# INFORMATION VISUALIZATION / EXPLORATORY DATA ANALYSIS

*Prof. Marti A. Hearst, UC Berkeley*

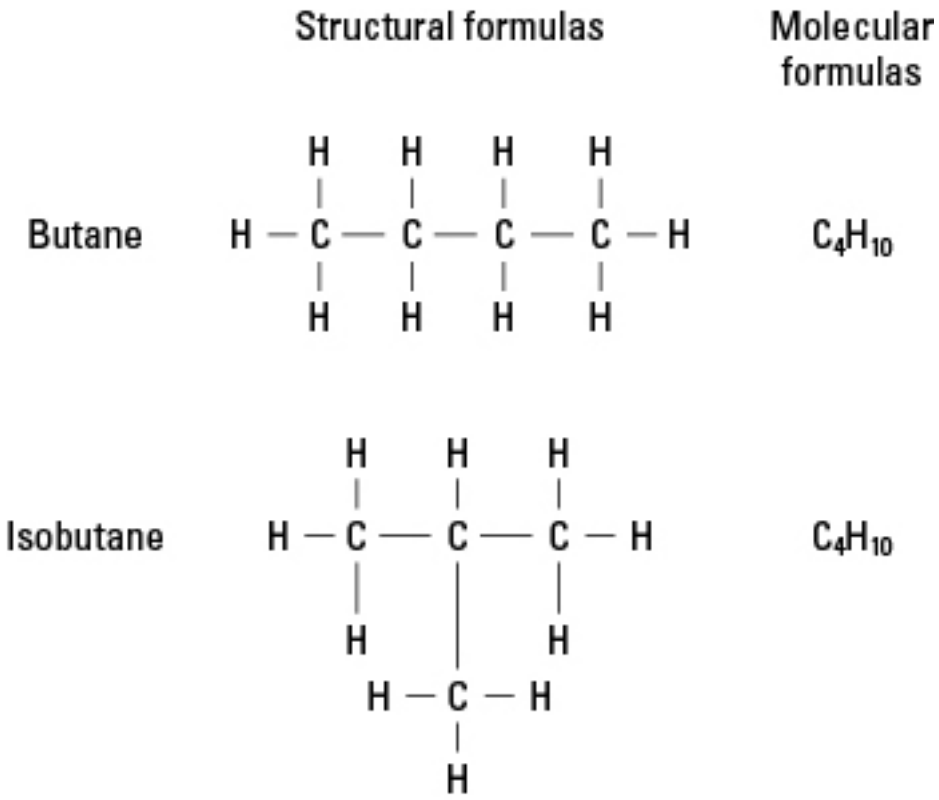
*Bay Area PHUSE Single Day Event, 9/26/2014*

# WHY INFOVIZ?

*Why not just statistics?*

# INSIGHT

# Same Formula; Different Structure




http://www.cliffsnotes.com/sciences/chemistry/chemistry/organic-compounds/structural-formulas

| Set A | | Set B | | Set C | | Set D | |
|---|---|---|---|---|---|---|---|
| X | Y | X | Y | X | Y | X | Y |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.11 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

**Summary Statistics**

$u_X = 9.0$ $\sigma_X = 3.317$

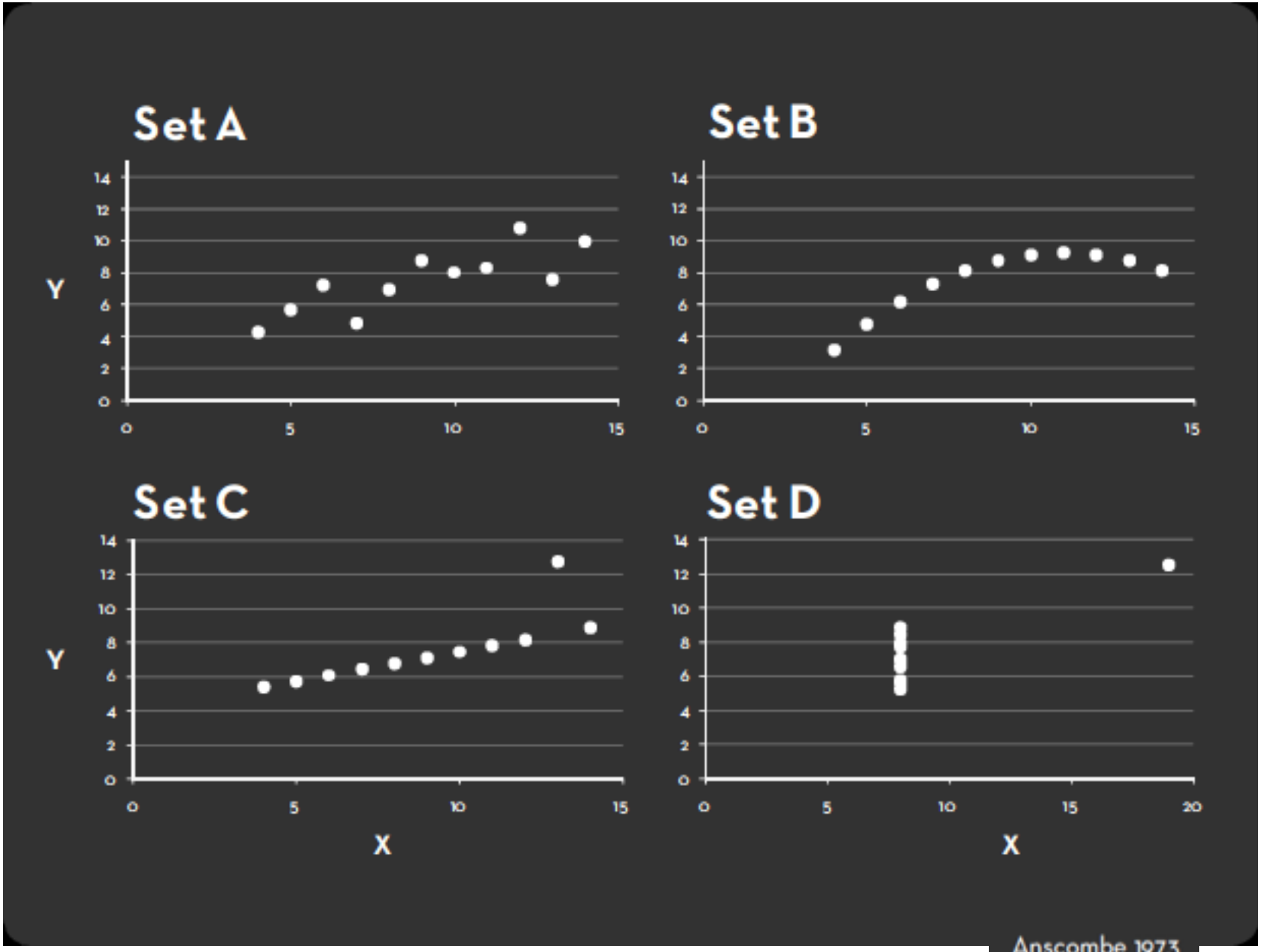
$u_Y = 7.5$ $\sigma_Y = 2.03$

**Linear Regression**

$Y = 3 + 0.5\,X$

$R^2 = 0.67$

Anscombe 1973


Set A
Y
14
12
10
8
6
4
2
0
0
5
10
15
Set B
14
12
10
8
6
4
2
0
0
5
10
15
Set C
Y
14
12
10
8
6
4
2
0
0
5
10
15
X
Set D
14
12
10
8
6
4
2
0
0
5
10
15
20
X
Anscombe 1973

PRESENTATION

ANALYSIS

# BRILLIANT INTERACTION

*NYTimes: Paths to the White House*
*infoviz students: TaleSpin Airline Accidents*

# INFOVIZ FINAL PROJECT DEMONSTRATIONS

FRIDAY, MAY 9TH, 2-4PM, 110 SOUTH HALL

## FINAL PROJECT REPORTS AND DEMONSTRATIONS 2014

Visualizing Clinical Trials

Colin Gerber, Anobel Odisho, Jason Ost

Visualizing Telegraph Ave

Lisa Jervis and Molly Robison

report

CoinViz Bitcoin Visualization

Arezu Aghaseyedjavadi, Brian Bloomer, Shaun Giudici

# BRILLIANT COMPRESSION

*NYTimes: Front Row to Fashion Week*

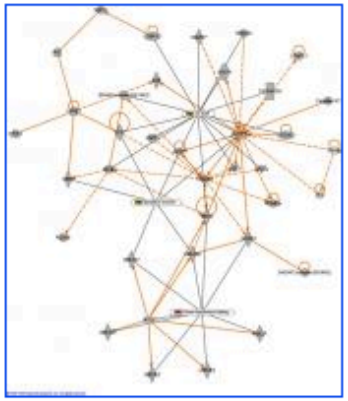
# NARRATIVE + EXPLORATION

*Infoviz class: Cal v. Stanford*
*Drought and Wildfires in California*

# BIOSCIENCE IS RICH IN VISUALIZATIONS

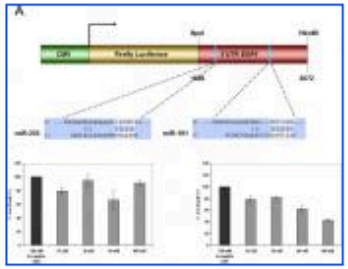
*Find them at biosearch.berkeley.edu*

# Query on “miRNA” at http://biosearch.berkeley.edu



View Fig. & Caption | View All Figs.

Figure 2. Functional relationships of miRNA target genes involved in the...

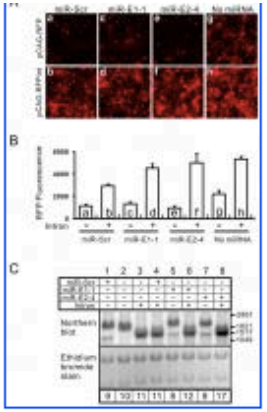


View Fig. & Caption | View All Figs.

Figure 4. MiRNA target gene validation of EGR1 using a luciferase reporter gene...

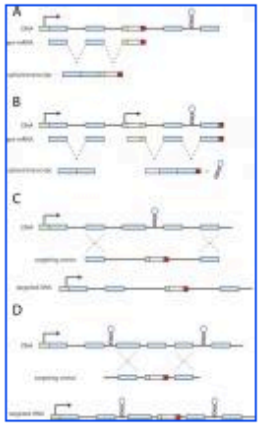
View Fig. & Caption | View All Figs.

Figure 1. Alignment of the Drosophila melanogaster let-7 miRNA to a cognate target...



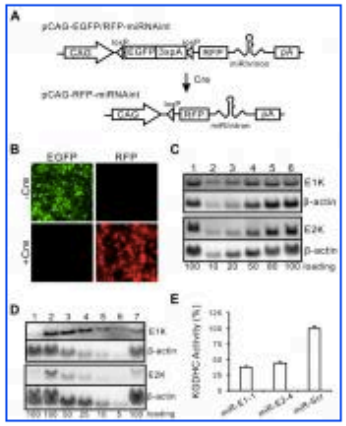
View Fig. & Caption | View All Figs.

Figure 3. Effects of pre-miRNA and intron on RFP expression. (A) Representative...



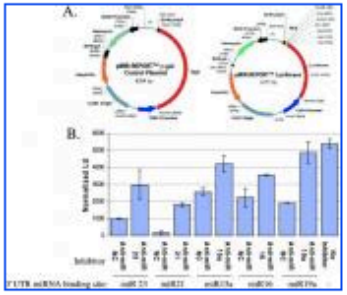
View Fig. & Caption | View All Figs.

Figure 1. Configurations of Protein:miRNA Gene Disruptions. Retroviral gene...



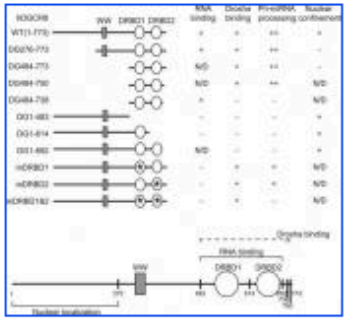
View Fig. & Caption | View All Figs.

Figure 5. Design and test of a Cre-inducible miRNA-expression construct with...



View Fig. & Caption | View All Figs.

Figure 1. (A) MiRNA reporter vectors used to analyze miRNA activity. The restriction sites for...



View Fig. & Caption | View All Figs.

Figure 6. Summary of the DGCR8 mutagenesis findings. Our DGCR8 mutants can be characterized into...

# WHAT IS EXPLORATORY DATA ANALYSIS?

# What is Exploratory Data Analysis?

- “looking at data to see what it seems to say” – Tukey

- the first step **before** confirmatory techniques
- getting to know the “lay of the land”
- detective work

# Why do we do EDA?

- Optimize a function
- Create a narrative, tell a story
- Plan a strategy
- Monitor a situation
- Look for causal relationships
- Solve a crime

# What do we do when we do EDA?

- Overviews
- Get to know the data: tendencies, trends
- Make discoveries
- Find anomolies / outliers / errors
- Formulate hypotheses
  - Support / refute / confirm them

# EDA Assignment

- Readings:
  - Steve Few's book *Now You See It*
  - Steve Few on Exploratory Vistas
  - Al Inselberg on Multidimensional Data Detective
  - Polaris paper (foundations of Tableau)
  - Short Tukey paper on EDA
  - Tableau Videos
- Demonstrations of Tableau on a data set for something we did in class
- Practice Assignment
  - Study EDA assignments by 2 other students
  - Write paragraphs about each
  - Instructor commented on these
    - Redirected students who were not on the right path

# EDA Assignment

- Assignment:
  - Select data set
  - Formulate hypotheses
  - Explore those hypotheses; let the data determine next steps
  - Verify or refute hypotheses; reformulate or formulate new ones
  - Create dashboards summarizing the results for each.

# EDA TOOL COMPARISON

*Saraiya, North, & Duca. "An evaluation of microarray visualization tools for biological insight.Infoviz 2004.*

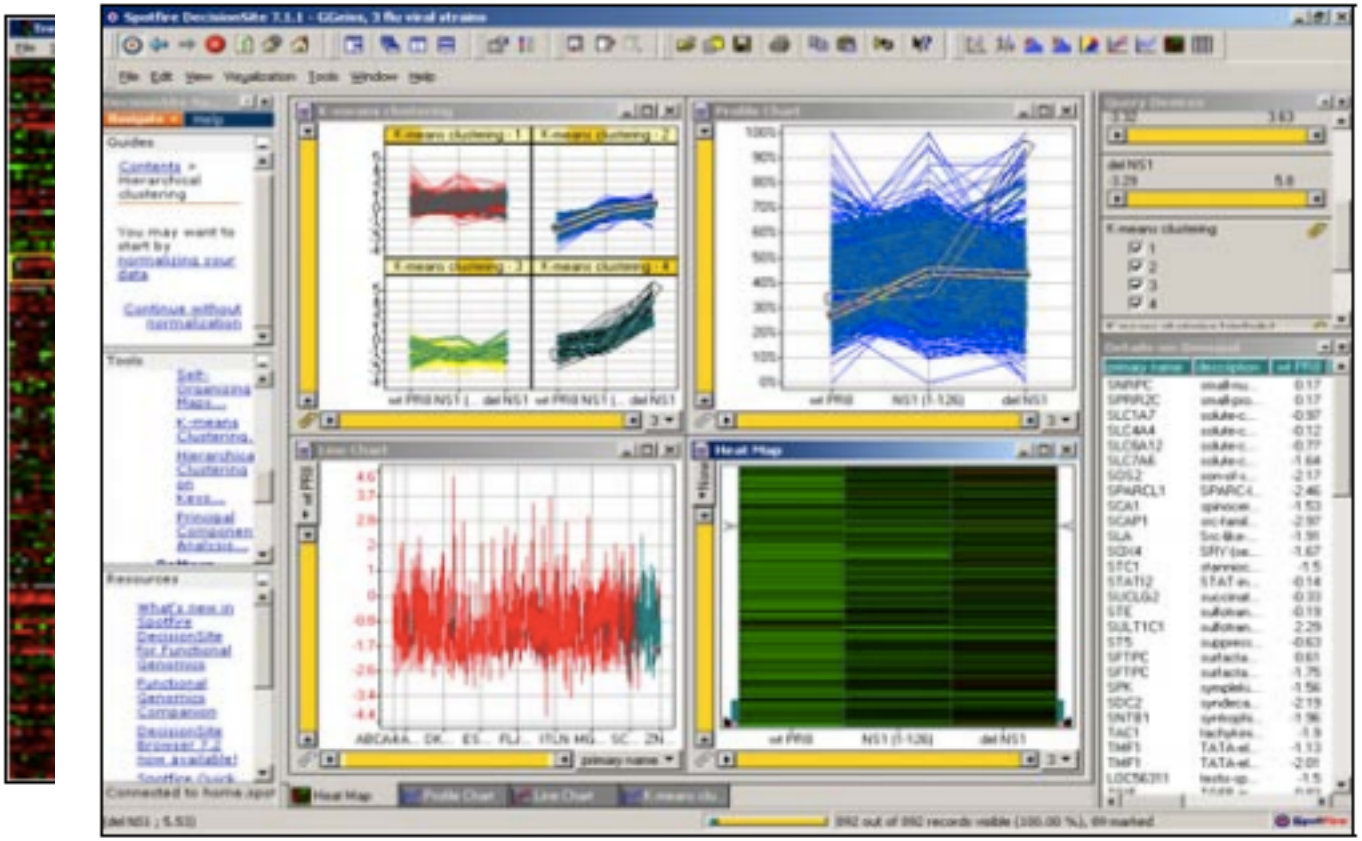


Figure 4: Spotfire® [30]

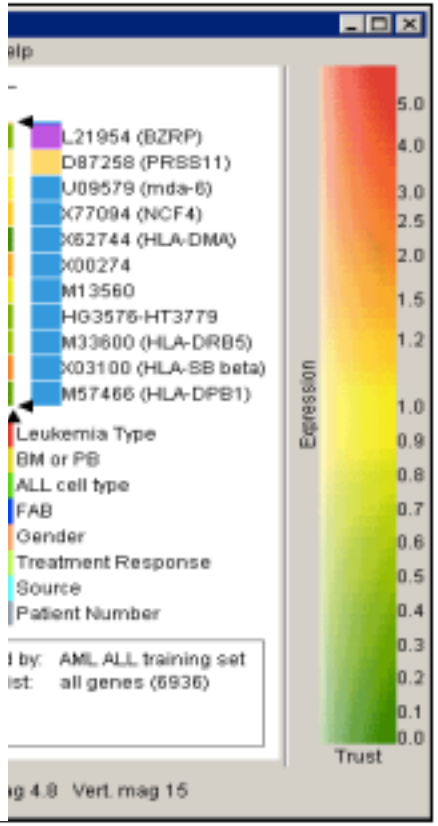
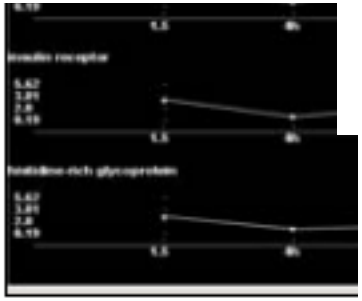



Figure 5: GeneSpring® [12]



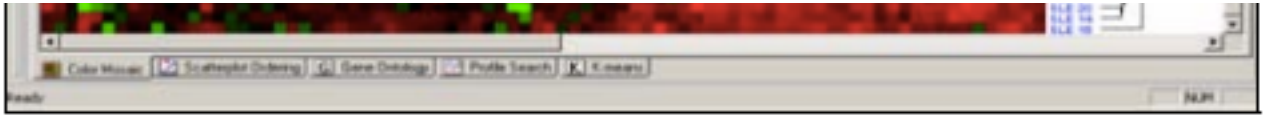


Figure 3: Hierarchical Clustering Explorer (HCE) [27]

Saraiya,

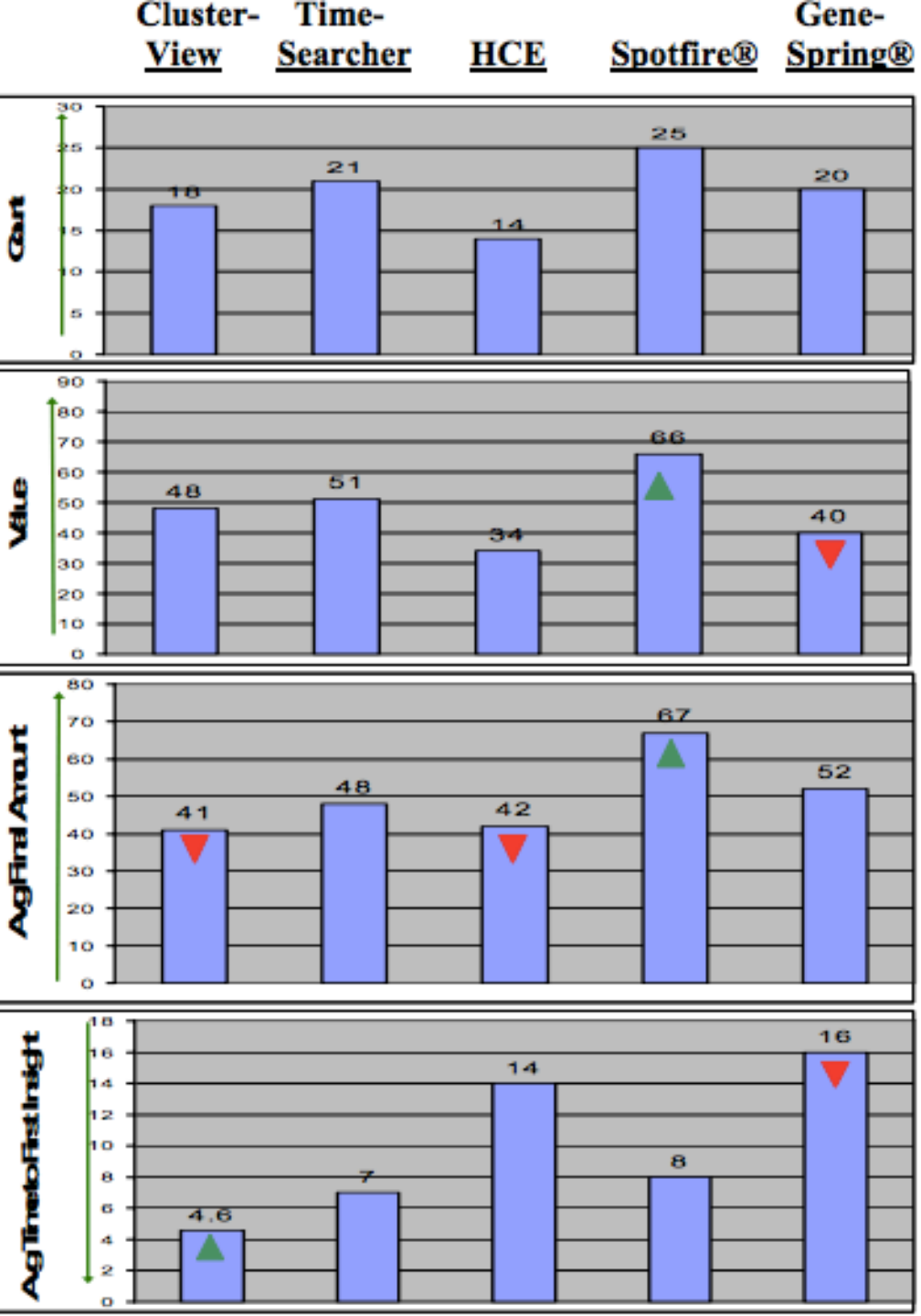



Table 6: Unexpected, hypotheses, and incorrect insights

| Visualization Tool | Unexpected Insights | Hypotheses Generated | Incorrect Insights |
|---|---|---|---|
| Clusterview | 3 | 2 | 0 |
| TimeSearcher | 3 | 1 | 0 |
| HCE | 5 | 1 | 2 |
| Spotfire® | 2 | 3 | 0 |
| GeneSpring® | 0 | 0 | 0 |

Table 7: Total number of insights in each category

| Tool | Overview | Patterns | Groups | Detail |
|---|---|---|---|---|
| Clusterview | 9 | 10 | 0 | 2 |
| Timesearcher | 10 | 8 | 0 | 3 |
| HCE | 6 | 5 | 0 | 1 |
| Spotfire® | 13 | 10 | 1 | 1 |
| GeneSpring® | 5 | 8 | 4 | 1 |

Saraiya, North, & Duca. "An evaluation of microarray visualization tools for biological insight.Infoviz 2004.

# EDA FUTURE DIRECTIONS

*We are researching how to improve the process.*

# THANK YOU.

*Prof. Marti A. Hearst, UC Berkeley*
*Bay Area PHUSE Single Day Event, 9/26/2014*