36-350 Data Mining

Tom Minka

Day 1

What is data mining?

A different kind of statistics course

- Exploratory data analysis
- Non-parametric methods
- Model fitting *to aid visualization*

Data mining

- Why it exists
- What it is
- Pitfalls
- Motivation for course

Deluge of data

- Business
  - Customer info
  - Register logs
  - Phone call logs
  - Bank transactions
  - Direct sales (amazon.com)
- Government
  - Population, crime
  - Employment, economy
- Science & Medicine
  - Astronomy surveys
  - Remote sensing
  - Neurological activity
  - Gene expression
  - Adverse drug reactions

Register logs

Customer	Bread	Milk	Eggs	Cereal	Coupon?
14		X		X	
15	X	X	X		X
16		X		X	
17	X		X		X
18	X	X			

Phone call detail

From	To	Date	Time	Length
555-5478	555-1280	1/3/99	10:50	5
555-2387	555-5478	1/6/99	03:35	17
555-5478	555-1280	1/7/99	14:51	25
555-4387	555-0902	1/16/99	7:17	9
555-0902	555-2387	1/22/99	12:06	3

What is this data used for?

- Usually, not for much
- Collected for billing and inventory
- Filed away in “data tombs”
- Undervalued, conveniently ignored
- Waiting for someone to analyze it

Business applications

- Targeted marketing
  - Who is likely to buy this product?
- Product/service recommendations
  - People who buy X often like Y as well
  - Grocery store coupons
- Assess loyalty of different groups
  - Who is likely to switch (“churn”)?

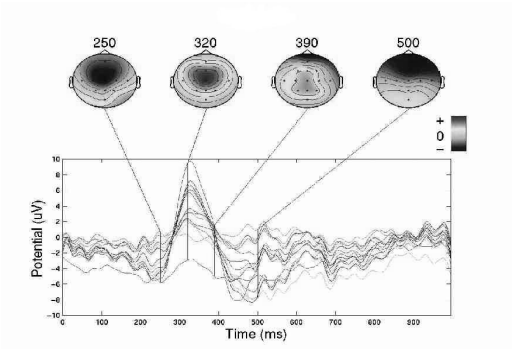
Business applications cont’d

- Assess credit risk, insurance risk of different groups
  - Who will repay a loan?
- Predict sales
  - Which items should we stock? In which stores?
- Identify fraud, inefficient practices
  - Anomalies, unexpected patterns

Scientific & medical applications

- New, interesting astronomical objects
- New geological formations, mineral deposits
- New insights into brain function, gene function
- Unexpected drug reactions

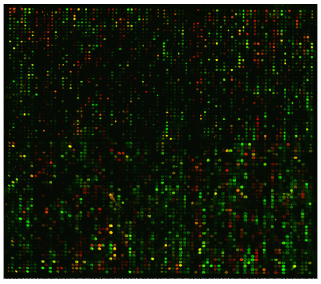
EEG traces



Gene expression microarrays

Red: above baseline

Green: below baseline



Problems with data

Most data is too complex for conventional tools:

- Too fragmented (phone calls)
- Too complex to model directly (neurons, genes, sky images)
- Too much spurious phenomena (register logs, drug reactions)
- Variation in amount of data per customer

Problems with statistical/machine learning algorithms

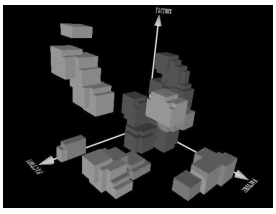
- Too focused, blind
- Require predefined goal, modeling strategy
- Precise, automated answer to a specific question
  - may overlook crucial aspects of data

Data mining

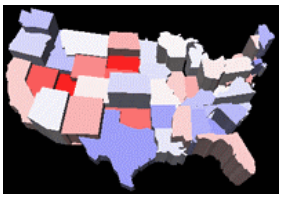
Utilization of statistics/machine learning methods within an exploratory framework

Emphasizes:
- Visualization
- Exploratory data analysis
- Non-parametric methods
- Serendipity

Visualization



VisualMine



MineSet

Data mining process

- Iteratively defocus one part to focus on another
- Defocus:
  - abstract data values (e.g. rounding, “large” vs. “small”)
  - summarize batches (e.g. median, quartiles)
- Focus:
  - Subdivide (trees, clustering)
  - Apply a model (curve fitting)
- Always keep your options open

Data mining

- Requires creativity
- Requires knowledge of option space
- Cannot be automated

Fundamental difficulties

- Poor data quality
  - Collected for different purposes
  - Needs “cleaning”
- Data is purely observational
  - Biased data selection
    - e.g. only people who shop here
  - True causes may be missing
  - Prolonged exploration may lead to overfitting

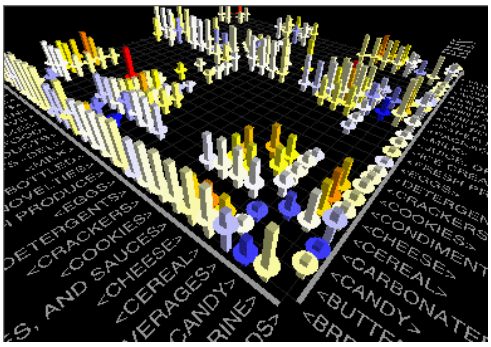
John W. Tukey

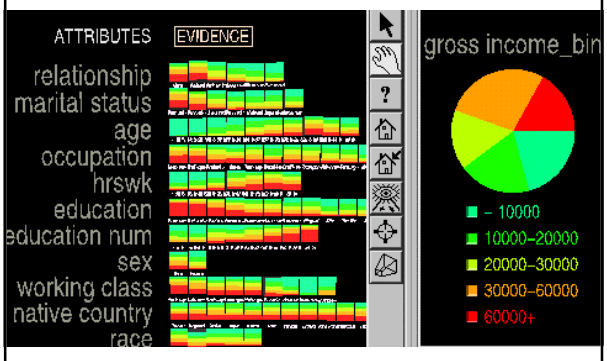
- Invented “Exploratory Data Analysis” (1977)
- Viewed data analysis as a unique field
  - Bigger than statistics, requiring more general methods
  - Esp. non-mathematical methods
- Some methods incorporated into statistics classes (stemplot, boxplot, two-way plot, etc)

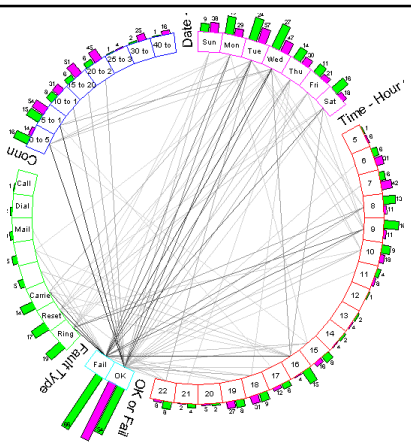
Why exploration is hard

- Modern software makes it easy to make lots of plots
- It does not give guidance about what you should plot
- Many (if not most) plots are useless, confusing
- Not same problem as graphic art

More data = better plot?





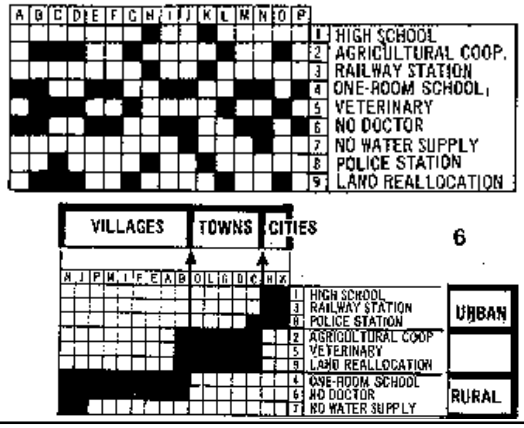


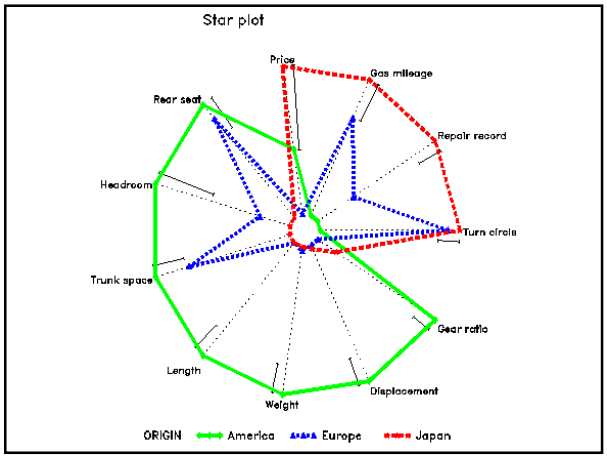
Why exploration is hard

It is hard to make a plot that:

- Isn't swamped by irrelevant influences, random variation
- Isn't misleading
- Clearly and fairly shows the size of an effect
- Answers an interesting question

Some good plots





Why so good?

Variables ordered by similarity, using linear model

Also:

- Variables standardized
- Error bars

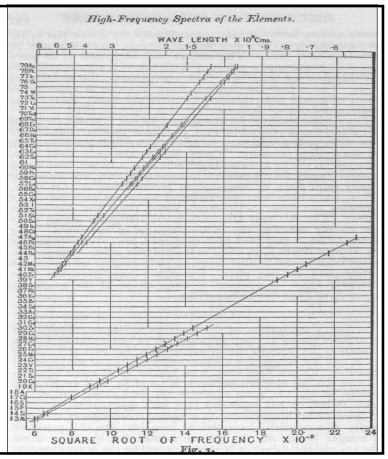
Good plots are guided by models

- Models suggest how data should be presented, simplified
- Models consolidate what we have seen in data, so that we may see farther
- Models define anomalies
- Even oversimplified models are useful

Moseley's plot:

• suggested a way of numbering the elements

• focused attention on deviations from the model



Balance

To do good visualization, you need modeling

To do good modeling, you need visualization

Structure of course

Wide variety of visualization and modeling techniques

- Overview of statistics and machine learning

Progression: Low dimension to high dimension

Alternate: Continuous vs. discrete values

Visualization and modeling together