Data Science

Its intellectual areas are all those that come into play in the analysis of data.

If a department has people who cover the areas, then the teaching will follow.

Divide & Recombine (D&R) for Large Complex Data

Data science is needed.

Our research group must cover its intellectual areas to succeed.

D&R: Two Goals

1. Deep Analysis

Comprehensive detailed analysis

Does not lose important information in the data

Achieving deep analysis for large complex data and small data requires both
- visualization analytic methods (output is visual)
- statistical (includes machine learning) analytic methods (output is numeric and categorical)

Visualization must be of
- detailed data at their finest granularity
- summary statistics

D&R: Two Goals

2. Interactive Language for Data Analysis (ILDA)

The data analyst and the methodologist developing methods for large data works within an ILDA

A well designed ILDA provides very time-efficient programming with the data and prototyping new methods and models

The ILDA must provide access to the 1000s of current methods of statistics, machine learning, and visualization

An example of such an ILDA is R

- provides very high efficiency of programming for the user
- as of recently: “Currently, the CRAN package repository features 4097 available packages.”

D&R Division

Parallelize the data

- division method
- divide the data into subsets

Visualization and statistical analytic methods are applied to each subset

- independently: no communication between the computations
- embarrassingly parallel

D&R Recombination

For each analytic method applied

- recombination method
- outputs across subsets are recombined

Visualization analytic methods

- to see data at finest granularity, applied to a sample of subsets based on a statistical sampling plan
- this is a data reduction, but a rigorous one
- recombination is a multipanel display, one panel per subset

Statistical analytic methods

- almost always applied to all subsets
- D&R result is the recombination
- replacement for the result that would have come from applying the method directly to all of the data

The Intellectual Areas

What is necessary to make D&R succeed?

Must work in all of the intellectual areas that come into play in the analysis of data because large, complex data challenge them all

i.e., Data Science

We know this because members of our research group analyze large complex data

- packet-level network data for situational awareness and attack attribution
- power grid data
- high-energy physics data
- Firefox longitudinal activity time series

i.e., Data Science includes data analysis because data analysis is one of the things that comes into play in data analysis

D&R for Visualization Analytic Methods

Sampling of subsets based on principles of statistical sampling results in rigorous data reduction

Visualization databases

Display design for rapid scan of large displays, based on principles of visual perception, and empirical study

D&R Research in Statistical Theory for Statistical Analytics Methods

The D&R result is almost always not the same as the direct all-data result

The division and recombination methods used for a statistical analytic method have a big impact on the accuracy

Best D&R methods: maximize statistical accuracy given data must be divided

Theoretical work

- seeks best D&R division methods
- determine how close best methods are to accuracy of direct, all-data result
- results so far are very promising

There is a lot of work to do because there are many statistical analytic methods

What About Parallel Algorithms for Statistical Analytic Methods?

There are parallel algorithms that seek the direct all-data result of a statistical analytic method

Compute on subsets like D&R

They must iterate, which slows things down, and results are not necessarily exact

Communications must communicate across subsets

D&R provides very fast embarrassingly parallel computation: no communication among states

If we can get high statistical accuracy, there is no downside

Many statistical analytic methods do not submit readily to parallel subset algorithms

D&R is ready to go; we do not have to rewrite the 4097 packages in R

What is D&R?

D&R is fundamentally

a statistical approach

to large complex data

that is able to exploit

parallel distributed

computational environments.

Why D&R Works

D&R leads to embarrassingly parallel computation

- the simplest parallel computation
- no communication among subset computations

Can exploit the distributed computational environment Hadoop

Can use cluster with heterogeneous nodes

Can program in R by merging R and Hadoop

Hadoop Distributed Computational Environment

Distributed file system (HDFS)

Distributed compute engine (Mapreduce)

Supported by Apache Software Foundation

Open source (Apache License)

History

- Google: released the architecture in a paper
- Yahoo: first public domain release of software
- Apache Software Foundation: signed on to develop and distribute

RHIPE: R and Hadoop Integrated Programming Environment

[hree-**pay**]: “in a moment” in Greek

First developed by Saptarshi Guha, formerly a Purdue Stat grad student, now at Mozilla

Written in C, Java, and R

Open source software, Apache license

It is an R package

Google discussion group: groups.google.com/group/rhipe

Code development site: Github

RHIPE EC2 available

See datadr.org for info on all D&R aspects: statistical and computational

RHIPE: What the R Analyst Does and Does Not Do

The analyst specifies analysis procedures wholly in R

- the division into subsets, each typically a single R object
- the analytic methods to apply to each subset
- the recombination

It’s like an analysis of one subset but just pushed out to all

Almost any R function or package can be applied to the data

RHIPE R commands enable communication with Hadoop

Hadoop spreads subsets of division across cluster

Hadoop manages all of the distributed parallel computation

Cluster Design and Setup for RHIPE

Configuration of hardware components, linux, Hadoop, and RHIPE

Seek optimization of performance

Guided by performance measurement

Redesign of R for Computational Efficiency

Interpreter

Vector computation

Saptarshi Guha, Mozilla (Formerly Purdue Stat grad student)

Ryan Hafen, Statistics, PNNL Stat (Formerly Purdue Stat grad student)

Pat Hanrahan, Stanford CS

Bill Cleveland, Purdue Statistics

Chuanhai Li, Purdue Statistics

John Gerth, Stanford CS

Jeff Heer, U. of Washington CS

Justin Talbot, Tableau (Formerly Stanford CS grad student)

Bowei Xi, Purdue Statistics

Tonglin Zhang, Purdue Statistics

Current Graduate Students, Purdue Statistics and Stanford CS

See datadr.org for more info.

Divide & Recombine for Detailed Visualization of Large Complex Data at their Finest Granularity

D&R Recombination

For each analytic method applied

- recombination method
- outputs across subsets are recombined

Visualization analytic methods

- to see data at their finest granularity, applied to a sample of subsets based on a statistical sampling plan
- this is a data reduction, but a rigorous one
- recombination is a multipanel display, one panel per subset

Statistical analytic methods

- almost always applied to all subsets
- D&R result is the recombination
- replacement for the result that would have come from applying the method directly to all of the data

One D&R Goal: Deep Analysis

Comprehensive detailed analysis

Does not lose important information in the data

Achieving deep analysis for large complex data and small data requires both

- visualization analytic methods (output is visual)
- statistical (includes machine learning) analytic methods (output is numeric and categorical)

Visualization must be of

- detailed data at their finest granularity
- summary statistics

D&R Visualization Framework

Apply a visualization method to each subset

Subsets have the detailed data at their finest granularity

Make a multi-panel, multi-page display

One panel per subset

But there can be millions of subsets

We must sample

Sampling Subsets

We can compute between-subset variables (BSVs)

Each BSV has one variable per subset

BSVs used to devise a sampling plan

The sampling is a data reduction

Sampling plans
- make the data reduction rigorous
- go a long way to help loss of information
- use the same ideas as sample survey design

Sampling Methods

Representative

Samples chosen to cover the joint region of values of a set of BSVs

Focused

Samples explore a particular sub-region of interest of the BSVs

Cognostic

A general notion of Tukey

Tailored to D&R

BSVs are developed that search for certain kinds of statistical behavior in a subset

Visualization Databases

We sample, but want to view as many subsets as possible

Large displays: it is possible to view 100s or even 1000s of panels

In an analysis of large complex data, there can be many such displays

This is a visualization database

Rapid-Scan Displays

To enable viewing many panels

Rapid scan displays enable viewing displays with a large number of subsets

Rapidly scan the panels of a large D&R visual display
- a controlled animation
- punctuated with thinking about the data

Requirement stated in 3 ways:

The display design must result in gestalt formation

Assessed patterns must form effortlessly (10s of ms)

Effect “hits you between the eyes”

Display designs that require attentive search to see patterns slow down scanning

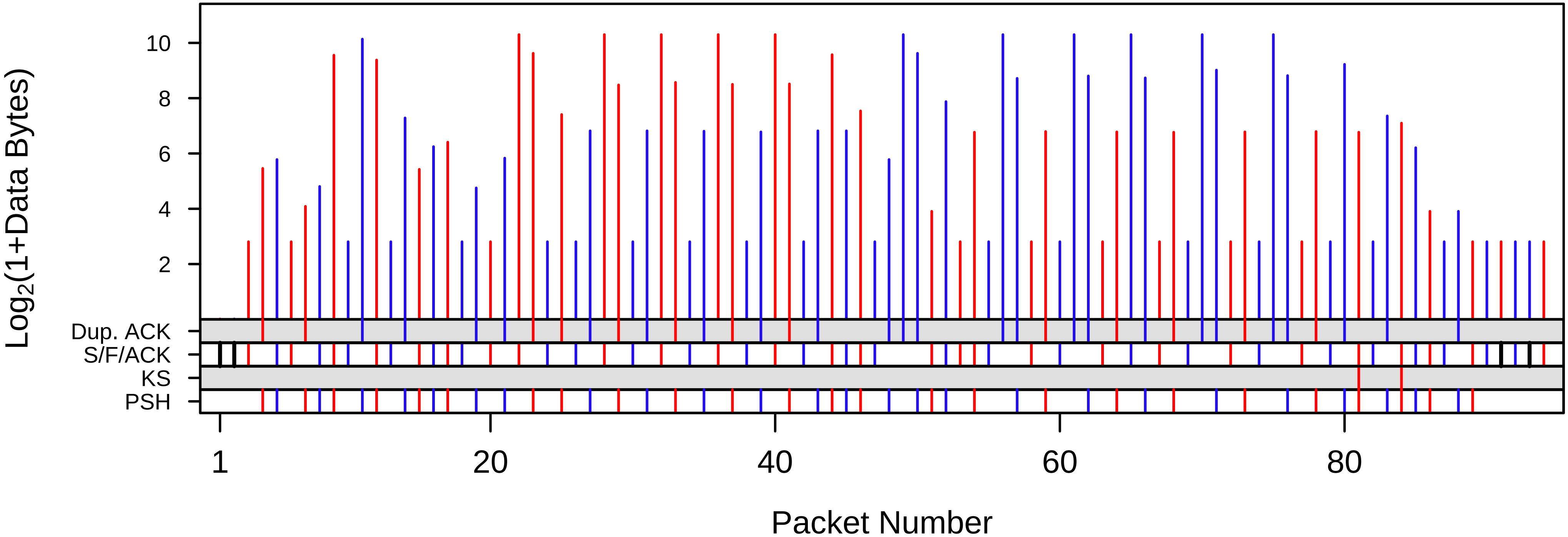
For many display methods, happens readily

For other display methods, must experiment with different designs to enhance gestalt formation

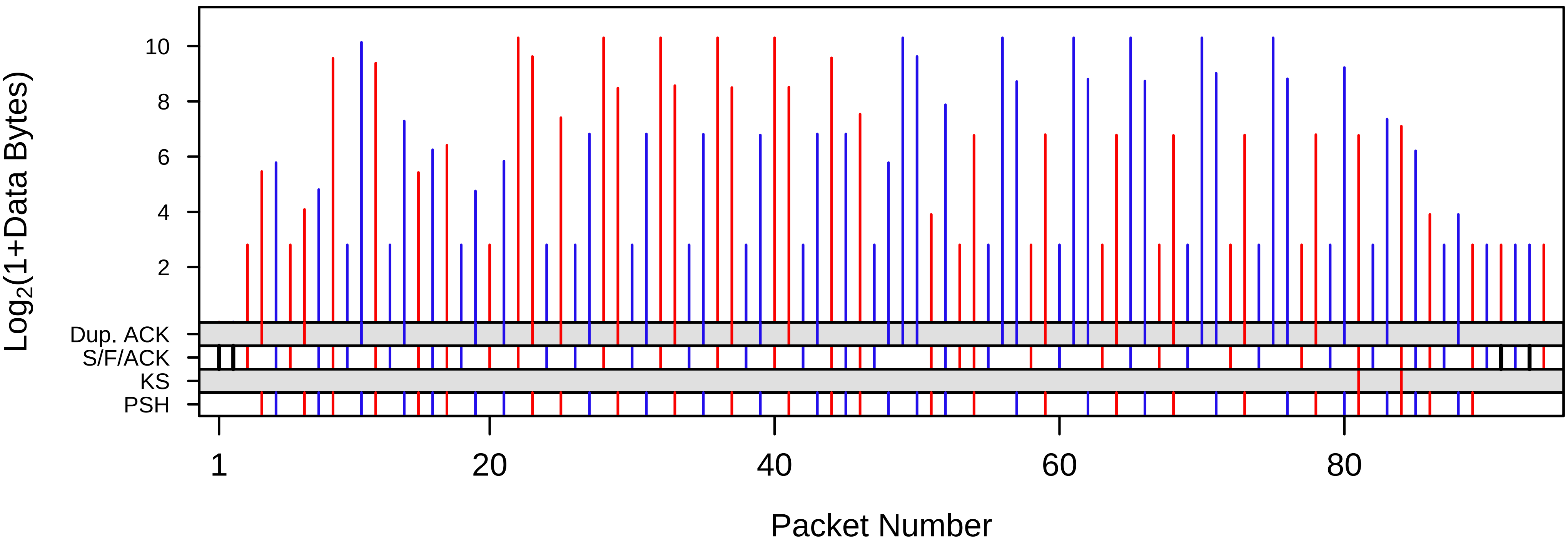
Issue of Gestalt Formation

Perceive time order of all packets, judging runs of packets in each group and the alternating patterns

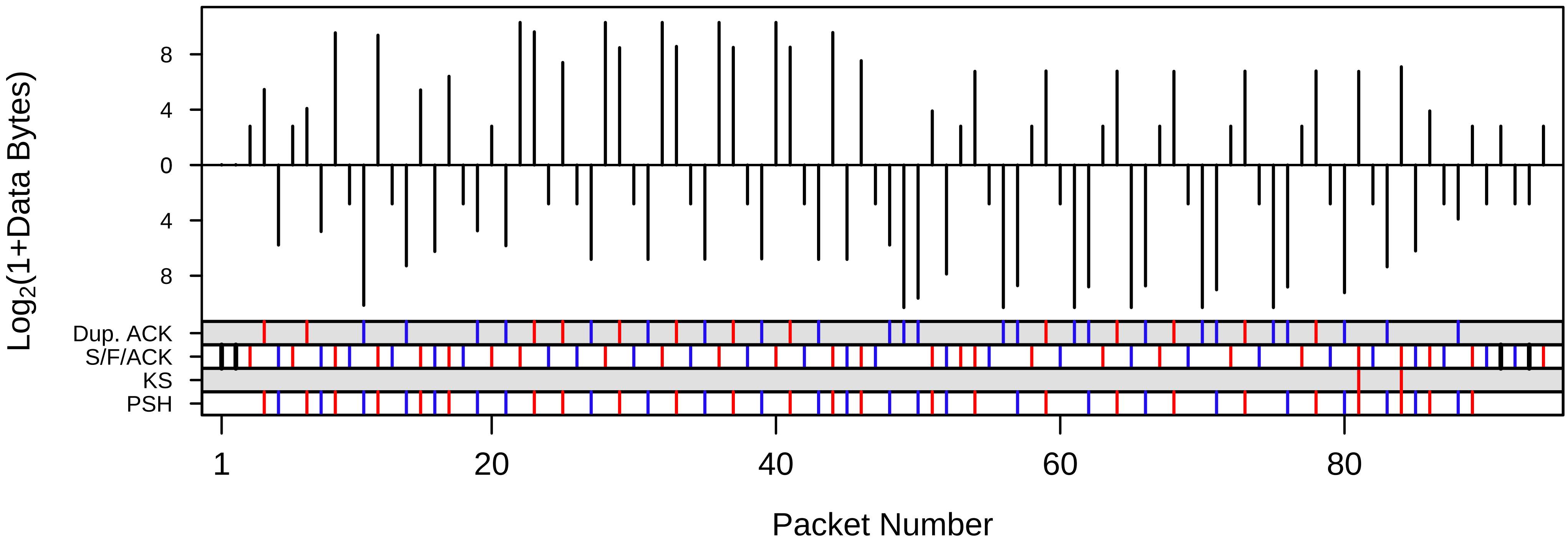
Judge relative lengths of line segments



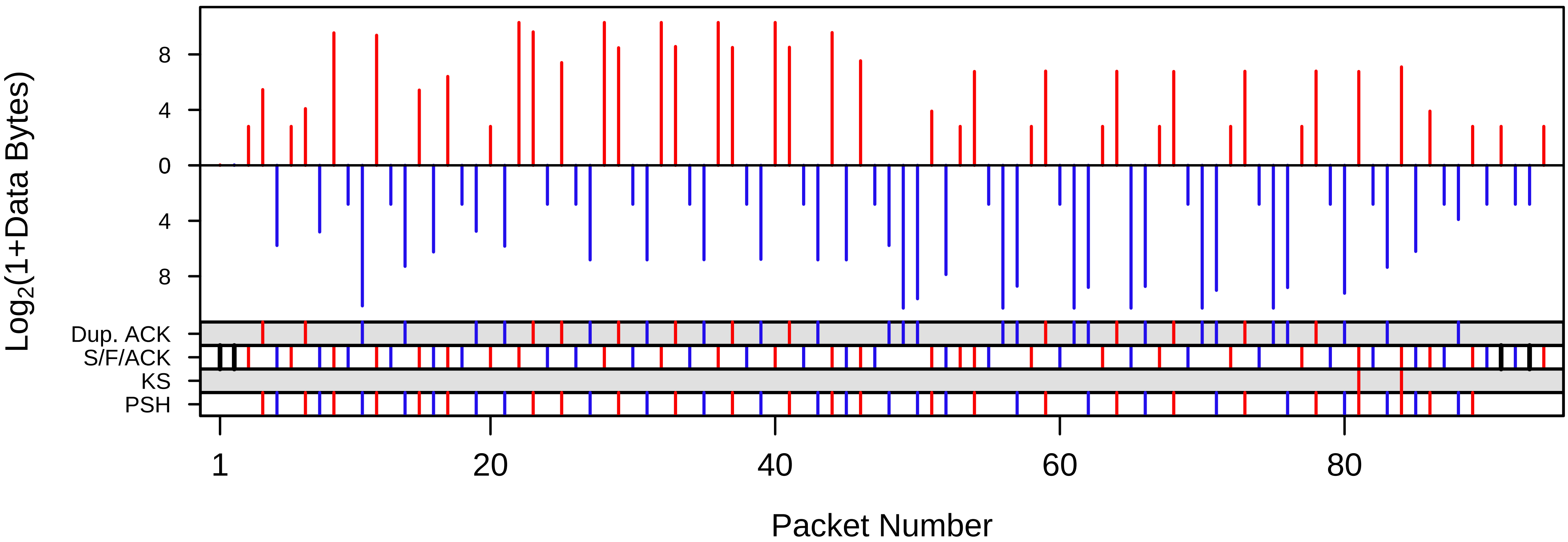
Gestalt Formation: Experiment 1 — Superposition and Color



Gestalt Formation: Experiment 2 — Juxtaposition and No Color



Gestalt Formation: Experiment 3: Juxtaposition and Color



VoIP

Semi-call: one direction of a call

Transmitted across the Internet in packets of 200 bytes with 20 ms of speech

Gateway to the Internet

- accumulates speech signal for 20 ms periods
- puts each speech period in a packet and sends it on its way
- intervals between packets 20 ms

VoIP Jitter

20 ms intervals perturbed as packets move from one router to the next

Jitter for two successive packets

- at a router on their path
- interarrival times at router - 20 ms
- jitter must not get too big

Silence suppression

- gateway detects drop in signal below a threshold
- stops sending packets
- creates alternating transmission and silence intervals
- jitter measured within a transmission interval

VoIP Dataset

Jitter measured on a router in New Jersey for 48 hr

1.237 billion values from 277540 semi-calls

27 sending gateways around the world

One division by transmission interval

14,161,628 subsets

Detailed data for each subset: J_i, the jitter and i, the is arrival order

Subset Sampling

Visualization method: plot J_i vs. i

A first look at the data

BSVs for sampling

- sending gateway, G
- arrival time of first packet of interval, T

Sampling plan: for each G, 48 intervals with values of T as equally spaced as possible

$27 \times 48 = 1296$ plots

Theoretical Investigations of D&R Division and Recombination Methods for Statistical Analytic Methods

The D&R result is almost always not exactly the same as the direct all-data result

Seek “best D&R” division methods and recombination methods

Best D&R: maximize statistical accuracy given data must be divided

The division and recombination methods used for a statistical analytic method have a big impact on the accuracy

Study how close best D&R result is to accuracy of direct, all-data result

An Example: Logistic Regression

Response:

$y_i, i = 1 \ldots n$: values of 0's and 1's

p Explanatory variables:

$X = [x_{ij}]$ is $n \times p$ matrix

Columns are n measurements of p variables

Model:

π_i = Pr$\left(y_i = 1\right)$

$\ell_i = \log(\pi_i/(1-\pi_i)) = \sum_{j=1}^{p} \beta_j x_{ij}$

Unknown parameters are $\beta = \left(\beta_1, ..., \beta_p\right)$

An Example: Logistic Regression

Maximum likelihood:

$\hat{\beta}$

Standard method of estimation of β

D&R estimation:

Suppose $n = rm$

Each subset is m rows of X and the corresponding values of Y

There are r subsets

r subset estimates are subset mle's: $\hat{\beta}_k, k = 1, \ldots, r$

One possible D&R recombination estimate of is the vector mean, $\ddot{\beta}$ of the $\hat{\beta}_k$

An Example: Logistic Regression

There are parallel algorithms that seek the direct all-data estimate

Compute on subsets like D&R

They must iterate, which slows things down, and results are not necessarily exact

D&R provides very fast embarrassingly parallel computation, and if we can get high statistical accuracy, there is no down side

Statistical Accuracy

Suppose we have two estimators for LR: $\dot{\beta}$ and $\ddot{\beta}$

Let $w_i, i = 1, \ldots, p$ be values of explanatory variables

Compare ratio of variances of $\sum_{i=1}^{p} \dot{\beta}_k w_k$ and $\sum_{i=1}^{p} \ddot{\beta}_k w_k$

General Concepts for High Statistical Accuracy

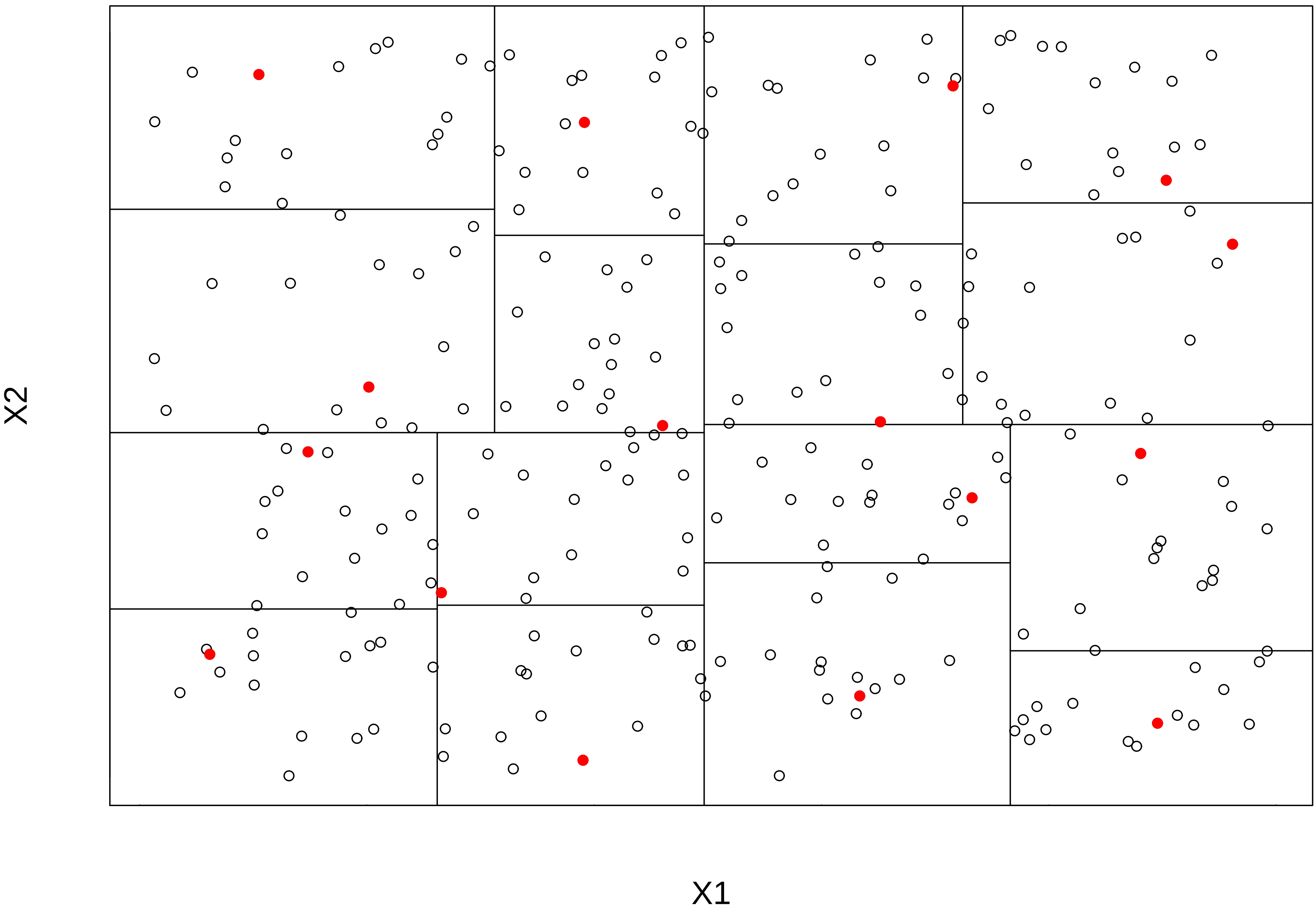
Division

Exact, near-exact, and random replicate division

Recombination

Covariance weighted recombination

Exact, Near-Exact and Random Replicate Division Methods



Covariance Weighted Estimators

D&R subset estimates $\hat{\beta}_k, k = 1, \ldots, r$

Let C_k be the covariance matrix of $\hat{\beta}_k$ or an estimate of it

The covariance weighted estimate is

$$\ddot{\beta} = \left(\sum_{k=1}^{r} C_k^{-1} \right)^{-1} \sum_{k=1}^{r} C_k^{-1} \beta_k$$

Covariance Weighted Estimators

Many estimation procedures are maximum likelihood

$$\ell(\beta) = \prod_{i=1}^{n} \ell_i(\beta)$$

Let $\tilde{\ell}_v$ be the product just for subset v so there are m

As m gets large

$\tilde{\ell}_v$ becomes well approximated by the normal and $\ddot{\beta}$ is the maximum likelihood estimate

D&R Statistical Theory for D&R

Statistical analytic methods now under study

1. Logistic regression as a start in the area of factor-response analysis

2. Autoregressive processes as a start in the area of time series

3. Non-parametric regression: the loess local regression method

Results have been excellent