The Data Deluge (Economist, March 2010)
sand + nuggets
sand + nuggets

nuggets
The Data Deluge  (Economist, March 2010)

sand + nuggets

filter

nuggets
The Data Deluge (Economist, March 2010)

data + information

analysis

insight

Brett Ryder
Data: Bivariate Relationships

Scatterplot
Data: Multivariate Relationships

Scatterplot matrix
Scatterplot matrix

- Can you really see something here?
- Don’t forget, *time is money*...
Household Plots
Household Plots

Can these convey multivariate relationships?
Insight = Patterns

From Economist:

Merely keeping up with this [data] flood, and storing the bits that might be useful, is difficult enough.

Analysing it, to spot patterns and extract useful information, is harder still.
Insight = Patterns

From Economist:

Merely keeping up with this [data] flood, and storing the bits that might be useful, is difficult enough.

Analysing it, to spot patterns and extract useful information, is harder still.

We need tools that allow users to spot patterns
  • supported by modern data analysis tools
Some Ground Rules

- Provide overview + detail on demand
- Provide focus + context
- Make tools familiar + intuitive
- Realize that there is no tool that does it all
Some Ground Rules

- Provide overview + detail on demand
- Provide focus + context
- Make tools familiar + intuitive
- Realize that there is no tool that does it all
Some Ground Rules

- Provide overview + detail on demand

- Provide focus + context

- Make tools familiar + intuitive (really the same)

- Realize that there is no tool that does it all (link them together)
Some Ground Rules

- Provide overview + detail on demand
- Provide focus + context
- Make tools familiar + intuitive (really the same)
- Realize that there is no tool that does it all (link them together)
Welcome to TripAdvisor<sup>ND</sup>

The tool that has it all
User-driven data reduction

- go from millions to ten-thousands
- density preserving
- outlier preserving
- multi-scale
TripAdvisor\textsuperscript{ND} Analysis Suite (1)

User-driven data reduction

- go from millions to ten-thousands
- density preserving
- outlier preserving
- multi-scale

User-driven dimension reduction

- go from hundreds to dozens
- significance preserving
- correlation preserving
- multi-scale
User-driven cluster analysis

- traditional clustering
- subspace clustering
- multi-scale
User-driven cluster analysis
  • traditional clustering
  • subspace clustering
  • multi-scale

User-driven cause + effect analysis
  • partial correlation
  • structural equation modeling
  • multi-scale
Data-Centric Overview

How can the N-D space conveniently be overviewed in 2D?

- we provide two solutions

Overall space layout

Sub-space layout
How can we show just the amount of detail needed?

- We make heavy use of illustrative abstraction
Detail on Demand

parallel coordinate lines
Detail on Demand

%Completed: 1
Leads: 2,730
Leads Won: 2,350
#Opportunity: 151
Pipe: 203,505
Expected ROI: 6
Actual Cost: 76,982
Cst/won Lead: 338
Planned Rev: 199,402
Planned: 12

blended lines
Detail on Demand

complete abstraction
Correlation is a prime metric to gauge relationships

- we provide a correlation-oriented layout linked with parallel coordinates (support numeric and categorical variables)
Example from Climate Science

Parallel Coordinate display

Correlation display

see how data relate

see how variables relate

see overall data trends

Low-vs. high pressure (dimension 1)
Visualization Mashups

For example: link geo-spatial with information displays

- Outlined areas
- Geographical display
- Information display

Variable selector:
- Cloud fraction
- Precipitation
- Liquid/ice radii
- Top height
- Base

Stream mapping:
- Area 2
- Area 1
- Area 3
Visualization Mashups

For example: link geo-spatial with information displays
Periodic Relationships

For example: view and compare energy usage in a household.
Practical Analysis in N-D Space

A customer is born...
How do sales teams compare?
What is a winning sales strategy?
What are the key variables?

(see demo next)
Proposals

#1: Work with CDDA members to extend our tools
   • growth happens only with collaborations and user studies
   • new data, new problems, new challenges

#2: Accelerate analysis and visualizations on GPUs
   • will greatly help interactivity with large data

#3: Develop in-software tutorials
   • will bring more complex visualization to the masses

#3: Put system in the cloud
   • enables fast responsive software updates
   • enables remote, browser-based, platform-neutral interactions
   • work out data security issues
Bonus: Healthcare Analytics

(see video)
Thanks

Support from NSF, NIH, DOE, BNL, PNL, CEWIT

Collaborators:
• Dr. Alla Zelenyuk, Dr. Dan Imre (formerly BNL, now PNL)
• Dr. IV Ramakrishan (Stony Brook University)
• Dr. Kevin McDonnell (Dowling College)

MS/PhD Students
• Peter Imrich, Yiping Han, Julia EunJu Nam, Supriya Garg, Hyunjung Lee, Zhiyuan Zhang