Scalable Visualization and Model Building

Pat Hanrahan and John Gerth
Department of Computer Science
Stanford University

William S. Cleveland
Department of Statistics
Department of Computer Science
Purdue University

The Target Audience for the Results of Our FODAVA Research

1. Research in fundamentals and certain methods of data visualization
   - millions of people around the world
   - access the vast number of mathematical and visual methods of data analysis developed in statistics, machine learning, and many subject matter areas
   - use software systems such as R, Systat, Matlab, SPSS, etc. to analyze data
   - our research team members are in this community

Examples of Mathematical and Visual Learning Methods and Models ● linear regression ● support vector machines ● trellis display ● principal components dimension reduction ● multidimensional scaling ● hidden Markov models ● orthogonal centroid dimension reduction ● scatterplot matrix ● Bayes networks ● long-range dependent time series models ● locally weighted regression ● normal quantile plot ● ARIMA models ● fractional ARIMA models ● cox regression model ● linked views ● power spectrum analysis ● Bayesian hierarchical models ● dot plots ● multifractal wavelet models ● analysis of variance . . .

The Target Audience for the Results of Our FODAVA Research

2. Research in visual analytics for cybersecurity

Cybersecurity analysts

Two such analysts are working on our project

Final answer

- tools cybersecurity analysts will use
- involve data reduction and visualization of the reduced data

Our research in getting to the final answer

- fundamentals: figure out how to carry out monitoring and forensics
- using mathematical and visualization methods and models from the above list
- starts with a data reduction carried out very gingerly
- reduction validated by our visualization of the raw data

Our Research: Based on Important Principle of Data Analysis

First effectively argued in the 1960s

John Tukey, Frank Anscombe, Cuthbert Daniel, and others

Part I

Visual displays of data are essential for

- understanding the patterns in a dataset
- determining which mathematical learning methods and models are appropriate for the data

Using mathematical methods and models, without understanding the patterns, risks

- missing important information in the data
- incorrect conclusions

Our Research is Based on an Important Principle of Data Analysis

Part II

But, one cannot get far with just visualization of the raw data

Need the mathematical learning methods at the outset as well

Fit mathematical structures to aid in visualizing the patterns in the data

Our Research is Based on an Important Principle of Data Analysis

Mathematical methods of data analysis and visualization methods are symbiotic. Both should be applied from the moment the data arrive.

Anscombe's Quartet

Observations of a response y_i and an explanatory variable x_i
Carry out a linear regression analysis and look at output.

Number of observations: n = 11

Mean of the x_i: $\bar{x} = 9.0$

Mean of the y_i: $\bar{y} = 7.5$

Least squares regression coefficient of y on x: $b_1 = 0.5$

Equation of the least-squared regression line: $y = 3 + 0.5x$

Sum of squares of x_i about $\bar{x}$: 110.0

Regression sum of squares: = 27.5 (1 df)

Residual sum of squares: = 13.75 (9 df)

Estimated standard error of b_1: 0.118

Multiple R^2: 0.667

What does the output of this mathematical modeling tell us about the data?

Anscombe's Quartet

11 observations of a response y_i and an explanatory variable x_i.
Carry out a linear regression analysis a look at output.

Number of observations: n = 11

Mean of the x_i: $\bar{x} = 9.0$

Mean of the y_i: $\bar{y} = 7.5$

Least squares regression coefficient of y on x: $b_1 = 0.5$

Equation of the least-squared regression line: $y = 3 + 0.5x$

Sum of squares of x_i about $\bar{x}$: 110.0

Regression sum of squares: 27.5 (1 df)

Residual sum of squares: 13.75 (9 df)

Estimated standard error of b_1: 0.118

Multiple R^2: 0.667

Without additional information about the data, the results tell us little

The Anscombe Quartet of Data Sets

Four data sets that have the values below.

Number of observations: n = 11

Mean of the x_i: $\bar{x} = 9.0$

Mean of the y_i: $\bar{y} = 7.5$

Least squares regression coefficient of y on x: $b_1 = 0.5$

Equation of the least-squared regression line: $y = 3 + 0.5x$

Sum of squares of x_i about $\bar{x}$: 110.0

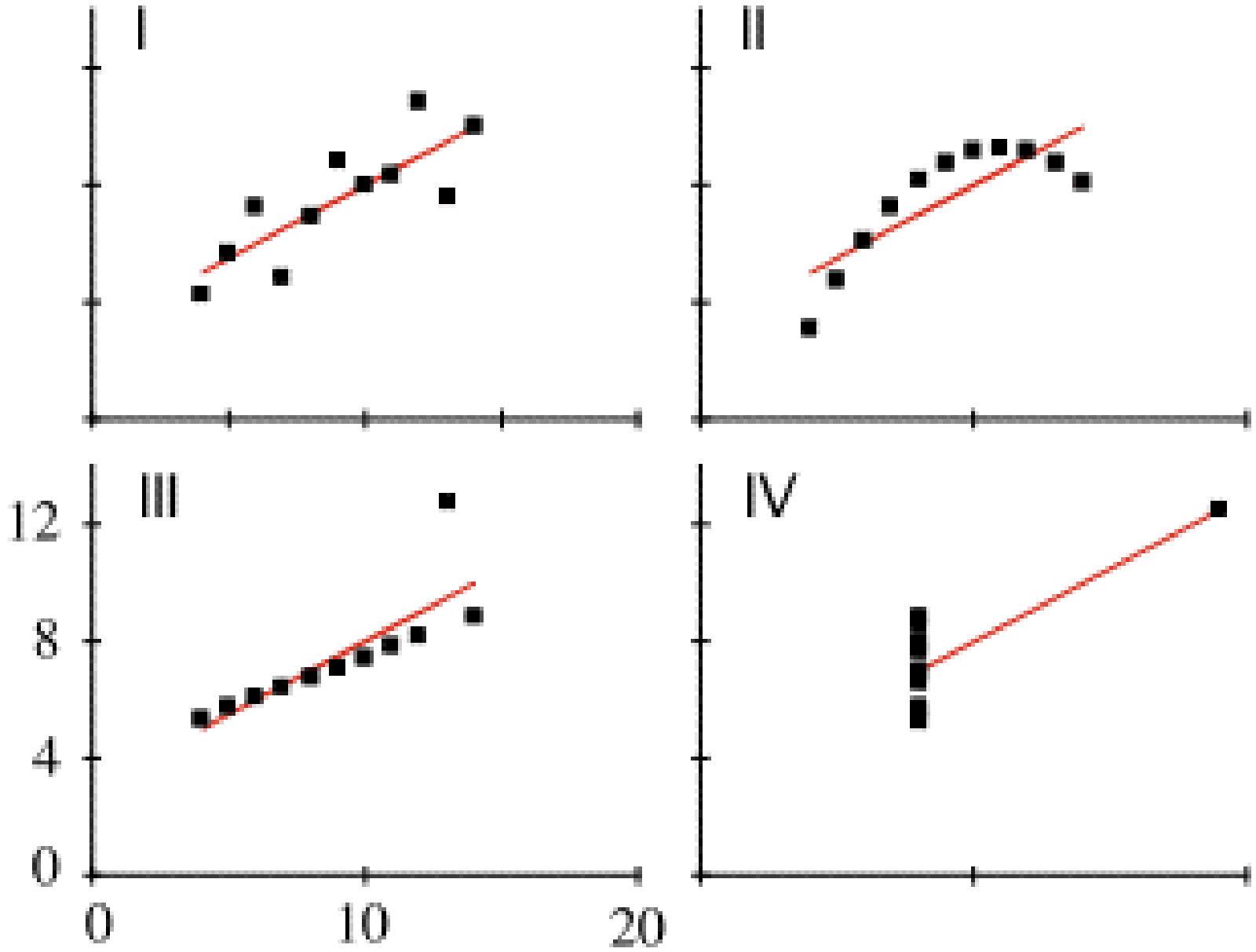
Regression sum of squares: 27.5 (1 df)

Residual sum of squares: 13.75 (9 df)

Estimated standard error of b_1: 0.118

Multiple R^2: 0.667

Anscombe's Quartet



A Common Objection to The Principle in the 1960s

Not consistent with “objective science”

Must bring hypotheses to the data determined a priori

Peeking at the data biases final results

The Common Objection of the 1960s

Dissipated as example after example showed the substantial inadequacies of analyses without an understanding of the patterns in the data

Some had to learn the hard way by having their data publicly re-analyzed

40 Years Later, in the 2000s: Very Large, Complex Datasets

The data analysis principle does not suddenly become false because the datasets have become very large

But the challenges to using visualization and mathematical methods to characterize the patterns in the data are substantial

A major component of our FODAVA research is directed at developing approaches, methods, and systems for visualizing large complex datasets

Cybersecurity

Analysis of large, complex Internet traffic datasets for cybersecurity command and control

Initial target: monitoring and forensics for enterprise networks

Cybersecurity: One Research Project

A Rules-Based Statistical Algorithm for Keystroke Detection (Streaming)

Detect TCP connections with keystroke packets under SSH

- detections go into cybersecurity database for further analysis
- add an asterisk when detected connection at a port where SSH is not known to be running

Information for detection: packet-level information for each connection from a monitor on a network link

- packet arrival timestamp at monitor
- TCP headers: source & destination port numbers, sequence numbers, and flags (SYN, FIN, ACK)
- IP headers: source & destination IP addresses (anonymized)
- no packet payload

First dataset (a trickle compared to latter collections that will occur)

- monitor on subnet of Purdue Statistics Department for 72 hours
- 600,000 connections
- 500,000,000 packets

Our General Approach for Analysis of Large, Complex Datasets

Divide and Conquer

Partition the data into subsets in one or more ways

Each subset is a small dataset (can apply all methods to it)

Sample the subsets of a partition (there can be many different sampling frames)

Apply mathematical methods or visualization methods or both to each subset of the sample

Typically, apply mathematical methods to much larger samples than the those for the visualization methods

For the cybersecurity data, each connection is a subset

This is a form of data reduction

- but we get to see the raw data directly
- can readily bring subject matter knowledge to bear on the reduction
- can oversample, and study sequentially until we are convinced of a pattern

One Example from Our Keystroke Algorithm

Shows packet classification at step 5 of the algorithm

RHIPE

Divide and Conquer: allows embarrassingly parallel computation

RHIPE (ml.stat.purdue.edu/rhipe)

- Saptarshi Guha, graduate student, Purdue Statistics
- R-Hadoop Integrated Processing Environment
- Greek for “in a moment”
- pronounced “hree pay”
- open source

A recent merging of

- the R interactive environment for data analysis (www.R-project.org)
- the Hadoop distributed file system and compute engine (hadoop.apache.org)

With partitioning, allows analysis of large, complex datasets

- all commands from within R (great saving of user time)
- RHIPE handles all reads and writes to the Hadoop distributed file system
- has taken us from the infeasible to the feasible for many datasets

Being picked up by organizations with large complex datasets to analyze: e.g., DoCoMo Research Labs, Lehman Brothers, Lawrence Livermore National Labs

Subset-Sample Visualization: Large Displays and Visualization Databases

Can get much insight by sampling partitions and applying a visualization method to each subset, even when the number of subsets in the sample is large

A single large display

- can have a large number of pages, each of which can have many panels
- each panel shows the visualization method for one subset
- total number of pages might be measured in hundreds for even a few thousand
- query on an as-needed basis

We create a visualization database: collection of displays, often with many displays that are large

Visualization Databases

A major component of our FODAVA research is a collection of topics to improve the effectiveness of visualization databases

Visualization Databases

A major component of our FODAVA research is a collection of topics to improve the effectiveness of visualization databases

Substantial success with just off-the-shelf methods

Still, we can do much better

Topic 1: Graph Design for Gestalt Formation to Enable Rapid Scanning

Gestalt

- pattern that forms effortlessly on the order of 10s of ms
- effect hits you between the eyes

Critical to data visualization generally

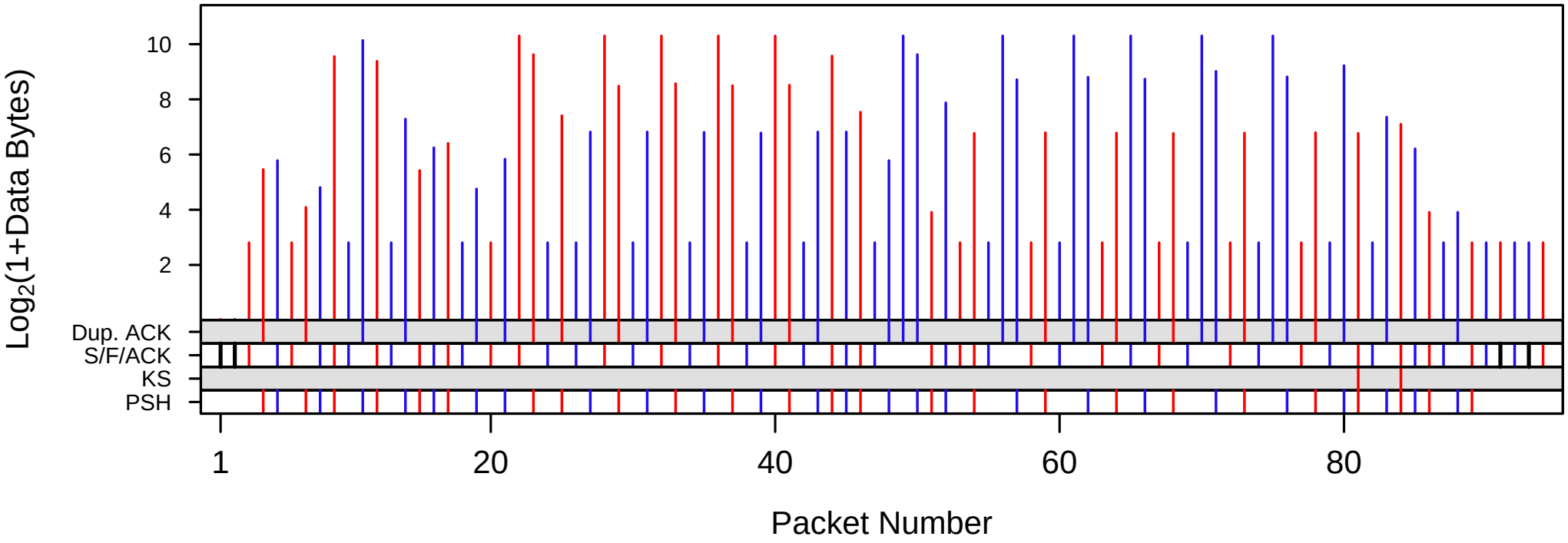
For large displays, enables rapid scanning

For many display methods, happens readily

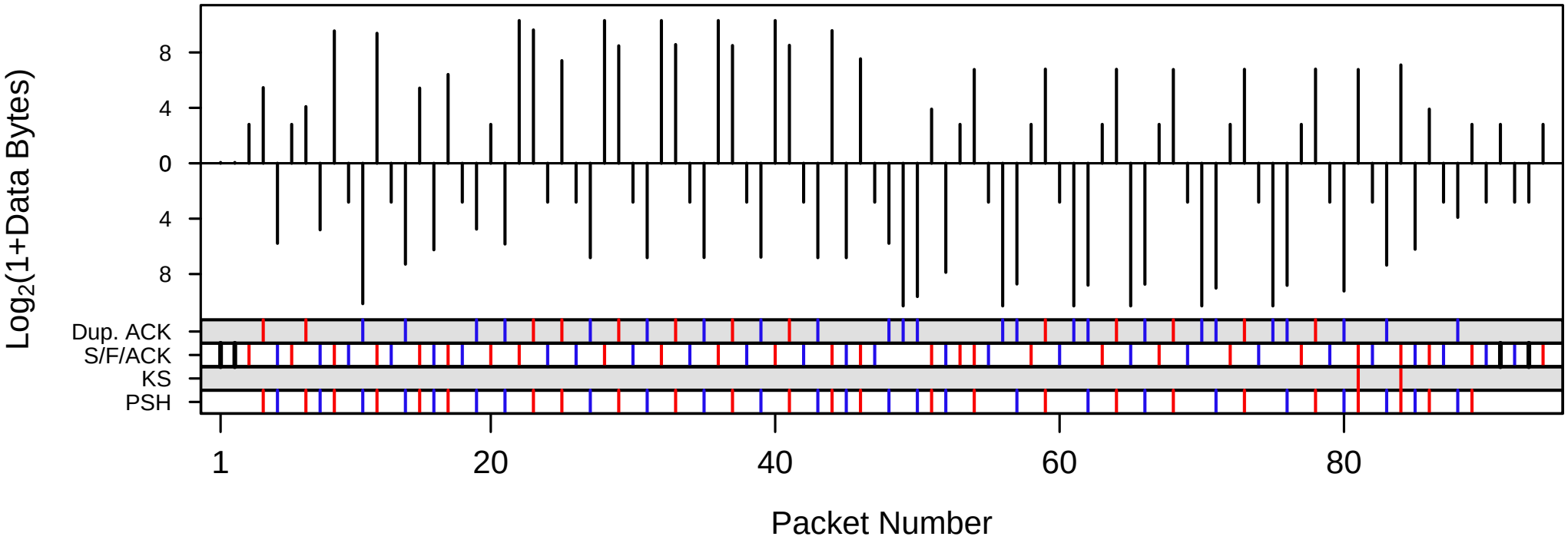
But for displays of complex data structures, must experiment with different designs of displays to effect this

Develop a large set of case studies, and seek to explain results using theory of gestalt psychology

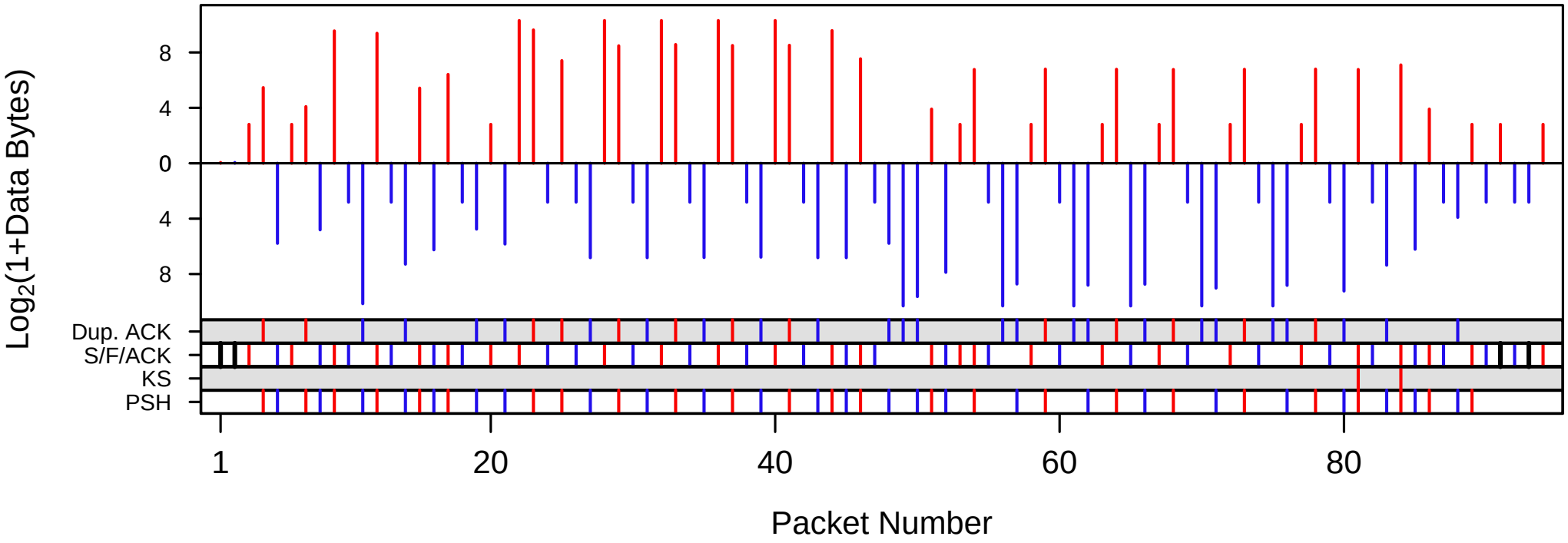
Gestalt Formation: Experiment 1: Color



Gestalt Formation: Experiment 2: Juxtaposition



Gestalt Formation: Experiment 3: Color and Juxtaposition



Topic 2: Sampling Partition Subsets

Representative sampling: in the spirit of statistical design of experiments and survey sampling

View-selection algorithms (a branch of Tukey's "cognostics"): algorithms that find "interesting" subsets

Topic 3: Automation Algorithms for Display Rendering

In making large displays we do not want to fuss over drawing details

- seek algorithms that make automated choices

Tick marks

Sizes of plotting symbols

Aspect ratio

Choosing scales across panels, white-space analysis

- 1. Different from one panel to the next to fill the data regions
- 2. Have the same number of units/cm, but otherwise vary scales to fill the data regions
- 3. The same

Topic 4: Large Display Viewer Design

Display resolution

- as a practical matter design so that each page of a large display resolves for 1024x768 pixels

Design for gestalt formation allows us to rapidly click through the pages with one-page in the visual field at a time

n pages on 3 adjacent large monitors

- e.g., each screen 2560x1050 pixels
- can put many pages in the visual field
- can view pages faster by a purely visual scan than by a click scan, and even make more effective comparisons

Designing viewers for multiple large monitors

Viewers can show multiple pages on two display devices, resulting in quite large values of n, a very big enhancement

This is big win

Topic 5: Interactive Modeling and Visual Analytics

A framework for the integration of visual analysis and statistical modeling for large datasets

Envision a system that facilities an iterative modeling process

The modeling cycle includes multiple stages

- descriptive visualization
- model selection
- model fitting
- diagnosis and evaluation
- iteration

Specific goal: build a highly interactive graphical environment that supports this process

A Rules-Based Statistical Algorithm for Keystroke Detection

Past approaches

- analyze summary statistics of the packet-level information
- a data reduction method at the outset
- classification task C: a connection has keystrokes or not

Our approach

- analyze packet-level data
- classification task P: a client packet has a keystroke or not
- classification task C: a connection has key partitioning: each connection is a subset

Outcome

- we have much higher accuracy in task C
- at the detailed packet level keystrokes have very distinctive patterns
- discovered this using data visualization tools