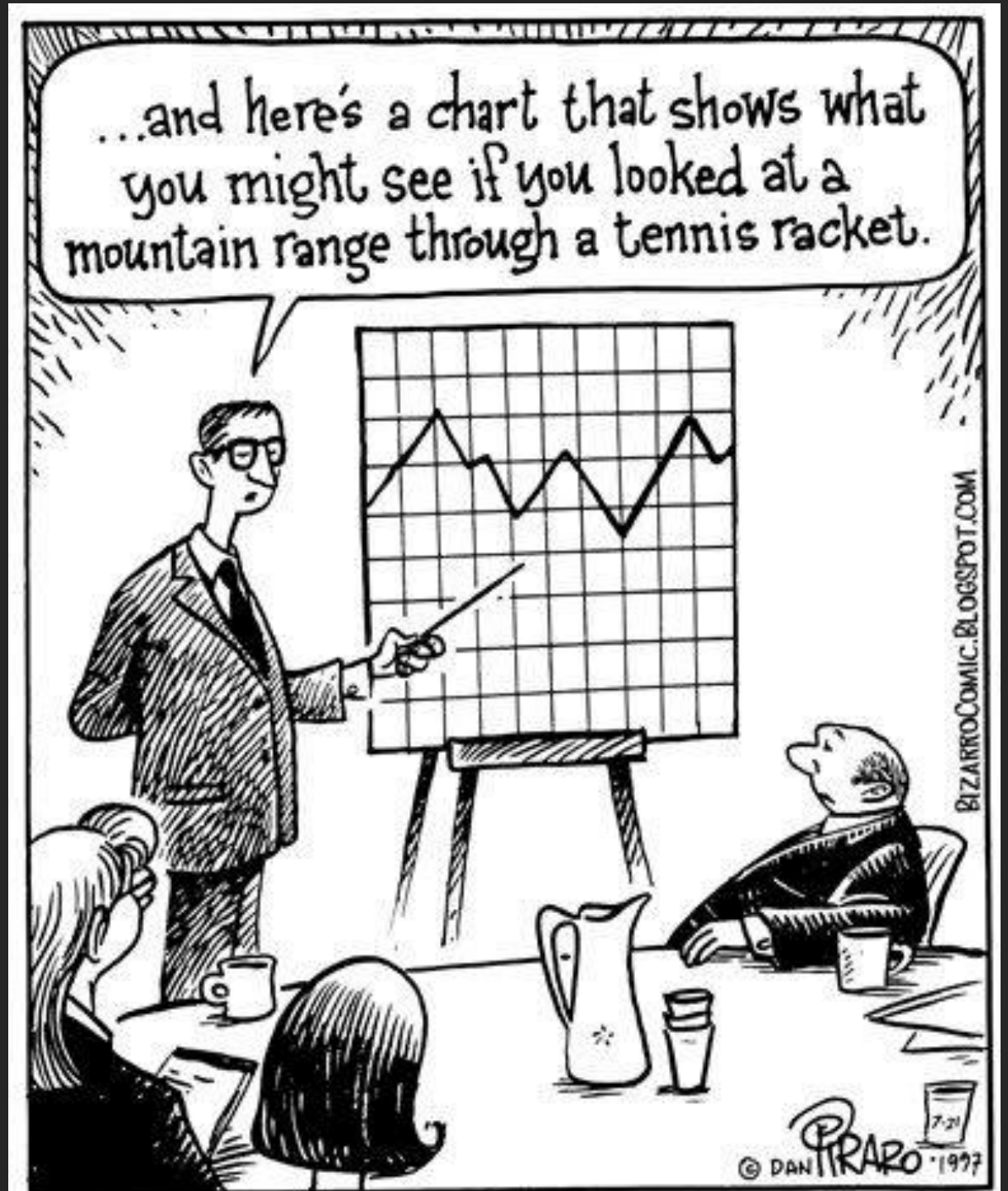
Evaluation
SI649

October 25, 2021

Some slides modified from Stasko ’09, Heer ’09, and Hearst ‘08

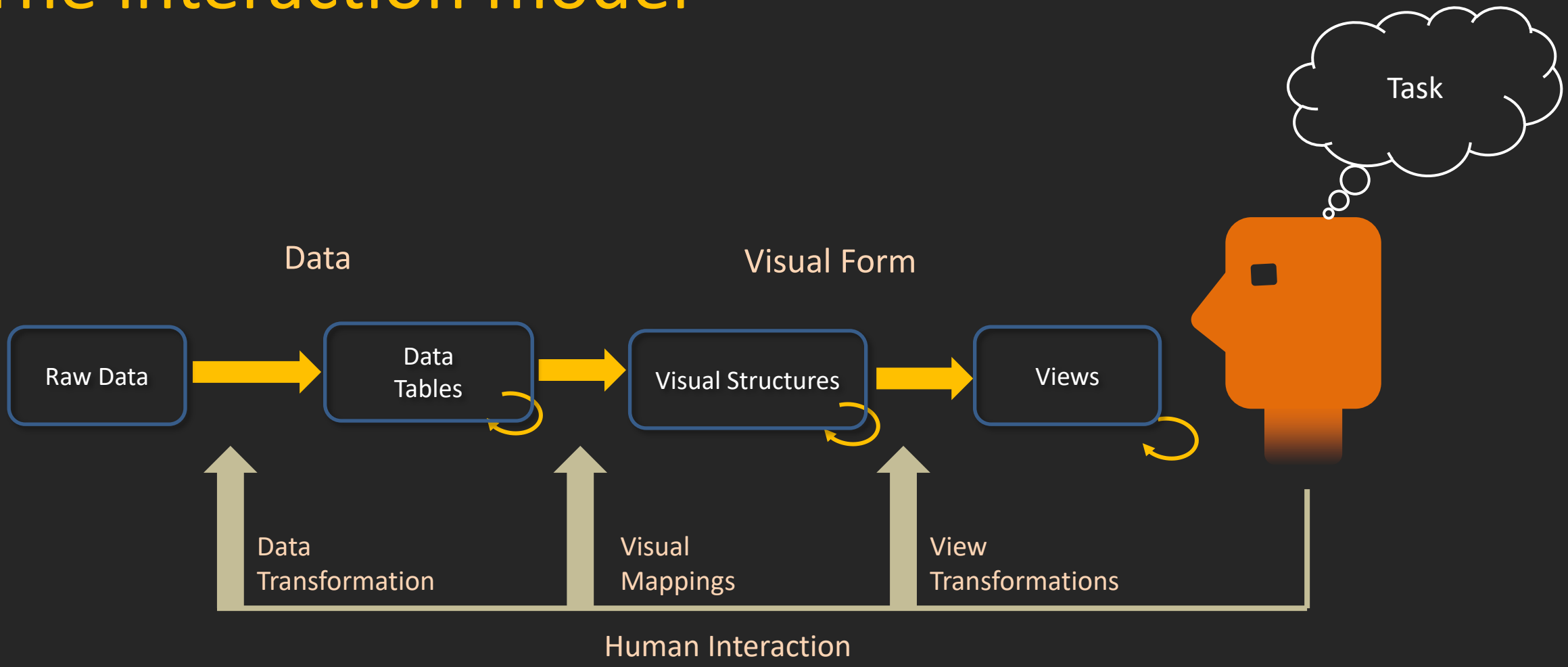


http://www.slido.com

event code **#C164**

Administrivia…

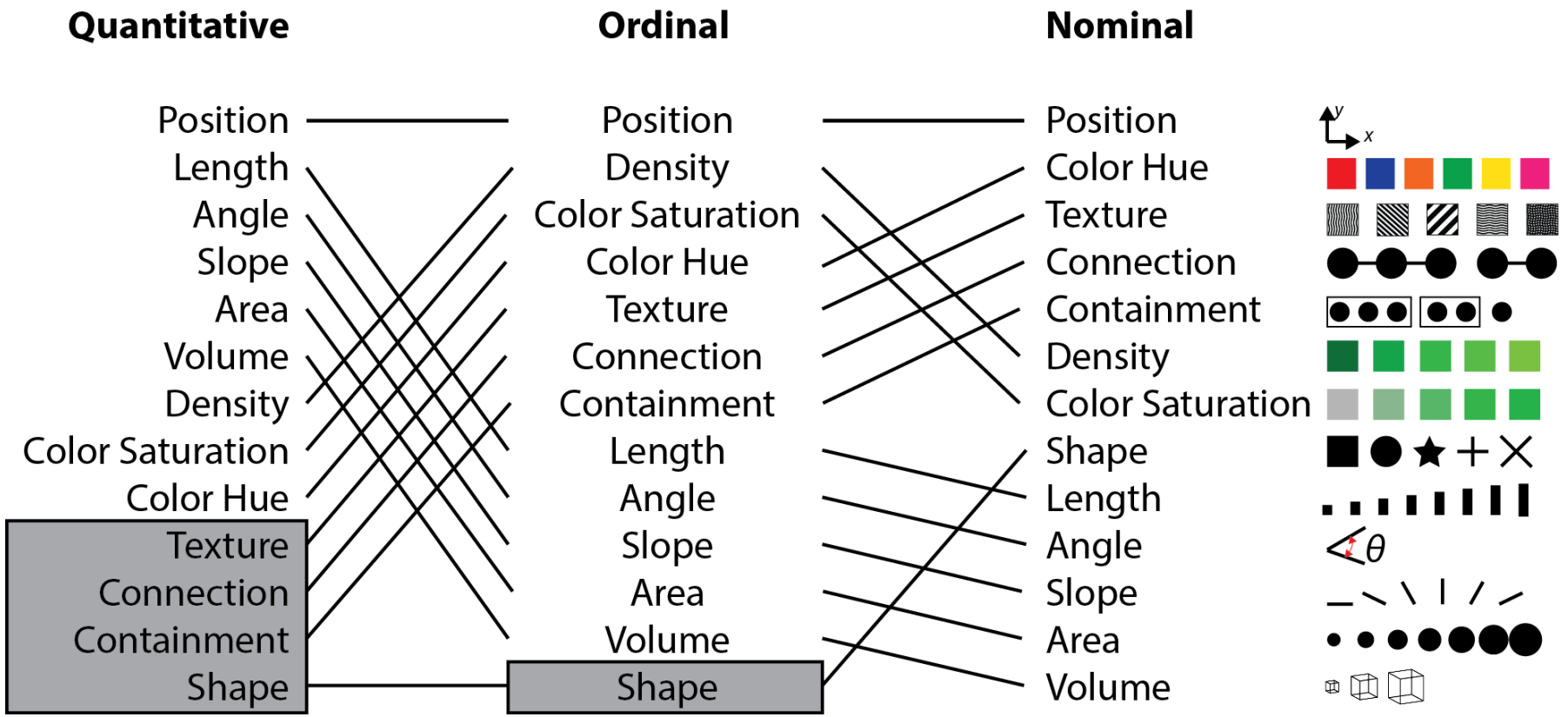
The interaction model



Lots of advice out there…

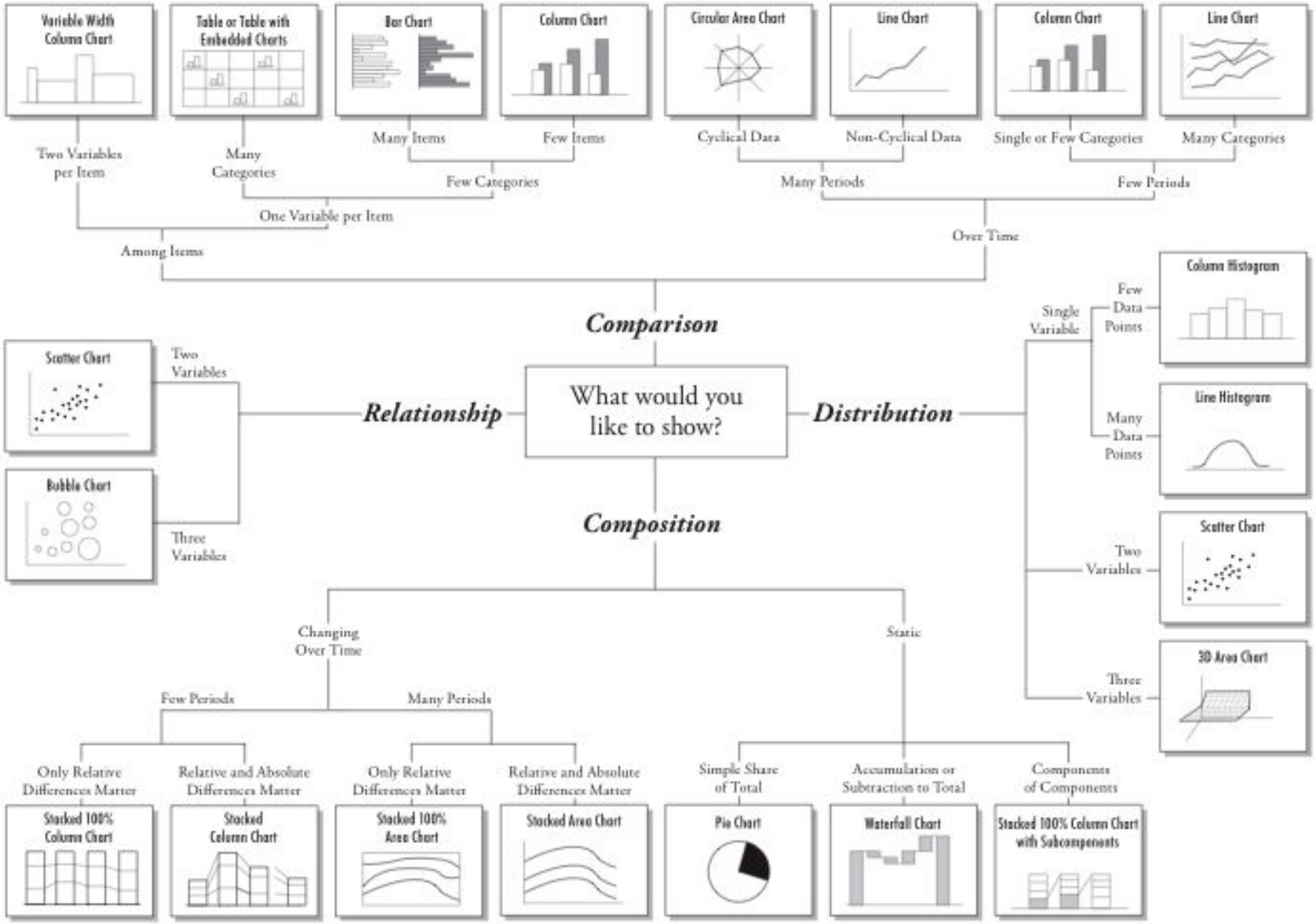
Be greedy:

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



From Mackinlay

Chart Suggestions—A Thought-Starter

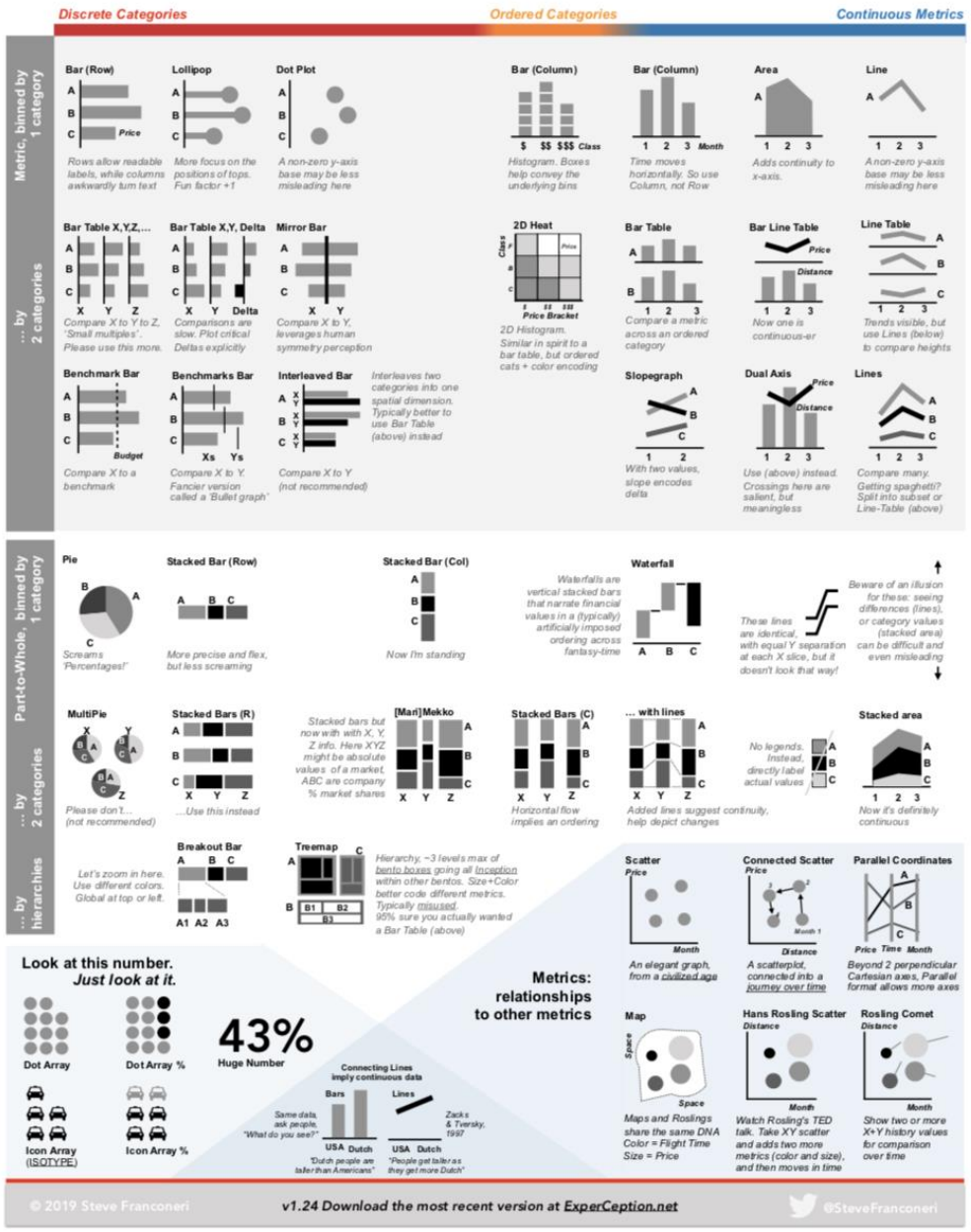


Which Visualization? A Quick Reference

You have the following data (sample):
Discrete Categories,
Ordered categories,
and Continuous Metrics

Here's how to plot them

Categories		Ordered Cats		Continuous Metrics			
City	Airline	Class	PriceBracket	Month	Distance	FlightTime	Price
Alphaville	XeroTrip	Coach	$	1	300	120	250
Betastan	YoloFly	Business	$$	2	500	185	1,525
Chicago	ZeusAir	First	$$$	3	650	240	4,023
...	...	...	...	...	...	...	...



and Continuous Metrics

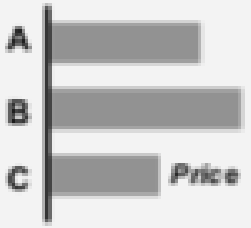
Here's how to plot them

Alphaville	XeroTrip
Betastan	YoloFly
Chicago	ZeusAir
...	...

Discrete Categories

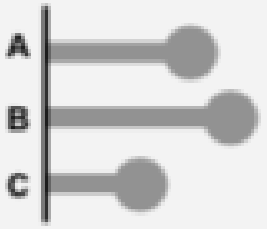
Metric, binned by 1 category

Bar (Row)



Rows allow readable labels, while columns awkwardly turn text

Lollipop



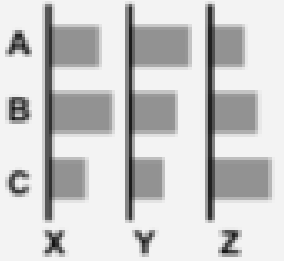
More focus on the positions of tops. Fun factor +1

Dot Plot

A non-zero y-axis base may be less misleading here

... by 2 categories

Bar Table X,Y,Z,...

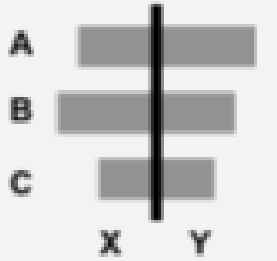


Compare X to Y to Z, "Small multiples". Please use this more.

Bar Table X,Y, Delta

Comparisons are slow. Plot critical Deltas explicitly

Mirror Bar



Compare X to Y, leverages human symmetry perception

QUALITATIVE CHART CHOOSER 3.0

BY JENNIFER LYONS & STEPHANIE EVERGREEN

TRY THIS VIZ

WHAT STORY ARE YOU TRYING TO TELL?

THE DATA ARE...

	QUANTIFICATION									NON-QUANTIFICATION										
	WAYS TO HIGHLIGHT A WORD	THEMATIC ANALYSIS								WAYS TO HIGHLIGHT A WORD	THEMATIC ANALYSIS									
	WORD CLOUD	SPECTRUM DISPLAY	HEATMAP	HISTOMAP	BUBBLE CHART	SPEEDOMETER	ICONS AND COLOR CODING	INDICATOR DOTS	VENN DIAGRAM	QUOTE AND PIC	CALLOUT BOX	BOLDED WORDS	CHANGE PHOTOS	CONCEPT MAPPING	NETWORK MAPPING	TIMELINE	JOURNEY MAP	DENDROGRAM	MIND MAPPING	FLOW/PATH DIAGRAM
HIERARCHY		●	●											●				●		
FLOW				●												●	●		●	●
COMPARISON	●	●	●			●	●	●	●				●							
CLUSTER	●	●			●									●	●		●	●	●	●
COMMENT/WORDS	●						●			●	●	●	●							

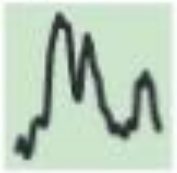
FOR MORE SEE STEPHANIEEVERGREEN.COM/BLOG PRESENTING DATA EFFECTIVELY EFFECTIVE DATA VISUALIZATION

Share price movements, economic time series, sectoral changes in a market

Commodity production, market capitalisation, volumes in general

national election results

Line



The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time - but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small

… how do we know it works?

Notes on current practice

- Current evaluation practice
  - Controlled experiments comparing design elements
  - Usability evaluation
  - Controlled experiments comparing two or more tools
  - Case studies of tools in realistic settings

[Plaisant]

Evaluation “tools”

Decision models

Validation

- Validate twice
  - Once *before* we make a decision (upstream)
  - Once *after* we made it (downstream)
- Example: domain
  - Threat: wrong problem
  - Upstream validation: observe in use
  - Downstream validation: observe adoption

Threat Wrong problem

Validate Observe and interview target users

Threat Wrong task/data abstraction

Threat Ineffective encoding/interaction idiom

Validate Justify encoding/interaction design

Threat Slow algorithm

Validate Analyze computational complexity

Implement system

Validate Measure system time/memory

Validate Qualitative/quantitative result image analysis

Test on any users, informal usability study

Validate Lab study, measure human time/errors for task

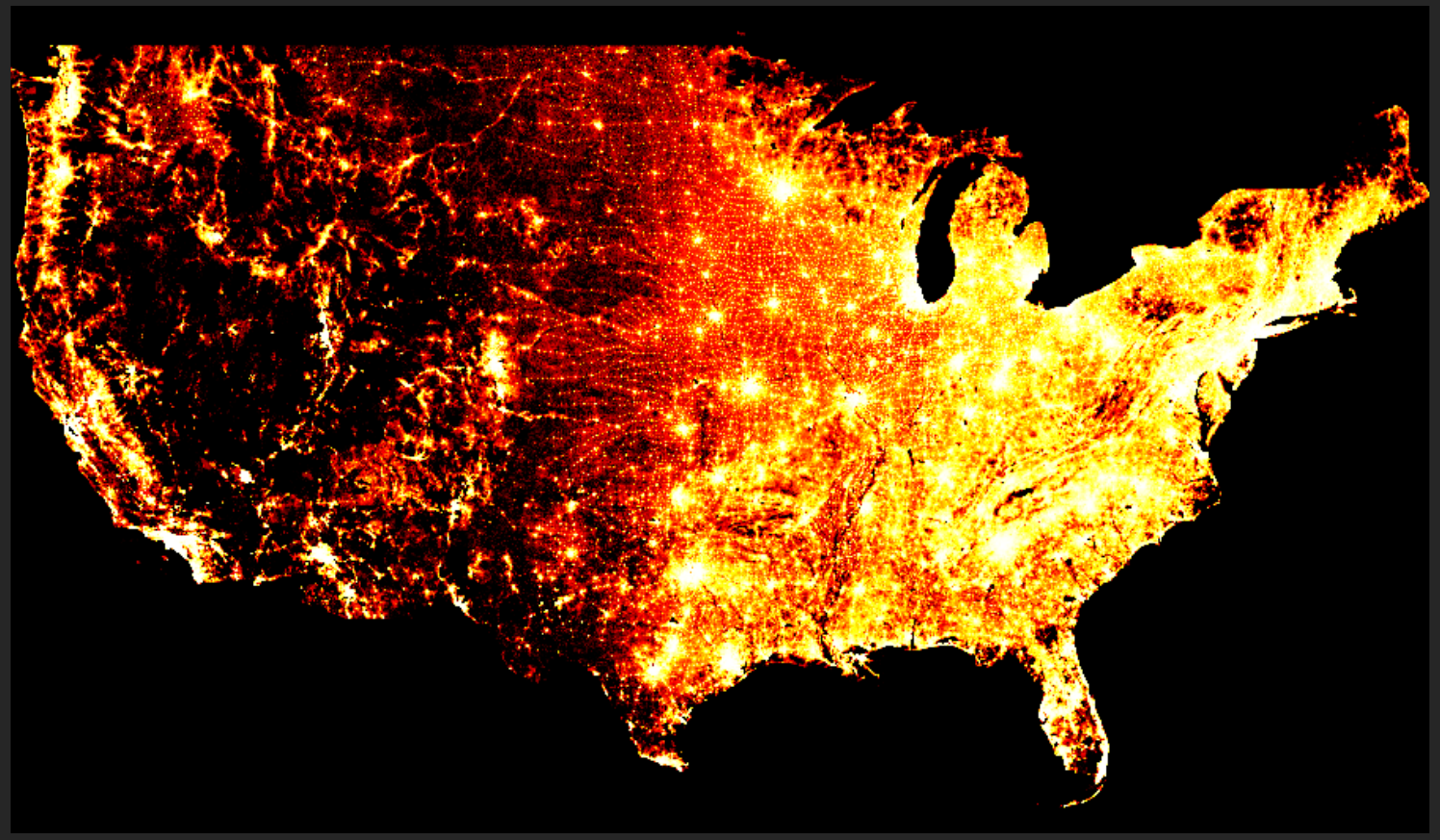
Validate Test on target users, collect anecdotal evidence of utility

Validate Field study, document human usage of deployed system

Validate Observe adoption rates

Algorithm

- Computational complexity



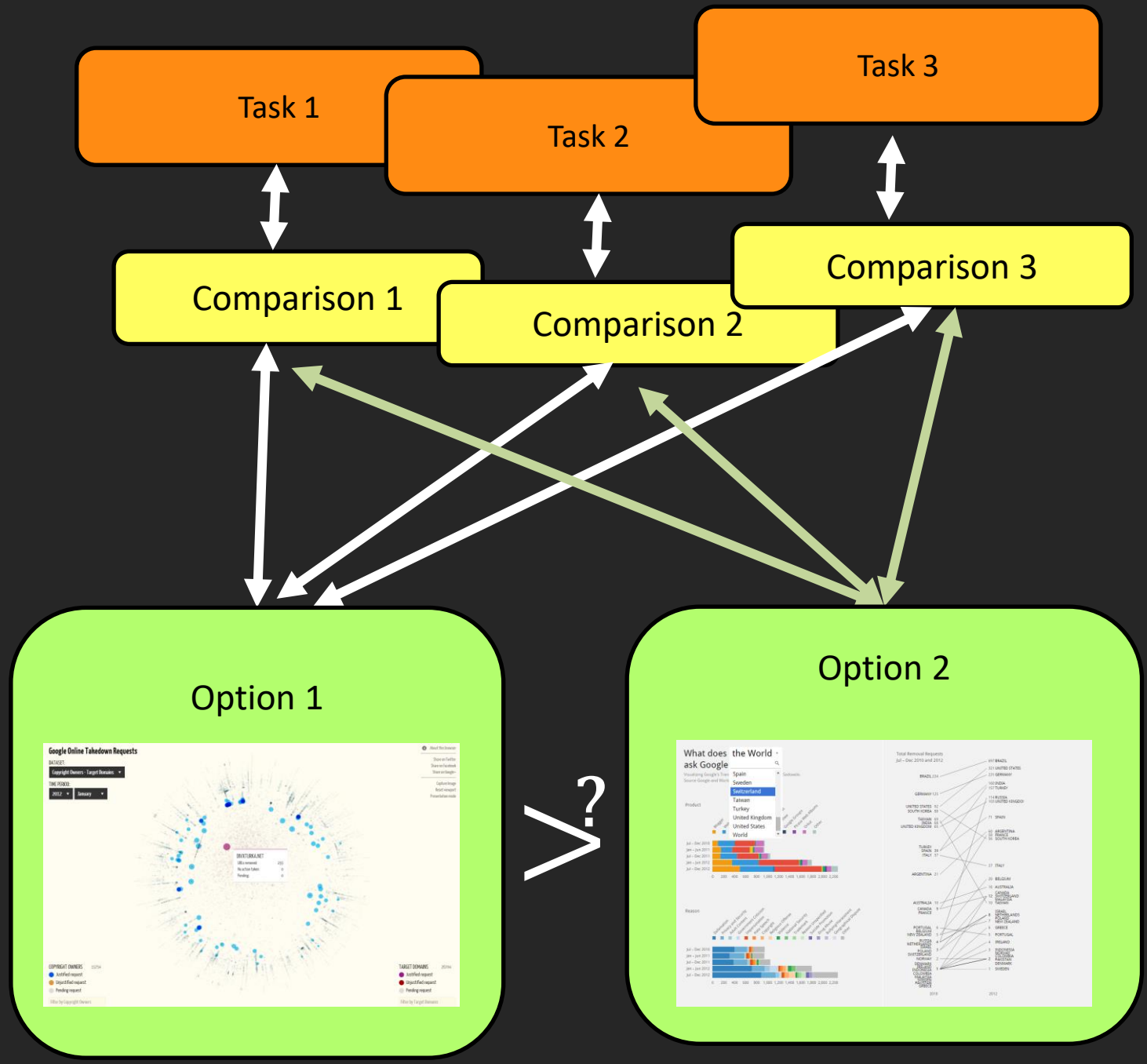
DataShader
https://examples.pyviz.org/census/census.html

InfoVis Design is often a
Wicked Problem

6 Characteristics of Wicked Problems

[Conklin]

- You don’t understand the problem until you have developed a solution
- There is no stopping rule
- Solutions are not right or wrong (but can be better/worse)
- Each is essentially unique and novel
- Every solution is a one-shot operation
- There is no given alternative solution


Task 1
Task 2
Task 3
Comparison 1
Comparison 2
Comparison 3
Option 1
Option 2
>?
Google Online Takedown Requests
What does the World ask Google
Spain
Sweden
Switzerland
Taiwan
Turkey
United Kingdom
United States
World

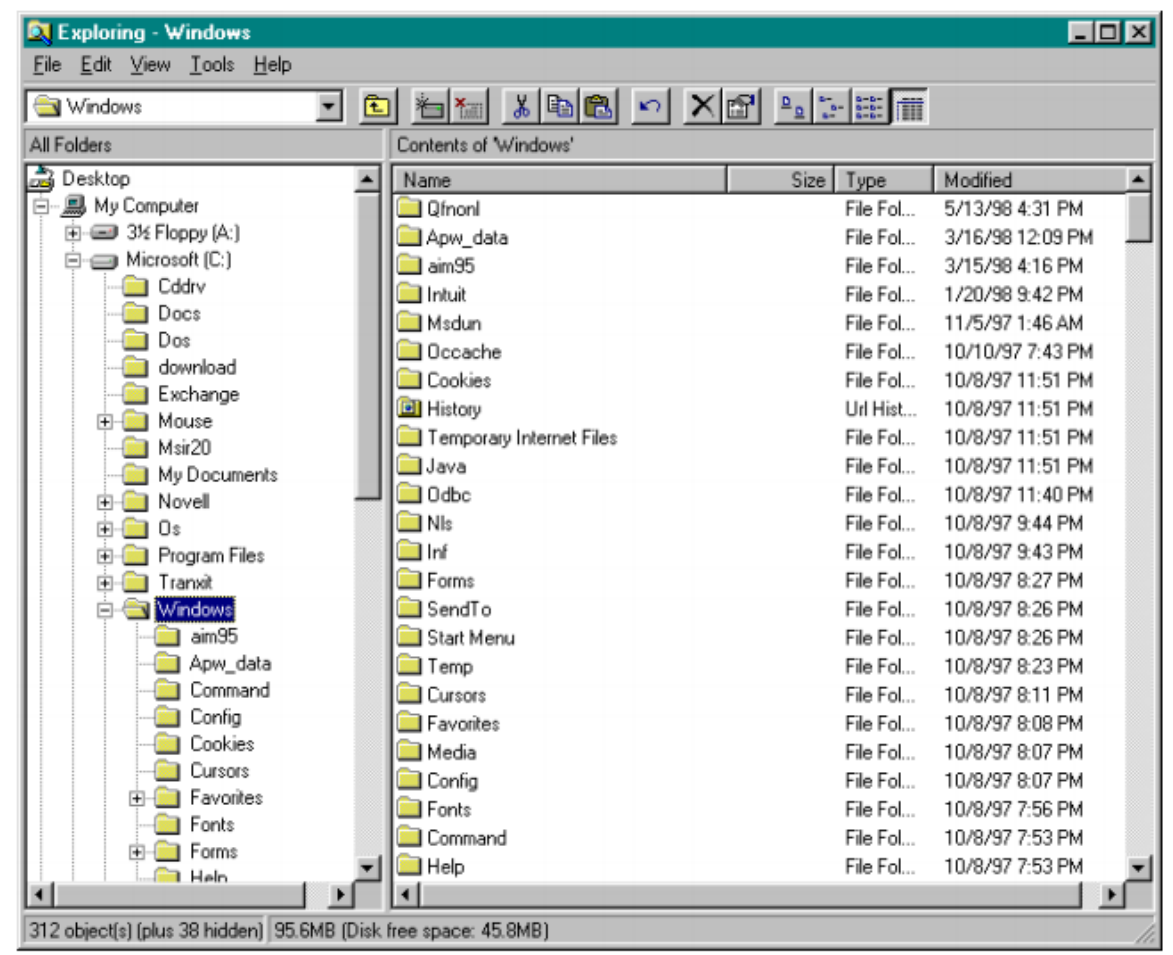
Lesson 1:
Vary experiment space to cover design space
(important when introducing new methods)

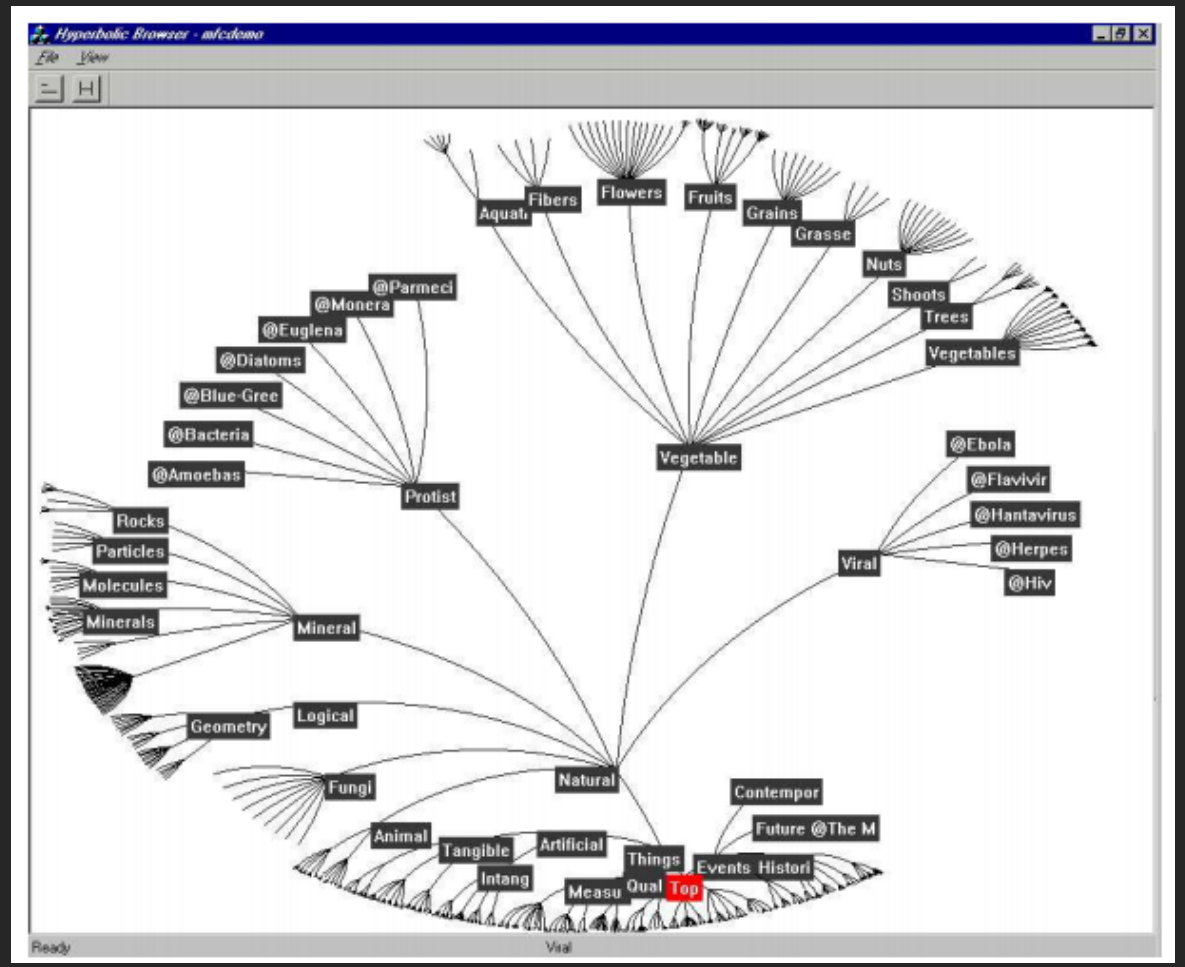
Lesson 2:
Test your solution in the context you want it used

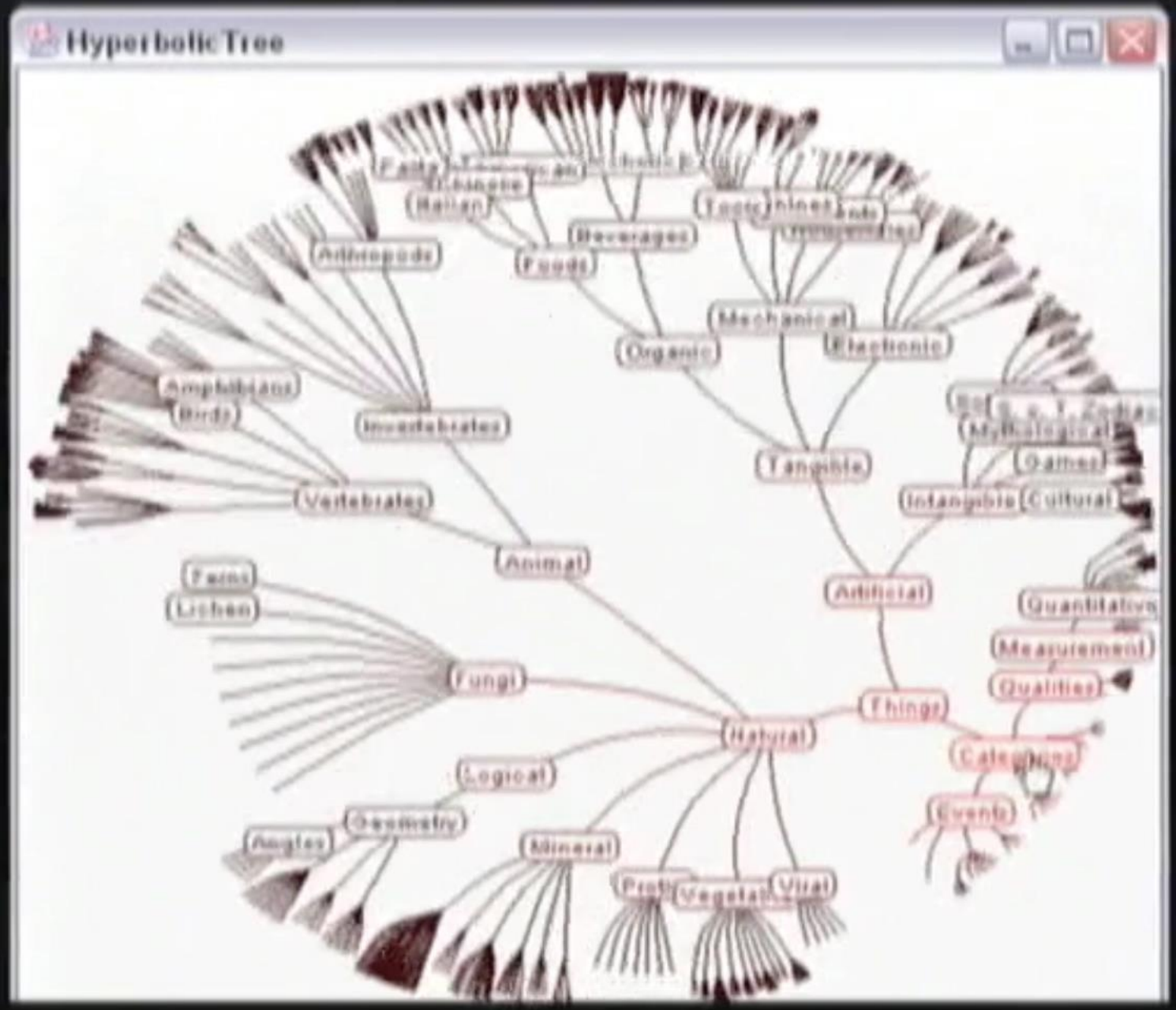
Lesson 3:
Context (the person, the training, the environment, the light, the culture, etc.) matters

The bake-off

CHI’97 Bakeoff





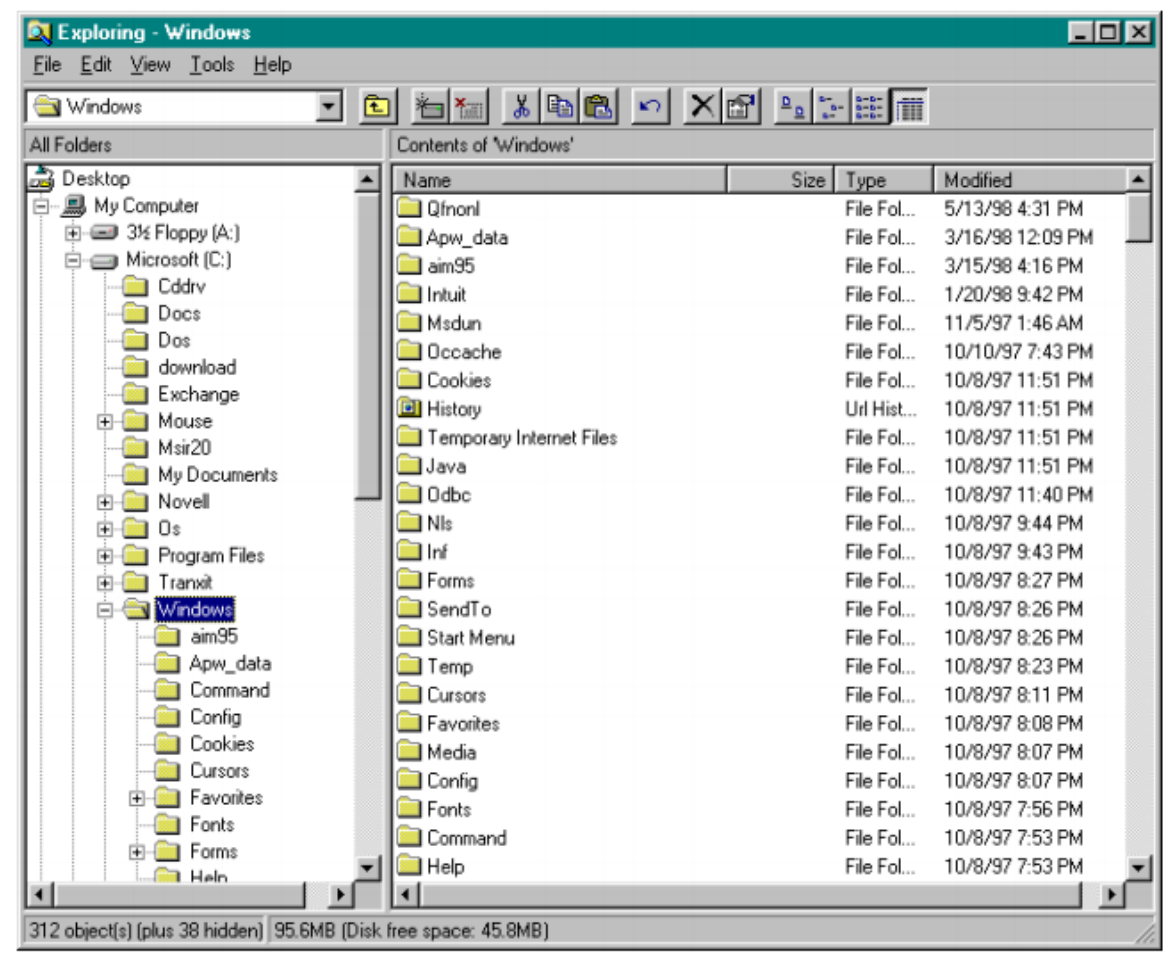

HyperbolicTree
Arthropods
Beverages
Foods
Mechanical
Electronic
Organic
Amphibians
Birds
Invertebrates
Tangible
Games
Intangible
Cultural
Vertebrates
Animal
Artificial
Plants
Lichen
Fungi
Things
Natural
Logical
Geometry
Angles
Mineral
Viral
Events

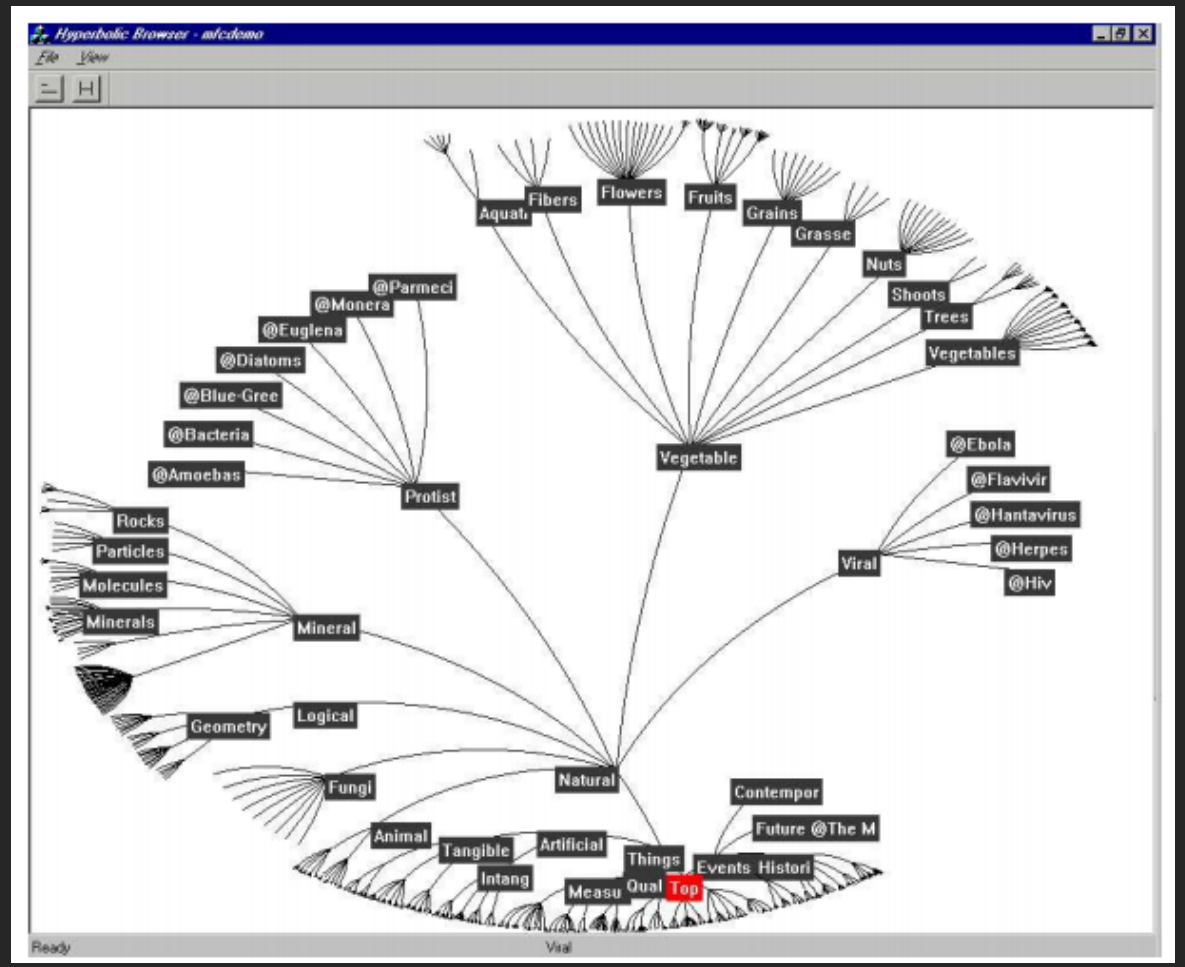
Evaluation

- Expert + Novice
  - “Expert and (for two systems) novice operators used the visualization and browsing tools to complete a set of generic retrieval tasks as quickly and accurately as possible within a large hierarchical data set.”
- Generic retrieval tasks
  - Speed/accuracy test

CHI'97 Bakeoff

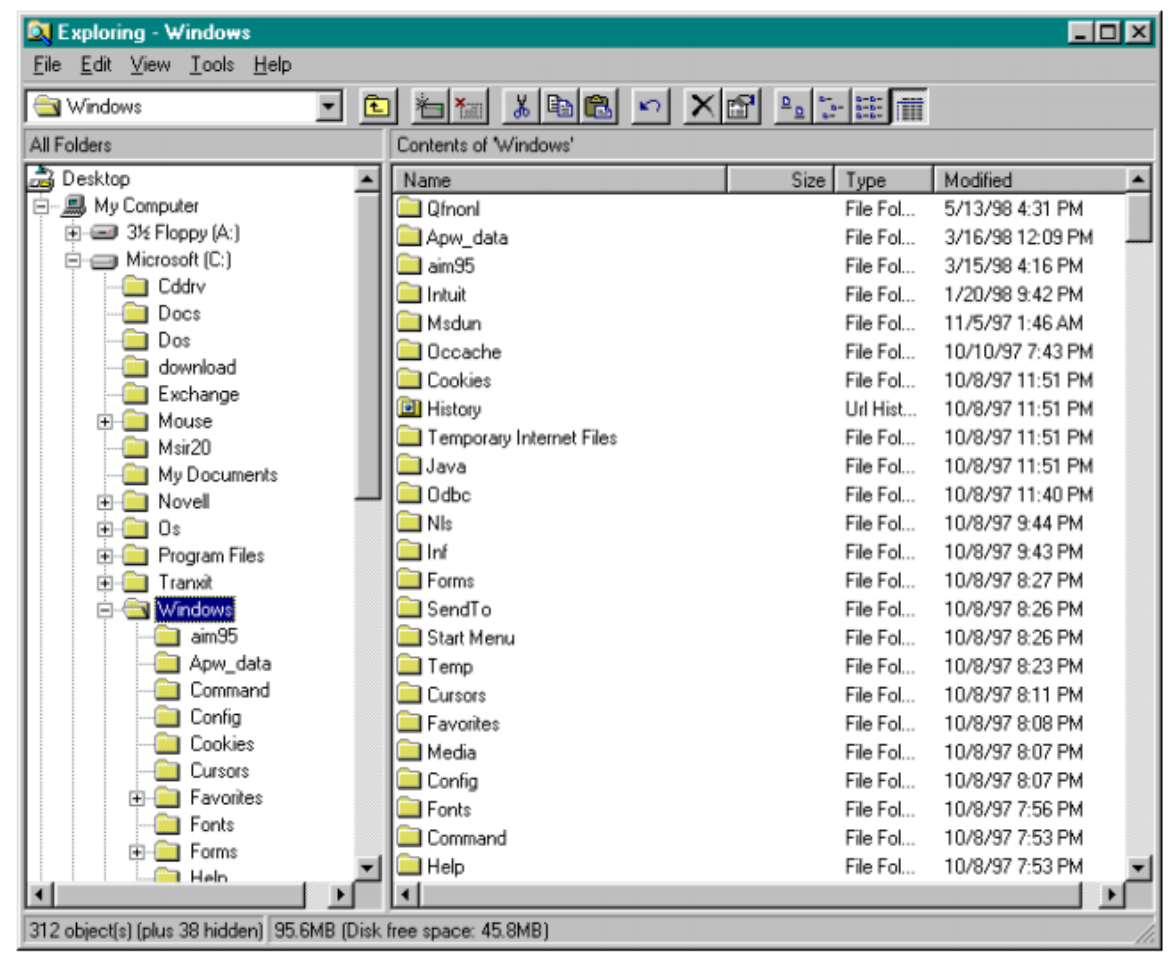
http://www.slido.com
event code **#C164**

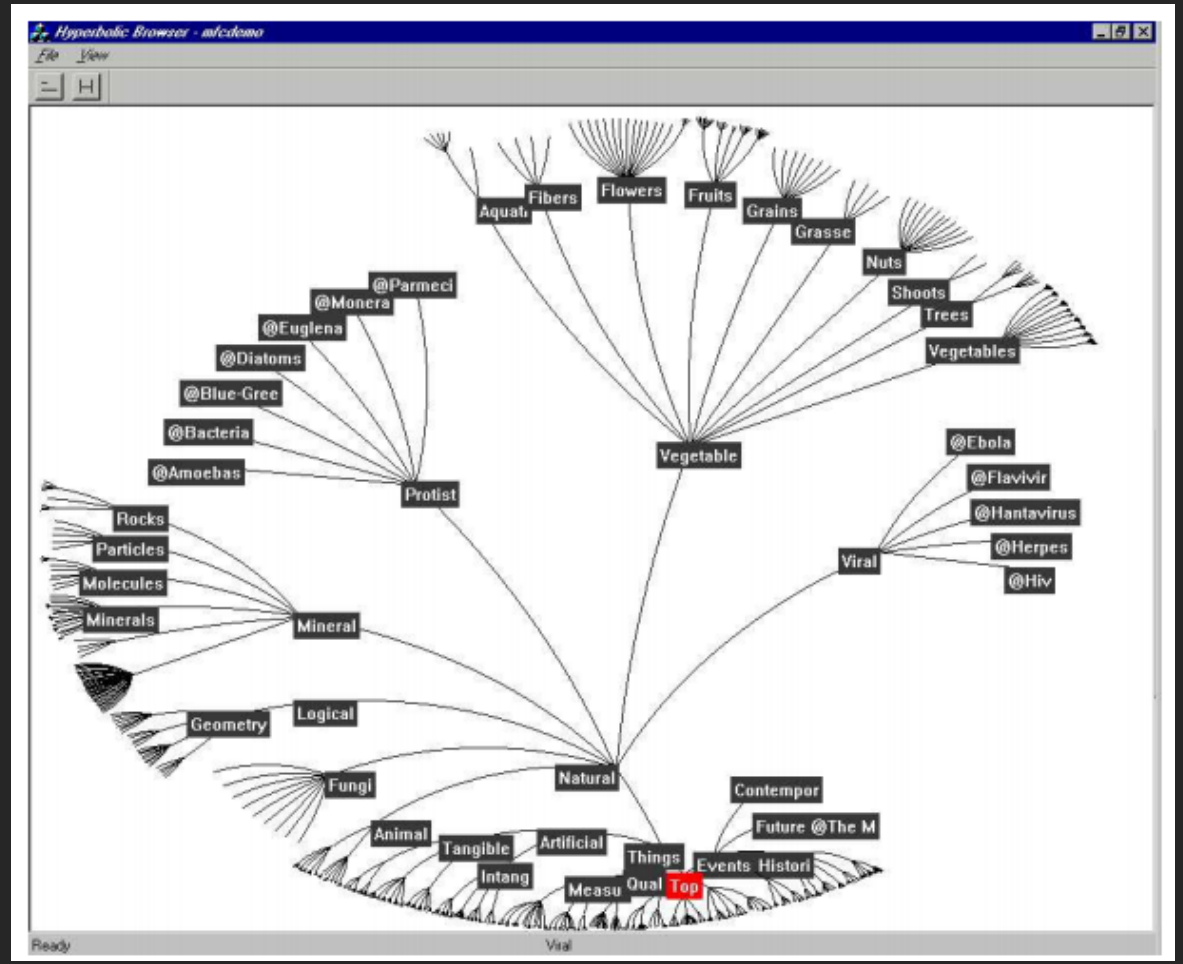




CHI’97 Bakeoff

Winner!





Group question: Can we conclude that the hyperbolic browser is the “best?” Why or why not?

http://www.slido.com
event code **#C164**

Some issues

- Different people driving each
  - "The Hyperbolic Tree proved itself to be extremely responsive, graphically efficient, and *devastatingly effective in the hands of a skilled operator using novel techniques like 'fanning'* the data in a focus-plus-context display."
- Tasks not ecologically valid

Group question: Construct a new experiment. What would you vary?

http://www.slido.com
event code **#C164**

New study [Pirolli et al. 2000]

- Task type 1
  - Retrieval
    - Simple – “find lake Victoria”
    - Complex – “which army is led by a generalissimo?”
  - Comparison
    - Local relational – “which religion has the most holidays?”
    - Complex relational – “Which Greek deity has the same name as a space mission?”

Table 1. Mean performance times in Experiment 1 by task type and browser.

Question Type	Browser	
	Explorer (sec)	Hyperbolic (sec)
Retrieval Tasks		
Simple	35.55	34.37
Complex	41.55	42.02
All retrieval	38.55	38.20
Comparison Tasks		
Local	42.78	41.91
Global	71.07	73.19
All comparison	56.93	57.55
All questions	47.74	47.87

Retrieval

Simple – "find lake Victoria"

Complex – "which army is led by a generalissimo?"

Comparison

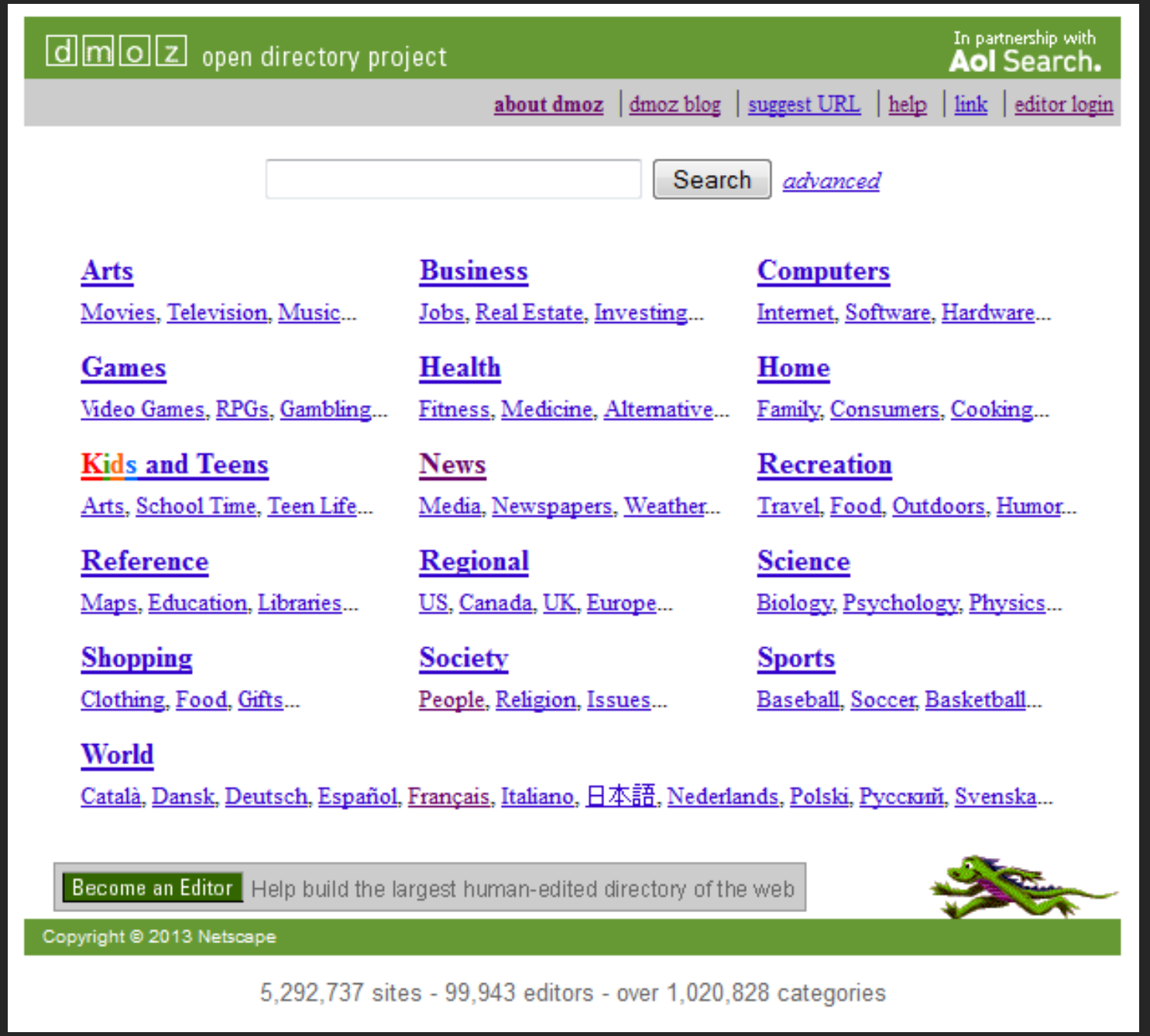
Local relational – "which religion has the most holidays?"

Complex relational – "Which Greek deity has the same name as a space mission?"

What gives?

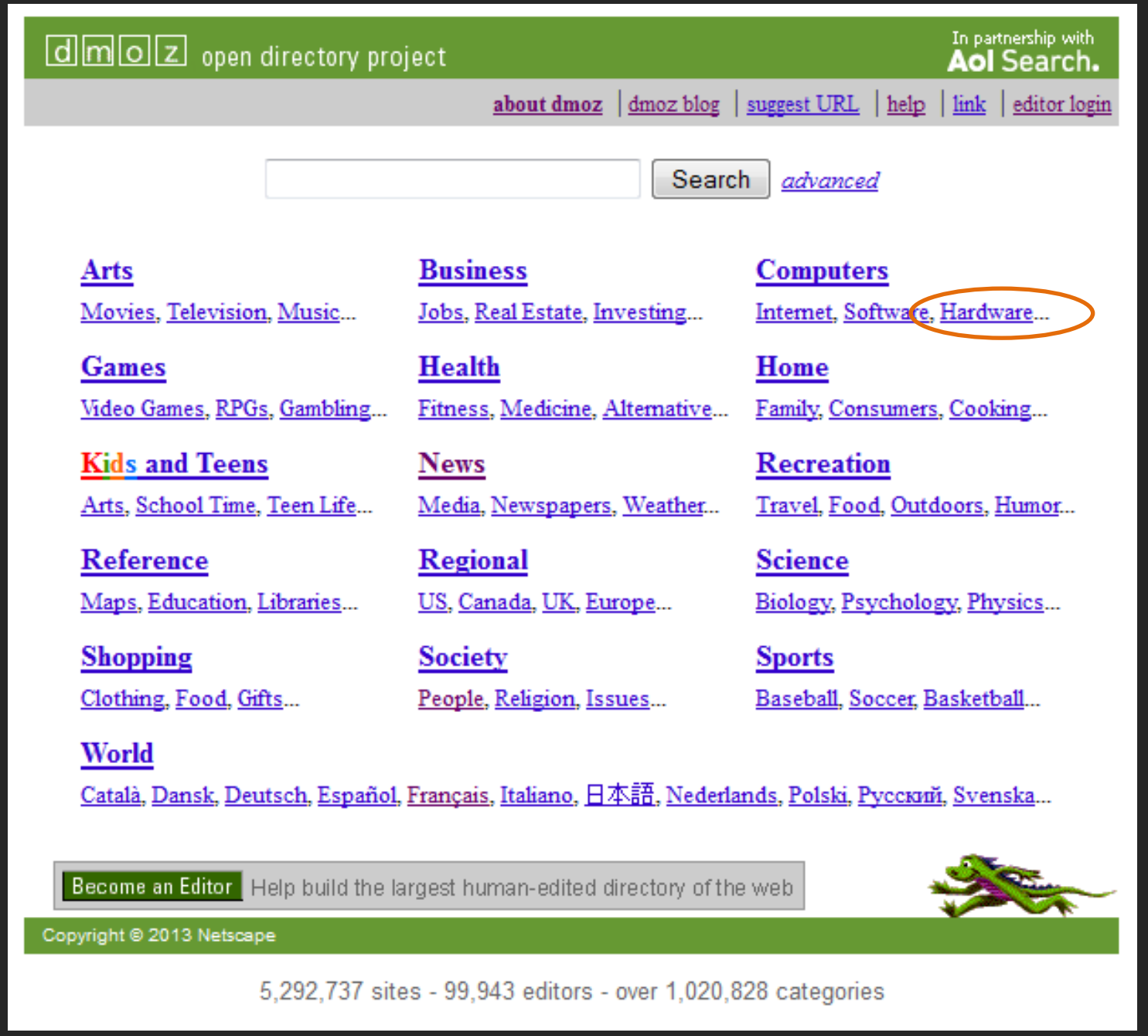
Where would you go to find “Hard drives”?

http://www.slido.com
event code **#C164**



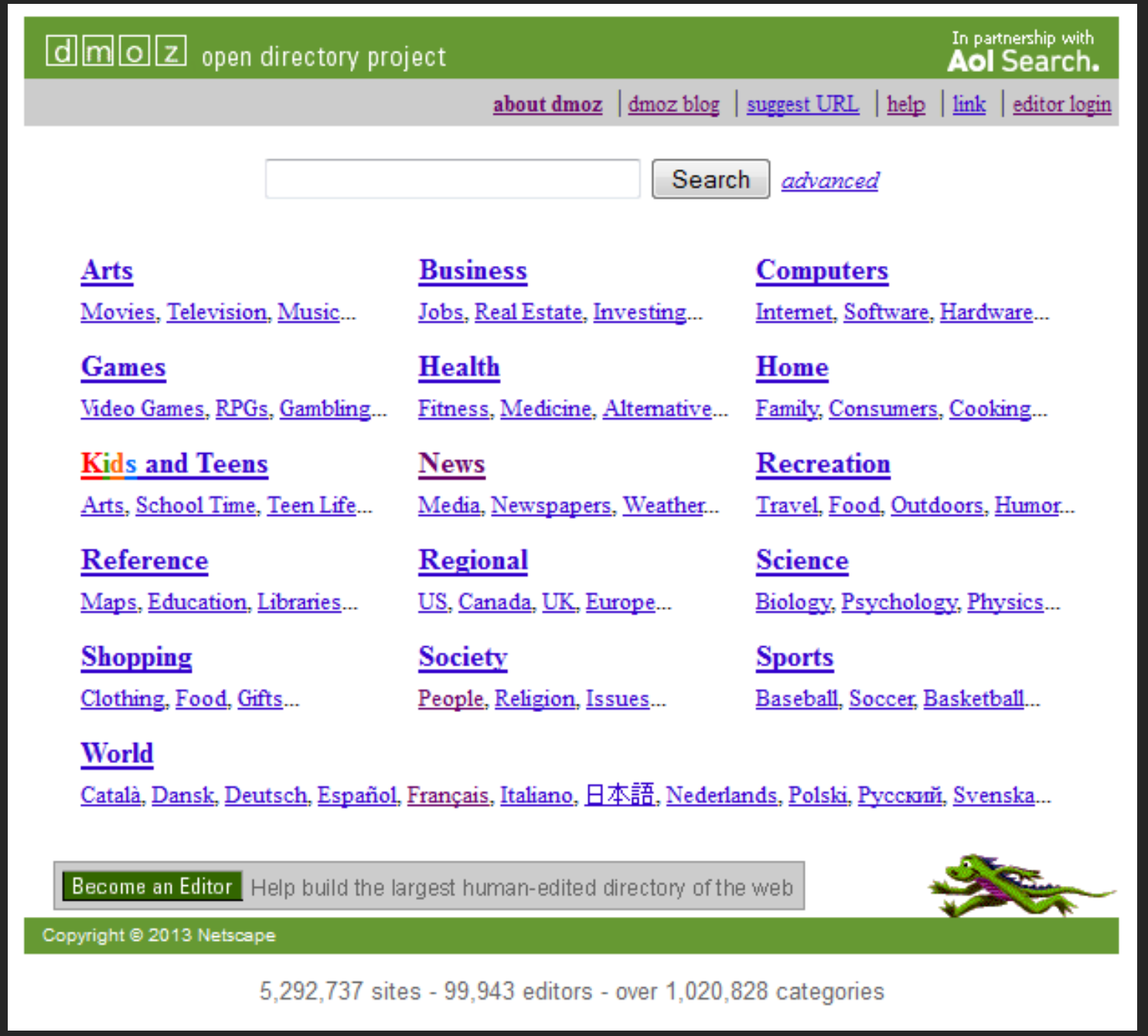
High scent

To find the “hard drive”, you’d click on “hardware”



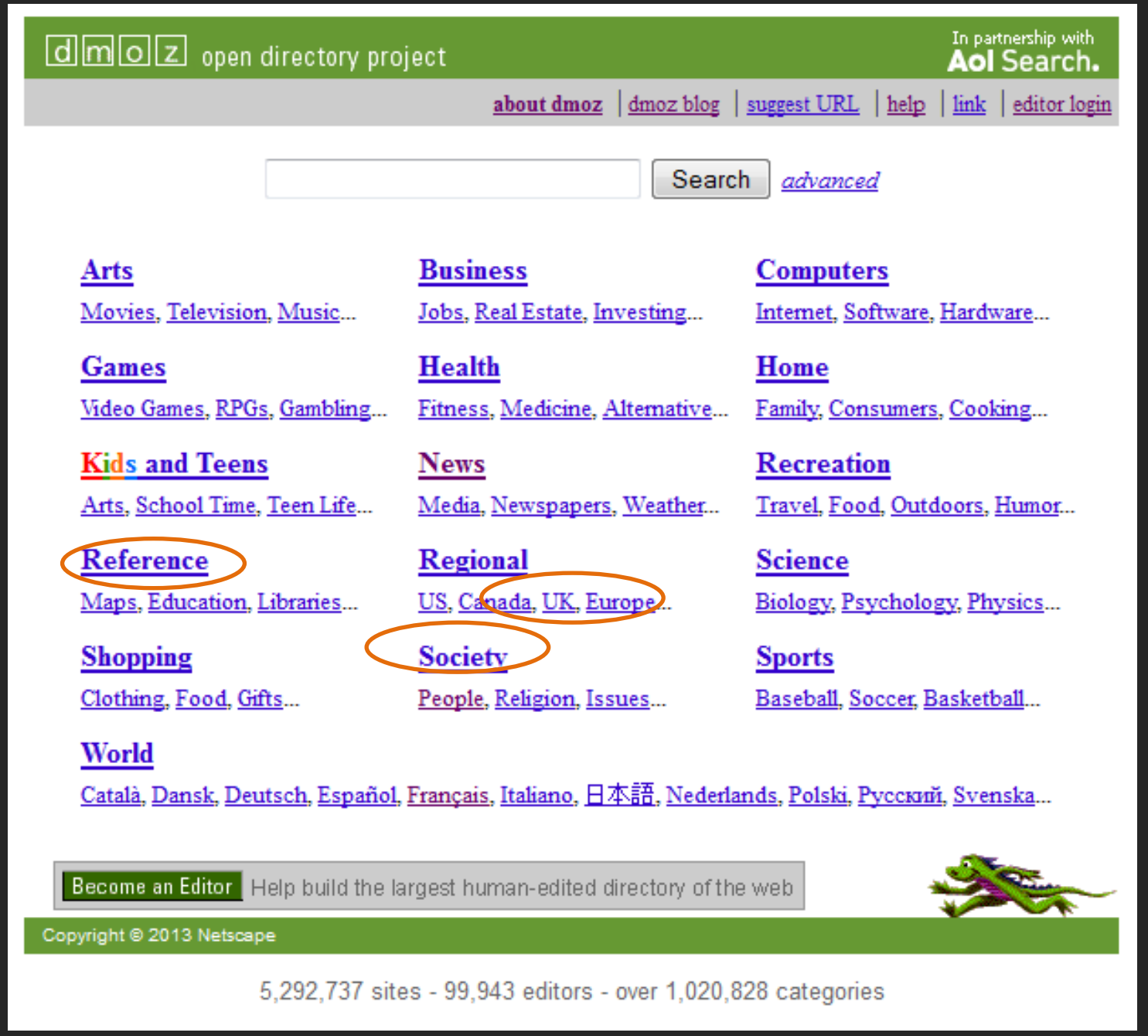
Where would you go to find “the highest rank in the British Royal Air Force”?

http://www.slido.com
event code **#C164**



Low scent

Where do you go to find the highest rank in the British Royal Air Force?

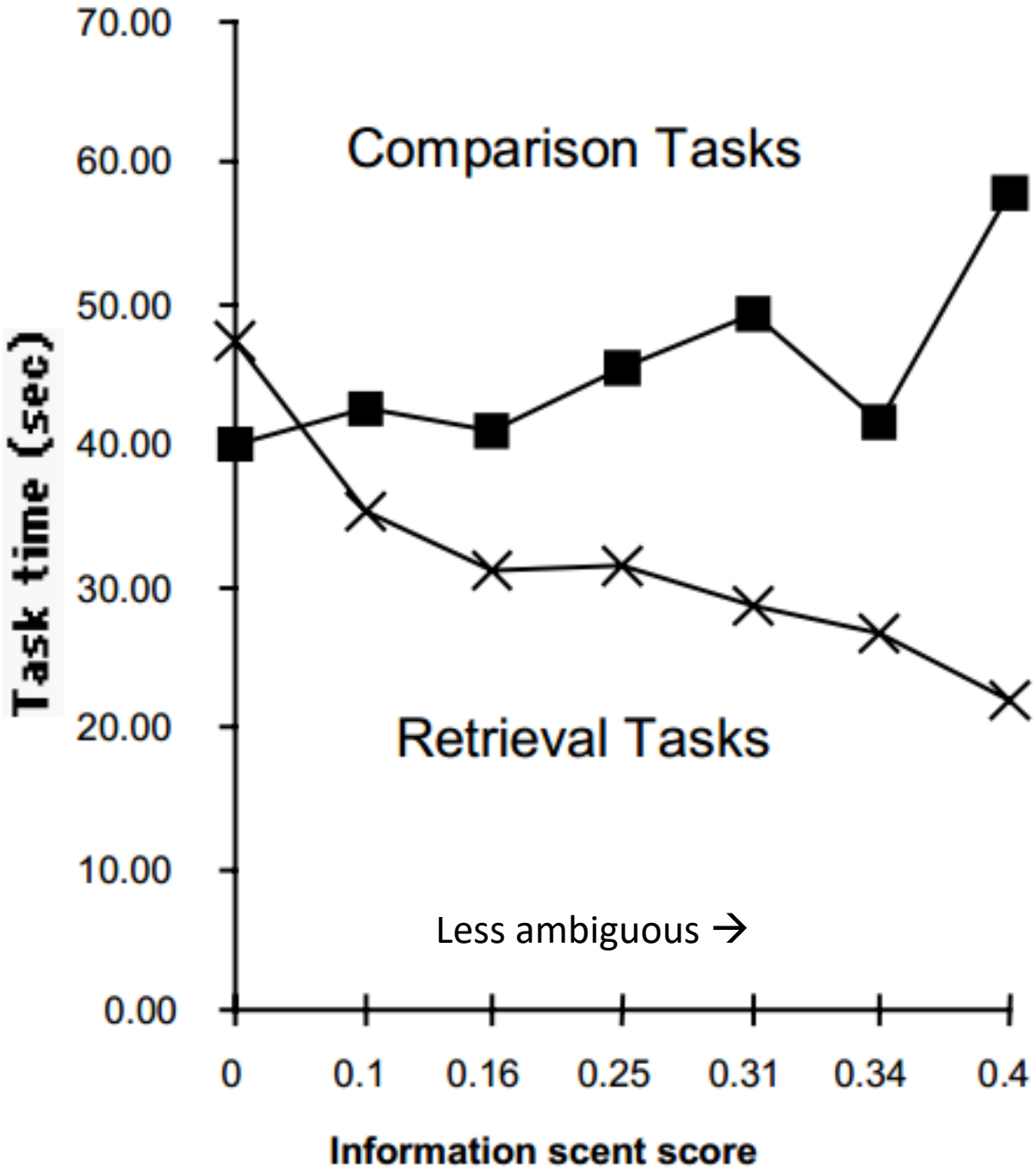


What gives?

- How useful is the text label on a node?
- *Information scent* = the proportion of participants who correctly identified the location of the task answer from looking at the upper branches in the tree

Ambiguity & Information Scent

- Task type 2
  - High scent
    - To find the “hard drive”, you’d click on “technology”
  - Low scent
    - What is the top level category to answer, ”What’s the highest rank in the British Royal Air Force?”



Comparison

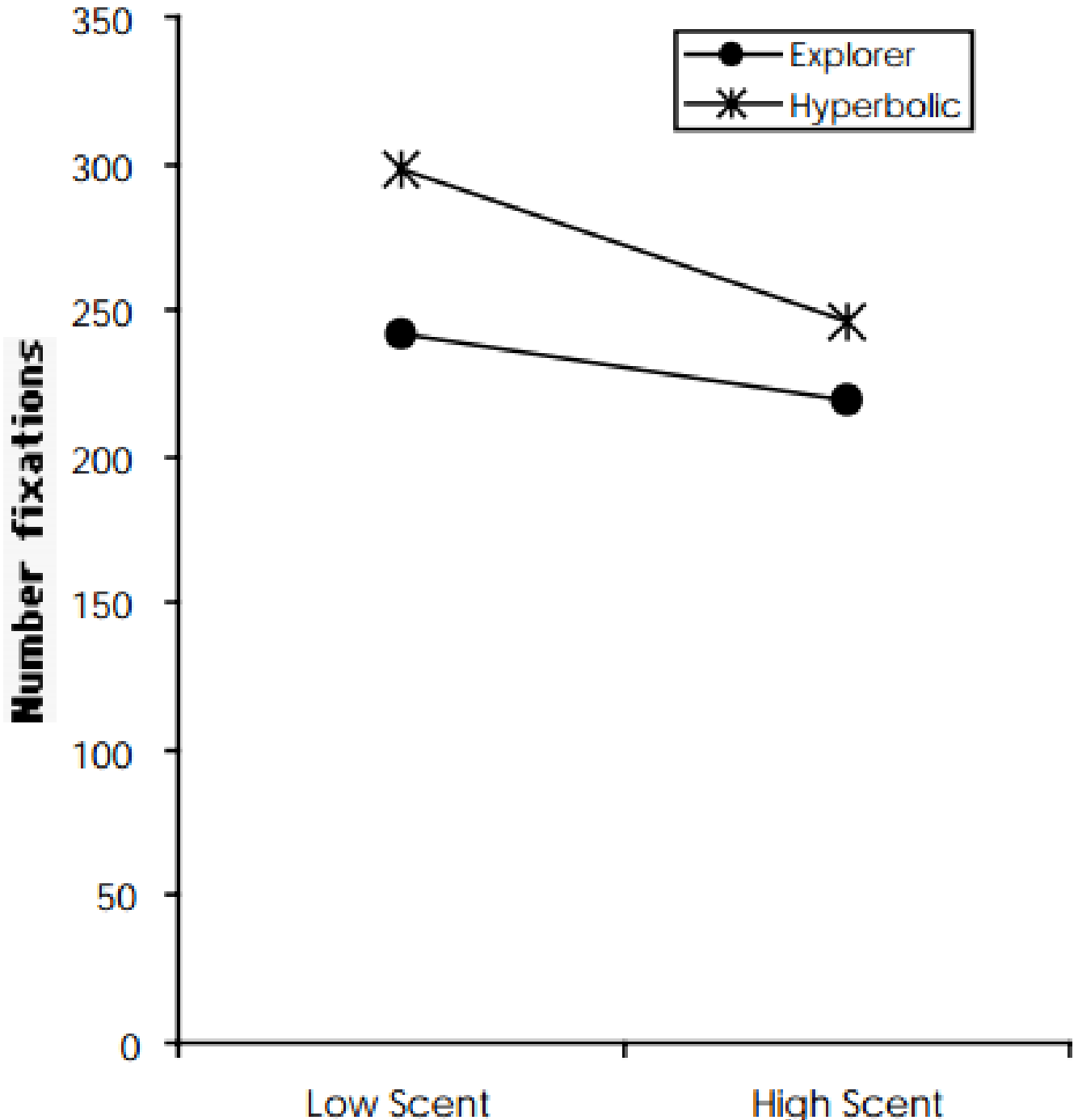
Local relational – “which religion has the most holidays?”

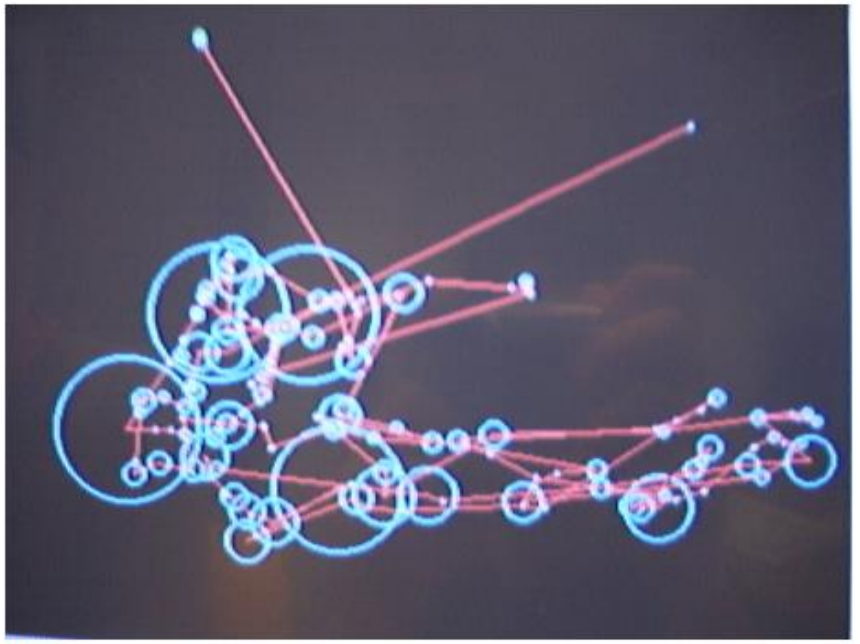
Complex relational – “Which Greek deity has the same name as a space mission?”

Retrieval

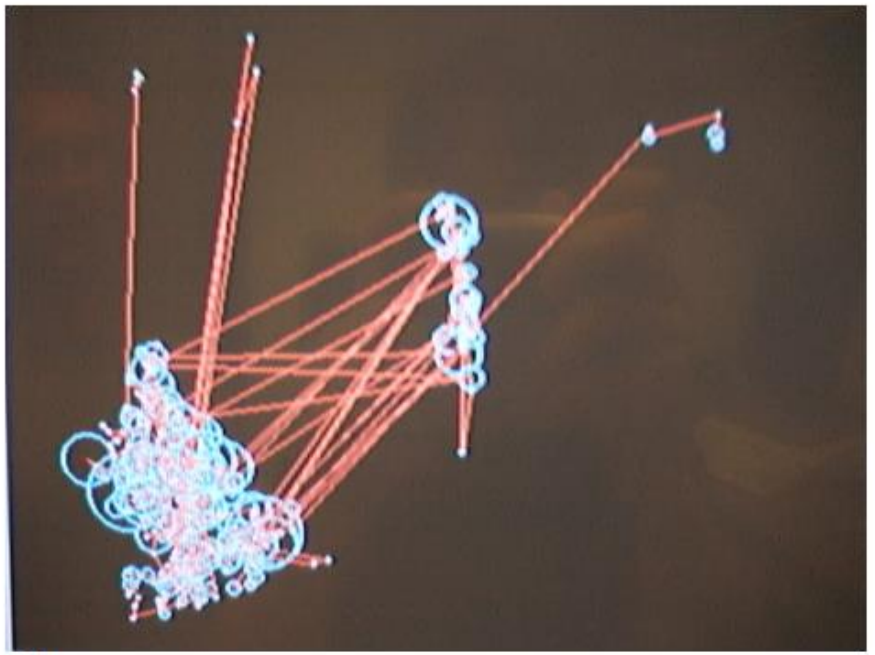
Simple – “find lake Victoria”

Complex – “which army is led by a generalissimo?”


Explorer
Hyperbolic
Number fixations
350
300
250
200
150
100
50
0
Low Scent
High Scent



(a)

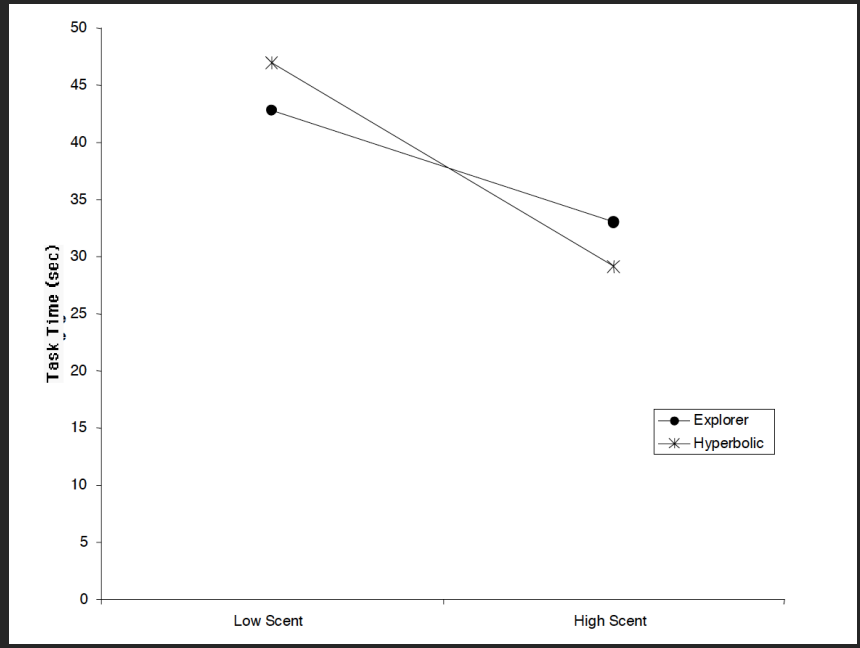


(b)

"Over all the question types, there *was no difference in the time required by the Hyperbolic and Explorer browser*…The two browsers differ however in the way they can take advantage of information scent."

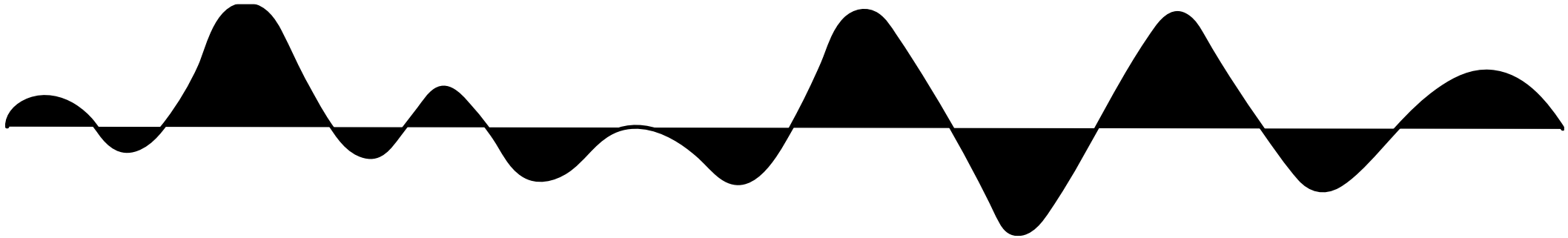
–Pirolli et al.

“When we examine users’ eye movements, *under high-scent conditions and for simple retrieval tasks, the Hyperbolic browser can traverse levels almost twice as fast* as the Explorer can. But it is *slower than the Explorer under low-scent conditions*. Additionally, participants using the Hyperbolic browser use more fixations to do the task, but their fixations are shorter.” –Pirolli et al.

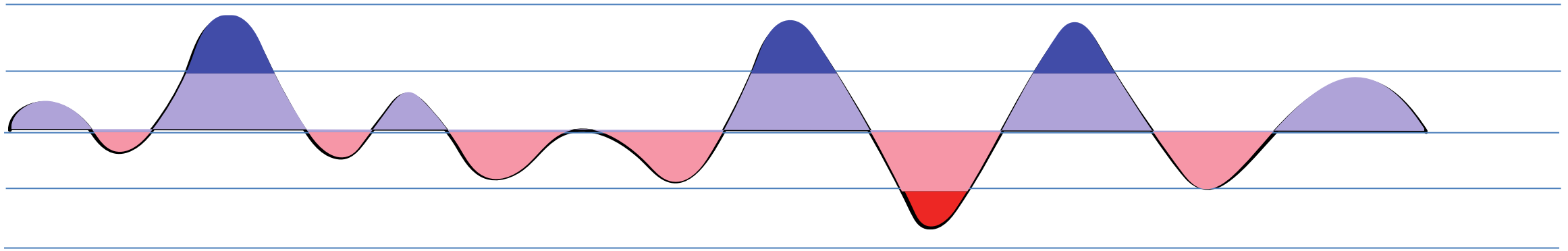


Another example…

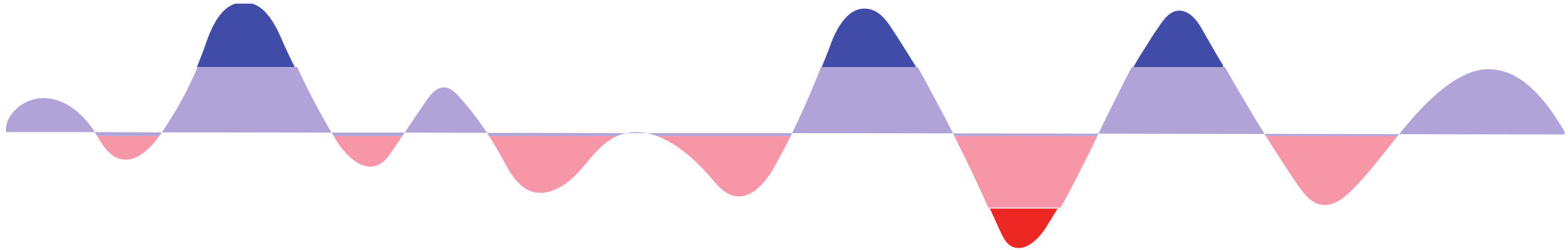
Horizon Charts



Horizon Charts

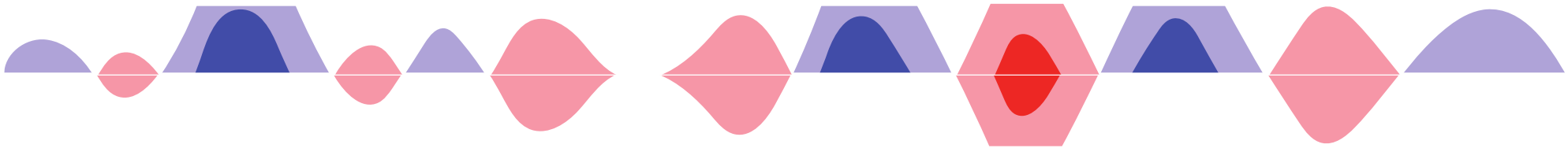


Horizon Charts

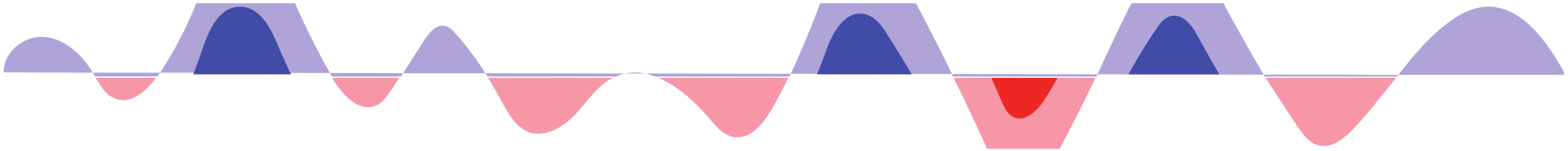


Horizon Charts

Mirrored

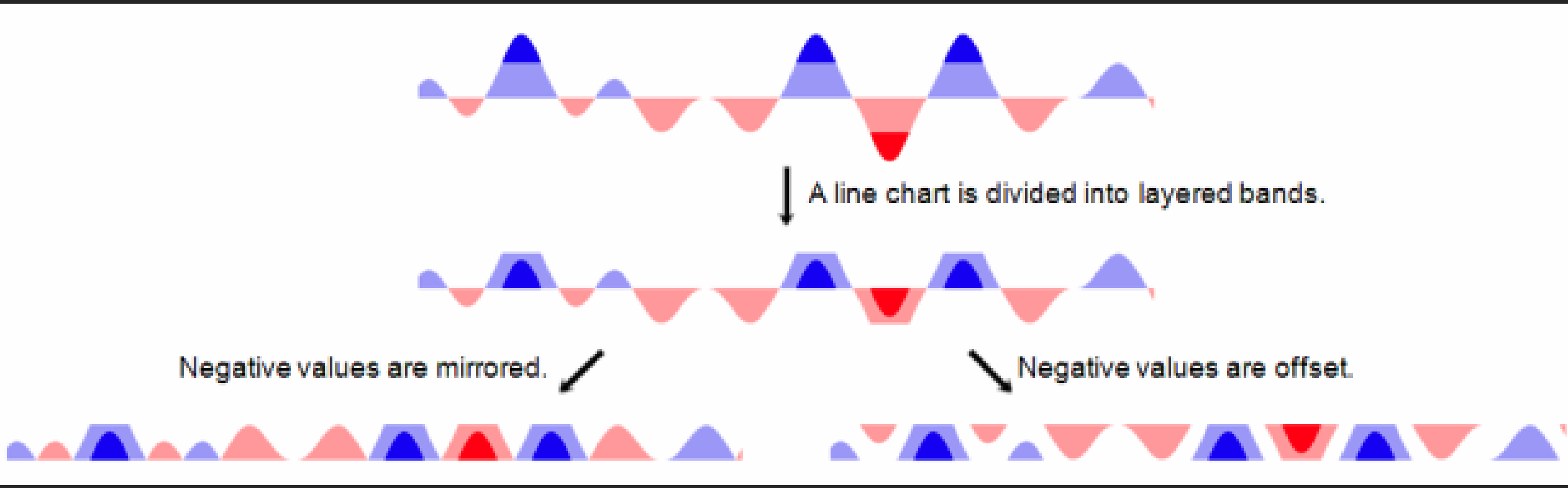


Offset



Horizon Charts

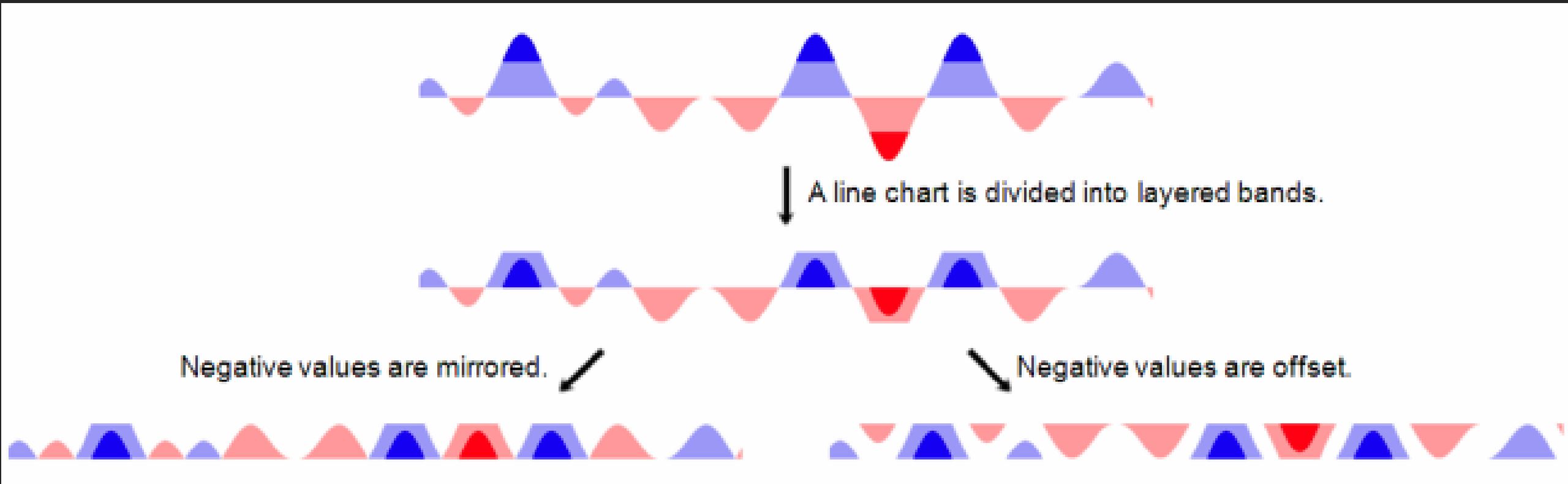
- Heer et al., CHI'09



Group question: What questions would you want to answer experimentally to understand how to use horizon charts?

Can you briefly outline a design to do so?

http://www.slido.com
event code **#C164**



Experiment 1

- How many bands? Mirrored or offset?
- Design: within subjects
  - 2 chart types: mirrored offset
  - 3 band counts: 2,3, 4
  - 16 trials per condition
  - 96 trials per subject
  - Given two values, which is higher and estimate the difference

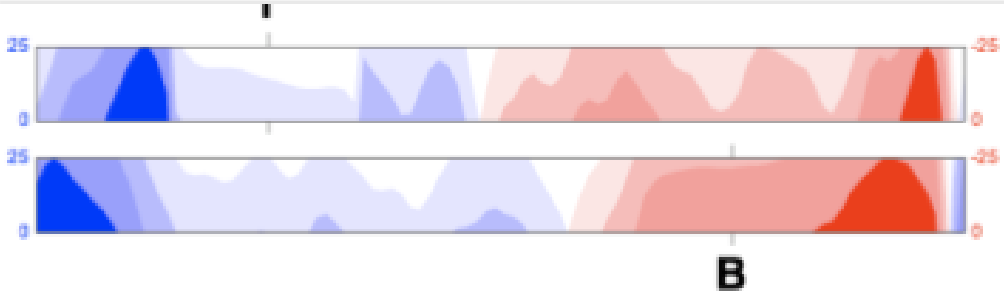


Figure 3. Example trial with a 4-band mirrored graph.
Each band covers 25 values; the total range is [-100, 100].
Subjects reported if T or B was larger, and by how much.

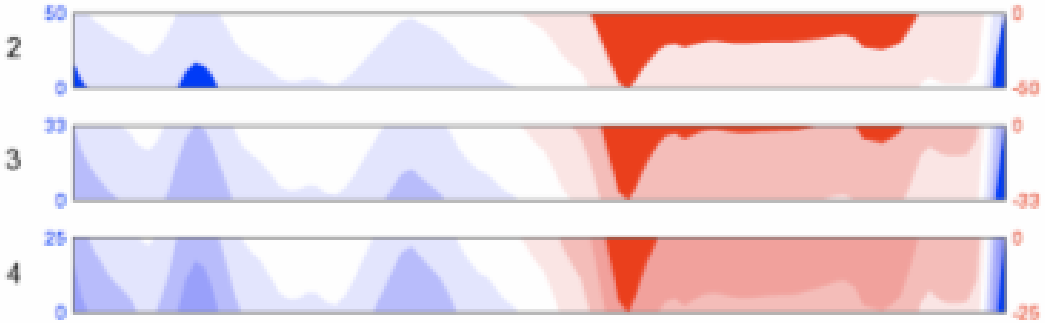
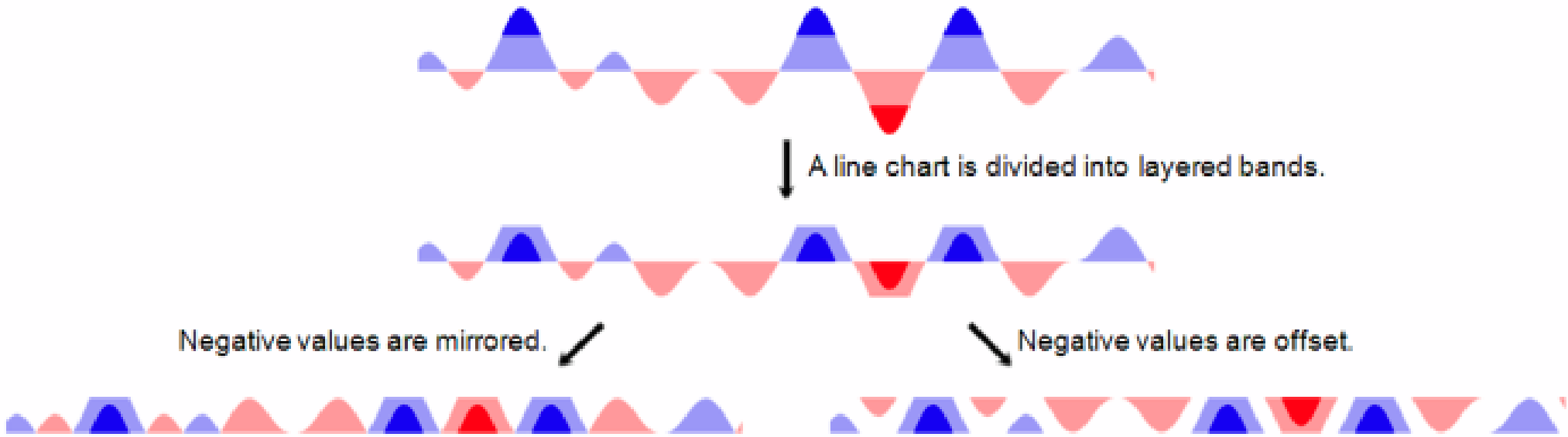


Figure 4. Offset horizon graphs with 2, 3, and 4 bands.

Experiment 1

http://www.slido.com
event code **#C164**

- Mirrored or offset?
- How many bands?



Experiment 2

- Line chart versus mirror/layer? Effect of size?

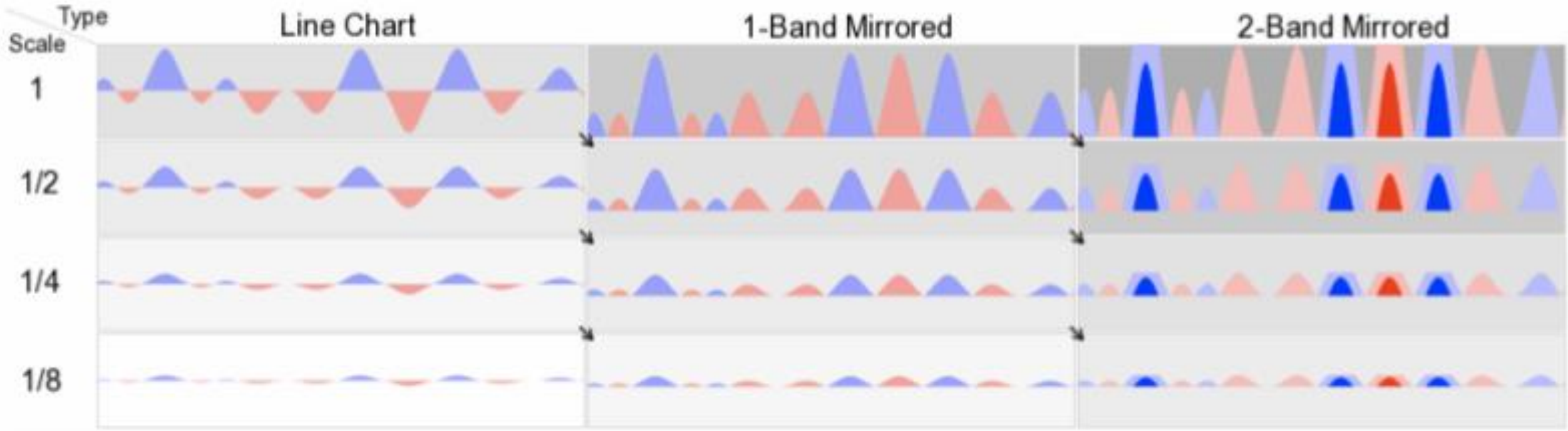


Figure 7. Chart Type and Scale Conditions in Experiment 2. We crossed 3 chart types and 4 chart heights. The diagonally adjacent cells indicated by arrows and shading have the same *virtual resolution*: the un-mirrored, un-layered size of the chart.

Experiment 2

- Design: 3 charts, 4 sizes, 10 trials per condition, 120 trials per subject

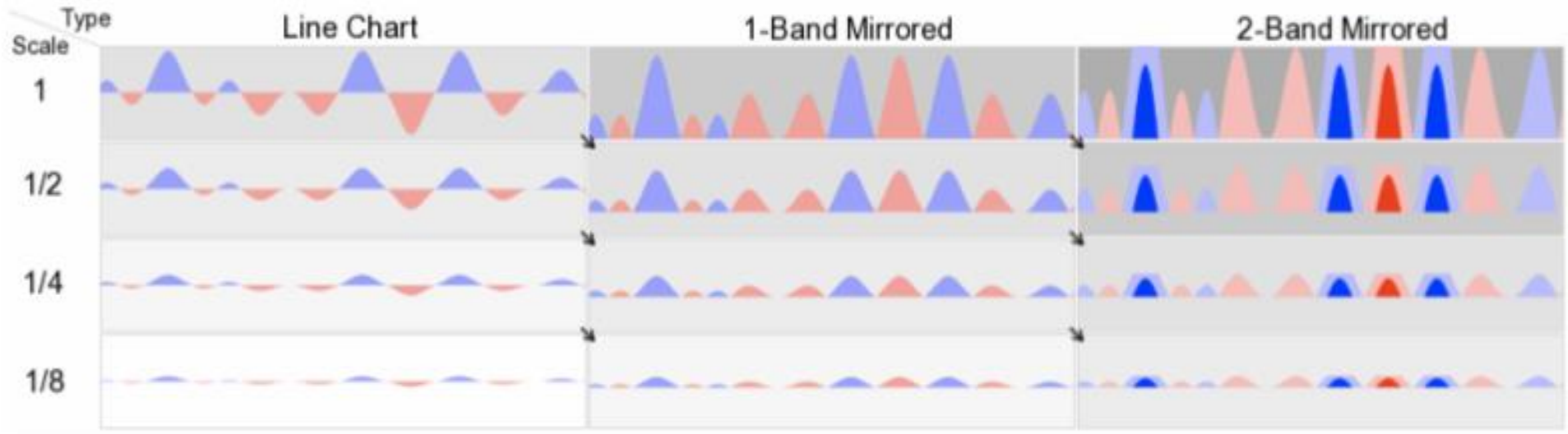


Figure 7. Chart Type and Scale Conditions in Experiment 2. We crossed 3 chart types and 4 chart heights. The diagonally adjacent cells indicated by arrows and shading have the same *virtual resolution*: the un-mirrored, un-layered size of the chart.

Results

- Found crossover point where 2-band better: 24 pixels
  - Virtual resolution: unmirrored, unlayered height
  - Line: 1x, 1 band: 2x, 2 band: 4x
- Guidelines
  - Mirroring safe
  - Up to 4 layers good (higher time, but more accurate)
  - 24 pixels good for line charts, 1 band mirrored
  - 12 or 6 pixels good for 2 band

Summary

- Evaluation key for “useful” and “usable” visualization systems
  - InfoVis is a little odd in this regards
- Many different evaluation techniques
  - Each with good and bad properties
  - Each requiring care in use
- Use tools to identify what you are evaluating, the design the right experiment for it