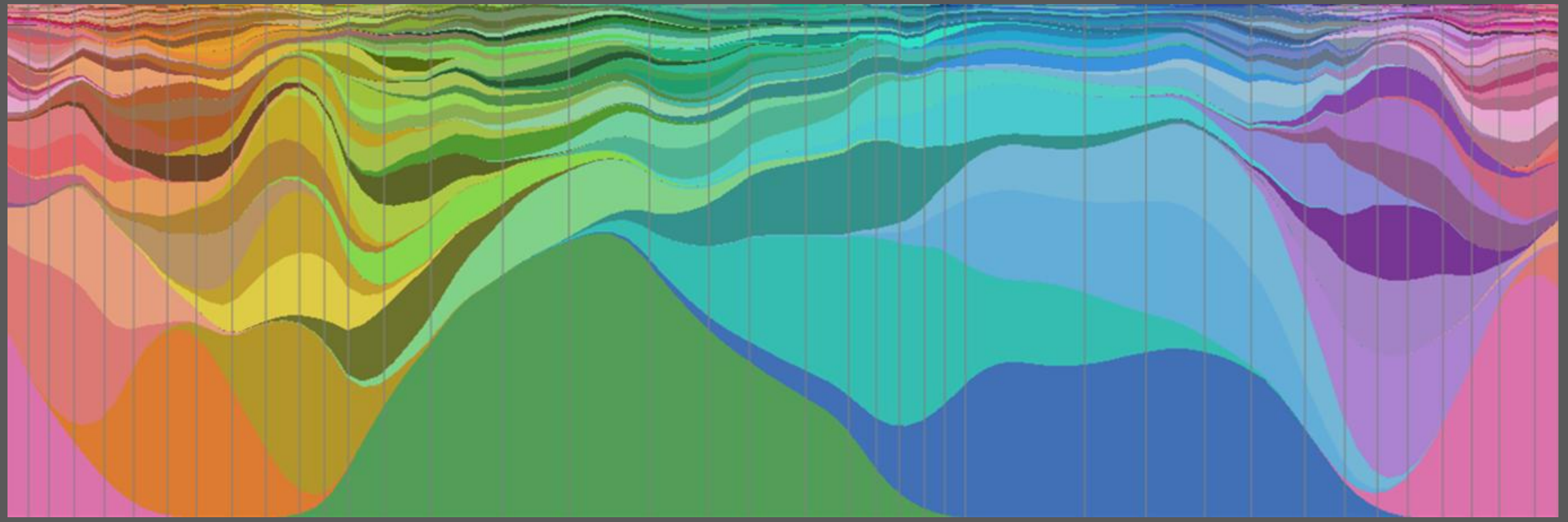
LDSCI5209 - Information Presentation and Visualisation

Week 3: Data Abstraction



Dimitris Mylonas

Northeastern University London

Recap: Mapping data to visual variables

Assign **data fields** (e.g., with *Nominal, Ordinal, Quantitative* types) to **visual channels** (x, y, colour, shape, size, ...) for a chosen **graphical mark** type (point, bar, line, ...).

Additional concerns include choosing appropriate **encoding parameters** (log scale, sorting, ...) and **data transformations** (bin, group, aggregate, ...).

Plan for today

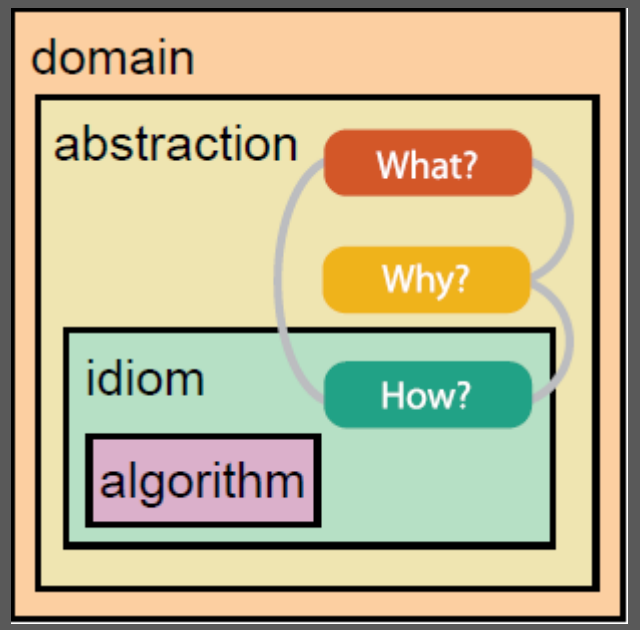
- Data and task abstraction
- Data types
- Tables

Data and task abstraction

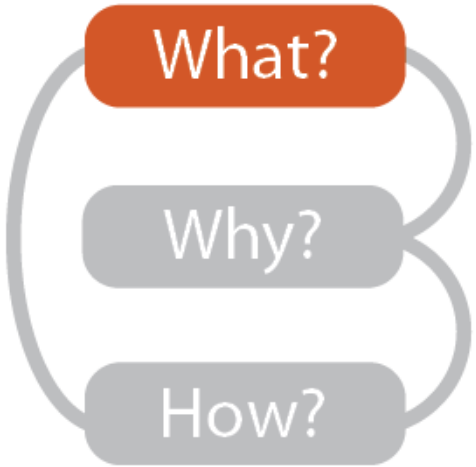
What, why, how?

Nested model of visualisation design

- Domain situation
  - who are the target users?
- Abstraction
  translate from specifics of domain to vocabulary of visualization
  - **what** is shown? **data abstraction**
  - **why** is the user looking at it? **task abstraction**
- Idiom
  how is it shown?
  - **visual encoding** idiom: how to draw
  - **interaction** idiom: how to manipulate
- Algorithm
  efficient computation



Munzner, 2009



What?

Datasets

Data Types

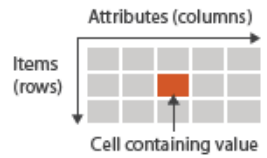
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

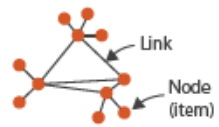
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

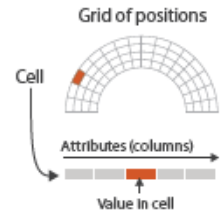
→ Tables



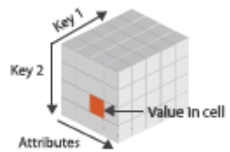
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

→ Geometry (Spatial)

Dataset Availability

→ Static

→ Dynamic

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

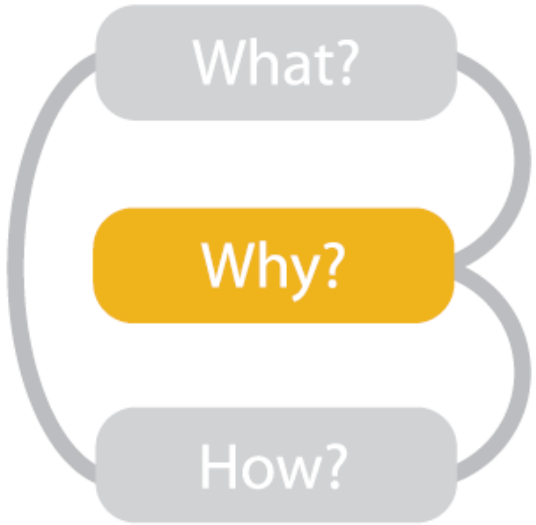
→ *Quantitative*

Ordering Direction

→ Sequential

→ Diverging

→ Cyclic



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

Why?

Actions

Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*

→ Produce

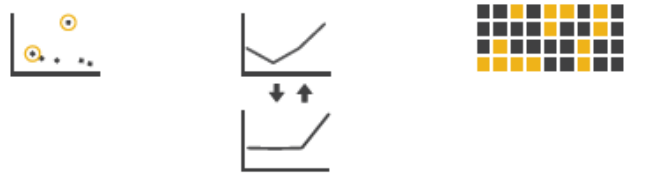
→ *Annotate* → *Record* → *Derive*

Search

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

Query

→ Identify → Compare → Summarize



Targets

All Data

→ Trends → Outliers → Features

Attributes

→ One → Many

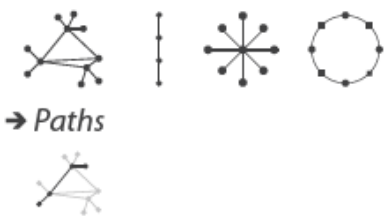
→ *Distribution* → *Dependency* → *Correlation* → *Similarity*

→ *Extremes*

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

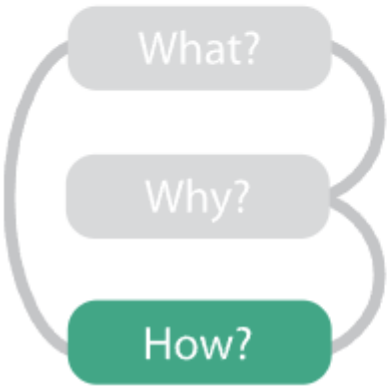
- Change
- Select
- Navigate

Facet

- Juxtapose
- Partition
- Superimpose

Reduce

- Filter
- Aggregate
- Embed



Operations of data abstraction

- Translate domain-specific language to generic visualisation language
- Identify dataset type(s), attribute types
- Identify cardinality
    - how many items in the dataset?
    - what is cardinality of each attribute?
        - number of levels for categorical data
        - range for quantitative data
- Consider whether to transform data
    - guided by understanding of task

Derived attributes

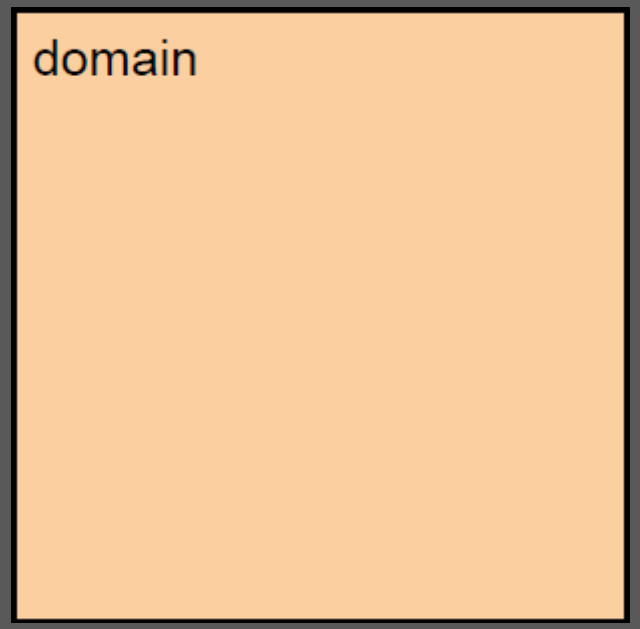
- Derived attribute: compute from originals

Original data

Derived data

Domain

- Details of an application domain
- Group of users, target domain & their data
  - varies wildly by domain
  - must be specific enough to get traction
- Domain questions/problems
  - break down into simpler abstract tasks

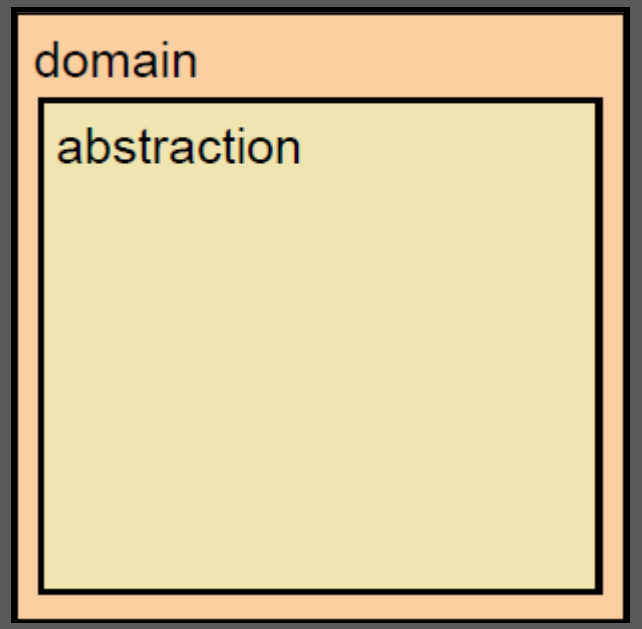


e.g., Find good movies?

- Identify movies in genres I like
- Domain: general population, movie enthusiasts

Task Abstraction

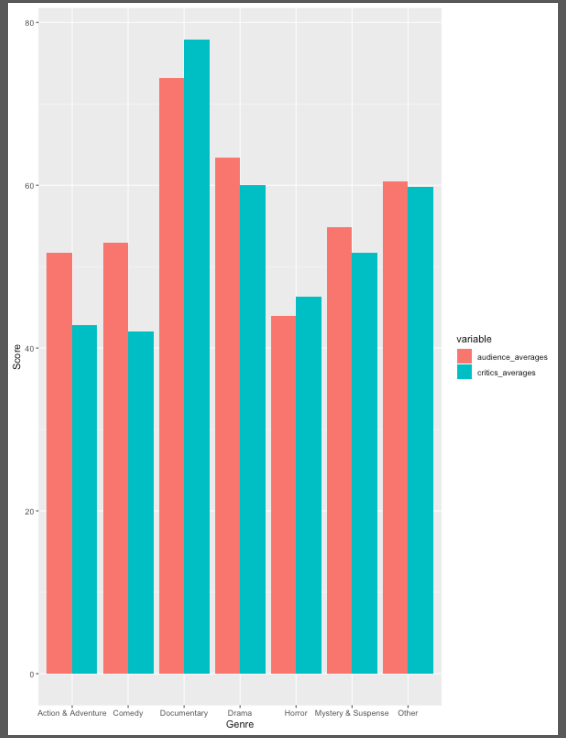
- map **what** and **why** into generalised terms
  - identify tasks that users wish to perform, or already do
  - find data types that will support those tasks
- possibly transform /derive if need be



Example of domain abstraction

Find good movies

- identify good movies in genres I like
- Domain: general population, movie enthusiasts
- Task: what is a good movie for me?
  - highly rated by critics?
  - highly rated by audiences?
  - successful at the box office?
  - similar to movies I liked?
  - matches specific genres?
- data: (is it available?)
  - yes! data sources IMDB, Rotten Tomatoes...
- How? e.g. stacked bar chart for audience and critic ratings



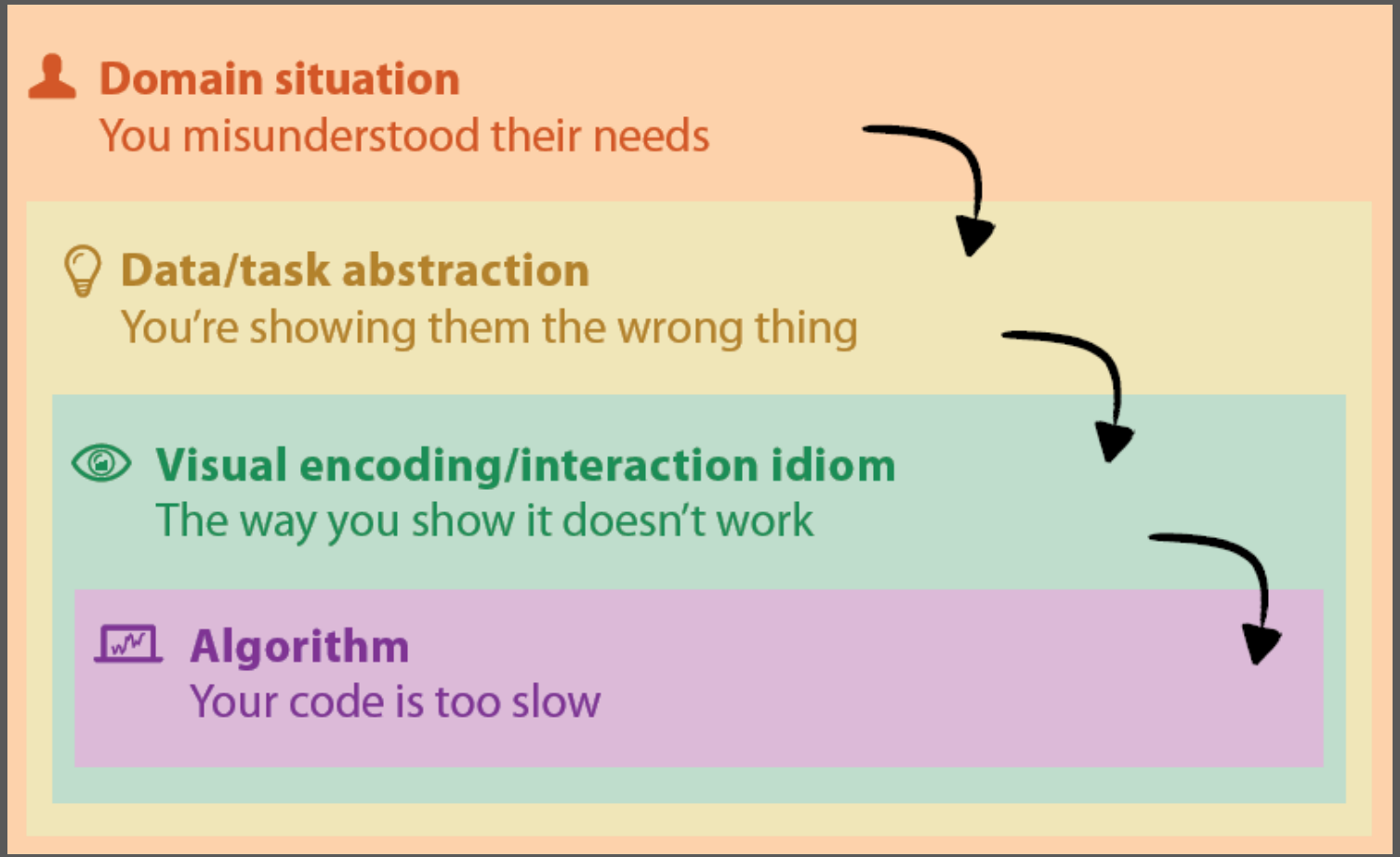
https://ucladatares.medium.com/movie-ratings-analysis-478c0de6c9f8

Analytic task taxonomy

- **Retrieve Value** : How long is the movie Gone with the Wind?
- **Filter**: What comedies have won awards?
- **Compute Derived Value**: How many awards have MGM studio won in total?
- **Find Extremum** : What director/film has won the most awards?
- **Sort** : Rank movies by most number of awards.
- **Determine Range** : What is the range of film lengths?
- **Characterize Distribution** : What is the age distribution of actors?
- **Find Anomalies** : Are there exceptions to the relationships?
- **Cluster** : Is there a cluster of typical film lengths?
- **Correlate** : Is there a trend of increasing film length over the years?

Amar et al., 2005

Validity threats at each level

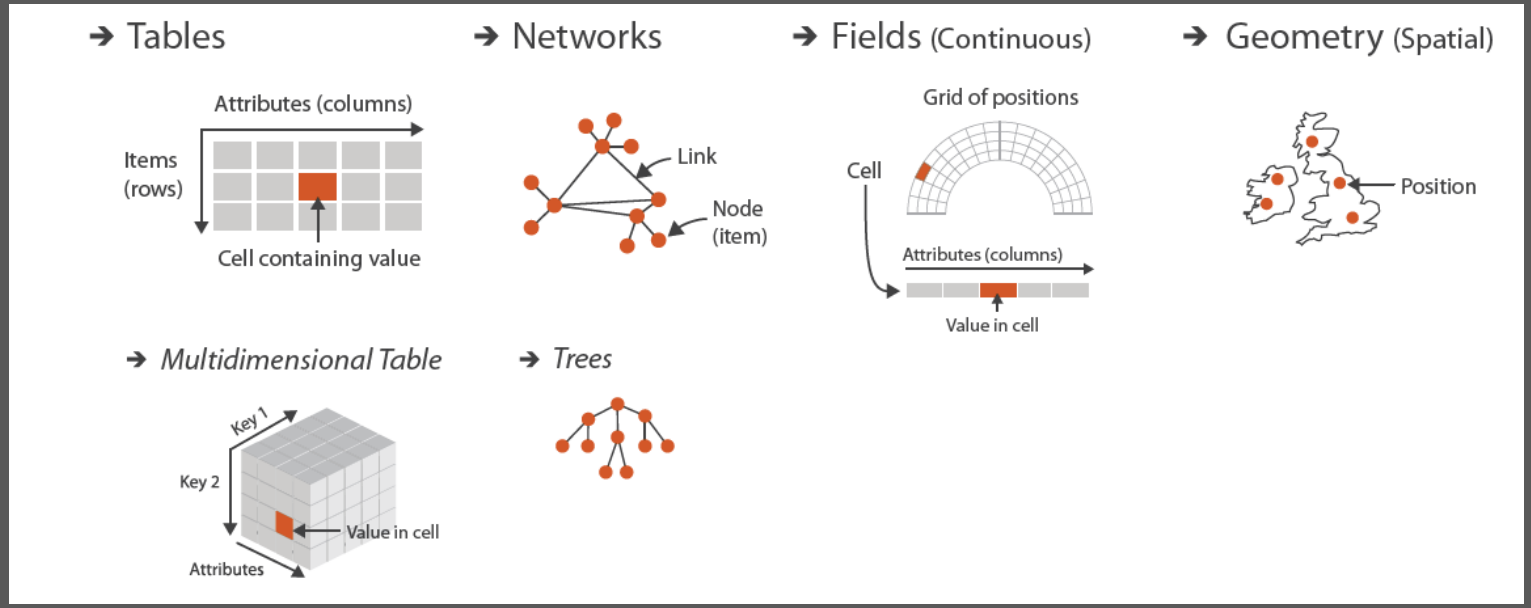


Data types

nD Tables, Networks, Fields, Geometry

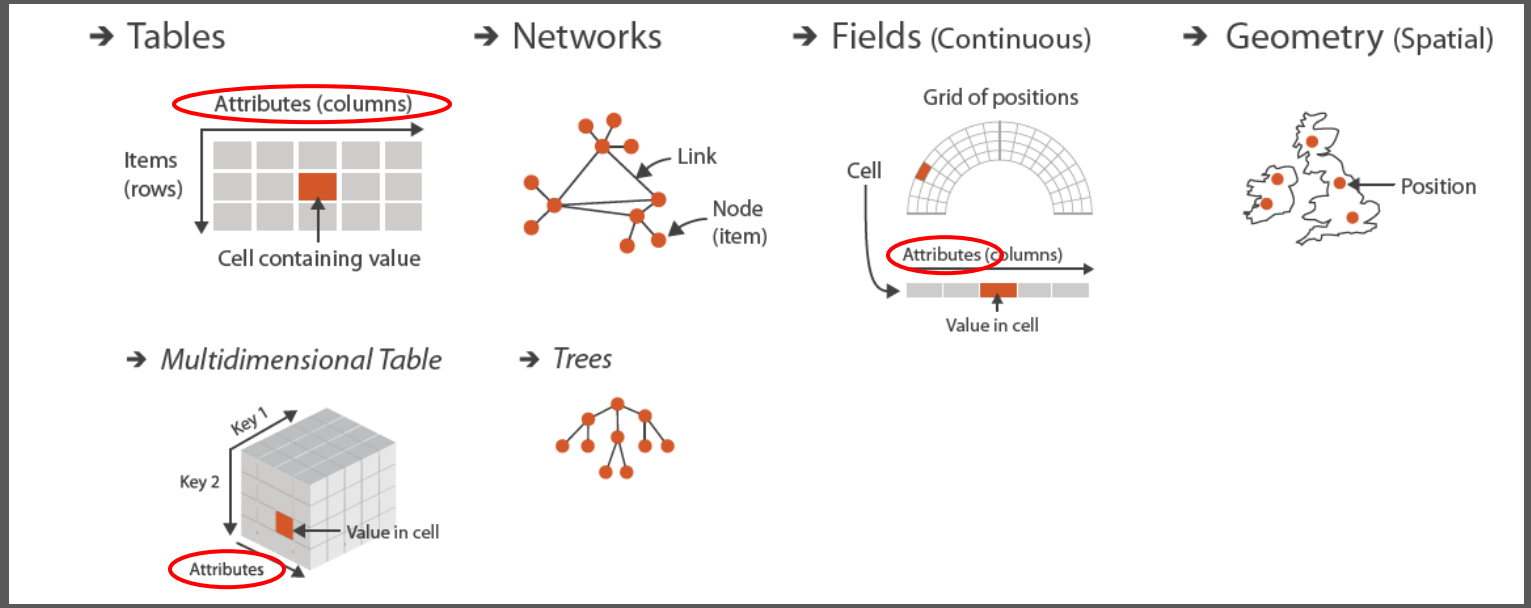
Dataset types

- DATASET = collection of information to be analysed
  - Tables, networks, trees fields, geometry
- TYPE = structural or mathematical interpretation of the data
  - Items, attributes, links, positions, grids

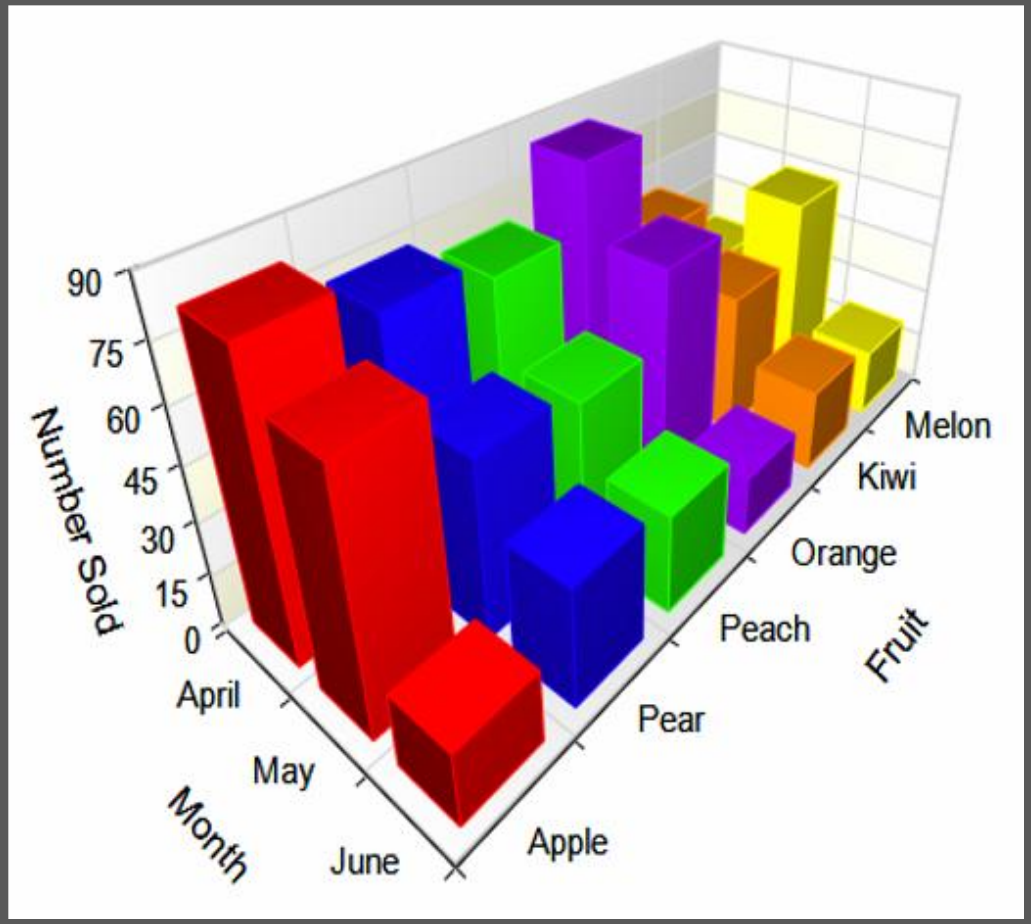


Dataset types

- DATASET = collection of information to be analysed
  - Tables, networks, trees fields, geometry
- TYPE = structural or mathematical interpretation of the data
  - Items, attributes, links, positions, grids

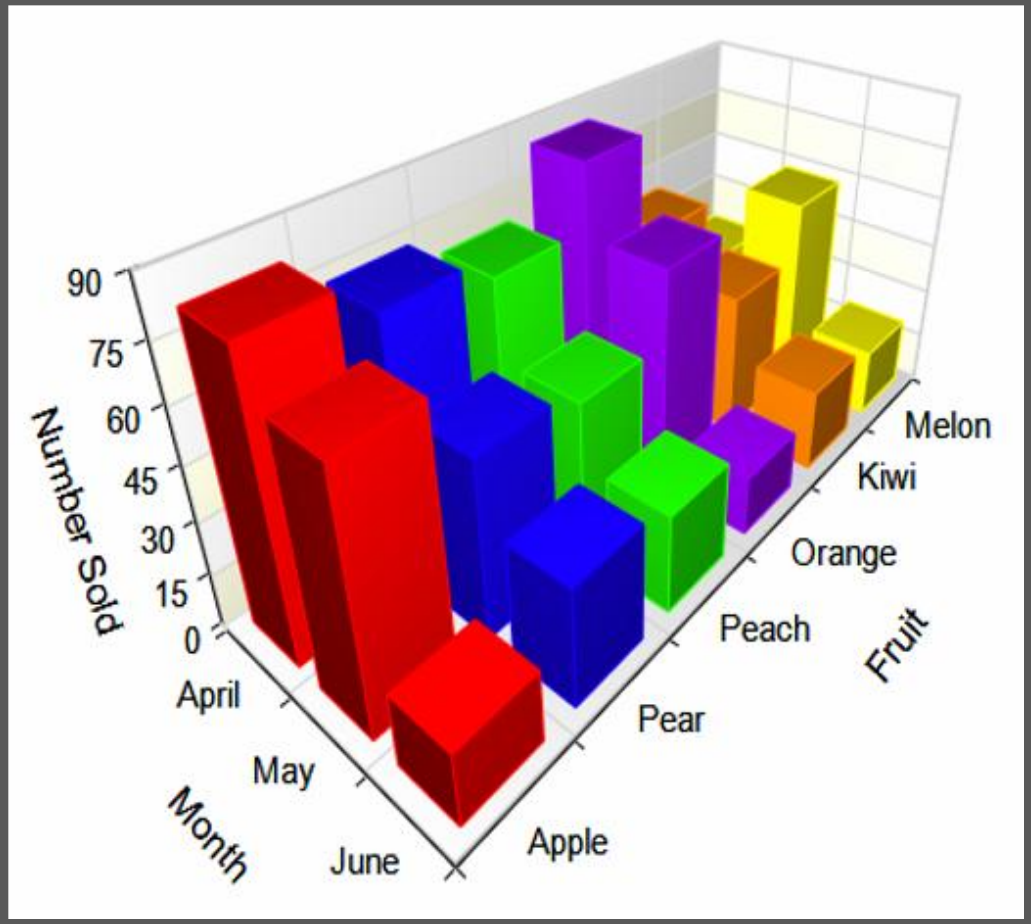


Attribute types: how many fruits?



Attribute types: how many fruits?

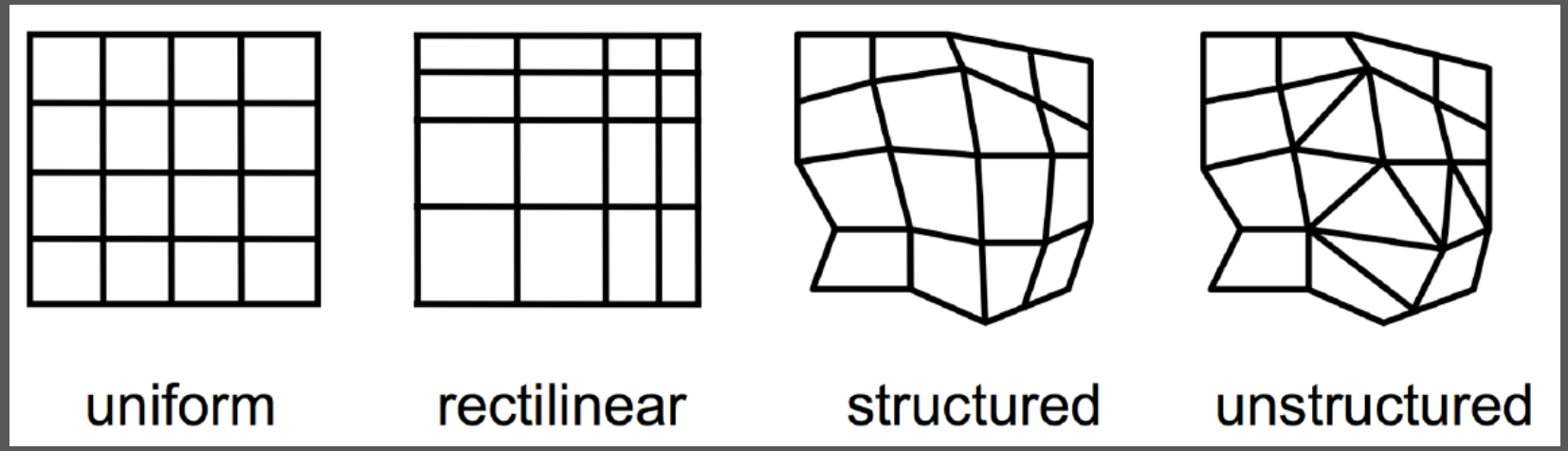
Quantitative



Categorical

Ordinal

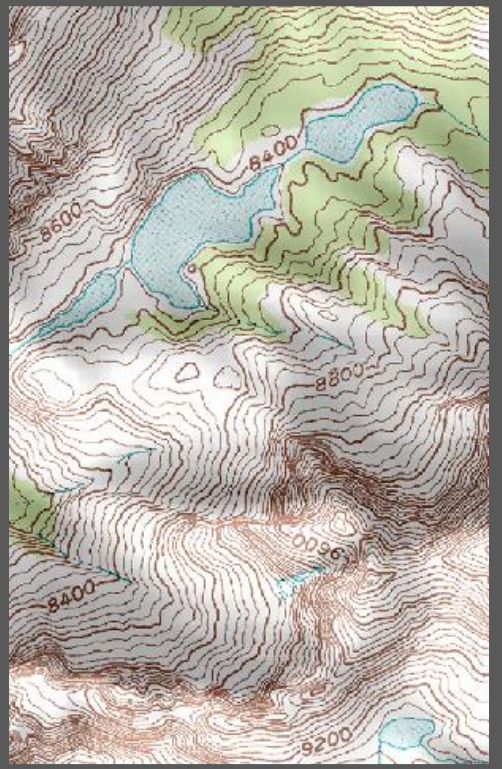
Grid types



https://en.wikipedia.org/wiki/Voronoi_diagram

Geometry

- shape of items
- explicit spatial positions
- points, lines, curves, surfaces, regions
  - (volumes outside scope of class)
- boundary between graphics and visualisation
  - graphics: geometry taken as given
  - vis: geometry is result of a design decision



Tables

1D, 2D, 3D, nD

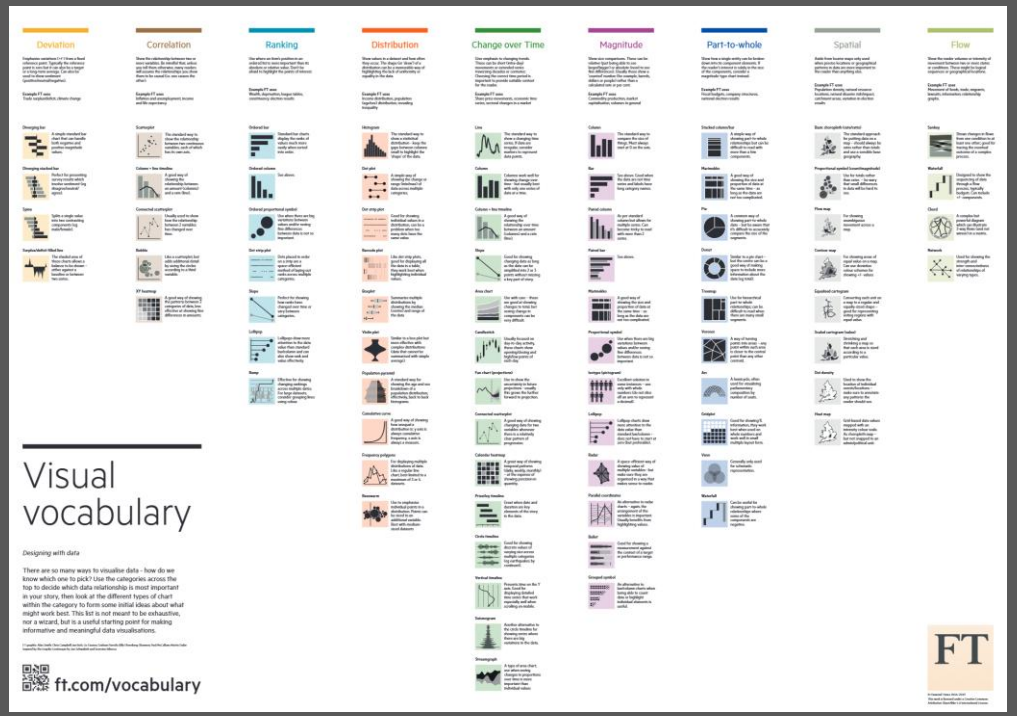
Working with tables

	Age	Gender	Height
Bob	19	M	176
Alice	25	F	168
Chris	26	M	185
Dan	22	M	191

- homogeneity
  - same data type? same scales?
- need different approaches based on scale
  - how many attributes?
    - up to ~50: tractable with direct visual encoding
    - thousands: need transformations / analytical methods
  - how many items?
    - up to 1K: tractable with direct visual encoding
    - >> 10K: need transformations / analytical methods

Tasks and techniques

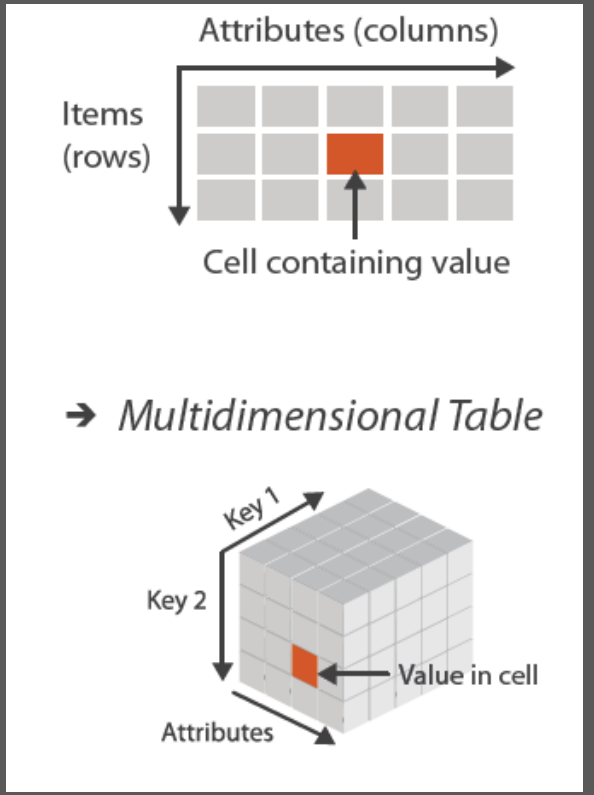
- Deviation
- Correlation
- Ranking
- Distribution
- Change-over-time
- Magnitude
- Part-to-whole
- Spatial
- Flow



https://github.com/Financial-Times/chart-doctor/blob/main/visual-vocabulary/Visual-vocabulary-en.pdf
https://gramener.github.io/visual-vocabulary-vega/#

Keys and values

- Keys
  - independent attribute
  - used as unique index to look up items
  - simple tables: 1 key
  - multidimensional tables: multiple keys
- Values
  - dependent attribute, value of cell
- classify arrangements by key count
  - 0, 1, 2, many...



0 Keys: Express values (magnitudes)

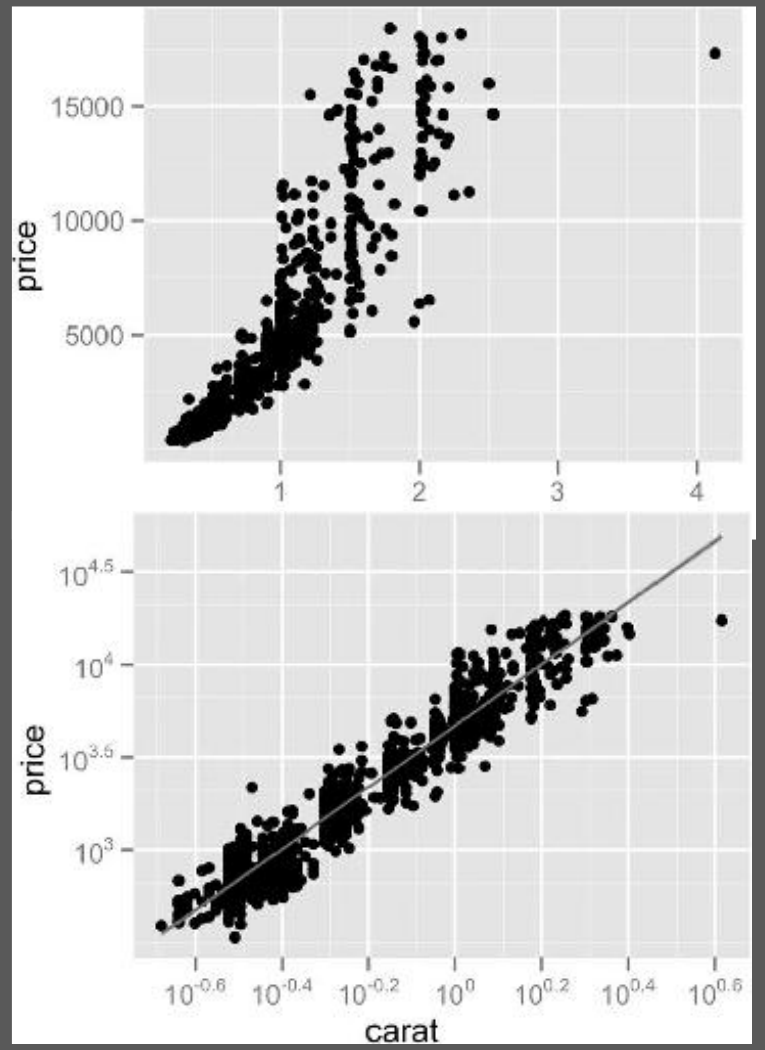
Data Visualisation Techniques

Scatter, Radial plots, Line, Bar, Pie Charts, Streamgraphs, Heatmaps

Scatter plots

No keys, only values

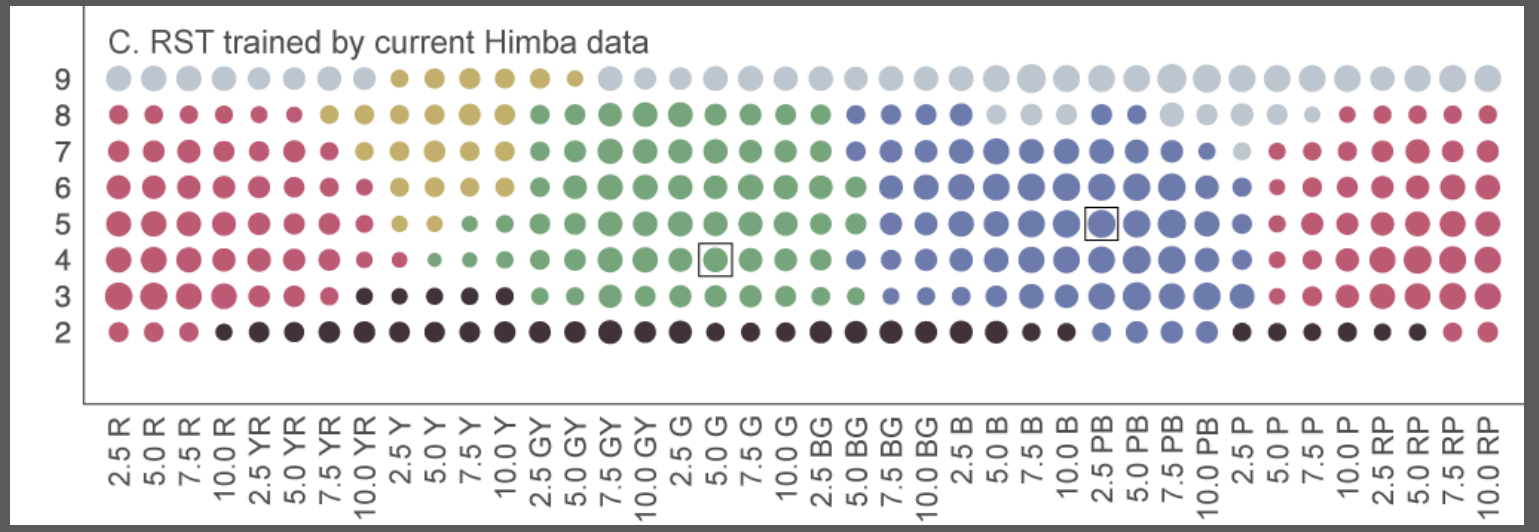
- Data: 2 quantitative
- Marks: points
- Channels: horizontal and vertical positions
- Tasks: find trends, outliers, correlation, distribution, clusters
- Scalability: hundreds of items



Scatter plots with additional channels

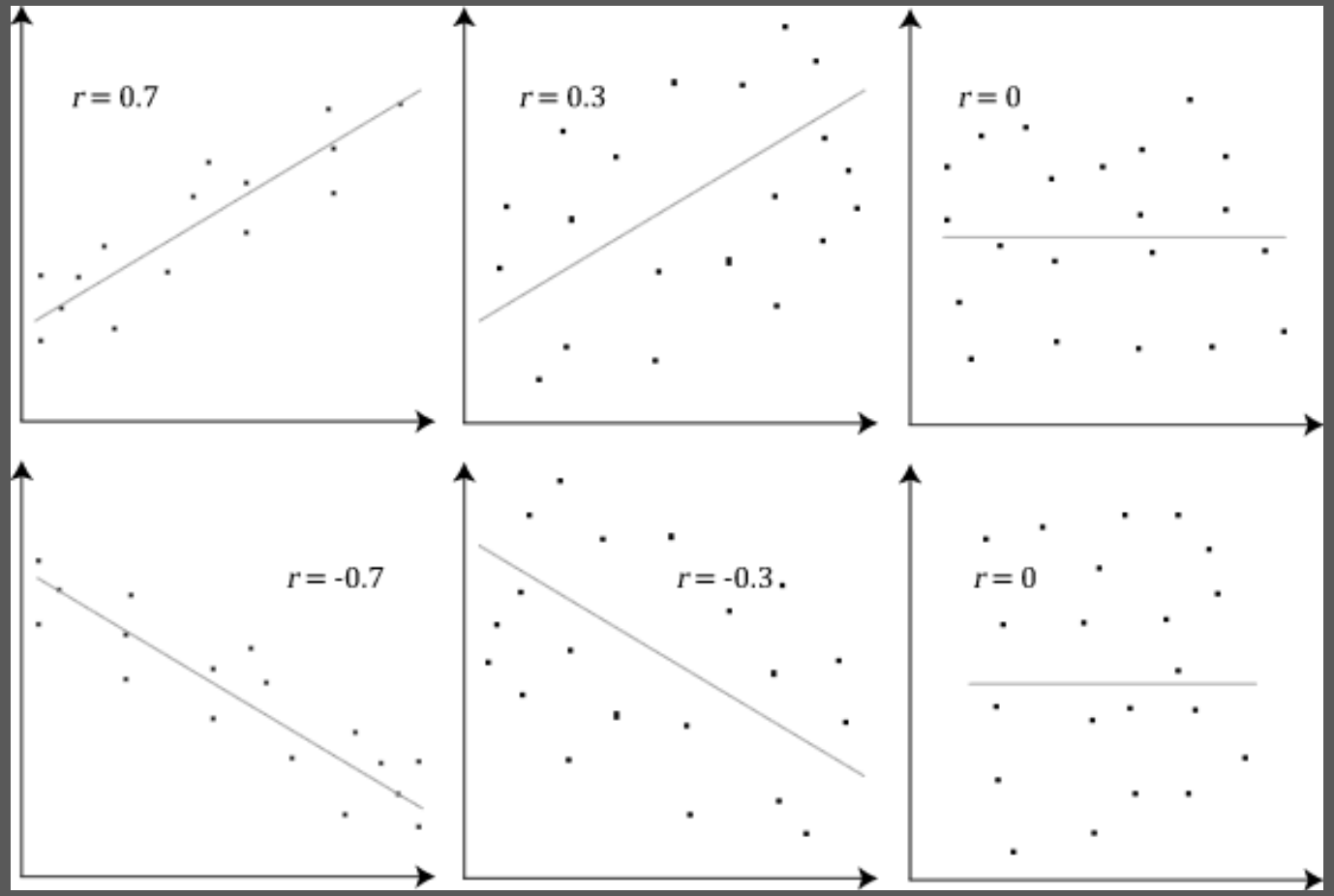
Additional encodings for point marks:

- Colour (clusters)
- Size (bubble plots)



Mylonas, Caparos & Davidoff, 2022)

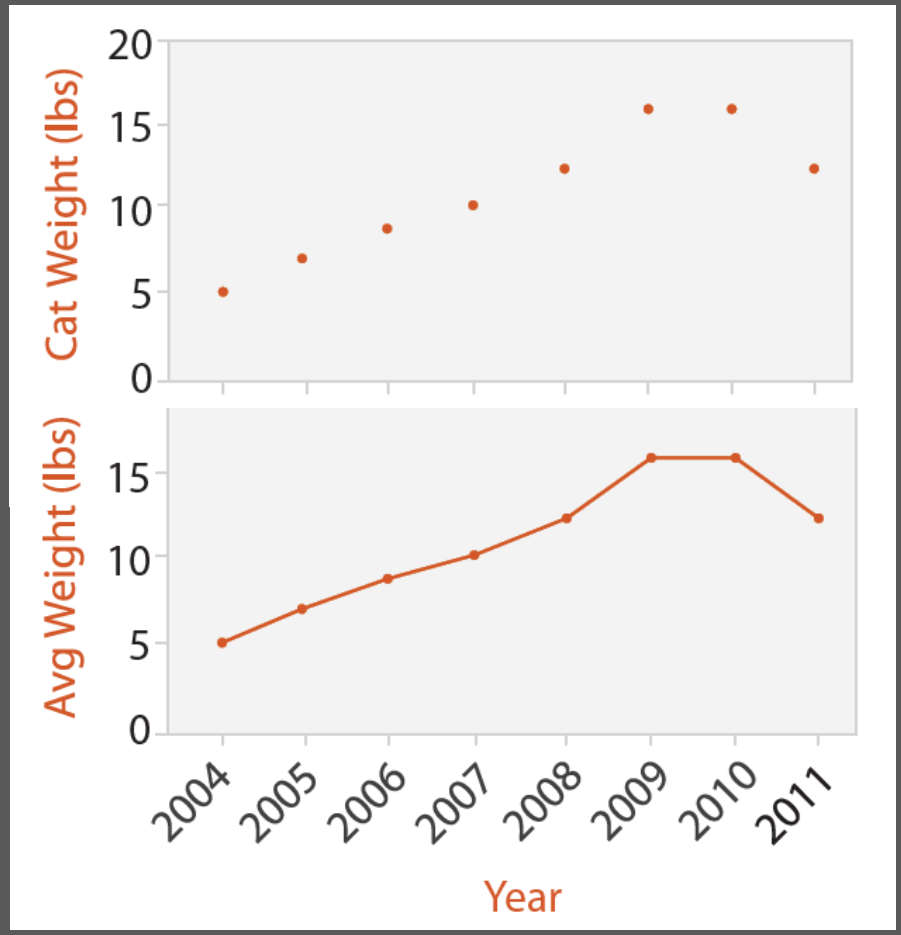
Scatter plot tasks (correlation)



Dot plots and line charts

One key, one value

- Data: 2 quantitative
- Marks: points and line
- Channels: position, length
- Task: find trend
- Scalability: hundreds key and value levels



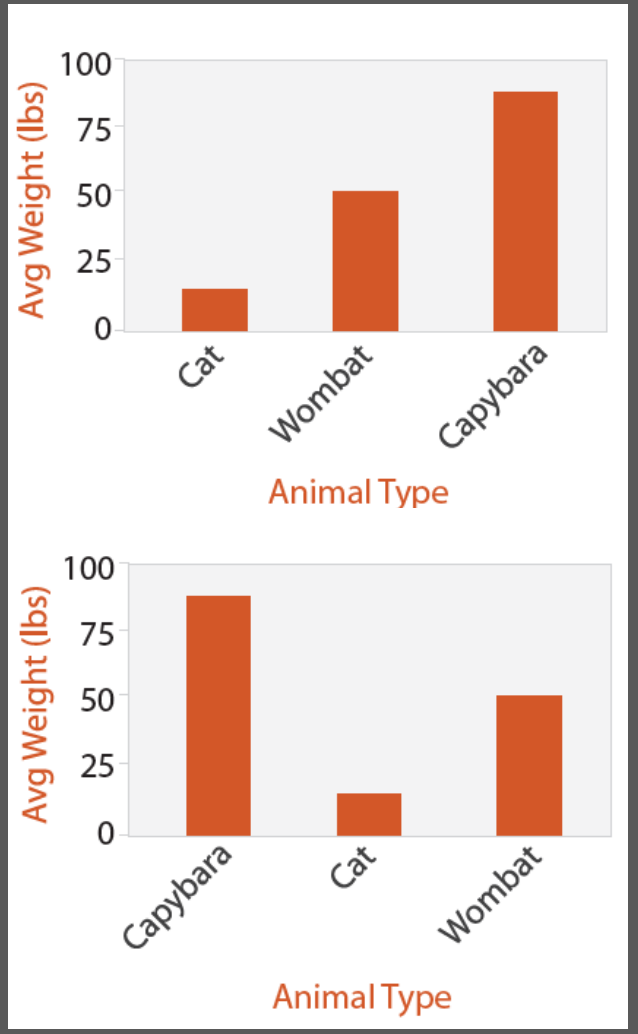
Bar charts

One key, one value

- Data: 1 categorical, 1 quantitative
- Marks: lines
- Channels: length, spatial regions

  - separated, aligned and ordered

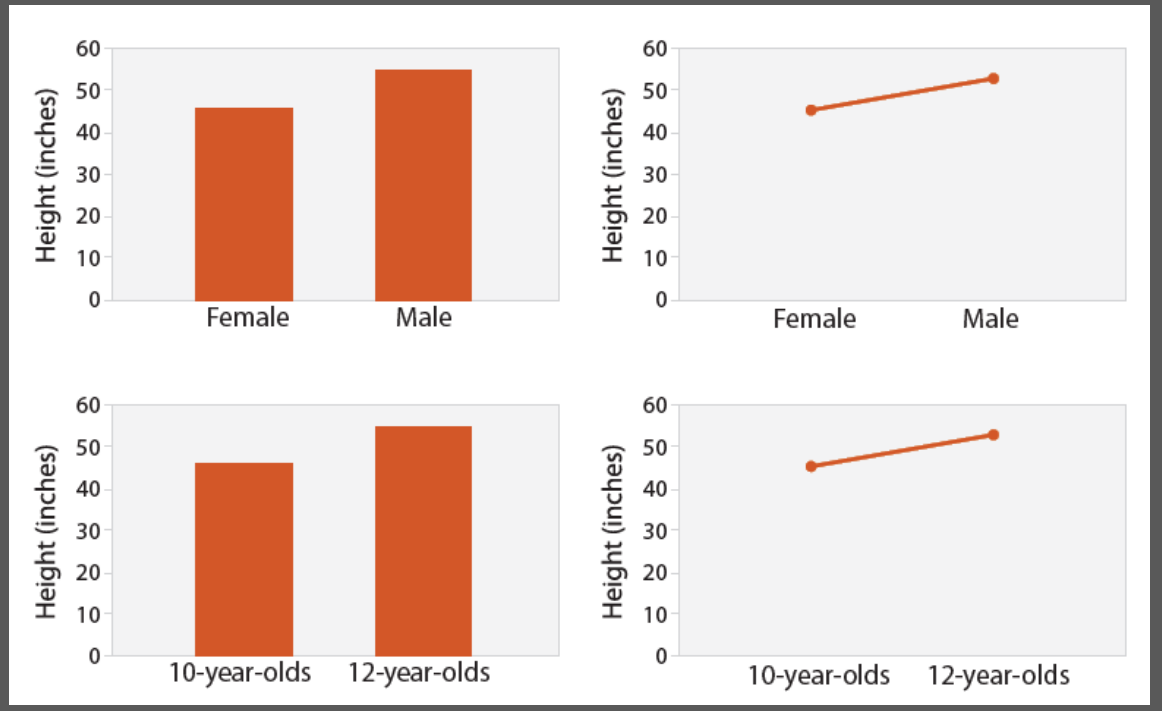
- Task: compare, look up values
- Scalability: dozens to hundreds of levels



Bar vs. line charts

Depends on type of key attribute

- Bar charts if categorical
- Line charts if ordered



Using line charts for categorical attributes violates expressiveness:
trend so strong that it overrides semantics

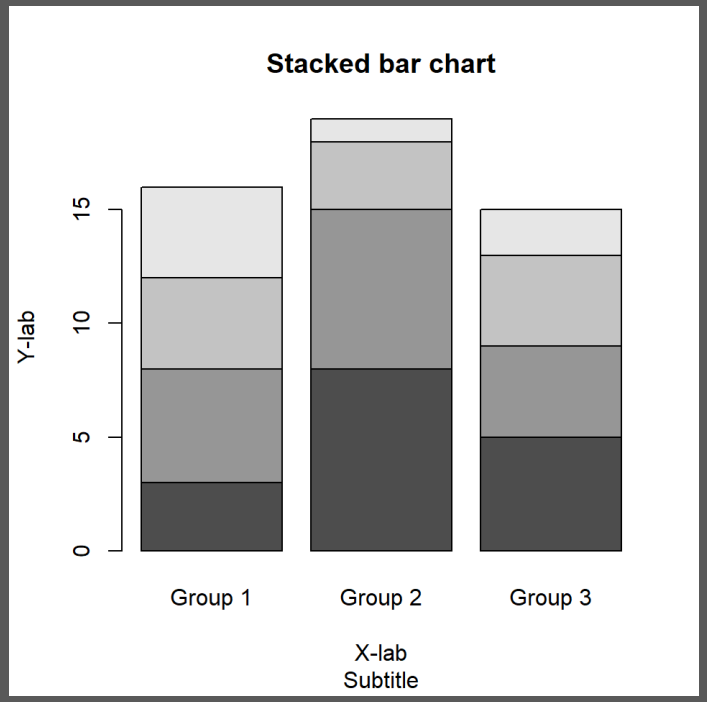
Stacked bar charts

One key, one value

- Data: 2 categorical, 1 quantitative
- Marks: vertical stack of lines

  glyph: composite object
- Channels: length, colour – lightness

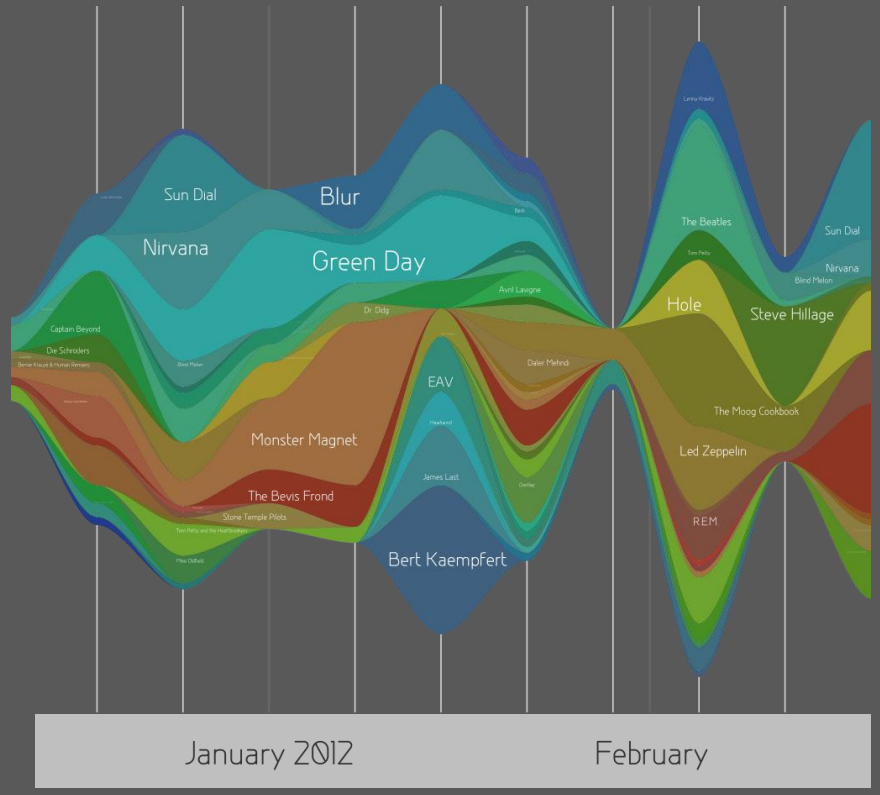
  - separated, aligned and unordered
- Task: part-to-whole relationship
- Scalability: up to one dozen levels



Streamgraphs

Generalized stacked graph

- Data: 1 categorical (bands),
  1 order key (dates)
  1 quantitative (counts)
- Derived data: geometry –height (counts)
  1 quantitative (layer ordering)
- Marks: lines and areas
- Channels: length, colour – hue
  - separated, aligned and ordered
- Task: change-over-time
- Scalability: hundreds of time keys and dozens to hundreds bands



Indexed line charts

One key, one value

- Data: 2 quantitative
- Derived data: 1 quantitative index instead of price
- Marks: lines
- Channels: length, colour – hue

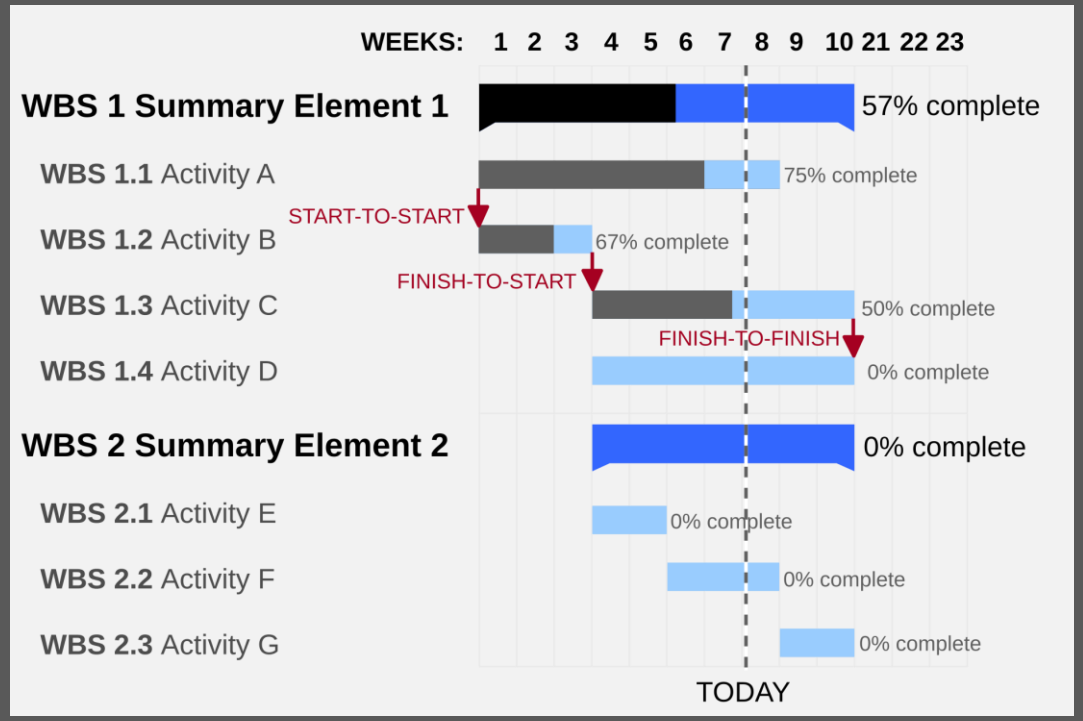
  - separated, aligned and ordered
- Task: change-over-time, normalised
- Scalability: hundreds key and value levels

Effect of DeepSeek to AI related shares 2025

Gantt charts

One key, two related values

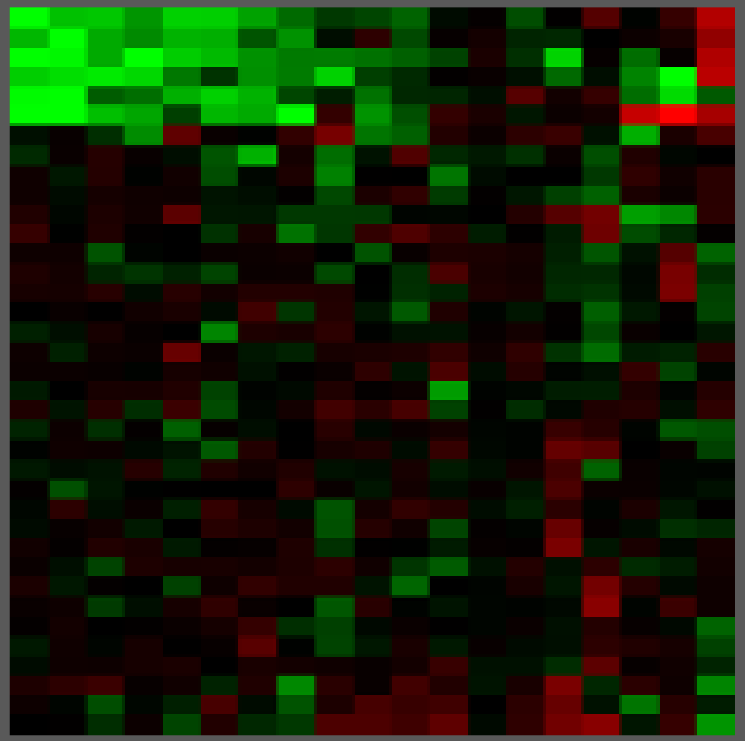
- Data: 1 categorical, 2 quantitative
- Marks: line (duration)
- Channels: position start-end time
- Task: Start/end, temporal overlaps
- Scalability: dozens of keys, hundreds of levels



Heatmaps

Two keys, one value

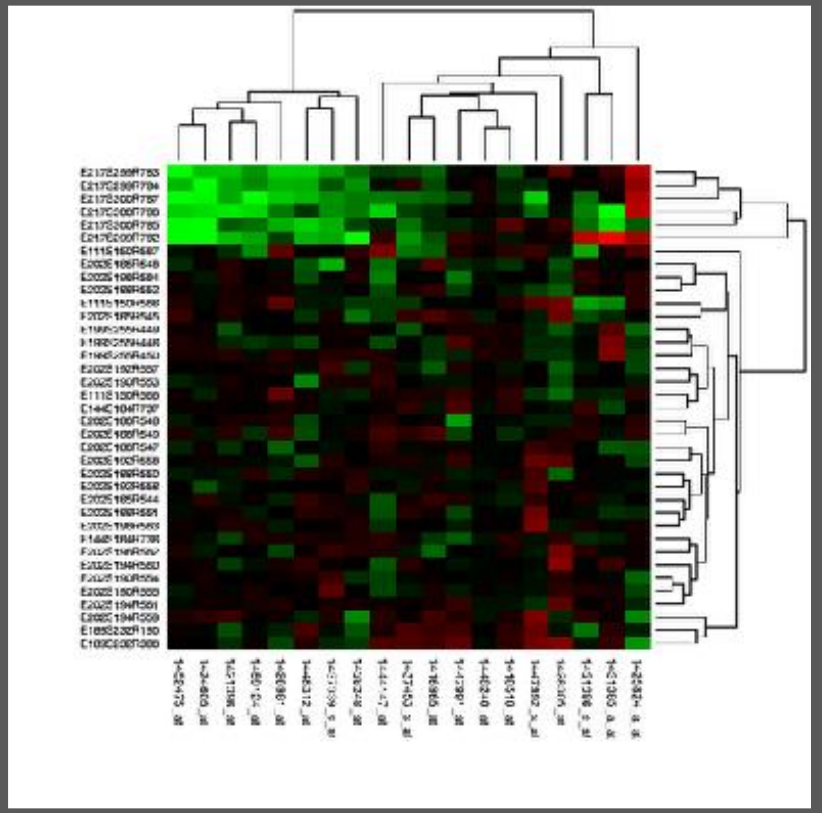
- Data: 2 categorical, 1 quantitative
- Marks: points
- Channels: colour – hue by quantity
- Task: find clusters, outliers
- Scalability: 1M items, 100s of categorical, ~10 quantitative levels



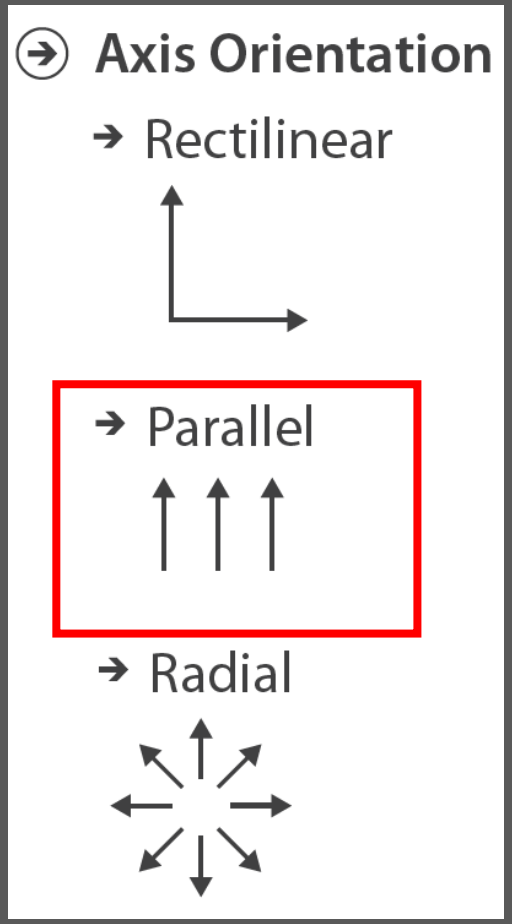
Cluster Heatmaps

In addition:

- Derived data: 2 cluster hierarchies
- Dendrogram: parent-child relationships
- Heatmap: marks re-ordered by clusters hierarchy traversal
- Task: assess quality of clusters found by automatic methods



Axis orientation

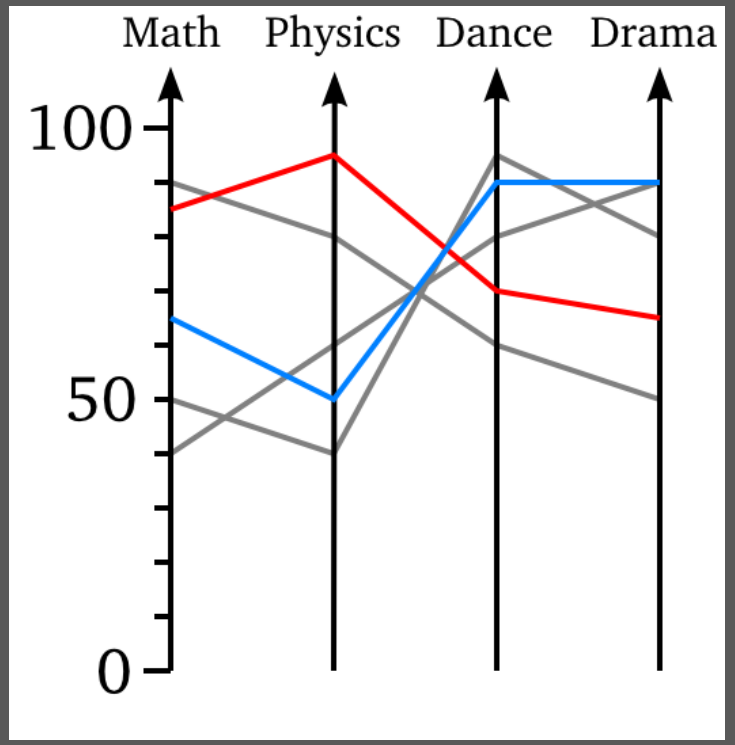


Parallel coordinates

Axes in parallel to show position

Data: 1 categorical, 1 quantitative

- Marks: lines (counts)
- Channels: colour, length, position
- Task: correlations, relationships
- Scalability: dozens of attributes and hundreds of items



Limitations of parallel coordinates

visible patterns only between neighboring axis pairs

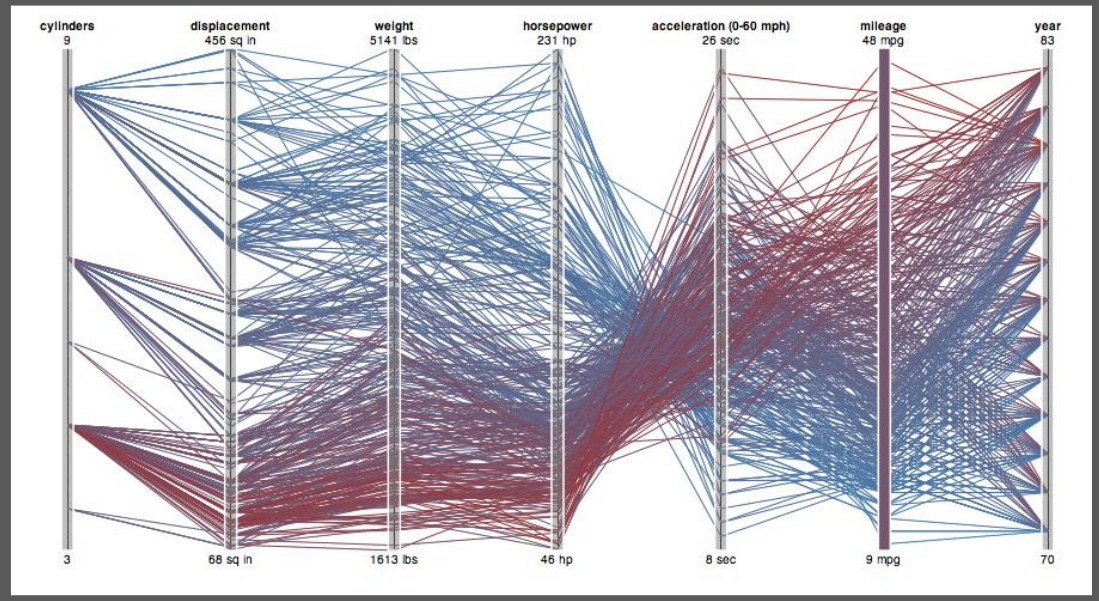
- how to pick axis order?

– usual solution: reorderable axes, interactive exploration

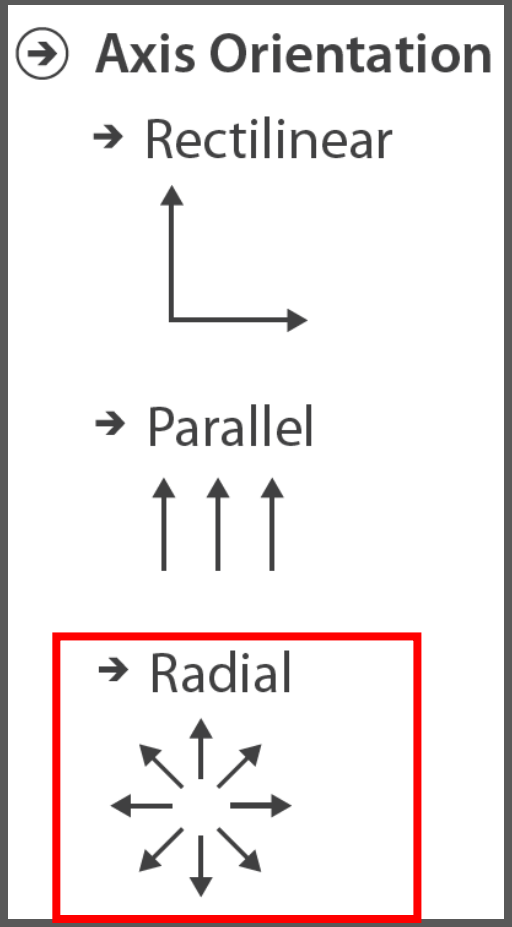
– same weakness as many other techniques

- downside of interaction: human-powered search

– some algorithms proposed, none fully solve



Axis orientation

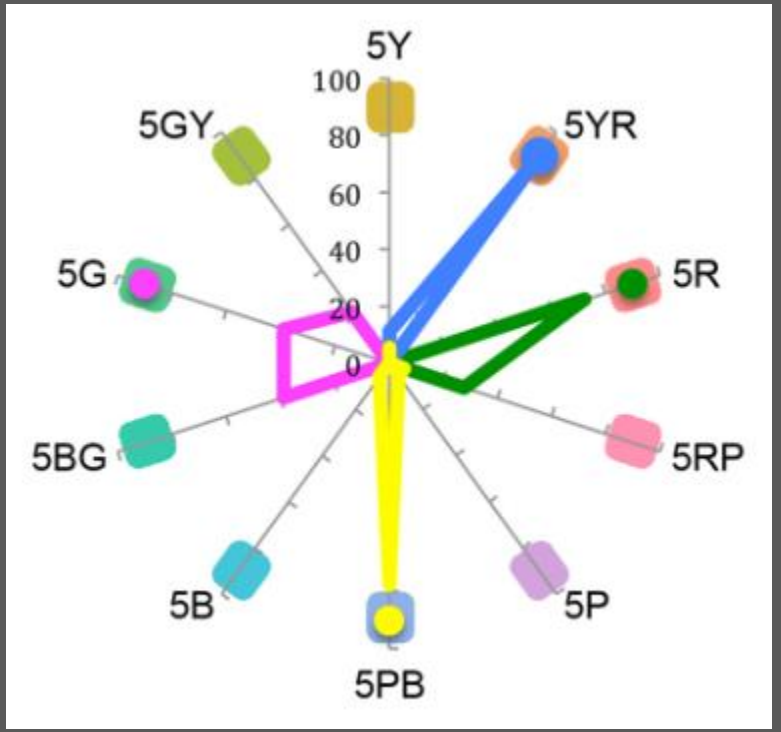


Radar plots

Radial axes meet at central ring

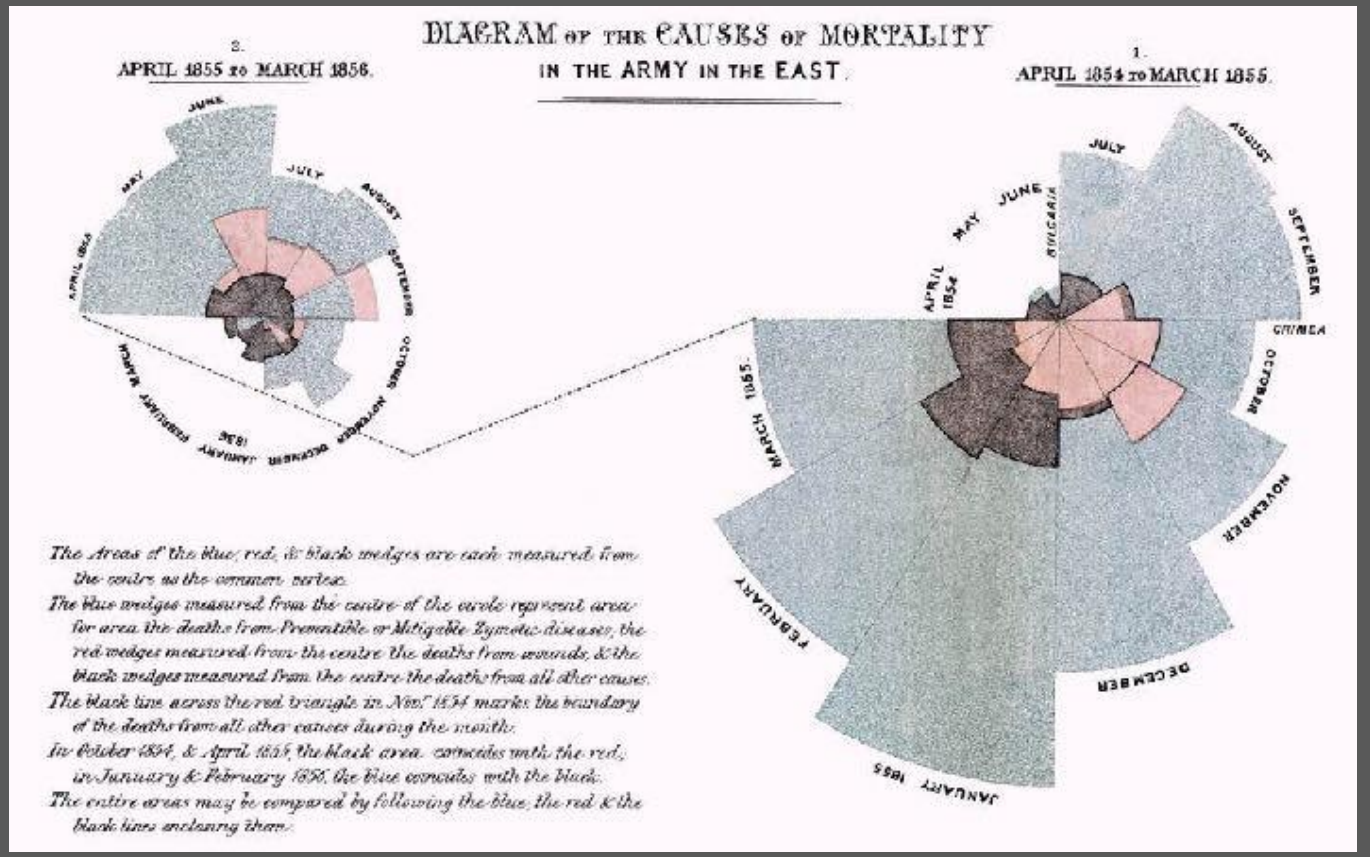
Data: 1 categorical, 1 quantitative

- Marks: line (counts)
- Channels: colour, length
- Task: orientation
- Scalability: dozens of keys and values



Zeki, Cheadle, Pepper and Mylonas (2017)

Diagram of the causes of mortality (1858)

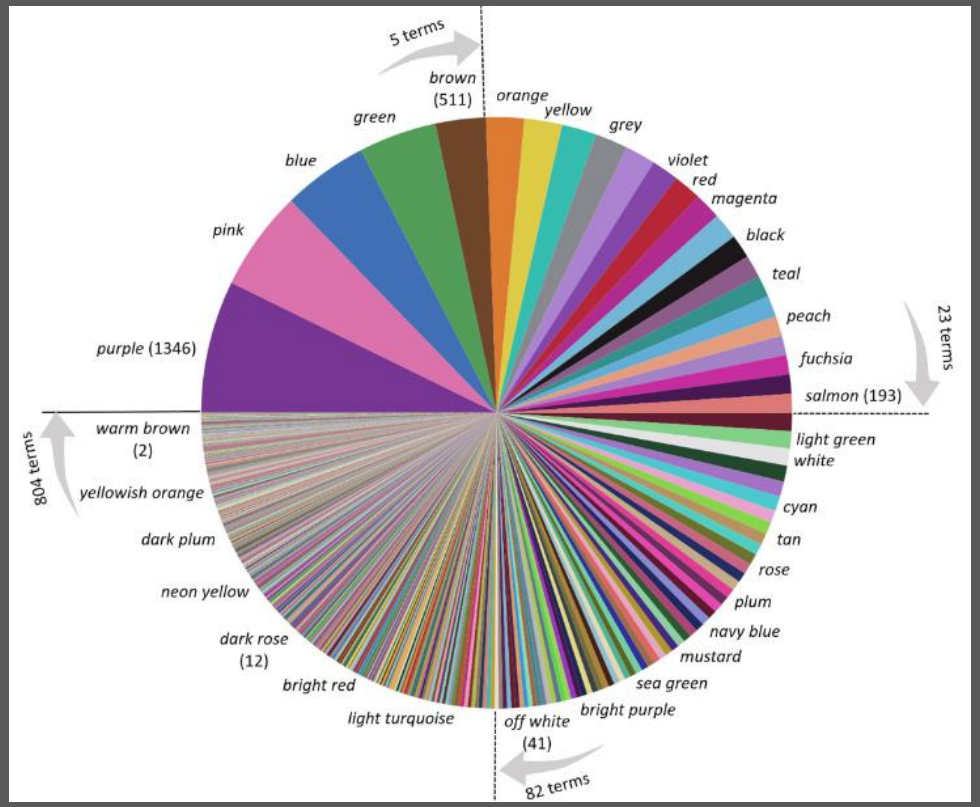


Nightingale and Farr (1858)

Pie charts

One key, one value

- Data: 1 categorical, 1 quantitative
- Marks: points
- Channels: lines with angle, colour by area
- Task: part-to-whole judgements
- Scalability: 2+ to hundreds of levels



Griffin and Mylonas (2019)

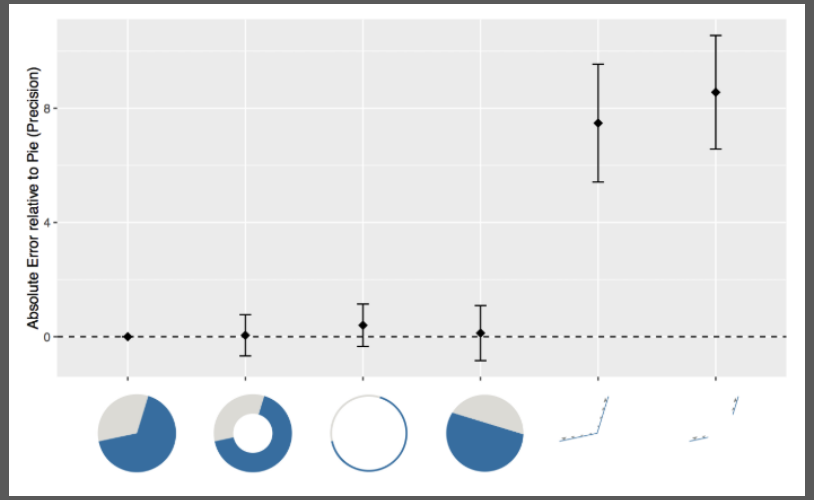
Criticism of pie charts

- Empirical evidence that people respond to arc length

– not angles

–maybe also areas?…

- donut charts no worse than pie charts



Questions?