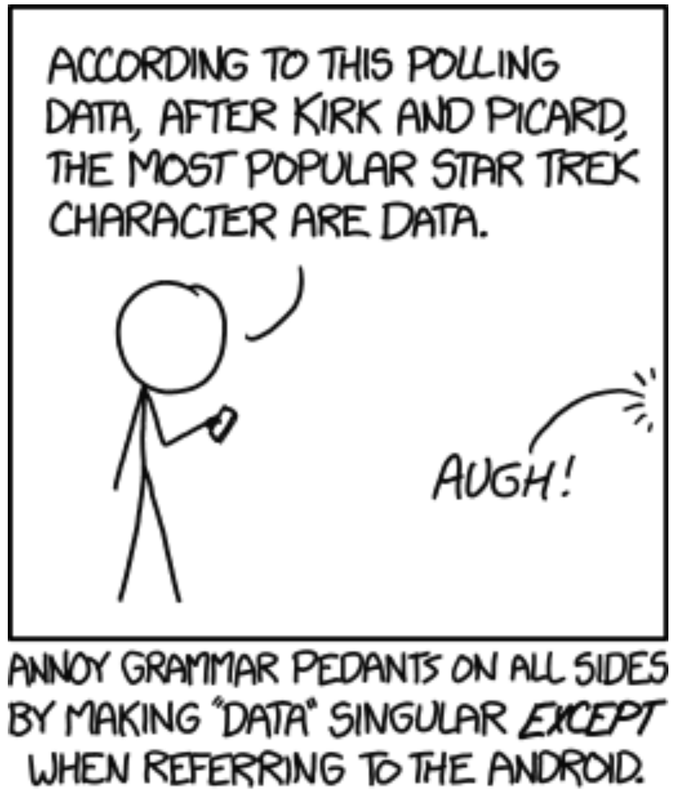
# CS-5630 / CS-6630
# Visualization for Data Science
# Data

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH



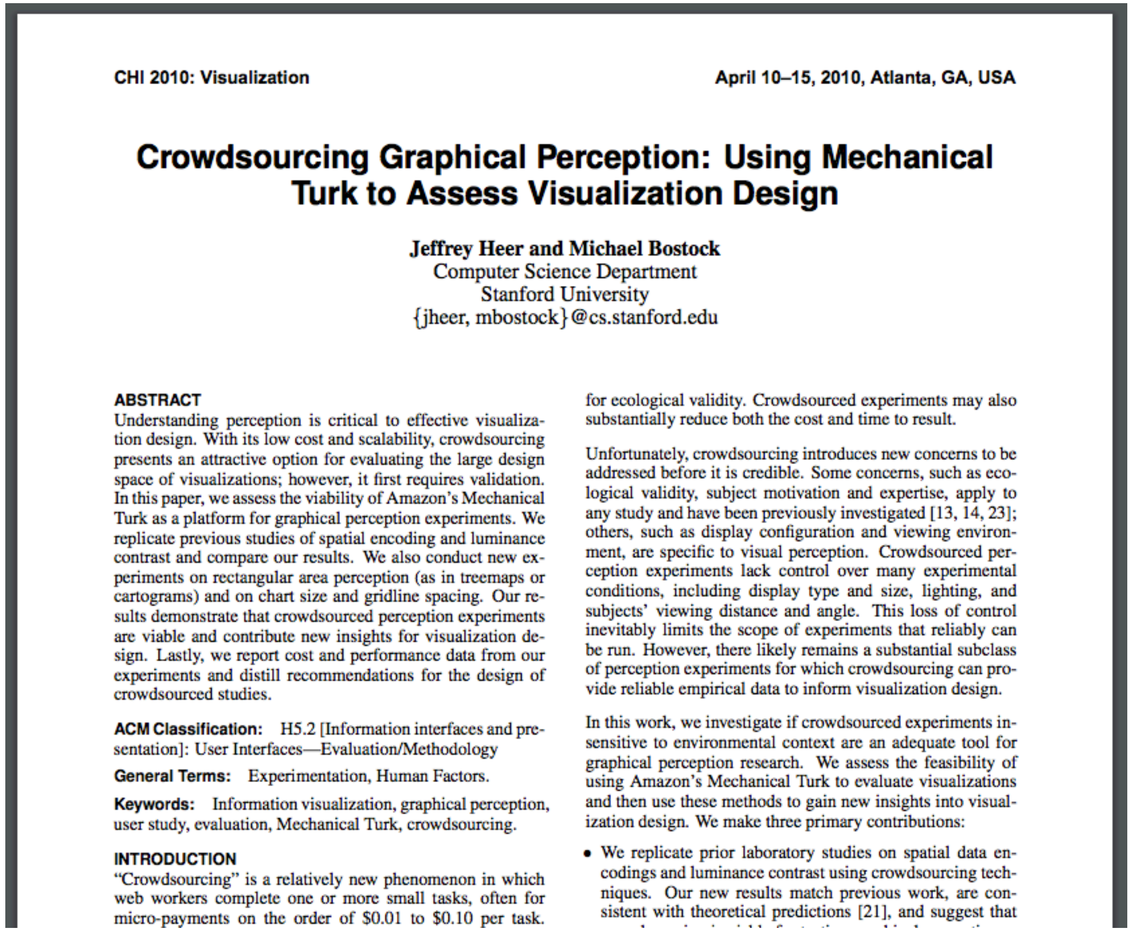

[xkcd]

# Next Week

Tuesday: JavaScript and D3 Intro

Thursday: Visualization Alphabet

Mandatory Reading: Crowdsourcing graphical perception: using mechanical turk to assess visualization design. Jeff Heer, Mike Bostock



CHI 2010: Visualization April 10–15, 2010, Atlanta, GA, USA

## Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design

**Jeffrey Heer and Michael Bostock**
Computer Science Department
Stanford University
{jheer, mbostock}@cs.stanford.edu

**ABSTRACT**
Understanding perception is critical to effective visualization design. With its low cost and scalability, crowdsourcing presents an attractive option for evaluating the large design space of visualizations; however, it first requires validation. In this paper, we assess the viability of Amazon's Mechanical Turk as a platform for graphical perception experiments. We replicate previous studies of spatial encoding and luminance contrast and compare our results. We also conduct new experiments on rectangular area perception (as in treemaps or cartograms) and on chart size and gridline spacing. Our results demonstrate that crowdsourced perception experiments are viable and contribute new insights for visualization design. Lastly, we report cost and performance data from our experiments and distill recommendations for the design of crowdsourced studies.

**ACM Classification:** H5.2 [Information interfaces and presentation]: User Interfaces—Evaluation/Methodology

**General Terms:** Experimentation, Human Factors.

**Keywords:** Information visualization, graphical perception, user study, evaluation, Mechanical Turk, crowdsourcing.

**INTRODUCTION**
"Crowdsourcing" is a relatively new phenomenon in which web workers complete one or more small tasks, often for micro-payments on the order of $0.01 to $0.10 per task. for ecological validity. Crowdsourced experiments may also substantially reduce both the cost and time to result.

Unfortunately, crowdsourcing introduces new concerns to be addressed before it is credible. Some concerns, such as ecological validity, subject motivation and expertise, apply to any study and have been previously investigated [13, 14, 23]; others, such as display configuration and viewing environment, are specific to visual perception. Crowdsourced perception experiments lack control over many experimental conditions, including display type and size, lighting, and subjects' viewing distance and angle. This loss of control inevitably limits the scope of experiments that reliably can be run. However, there likely remains a substantial subclass of perception experiments for which crowdsourcing can provide reliable empirical data to inform visualization design.

In this work, we investigate if crowdsourced experiments insensitive to environmental context are an adequate tool for graphical perception research. We assess the feasibility of using Amazon's Mechanical Turk to evaluate visualizations and then use these methods to gain new insights into visualization design. We make three primary contributions:

- We replicate prior laboratory studies on spatial data encodings and luminance contrast using crowdsourcing techniques. Our new results match previous work, are consistent with theoretical predictions [21], and suggest that

# Terms

## Dataset Types

what can be visualized?

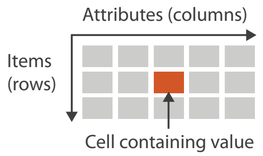
## Data Types

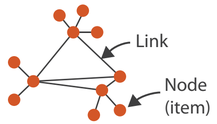
fundamental units

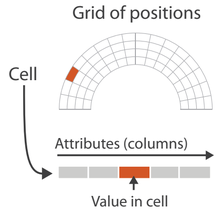
combinations make up Dataset Types

**Dataset Types**

- Tables
- Networks
- Fields (Continuous)
- Geometry (Spatial)

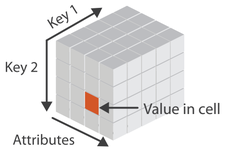




- *Multidimensional Table*
- *Trees*



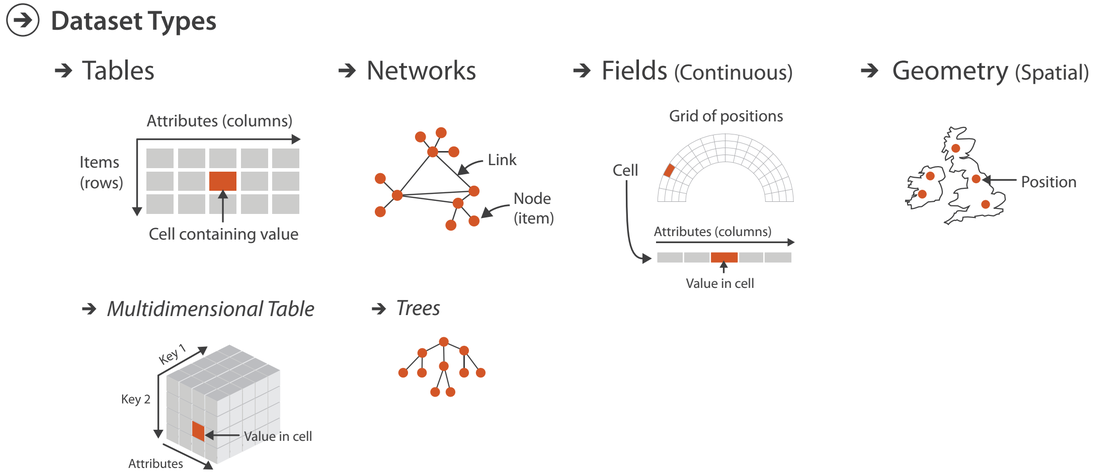

**Data Types**

- Items
- Attributes
- Links
- Positions
- Grids

# Structure

## Structured Data

known data types, semantics




## Unstructured Data

no predefined data model

text-heavy, interspersed with facts (dates, times, locations)

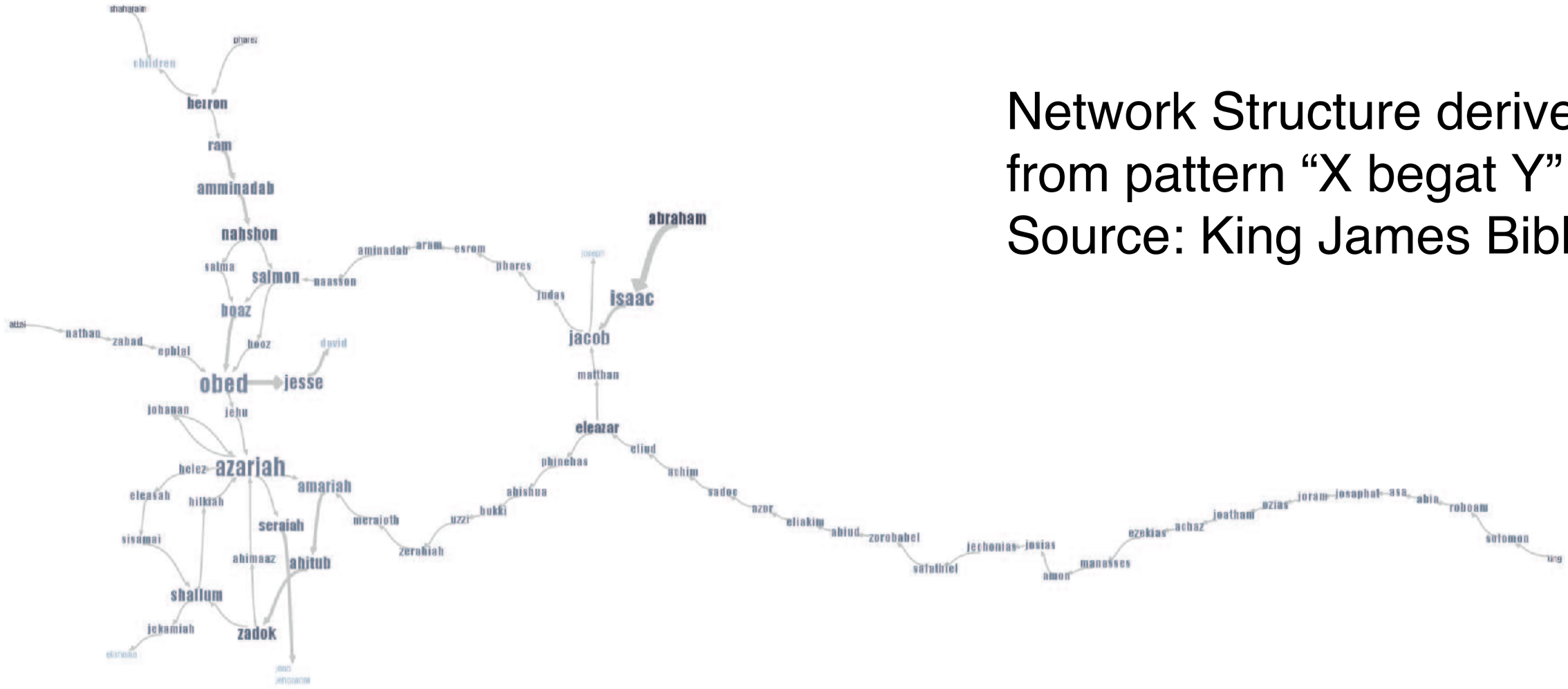
video, images

Translate into structured data

Natural Language Processing

Text mining (sentiment, keywords, concepts, categories)

# Text Example: Phrase Net



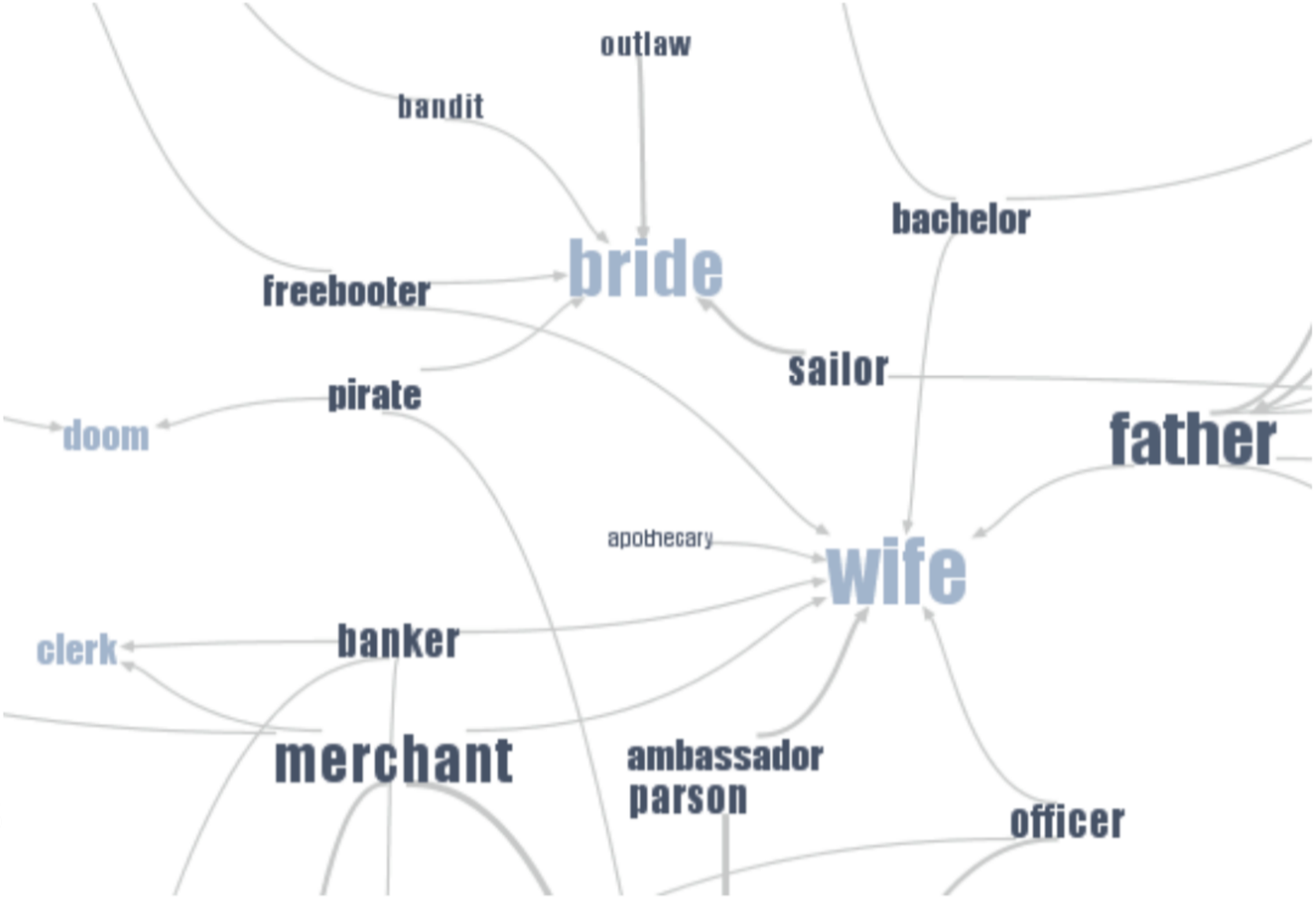

Network Structure derived from pattern “X begat Y”
Source: King James Bible

[van Ham, InfoVis 2009]

# Example: Phrase Net

Pattern: “X’s Y”

18th & 19th century novels




**More in Lecture**
**Text & Document Vis**

[van Ham, InfoVis 2009]

# Data Semantics

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

Basil, 7, S, Pear

What does it mean?

**Semantics:** real world meaning

- Name? City? Fruit? Height? Age? Day of Month?
- Metadata

# Data Types

structural or mathematical interpretation of data

**Item, Link, Attribute, Position, Grid**

Different from data types in programming!

# Items & Attributes

Item: individual entity, discrete

e.g., Patient, Car, Stock, City

"independent variable"

Attribute: measured, observed, logged property

e.g., Patient: height, blood pressure
Car: horsepower, make

"dependent variable"

Item: Person  Attributes

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

Cell

# Other Data Types

Links

- Express relationship between two items
- Friendship on Facebook, Interaction between proteins

Positions

- Spatial data -> location in 2D or 3D
- Pixels in photo, Voxels in MRI scan, latitude/longitude

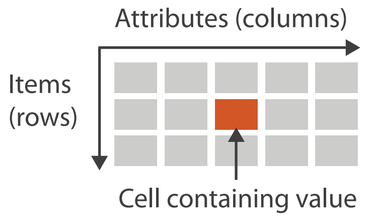
Grids

- Sampling strategy for continuous data
- How many Voxels in MRI scan, positions of weather stations in the US

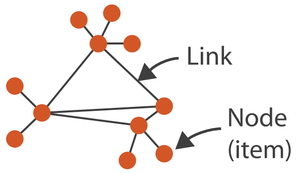
# Dataset Types

## Dataset Types

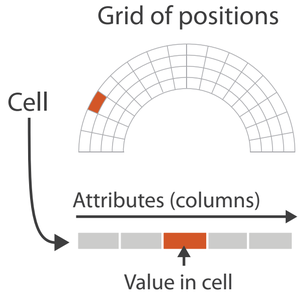
### Tables




### Networks

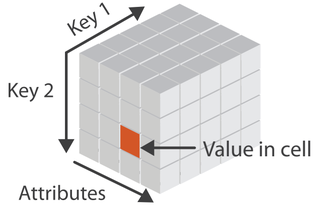



### Fields (Continuous)

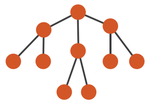



### Geometry (Spatial)


### *Multidimensional Table*




### *Trees*



# Tables

## Flat Table

- one item per row
- each column is attribute
- unique (implicit) **key**
- no duplicates

## Multidimensional Table

- indexing based on multiple keys

Attributes

Keys | Values

Item

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

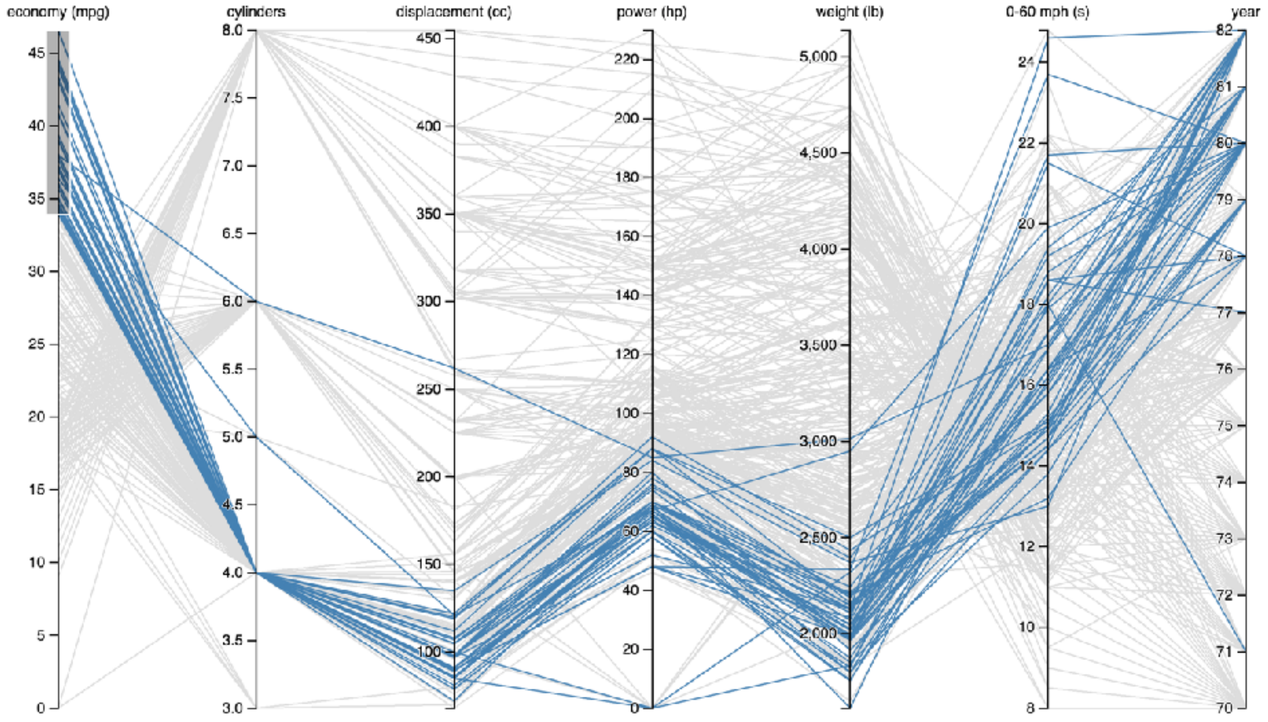
# Multidimensional Tables

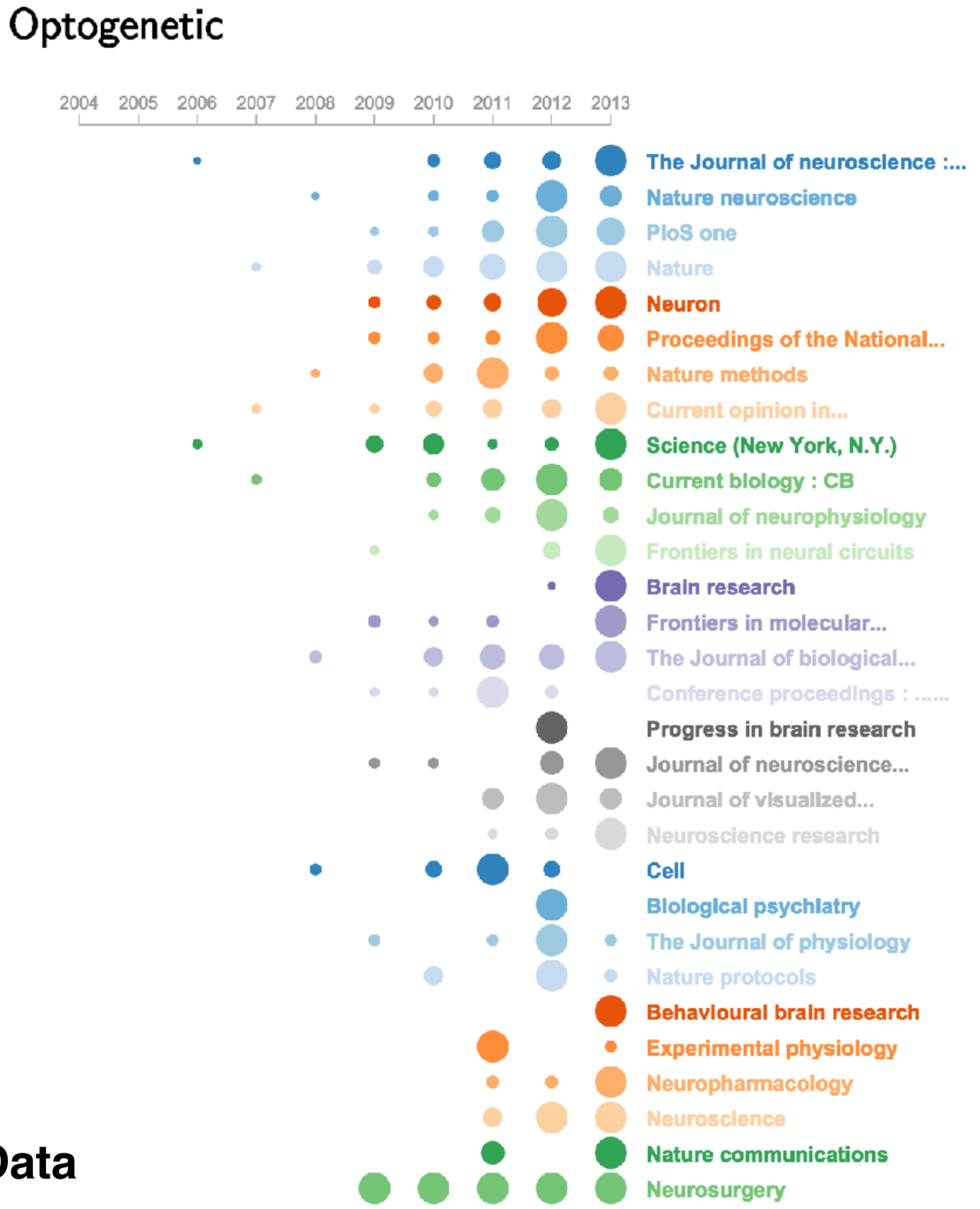
Keys: Genes

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | #1.2 | | | | |
| 2 | 1500 | 529 | | | |
| 3 | GeneName | DESCRIPTION | TCGA-02-0001-01C-01R-0177-01 | TCGA-02-0003-01A-01R-0177-01 | TCGA-02-0004-01A-01R-0298-01 |
| 4 | LTF | LTF | -1.265728057 | 2.377012066 | 4.123979585 |
| 5 | POSTN | POSTN | 2.662411805 | 3.932400324 | 5.031585377 |
| 6 | TMSL8 | TMSL8 | -3.082217838 | -2.243148513 | -0.02313681 |
| 7 | HLA-DQA1 | HLA-DQA1 | -1.739664398 | 4.577962344 | 3.127744964 |
| 8 | RP11-35N6.1 | RP11-35N6.1 | -3.346352968 | -2.895400157 | -3.473035067 |
| 9 | STMN2 | STMN2 | -2.578511106 | -3.051605144 | -1.729892888 |
| 10 | DCX | DCX | -2.26078976 | -2.529795801 | -2.844966278 |
| 11 | AGXT2L1 | AGXT2L1 | -2.639493611 | -3.113204863 | -0.403975027 |
| 12 | IL13RA2 | IL13RA2 | -2.93596915 | -1.873600916 | 2.976256911 |
| 13 | SLN | SLN | -2.466718221 | -2.208406749 | 1.025827904 |
| 14 | MEOX2 | MEOX2 | -2.395054066 | -1.062676046 | 1.783235317 |
| 15 | COL11A1 | COL11A1 | 1.211934832 | -0.399392588 | 4.733608974 |
| 16 | NNMT | NNMT | 0.703745164 | 0.664082419 | 3.069030715 |
| 17 | F13A1 | F13A1 | -0.224094042 | 2.222197544 | 1.171354775 |
| 18 | CXCL14 | CXCL14 | -3.1309694 | -1.395056071 | 2.569540659 |
| 19 | MBP | MBP | -1.906390566 | -2.037626447 | -2.935744906 |
| 20 | TF | TF | -4.334123292 | -4.680680246 | -2.975788866 |
| 21 | KCND2 | KCND2 | -1.777692395 | -2.100362021 | -1.996306032 |
| 22 | GABRB1 | GABRB1 | -2.214760175 | -3.022654105 | -3.185499425 |

Keys: Patients

# Visualizing Tables




**More in Lecture on Tables & High-Dimensional Data**

# Graphs/Networks

A graph G(V,E) consists of a set of **vertices (nodes)** V and a set of **edges (links)** E connecting these vertices.

Diagrammatic Example

# Graphs/Networks

A simple graph is a graph which contains

- No multi-edges
- No loops

Not a simple graph!
→ A ***general graph***

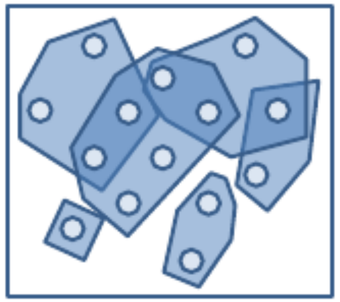
# Special Graphs

A ***tree*** is a graph with *no cycles*

Tree

A ***directed graph*** (digraph) is a graph that distinguishes between edges A-> B and A <- B

A ***hypergraph*** is a graph with edges connecting any number of vertices

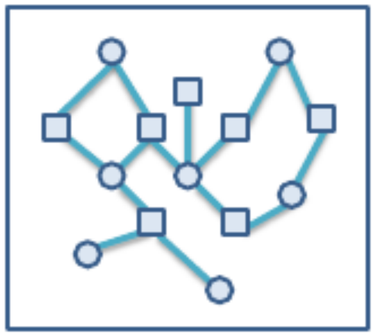


Hypergraph Example

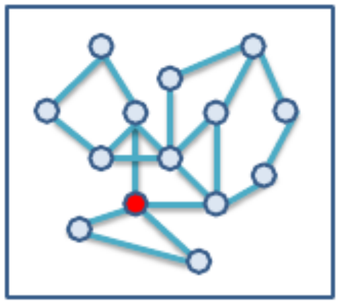
# Special Graphs

A ***bipartite graph*** has vertices that can be partitioned into two independent sets

An ***articulation point*** is a Vertex, which if deleted from the graph would break up a ***connected graph*** into multiple graphs,or an ***unconnected graph***



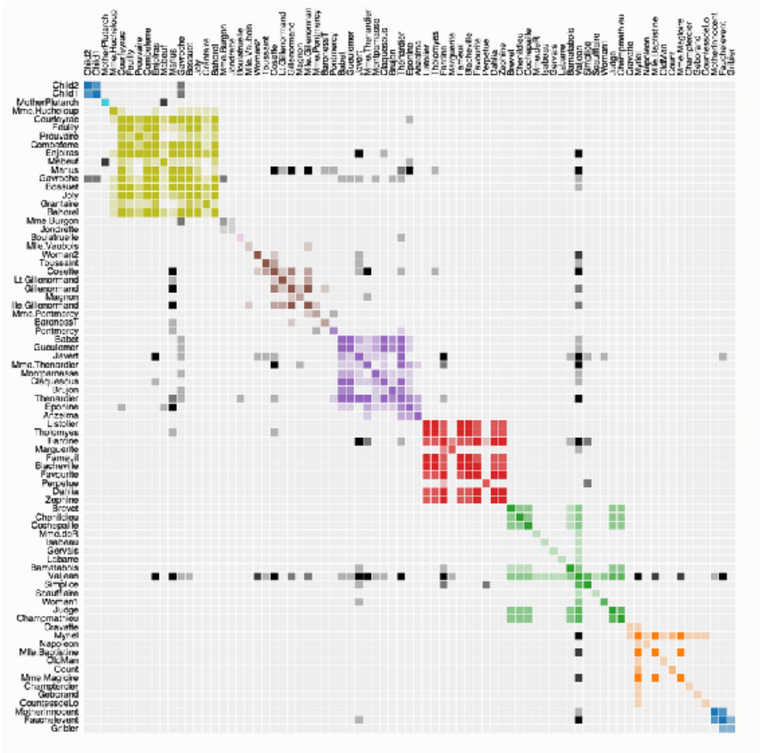
Bipartite Graph



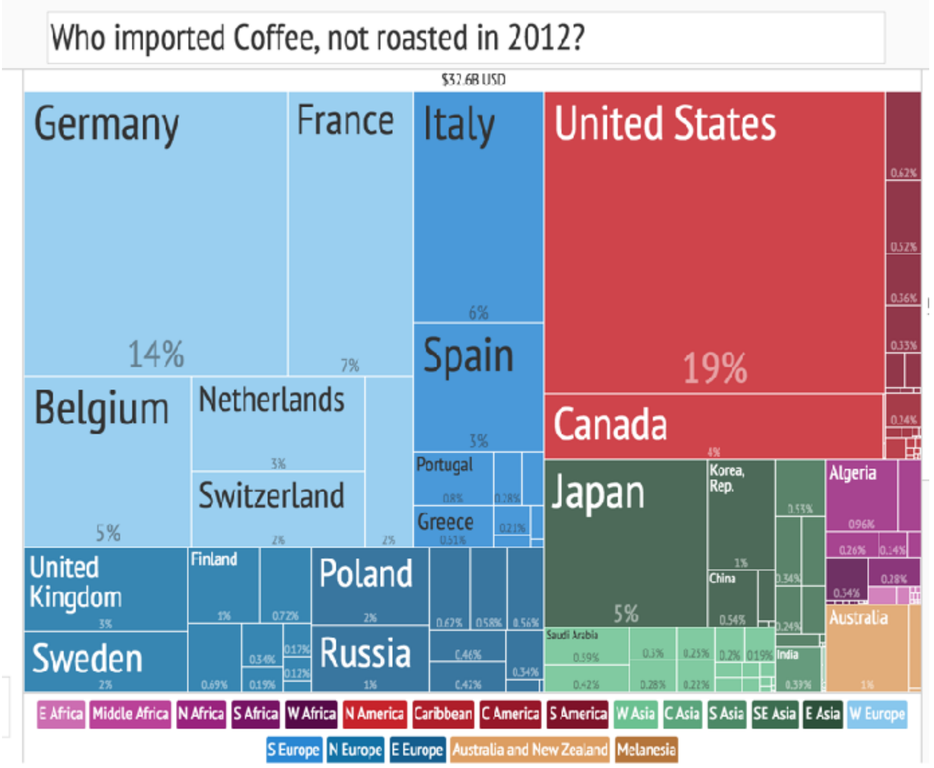
Articulation Point (red)

# Visualizing Graphs

Node-Link Diagram



Matrix




Treemap (Implicit Tree Visualization)

**More in Lecture on Graphs & Trees**

# Fields

Attribute values associated with cells

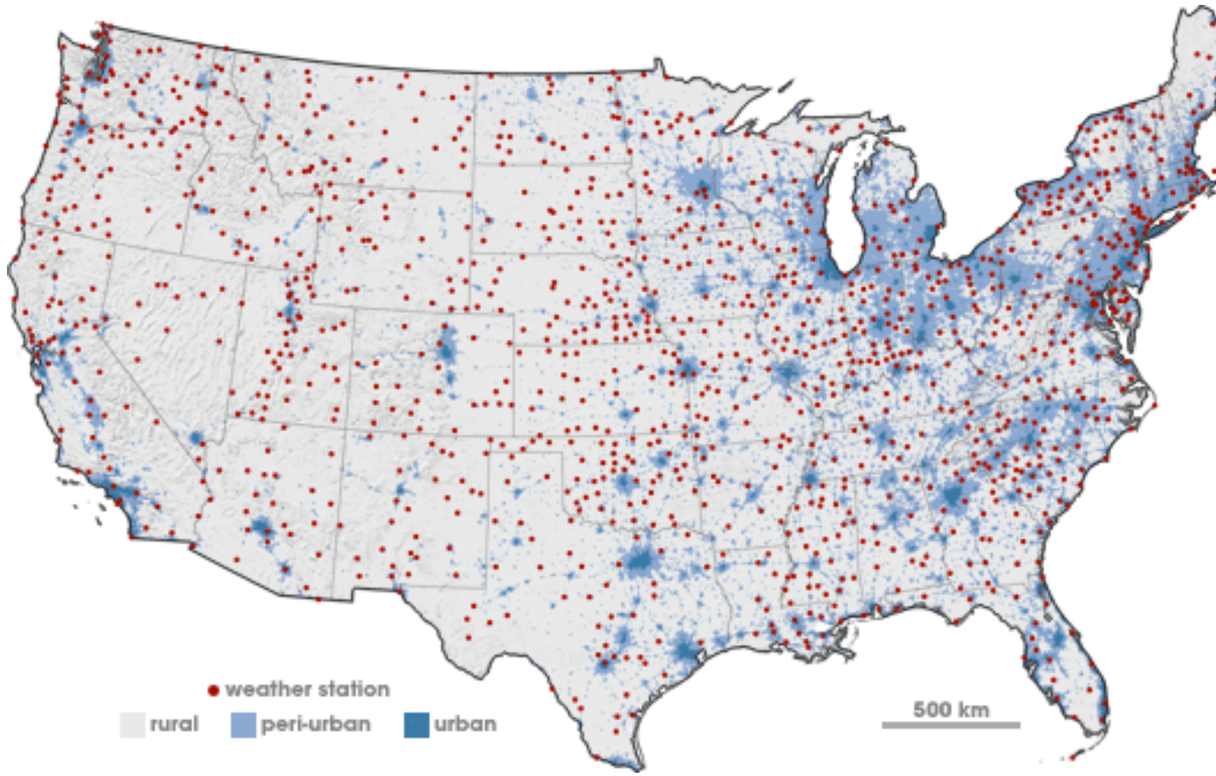
Cell contains data from continuous domain

- Temperature, pressure, wind velocity

Measured or simulated

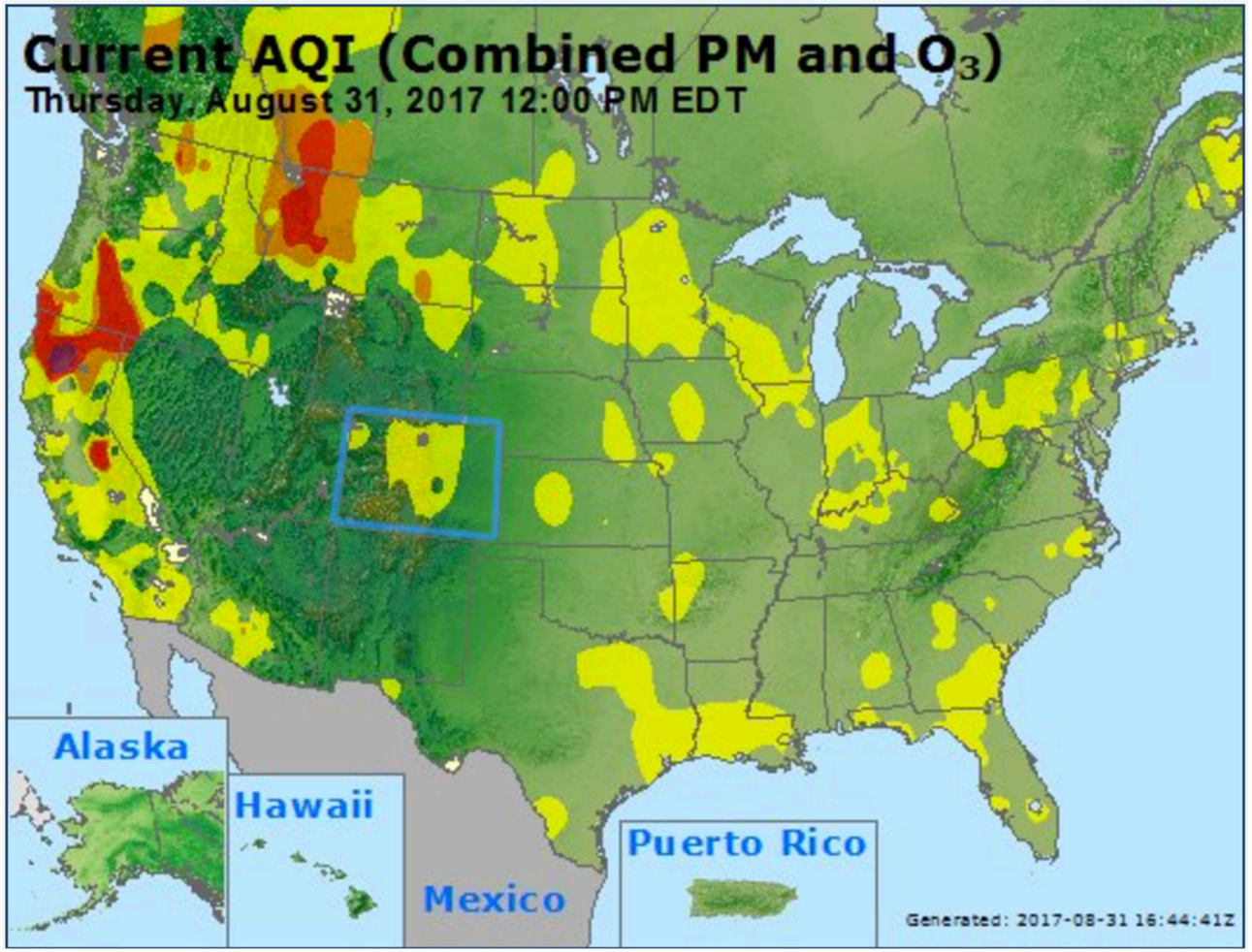
Sampling & Interpolation

- Signal processing & stats

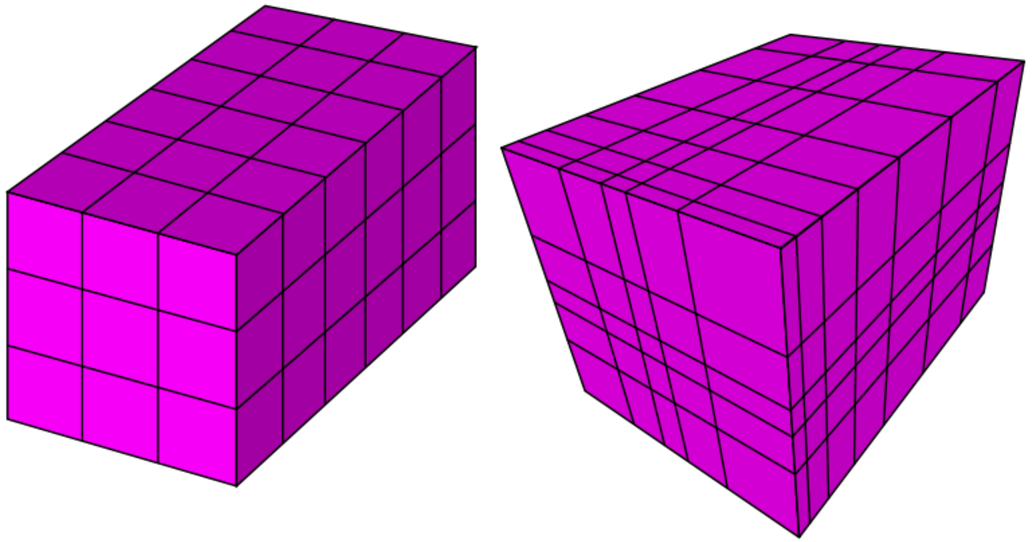



Weather Stations in the US. Source: NASA

# Field Example: Air Quality



# Fields: Grid Types



Uniform Grid

- Geometry & topology can be computed

Rectilinear Grid

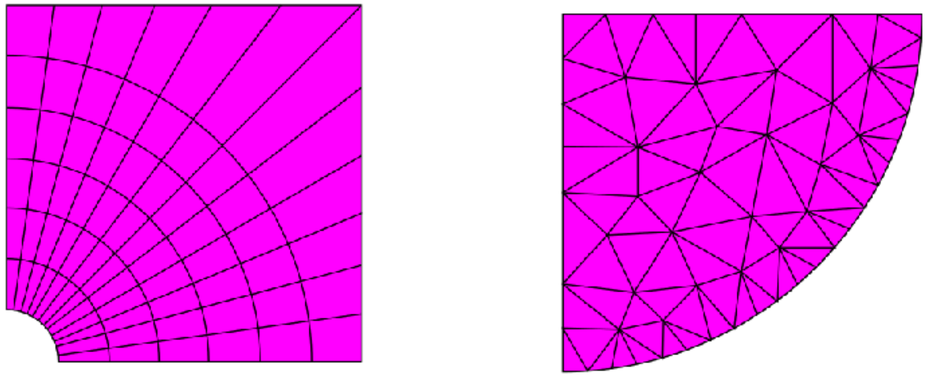
- Nonuniform sampling

Structured Grid

- allows curvilinear grids

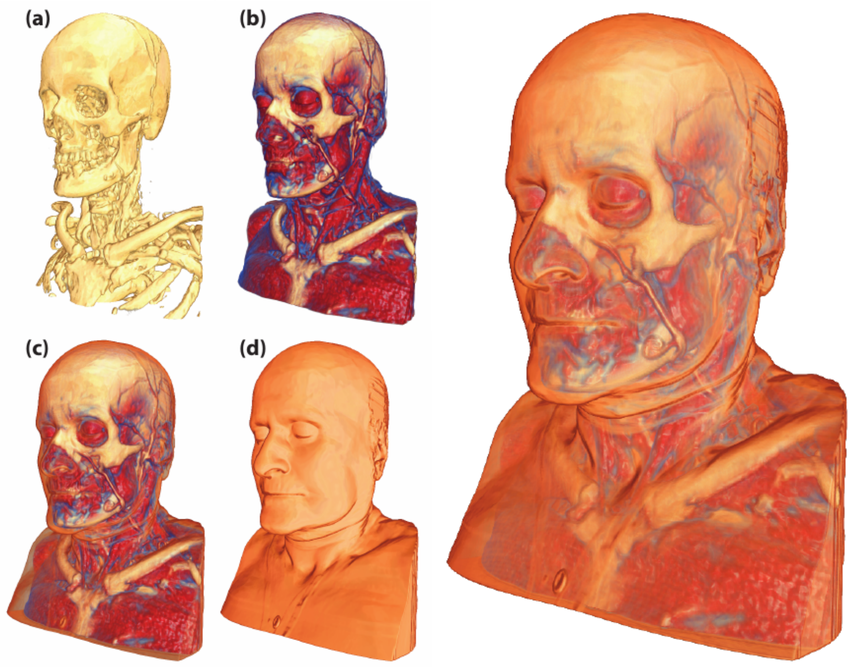
Unstructured Grid

- full flexibility, store position and connection

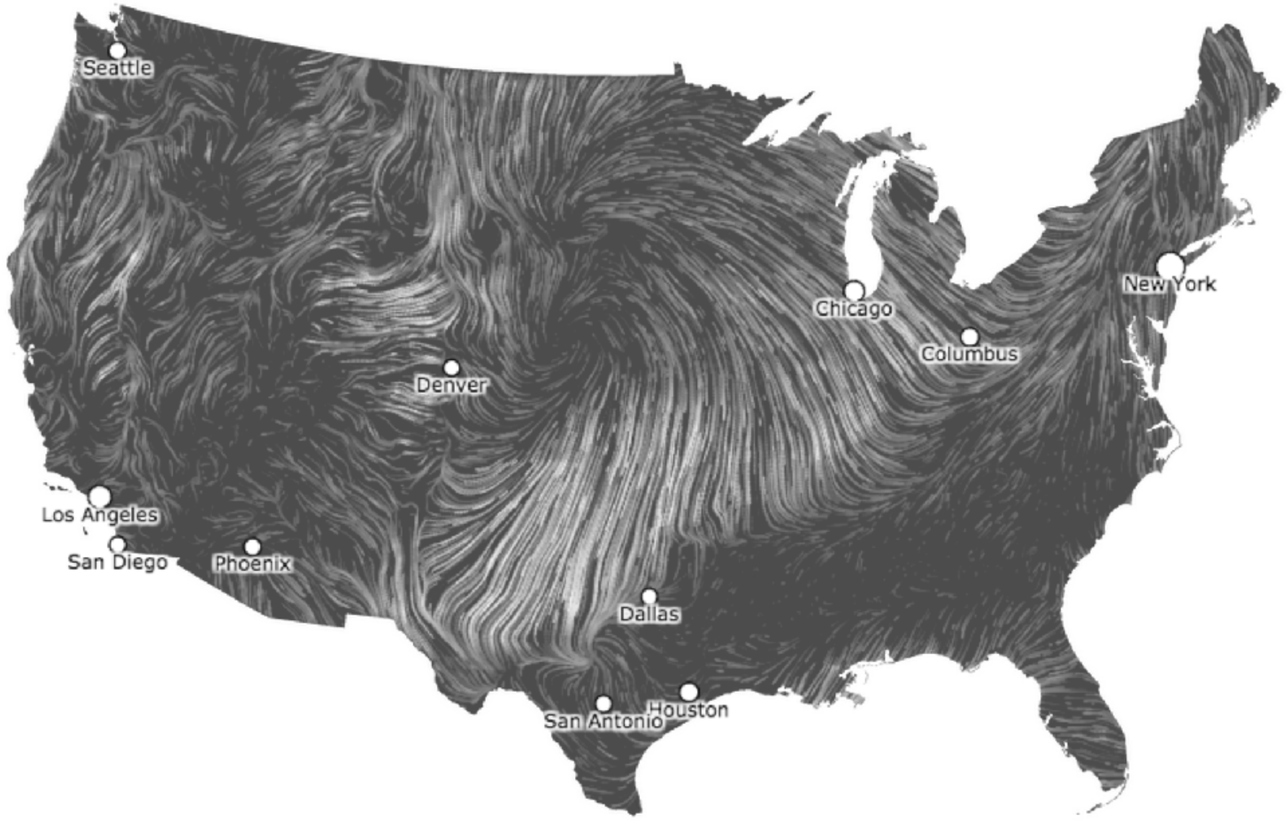


[Wikipedia]

# Visualizing Fields




[Bruckner 2007]




**More in Maps, CS 5635 / 6635 - Visualization for Scientific Data**

# Side Note: Academic Subfields

**Information Vis**

"Abstract Data"

- Tables, Graphs

Free to choose spatial layout

**Visual Analytics**

InfoVis + Stats + Machine learning

Applied Work

Funding buzzword

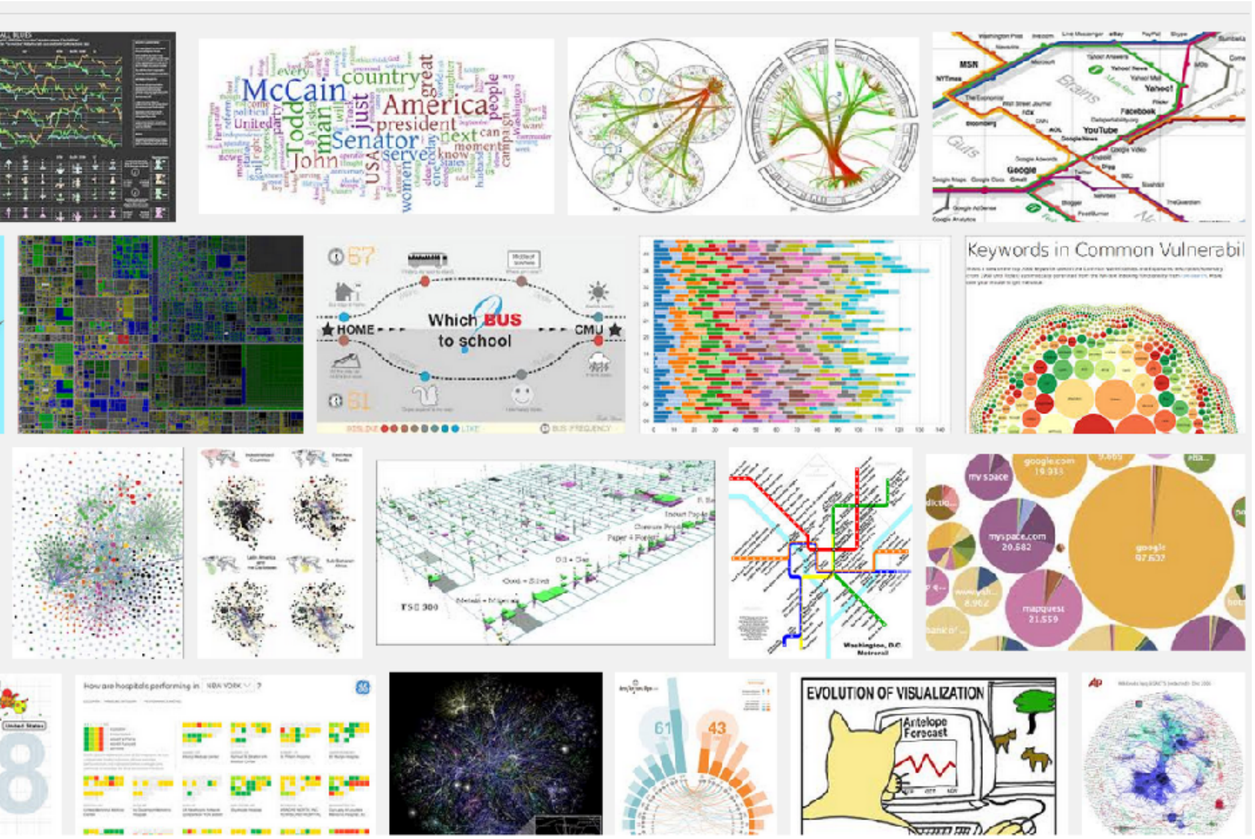
**Scientific Vis**

"Spatial Data" (Fields)

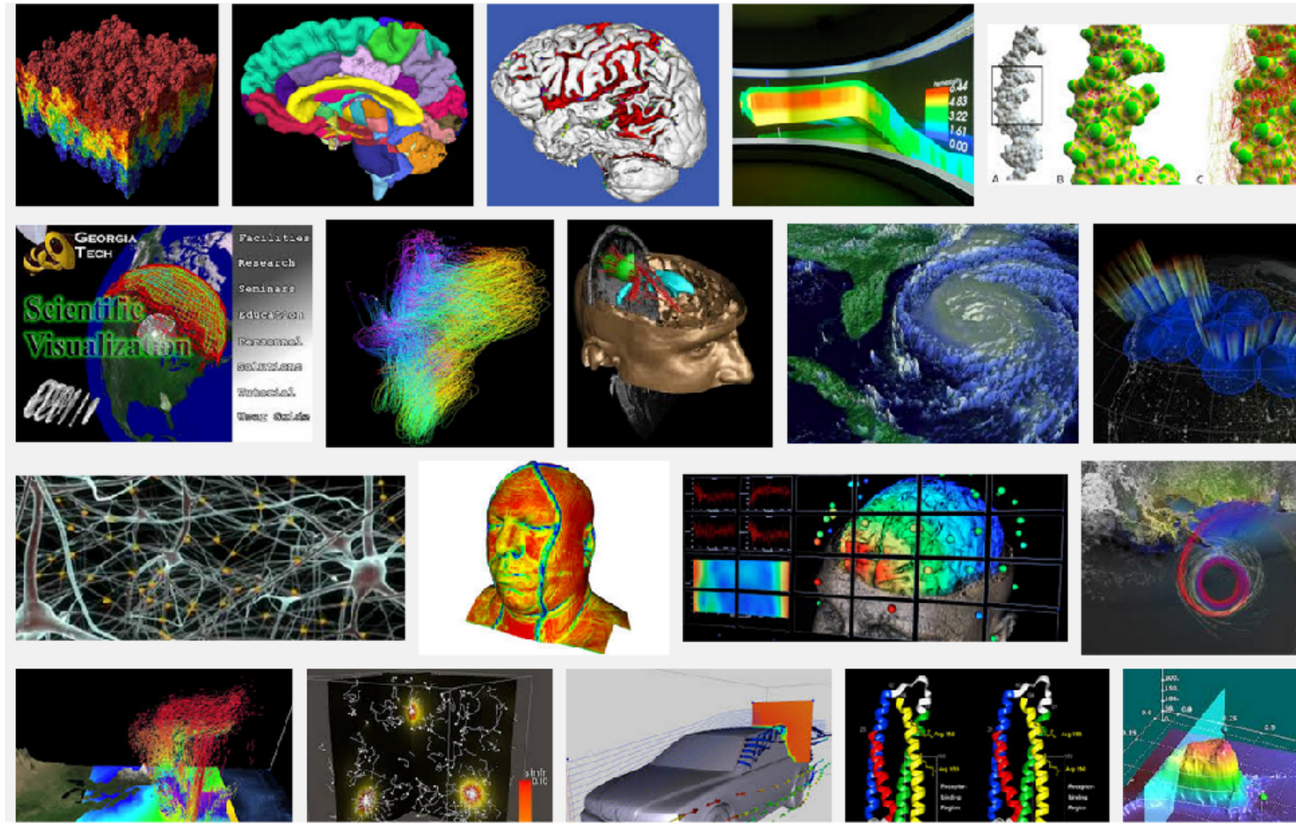
Not free to choose spatial layout

Find best way to depict reality

# InfoVis or SciVis?

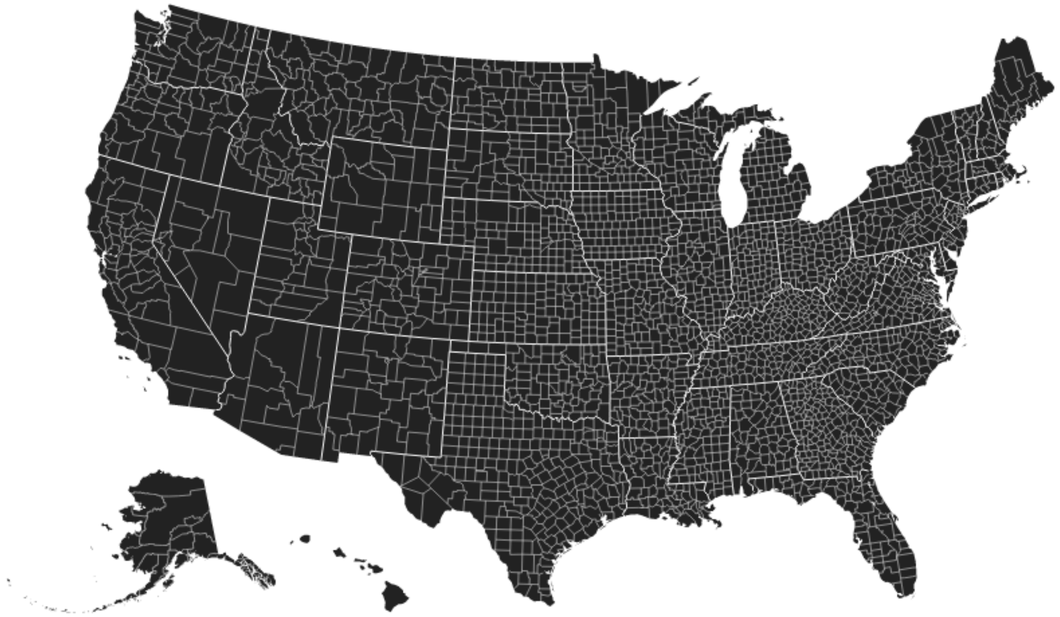



InfoVis: White Background




SciVis: Black Background

# Geometry



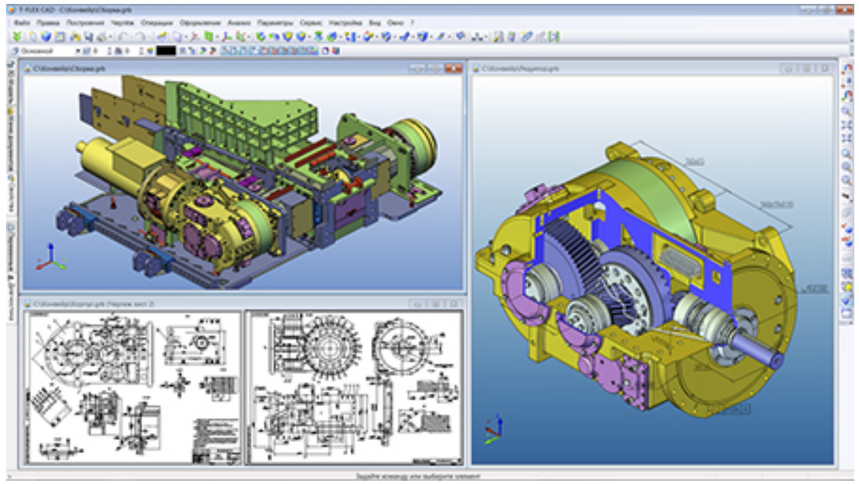
Shape of items

Explicit spatial positions

Points, lines, curves, surfaces, regions, volumes

Important in Computer Graphics, CAD, …

Not a core Vis topic



# Other Collections

## Sets

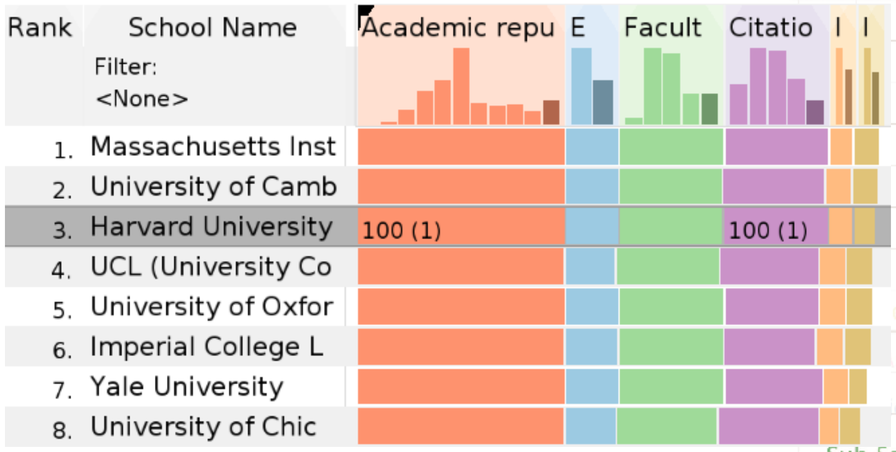
Unique items, unordered

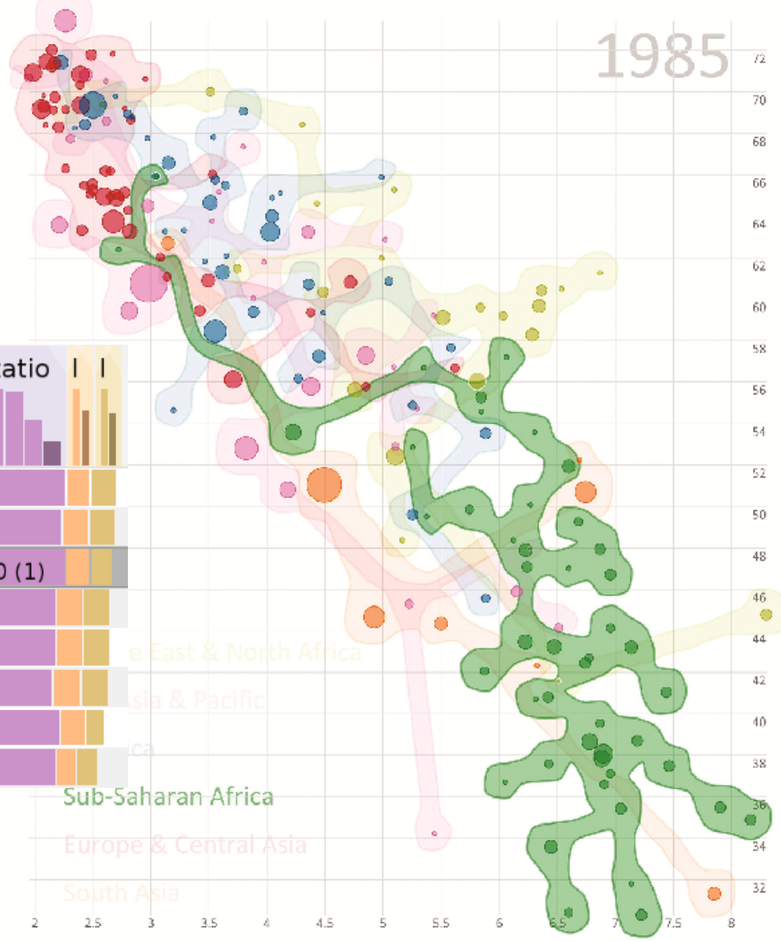
## Lists

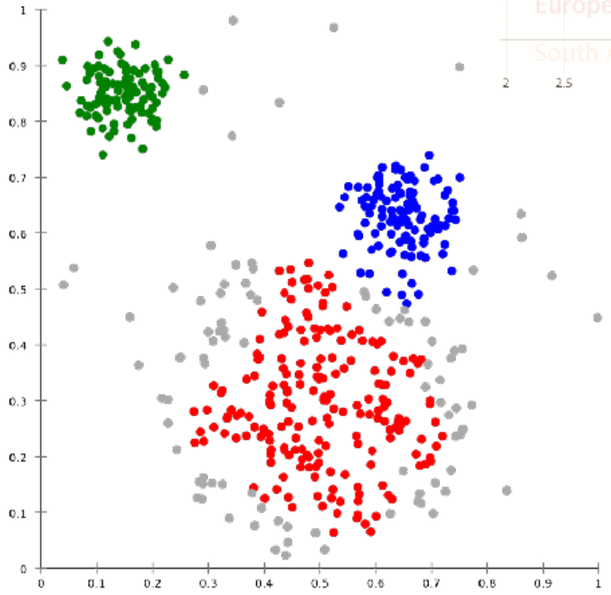
Ordered, duplicates allowed

## Clusters

Groups of similar items



# Attribute Types

Which classes of values & measurements are there?

Categorical (nominal)

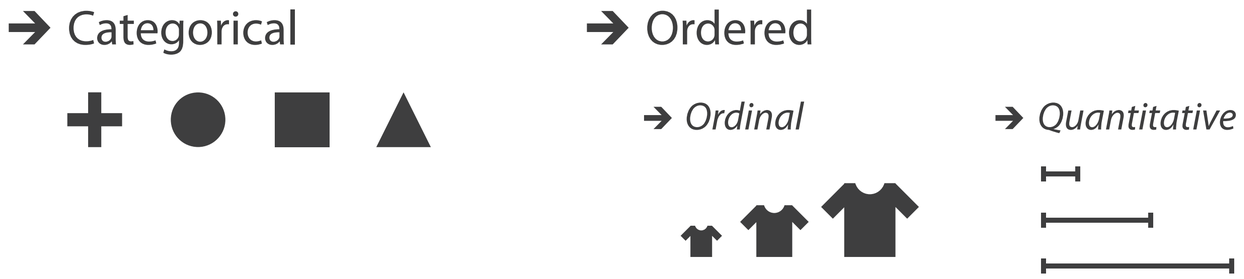
- Compare equality
- *Fruit, Gender, Movie Genres, File Types*

Ordered

- Ordinal
  - Great/Less than defined
  - *Shirt size, Rankings*

Quantitative

- Arithmetic possible
- *Length, Weight, Count*

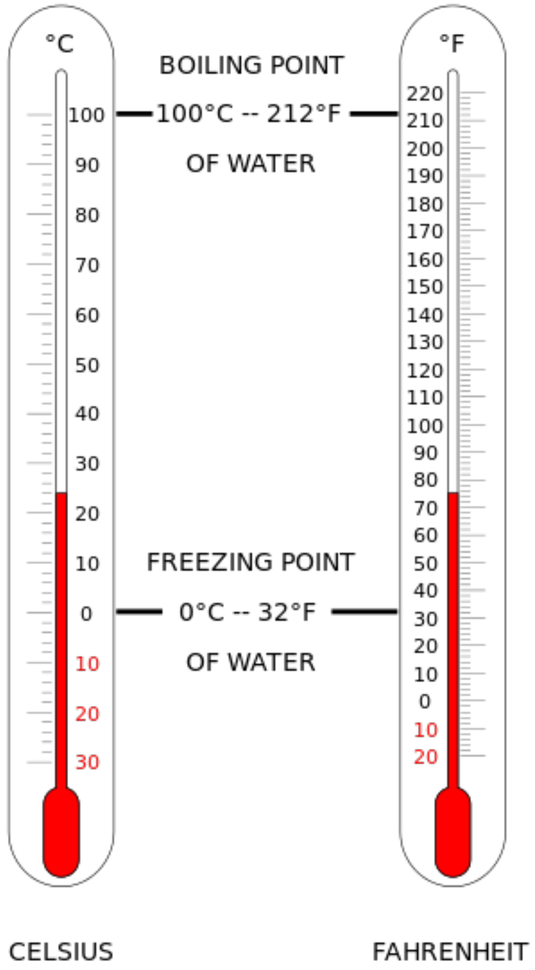


# Quantitative Data Type: Interval

There are equal differences between successive points on the scale but the position of zero is arbitrary.

Does Zero  mean none?

- Dates:  Jan 19;  Location:  (Lat, Long)
- Cannot compare directly. Temp in C & F
- Only differences (i.e., intervals) can be compared



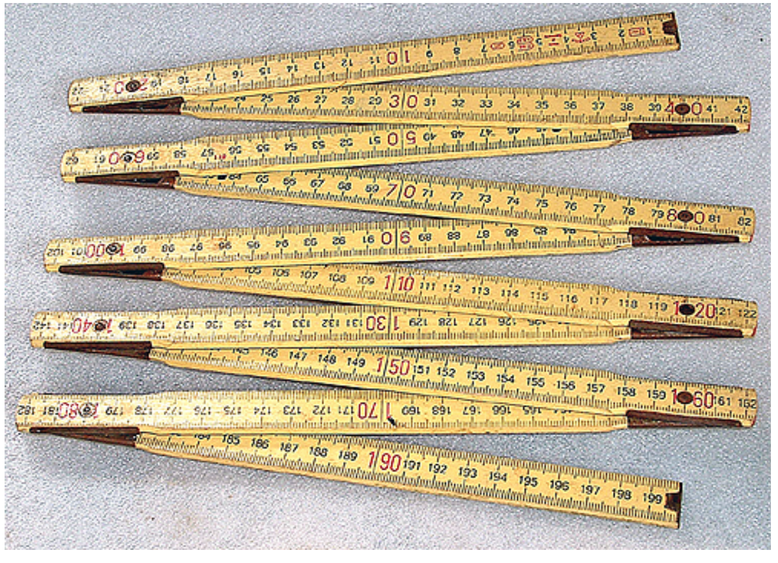
# Quantitative Data Types: Ratio

The relative magnitudes of scores and the differences between them matter. The position of zero is fixed.

Zero: there is nothing of the measured entity observed

Measurements: Length, Mass, Age, Weight

Can measure ratios & proportions



SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

# On the Theory of Scales of Measurement

S. S. Stevens

*Director, Psycho-Acoustic Laboratory, Harvard University*

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

## A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads

| Scale | Basic Empirical Operations | Mathematical Group Structure | Permissible Statistics (invariantive) |
|---|---|---|---|
| NOMINAL | Determination of equality | *Permutation group*<br>$x' = f(x)$<br>$f(x)$ means any one-to-one substitution | Number of cases<br>Mode<br>Contingency correlation |
| ORDINAL | Determination of greater or less | *Isotonic group*<br>$x' = f(x)$<br>$f(x)$ means any monotonic increasing function | Median<br>Percentiles |
| INTERVAL | Determination of equality of intervals or differences | *General linear group*<br>$x' = ax + b$ | Mean<br>Standard deviation<br>Rank-order correlation<br>Product-moment correlation |
| RATIO | Determination of equality of ratios | *Similarity group*<br>$x' = ax$ | Coefficient of variation |

On the theory of scales and measurements [S. Stevens, 46]

# Data Types

Nominal (categories, labels)

Operations: =, ≠

Ordinal (ordered)

Operations: =, ≠, >, <

Interval (location of zero arbitrary)

Operations: =, ≠, >, <, +, – (distance)

Ratio (zero fixed)

Operations: =, ≠, >, <, +, –,×, ÷ (proportions)

On the theory of scales and measurements [S. Stevens, 46]

# Quiz!

What type of variable (Nominal, Ordinal, Interval, or Ratio) are the following:

1. 50 meter race times
2. College major
3. Amazon rating for a product
4. IQ Score
5. Product Name

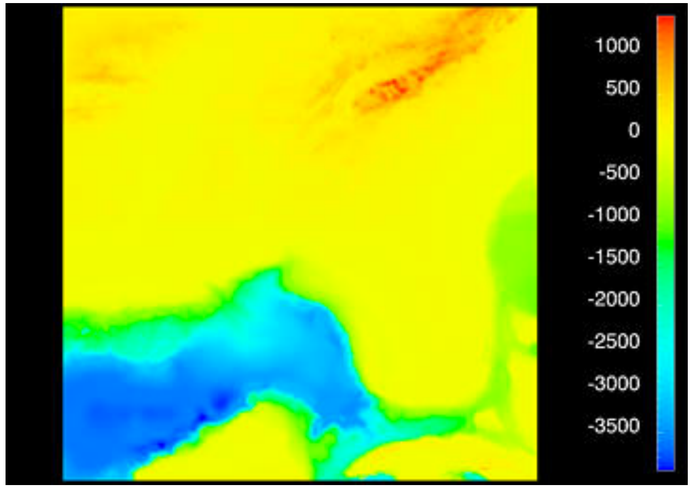
# Sequential & Diverging Data

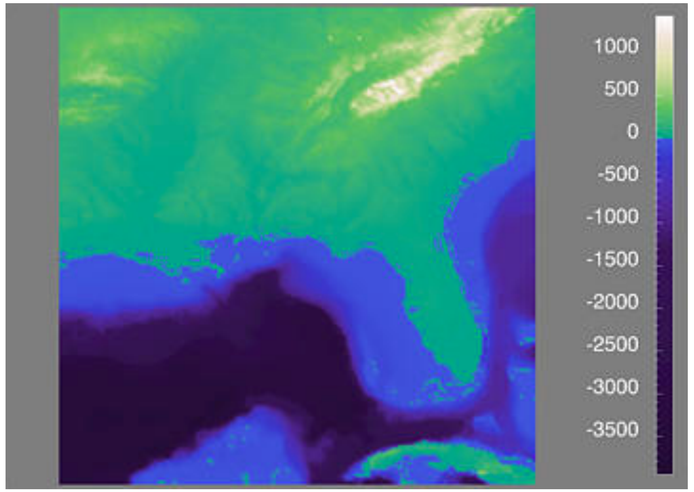
Sequential:

- homogeneous from min to max
- # people in countries

Diverging:

- two or multiple sequences that meet
- Elevation dataset: above sea level & below sea level



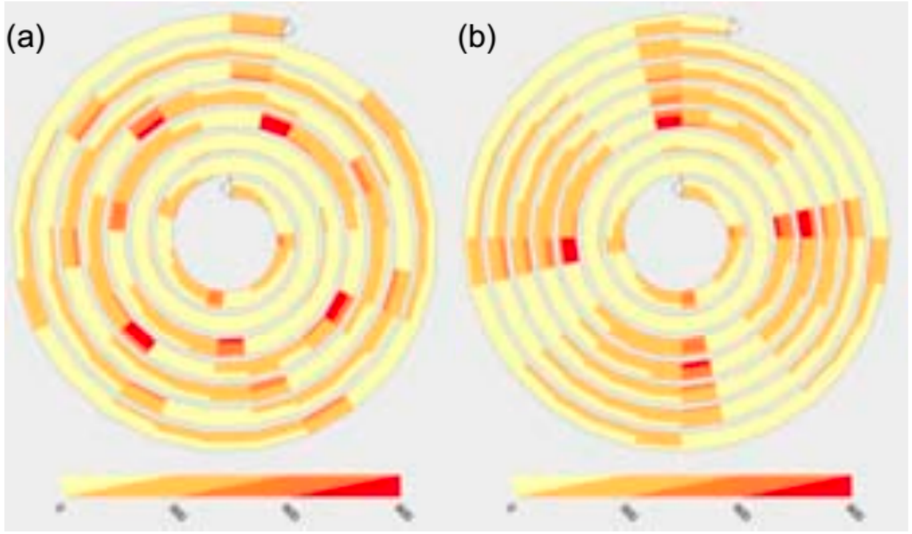
# Other Structure

Cyclic data

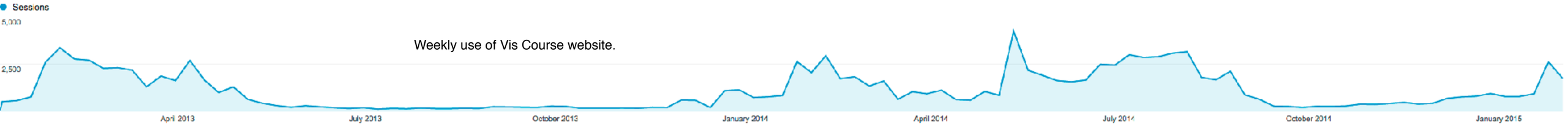
time (hours, week, month, year)

Aggregation

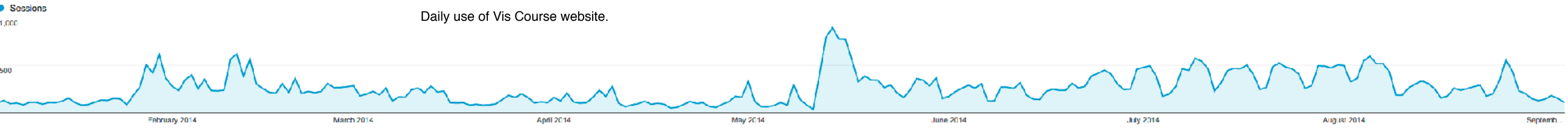
might be patterns on multiple levels




Respiratory disease cases.
Left: 25 day pattern
Right: 28 day pattern
[Tominski 2008]




Weekly use of Vis Course website.




Daily use of Vis Course website.

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | 0.55 | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | 0.79 | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | 0.72 | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | 0.6 | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | 0.65 | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | 0.6 | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | 0.6 | 6/12/06 |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | 0.69 | 6/14/06 |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | 0.82 | 5/3/08 |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | 0.64 | 10/23/07 |
| 27 | 166 | 9/12/07 | 2-High | Small Box | 0.55 | 9/14/07 |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | 0.57 | 8/10/06 |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | 0.42 | 4/7/08 |

Item/Element/
(Independent)
Variable

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | 0.44 | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | 0.6 | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | 0.59 | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | 0.82 | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | 0.55 | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | 0.6 | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | 0.6 | 6/12/06 |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | 0.69 | 6/14/06 |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | 0.82 | 5/3/08 |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | 0.64 | 10/23/07 |
| 27 | 166 | 9/12/07 | 2-High | Small Box | 0.55 | 9/14/07 |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | 0.57 | 8/10/06 |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | 0.42 | 4/7/08 |

Attribute/
Dimension/
(Dependent)
Variable/
Feature

Sheets Charts SmartArt Graphics WordArt

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | 0.5 | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | 0.9 | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | 0.72 | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | 0.6 | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | 0.65 | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | 0.52 | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | 0.58 | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | 0.55 | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | 0.49 | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | 0.56 | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | 0.44 | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | 0.6 | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | 0.59 | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | 0.82 | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | 0.55 | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | 0.6 | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | 0.6 | 6/12/06 |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | 0.69 | 6/14/06 |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | 0.82 | 5/3/08 |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | 0.64 | 10/23/07 |
| 27 | 166 | 9/12/07 | 2-High | Small Box | 0.55 | 9/14/07 |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | 0.57 | 8/10/06 |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | 0.42 | 4/7/08 |

Semantics

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | 0.55 | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | 0.79 | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | 0.72 | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | 0.6 | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | 0.58 | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | 0.55 | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | 0.49 | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | 0.56 | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | 0.44 | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | 0.6 | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | 0.59 | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | 0.82 | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | 0.55 | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | 0.6 | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | 0.6 | 6/12/06 |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | 0.69 | 6/14/06 |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | 0.82 | 5/3/08 |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | 0.64 | 10/23/07 |
| 27 | 166 | 9/12/07 | 2-High | Small Box | 0.55 | 9/14/07 |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | 0.57 | 8/10/06 |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | 0.42 | 4/7/08 |

Keys?

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | 0.55 | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | 0.79 | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | 0.72 | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | 0.6 | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | 0.65 | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | 0.52 | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | 0.58 | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | 0.55 | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | 0.49 | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | 0.56 | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | 0.44 | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | 0.6 | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | 0.59 | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | 0.82 | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | 0.55 | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | | 6/12/06 |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | | 6/14/06 |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | | 5/3/08 |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | | 10/23/07 |
| 27 | 166 | 9/12/07 | 2-High | Small Box | 0.55 | 9/14/07 |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | 0.57 | 8/10/06 |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | 0.42 | 4/7/08 |

Attribute Types?

| | A | B | C | S | T | U |
|---|---|---|---|---|---|---|
| 1 | Order ID | Order Date | Order Priority | Product Container | Product Base Margin | Ship Date |
| 2 | 3 | 10/14/06 | 5-Low | Large Box | 0.8 | 10/21/06 |
| 3 | 6 | 2/21/08 | 4-Not Specified | Small Pack | 0.55 | 2/22/08 |
| 4 | 32 | 7/16/07 | 2-High | Small Pack | 0.79 | 7/17/07 |
| 5 | 32 | 7/16/07 | 2-High | Jumbo Box | 0.72 | 7/17/07 |
| 6 | 32 | 7/16/07 | 2-High | Medium Box | 0.6 | 7/18/07 |
| 7 | 32 | 7/16/07 | 2-High | Medium Box | 0.65 | 7/18/07 |
| 8 | 35 | 10/23/07 | 4-Not Specified | Wrap Bag | 0.52 | 10/24/07 |
| 9 | 35 | 10/23/07 | 4-Not Specified | Small Box | 0.58 | 10/25/07 |
| 10 | 36 | 11/3/07 | 1-Urgent | Small Box | 0.55 | 11/3/07 |
| 11 | 65 | 3/18/07 | 1-Urgent | Small Pack | 0.49 | 3/19/07 |
| 12 | 66 | 1/20/05 | 5-Low | Wrap Bag | 0.56 | 1/20/05 |
| 13 | 69 | 6/4/05 | 4-Not Specified | Small Pack | 0.44 | 6/6/05 |
| 14 | 69 | 6/4/05 | 4-Not Specified | Wrap Bag | 0.6 | 6/6/05 |
| 15 | 70 | 12/18/06 | 5-Low | Small Box | 0.59 | 12/23/06 |
| 16 | 70 | 12/18/06 | 5-Low | Wrap Bag | 0.82 | 12/23/06 |
| 17 | 96 | 4/17/05 | 2-High | Small Box | 0.55 | 4/19/05 |
| 18 | 97 | 1/29/06 | 3-Medium | Small Box | 0.38 | 1/30/06 |
| 19 | 129 | 11/19/08 | 5-Low | Small Box | 0.37 | 11/28/08 |
| 20 | 130 | 5/8/08 | 2-High | Small Box | 0.37 | 5/9/08 |
| 21 | 130 | 5/8/08 | 2-High | Medium Box | 0.38 | 5/10/08 |
| 22 | 130 | 5/8/08 | 2-High | Small Box | 0.6 | 5/11/08 |
| 23 | 132 | 6/11/06 | 3-Medium | Medium Box | | |
| 24 | 132 | 6/11/06 | 3-Medium | Jumbo Box | | |
| 25 | 134 | 5/1/08 | 4-Not Specified | Large Box | | |
| 26 | 135 | 10/21/07 | 4-Not Specified | Small Pack | | |
| 27 | 166 | 9/12/07 | 2-High | Small Box | | |
| 28 | 193 | 8/8/06 | 1-Urgent | Medium Box | | |
| 29 | 194 | 4/5/08 | 3-Medium | Wrap Bag | | |

Categorical

Ordinal

Quantitative

# Data vs. Conceptual Model

Data Model:  Low-level description of the data

Set with operations, e.g., floats with +, -, /, *

Conceptual Model:  Mental construction

Includes semantics, supports reasoning

| Data | Conceptual |
| --- | --- |
| 1D floats | temperature |
| 3D vector of floats | space |

# Data vs. Conceptual Model

From data model...

32.5, 54.0, -17.3, … (floats)

using conceptual model...

Temperature

to data type

Continuous to 4 significant digits (Q)

Hot, warm, cold (O)

Burned vs. Not burned (N)

# Combinations, Derived Data

Networks can have attributes

Attributes have hierarchies

Data types can be transformed

Real life is complicated…