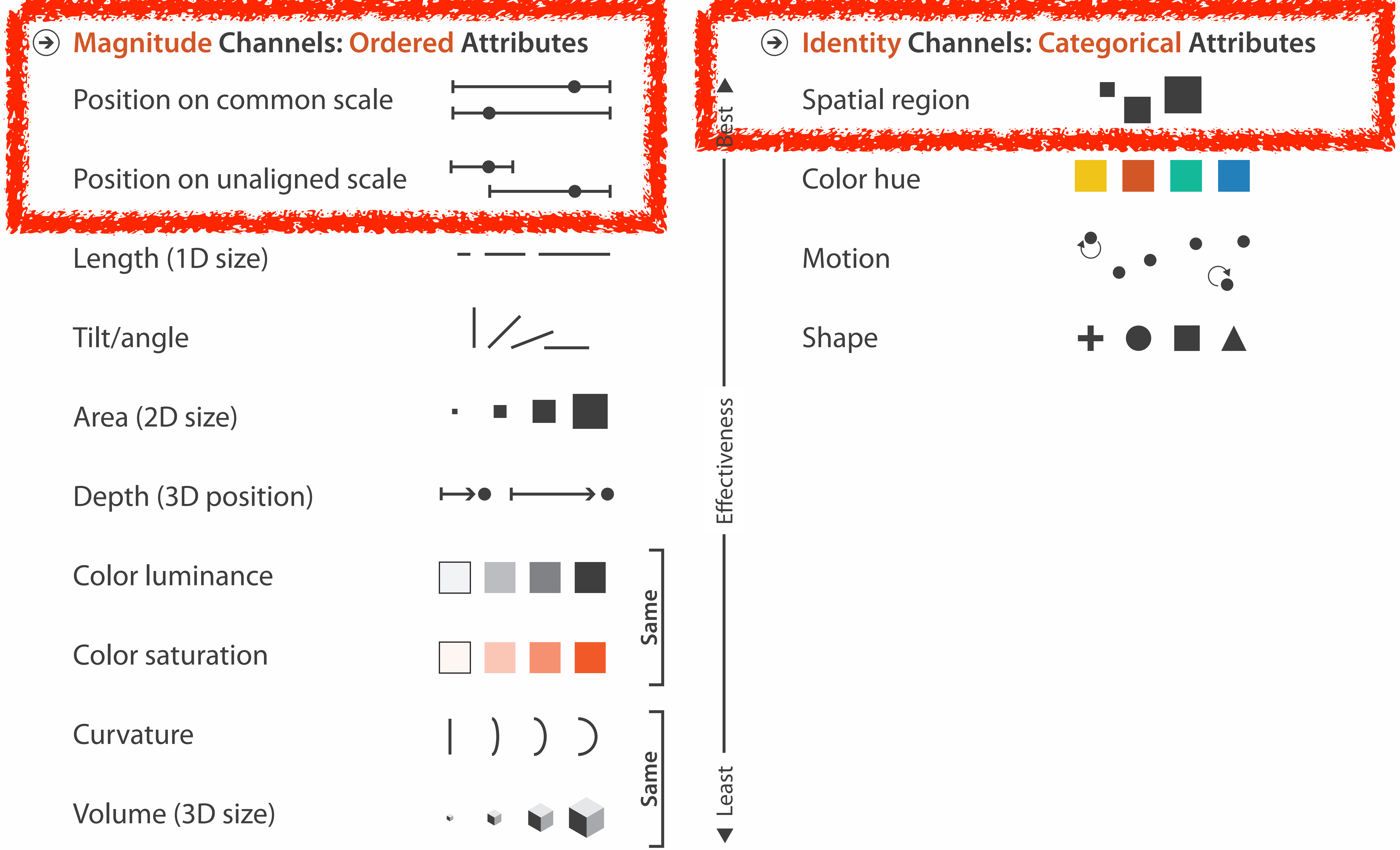
mark: point



4:
vertical position
horizontal position
color hue
size (area)

mark: point

Channels: Expressiveness types and effectiveness rankings



Effectiveness and expressiveness principles

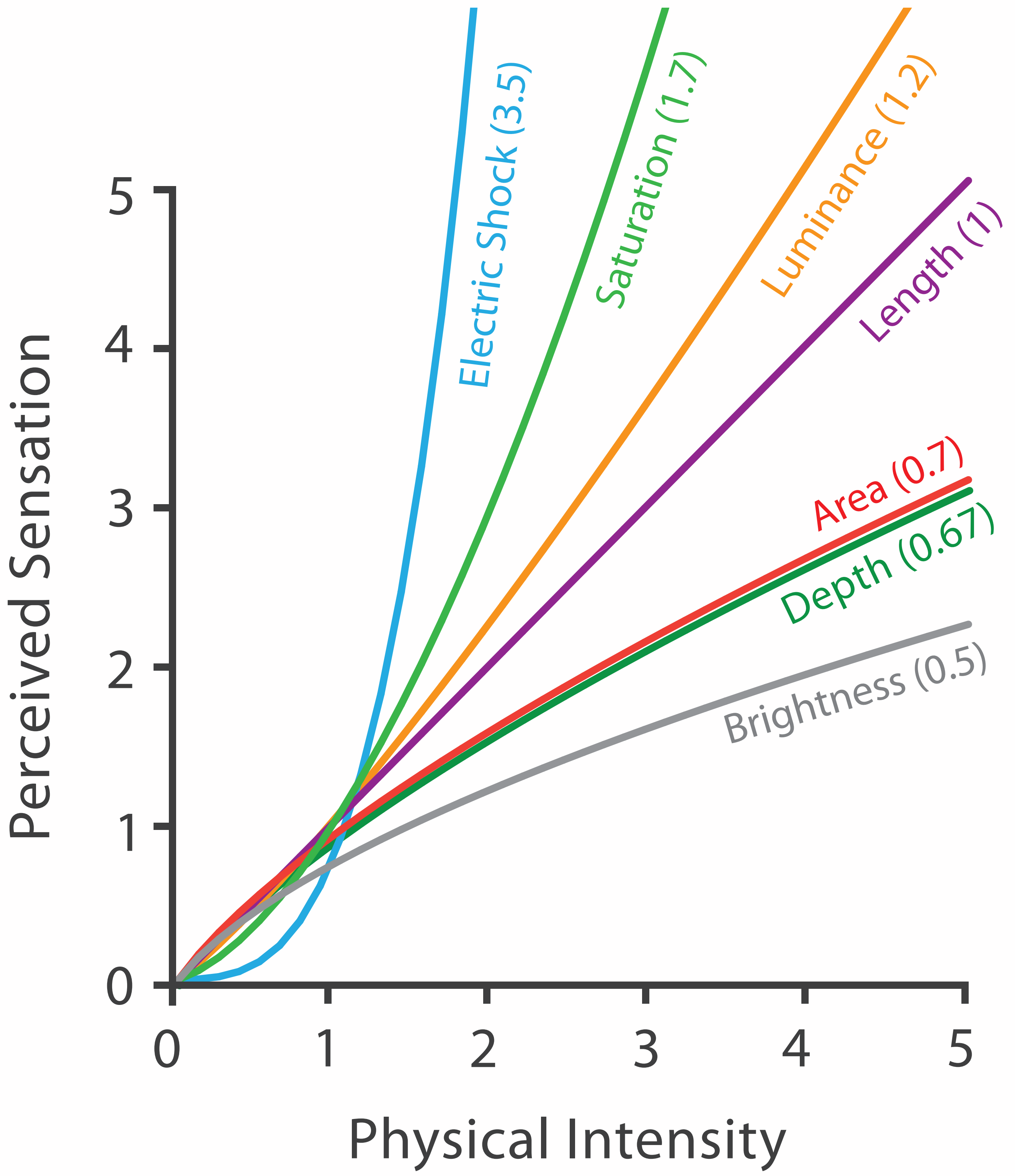
- effectiveness principle
  - encode most important attributes with highest ranked channels
- expressiveness principle
  - match channel and data characteristics

[*Automating the Design of Graphical Presentations of Relational Information.* Mackinlay. ACM Trans. on Graphics (TOG) 5:2 (1986), 110–141.]

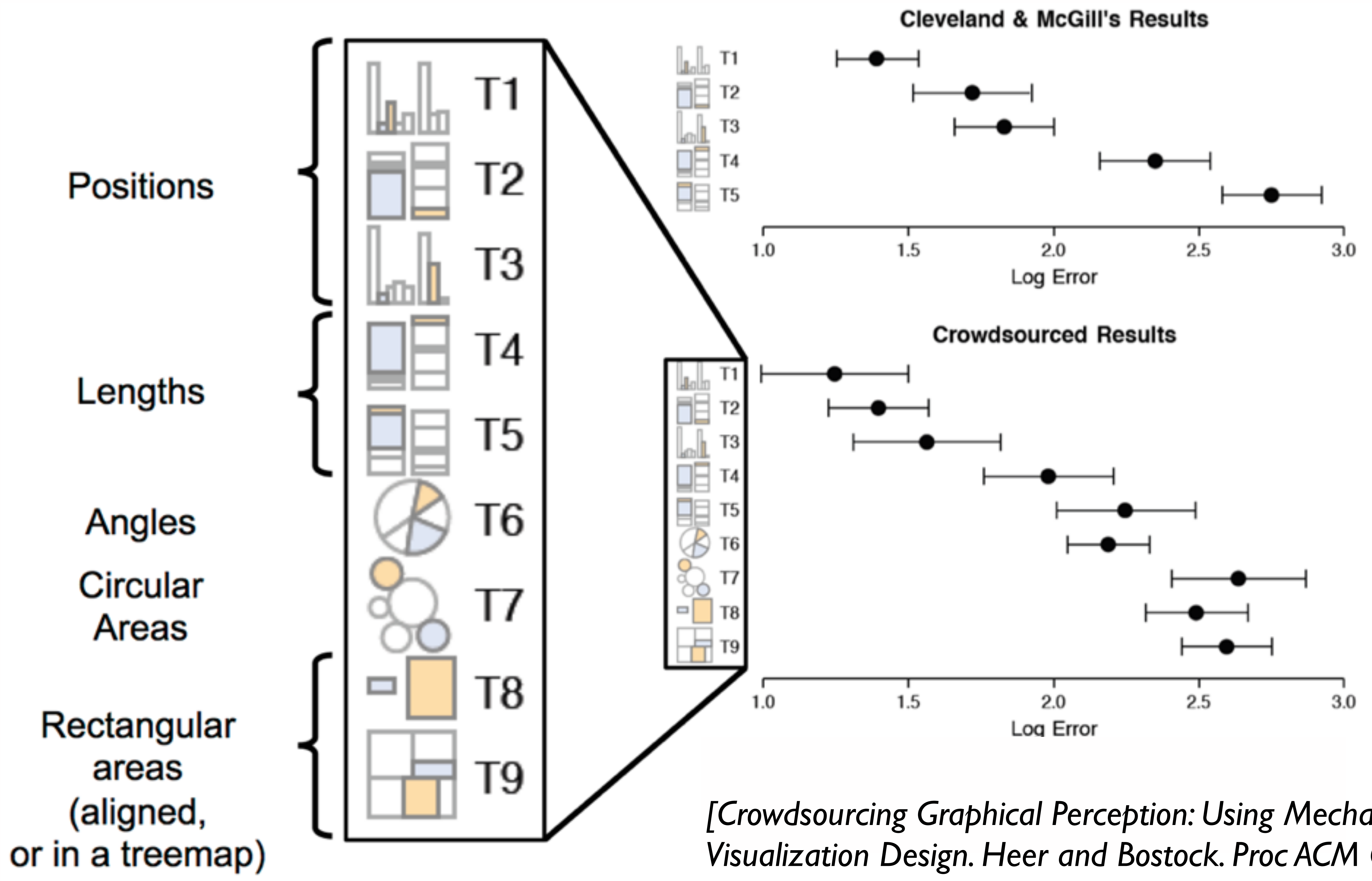
- rankings: where do they come from?
  - accuracy
  - discriminability
  - separability
  - popout

Accuracy: Fundamental Theory

Steven's Psychophysical Power Law: $S = I^N$



Accuracy: Vis experiments

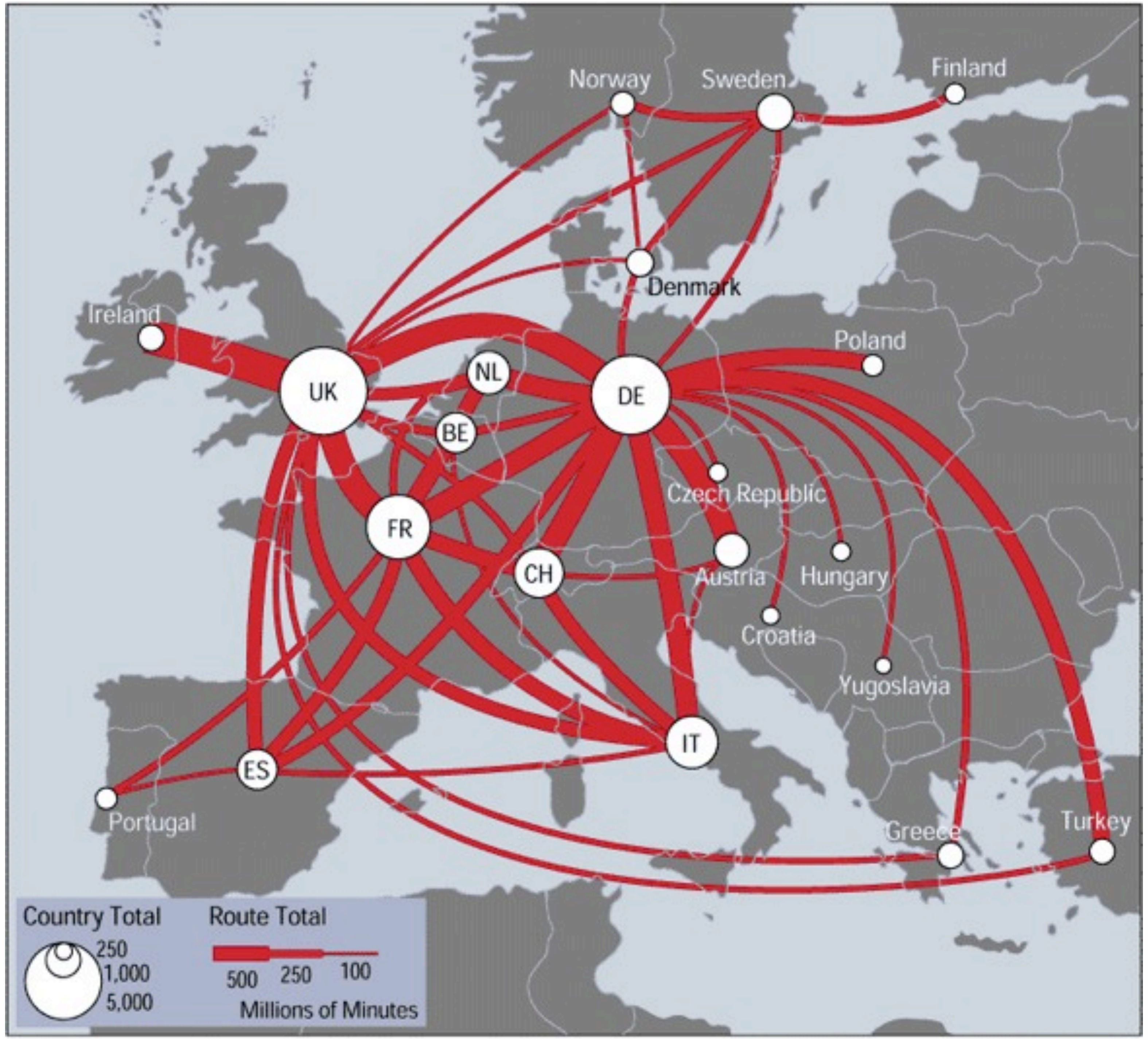


[Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. Heer and Bostock. Proc ACM Conf. Human Factors in Computing Systems (CHI) 2010, p. 203–212.]

after Michael McGuffin course slides, http://profs.etsmtl.ca/mmcguffin/

Discriminability: How many usable steps?

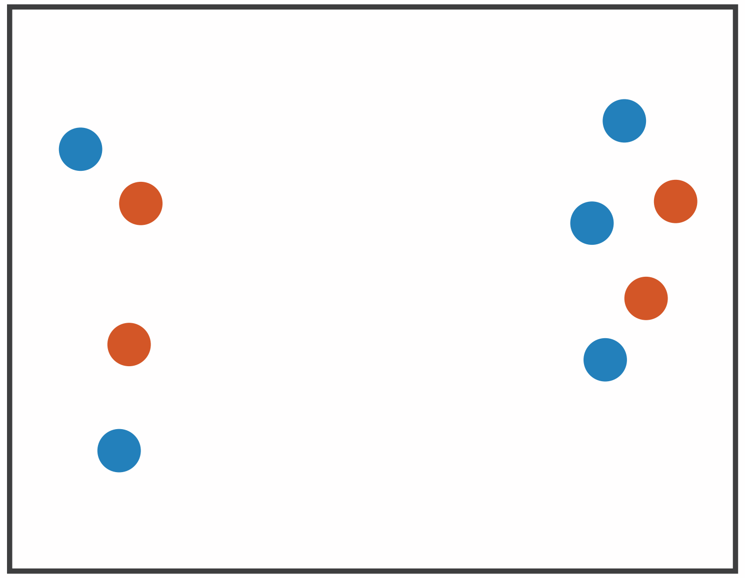
- linewidth: only a few



[mappa.mundi.net/maps/maps 014/telegeography.html]

Separability vs. Integrality

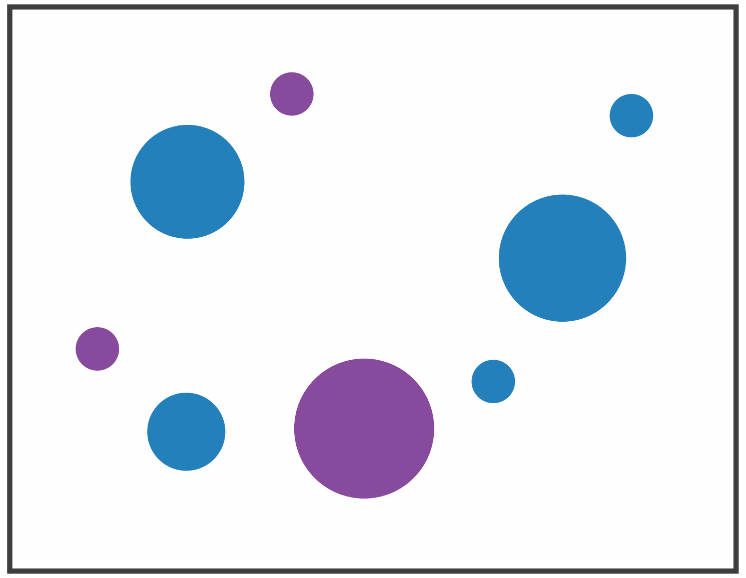
Position
+ Hue (Color)



Fully separable

2 groups each

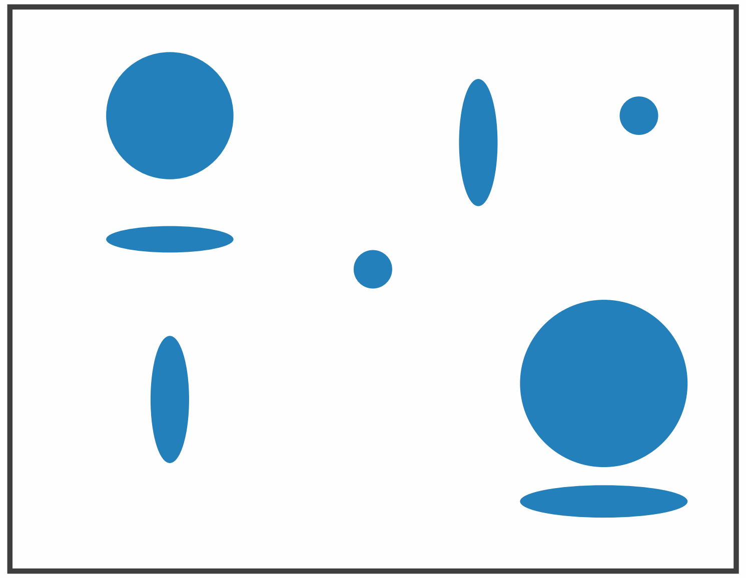
Size
+ Hue (Color)



Some interference

2 groups each

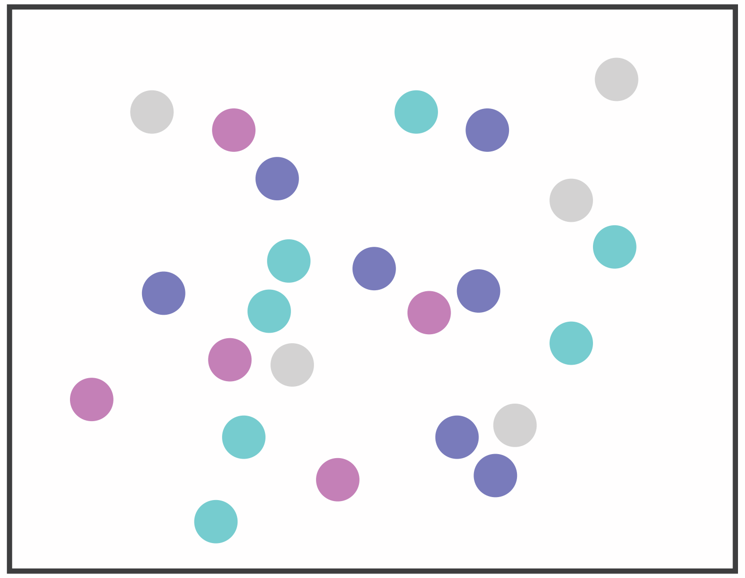
Width
+ Height



Some/significant interference

3 groups total: integral area

Red
+ Green

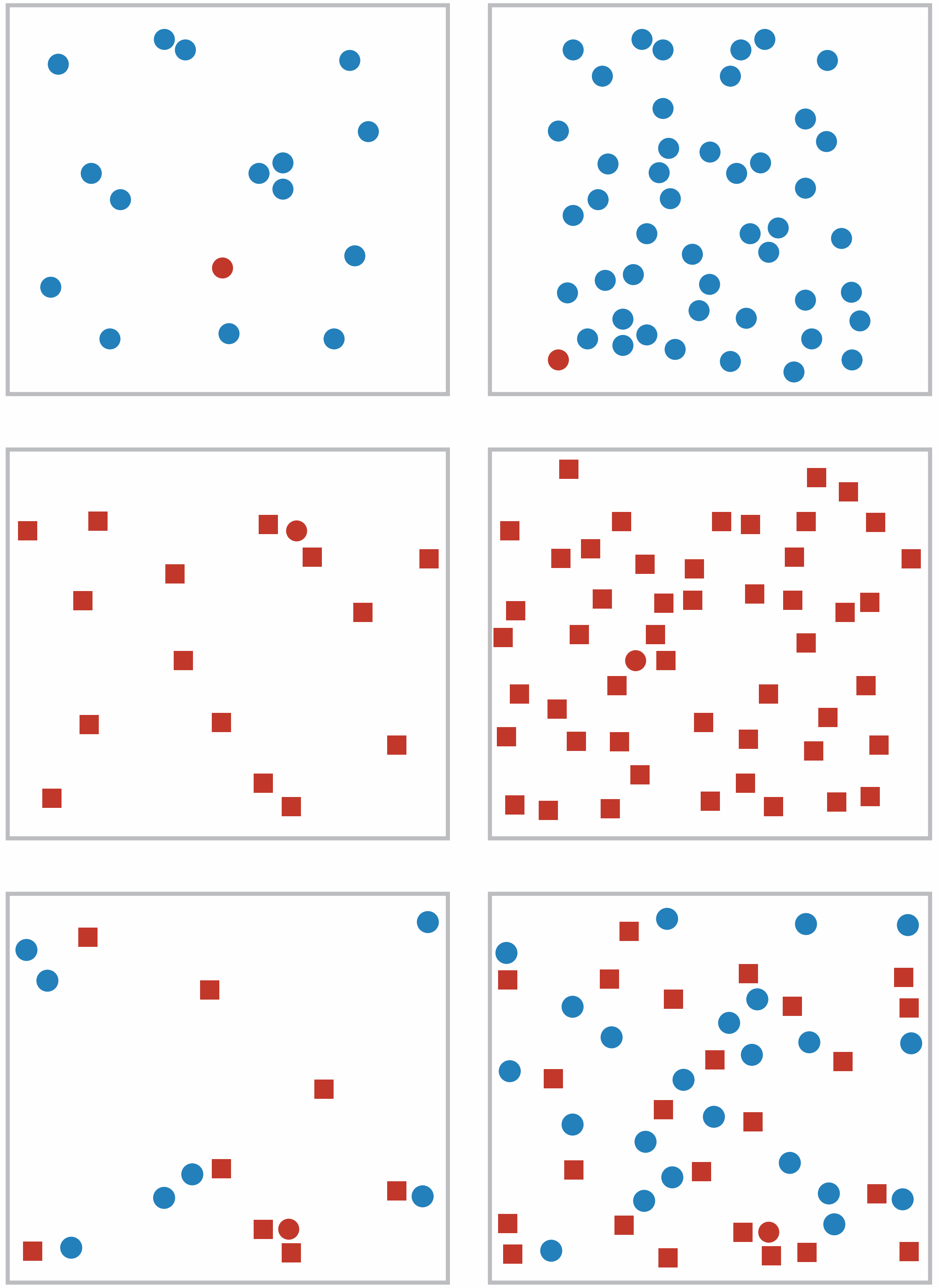


Major interference

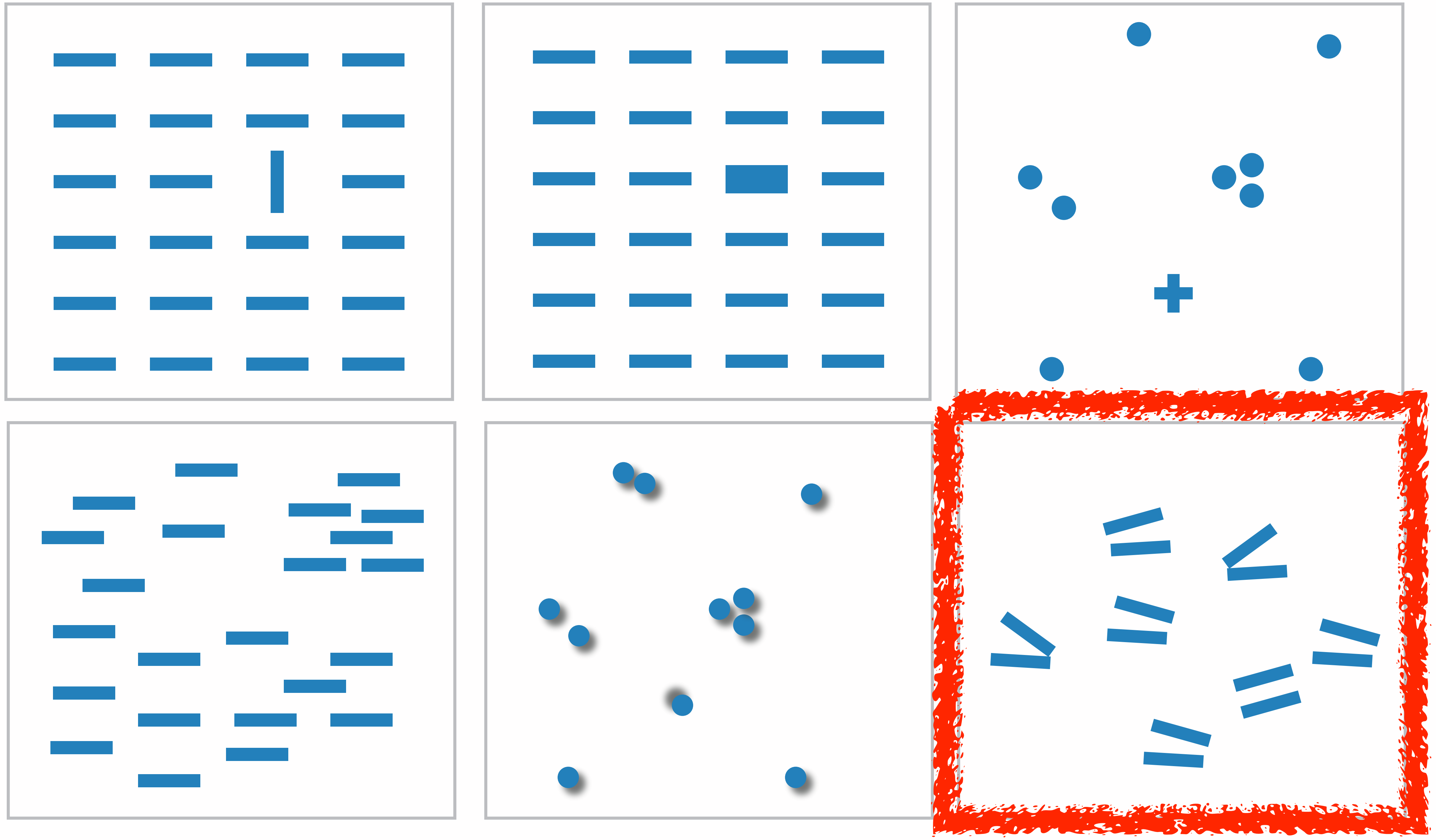
4 groups total: integral hue

Popout

- find the red dot
  - how long does it take?
- parallel processing on many individual channels
  - speed independent of distractor count
  - speed depends on channel and amount of difference from distractors
- serial search for (almost all) combinations
  - speed depends on number of distractors



Popout

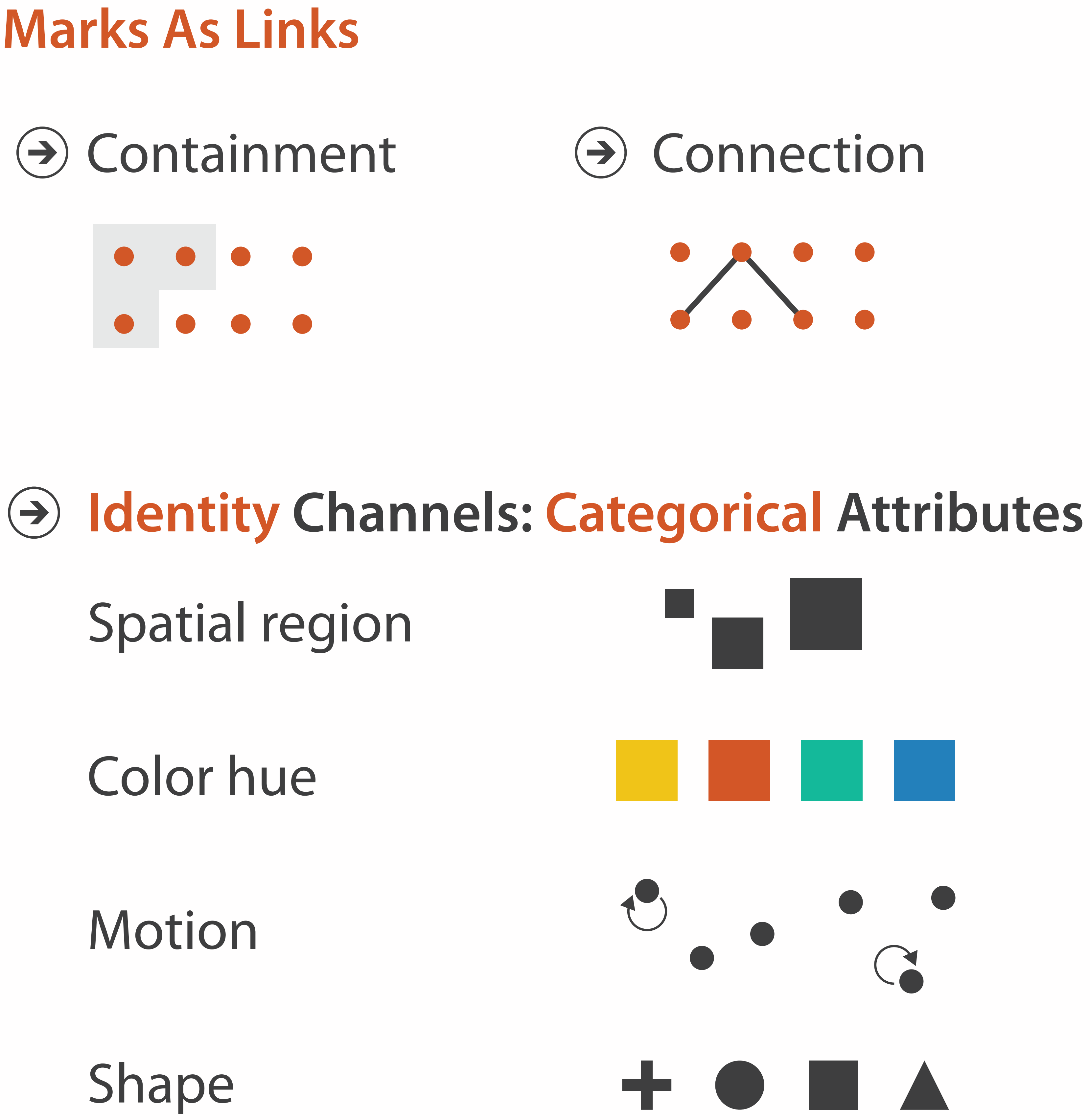


- many channels: tilt, size, shape, proximity, shadow direction, ...
- but not all! parallel line pairs do not pop out from tilted pairs

Grouping

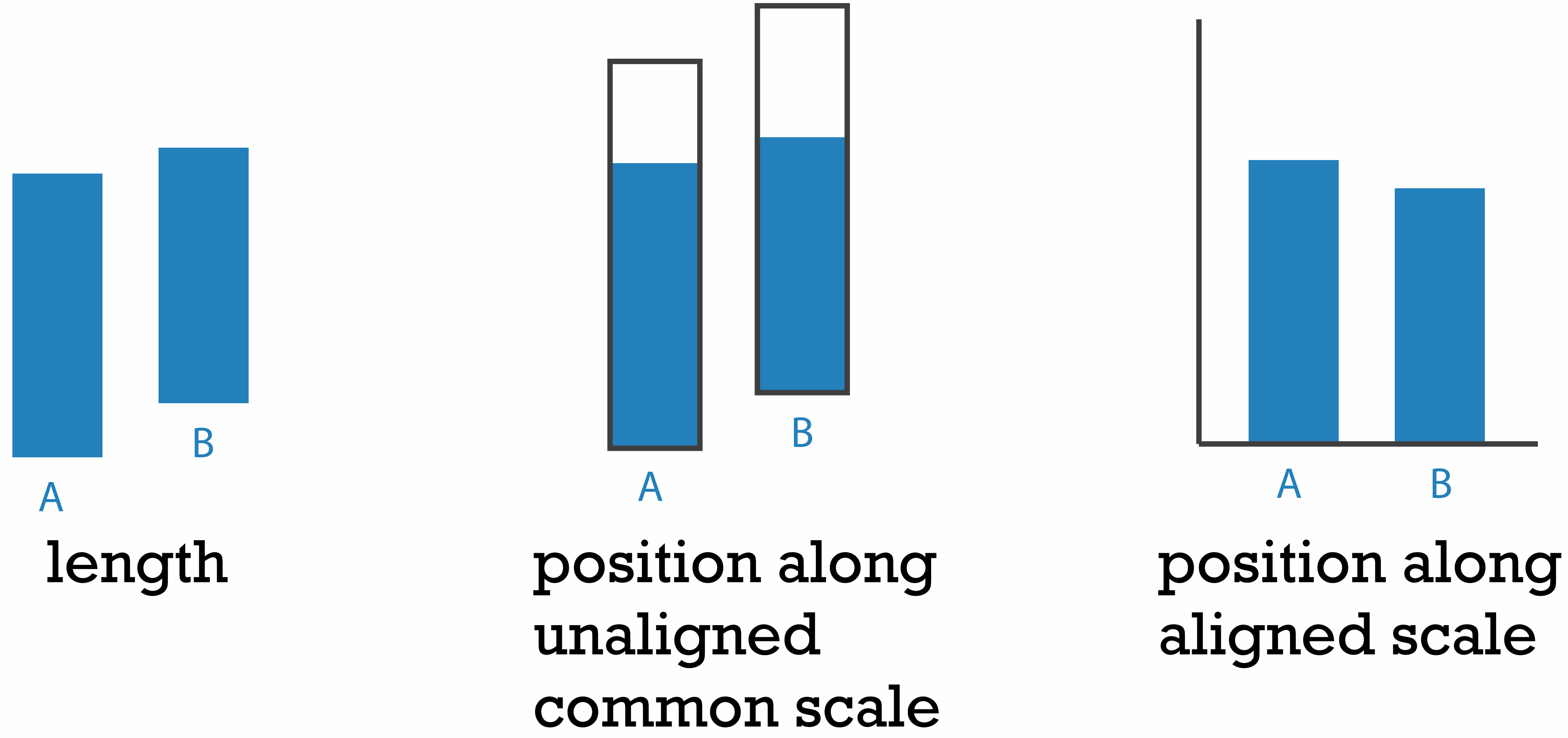
- containment
- connection

- proximity
  - same spatial region
- similarity
  - same values as other categorical channels



Relative vs. absolute judgements

- perceptual system mostly operates with relative judgements, not absolute
  - that's why accuracy increases with common frame/scale and alignment
  - Weber's Law: ratio of increment to background is constant
    - filled rectangles differ in length by 1:9, difficult judgement
    - white rectangles differ in length by 1:2, easy judgement



after [Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Cleveland and McGill. Journ. American Statistical Association 79:387 (1984), 531–554.]

Further reading

- Visualization Analysis and Design. Munzner. AK Peters / CRC Press, Oct 2014.
  - *Chap 5: Marks and Channels*
- *On the Theory of Scales of Measurement.* Stevens. Science 103:2684 (1946), 677–680.
- Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. Stevens. Wiley, 1975.
- *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods.* Cleveland and McGill. Journ. American Statistical Association 79:387 (1984), 531–554.
- *Perception in Vision*. Healey. http://www.csc.ncsu.edu/faculty/healey/PP
- Visual Thinking for Design. Ware. Morgan Kaufmann, 2008.
- Information Visualization: Perception for Design, 3rd edition. Ware. Morgan Kaufmann /Academic Press, 2004.