CSE 590
Data Science Fundamentals

Intro to Visual Analytics

Klaus Mueller

Computer Science Department
Stony Brook University and SUNY Korea
VISUAL ANALYTICS (Layman’s View)
Visual Analytics (Layman’s View)
Visual Analytics (Layman’s View)
Visual Analytics (Layman’s View)
Visual Analytics

Visual Interface

Computer

Data

Human
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms

Data

manage

Human
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms

Data

Human
- pattern recognition
- creative thought

manage
VISUAL ANALYTICS

Visual Interface

Computer
- computing hardware
- algorithms

Data

manage

Human
- pattern recognition
- creative thought
- mental model
- abstracted knowledge
Visual Analytics

Visual Interface

Computer
- computing hardware
- algorithms
- formal model
- formatted knowledge

Data

Human
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

manage
**Visual Analytics**

**Computer**
- computing hardware
- algorithms
- formal model
- formatted knowledge

**Human**
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

**Visual Interface**

**Data**

formalized insight
Visual Analytics

**Computer**
- computing hardware
- algorithms
- formal model
- formatted knowledge

**Data**

**Visual Interface**

**Human**
- pattern recognition
- creative thought
- mental model
- abstracted knowledge

update

visualize

manage
**Visualization Analytics**

**Computer**
- Computing hardware
- Algorithms
- Apply/update
- Formal model
- Formatted knowledge

**Human**
- Pattern recognition
- Creative thought
- Mental model
- Abstracted knowledge

**Visual Interface**
- Interact
- Learn
- Manage

**Data**
- Learn
- Manage
Visual Analytics

**Computer**
- computing hardware
- algorithms
- apply/update
- formal model
- formatted knowledge

**Human**
- pattern recognition
- creative thought
- apply/update
- mental model
- abstracted knowledge

**Visual Interface**
- visual communication

**Data**
- update
- learn
- manage
- interact
- visualize
First a Primer on Multivariate data Visualization
Consider the salient features of a car (not really big data):

- miles per gallon (MPG)
- top speed
- acceleration
- number of cylinders
- horsepower
- weight
- year
- country origin
- brand
- number of seats
- number of doors
- reliability (average number of breakdowns)
- and so on...
Hard to see patterns...
a car as a 7-dimensional data point
a database of cars
after some clustering
PARALLEL COORDINATES

with mean trend
PC WITH ILLUSTRATIVE ABSTRACTION

%Completed 1

#Leads 2,730

Leads Won 2,350

#Opportunity 151

PipeRev 203,505

ExpectedROI 6

Actual Cost 76,982

Cst/wonLead 338

PlannedRev 199,402

PlannedROI 2

individual polylines
completely abstracted away
PC WITH ILLUSTRATIVE ABSTRACTION

blended partially
all put together – three clusters
Interaction is Key

Interaction in Parallel Coordinate
Case Study: Sales Strategy Analysis
Anatomy of a Sales Pipeline

Lead generator

Lead
Qualified lead
Opportunity
Opportunity
Opportunity
Opportunity
Opportunity
Customer

Prospect

Responds
Responds
Requests info (RFI)
Requests pricing (RFP)
Shapes deal
Signs/buys

Lead pool
Scene:
- a meeting of sales executives of a large corporation, Vandelay Industries

Mission:
- review the strategies of their various sales teams

Evidence:
- data of three sales teams with a couple of hundred sales people in each team
Meet Jim, one of the sales strategy analysts.

- He begins and constructs the following correlation graph.
He asks the TSP to compute an initial route
It gives rise to this parallel coordinate display
He asks the TSP to compute an initial route.
It gives rise to this parallel coordinate display.
He asks the TSP to compute an initial route
It gives rise to this parallel coordinate display
Now meet Kate, another sales analyst in the meeting room:

“Hey, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads”
"Let’s go and make a more revealing route!"

- so she uses the mouse and designs the route shown
“Let’s go and make a more revealing route!”

- so she uses the mouse and designs the route shown
It is now immediately obvious:

- the blue team employs a very different strategy than the green and the red teams.
- it generates far fewer leads but spends much more resources on each → this gives it an advantage in the final outcome.
- the blue team is also much more consistent than the other teams, as indicated by the much narrower band
Kate notices something else:

- now looking at the red team
- there seems to be a spread in effectiveness among the team
- the team splits into three distinct groups

She recommends: "Maybe fire the least effective group or at least retrain them"
Scenario 2
Let us again imagine a meeting of company executives at Vandelay Industries who now would like to make sales policies for the next year based on their three sales teams’ behaviours of this year.

Again, the data contains three sales teams of a large corporation with a couple of hundred sales people in each team.
Now meet John from the market planning department who always wants to create more opportunities.

"Since cost does not have strong correlations with other variables, the company can make any strategies for the other variables, and it will not influence the actual cost."
“The company should raise efforts to create more *opportunities* for the next year without considering the money issue.”
Emily, from the financial department, believes that there must be something wrong with this statement since *cost should* play an important role in the sales pipeline.

She notices these three sales teams behave quite differently. It is likely a mistake if they consider the three teams altogether.
She suggests that they plot the correlation maps for the three teams separately.

The three teams have quite similar correlation patterns, which is consistent with her expertise that there must be some marketing model that guides the sales behaviors and the model should involve cost in it.
There are 7 variables involved in the pattern: *opportunities*, *cost*, *cost per won lead*, *# leads*, *# leads won*, *expected ROI*, and *pipeline revenue*

- the other variables are not highly related
- so these 7 variables should be focused on as references to make decisions
1 – increase the # opportunities (John’s idea)
   - cost per won lead will increase
   - the overall cost will increase
   - expected ROI will be reduced

2 – reduce cost (her idea)
   - cost per won lead will have to be reduced
   - opportunities will then be reduced
   - expected ROI will be increased
CEO Tom weighs in:

- Increasing *cost* is **not** preferred because this year’s expense already exceeds the budget
- Although the number of *opportunities* is reduced, the *expected ROI* will go up

So Tom decides on business strategy #2

- Sorry, John