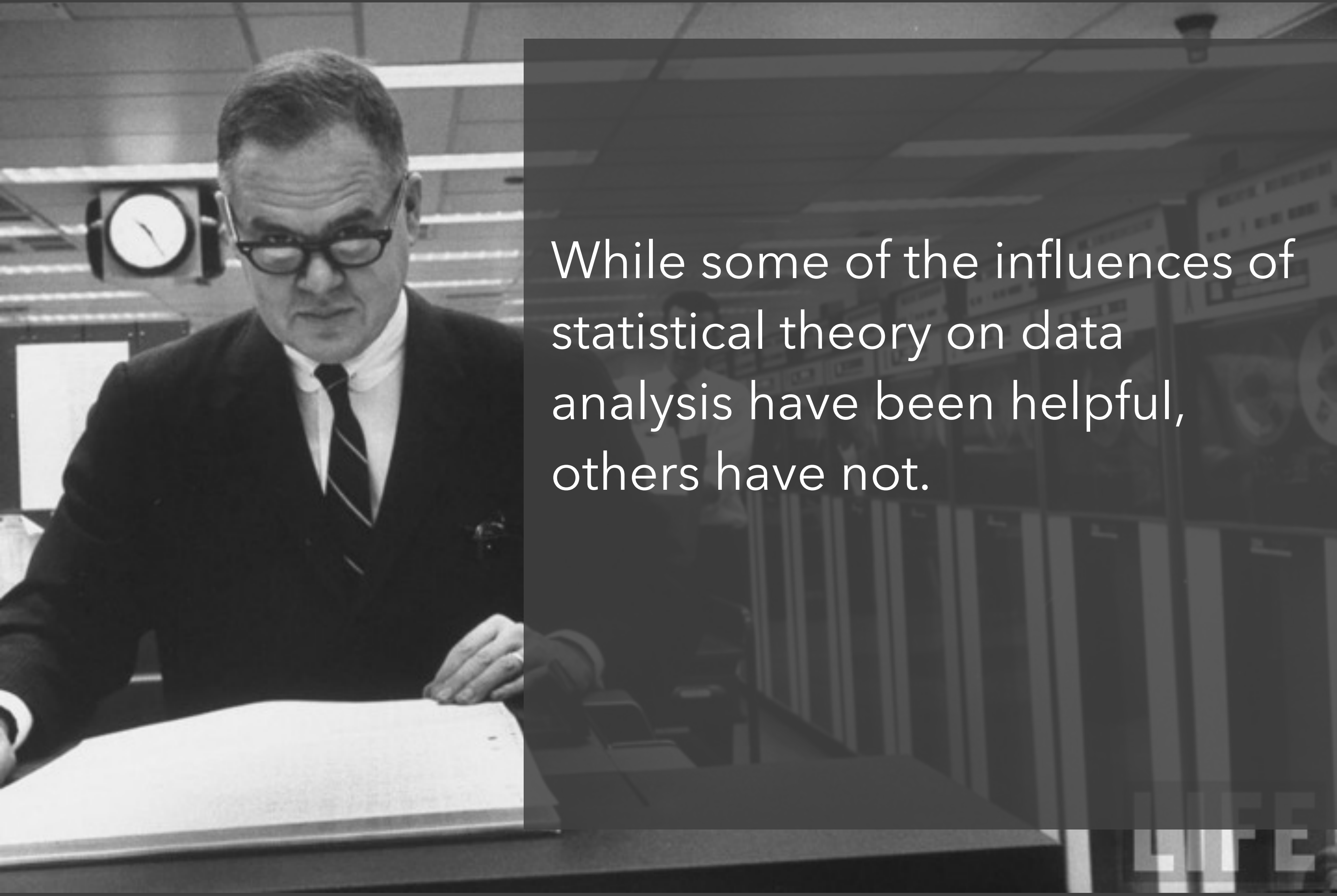
Principles of Data Visualization

Jeffrey Heer @jeffrey_heer
University of Washington

Data Analysis & Statistics, Tukey & Wilk 1965

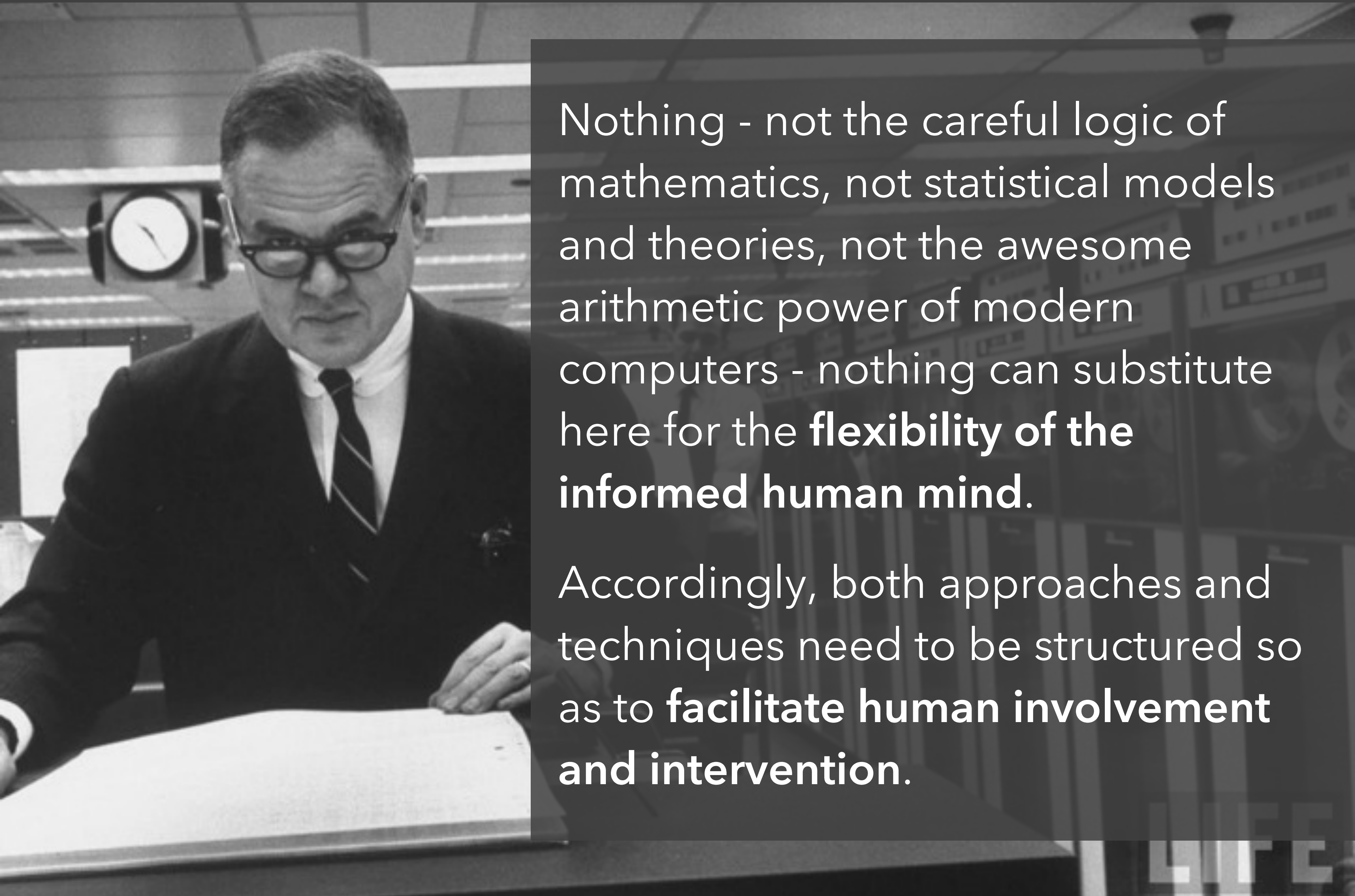
Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.



While some of the influences of statistical theory on data analysis have been helpful, others have not.

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.



Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind**.

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

Some implications for effective data analysis are: (1) that it is essential to have convenience of **interaction of people and intermediate results** and (2) that at all stages of data analysis, the nature and detail of output, both actual and potential, need to be **matched to the capabilities of the people who use it and want it.**

Our Focus: Visual Encoding

task
questions, goals
assumptions

data
physical data type
abstract data type

domain
metadata
semantics
conventions

processing
algorithms

mapping
visual encoding

image
visual channel
graphical marks

Visual Language is a Sign System

Jacques Bertin

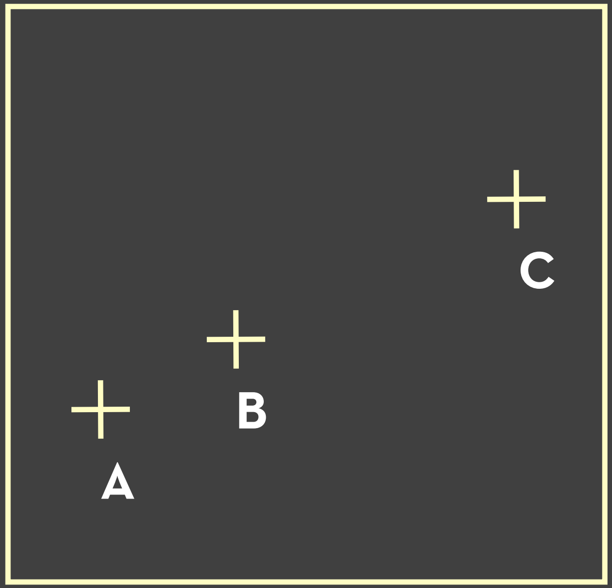
Images perceived as a set of signs

Sender encodes information in signs

Receiver decodes information from signs

Sémiologie Graphique, 1967

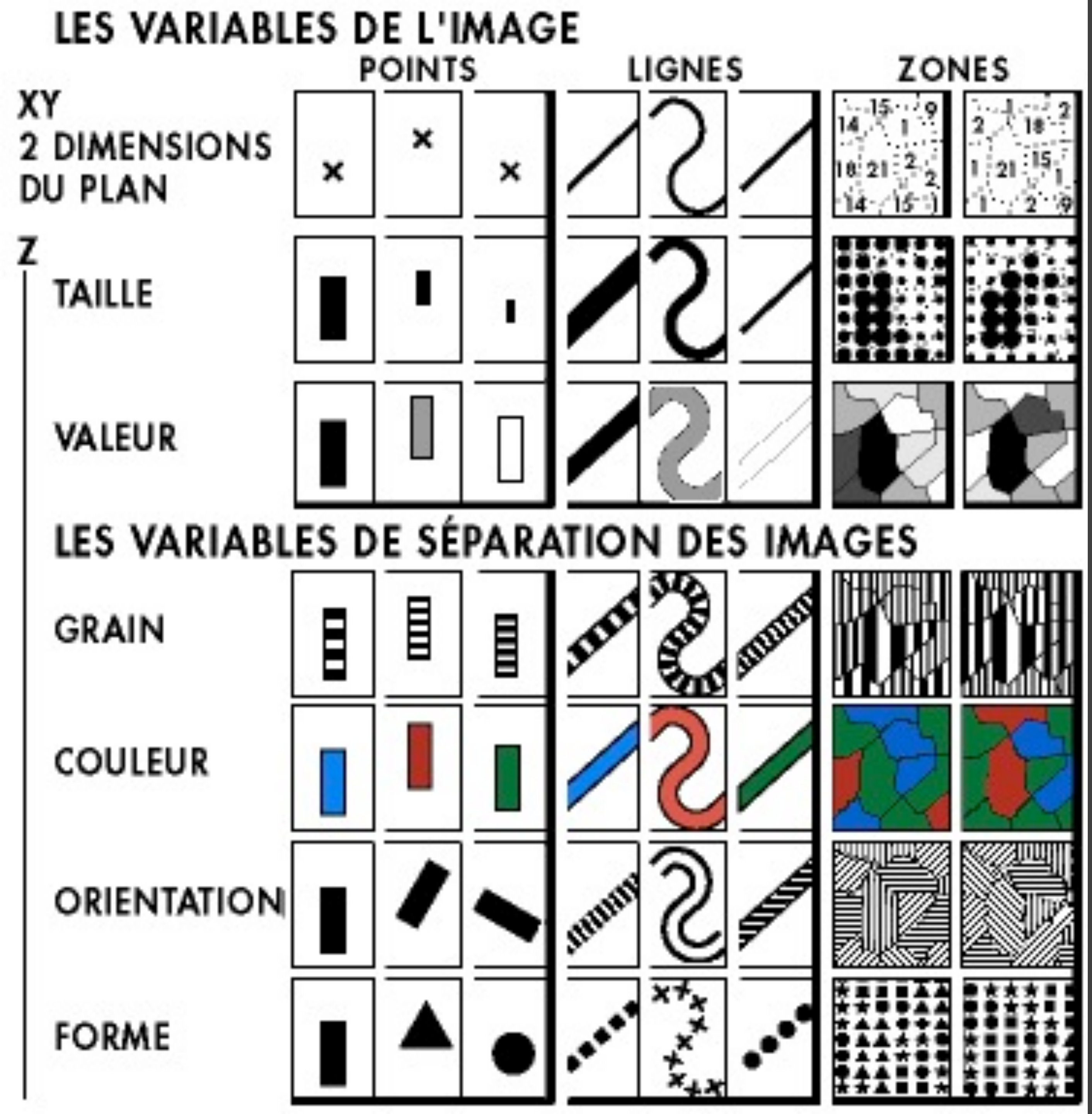
Bertin's Semiology of Graphics



1. A, B, C are distinguishable
2. B is between A and C.
3. BC is twice as long as AB.

∴ Encode quantitative variables

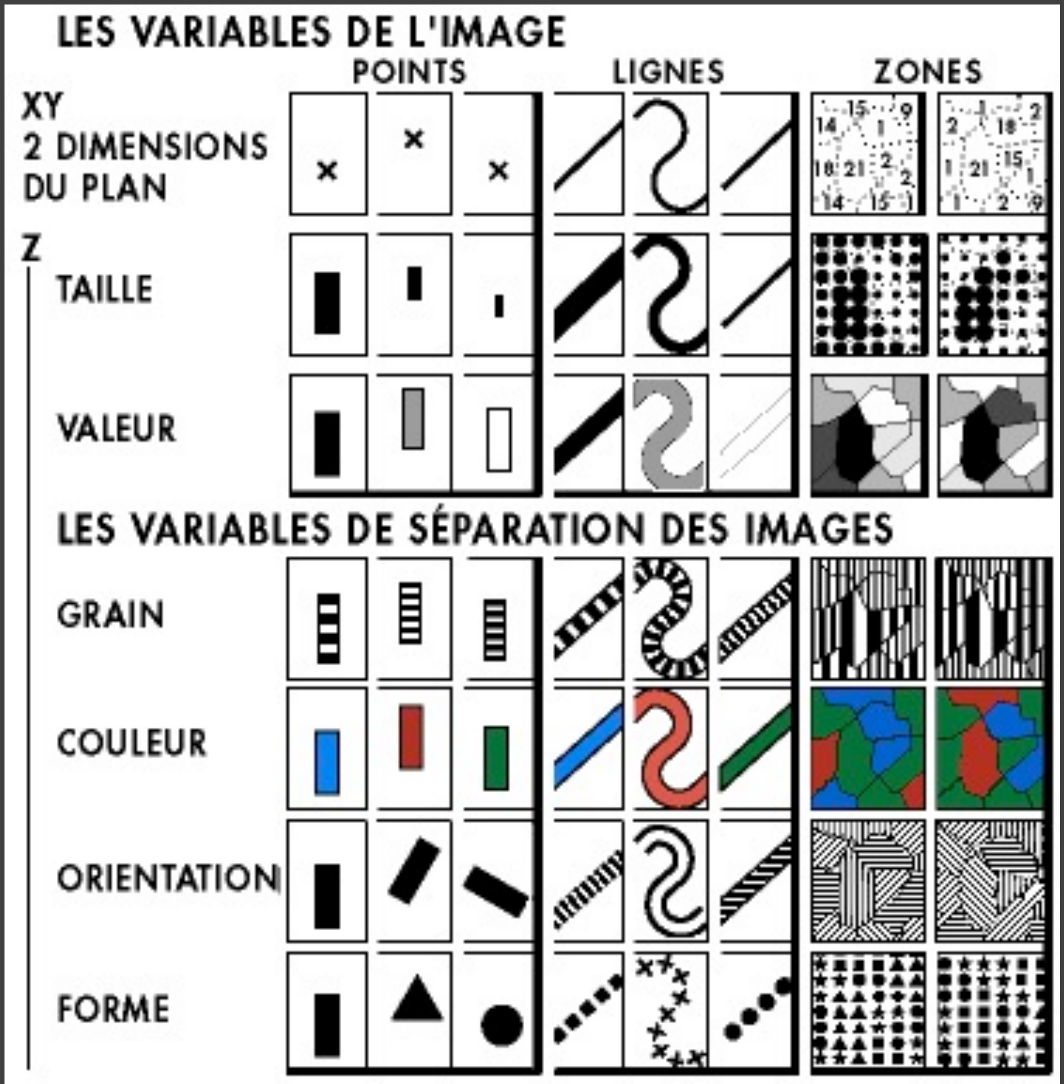
"Resemblance, order and proportion are the three signfields in graphics." - Bertin



LES VARIABLES DE L'IMAGE
POINTS
LIGNES
ZONES
XY
2 DIMENSIONS
DU PLAN
Z
TAILLE
VALEUR
LES VARIABLES DE SÉPARATION DES IMAGES
GRAIN
COULEUR
ORIENTATION
FORME

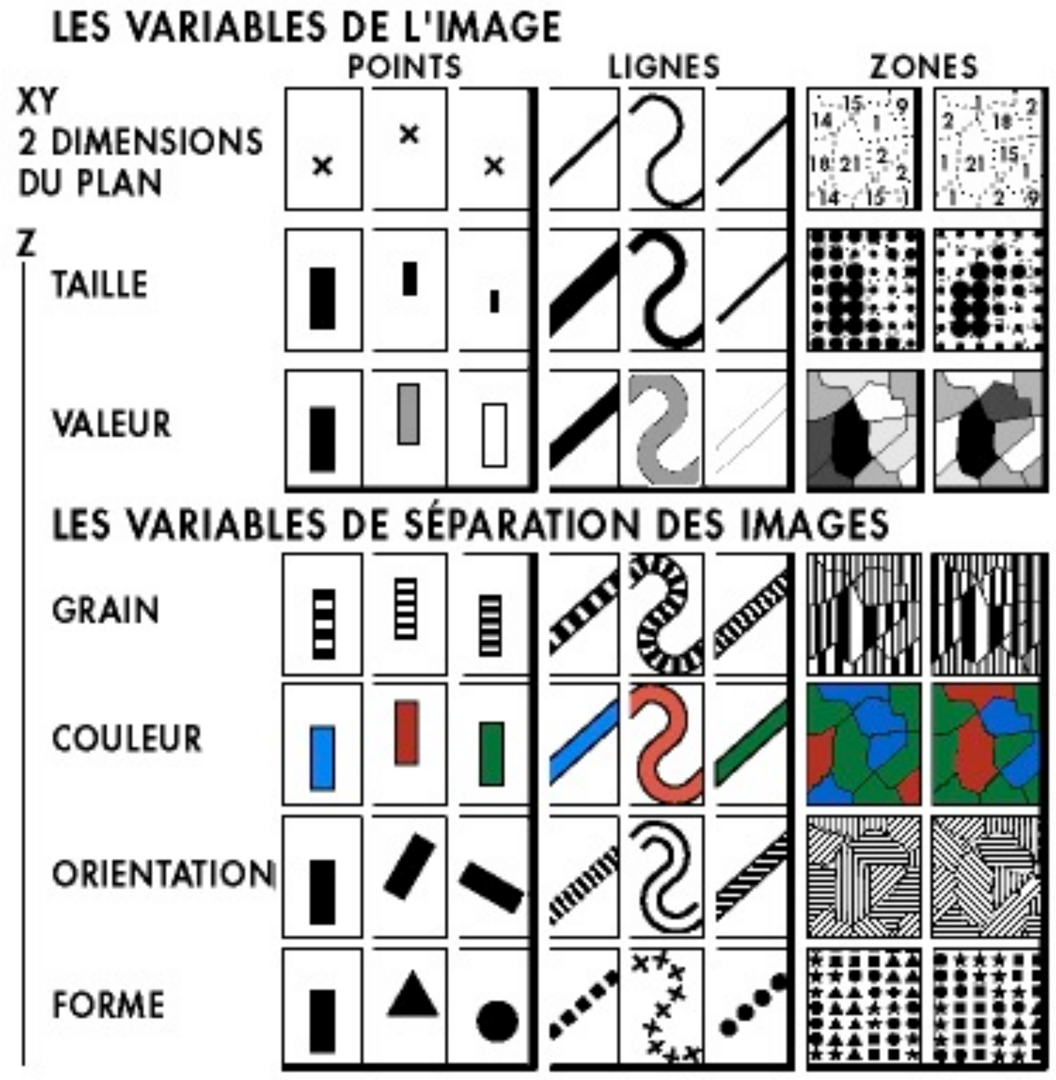
Visual Encoding Variables

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape



Visual Encoding Variables

Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape
Transparency
Blur / Focus …



What makes a visualization **“good”**?

Design Principles [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Principles [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Expresses facts not in the data

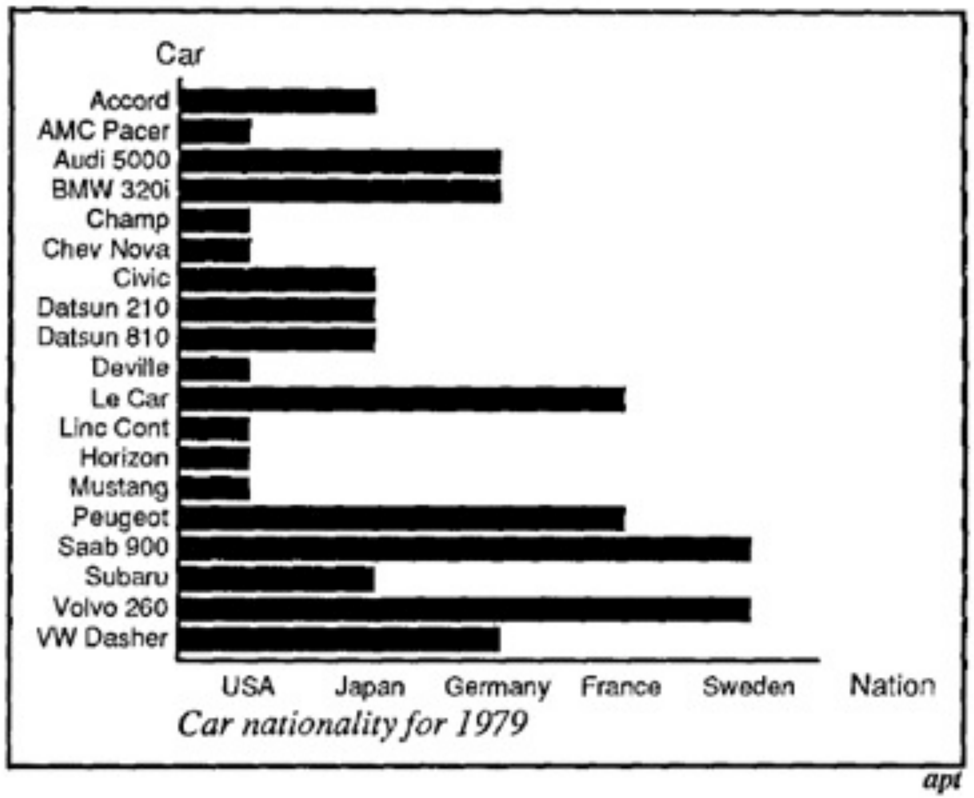


Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

A length is interpreted as a quantitative value.

Design Principles [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Principles [Mackinlay 86]

Expressiveness
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness
A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Principles [Tversky 02]

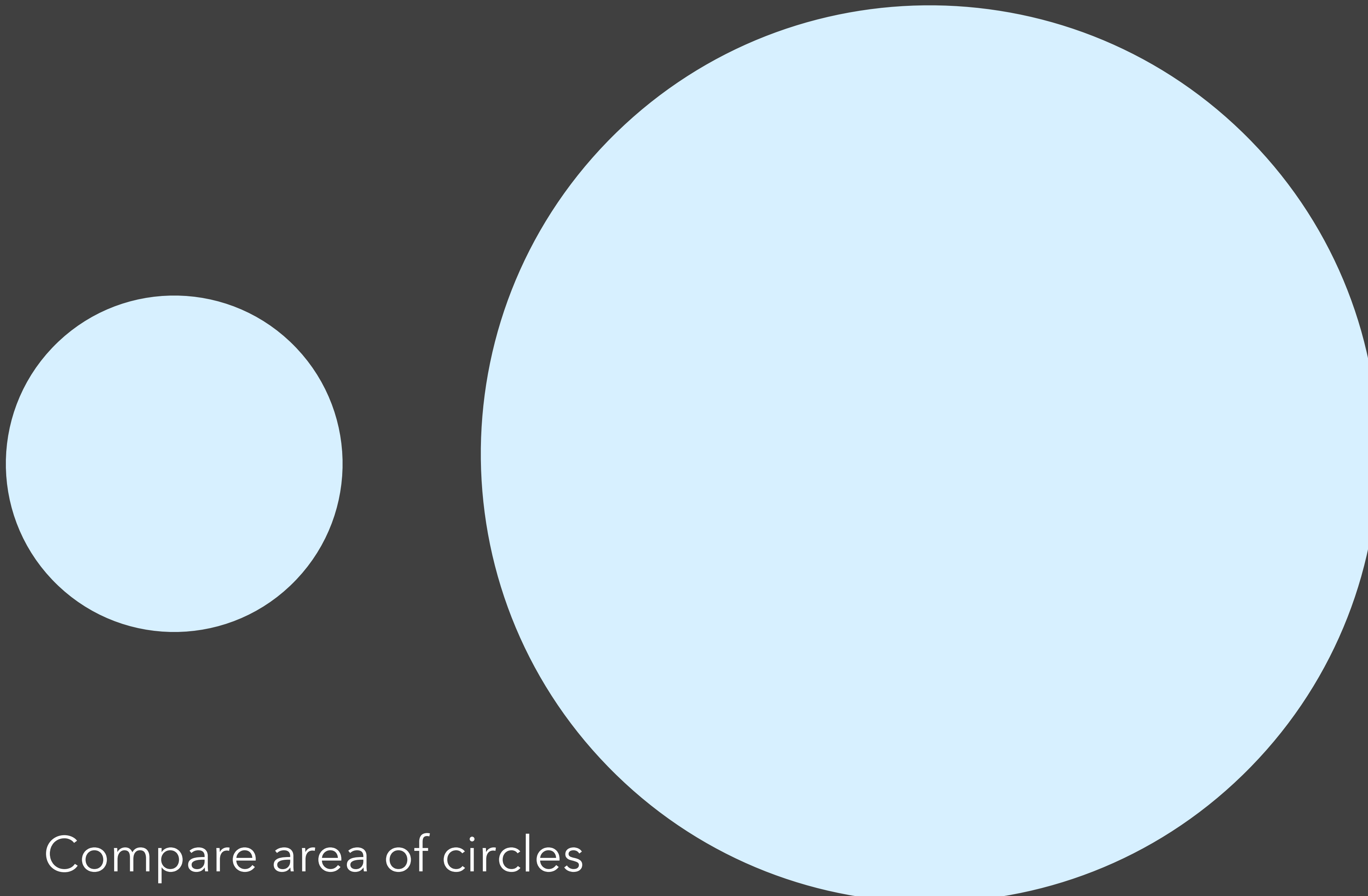
Congruence
The structure and content of the external representation should correspond to the desired structure and content of the internal representation.

Apprehension
The structure and content of the external representation should be readily and accurately perceived and comprehended.

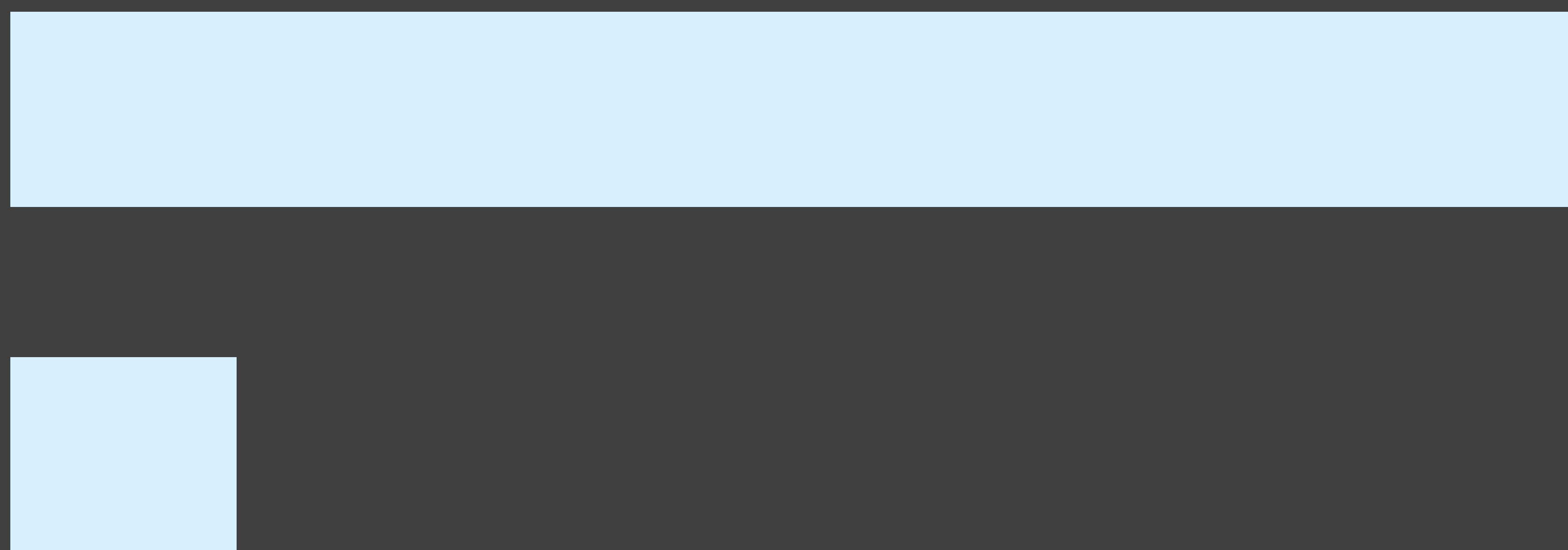
Design Principles *Translated*

Tell the truth and nothing but the truth
(don't lie, and don't lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)



Compare area of circles



Compare length of bars

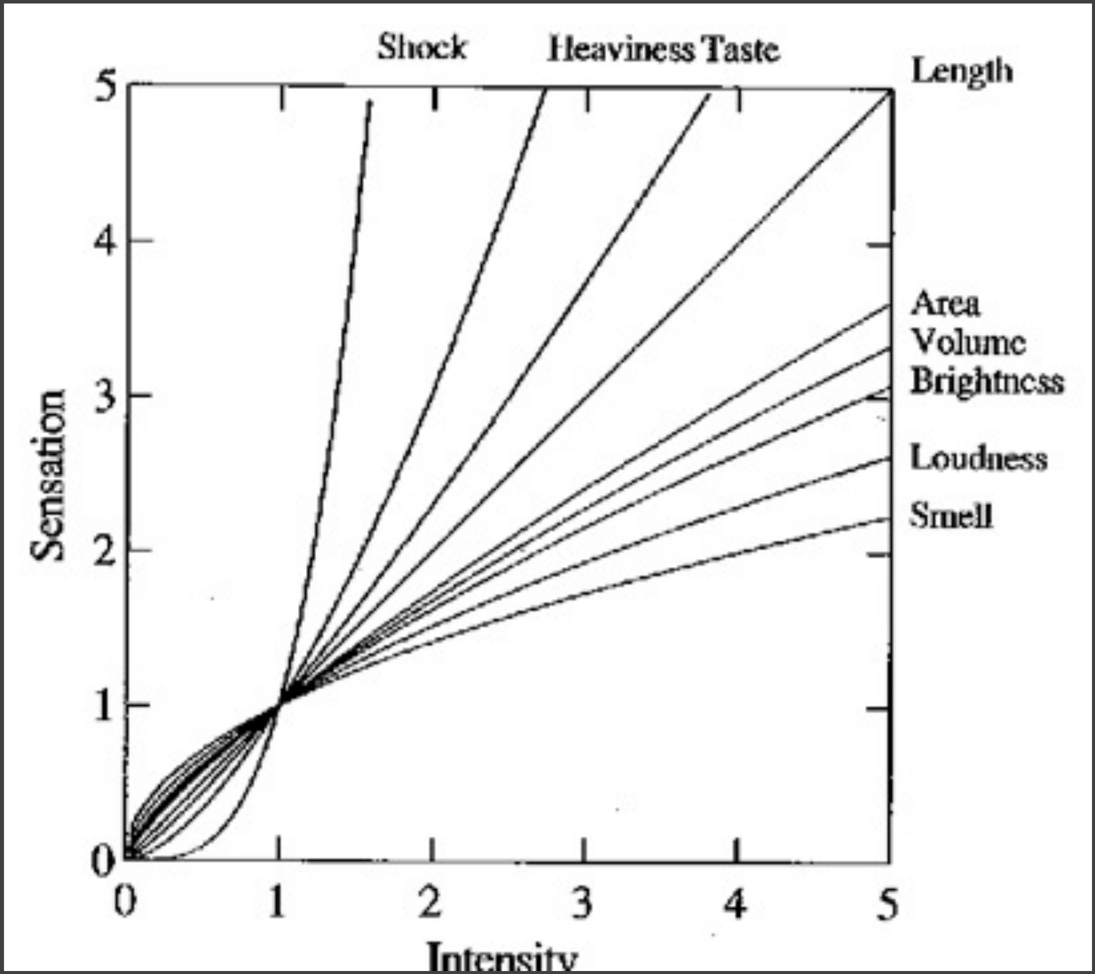
Steven's Power Law

Exponent
(Empirically Determined)

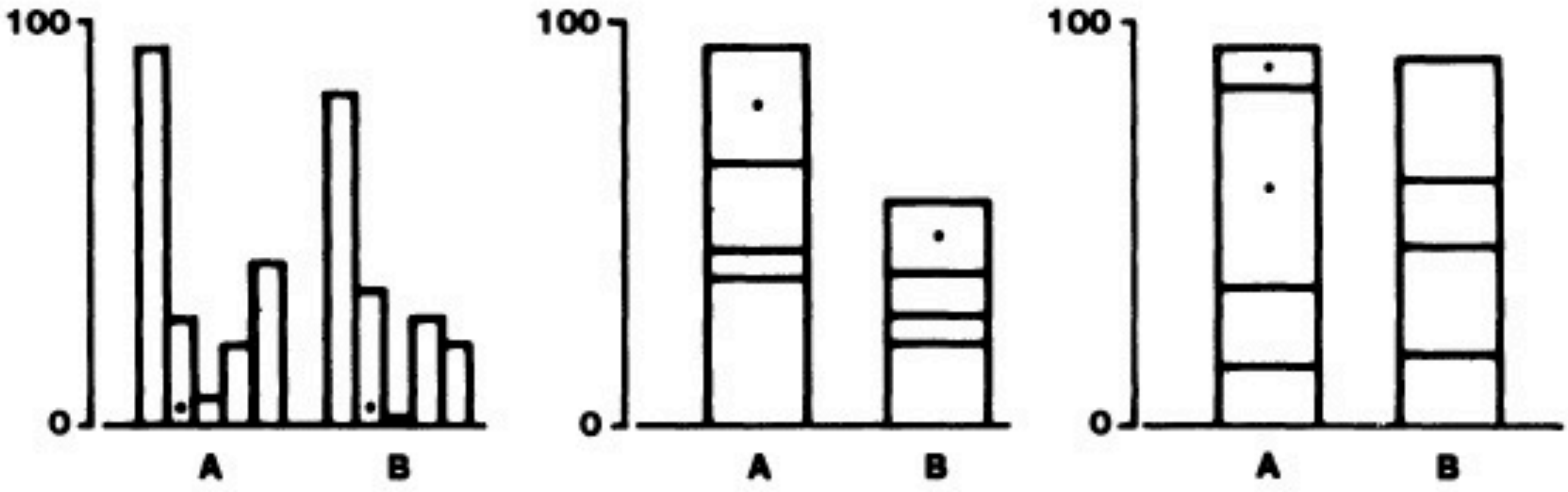
$$S = I^p$$

Perceived Sensation — Physical Intensity

Predicts bias, not necessarily accuracy!



Graph from Wilkinson 99, based on Stevens 61



Graphical Perception [Cleveland & McGill 84]

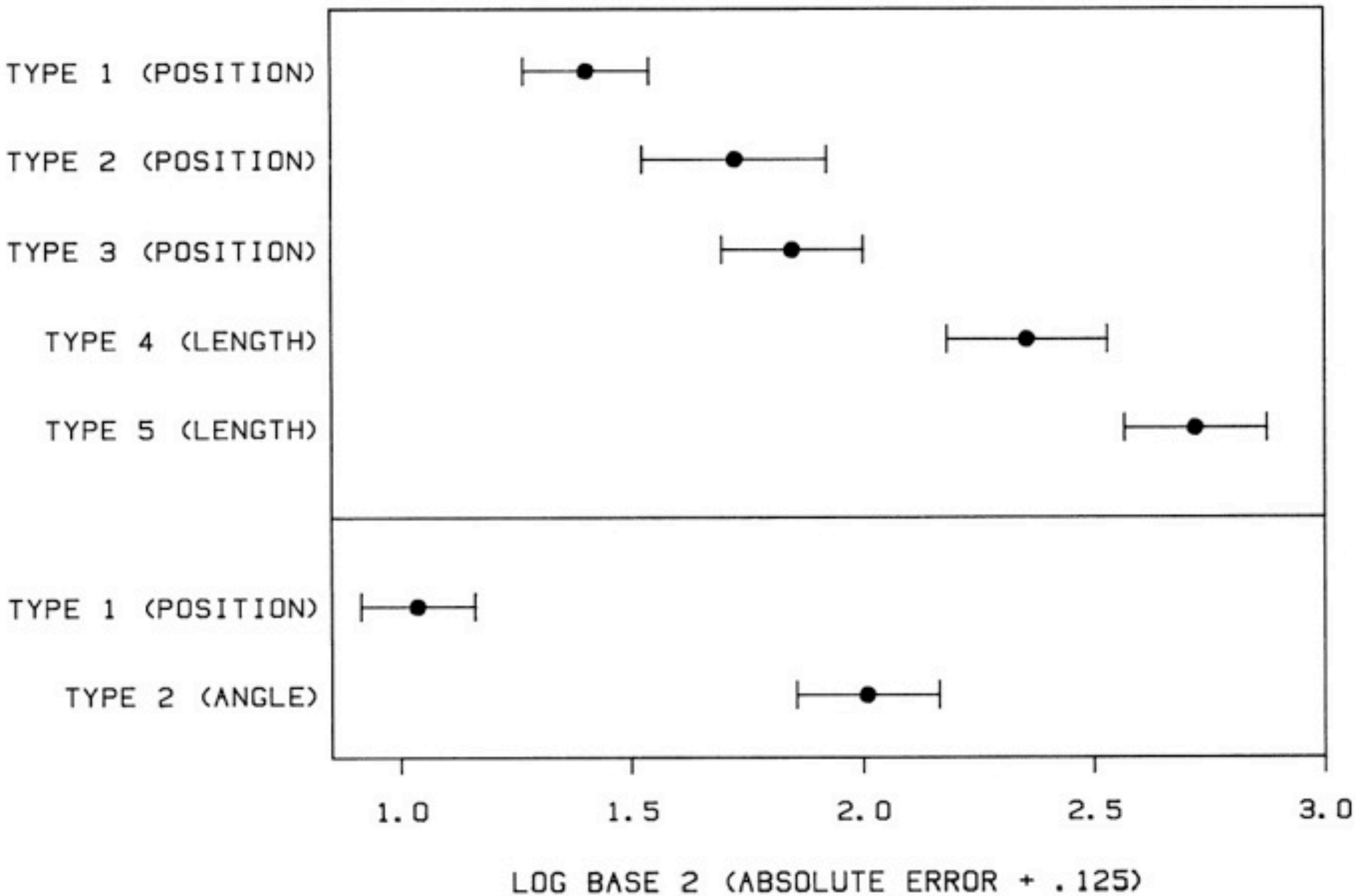
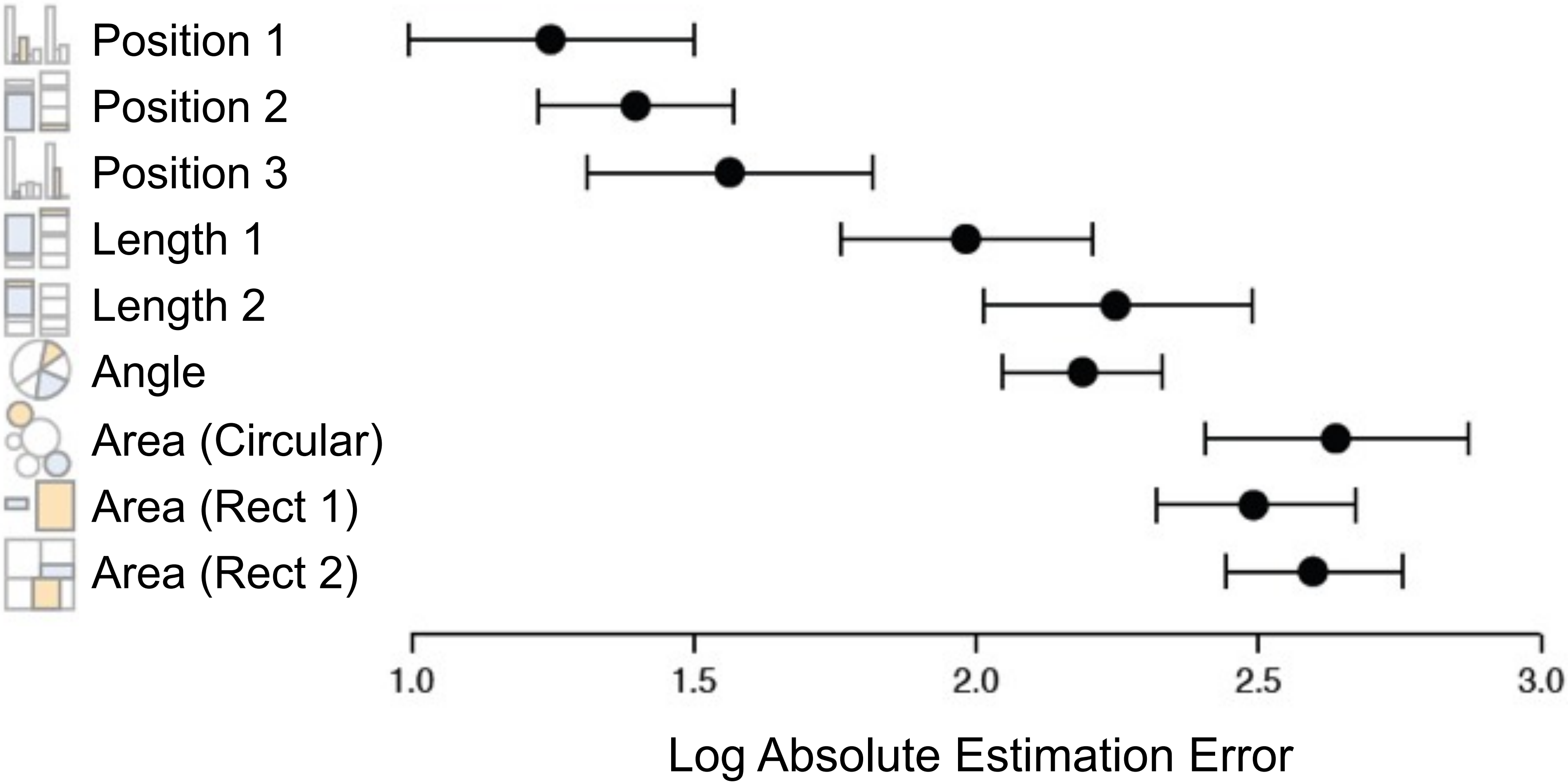


Figure 16. Log absolute error means and 95% confidence intervals for judgment types in position–length experiment (top) and position–angle experiment (bottom).



Graphical Perception Experiments

Empirical estimates of encoding effectiveness

Relative Magnitude Estimation

Most accurate

Position (common) scale

Position (non-aligned) scale

Length

Slope

Angle

Area

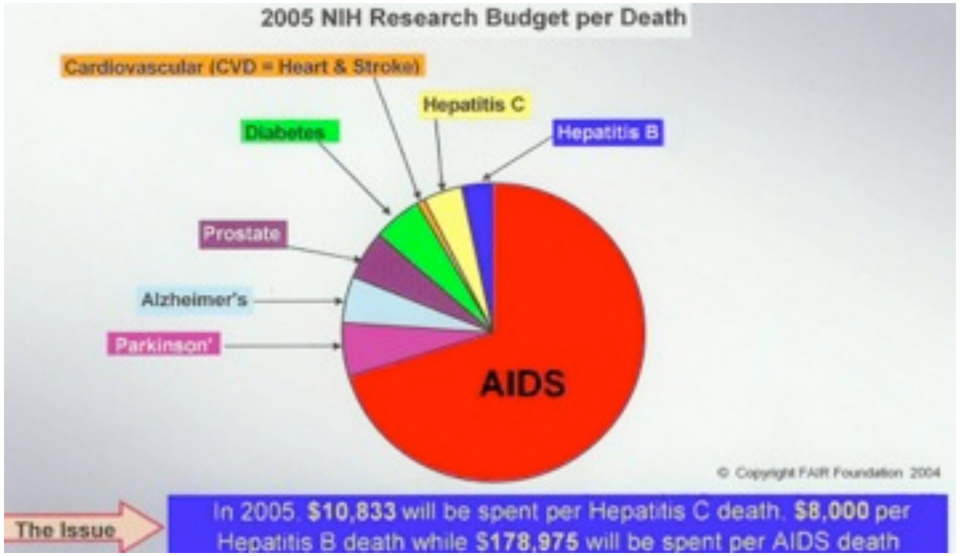
Volume

Color hue-saturation-density

Least accurate

Effectiveness Rankings [Mackinlay 86]

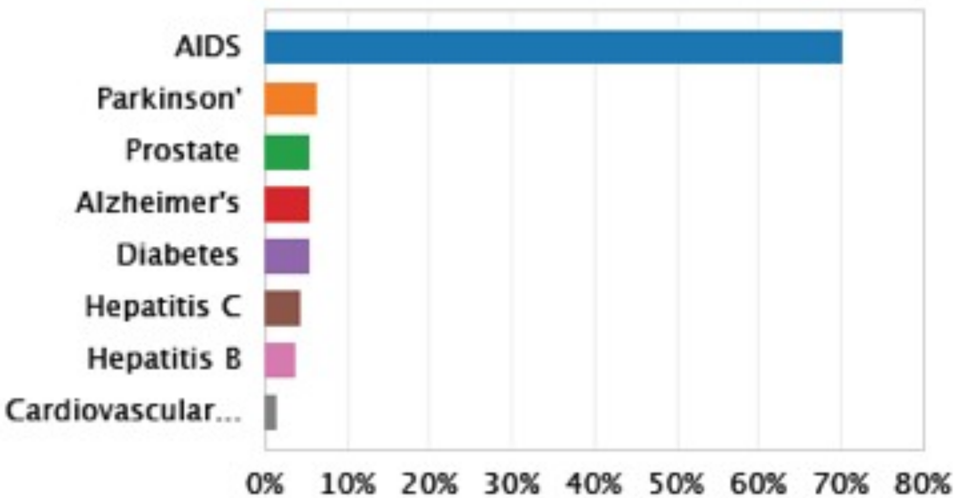
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture	Slope	Angle
Connection	Area (Size)	Slope
Containment	Volume	Area
Shape	Shape	Volume



Type: Pie

Data Table

AIDS	70.0%
Parkinson'	6.0%
Prostate	5.2%
Alzheimer's	5.1%
Diabetes	5.1%
Hepatitis C	4.0%
Hepatitis B	3.5%
Cardiovasc…	1.1%

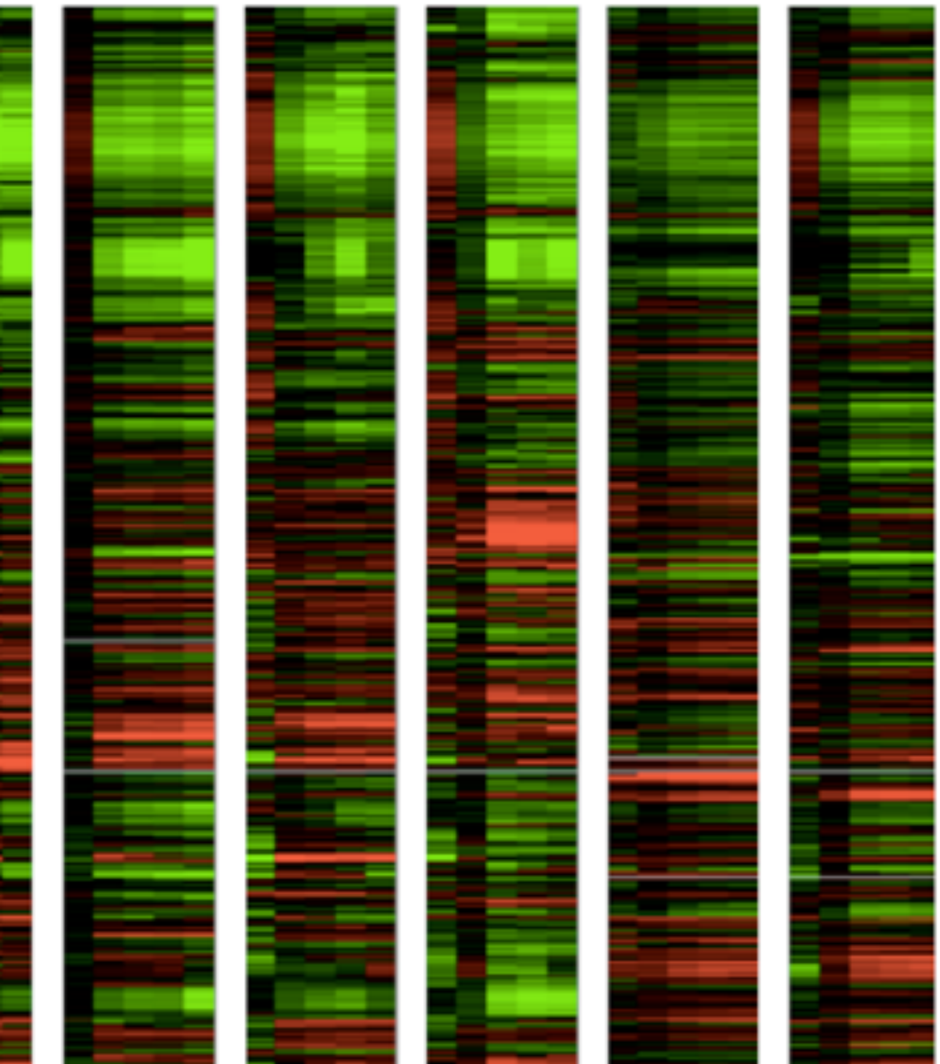


ReVision: Automated Chart Interpretation

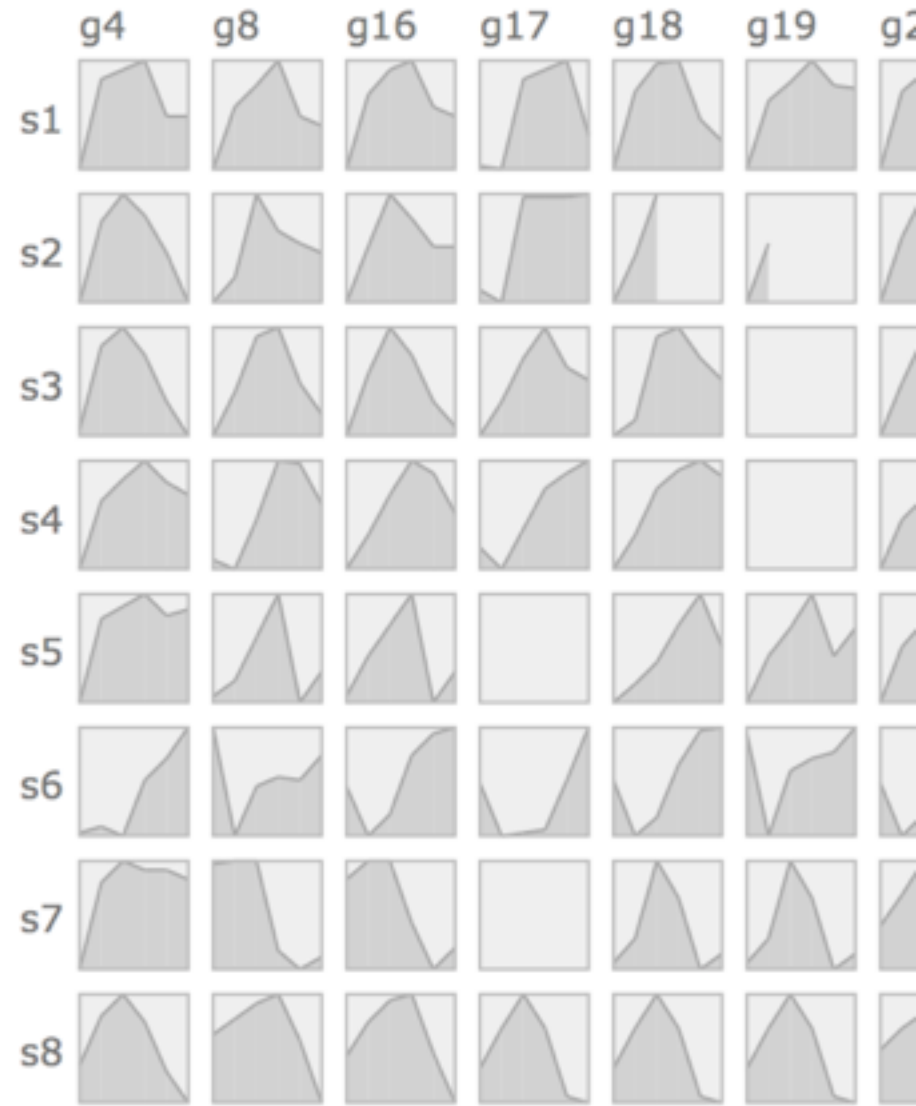
Analysis and redesign of chart images [Savva et al 2011]

Gene Expression Time-Series [Meyer et al 11]

Color Encoding



Position Encoding

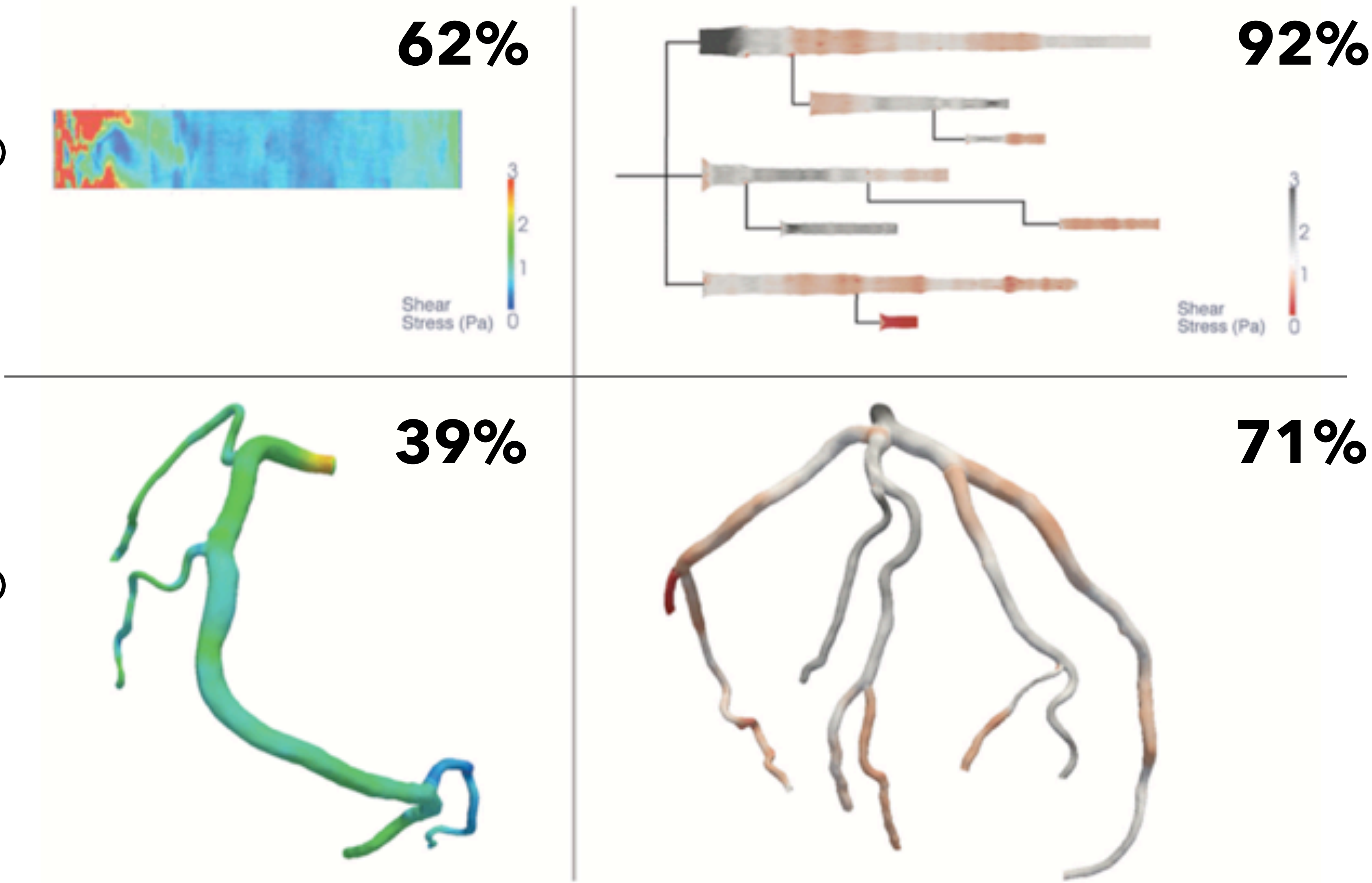


Artery Visualization [Borkin et al 11]

Rainbow Palette | Diverging Palette

2D

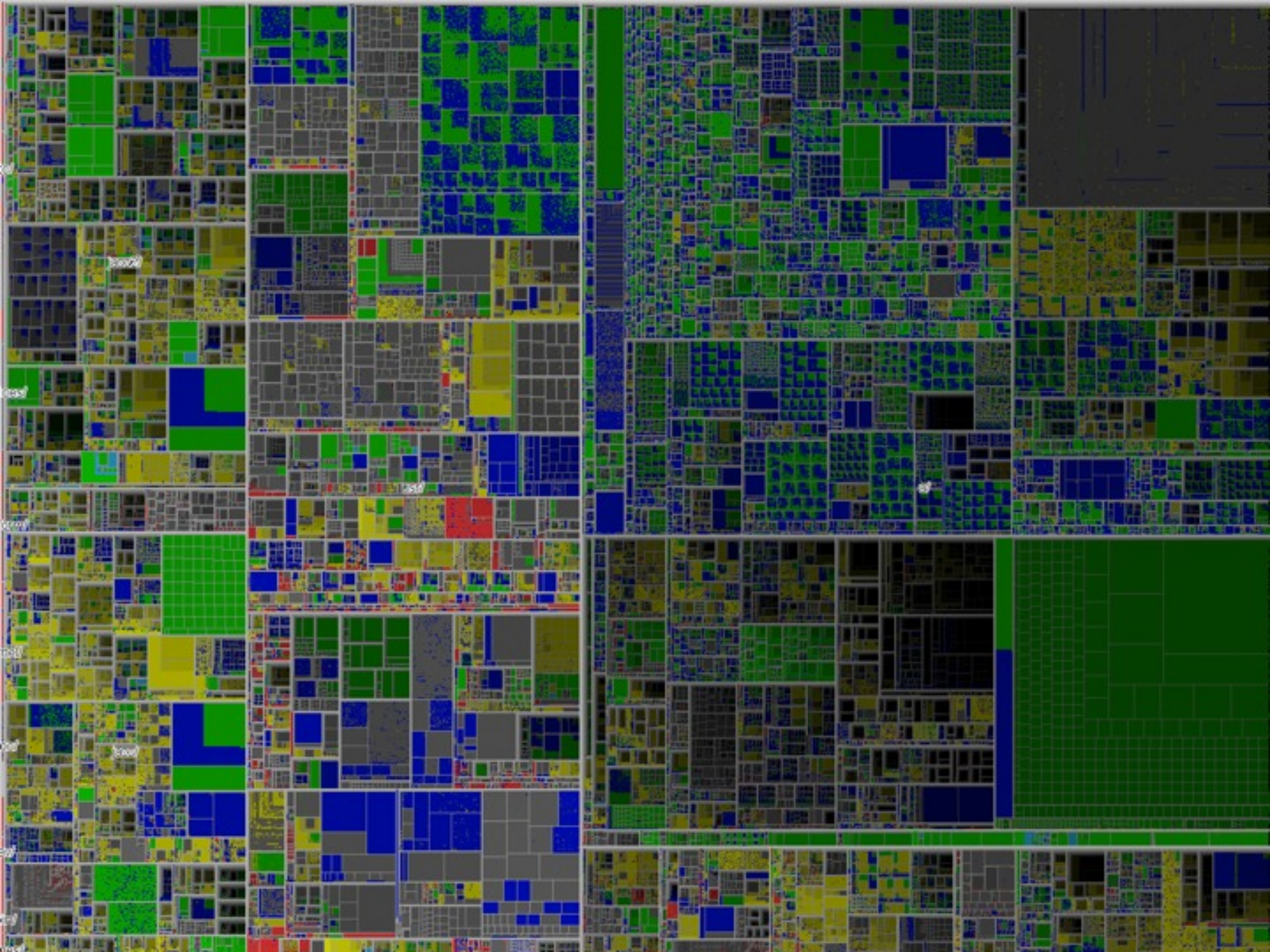
3D



FOLLOW-UP QUESTION:

What about **interactions** between encodings?

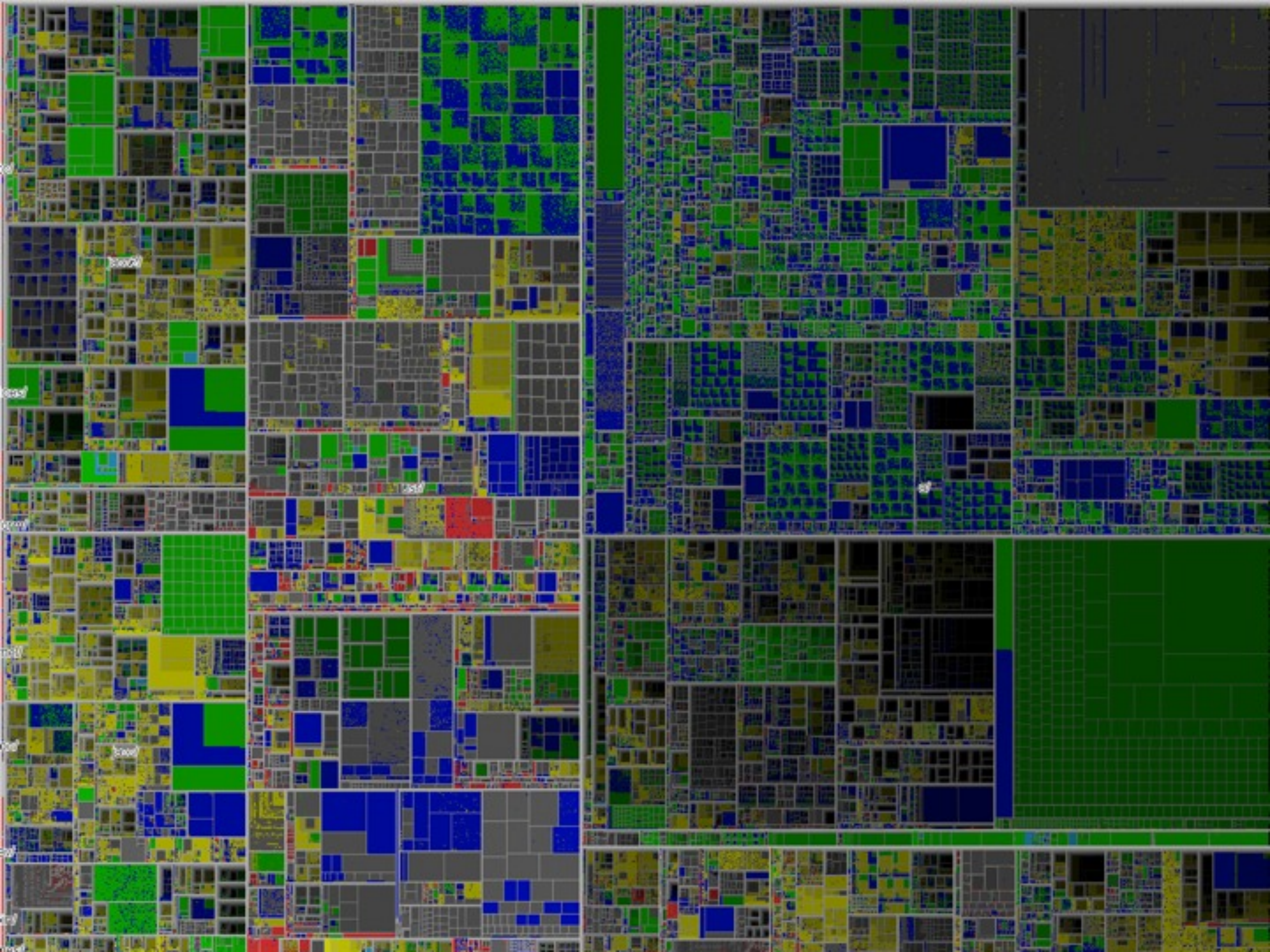
Data Density

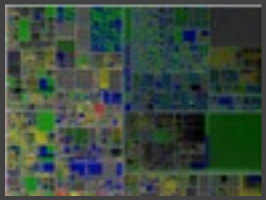


$$\text{Data Density} = \frac{(\text{\# entries in data})}{(\text{area of graphic})}$$

"Graphical excellence… gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space"

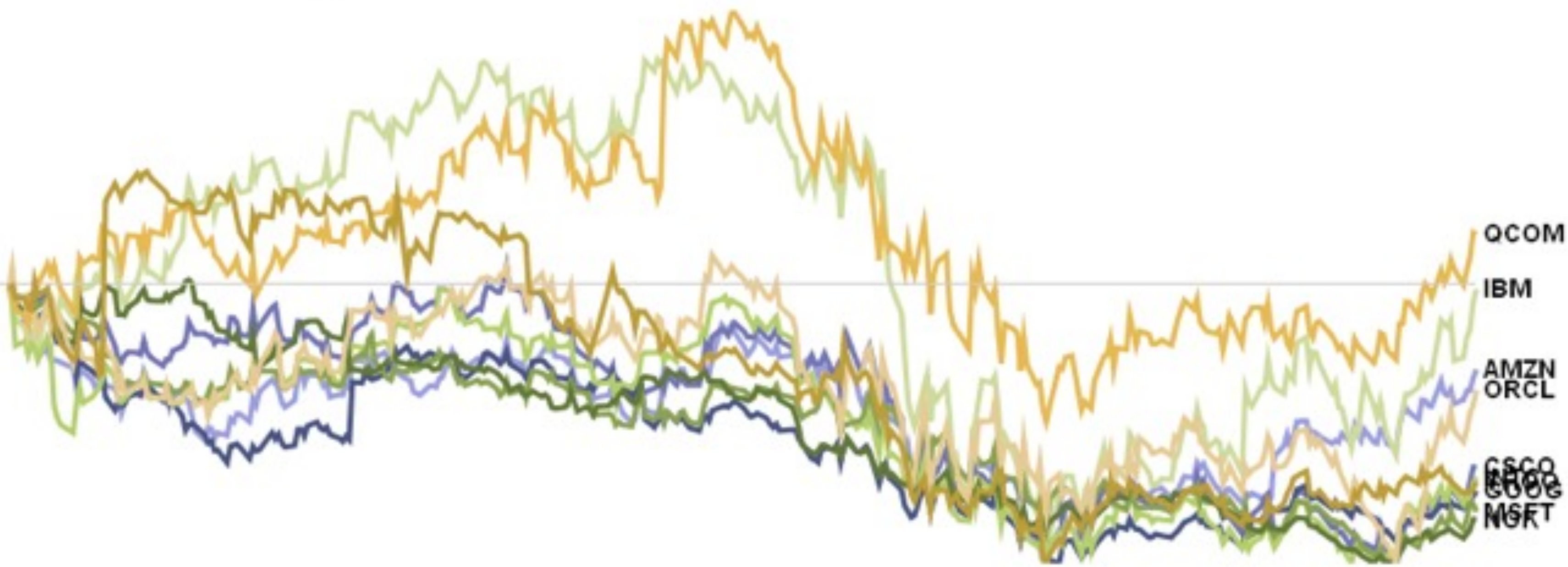
[Tufte 83]



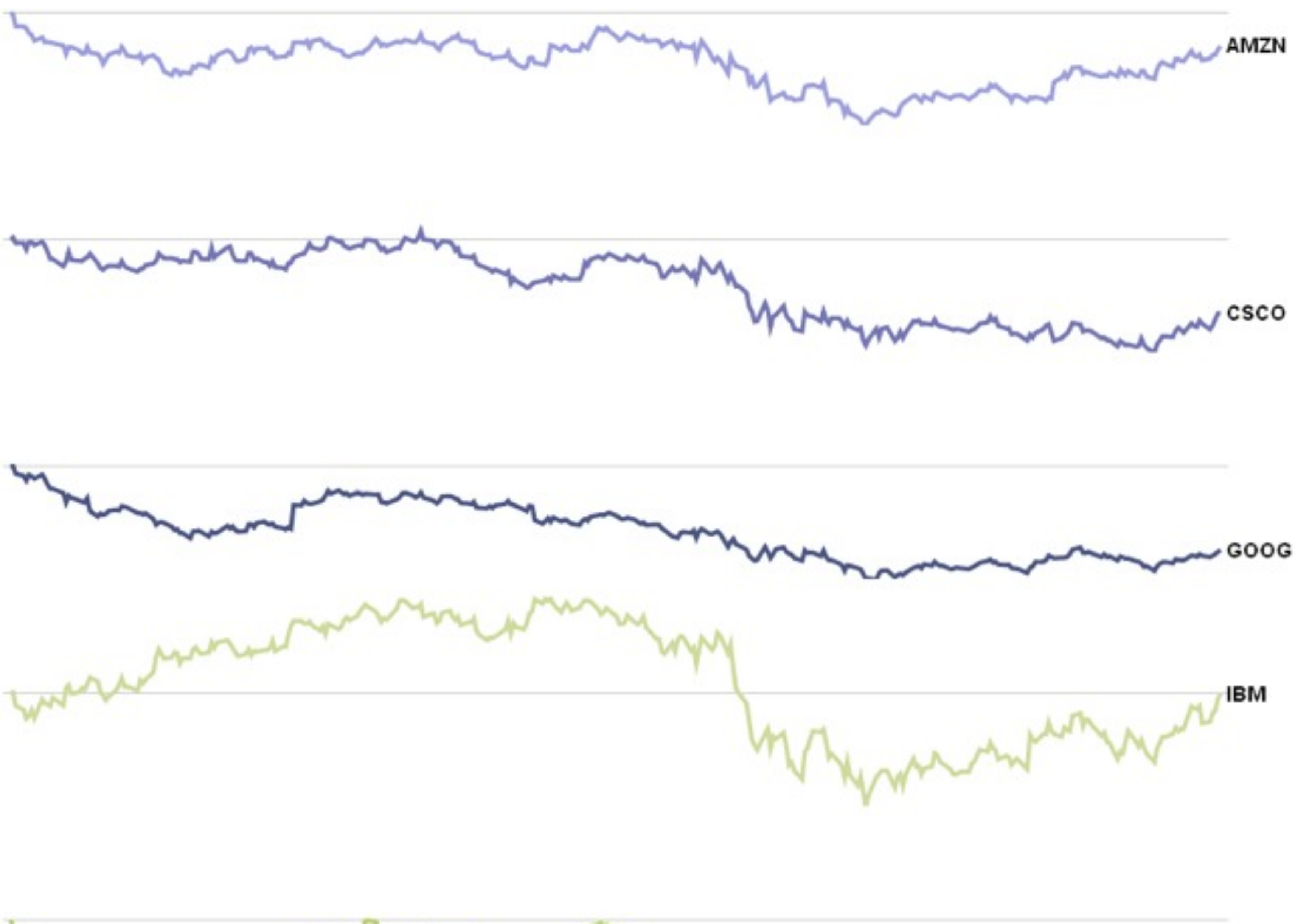


Relative Technology Stock Performance: Jan 2008 - Present

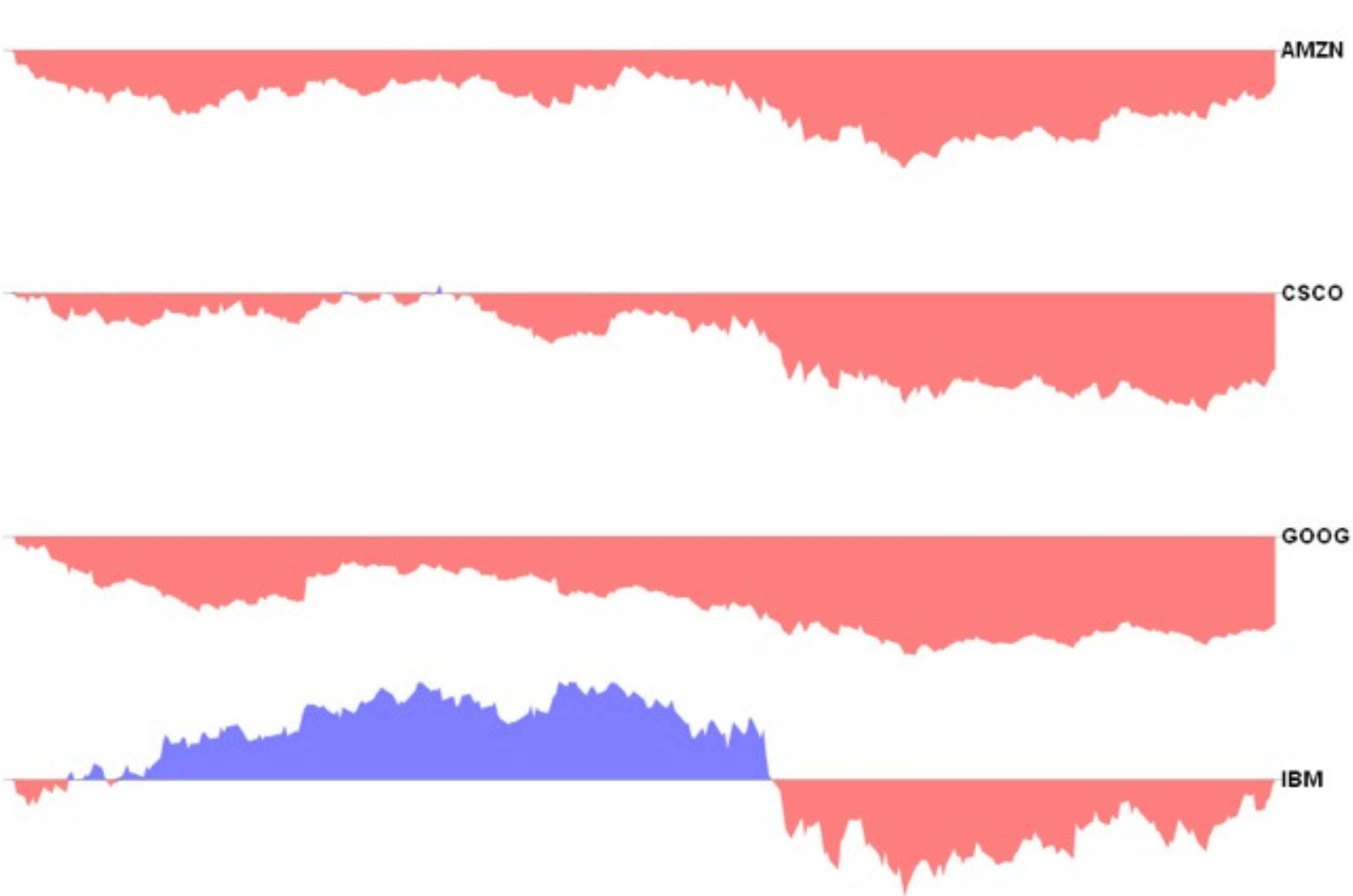
Relative Technology Stock Performance: Jan 2008 - Present



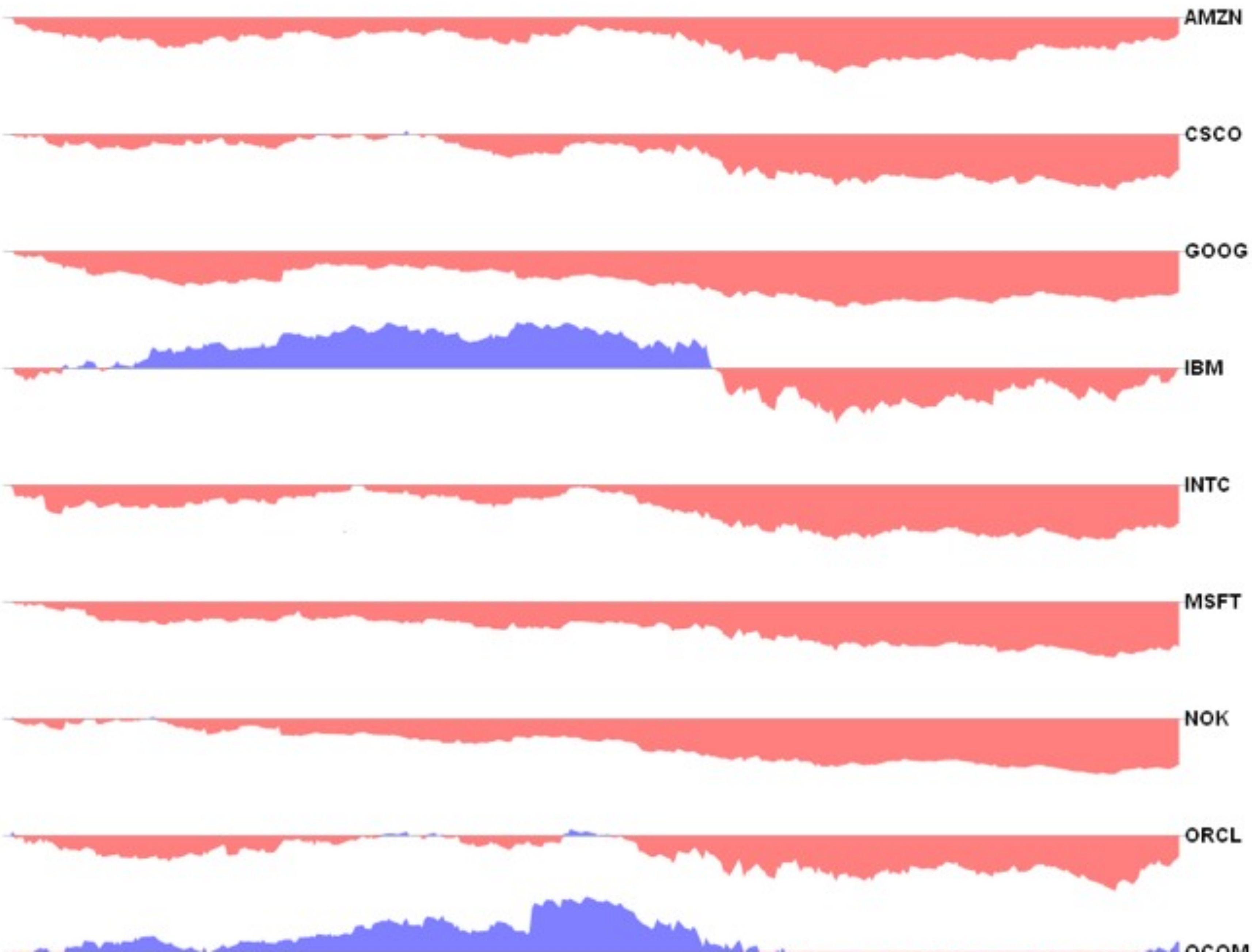
Relative Technology Stock Performance: Jan 2008 - Present



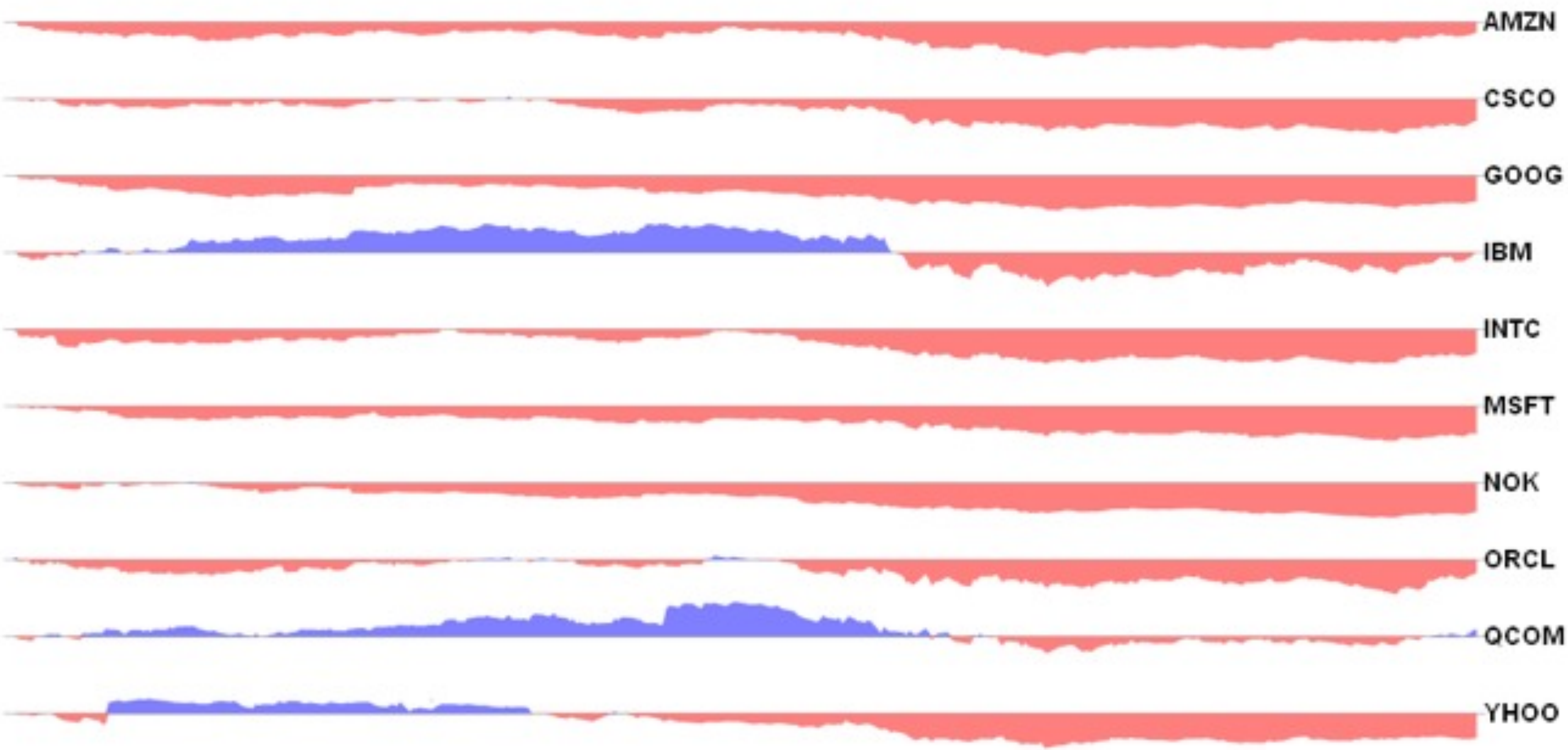
Relative Technology Stock Performance: Jan 2008 - Present



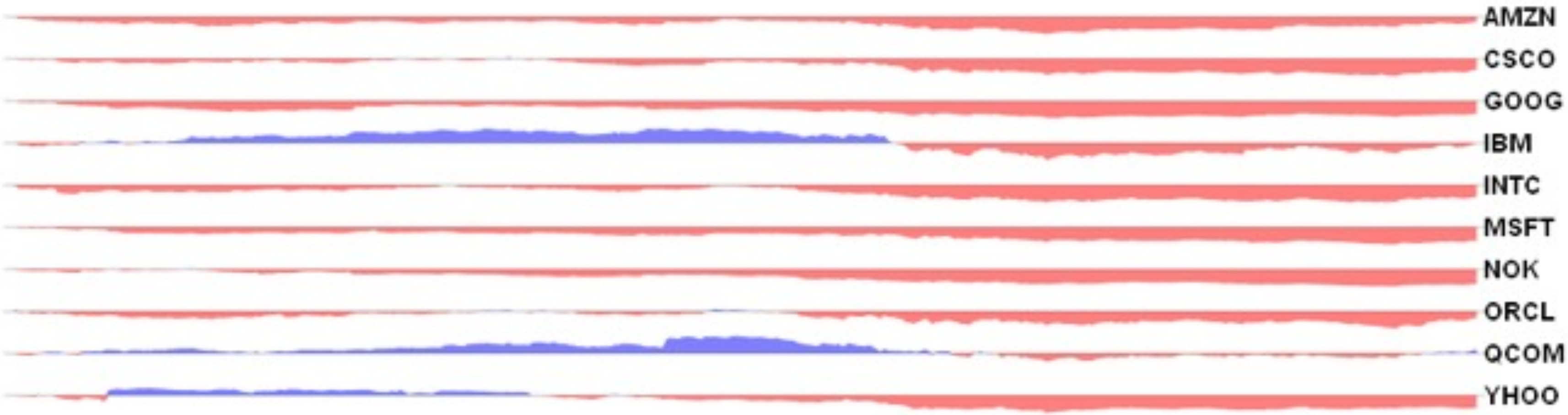
Relative Technology Stock Performance: Jan 2008 - Present



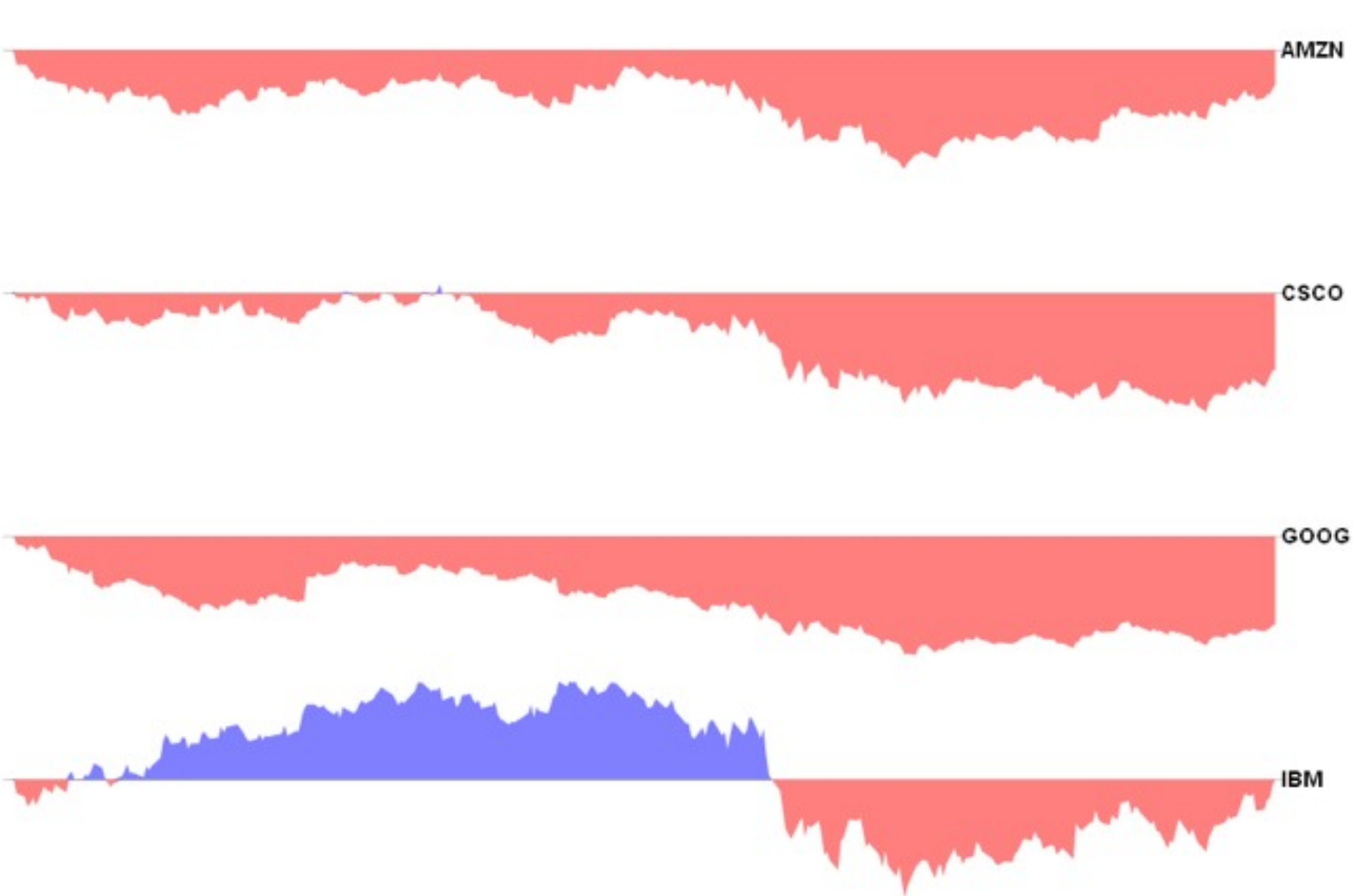
Relative Technology Stock Performance: Jan 2008 - Present



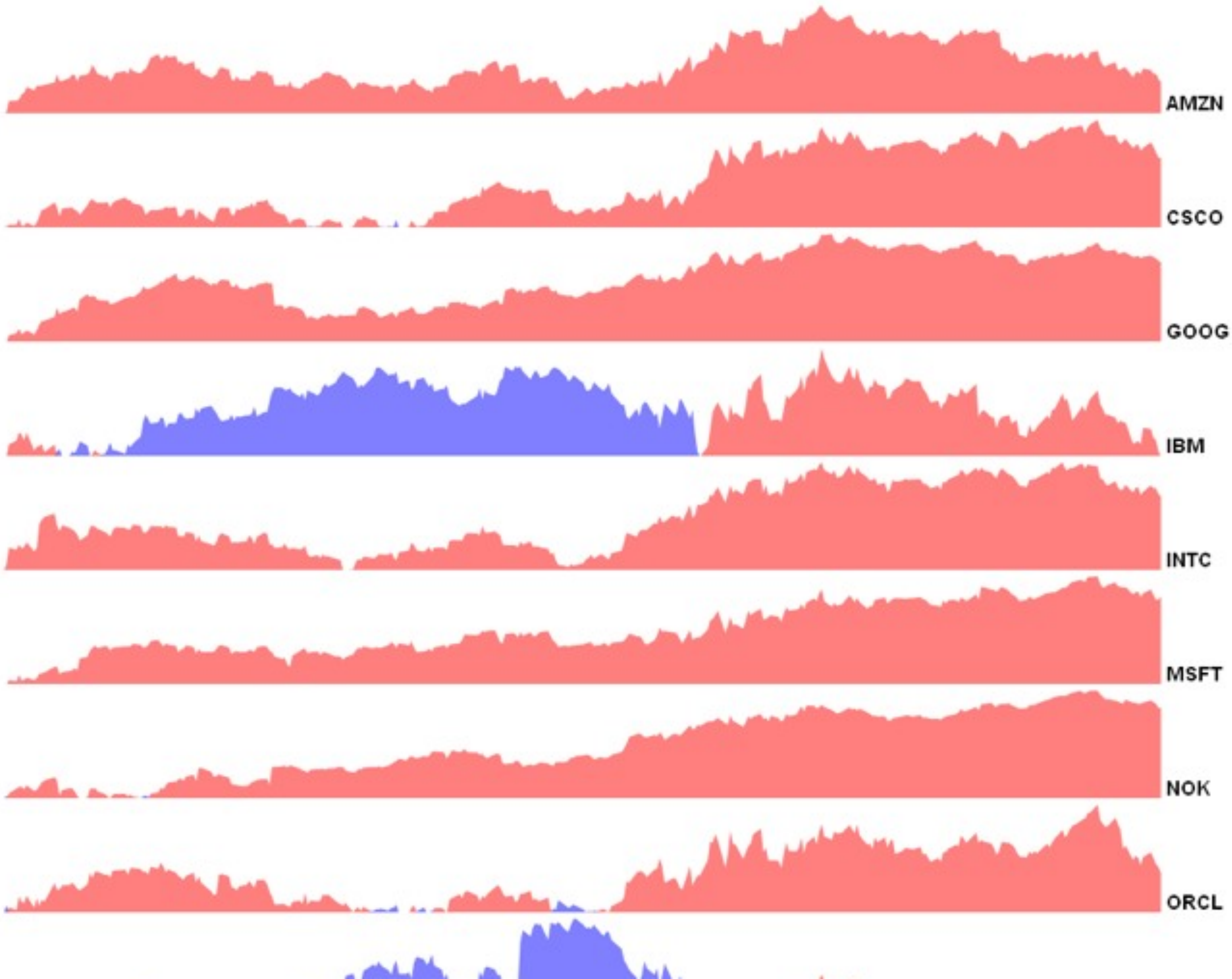
Relative Technology Stock Performance: Jan 2008 - Present



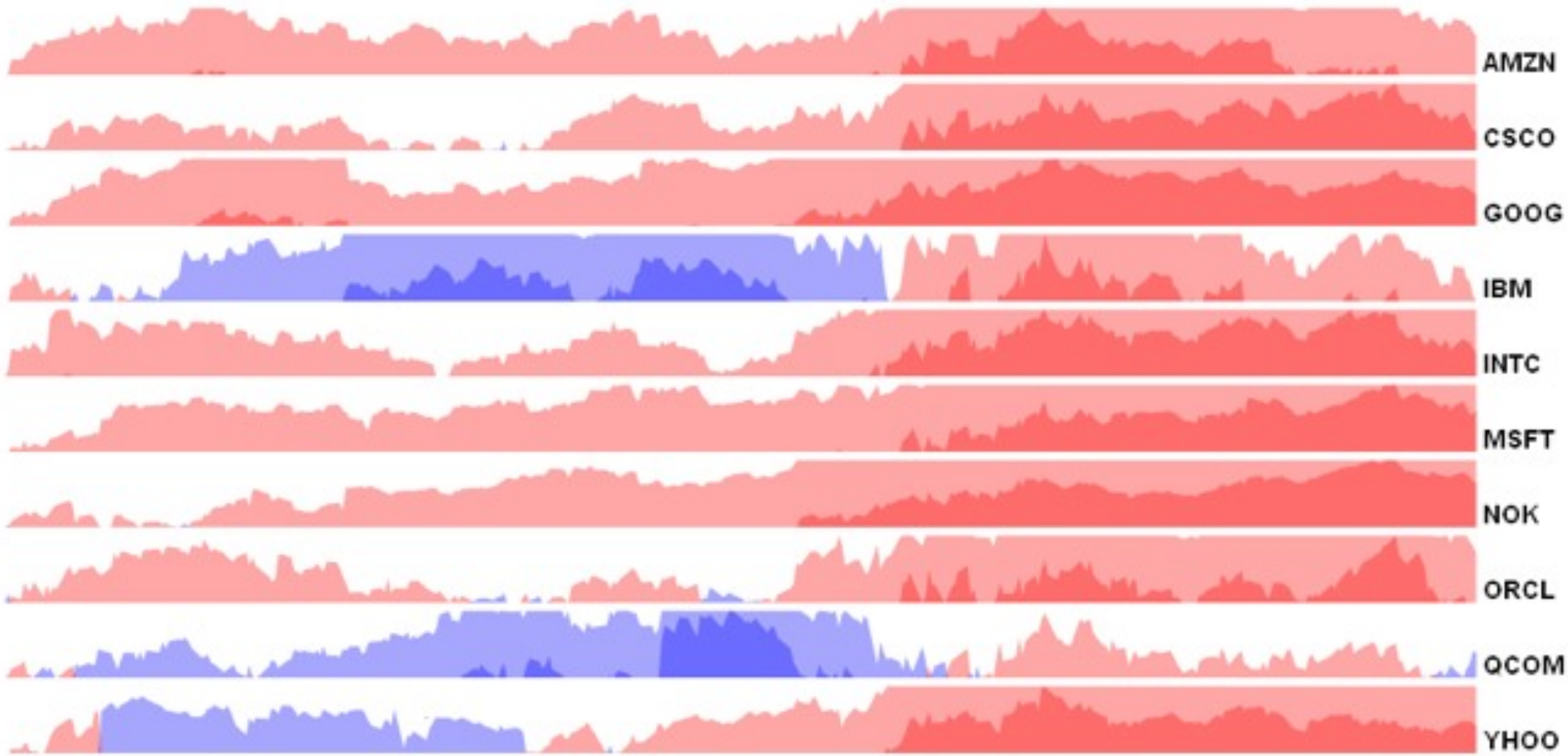
Relative Technology Stock Performance: Jan 2008 - Present



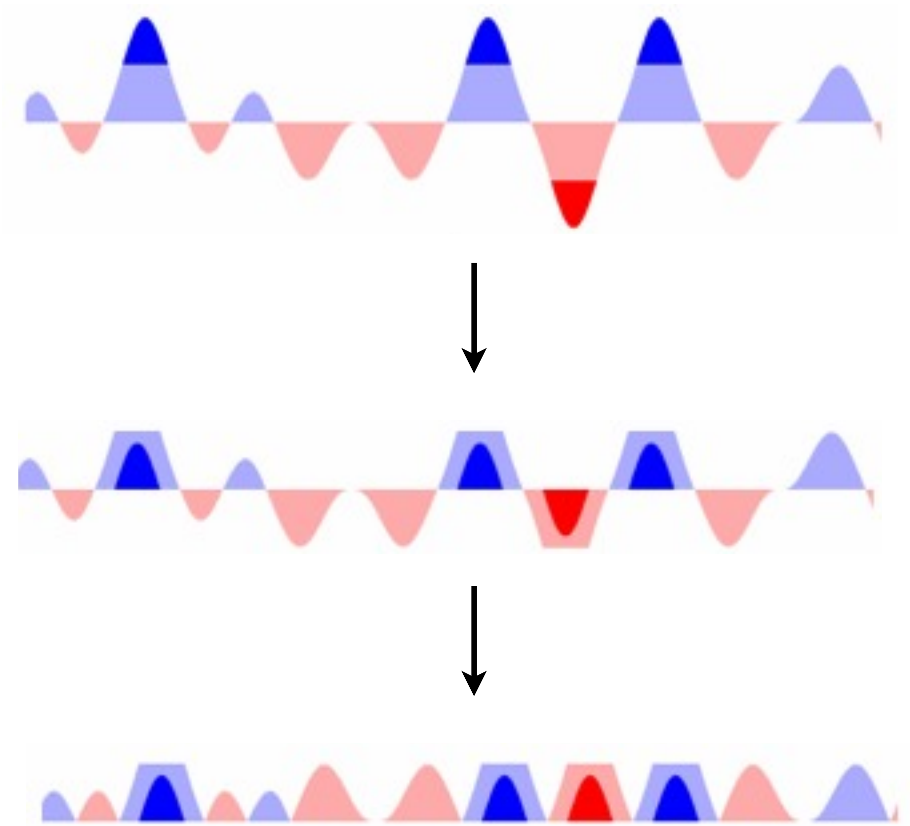
Relative Technology Stock Performance: Jan 2008 - Present



Relative Technology Stock Performance: Jan 2008 - Present



Horizon Graphs

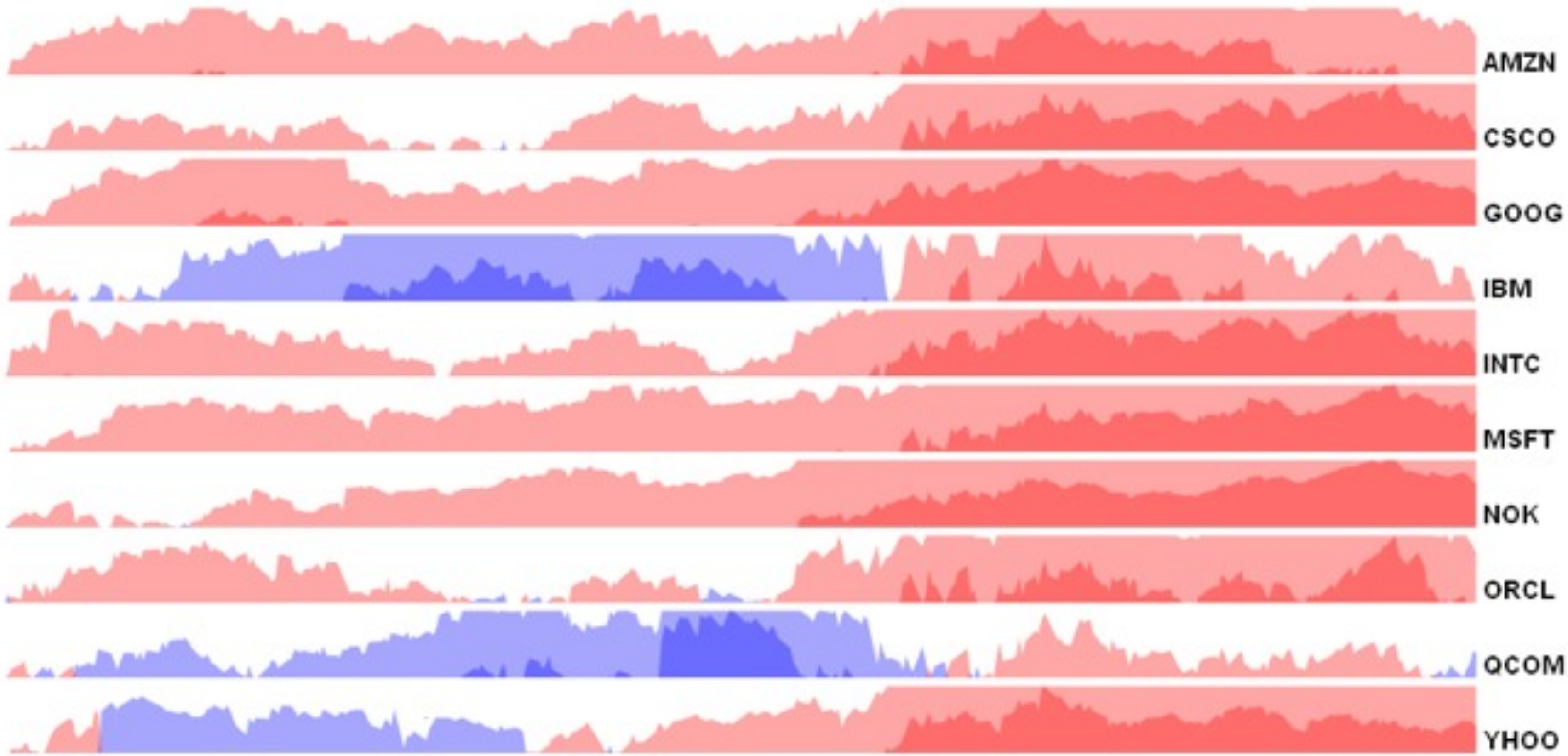


Segment Peaks

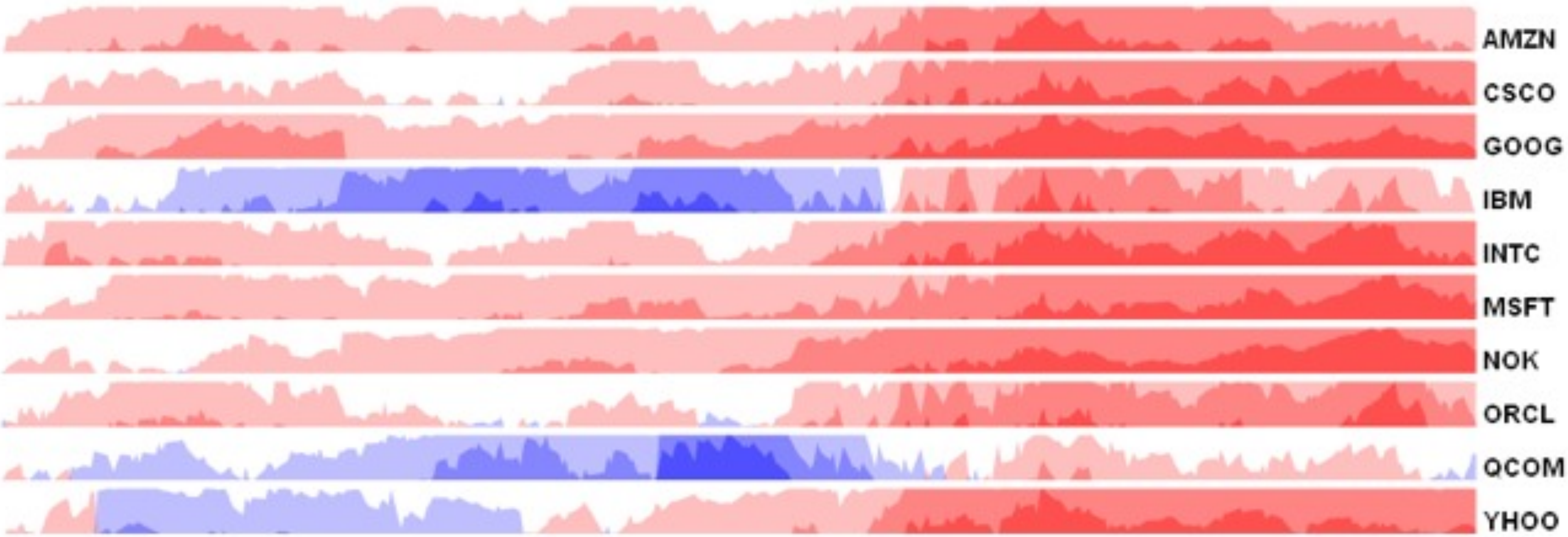
Layer Segments

Mirror Negative Values

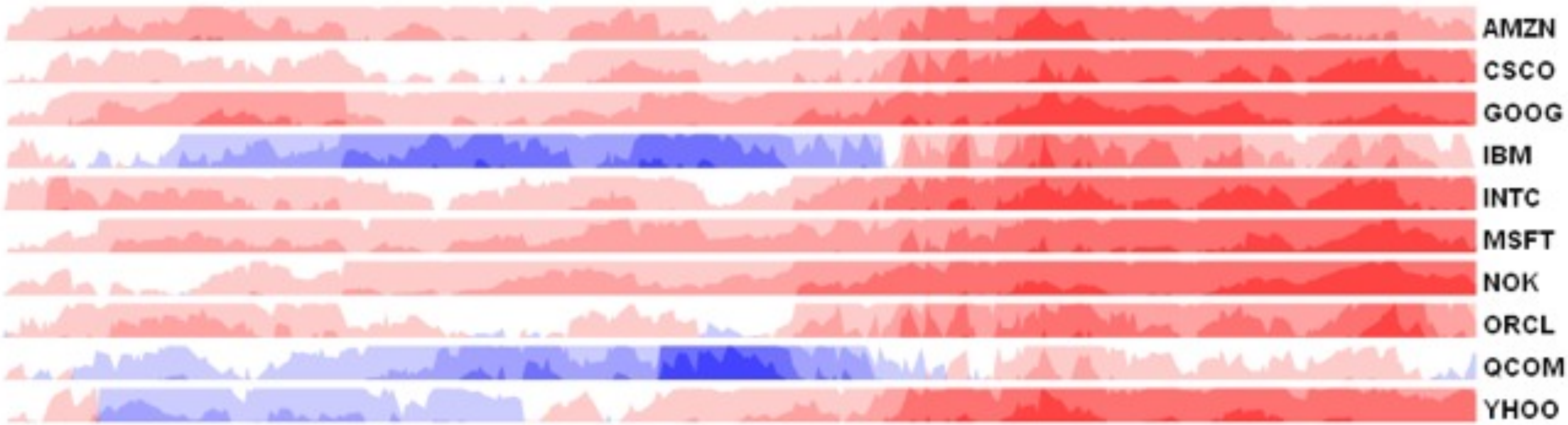
Relative Technology Stock Performance: Jan 2008 - Present



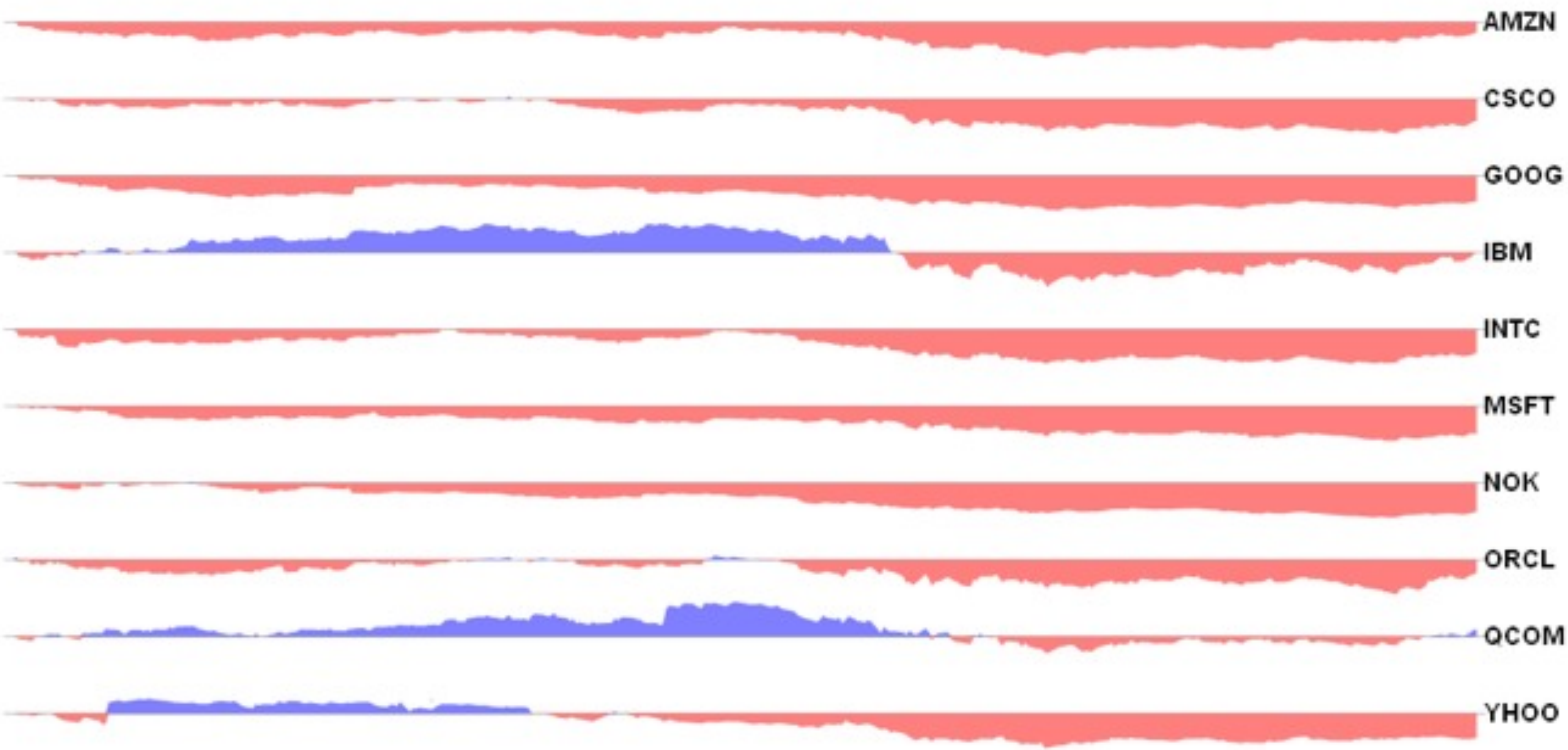
Relative Technology Stock Performance: Jan 2008 - Present



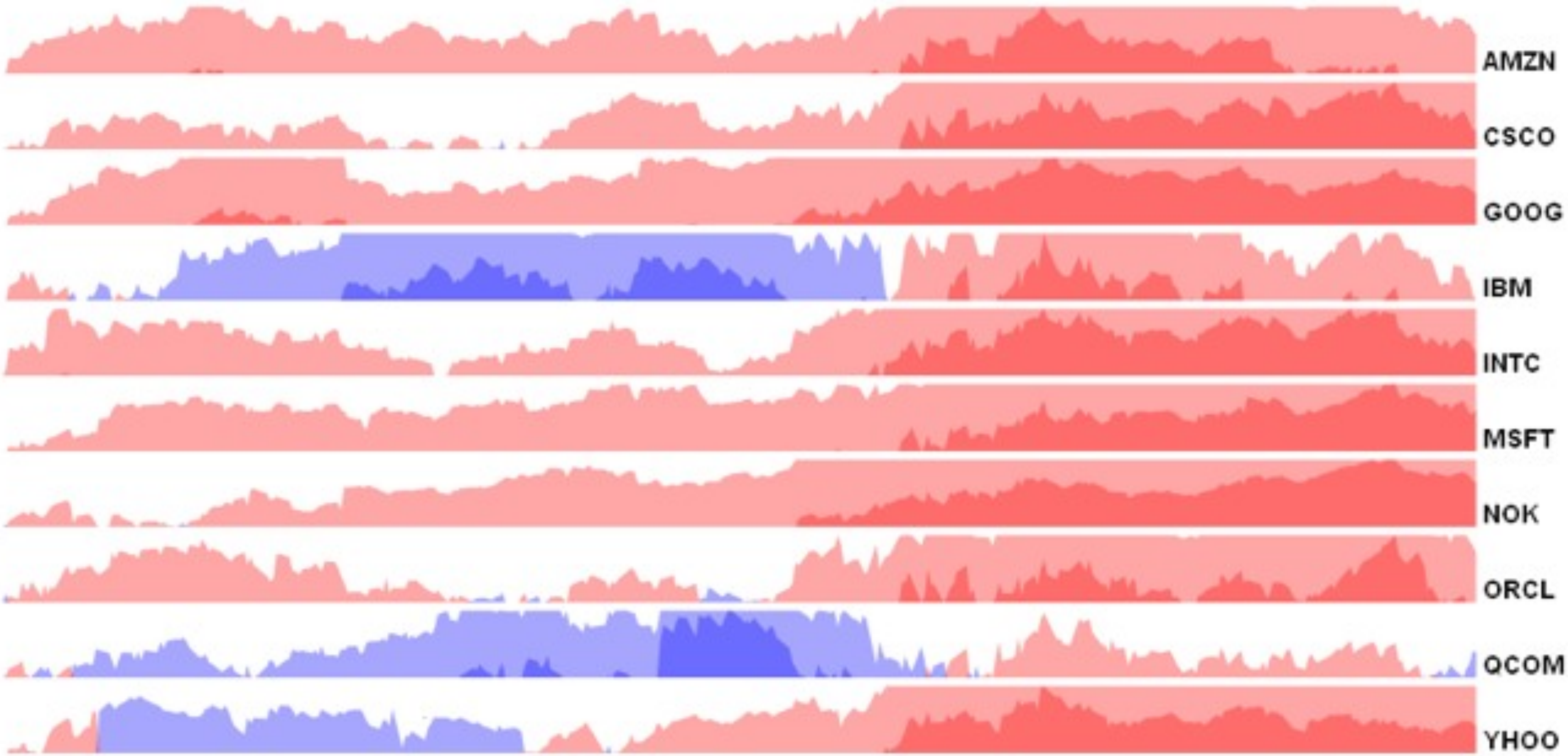
Relative Technology Stock Performance: Jan 2008 - Present

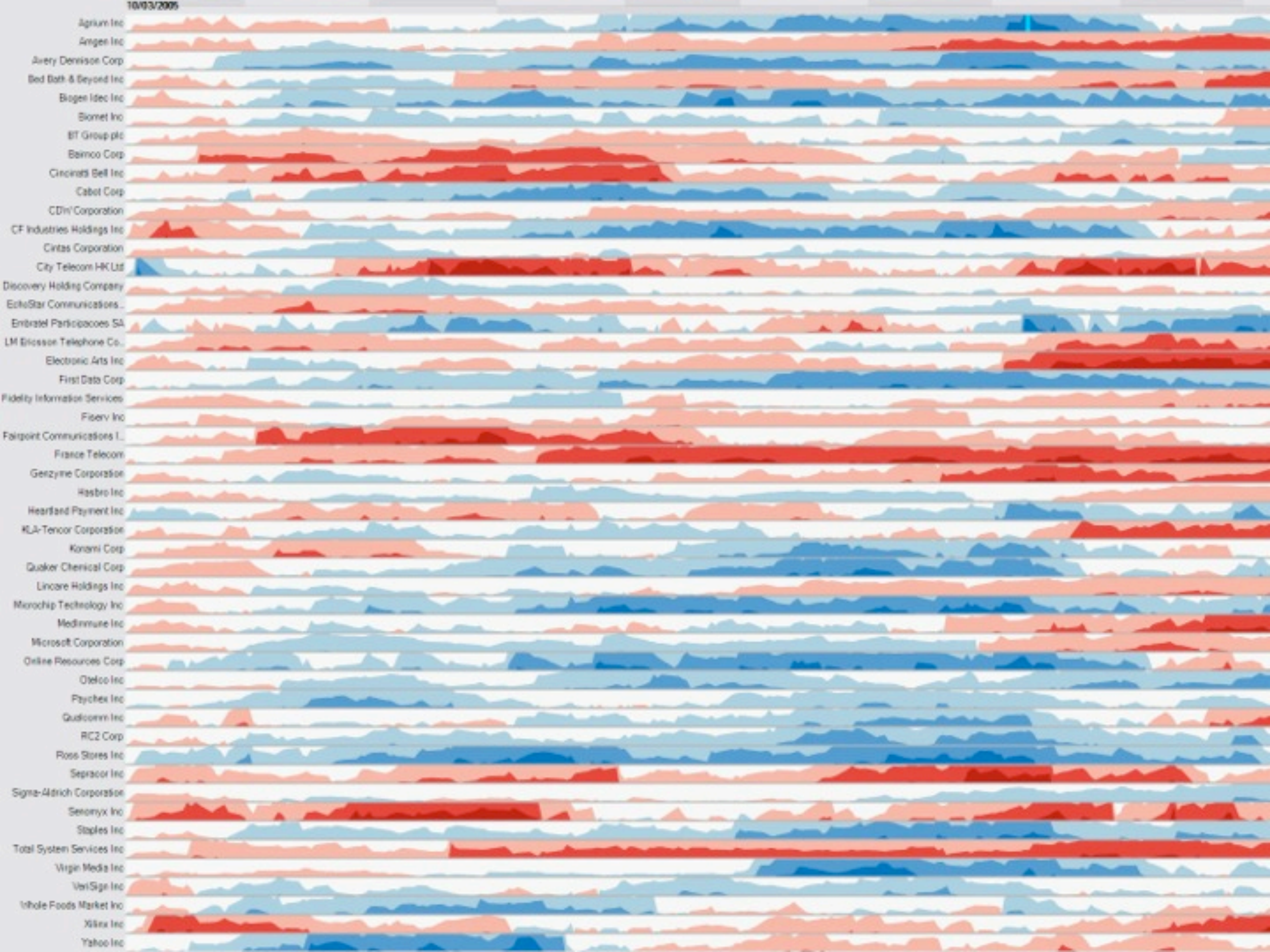


Relative Technology Stock Performance: Jan 2008 - Present



Relative Technology Stock Performance: Jan 2008 - Present

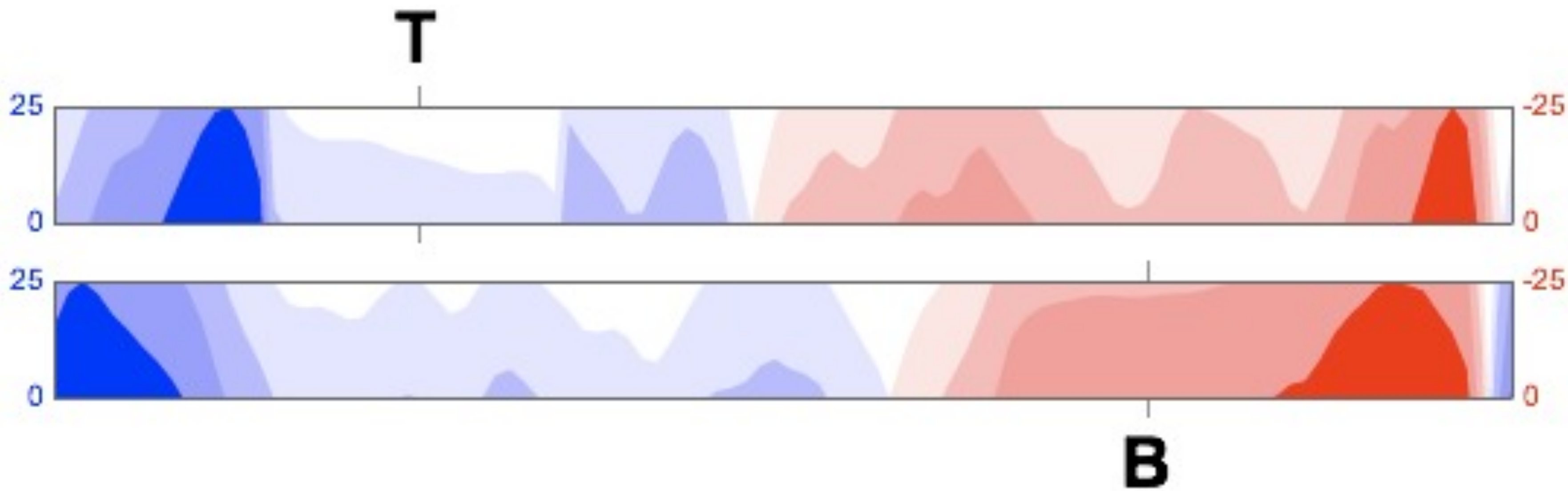



10/03/2005
Agrium Inc
Amgen Inc
Avery Dennison Corp
Bed Bath & Beyond Inc
Biogen Idec Inc
Biomet Inc
BT Group plc
Bemco Corp
Cincinnati Bell Inc
Cabot Corp
CDI Corporation
CF Industries Holdings Inc
Cintas Corporation
City Telecom HK Ltd
Discovery Holding Company
EchoStar Communications...
Embratel Participacoes SA
LM Ericsson Telephone Co...
Electronic Arts Inc
First Data Corp
Fidelity Information Services
Fiserv Inc
Fairpoint Communications I...
France Telecom
Genzyme Corporation
Hasbro Inc
Heartland Payment Inc
KLA-Tencor Corporation
Konami Corp
Quaker Chemical Corp
Lincare Holdings Inc
Microchip Technology Inc
MedImmune Inc
Microsoft Corporation
Online Resources Corp
Otelco Inc
Paychex Inc
Qualcomm Inc
RC2 Corp
Ross Stores Inc
Sepracor Inc
Sigma-Aldrich Corporation
Senomyx Inc
Staples Inc
Total System Services Inc
Virgin Media Inc
VeriSign Inc
Whole Foods Market Inc
Xilinx Inc
Yahoo Inc

Experiment: Chart Type & Size

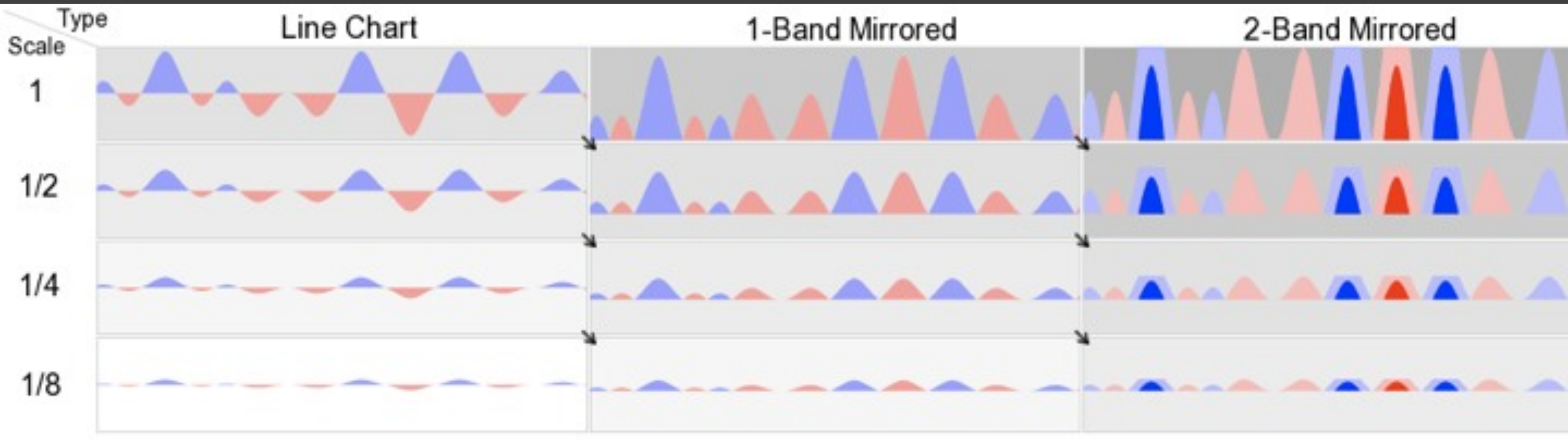
Q1: How do mirroring and layering affect estimation time and accuracy compared to line charts?

Q2: How does chart size affect estimation time and accuracy?



Estimate the difference between T and B (0-200) to within 5 values.

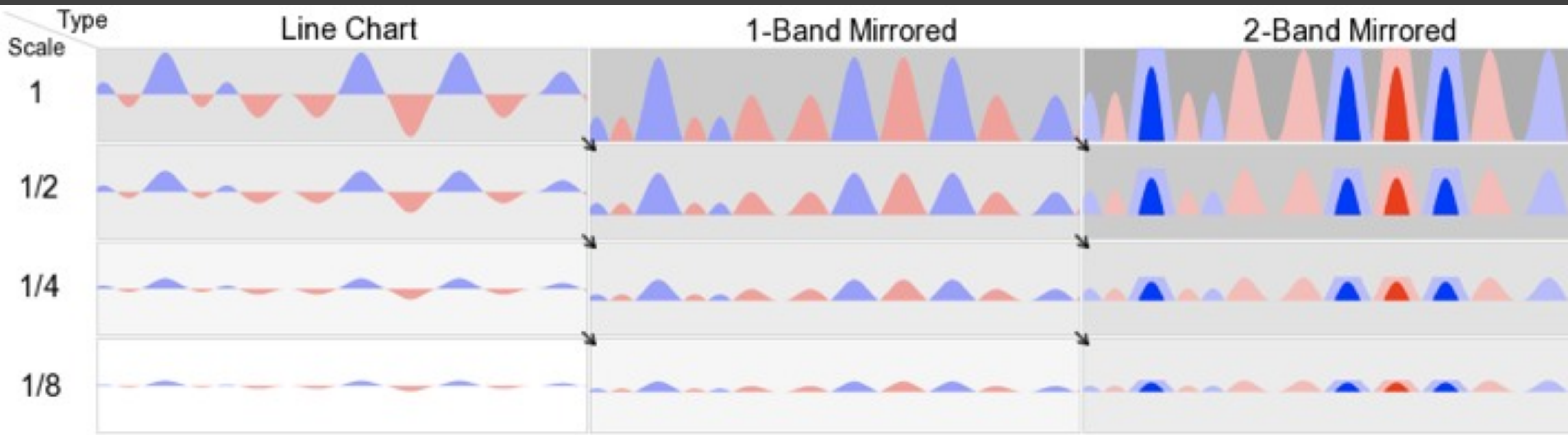
Experiment Design



3 (chart type) x 4 (size) within-subjects design

- N = 30 (17 male, 13 female), undergrads
- 14.1 inch LCD display, 1024 x 768 resolution
- At scale = 1, chart is 13.9 x 1.35 cm (48 px)

Experiment Design

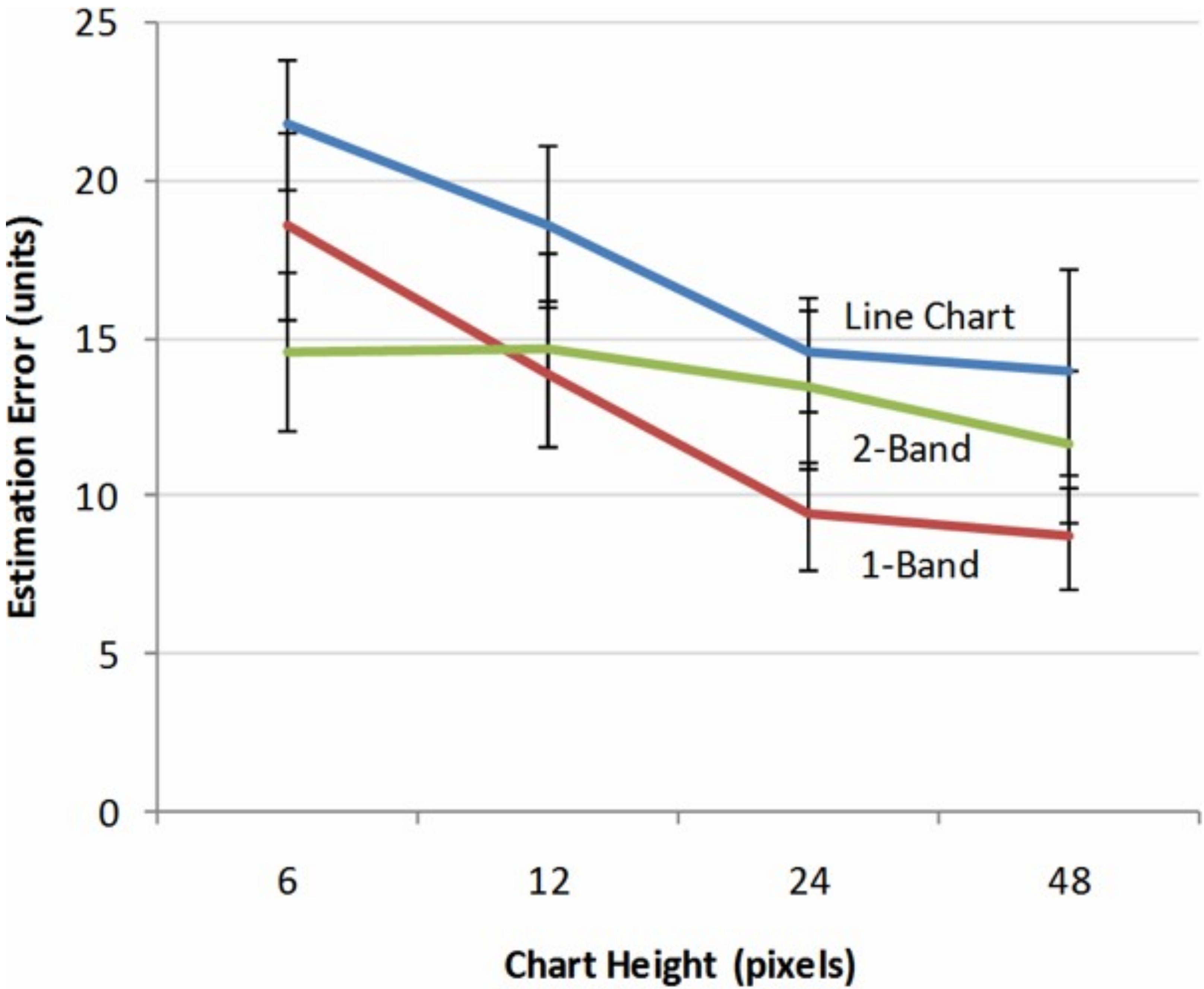


3 (type) x 4 (size) within-subjects design

N = 30 (17 male, 13 female), undergrads

2 (type) x 3 (size:1/8, 1/12, 1/24) follow-up

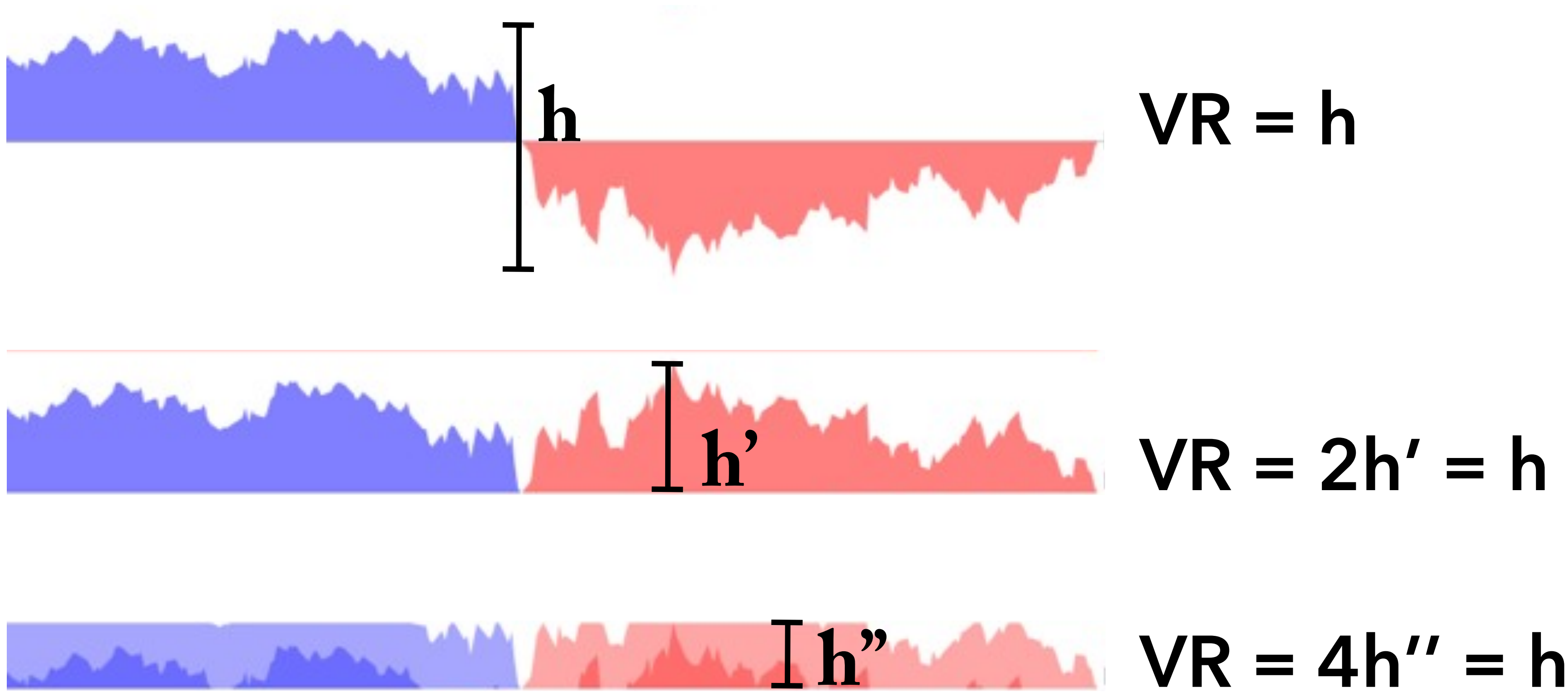
N = 8 (6 male, 2 female), engineering grads

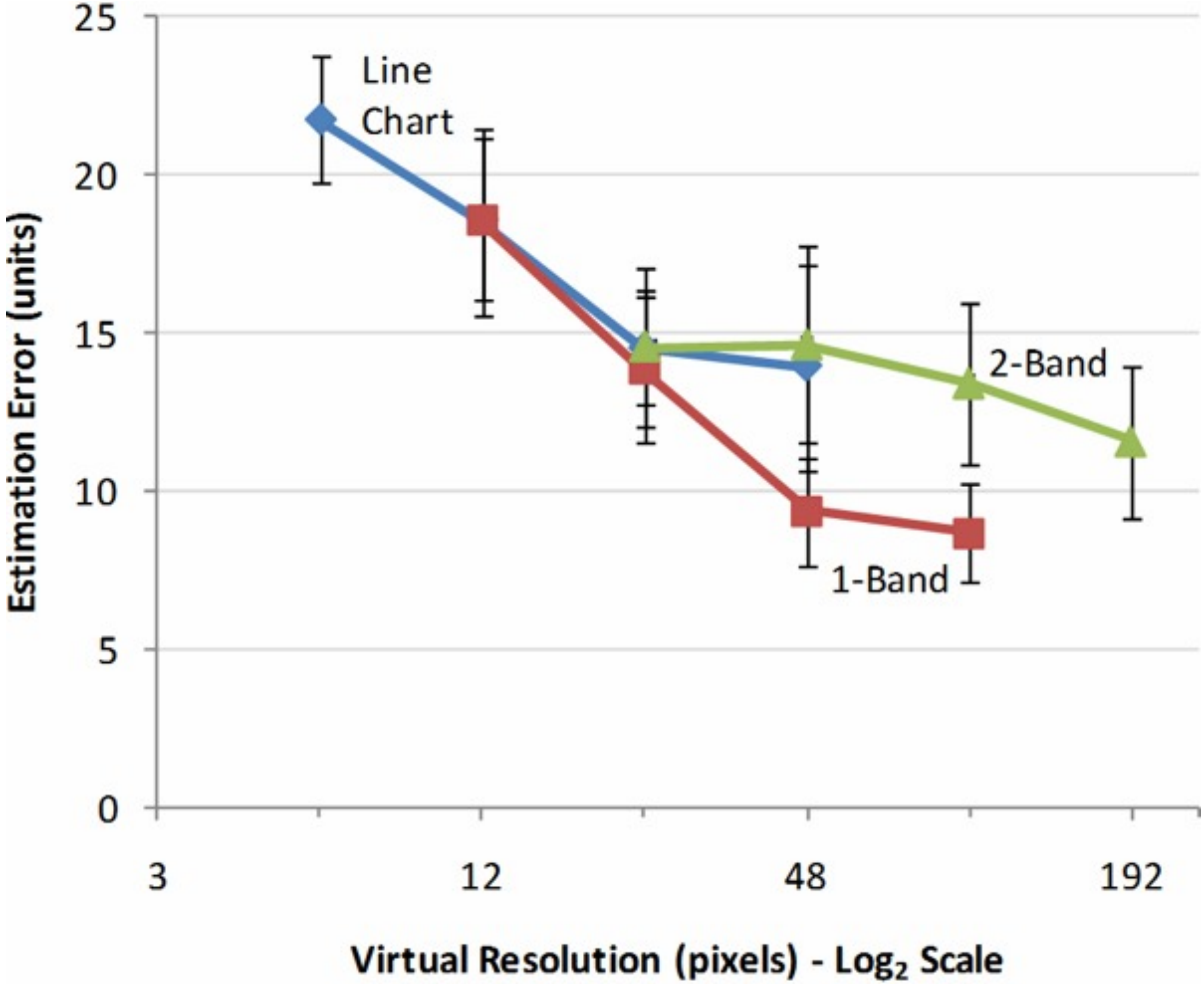


25
20
15
10
5
0
Estimation Error (units)
Line Chart
2-Band
1-Band
6
12
24
48
Chart Height (pixels)

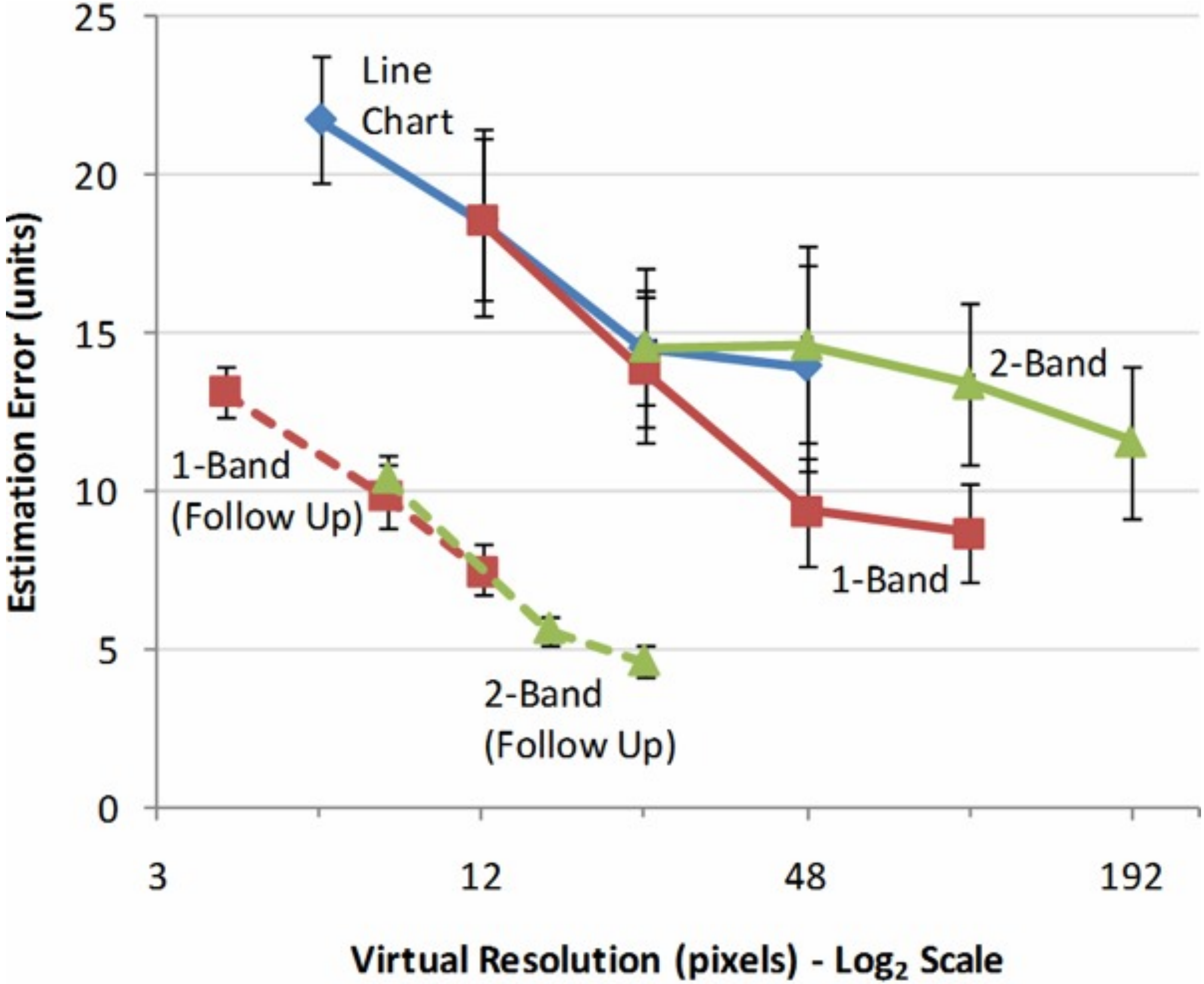
Virtual Resolution (VR)

The un-mirrored, un-layered height of a chart

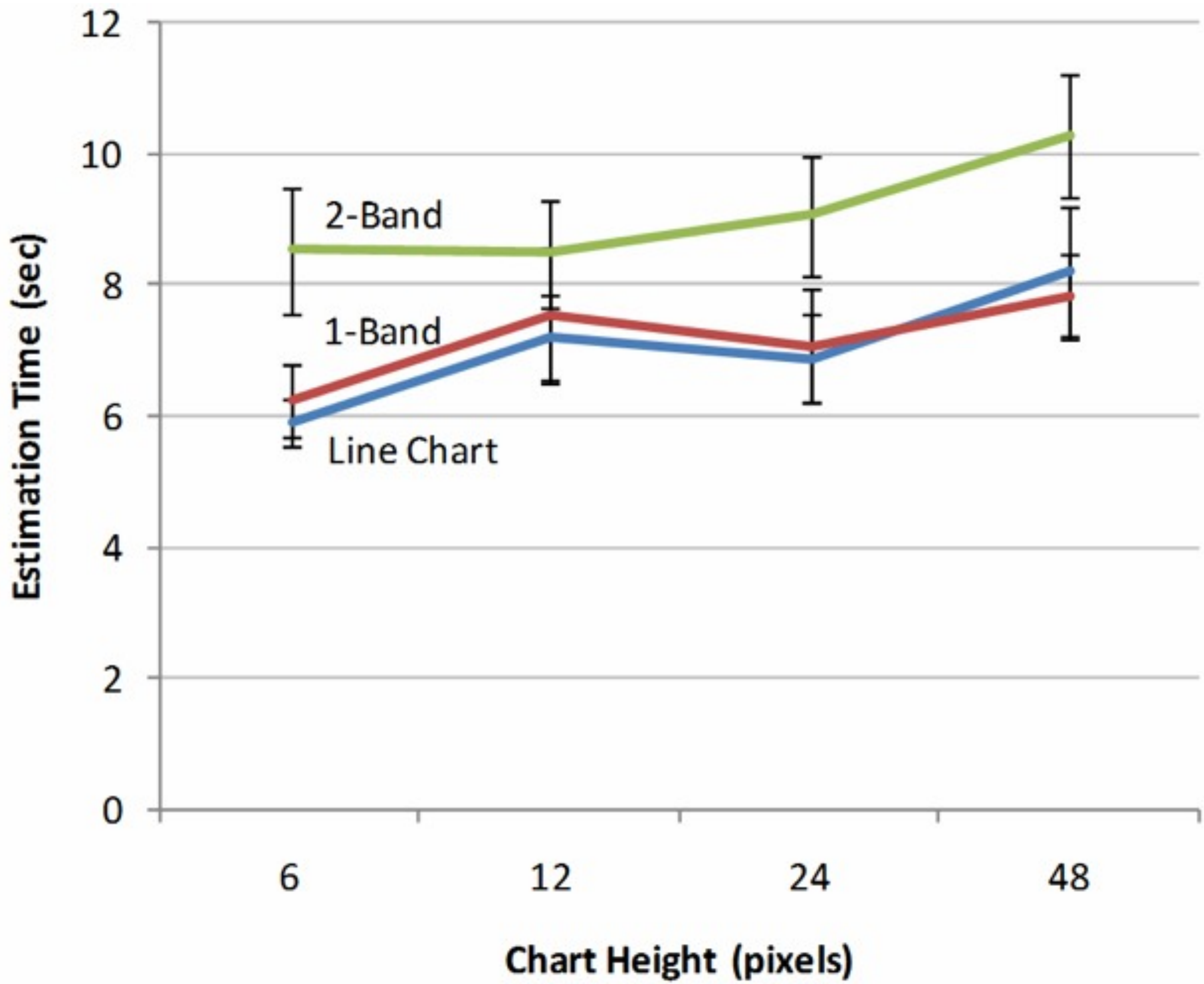




Estimation Error (units)
25
20
15
10
5
0
Line Chart
2-Band
1-Band
3
12
48
192
Virtual Resolution (pixels) - Log2 Scale



Line Chart
2-Band
1-Band (Follow Up)
1-Band
2-Band (Follow Up)
Estimation Error (units)
25
20
15
10
5
0
3
12
48
192
Virtual Resolution (pixels) - Log2 Scale


12
10
8
6
4
2
0
Estimation Time (sec)
2-Band
1-Band
Line Chart
6
12
24
48
Chart Height (pixels)

Experiment Results

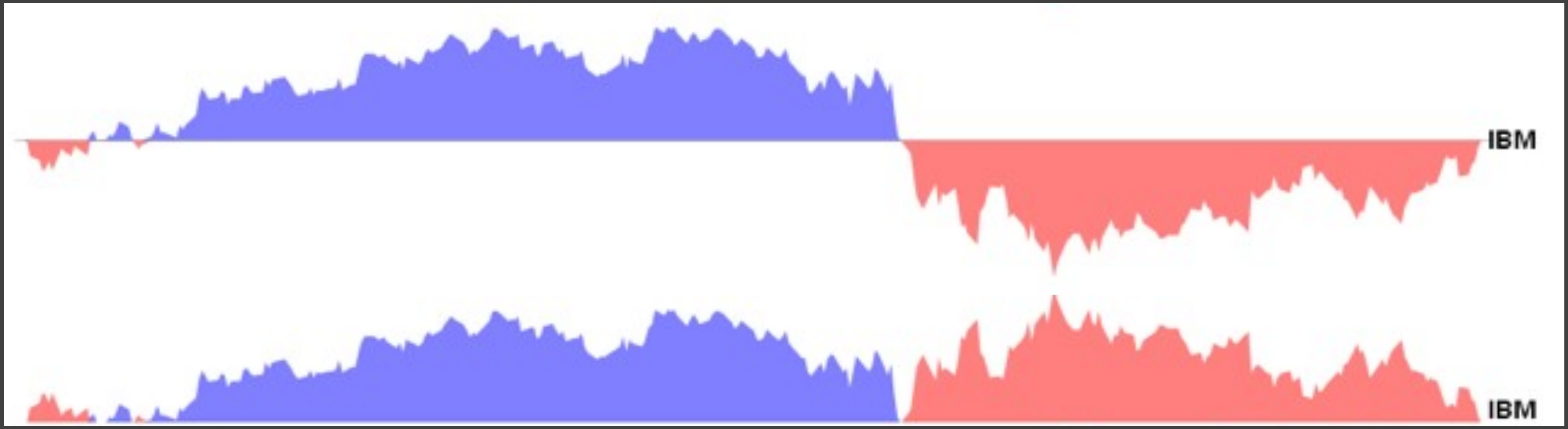
Q1: 2-band horizon graph (but not mirrored graph) has higher baseline estimation time and error.

Q2: Estimation error increases as the *virtual resolution* decreases.

Estimation time decreases as the *physical height* decreases.

Design Guidelines

Mirroring does not hamper perception

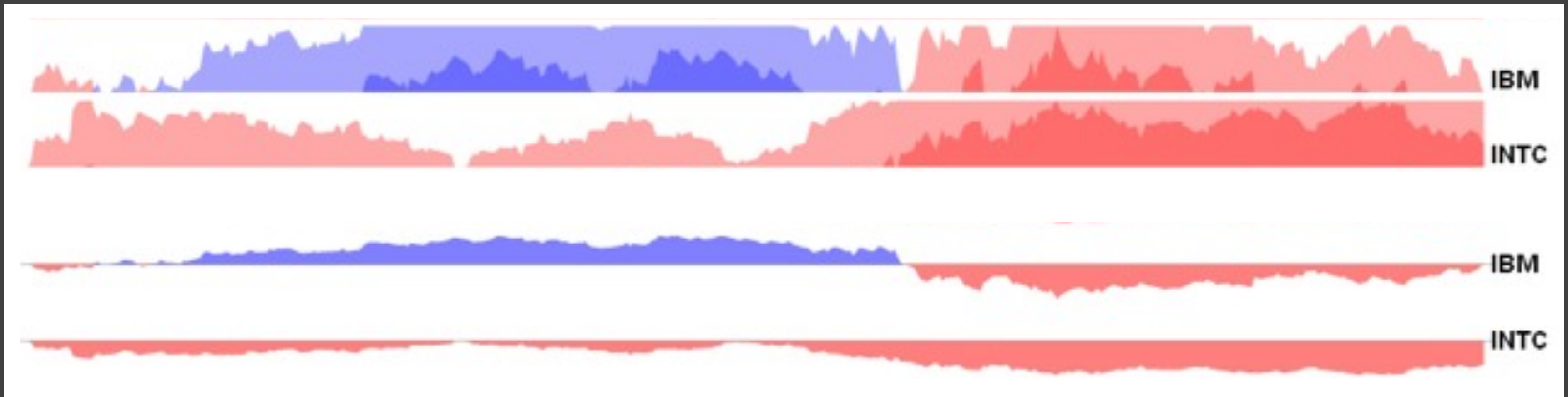


Design Guidelines

Mirroring does not hamper perception

Layered bands beneficial for smaller charts

- **2-band mirror charts** more accurate for heights under 6.8mm (24 pixels @ 1024x768)
- Predict benefits for 3 bands under 1.7mm (6 px)



Design Guidelines

Mirroring does not hamper perception

Layered bands beneficial for smaller charts

Optimal chart sizing

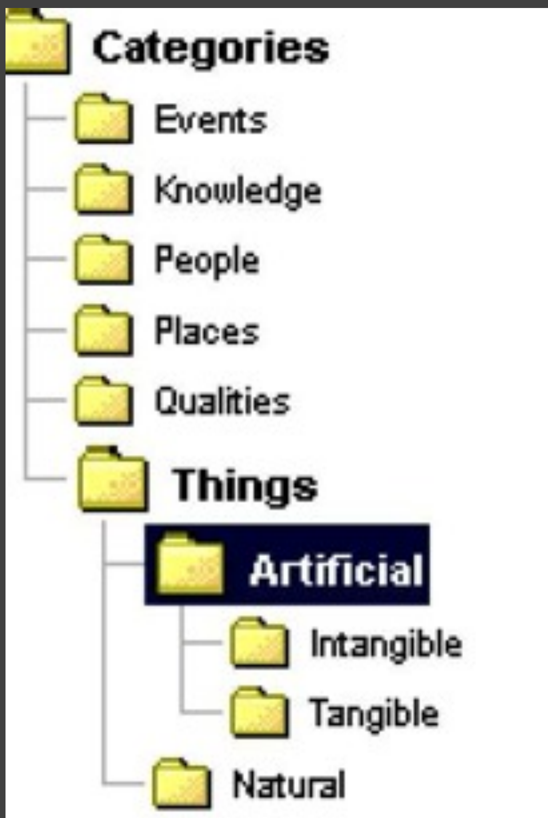
- **Sweet spots** in time/error curves
- 6.8mm (24 px) for line chart & mirrored chart
- 3.4mm (12 px) for 2-band horizon graph

FOLLOW-UP QUESTION:

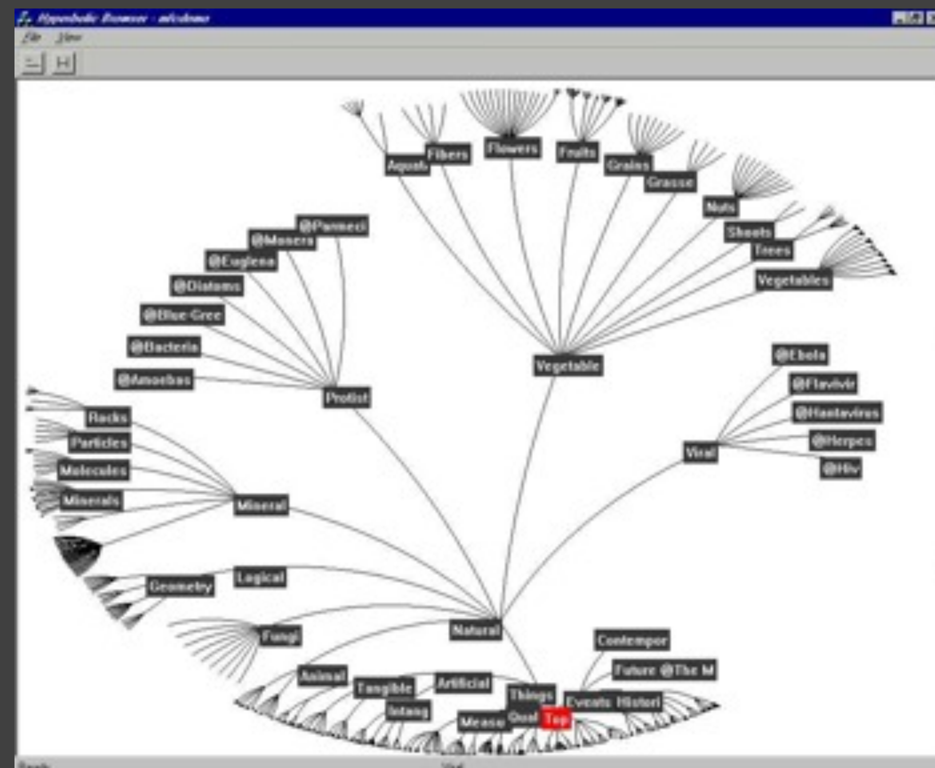
What other **tasks** and **performance measures** should one test?

Trees

The Great Browse-Off! [CHI 97]



vs.



Microsoft File Explorer

Xerox PARC Hyperbolic Tree

Which visualization is better?

Which visualization is better?

Xerox PARC researchers ran eye-tracking studies to investigate... [Pirolli et al 00]

Which visualization is better?

Xerox PARC researchers ran eye-tracking studies to investigate... [Pirolli et al 00]

Subjects performed both retrieval and comparison tasks of varying complexity.

Which visualization is better?

Xerox PARC researchers ran eye-tracking studies to investigate... [Pirolli et al 00]

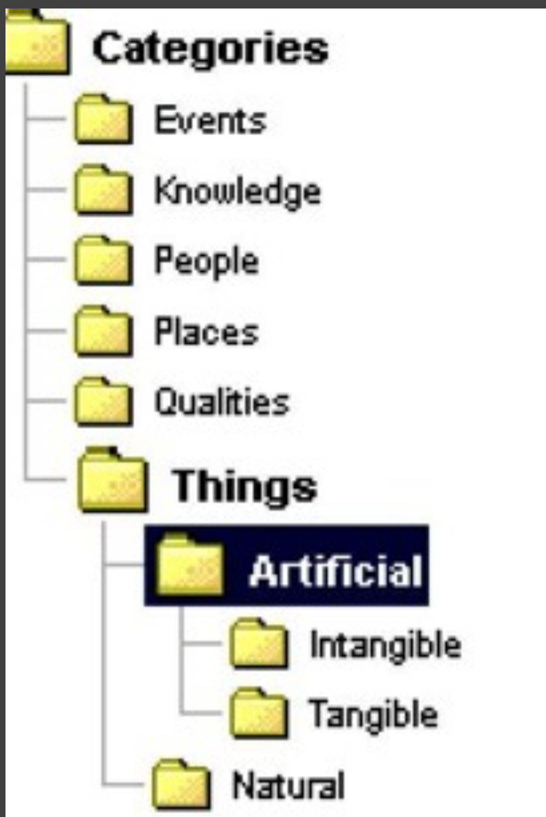
Subjects performed both retrieval and comparison tasks of varying complexity.

No significant performance differences were found across task conditions.

How do users navigate the tree?

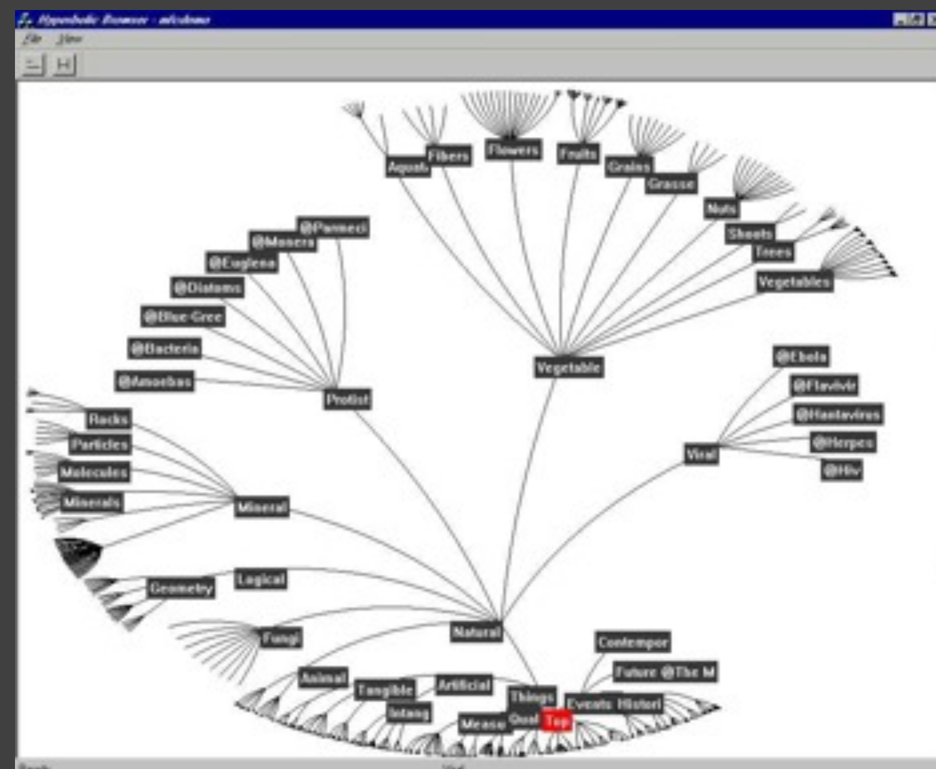
How do users navigate the tree?

They read the labels!



Microsoft File Explorer

vs.



Xerox PARC Hyperbolic Tree

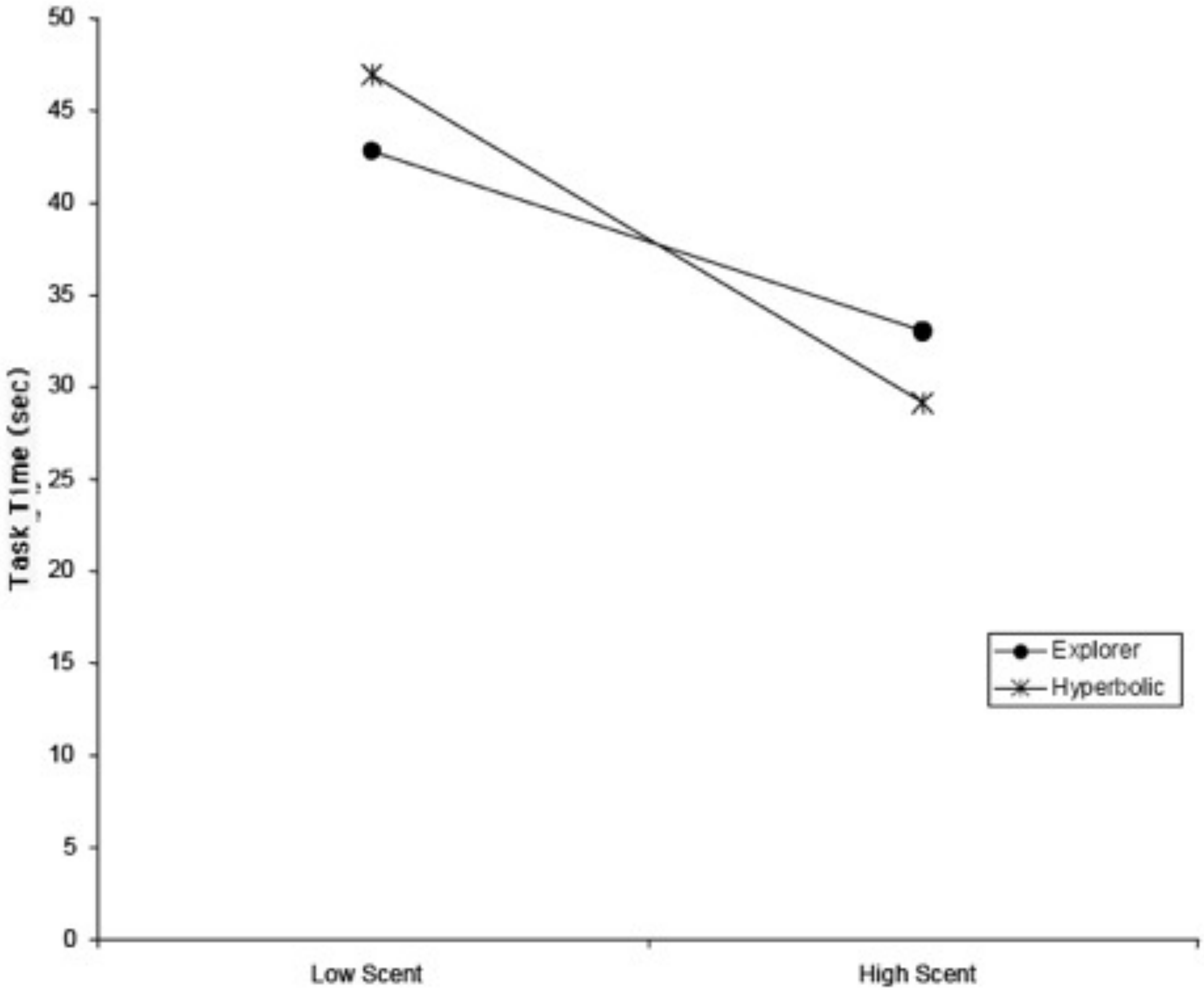
How do users navigate the tree?

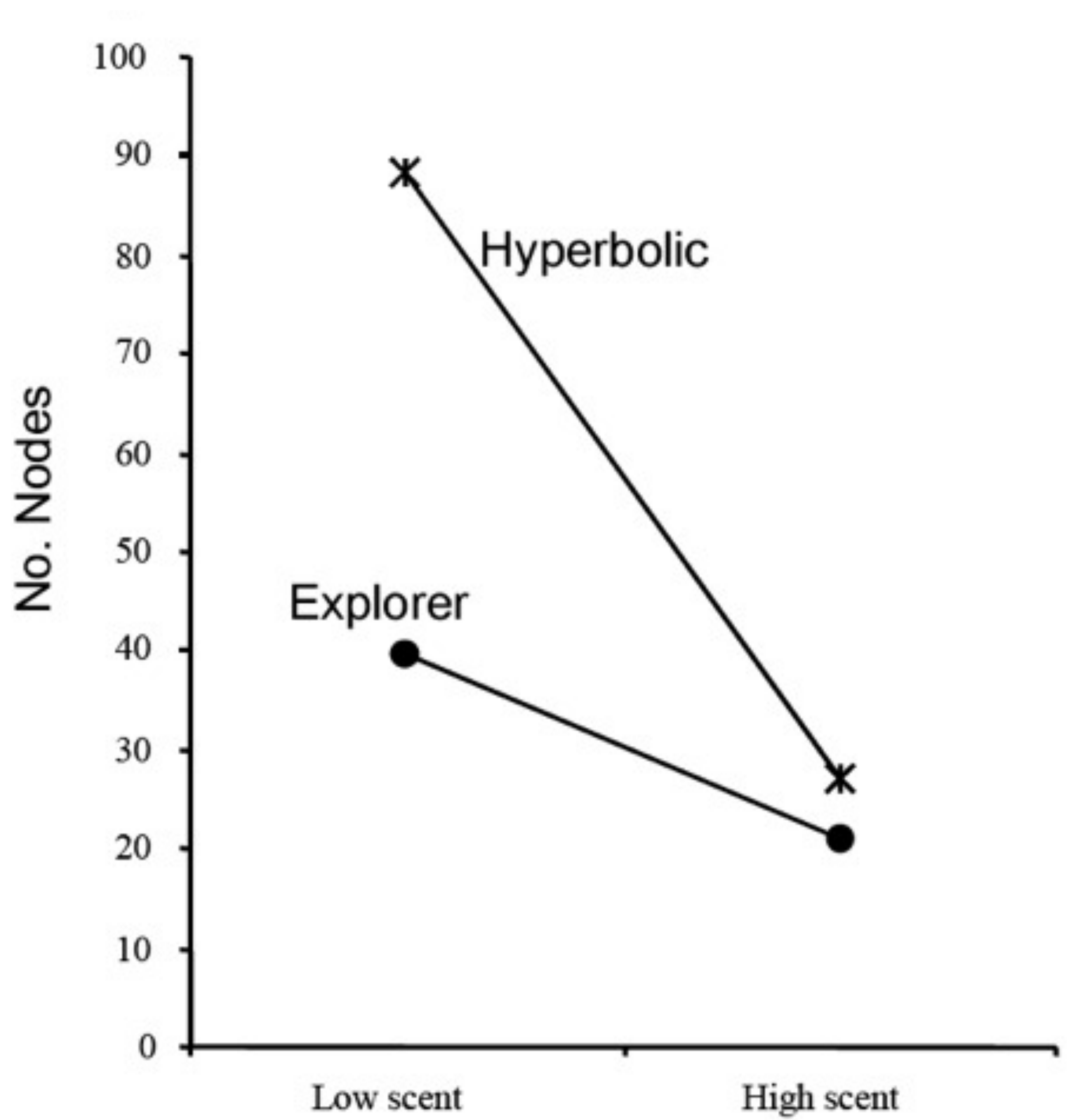
Information Scent: A user's (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues. [Pirolli & Card 99]

How do users navigate the tree?

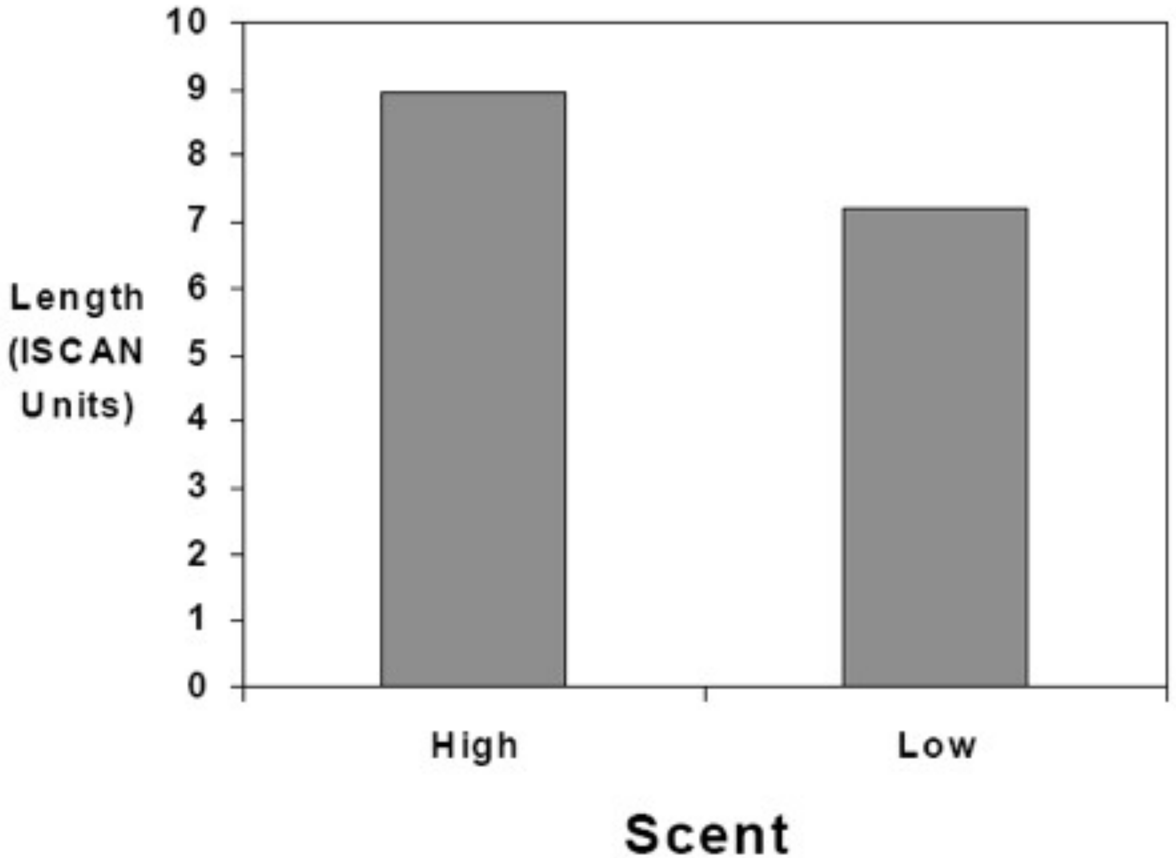
Information Scent: A user's (imperfect) perception of the value, cost, or access path of information sources obtained from proximal cues. [Pirolli & Card 99]

Operationalize as: the proportion of participants who correctly identified the location of the task answer from looking at upper branches in the tree.

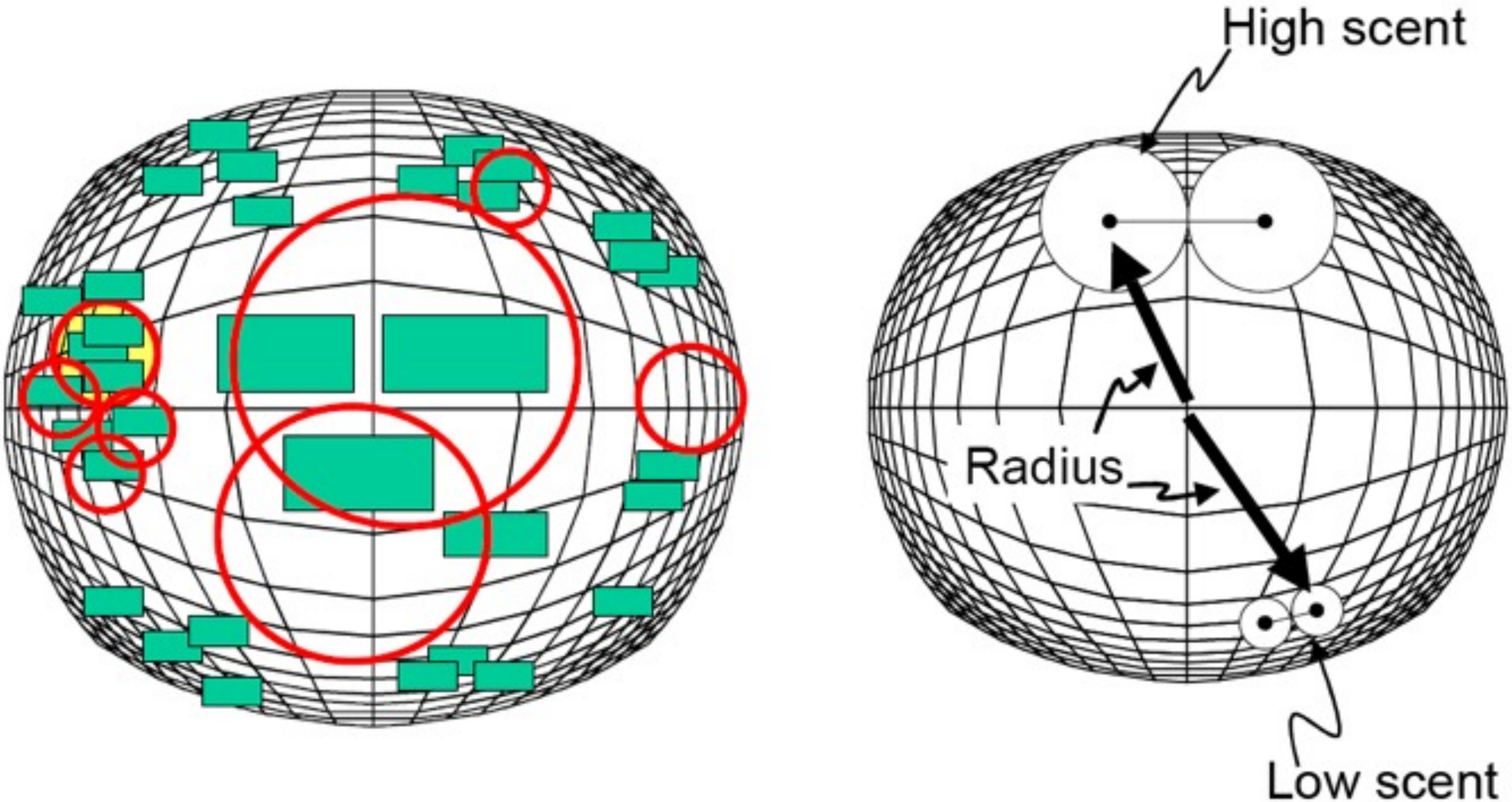

Task Time (sec)
50
45
40
35
30
25
20
15
10
5
0
Low Scent
High Scent
Explorer
Hyperbolic


No. Nodes
100
90
80
70
60
50
40
30
20
10
0
Hyperbolic
Explorer
Low scent
High scent

Length of Eye Movements



An Adaptive Field of View?

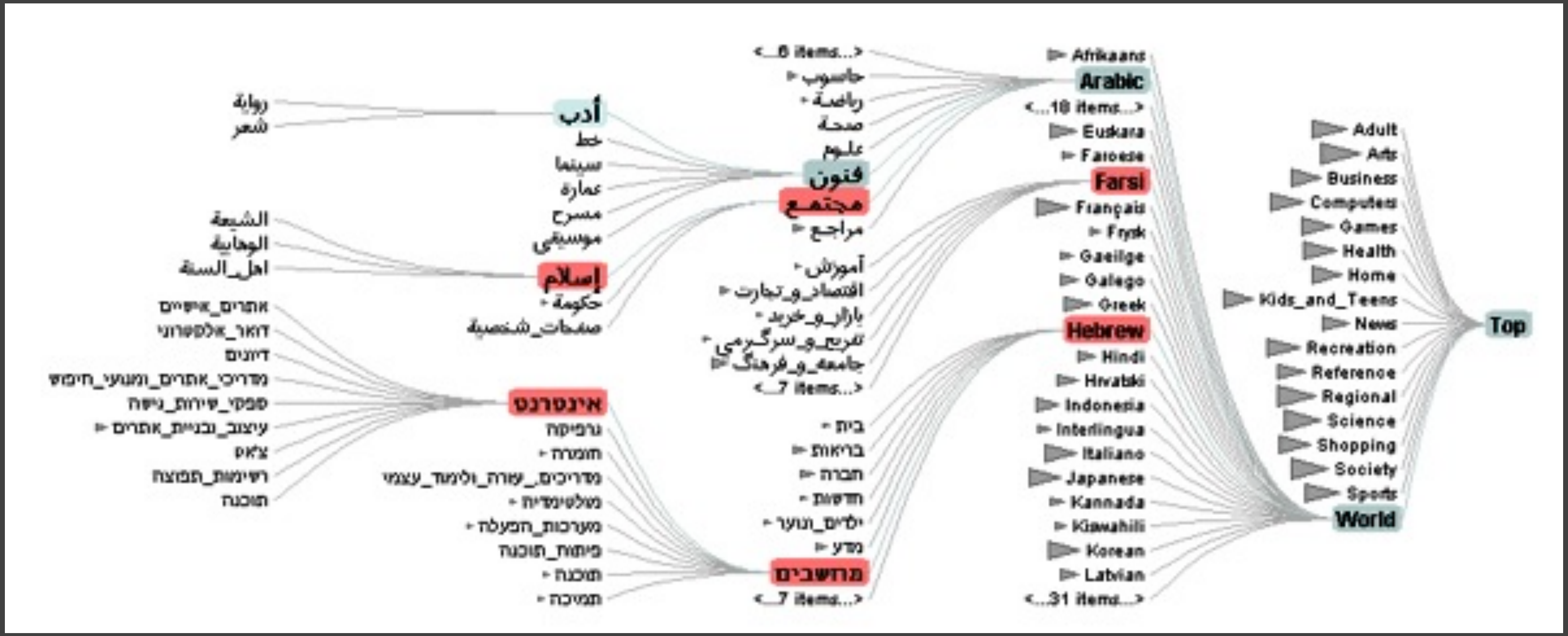


Design Guidelines

Design Guidelines

Support rapid visual scanning

Most people **don't read in circles!**



Degree of Interest Trees [Heer & Card 04]

Design Guidelines

People don't read in circles!

Showing more is not always better

- **Distractors** can decrease task performance
- Interaction with quality of **information scent**

Design Guidelines

- People don’t read in circles!
- Showing more is not always better
- **Navigation cues critical to search**
  - **Informative labels** or landmarks needed
  - Poor **information scent** undermines search

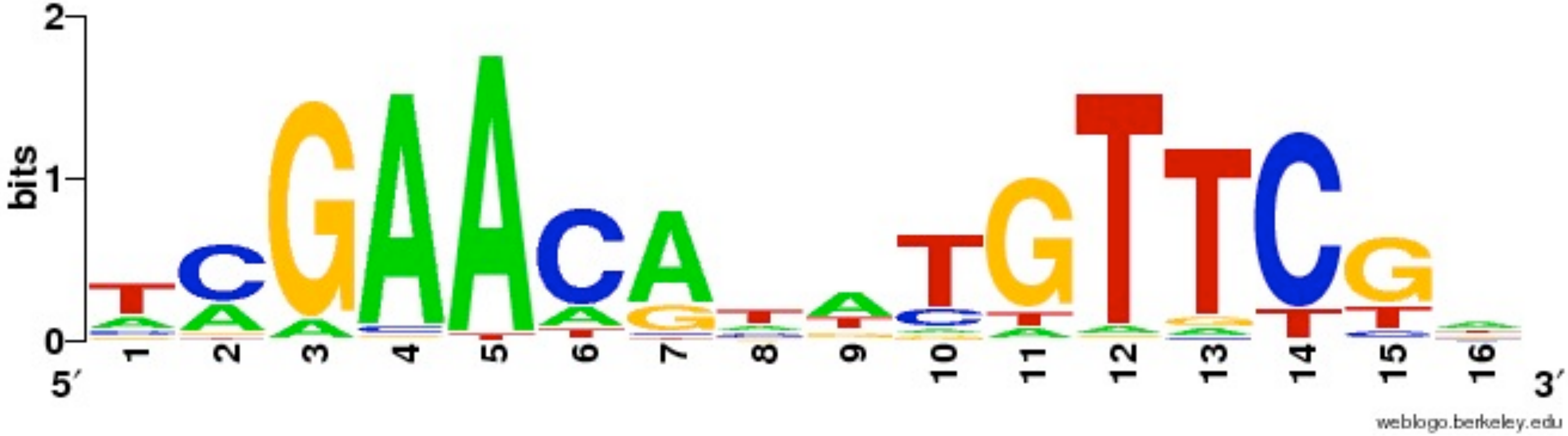
Lessons Learned

Both **task** and **data properties** (here, *information scent*) may interact with the visualization type in unexpected ways.

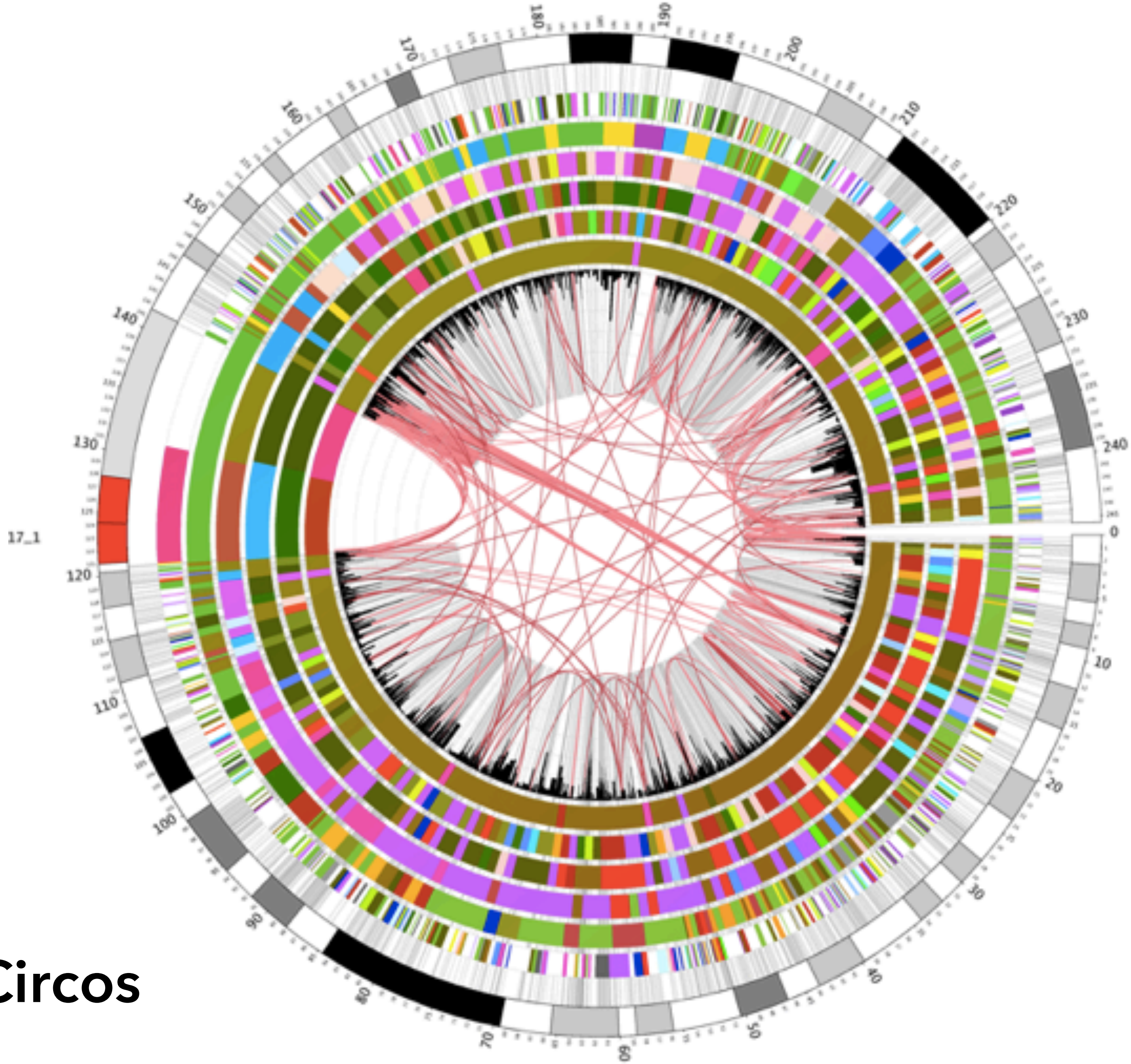
Equal **performance** in terms of accuracy or response time is **not the whole picture**. We often require more detailed study!

FOLLOW-UP QUESTION:

Which **bio-visualizations** should we evaluate?



Sequence Logos



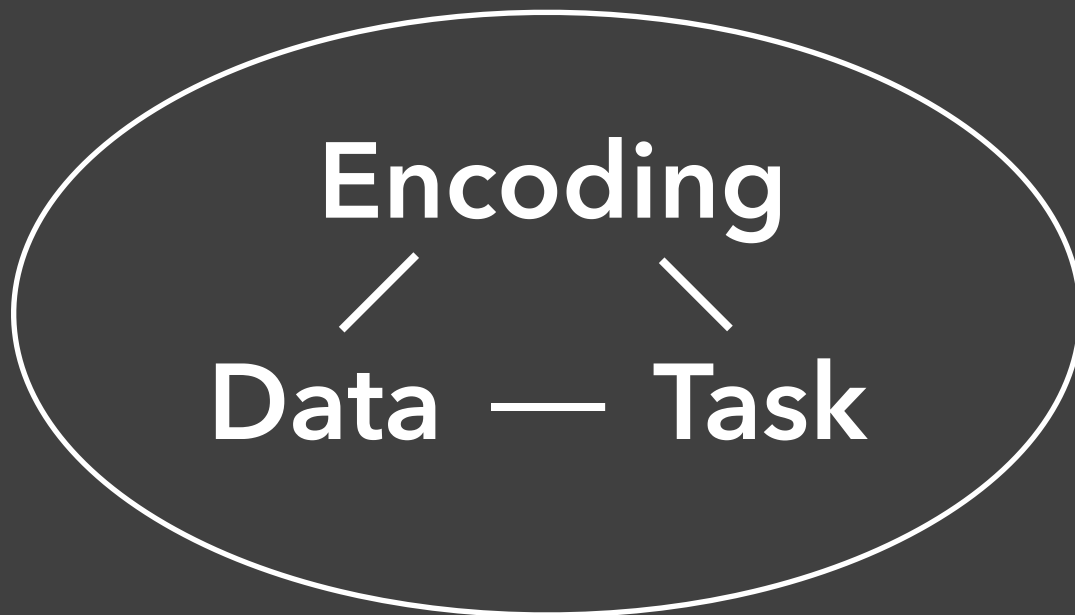
17_1
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240

Circos

Bio & Vis: Fellow Travelers

Biological data analysis and the study of visualization fundamentals should be **mutually reinforcing efforts**.

What **new principles** can we establish through the **design** and **evaluation** of biological data visualizations?



Users & Domain

Additional Resources

Perception for Design. Colin Ware.

How Maps Work. Alan MacEachren.

Graphical Perception. Cleveland & McGill.

A Nested Model for Visualization Design & Evaluation. Tamara Munzer.

Principles of Data Visualization

Jeffrey Heer @jeffrey_heer
http://idl.cs.washington.edu