CSE 332
Introduction to Visualization

Data Preparation, Representation, & reduction

Klaus Mueller

Computer Science Department
Stony Brook University

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, data, and basic tasks	
3	Data preparation, representation, and reduction	Project 1 out
4	Visual perception and cognition	
5	Introduction to D3, basic vis techniques for non-spatial data	
6	Visual design and aesthetics	
7	Foundations of statistics	Project 2 out
8	Data types, notion of similarity and distance	
9	Data mining techniques: clusters, text, patterns, classifiers	
10	Data mining techniques: clusters, text, patterns, classifiers	
11	High-dimensional data, dimensionality reduction	
12	Computer graphics and volume rendering	Project 3 out
13	Techniques to visualize spatial (3D) data	
14	Scientific and medical visualization	
15	Scientific and medical visualization	
16	Non-photorealistic rendering	
17	Midterm	
18	Principles of interaction	Project 4 out
19	Visual analytics and the visual sense making process	
20	Correlation and causal modeling	
21	Big data: data reduction, summarization	
22	Visualization of graphs and hierarchies	
23	Visualization of text data	Project 5 out
24	Visualization of time-varying and time-series data	
25	Memorable visualizations, visual embellishments	
26	Evaluation and user studies	
27	Narrative visualization and storytelling	
28	Data journalism	

Rectangular Dataset

One data item

The variables
→ the attributes or properties we measured

The data items
→ the samples (observations) we obtained from the population of all instances

	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100

Rectangular Dataset

Also called the *Data Matrix*

Car performance metrics

or Survey question responses

or Patient characteristics

....

One data item

Car models

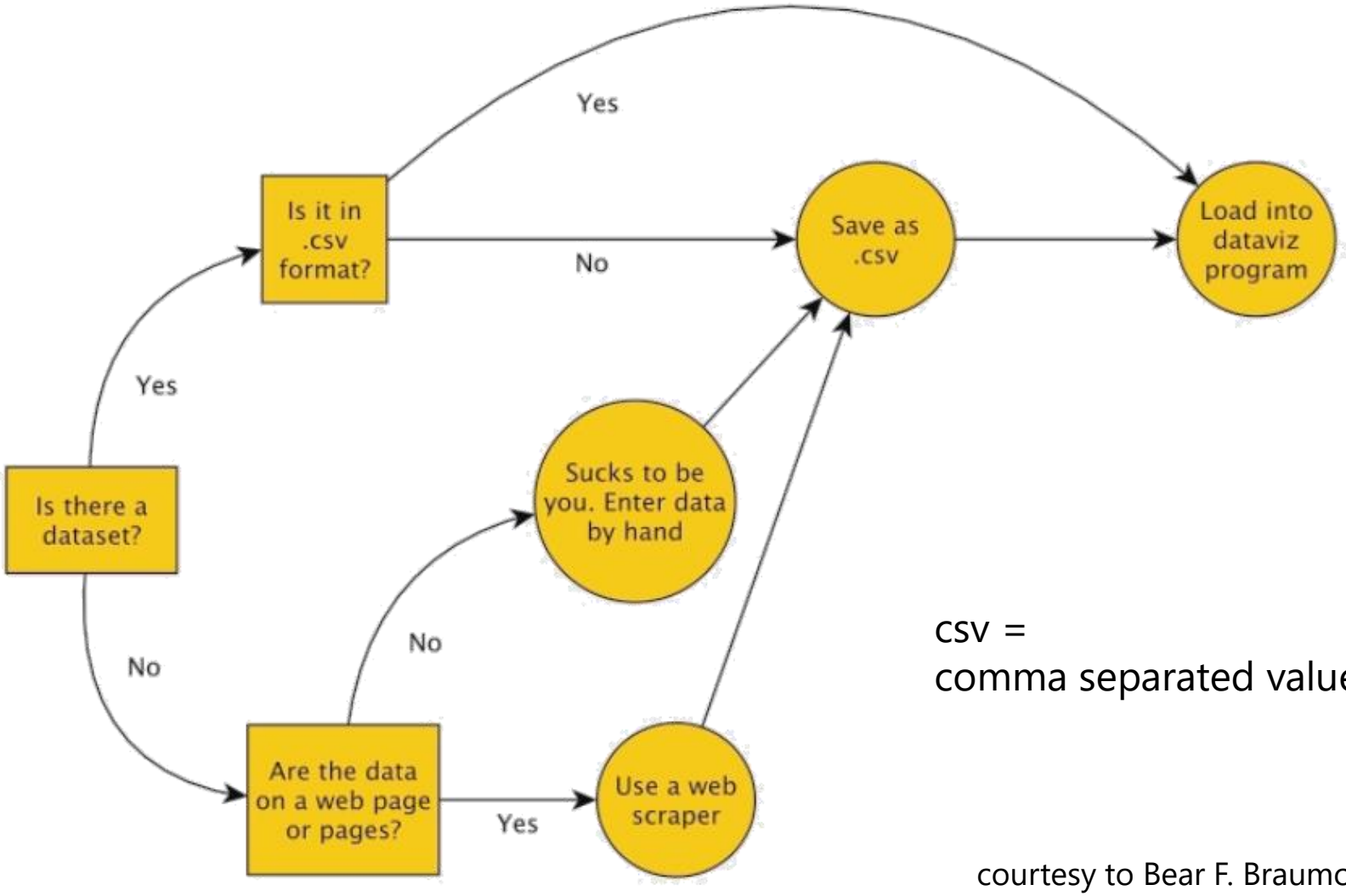
or Survey respondents

or Patients

....

	A	B	C	D	E	F	
1	**Name**	**Country**	**Miles Per Gallon**	**Accceleration**	**Horsepower**	**weight**	**cylir**
2	Volkswagen Rabbit DI	Germany	43,1	21,5	48	1985	
3	Ford Fiesta	Germany	36,1	14,4	66	1800	
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	
8	Dodge Diplomat	USA	19,4	13,2	140	3735	
9	Mercury Monarch	USA	20,2	12,8	139	3570	
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	
12	Ford Fairmont A	USA	20,2	15,8	85	2965	
13	Ford Fairmont M	USA	25,1	15,4	88	2720	
14	Plymouth Volare	USA	20,5	17,2	100	3430	
15	AMC Concord	USA	19,4	17,2	90	3210	
16	Buick Century	USA	20,6	15,8	105	3380	

How To Import Data?



csv =
comma separated values file

courtesy to Bear F. Braumoeller

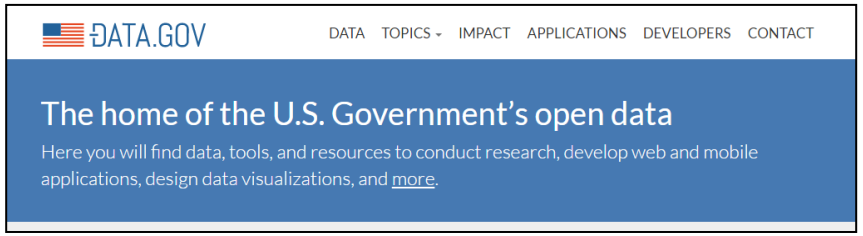
How To Get data? (1)

Use Google

- type the topic you like and perhaps ‘data”, ‘’database’. csv’, etc.

Other sources:

- https://www.data.gov/



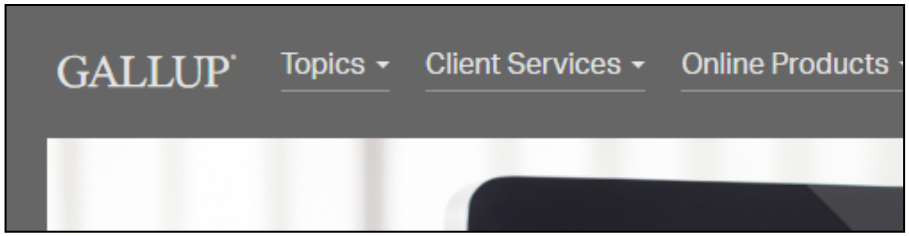
- https://fedstats.sites.usa.gov/

- http://data.worldbank.org/

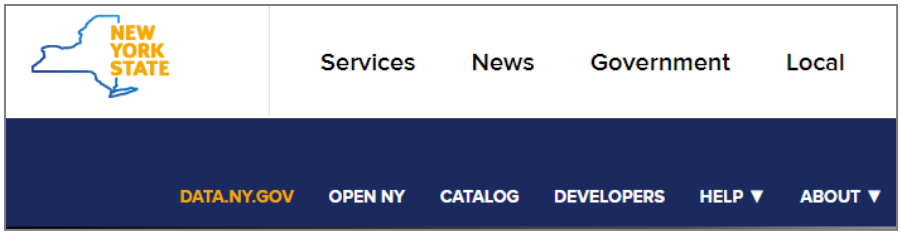
How To Get data? (2)

Other sources:

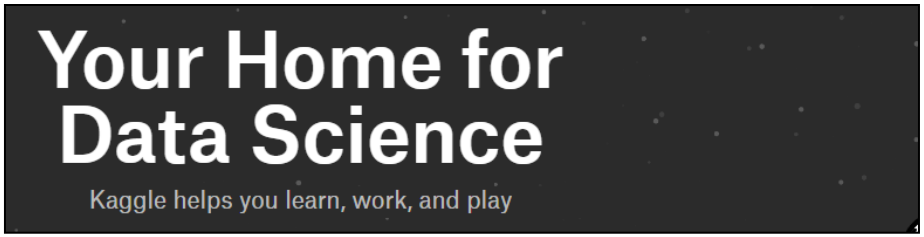
- http://www.gallup.com/products/184157/gallup-analytics-universities-colleges.aspx



- https://data.ny.gov/



- https://www.kaggle.com/



Project #1

Use these means to find some interesting data on the web

- something that challenges and interests you
- there are many data sources on the web
- use google and some imagination

Criteria for selection

- more than 500 data points (observations)
- more than 10 attributes
- the more the better (you can always reduce it)

Deliverables

- 2-page report that describes the data and justifies your choice
- a URL to the data source

Do NOT mention your name on the report

Due date

- Tuesday, Sept. 20, 11:59pm (submission procedure – stay tuned)

Project #1: Dataset Example

Multivariate - Quantitative data and Categorical data

Data Items

	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100
25	Toyota Corona	Japan	27,5	14,2	95	2560	4	78	2975

Categorical | **Quantitative** | **Categorical (Ordinal)** **Quantitative**

Data types

Quantitative (Numerical)

Categorical (Ordinal)

Project #1: Notes on Dataset

Other data types are OK

- text, images, video, logs, etc.
- just convert them to numbers via appropriate mechanism as discussed in class
- must produce a spreadsheet of rows (data items) and attributes (columns)

Categorical data

- color, brand, country, etc.
- concert into numbers by assigning a numerical ID

After Downloading The Data ...

Do you think data are always clean and perfect?

Think again

Real world data are dirty

Data cleaning (wrangling)

- fill in **missing values**
- smooth **noisy data**
- identify or remove **outliers**
- resolve **inconsistencies**
- **standardize/normalize** data
- **fuse/merge** disjoint data

Missing Values

Data is not always available

- e. g, many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- many more reasons

Missing Data – Example

Assume you get these baseball fan data

Age	Income	Team	Gender
23	24,200	Mets	M
39	50,245	Yankees	F
45	45,390	Yankees	F
22	32,300	Mets	M
52		Yankees	F
27	28,300	Mets	F
48	53,100	Yankees	M

- How would you estimate the missing value for income?

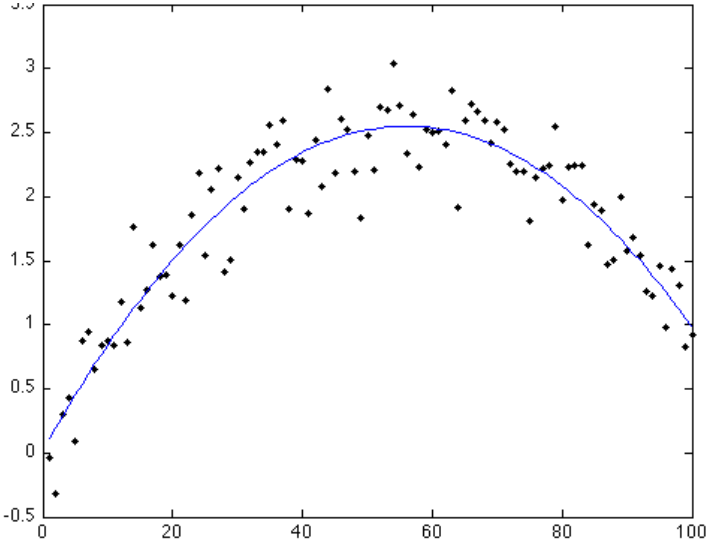
Missing Data – Example

Assume you get these baseball fan data

Age	Income	Team	Gender
23	24,200	Mets	M
39	50,245	Yankees	F
45	45,390	Yankees	F
22	32,300	Mets	M
52		Yankees	F
27	28,300	Mets	F
48	53,100	Yankees	M

- How would you estimate the missing value for income?
  - ignore or put in a default value (will decimate the usable data)
  - manually fill in (can be tedious or infeasible for large data)
  - average over all incomes
  - average over incomes of Yankee fans
  - average over incomes of female Yankees fans
  - use a probabilistic method (regression. Bayesian, decision tree)

Noisy Data



Noise = Random error in a measured variable

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

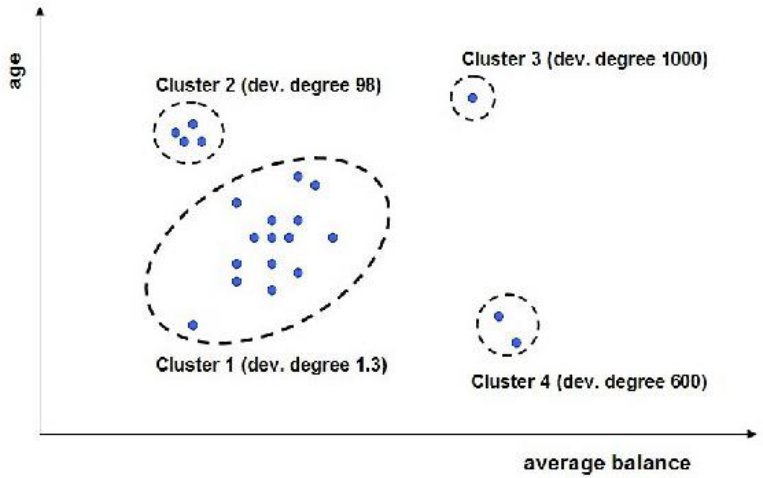
Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data

Noisy Data – What To Do

Binning method

- discussed last lecture



Clustering

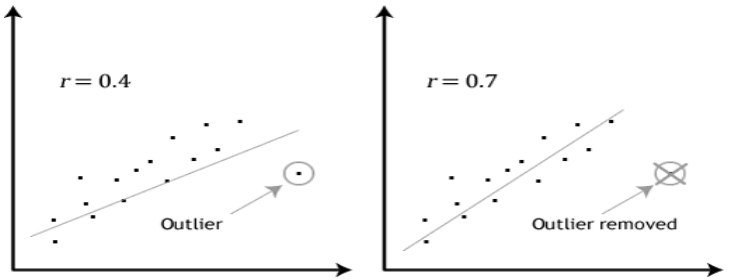
- detect and remove outliers

Semi-automated method

- combined computer and human inspection
- detect suspicious values and check manually (need visualization)

Regression

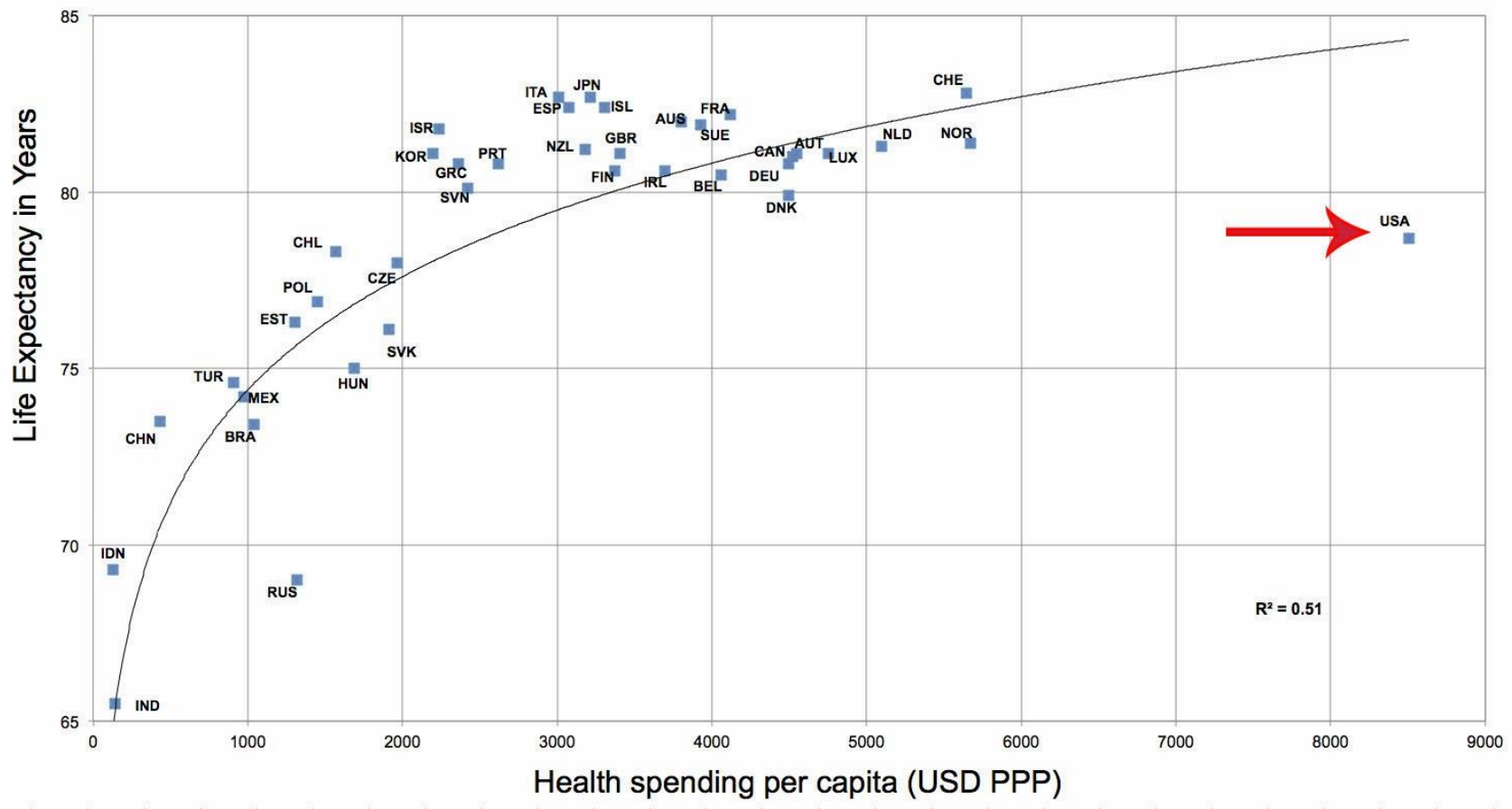
- smooth by fitting the data to a regression function



Noise Removal – A Word Of caution

An outlier may not be noise

- it may be an anomaly that is very valuable (e.g., the Higgs particle)



Resolve Inconsistencies

Inconsistencies in naming conventions or data codes

- e.g., 2/5/2002 could be 2 May 2002 or 5 Feb 2002

Redundant data

- duplicate tuples, which were received twice should be removed

Data Transformation

Can help reduce influence of extreme values

See our discussion last lecture

Data Normalization

Sometimes we like to have all variables on the same scale

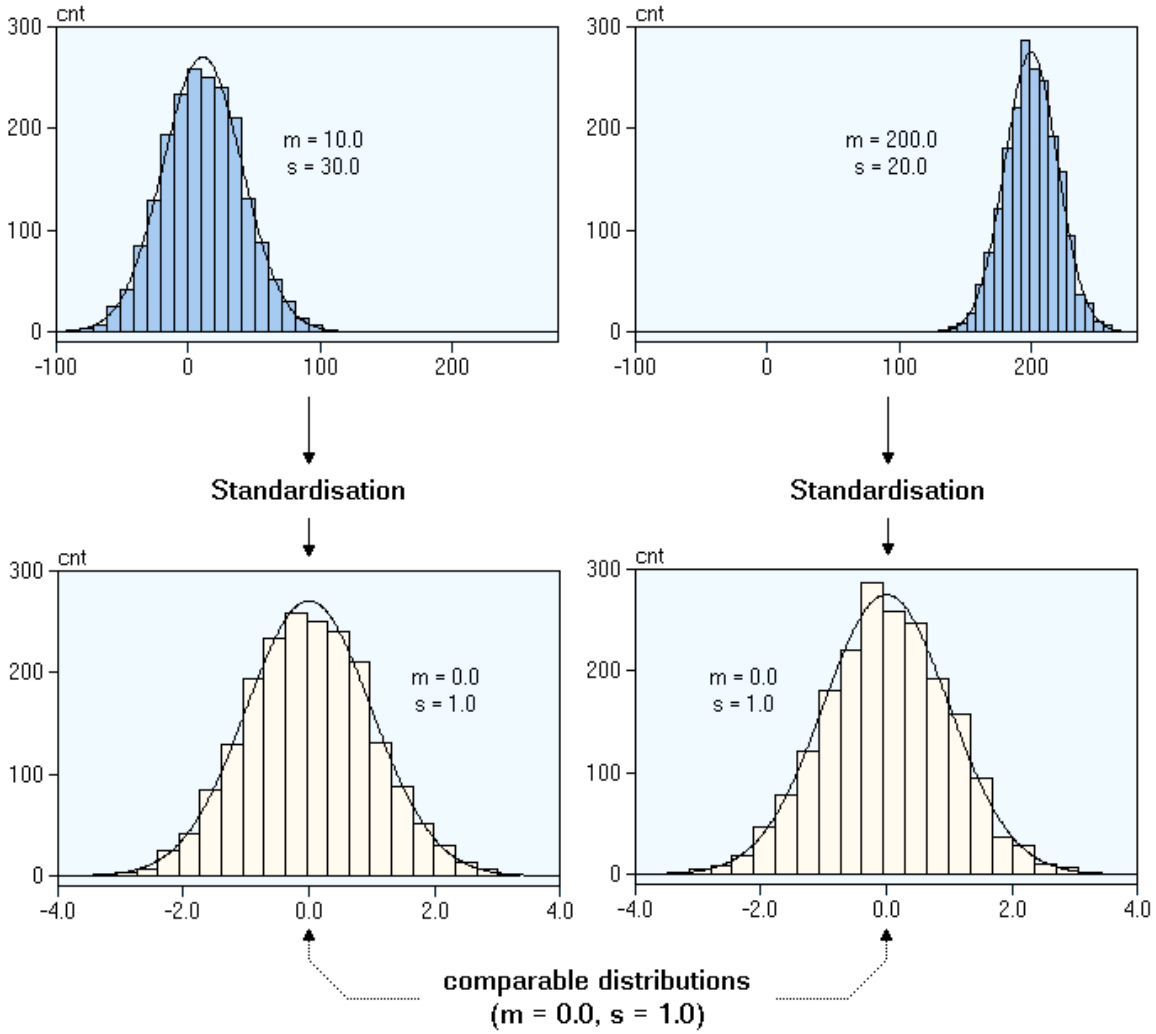
- min-max normalization

$$v' = \frac{v - min}{max - min}$$

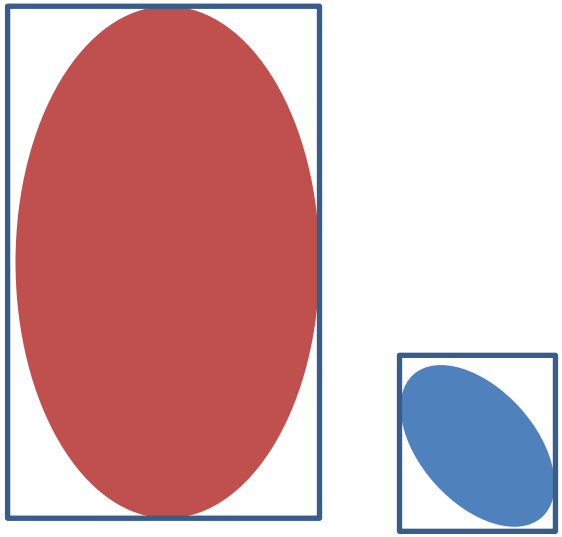
- standardization / z-score normalization

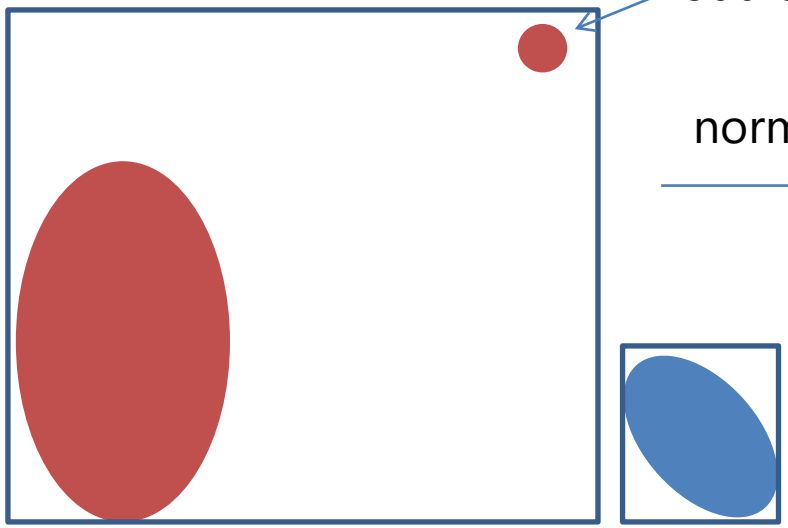
$$v' = \frac{v - \overline{v}}{\sigma_v}$$

Standardization



Normalization

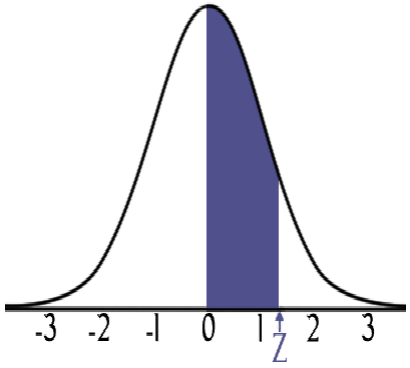




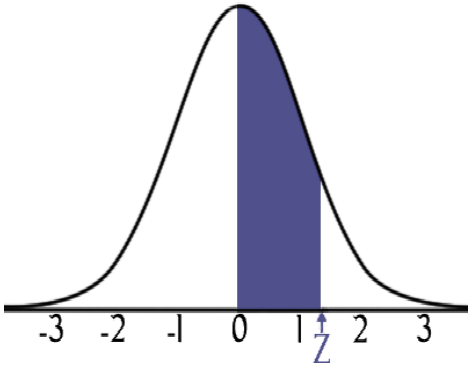
Standardization

Is standardization less or more sensitive to outliers?

socrative
by MasteryConnect



without outlier



with outlier (just slightly extended)

Data Integration

Data integration/fusion

- multiple databases
- data cubes
- files
- notes

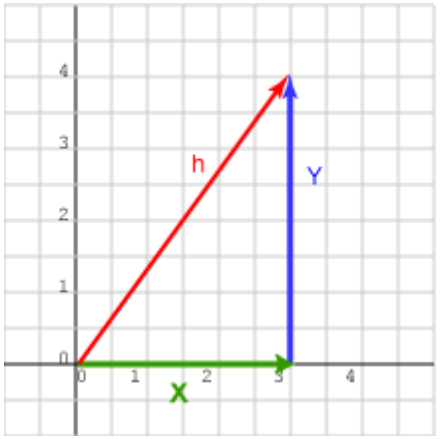
Produces new opportunities

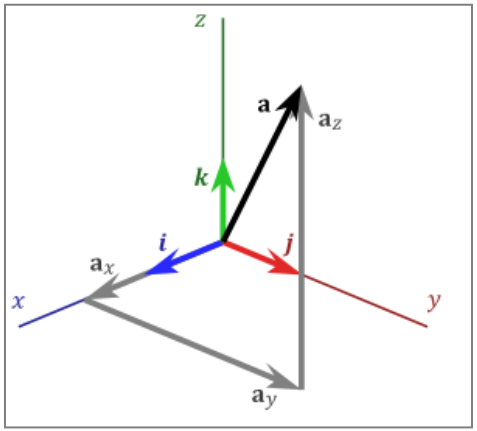
- can gain more comprehensive insight (value > sum of parts)
- but watch out for *synonymy and polysemy*
- attributes with different labels may have the same meaning
  – "comical" and "hilarious"
- attributes with the same label may have different meaning
  – "jaguar" can be a cat or a car

Representation

Each data item is an N-dimensional vector (N variables)

- recall 2D and 3D vectors in 2D and 3D space, respectively

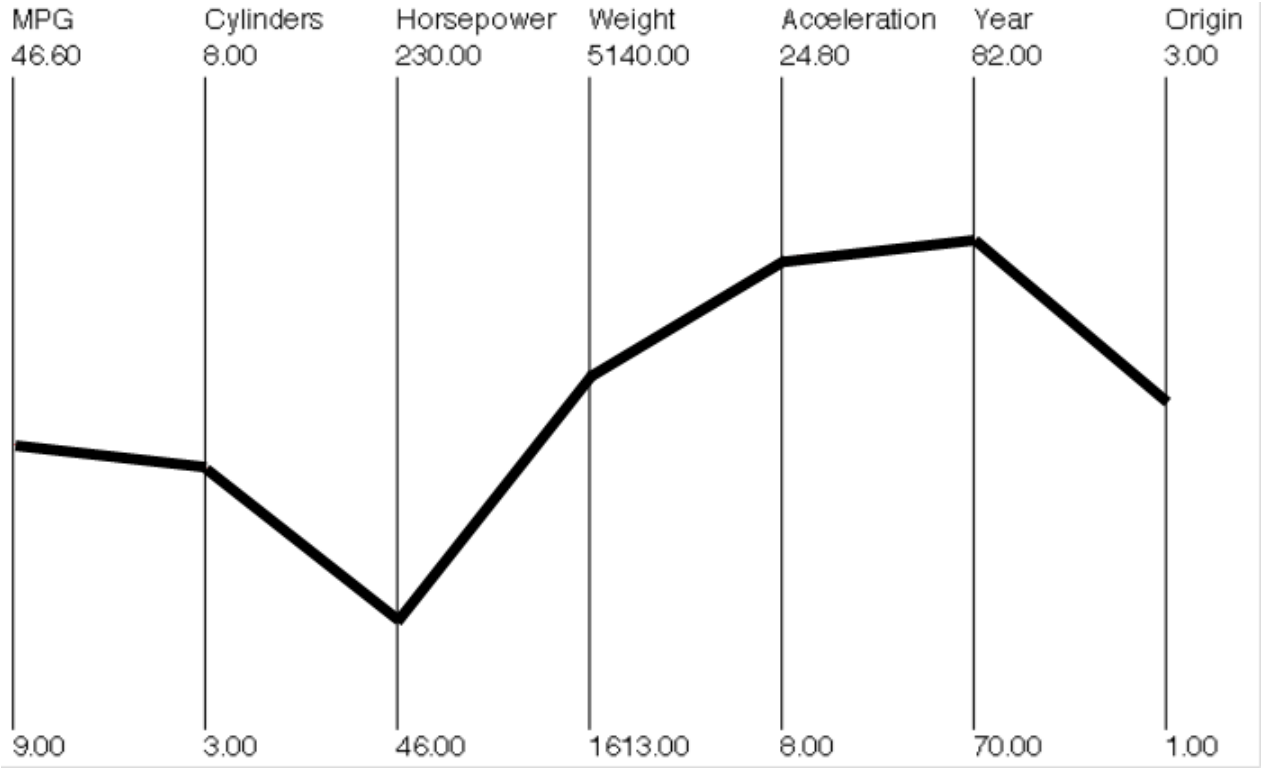




Now we have N-D attribute space

- now the data axes extend into more than 3 orthogonal directions
- hard to imagine?
- that's why need good visualization methods

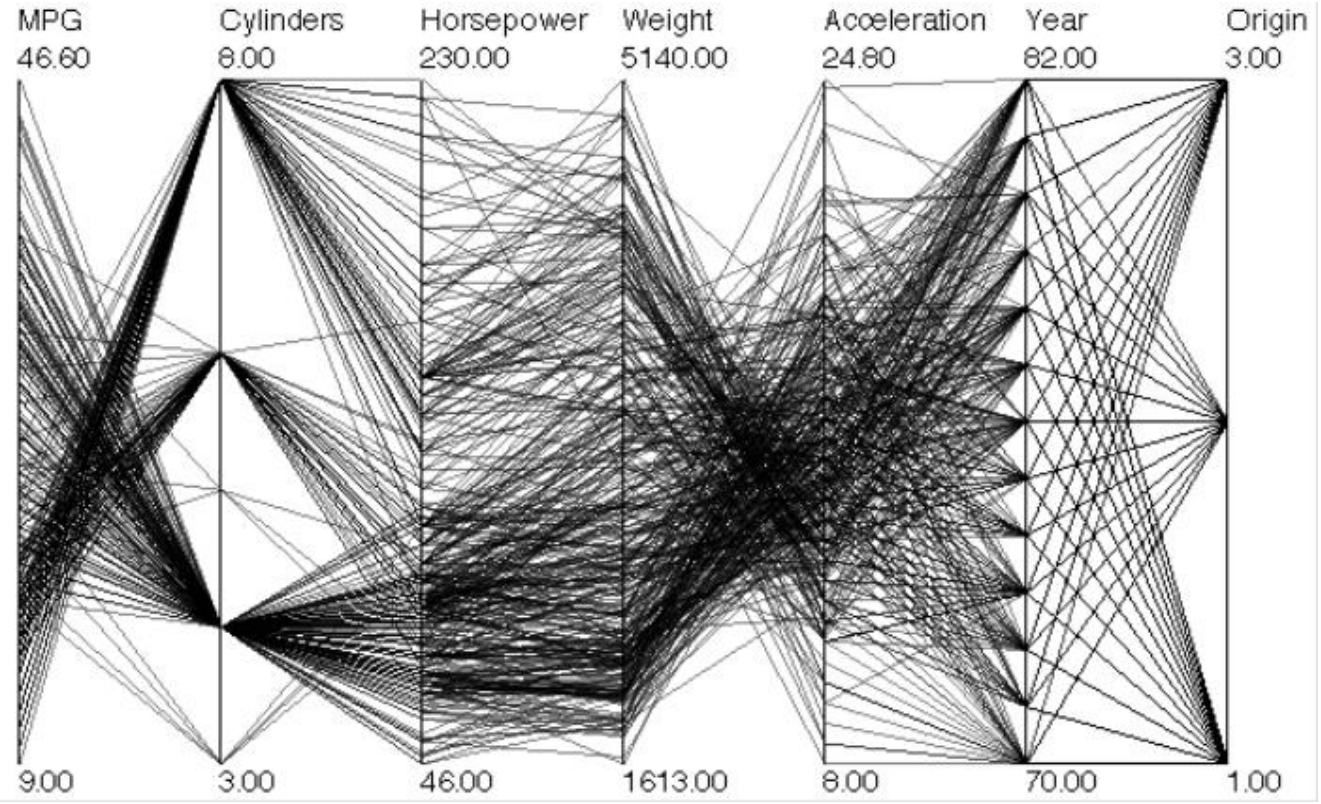
Parallel Coordinates – 1 Car



The N=7 data axes are arranged side by side

- in parallel

Parallel Coordinates – 100 Cars

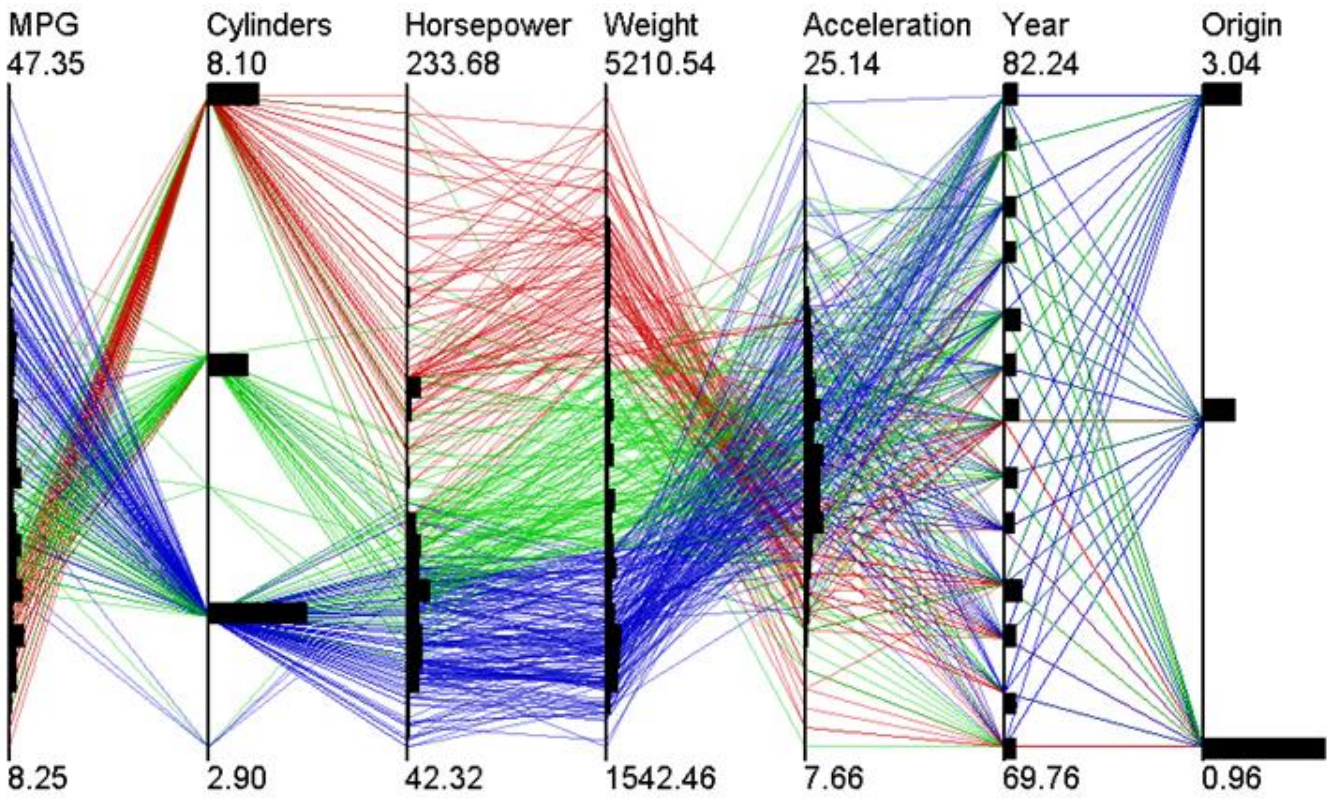


Hard to see the individual cars?

- what can we do?

socrative
by MasteryConnect

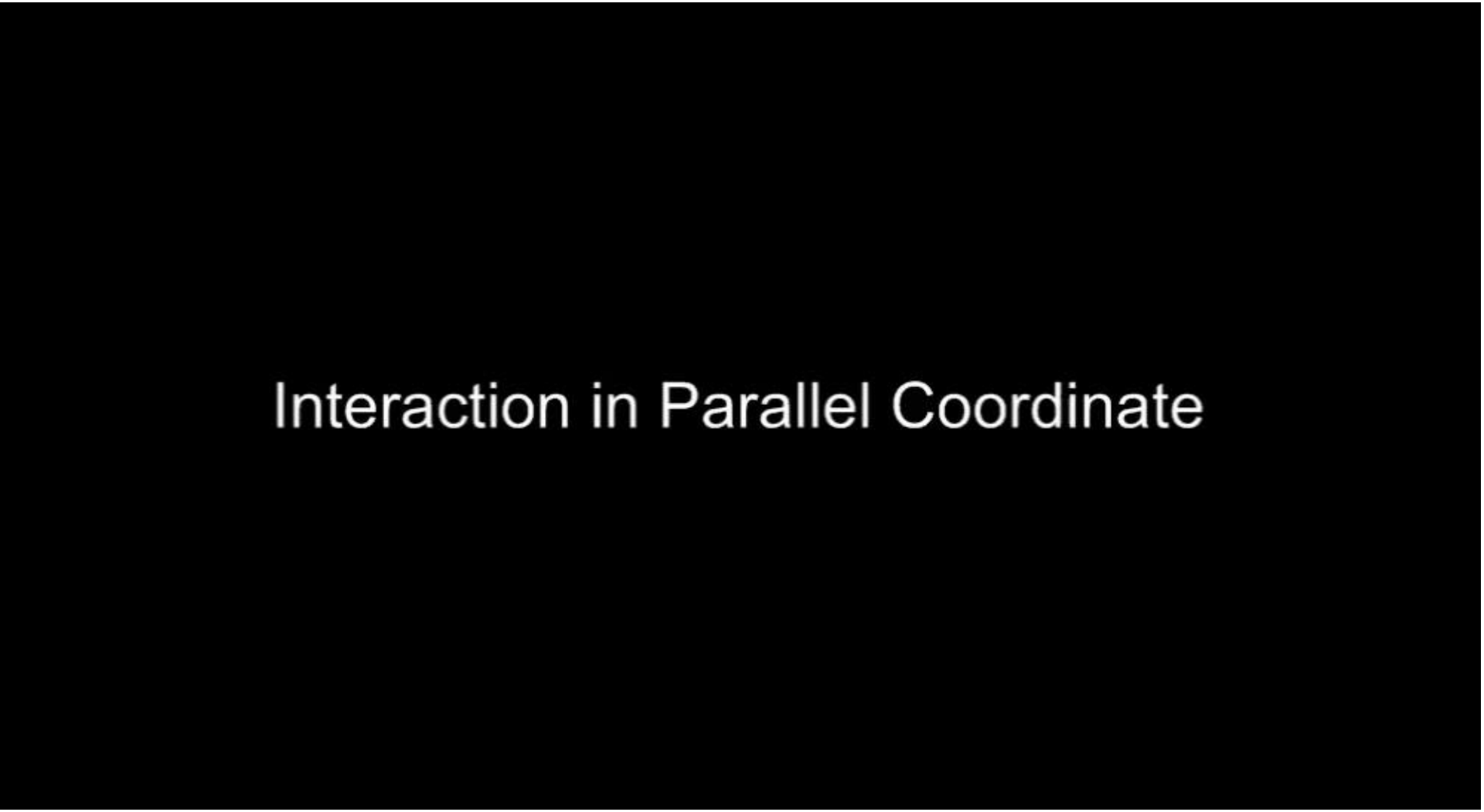
Parallel Coordinates – 100 Cars



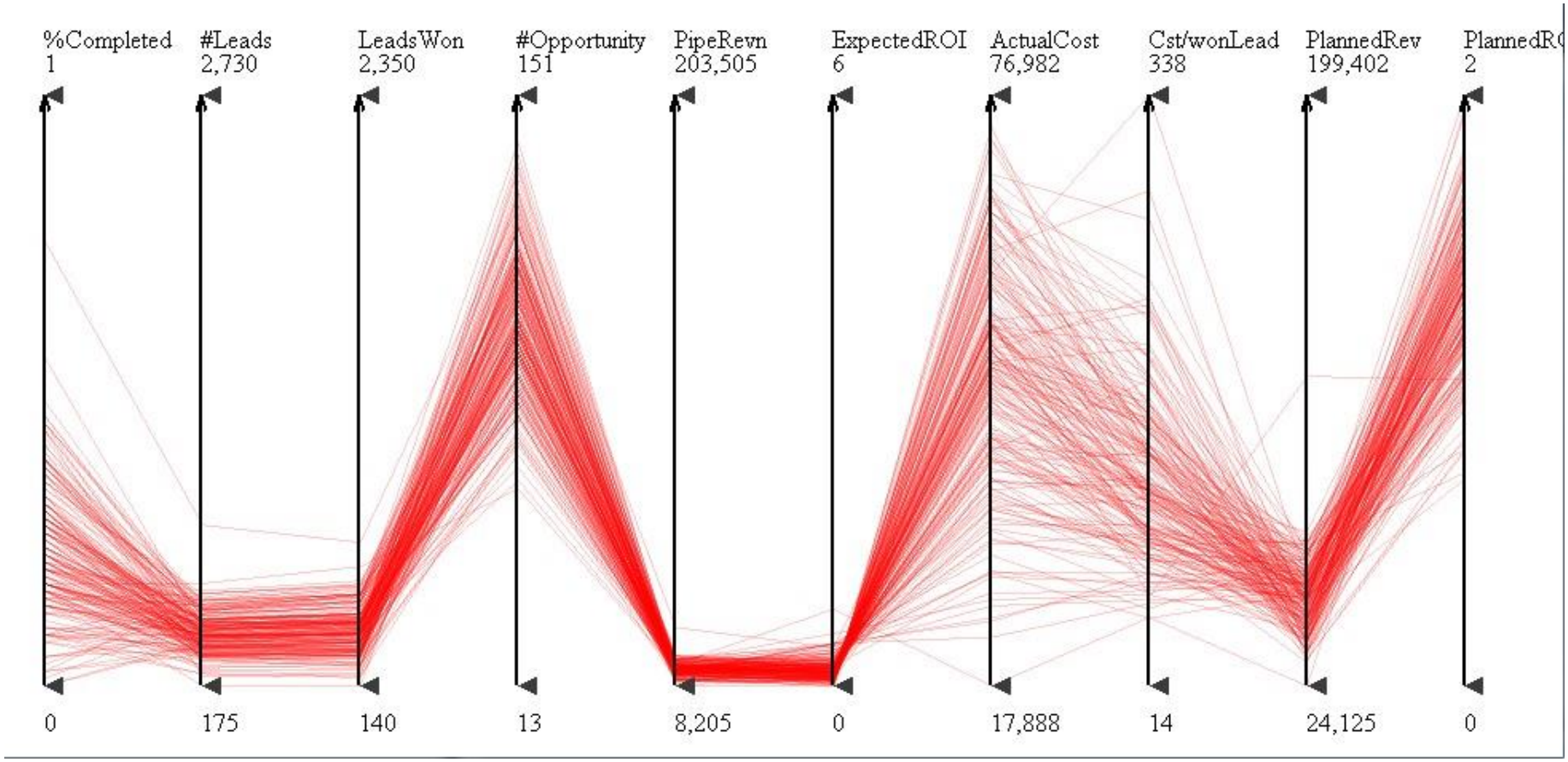
Grouping the cars into sub-populations

- this is called *clustering*
- an be automated or interactive (put the user in charge)

Interactive Clustering With parallel Coordinates

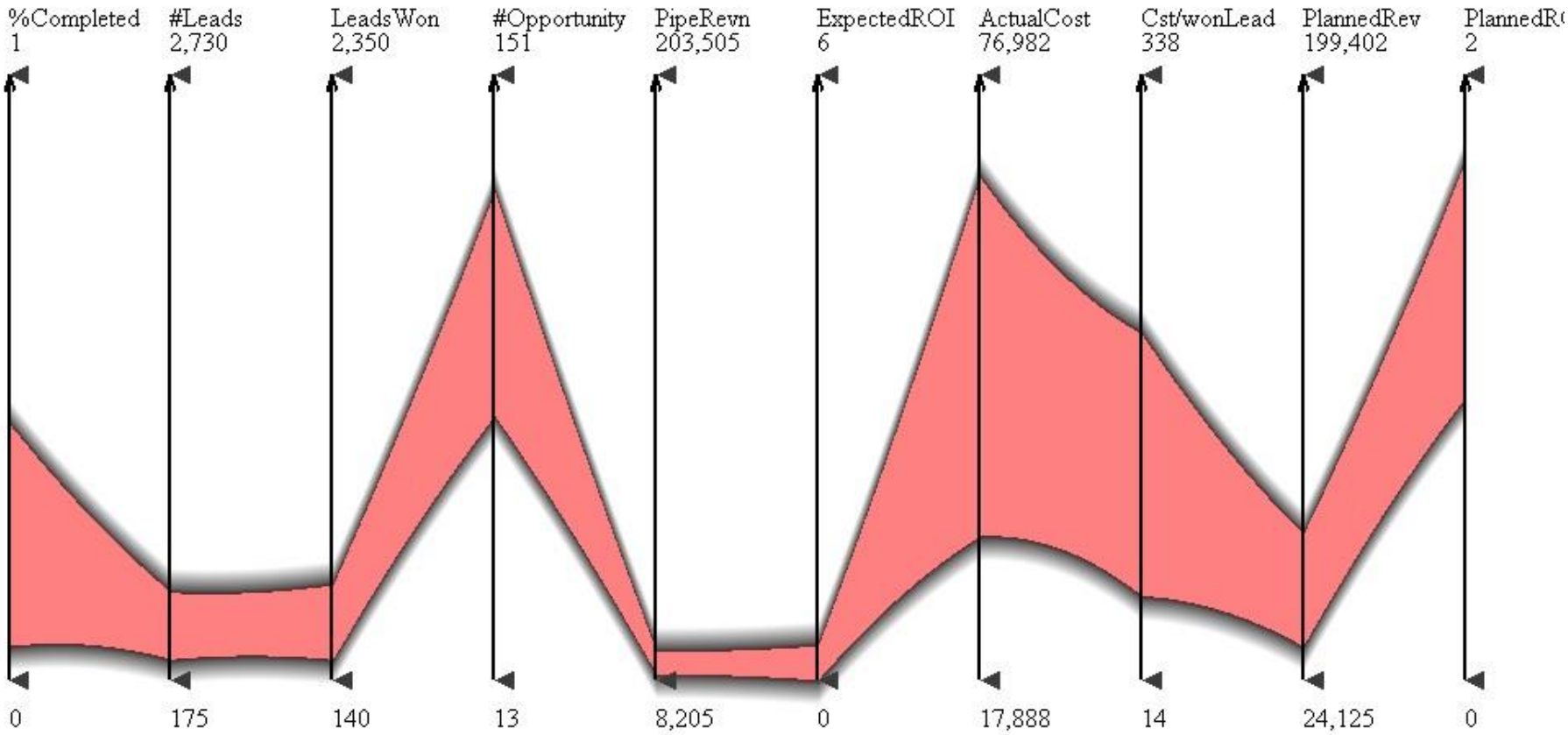


Illustrative Abstraction



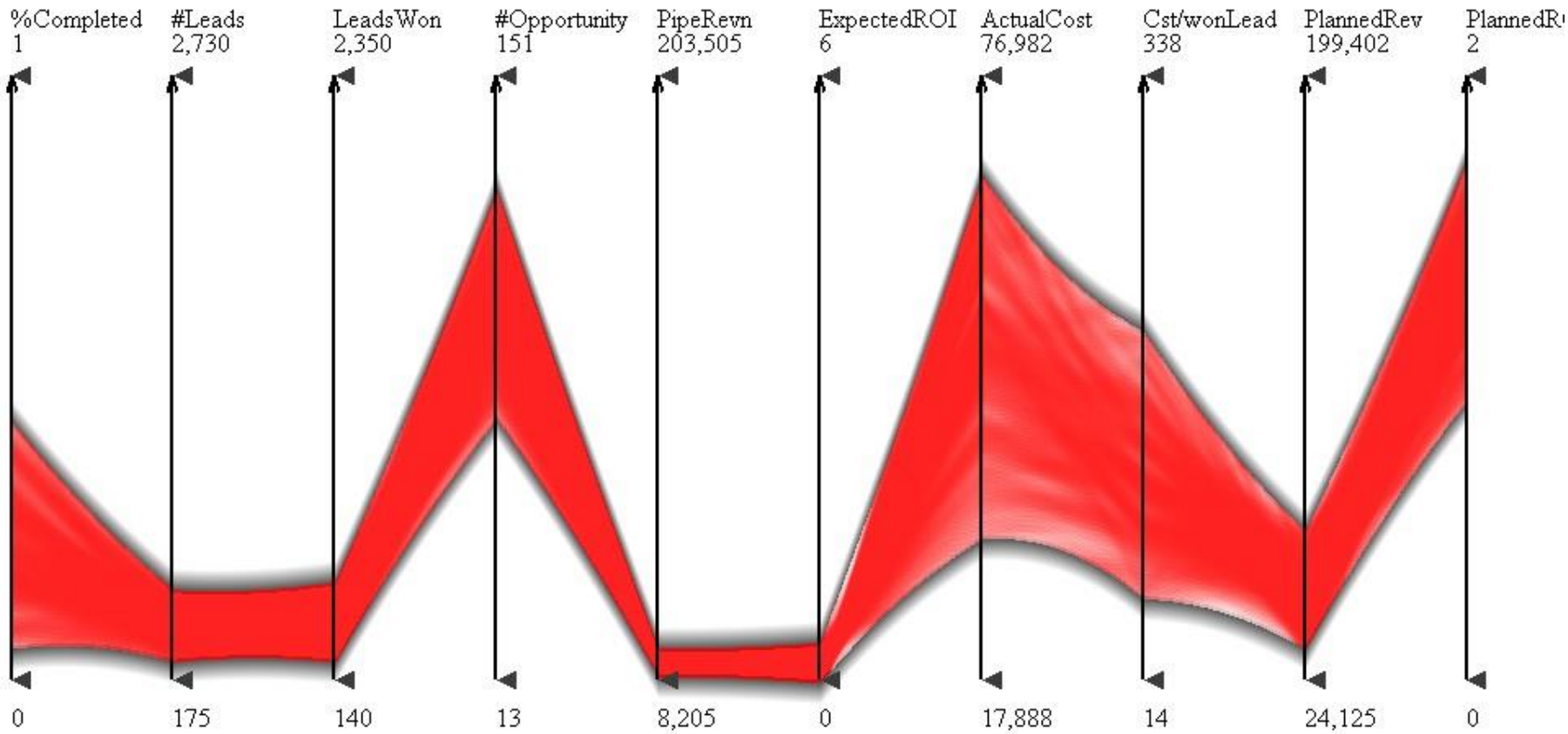
individual polylines

PC With Illustrative Abstraction



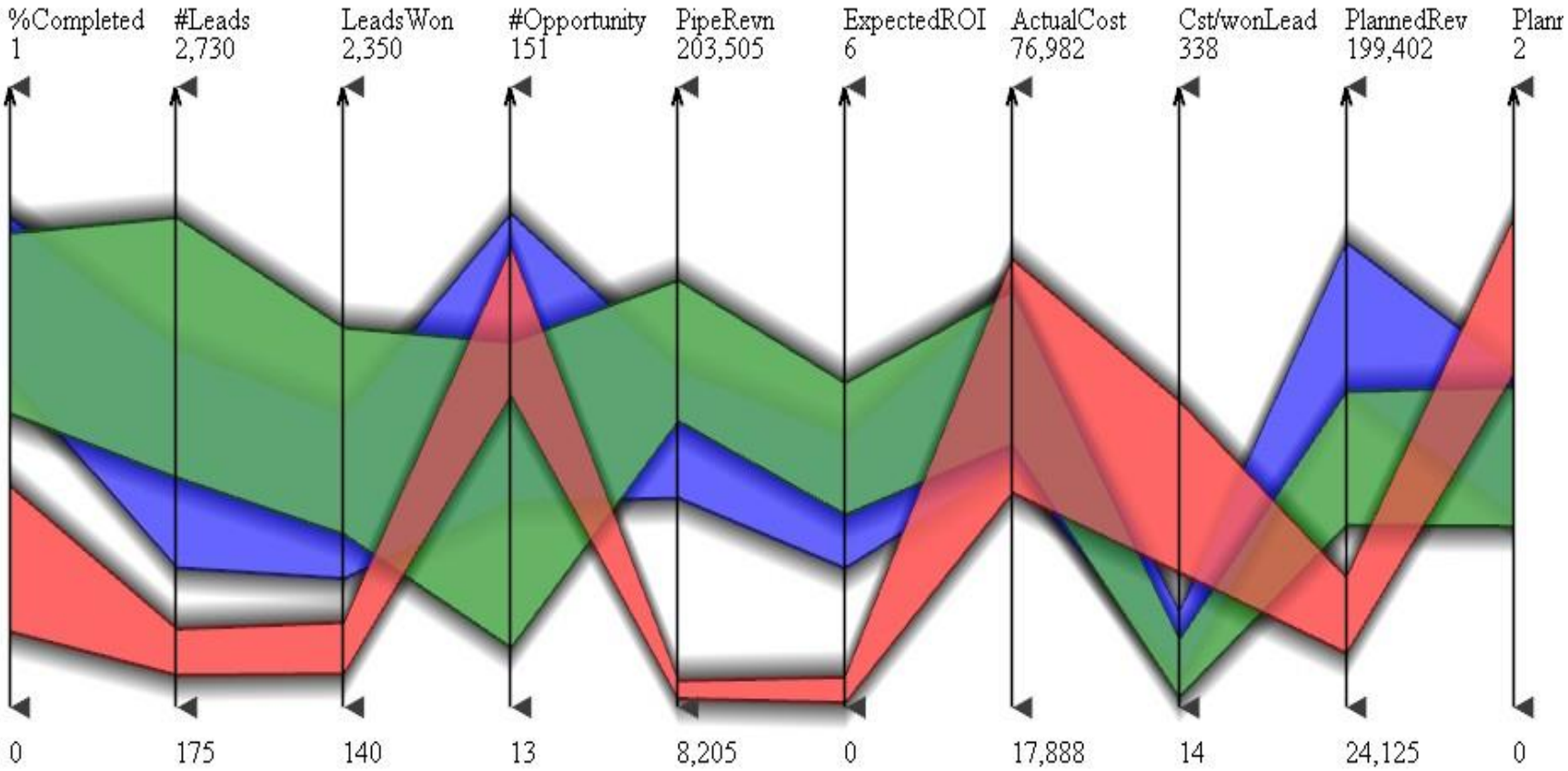
completely abstracted away

PC With Illustrative Abstraction



blended partially

PC With Illustrative Abstraction



all put together – three clusters

Story Telling with Data

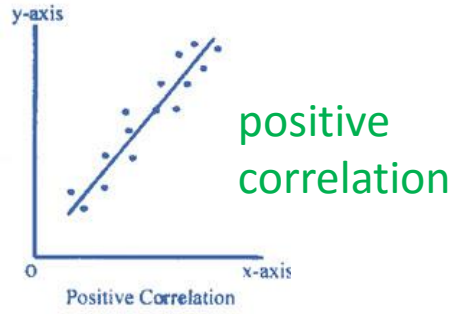
Also called *data narratives*

- parallel coordinates are well suited for this
- next is an example from the business world

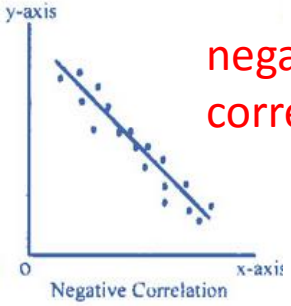
Interlude – Correlation

Our example make use of *correlation*

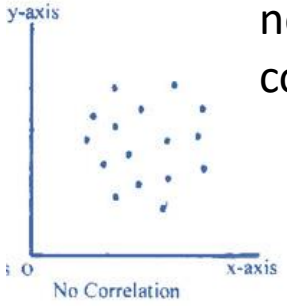
- correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together
- a positive correlation indicates the extent to which those variables increase or decrease in parallel
- a negative correlation indicates the extent to which one variable increases as the other decreases



positive correlation



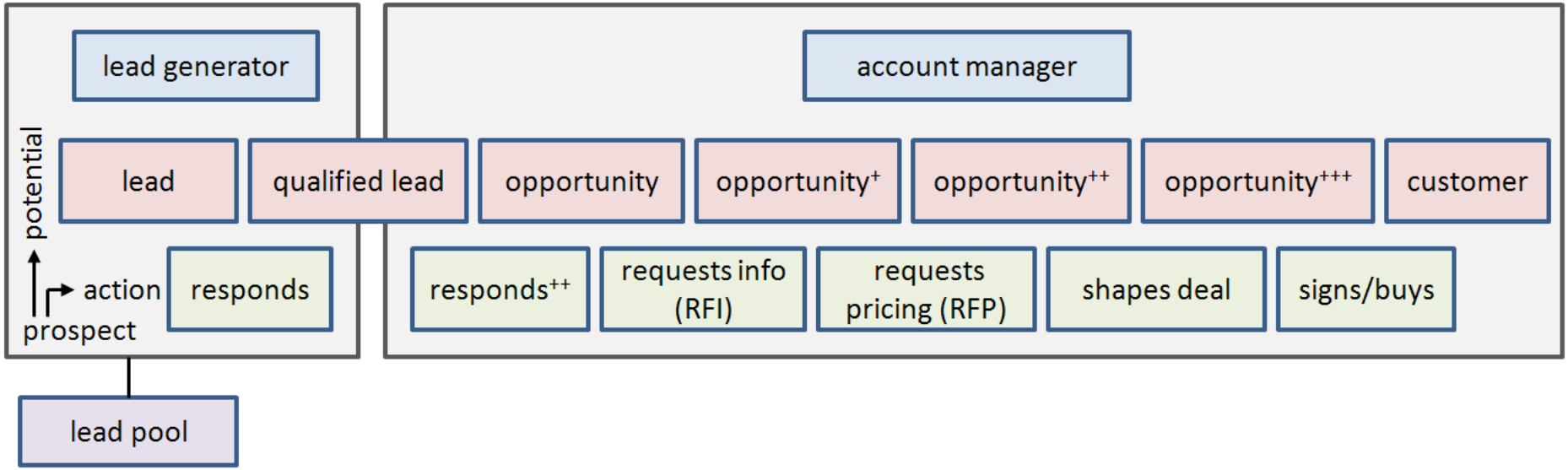
negative correlation



no correlation

spatial proximity representation

Background – Anatomy of a Sales Pipeline



The Setup

Scene:

- a meeting of sales executives of a large corporation, Vandelay Industries

Mission:

- review the strategies of their various sales teams

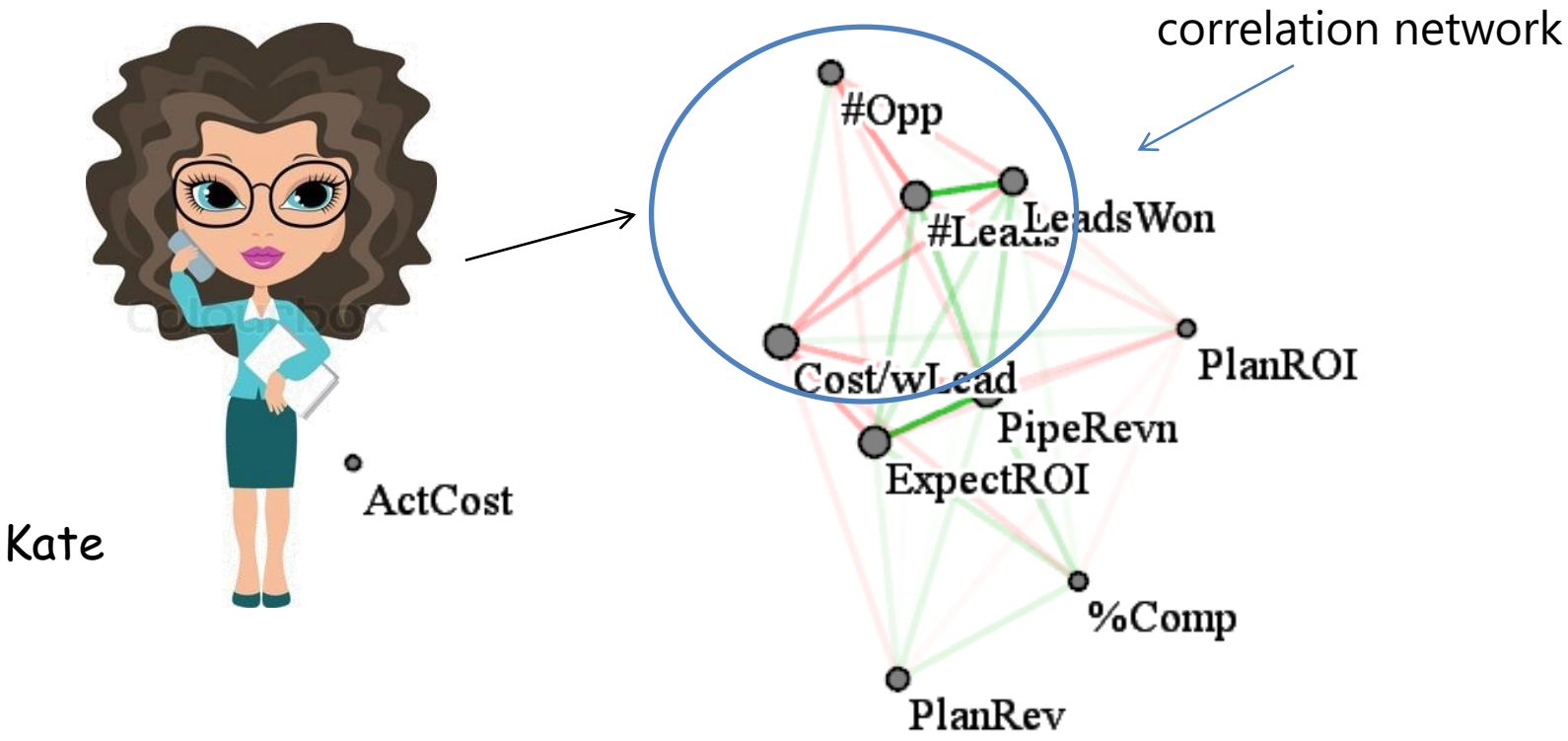
Evidence:

- data of three sales teams with a couple of hundred sales people in each team

Kate Explains It All

Meet Kate, a sales analyst in the meeting room:

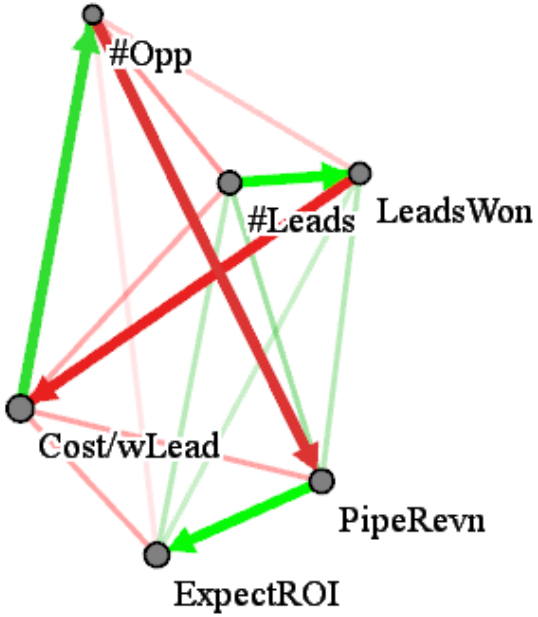
"OK…let's see, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads"

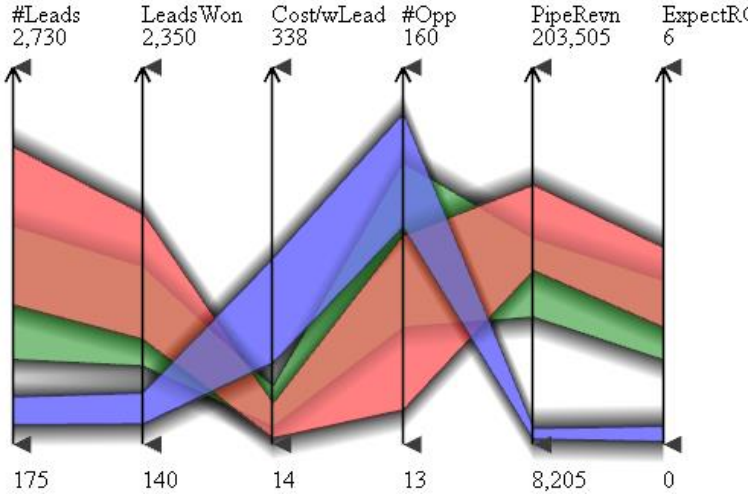


Kate Designs the Narration

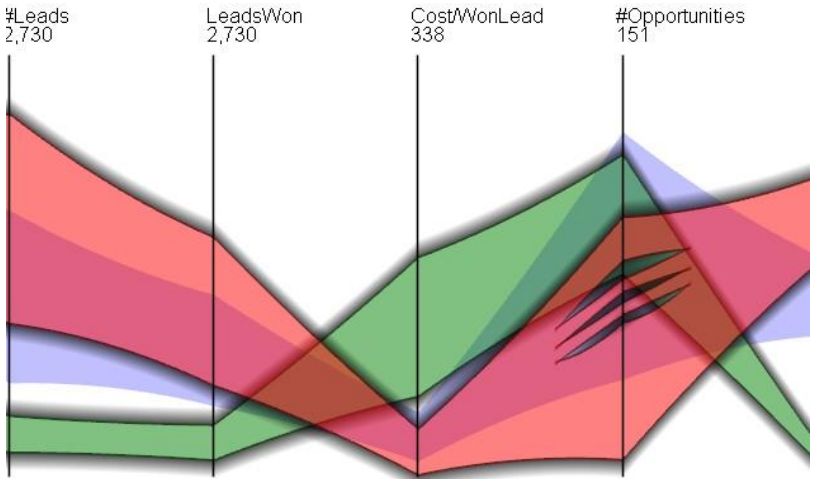
"Let's go and make a revealing route!"

- she uses the mouse and designs the route shown
- she starts explaining the data like a story ...





Further Insight



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”

The Need for Data Reduction

Purpose

- reduce the data to a size that can be feasibly stored
- reduce the data so a mining algorithm can be feasibly run

Alternatives

- buy more storage
- buy more computers or faster ones
- develop more efficient algorithms (look beyond O-notation)

In practice, all of this is happening at the same time

- but the growth of data and complexities is faster
- and so data reduction is important

Data Reduction

Sampling

- random
- stratified

Data summarization

- binning (already discussed)
- clustering (see future a lecture)
- dimension reduction (see next lecture)

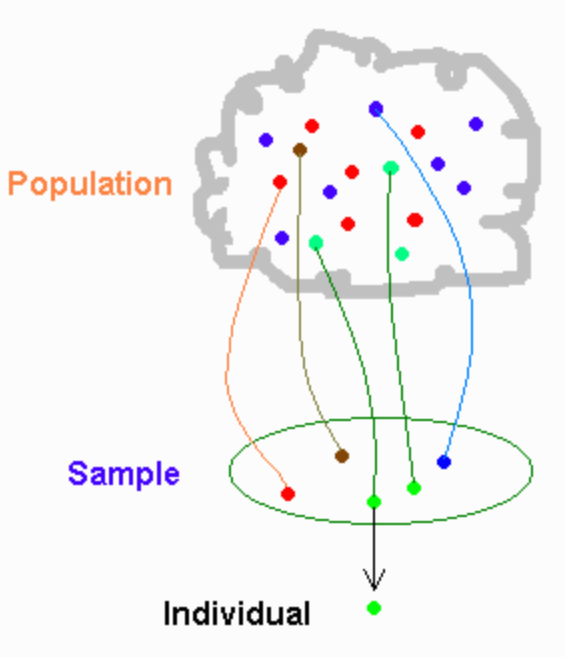
Sampling

The goal

- pick a representative subset of the data

Random sampling

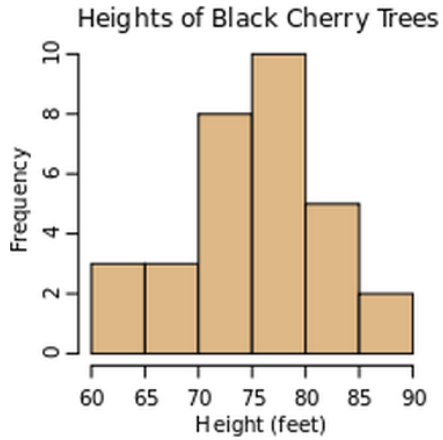
- pick sample points at random
- will work if the points are distributed uniformly
- this is usually not the case
- outliers will likely be missed
- so the sample will not be representative

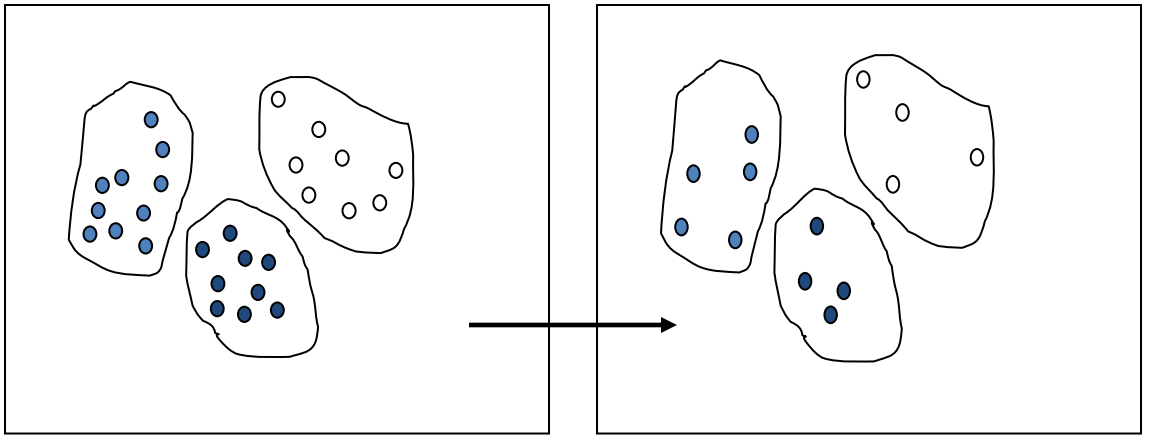


Adaptive Sampling

Pick the samples according to some knowledge of the data distribution

- create a binning of some sort (outliers will form bins as well)
- also called *strata* (stratified sampling)
- the size of each bin represents its percentage in the population
- it guides the number of samples – bigger bins get more samples





sampling rate ~ bin height

sampling rate ~ cluster size