数据可视化

Lecture 02

Data visualization design (I)

虞思逸

数据可视化设计

I.可视化分析框架
- 引入：定义
- 分析：是什么？为什么？怎么做？
- 标志和通道

II.空间布局
- 编排表格
- 编排空间数据
- 编排网络和树

III.颜色和交互
- 地图颜色
- 操作：改变、选择、导航
- 并列（juxtapose），隔离（partition），重叠（superimpose）

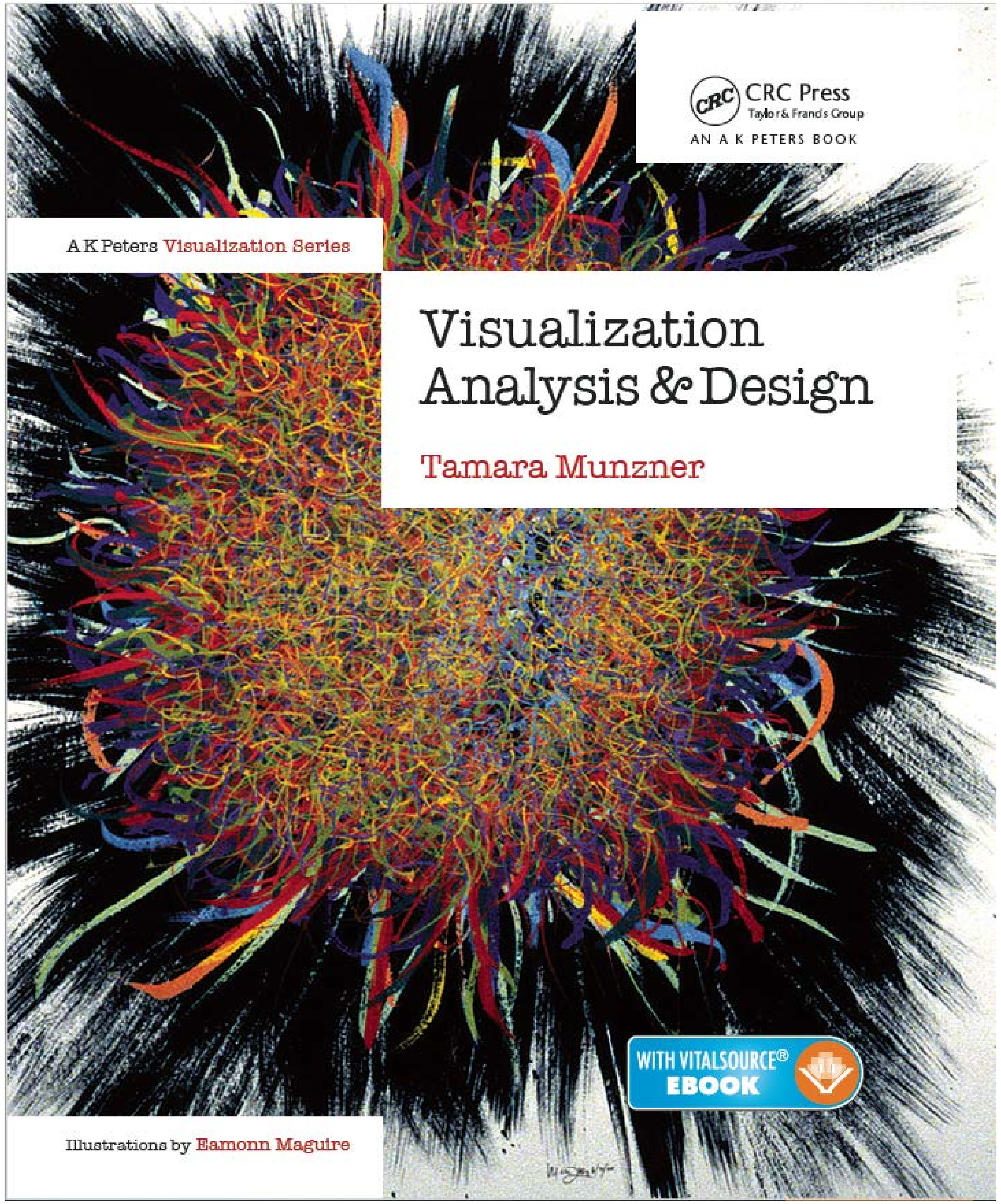
IV.GUIDELINES & METHODS
- 数据的减少：过滤、总计
- 经验法则（Rules of thumb）
- 设计学习方法

CRC Press
Taylor & Francis Group
AN A K PETERS BOOK

A K Peters Visualization Series

Visualization Analysis & Design

Tamara Munzner



WITH VITALSOURCE® EBOOK

Illustrations by Eamonn Maguire

可视化分析框架的
定义

可视化分析框架的
分析框架
What？ Why？ How？

可视化分析框架的
标志与通道

可视化分析框架的
定义

可视化设计的定义

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

可视化设计的应用场景

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

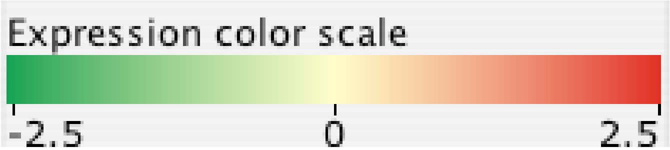
Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

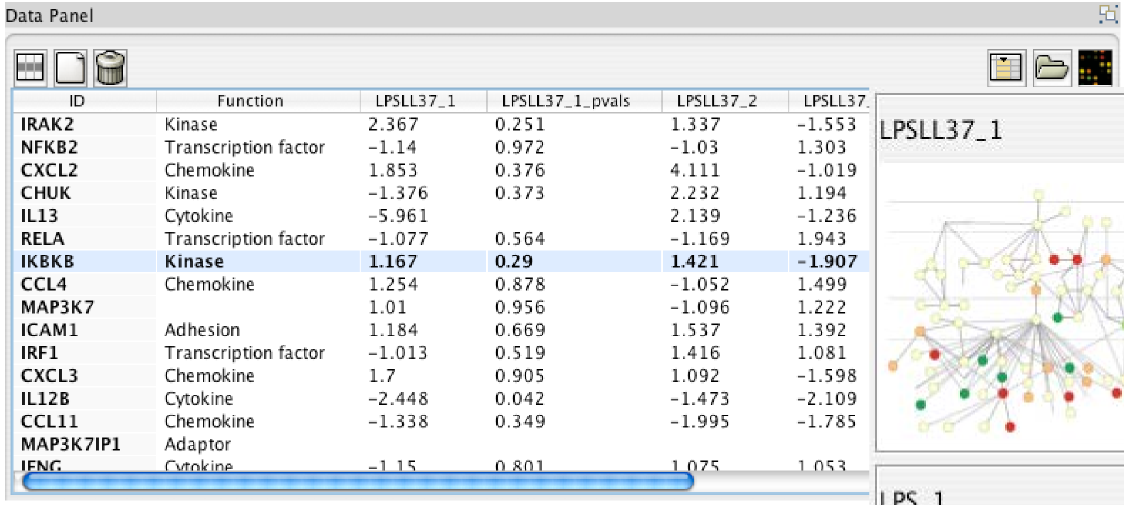
- 提供人类认识过程中所需细节
- 尚未有得到可信结论
  - doesn't know exactly what questions to ask in advance
  - exploratory data analysis
    - ***speed up*** through human-in-the-loop visual data analysis
  - present known results to others
  - stepping stone towards automation
    - before model creation to provide understanding
    - during algorithm creation to refine, debug, set parameters
    - before or during deployment to build trust and monitor

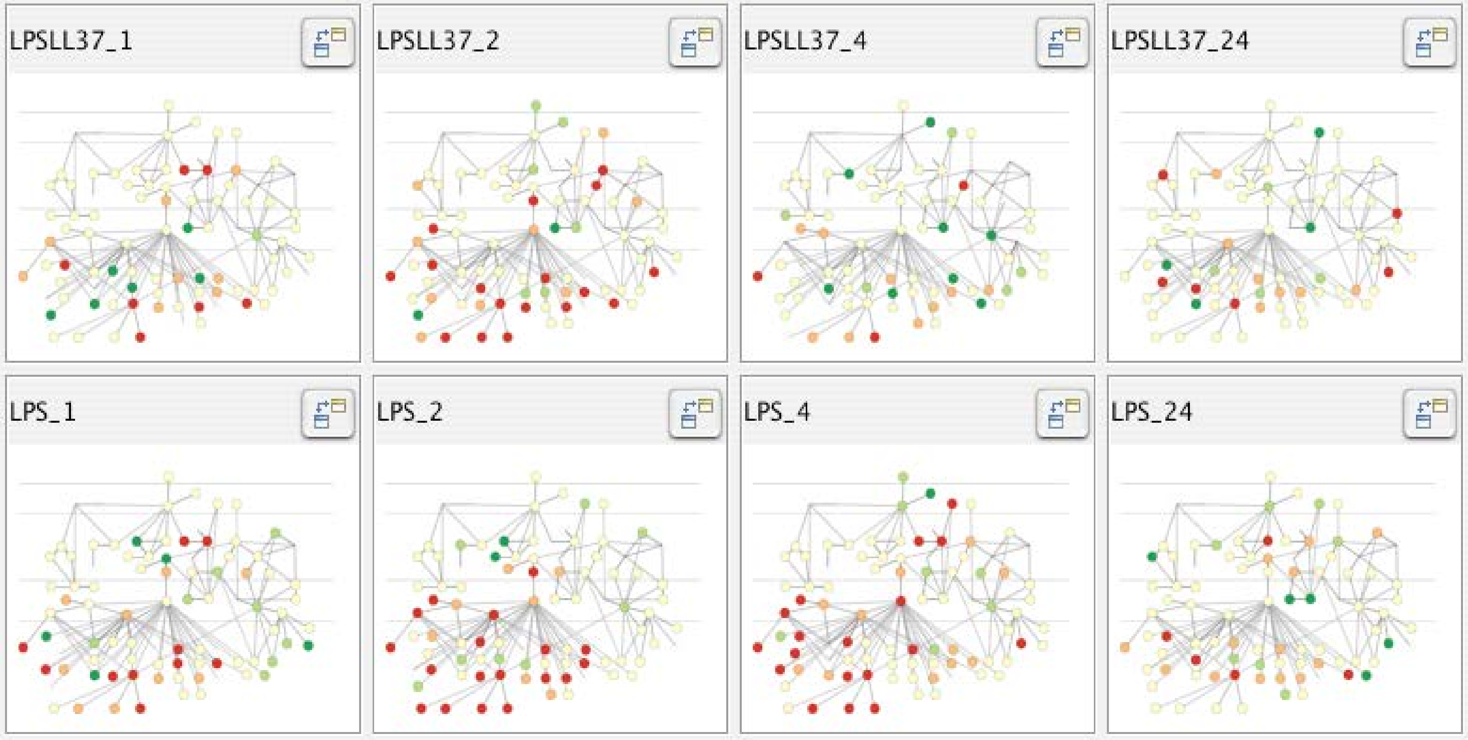
为什么要进行数据的外在可视化?

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

- external representation: 将认知变为感知







[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.]

为什么有时候需要展示所有数据?

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

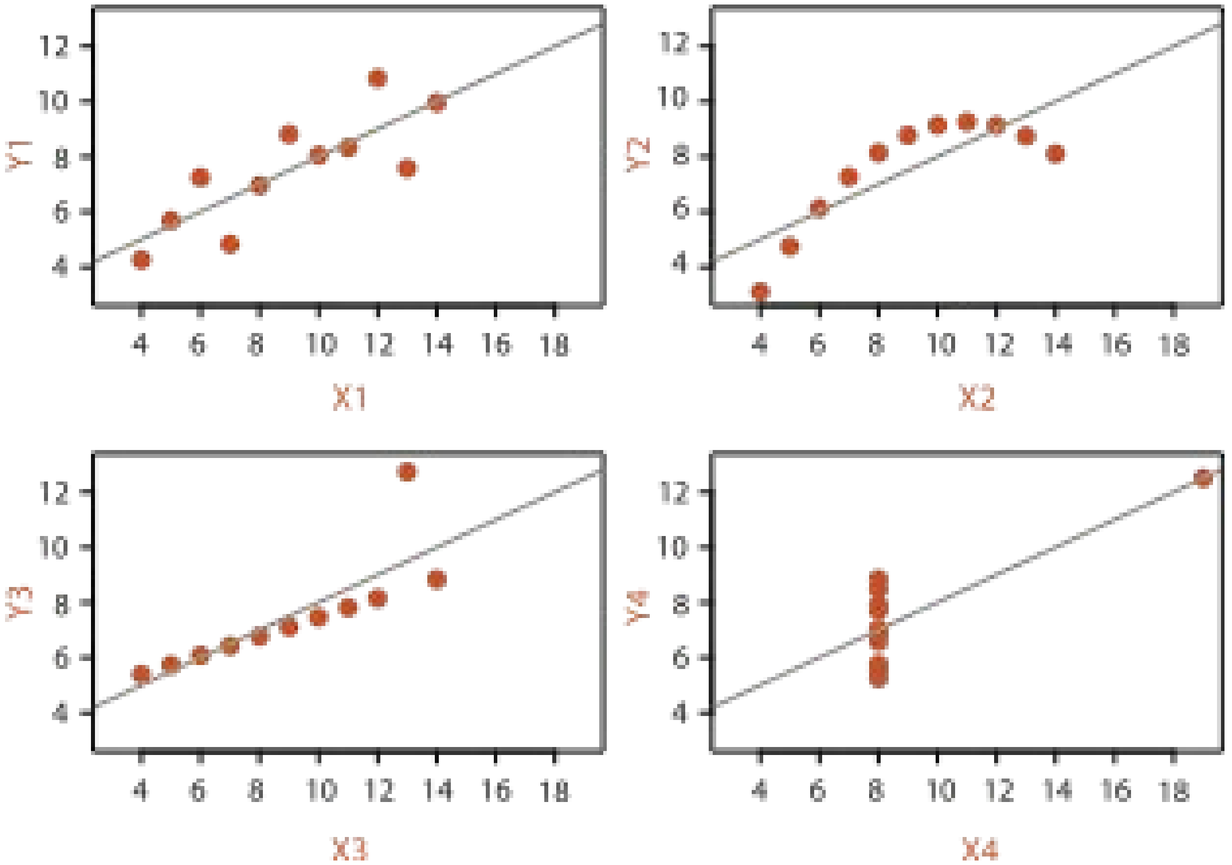
- summaries lose information, details matter
  - confirm expected and find unexpected patterns
  - assess validity of statistical model

Anscombe's Quartet

Identical statistics	
x mean	9
x variance	10
y mean	7.5
y variance	3.75
x/y correlation	0.816

https://www.youtube.com/watch?v=DbJyPELmhJc

Same Stats, Different Graphs



关于任务与效率

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

- effectiveness requires match between data/task and representation
  - set of representations is huge
  - many are ineffective mismatch for specific data/task combo
  - 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
- how to validate effectiveness
  - many methods, must pick appropriate one for your context

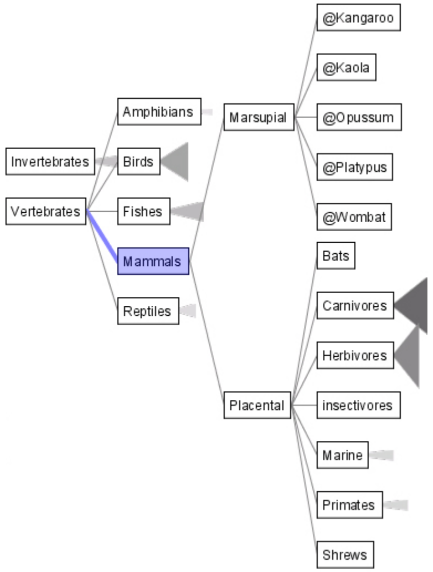
可视化设计过程中的局限性

- 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

数据可视化分析的目的

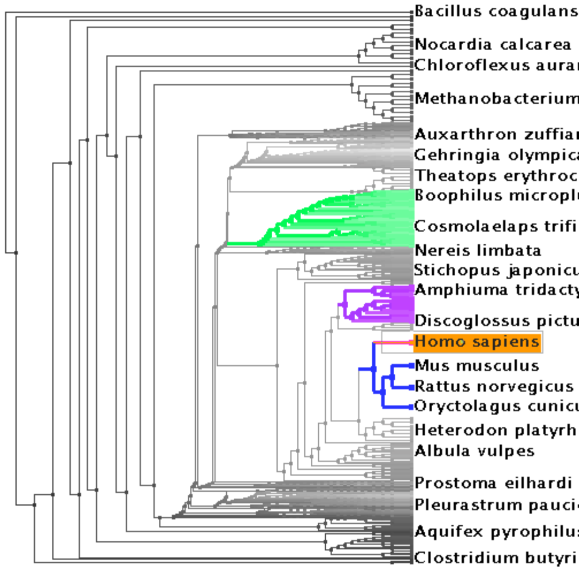
- imposes structure on huge design space
  - –scaffold to help you think systematically about choices
  - –analyzing existing as stepping stone to designing new
  - –most possibilities ineffective for particular task/data combination

SpaceTree



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

TreeJuxtaposer



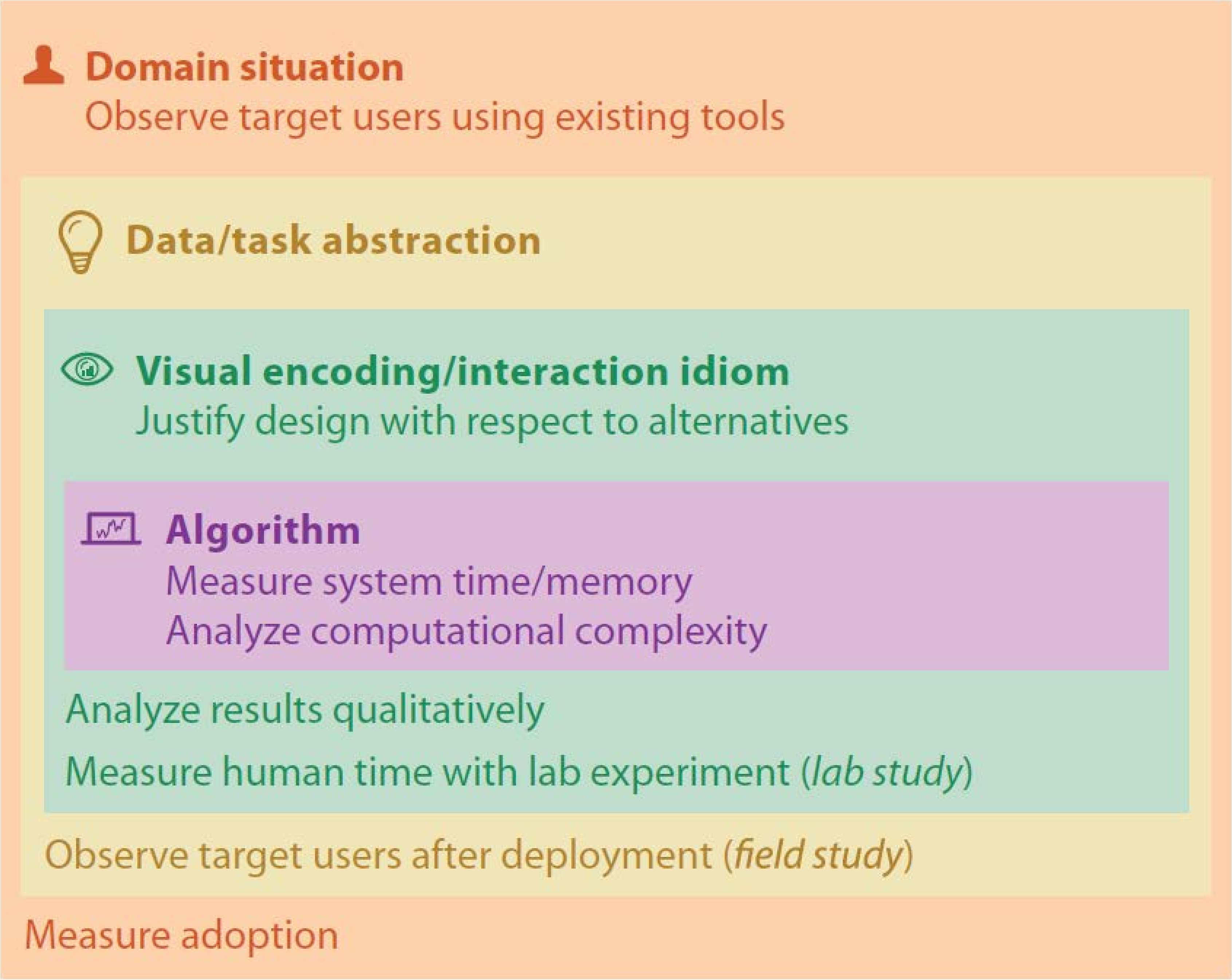
[TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context With Guaranteed Visibility. ACM Trans. on Graphics (Proc. SIGGRAPH) 22:453– 462, 2003.]

可视化分析框架的

分析框架

What? Why? How?

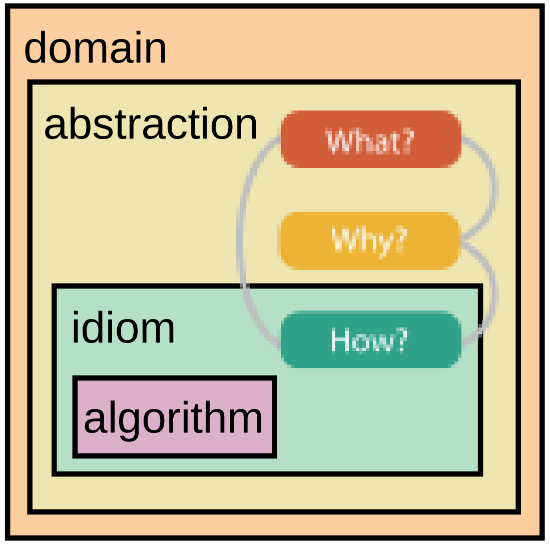
分析框架：
四个层次，
三个问题



分析框架：四个层次，三个问题

[A Nested Model of Visualization Design and Validation.
Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

- *domain situation 领域形势*
  - who are the target users?
- *abstraction 抽象*
  - translate from specifics of domain to vocabulary of vis
    - **what** is shown? **data** abstraction
    - **why** is the user looking at it? **task** abstraction
- *idiom 风格*
  - **how** is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm算法*
  - efficient computation



[A Multi-Level Typology of Abstract Visualization Tasks
Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

Why is validation difficult?

- solution: use methods from different fields at each level

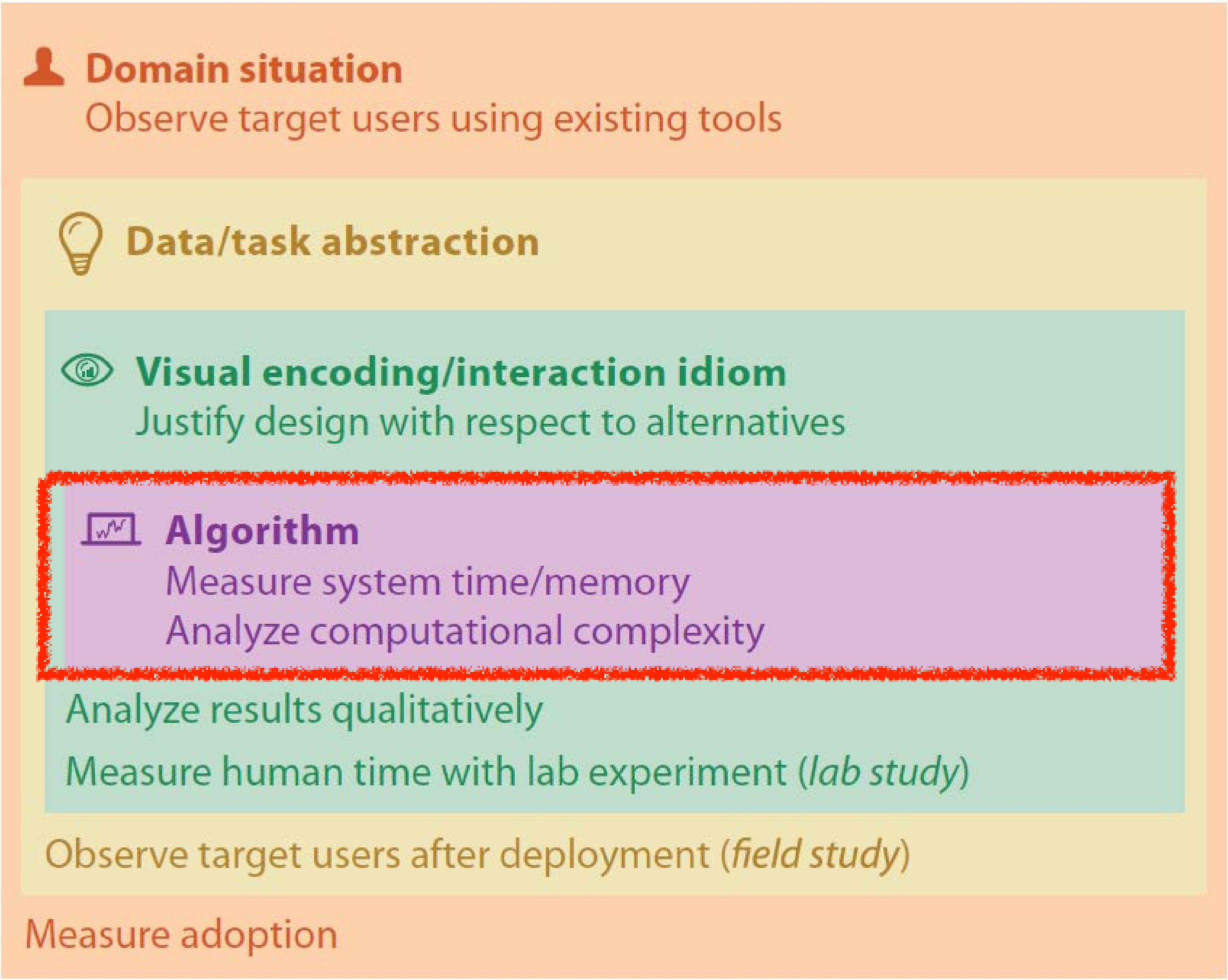
anthropology/ ethnography

cognitive psychology

design

computer science

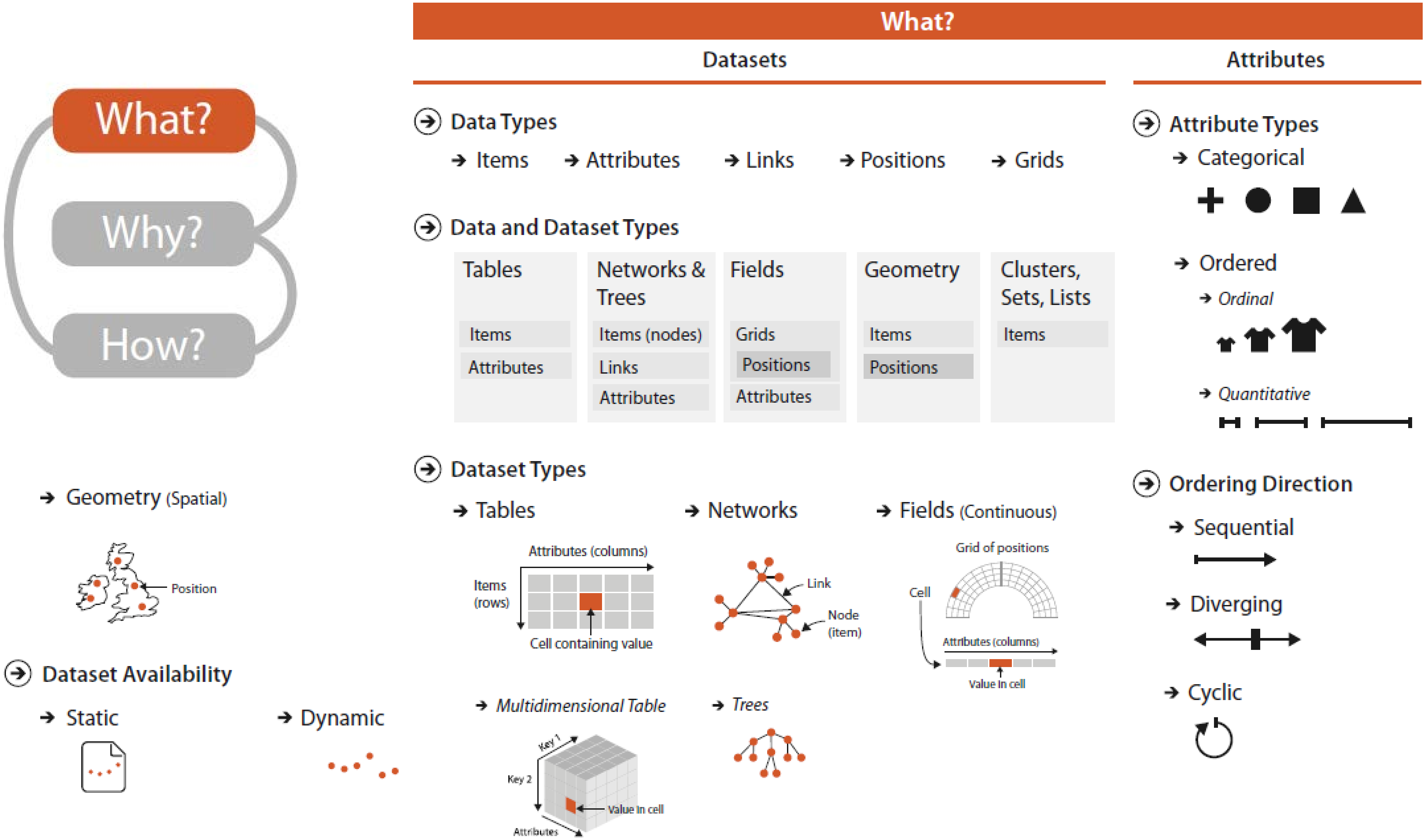
anthropology/ ethnography



problem-driven work

technique-driven work

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

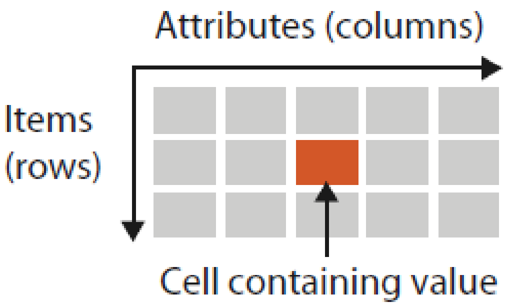


What?
Why?
How?
What?
Datasets
Attributes
Data Types
Items
Attributes
Links
Positions
Grids
Data and Dataset Types
Tables
Items
Attributes
Networks & Trees
Items (nodes)
Links
Attributes
Fields
Grids
Positions
Attributes
Geometry
Items
Positions
Clusters, Sets, Lists
Items
Dataset Types
Tables
Attributes (columns)
Items (rows)
Cell containing value
Networks
Link
Node (item)
Fields (Continuous)
Grid of positions
Cell
Attributes (columns)
Value in cell
Multidimensional Table
Key 1
Key 2
Attributes
Value in cell
Trees
Geometry (Spatial)
Position
Dataset Availability
Static
Dynamic
Attribute Types
Categorical
Ordered
Ordinal
Quantitative
Ordering Direction
Sequential
Diverging
Cyclic

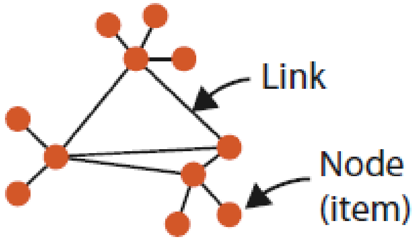
四种数据类型

➔ Dataset Types

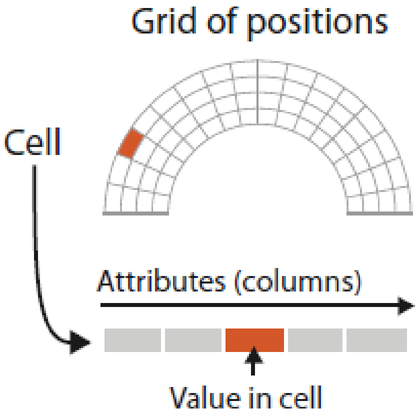
➔ Tables



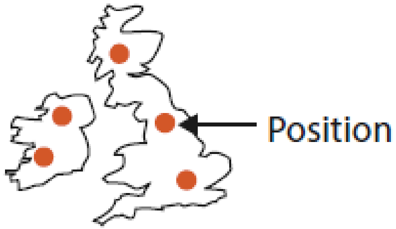
➔ Networks



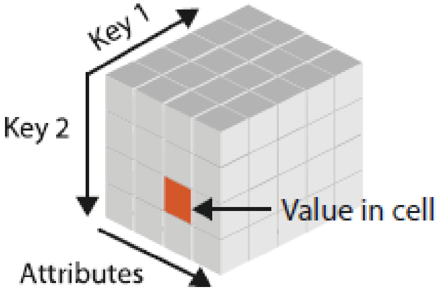
➔ Fields (Continuous)



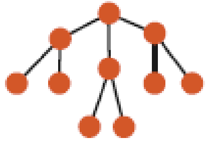
➔ Geometry (Spatial)



➔ *Multidimensional Table*



➔ *Trees*



- visualization vs computer graphics
  – geometry is design decision

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		

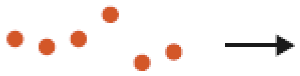
Data Types

- Items
- Attributes
- Links
- Positions
- Grids

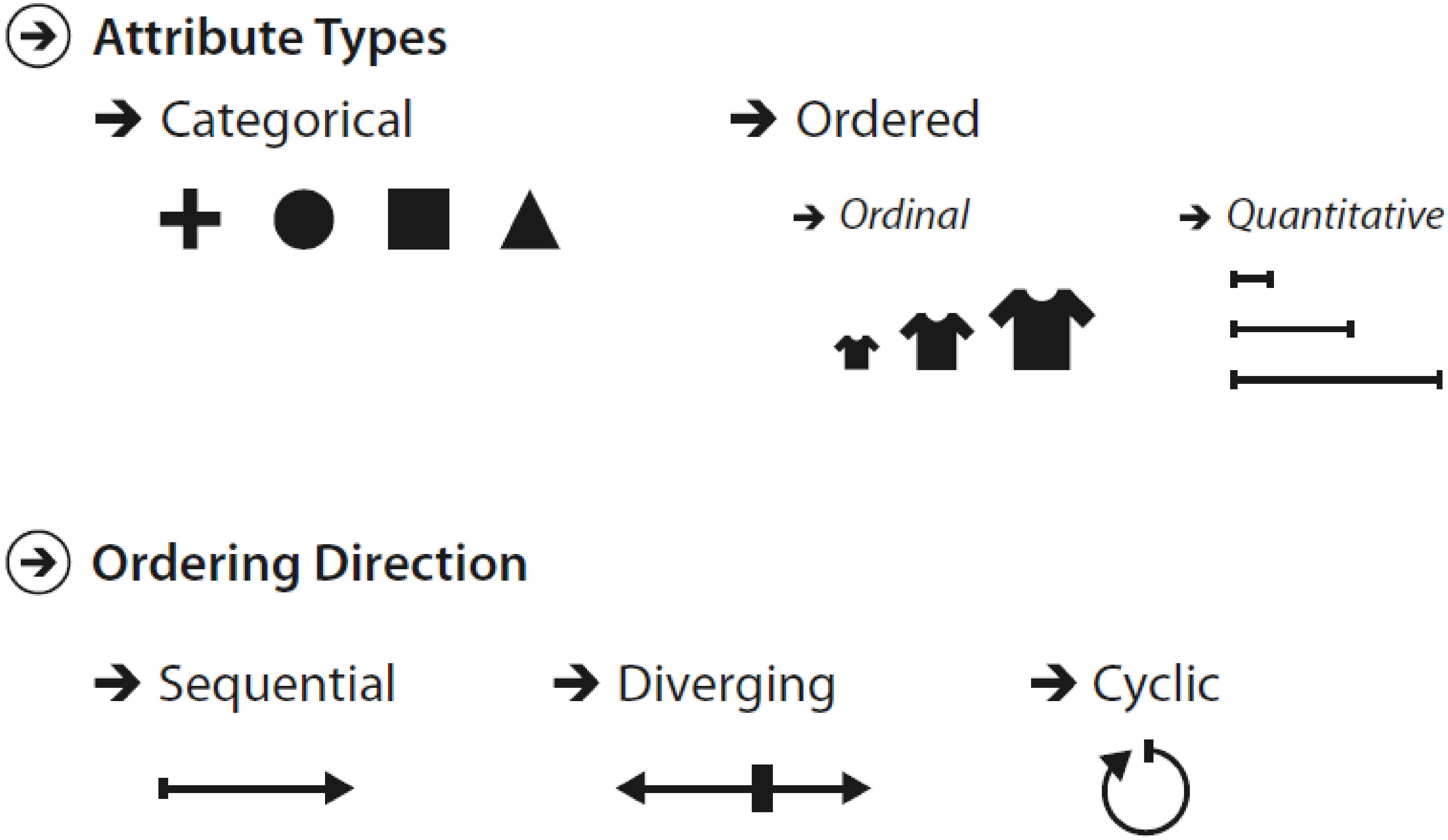
Dataset Availability

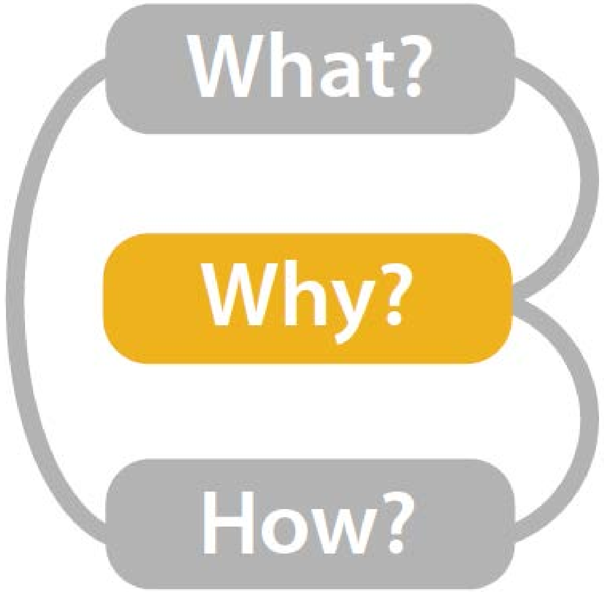
- Static

- Dynamic

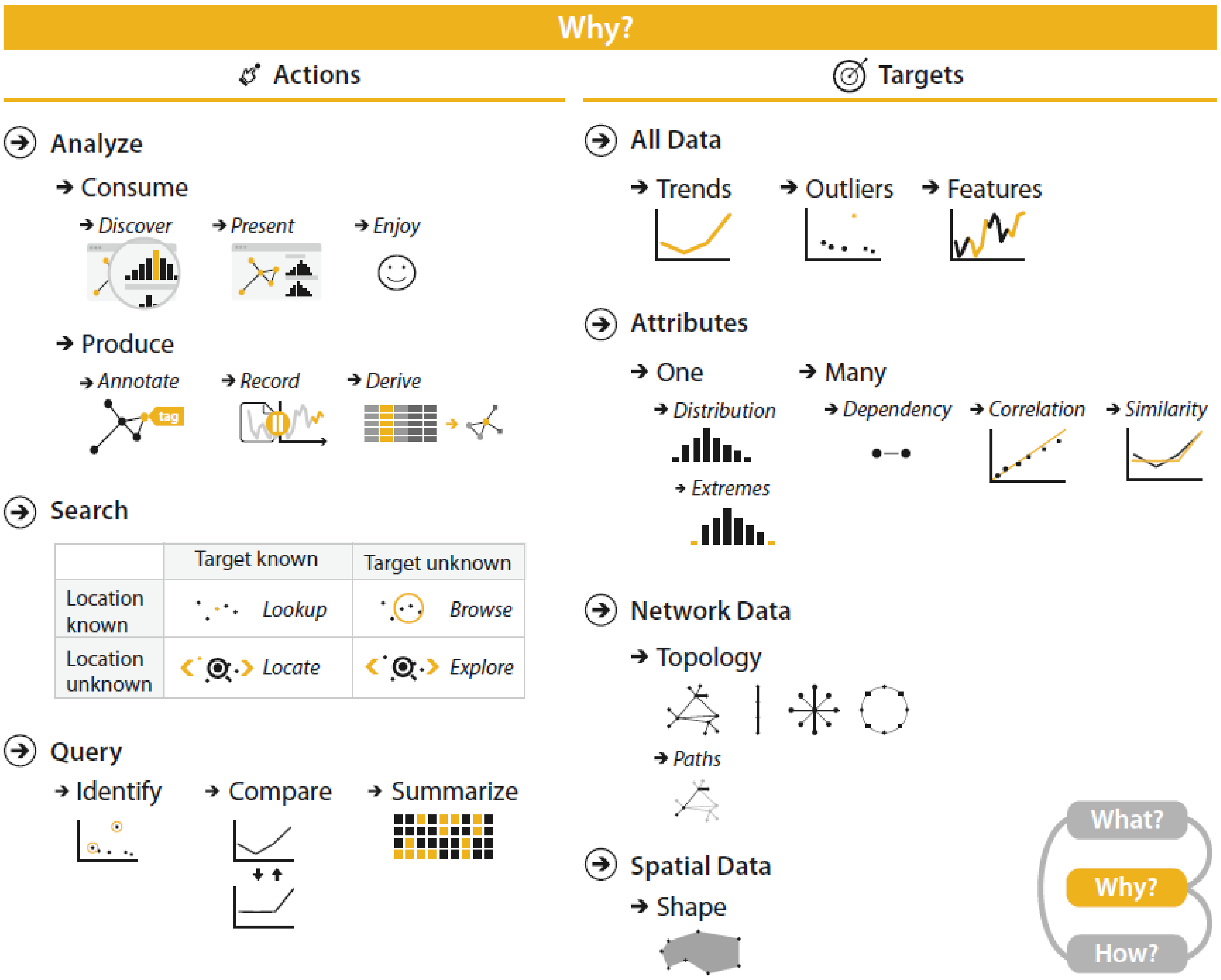


属性类型


Attribute Types
Categorical
Ordered
Ordinal
Quantitative
Ordering Direction
Sequential
Diverging
Cyclic

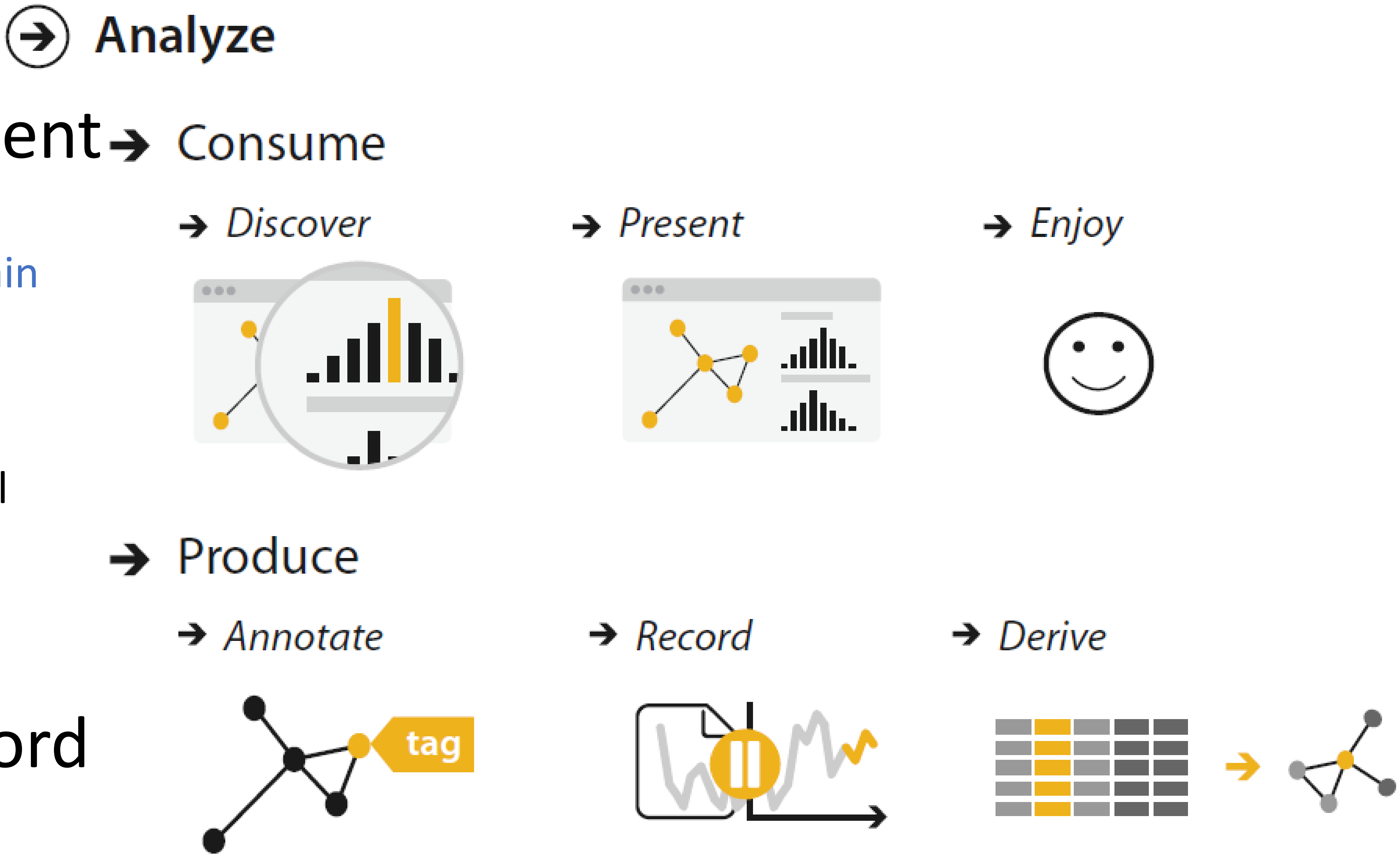


- {action, target} pairs
  - –*discover distribution*
  - –*compare trends*
  - –*locate outliers*
  - –*browse topology*



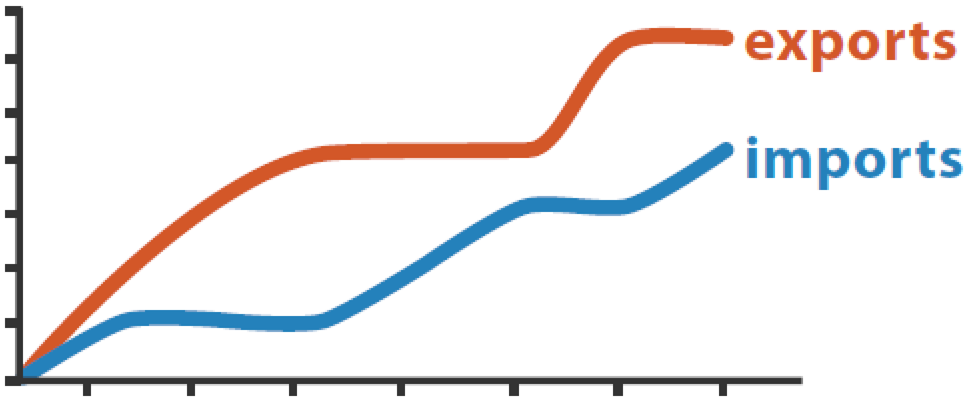
Actions: Analyze

- consume
  - –Discover/present
    - classic split
    - aka explore vs explain
  - –enjoy
    - newcomer
    - aka casual, social
- produce
  - –annotate, record
  - –derive
    - crucial design choice

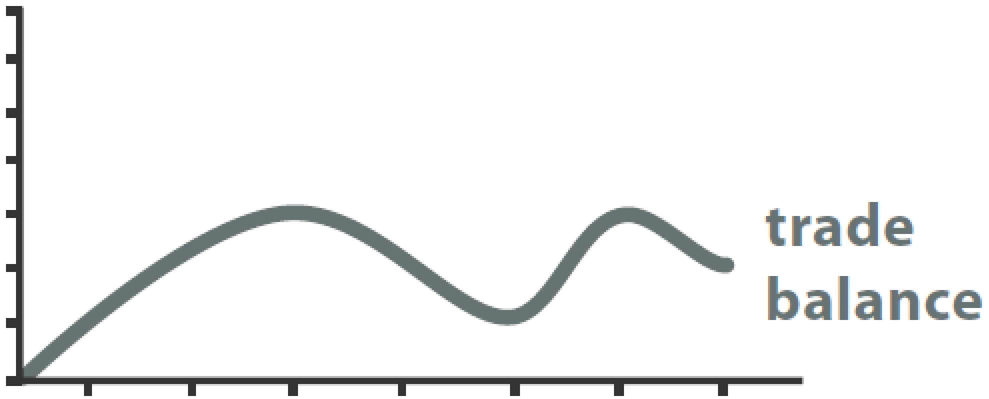


Derive

- don't just draw what you're given!
  - decide what the right thing to show is
  - create it with a series of transformations from the original dataset
  - draw that
- one of the four major strategies for handling complexity



Original Data



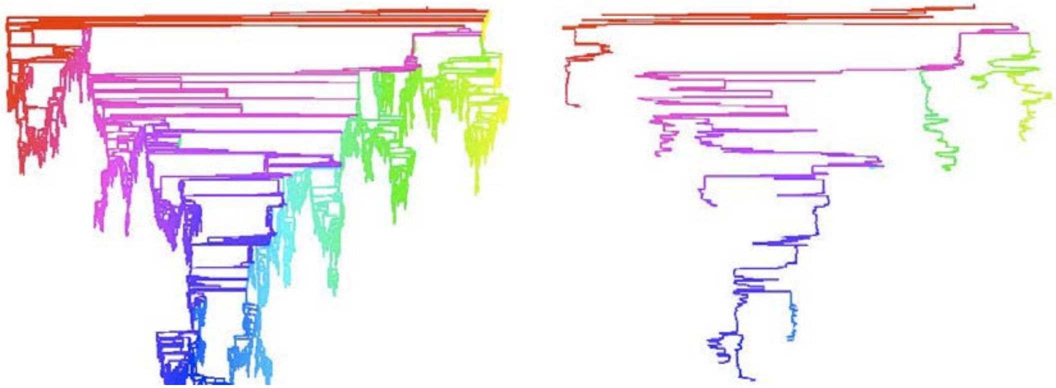
trade balance = exports − imports

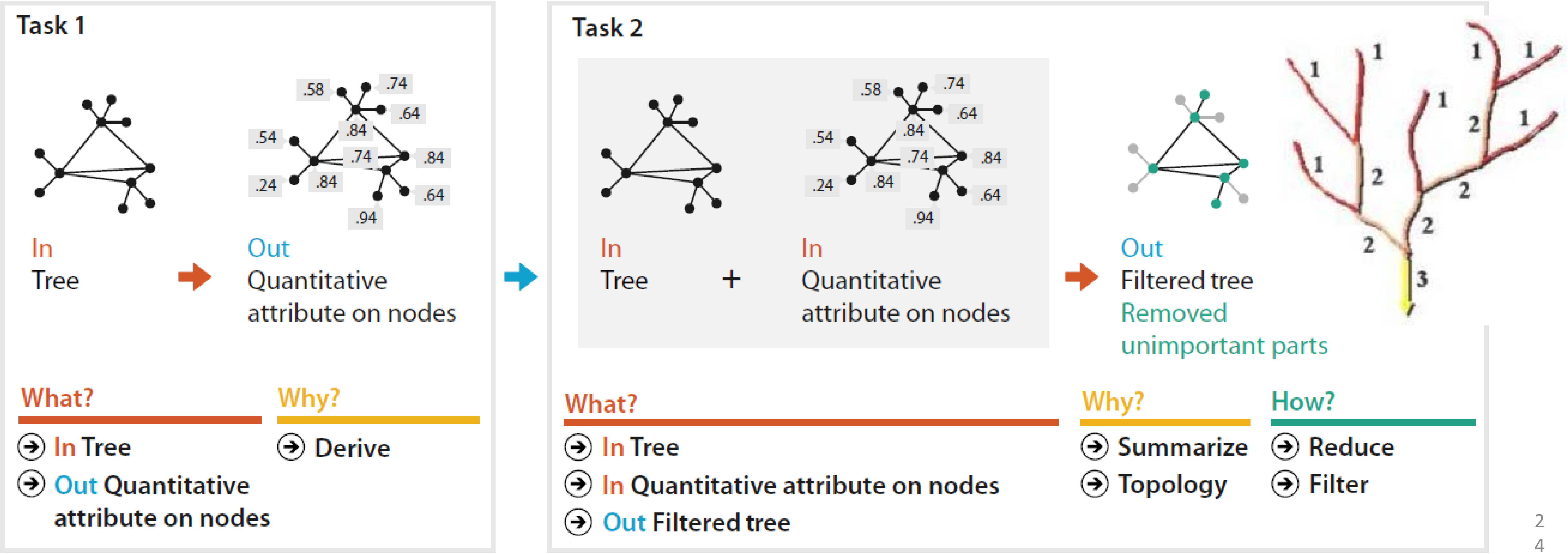
Derived Data

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.





Actions: Search, query

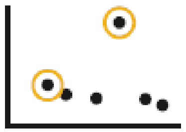
- what does user know?
  - –target, location
- how much of the data matters?
  - –one, some, all
- independent choices for each of these three levels
  - –analyze, search, query
  - –mix and match

Search

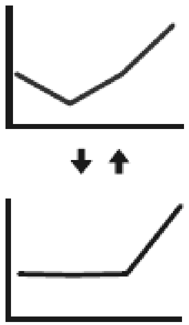
	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

Query

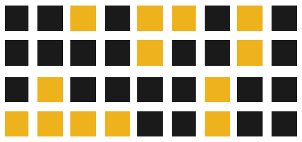
→ Identify



→ Compare



→ Summarize

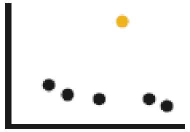


Why: Targets

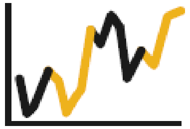
All Data

- Trends

- Outliers

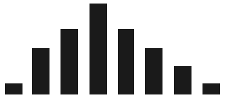


- Features

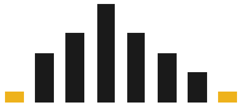


Attributes

- One
  - *Distribution*

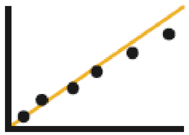


  - *Extremes*

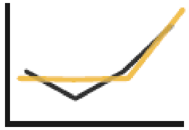


- Many
  - *Dependency*

  - *Correlation*

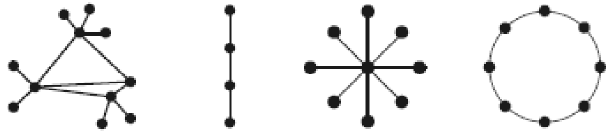


  - *Similarity*

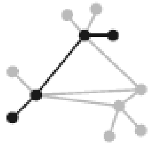


Network Data

- Topology

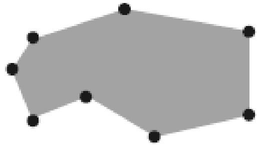


  - *Paths*

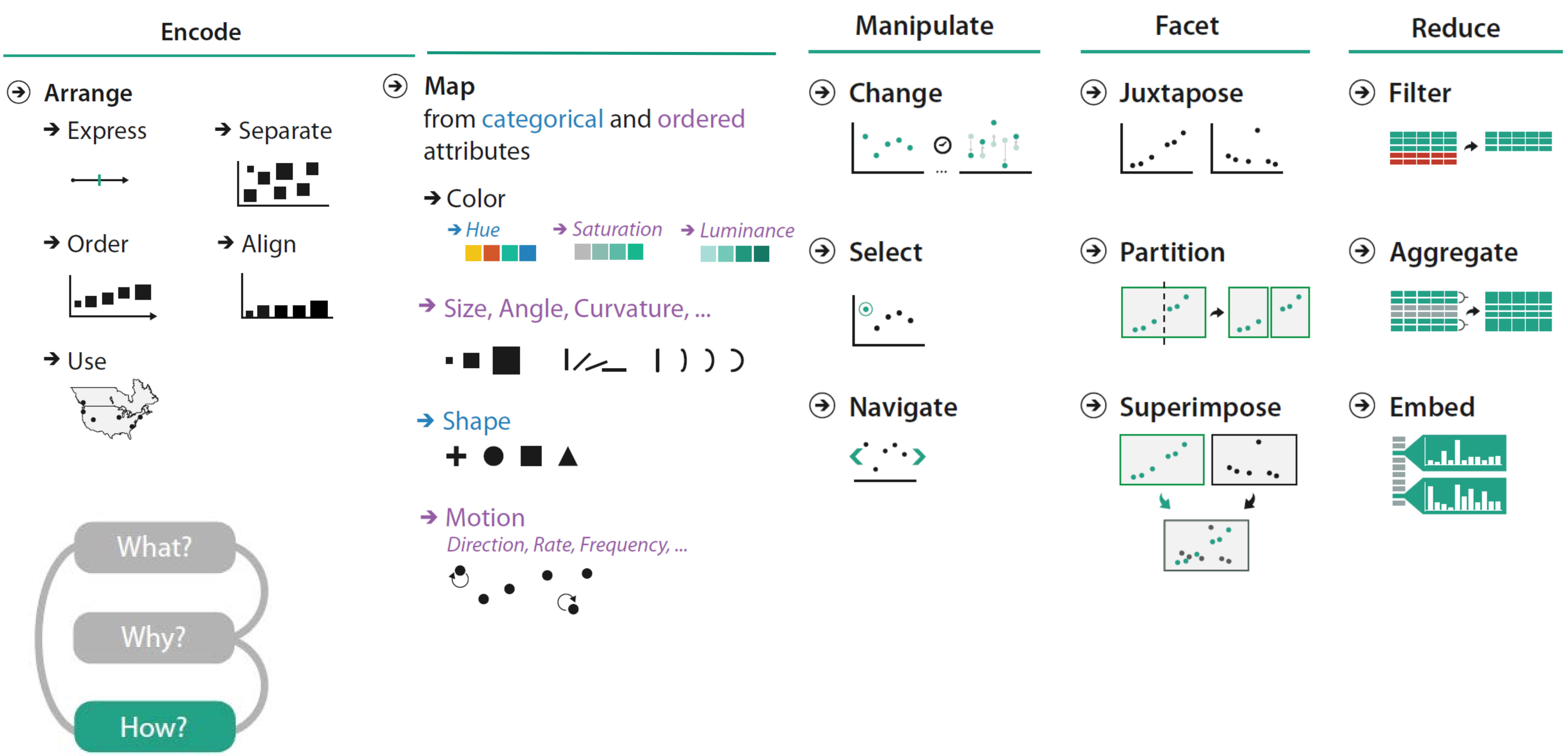


Spatial Data

- Shape



How?


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

Further reading

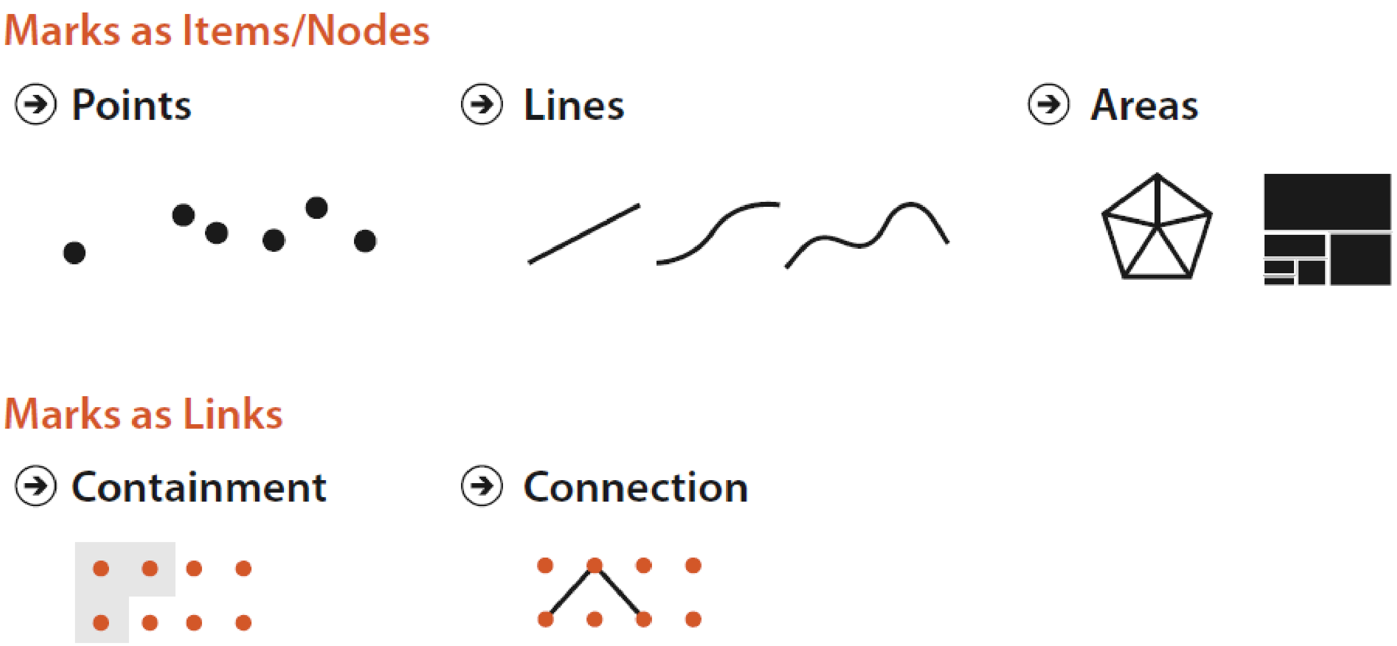
- Visualization Analysis and Design. Munzner. AK Peters Visualization Series, CRC Press, 2014.
  - *Chap 2: What: Data Abstraction*
  - *Chap 3: Why: Task Abstraction*
- *A Multi-Level Typology of Abstract Visualization Tasks.* Brehmer and Munzner. IEEE Trans. Visualization and Computer Graphics (Proc. 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.

可视化分析框架的

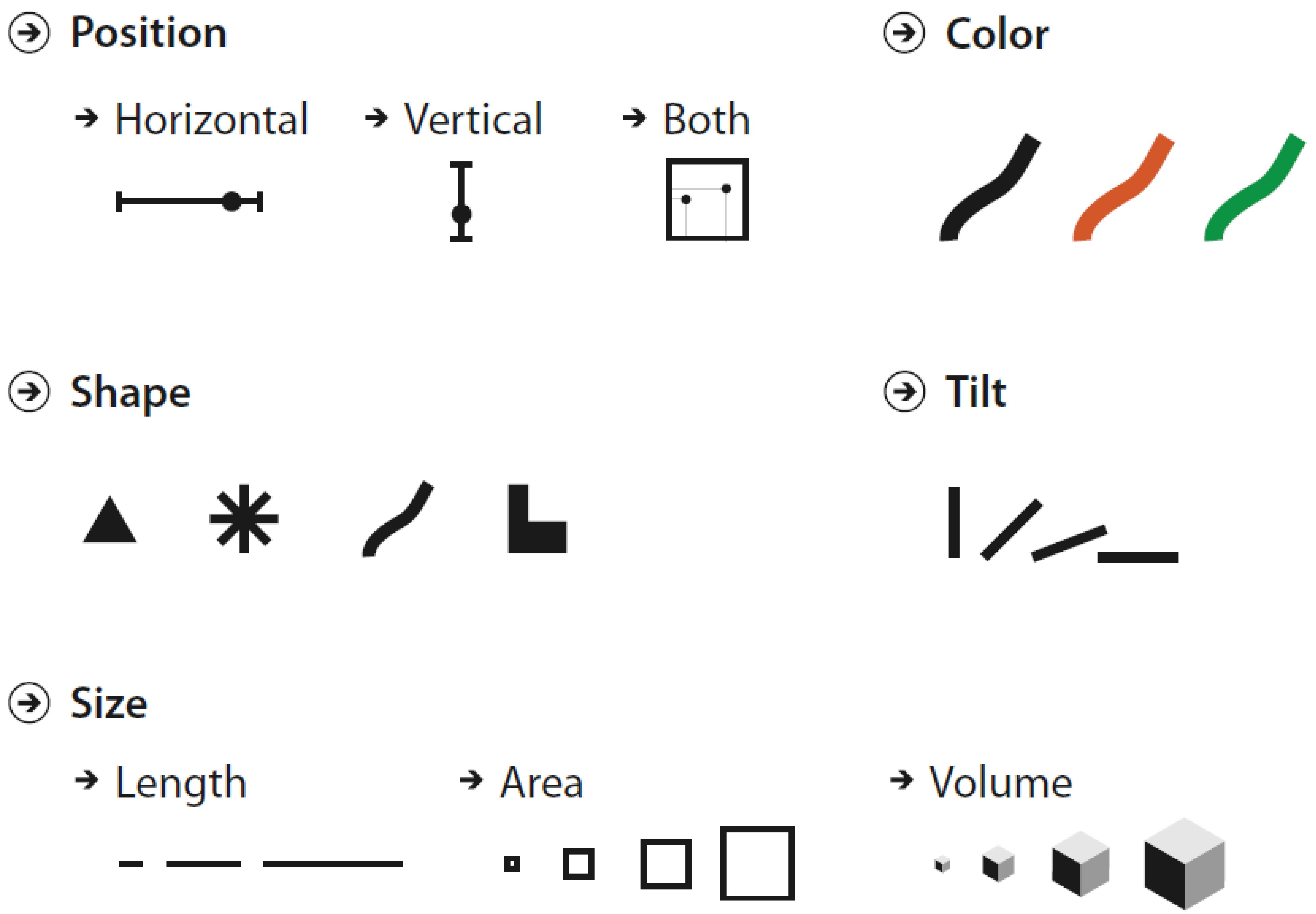
标志与通道

Definitions: Marks and channels

- marks
  - geometric primitives

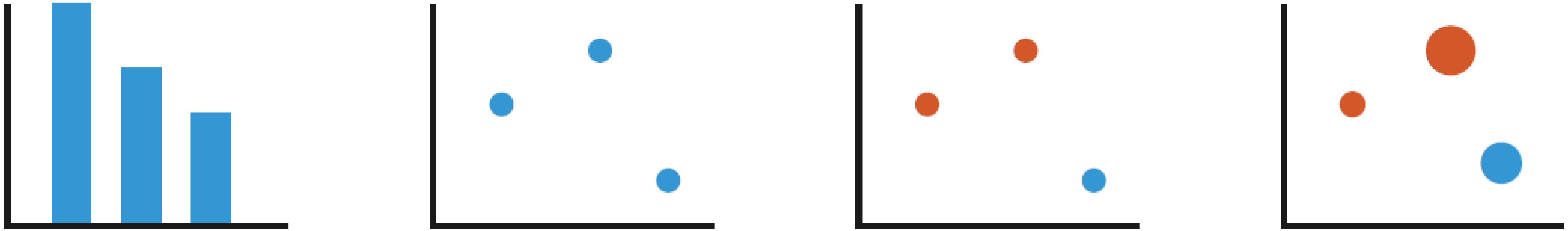


- channels
  - –control appearance of marks
  - –can redundantly code with multiple channels

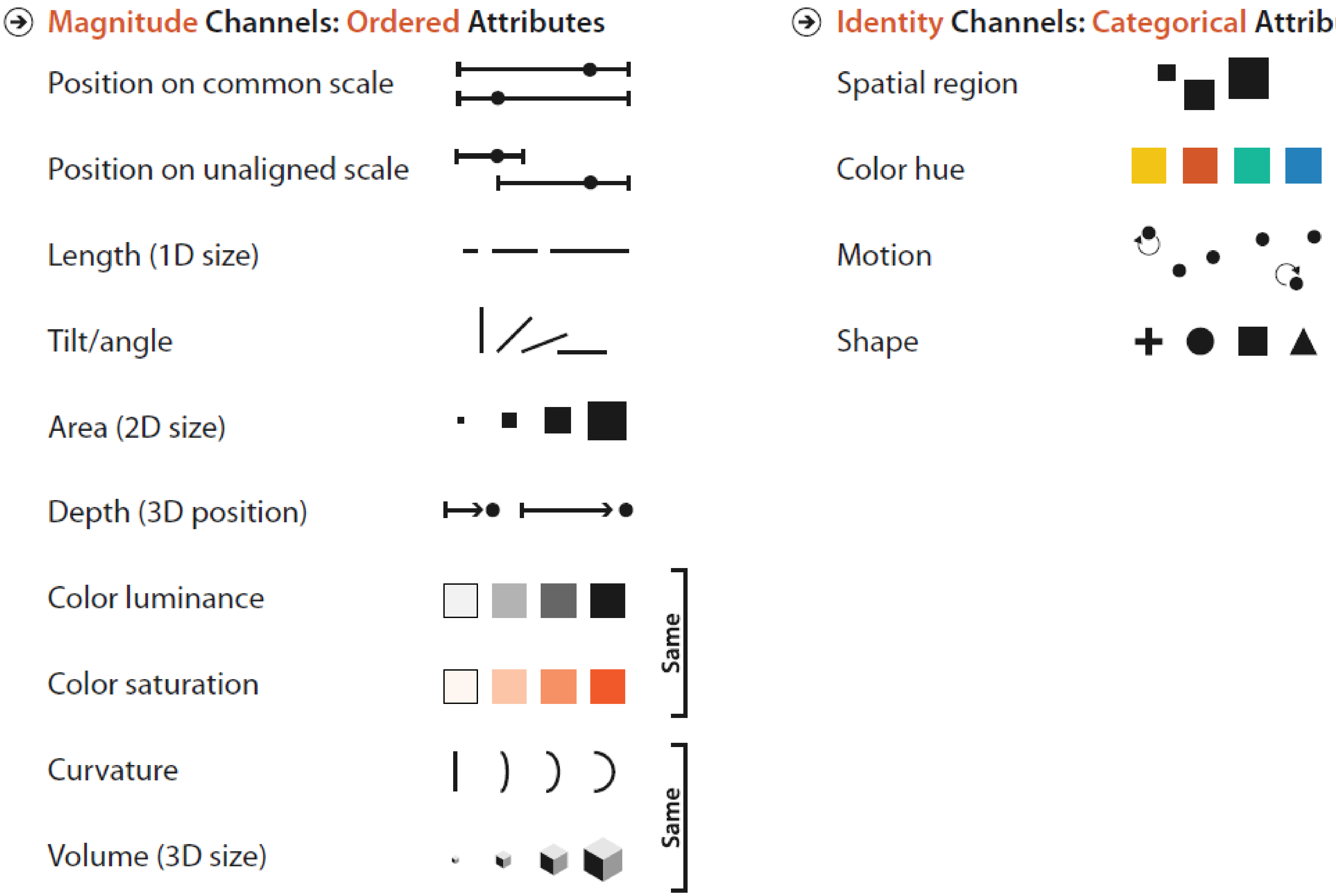


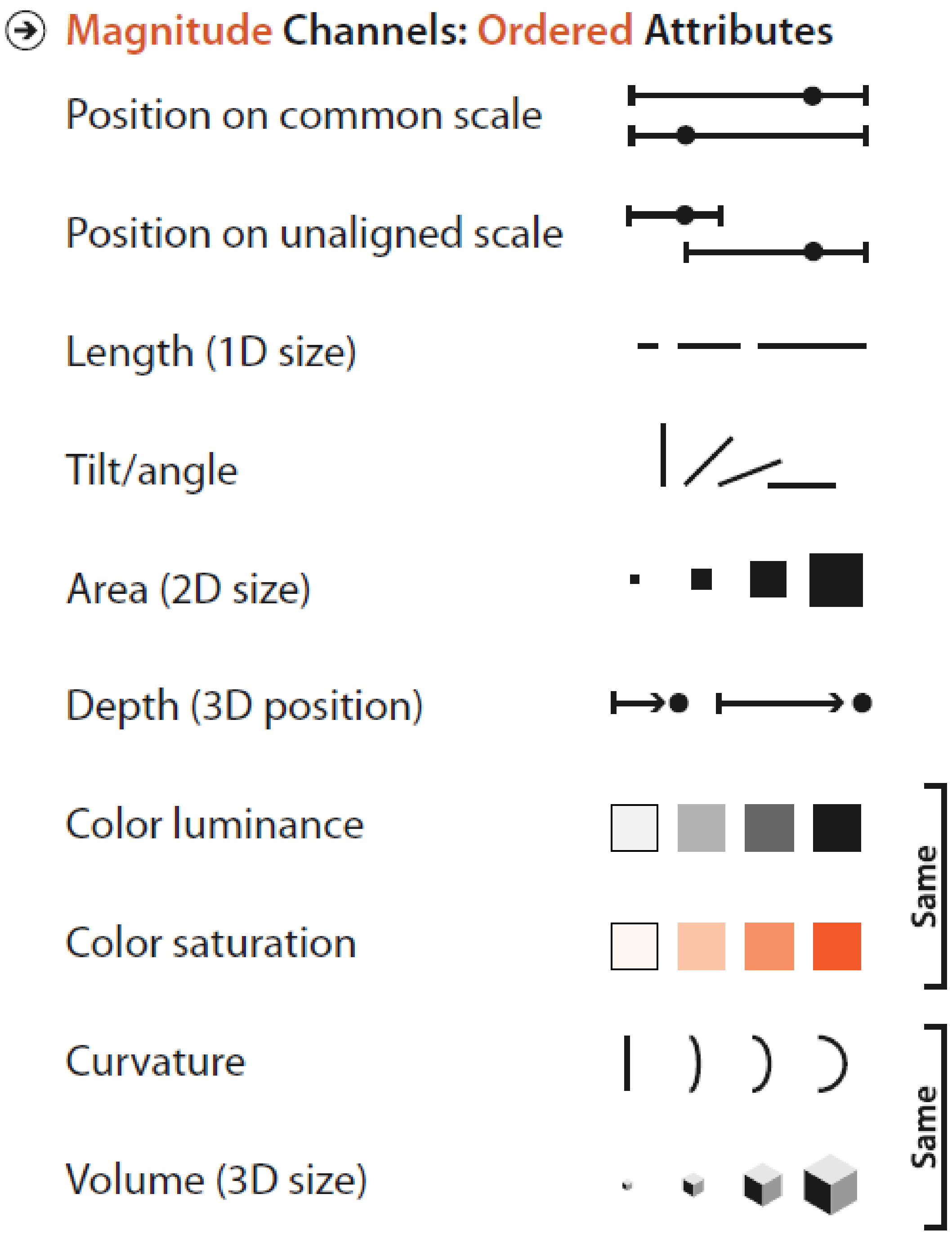
Visual encoding

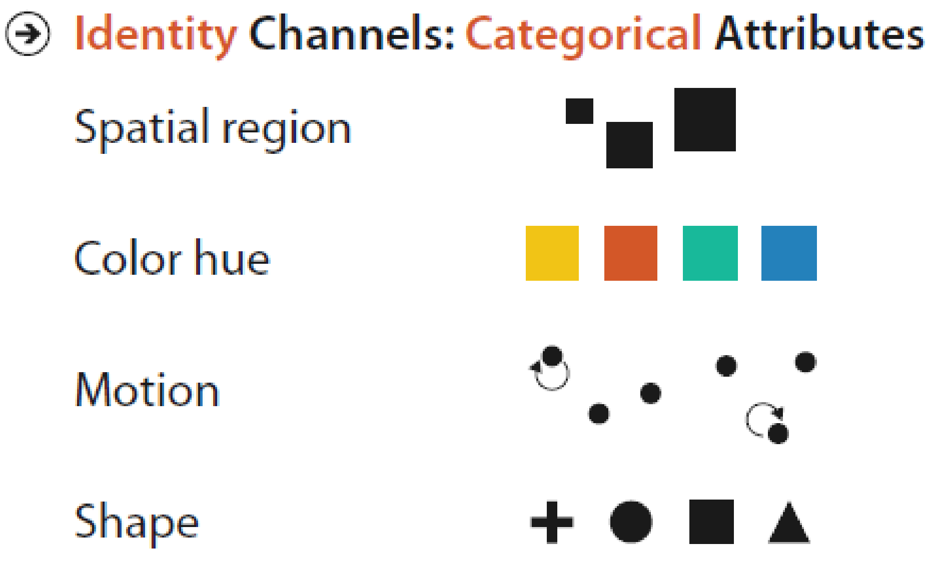
- analyze idiom structure
  - as combination of marks and channels



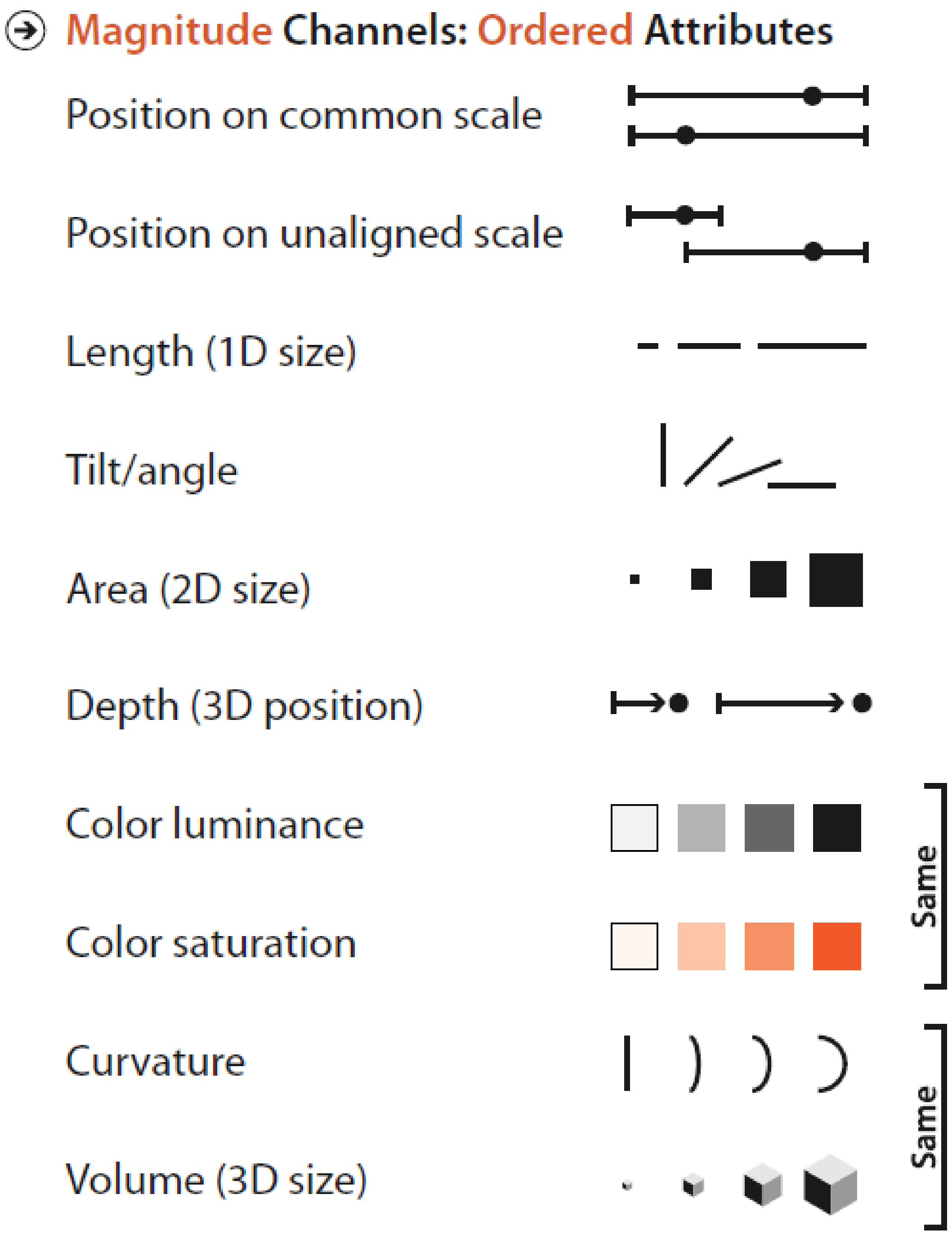
(a)	(b)	(c)	(d)
vertical position	vertical position horizontal position	vertical position horizontal position color hue	vertical position horizontal position color hue size (area)
mark: line	mark: point	mark: point	mark: point

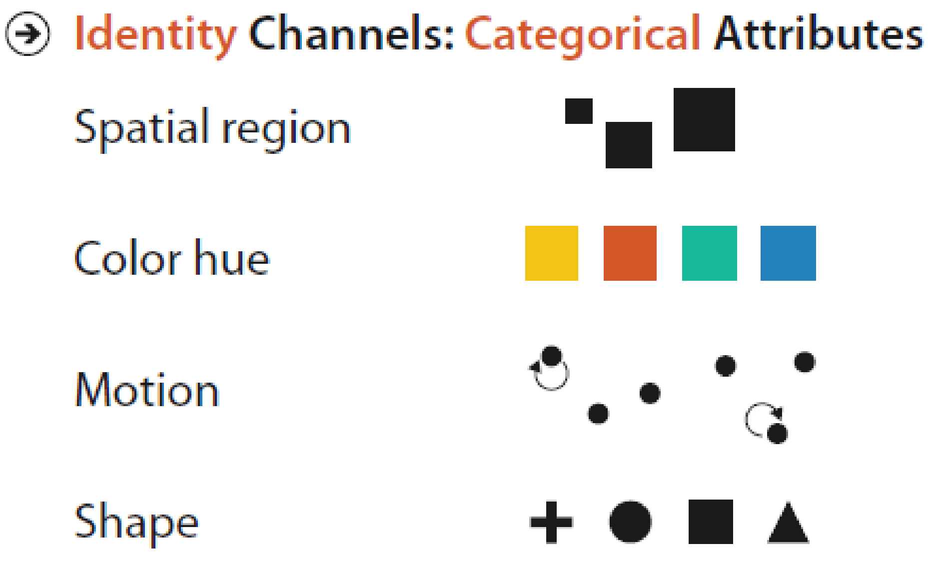

Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape



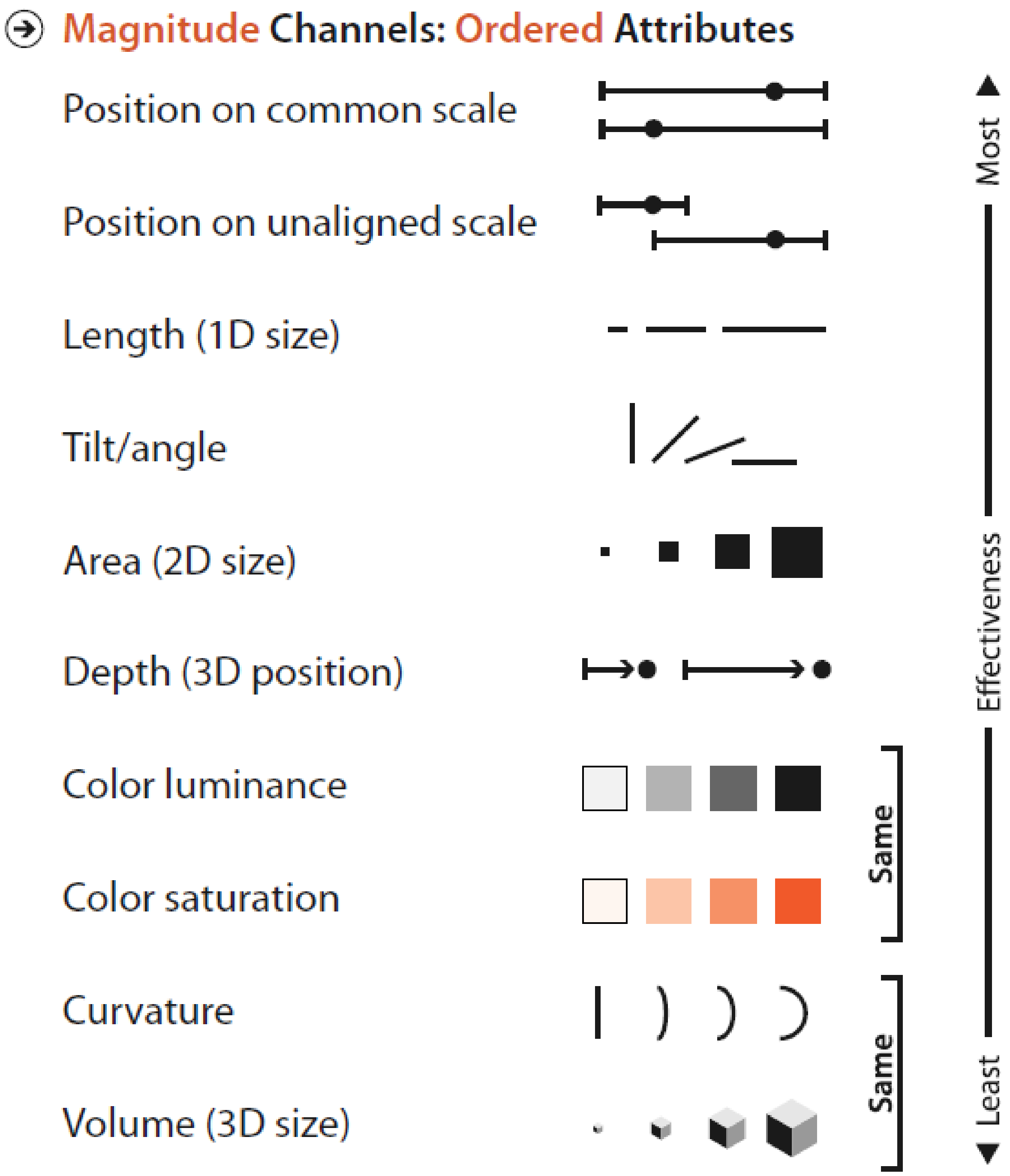


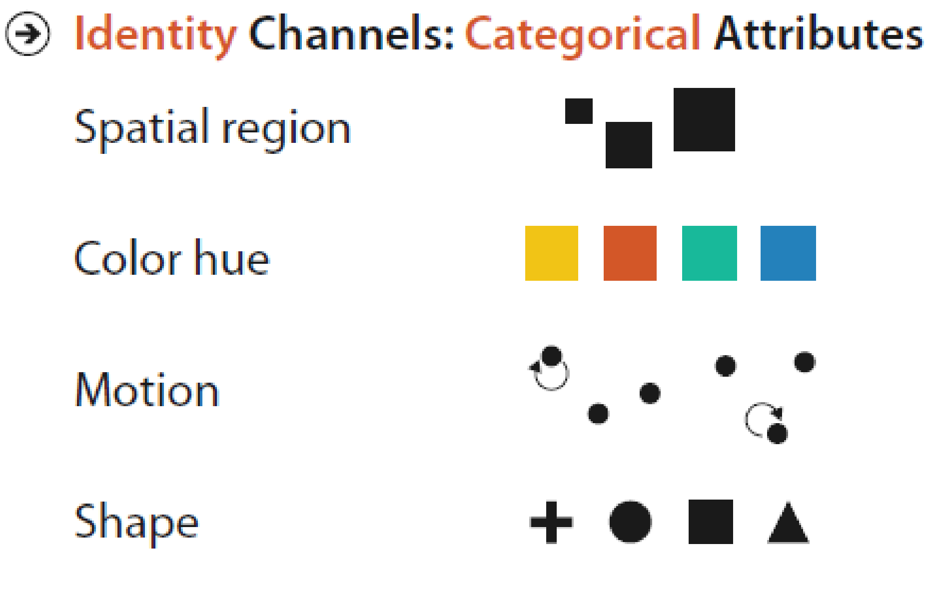
- expressiveness principle
  - **match** channel and data characteristics





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

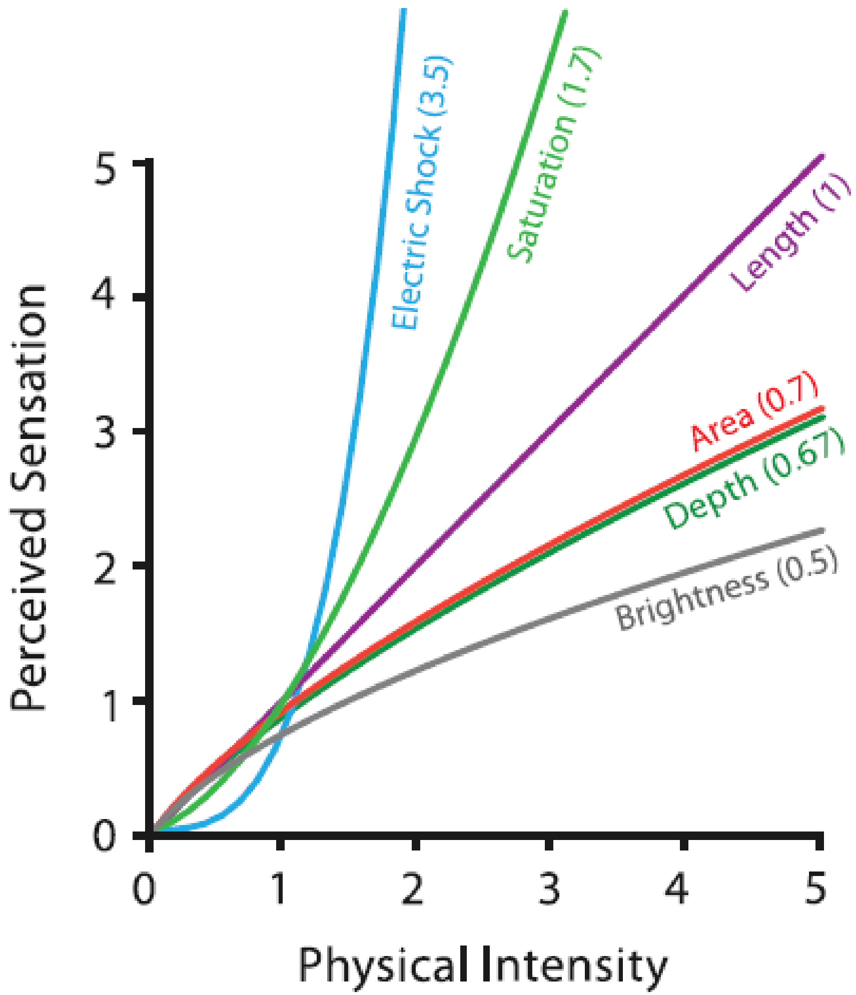


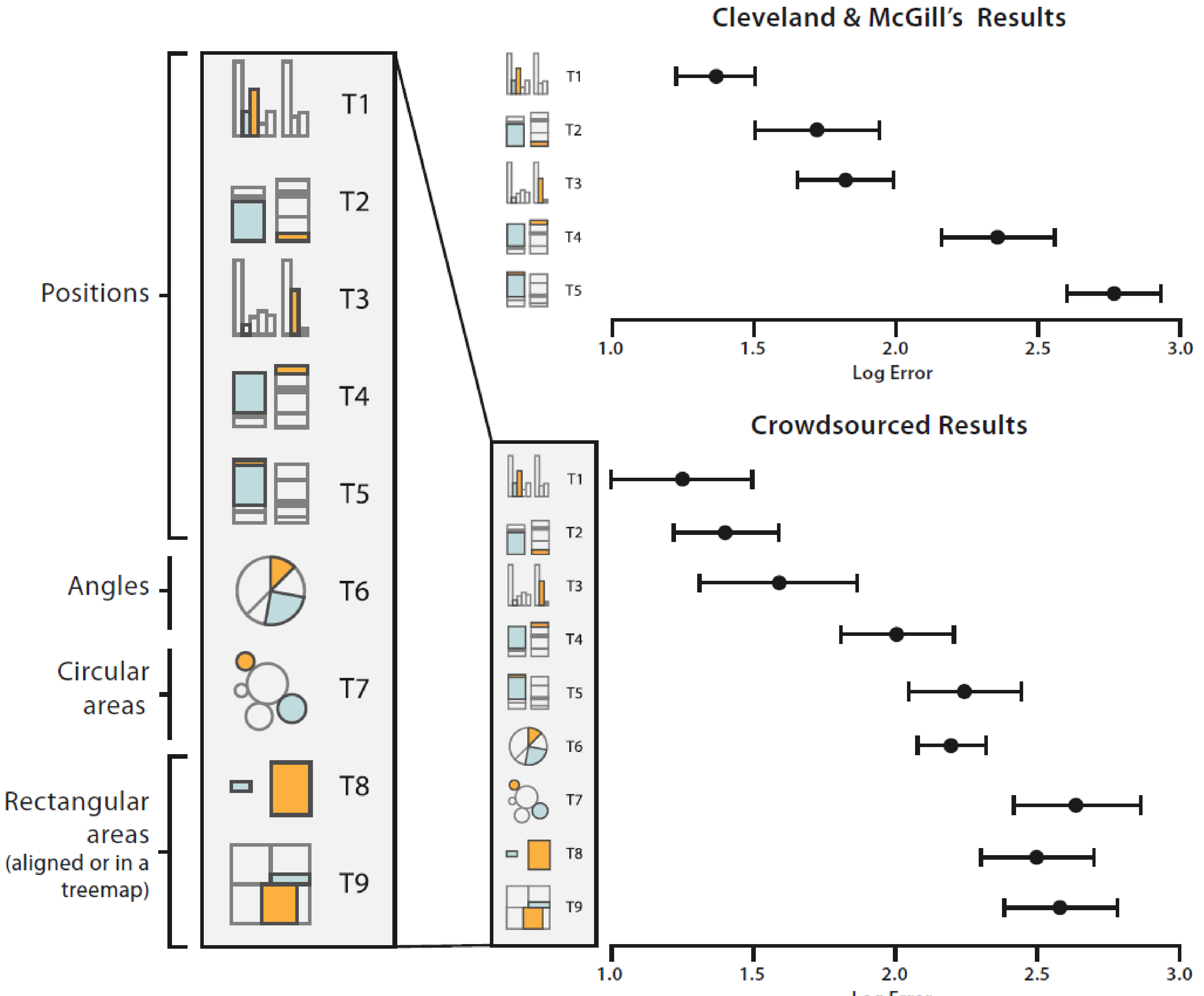


- expressiveness principle
  - match channel and data characteristics
- effectiveness principle
  - encode most important attributes with highest ranked channels
  - spatial position ranks high for both
- distinguishability
  - **enough levels** in channel to match data

Accuracy: Fundamental Theory

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

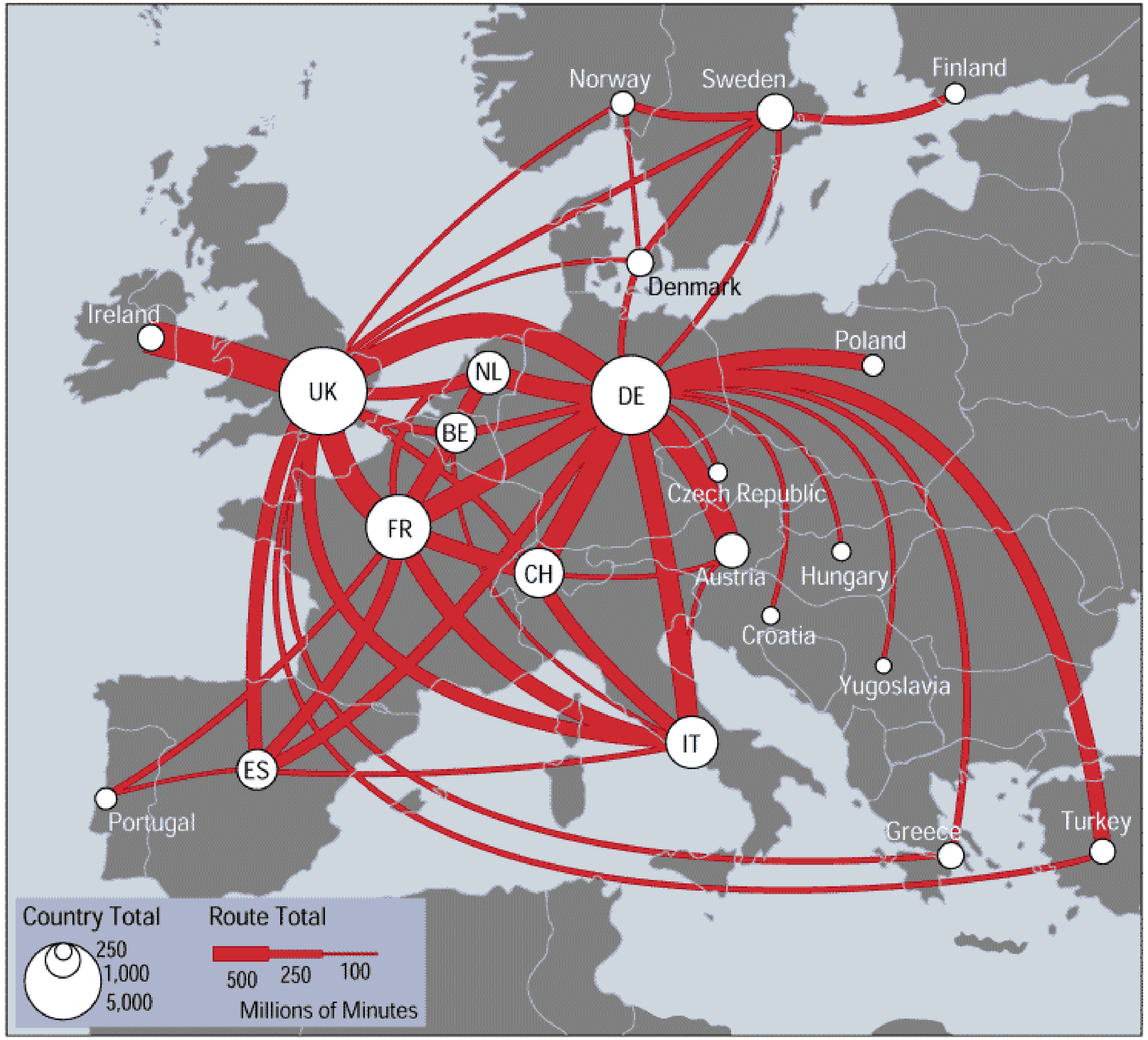




[Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. Heer and Bostock. Proc ACM Conf. Human Factors in Computing Systems (CHI) 2010, p. 203–212.]

Discriminability: How many usable steps?

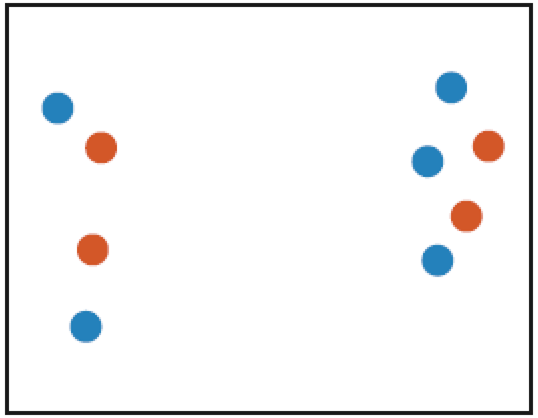
- must be sufficient for number of attribute levels to show
  - linewidth: few bins



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

Separability vs. Integrality

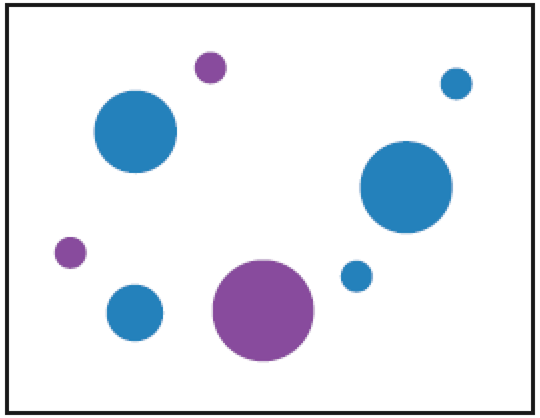
Position
+ Hue (Color)



Fully separable

2 groups each

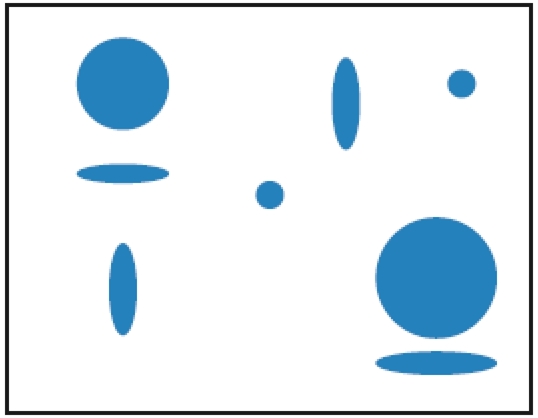
Size
+ Hue (Color)



Some interference

2 groups each

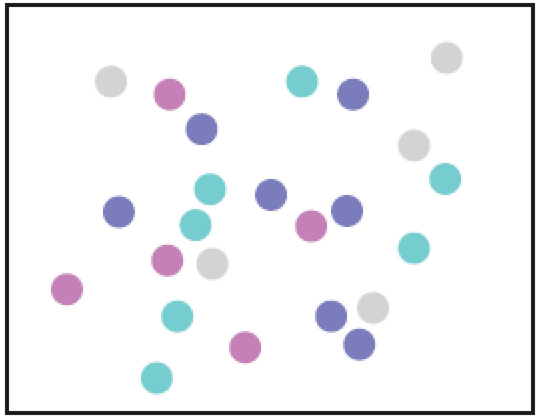
Width
+ Height



Some/significant interference

3 groups total: integral area

Red
+ Green

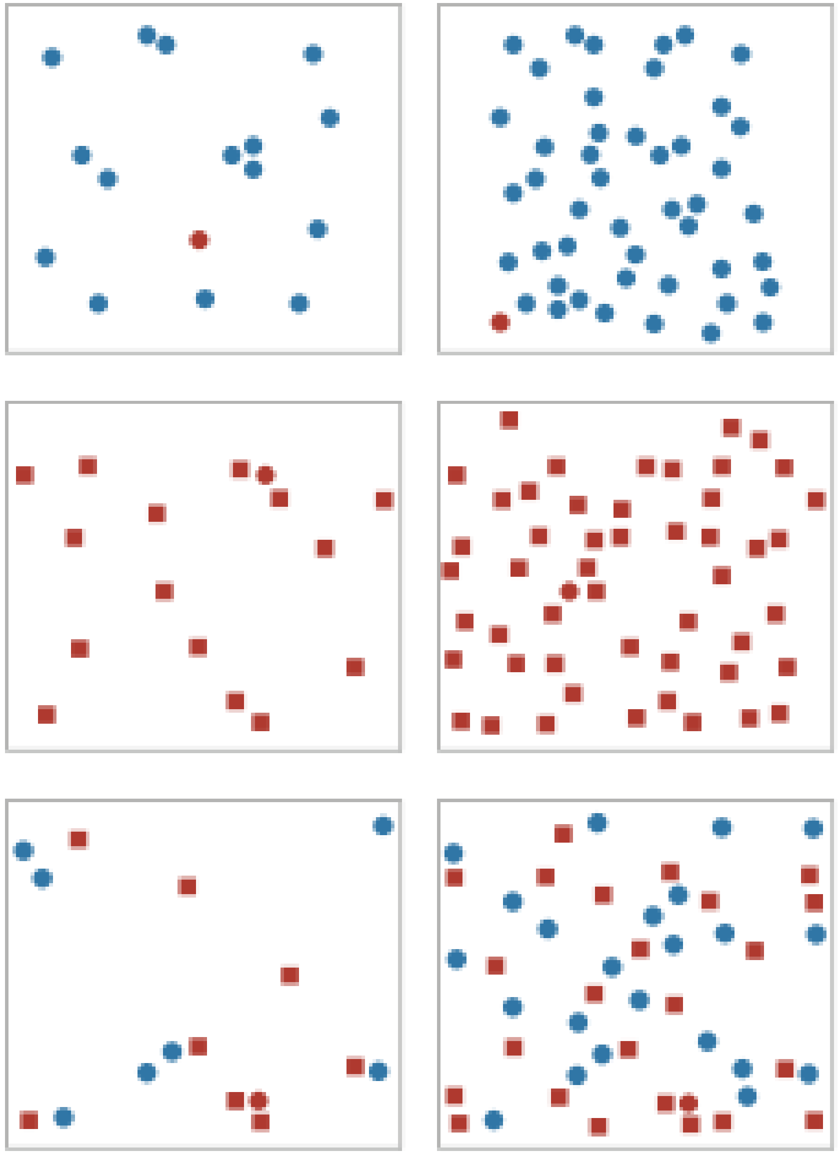


Major interference

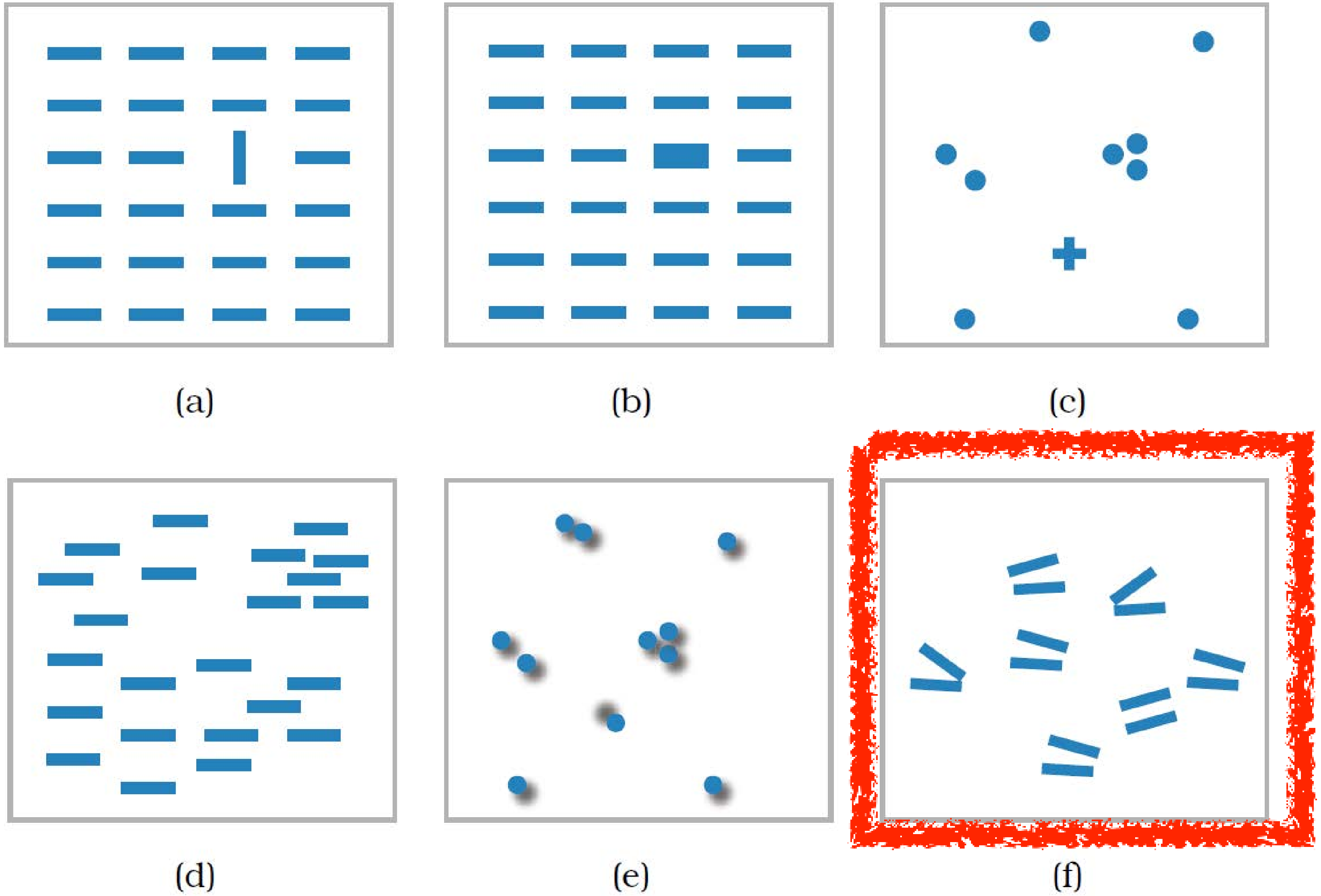
4 groups total: integral hue

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

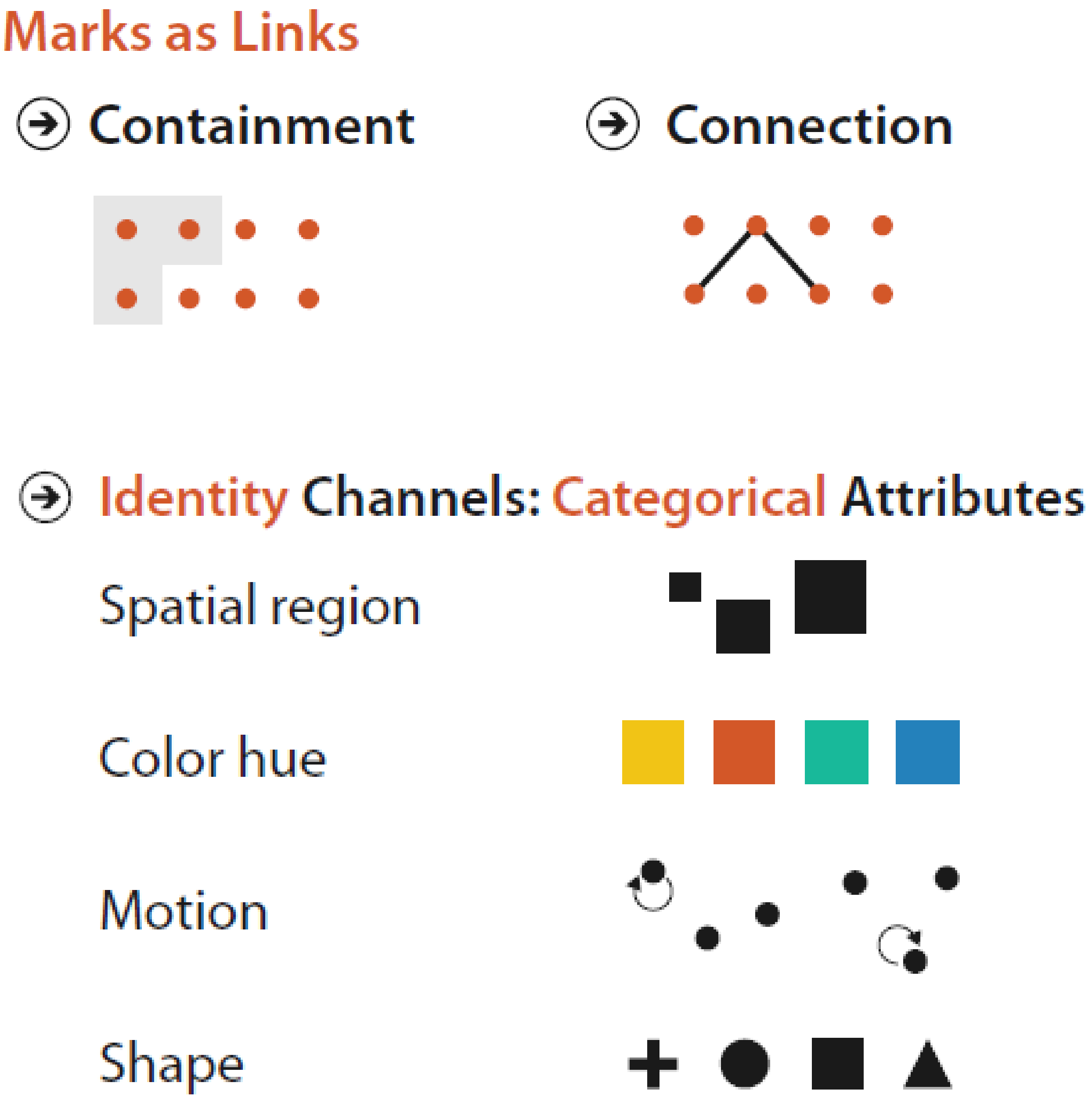


- many channels: tilt, size, shape, proximity, shadow direction, ...
- but not all! parallel line pairs do not pop out from tilted pairs

Grouping

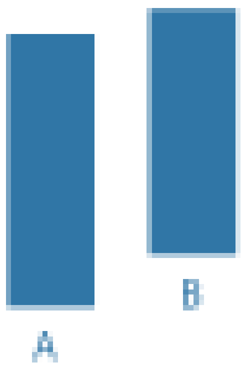
- containment
- connection

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

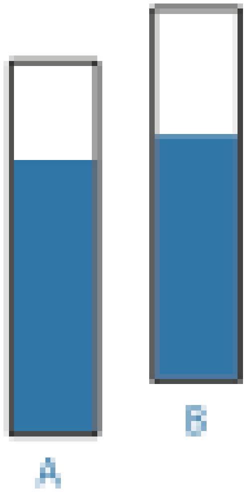


Relative vs. absolute judgements

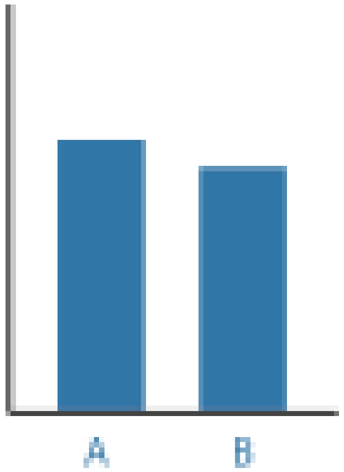
- perceptual system mostly operates with relative judgements, not absolute
  - that’s why accuracy increases with common frame/scale and alignment
  - Weber’s Law: ratio of increment to background is constant
    - filled rectangles differ in length by 1:9, difficult judgement
    - white rectangles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



length



position along unaligned common scale

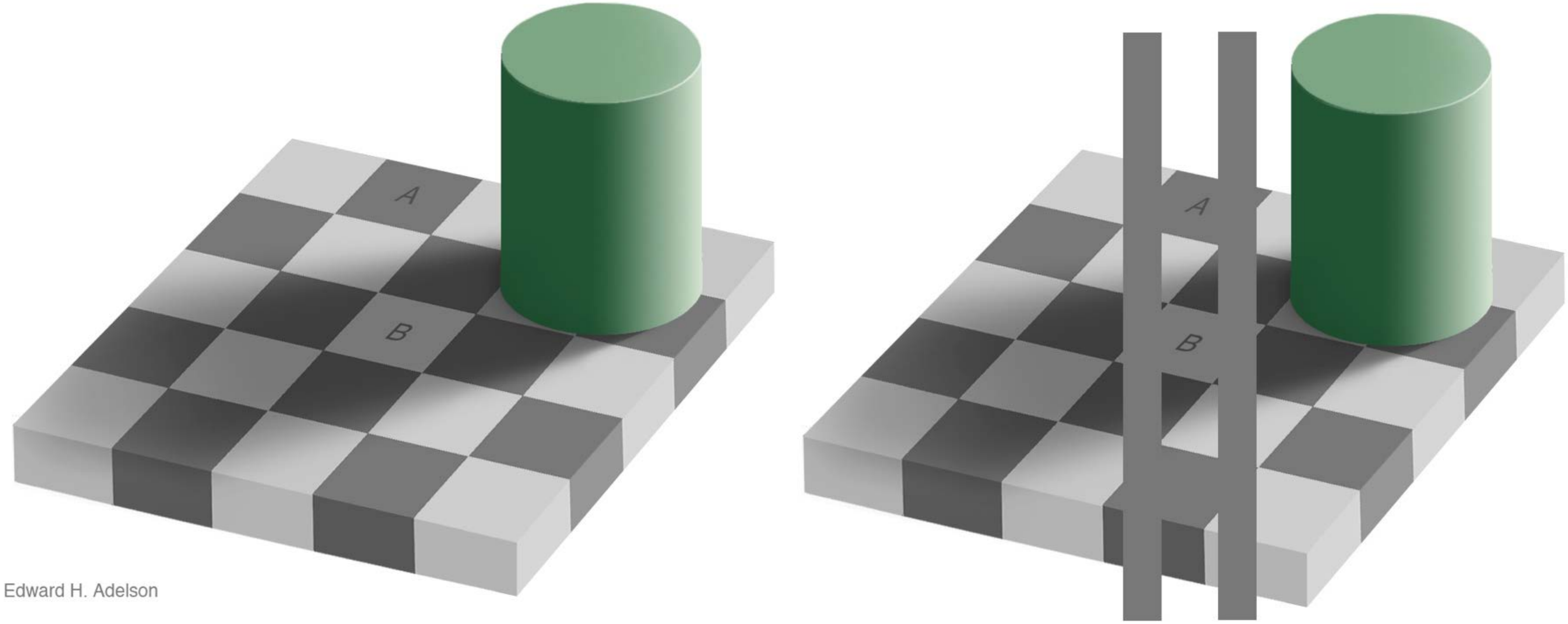


position along aligned scale

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

Relative luminance judgements

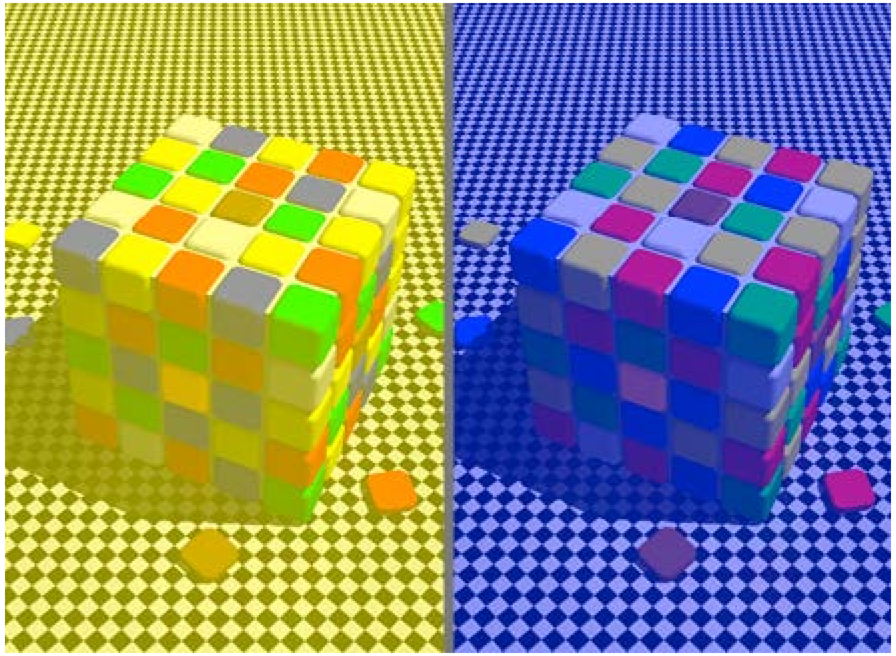
- perception of luminance is contextual based on contrast with surroundings

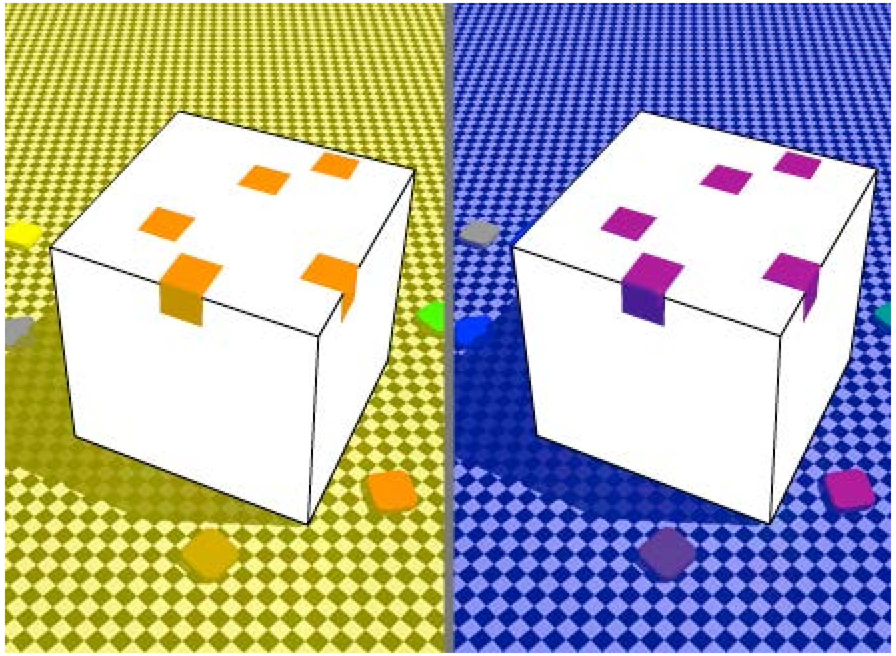


http://persci.mit.edu/gallery/checkershadow

Relative color judgements

- color constancy across broad range of illumination conditions





http://www.purveslab.net/seeforyourself/

参考

- **书籍网站：**[Visualization Analysis and Design](https://www.cs.ubc.ca/~tmm/vadbook/)
  https://www.cs.ubc.ca/~tmm/vadbook/
- **作者：**[Tamara Munzner](https://www.cs.ubc.ca/~tmm/)

  https://www.cs.ubc.ca/~tmm/