11A – DATA MINING

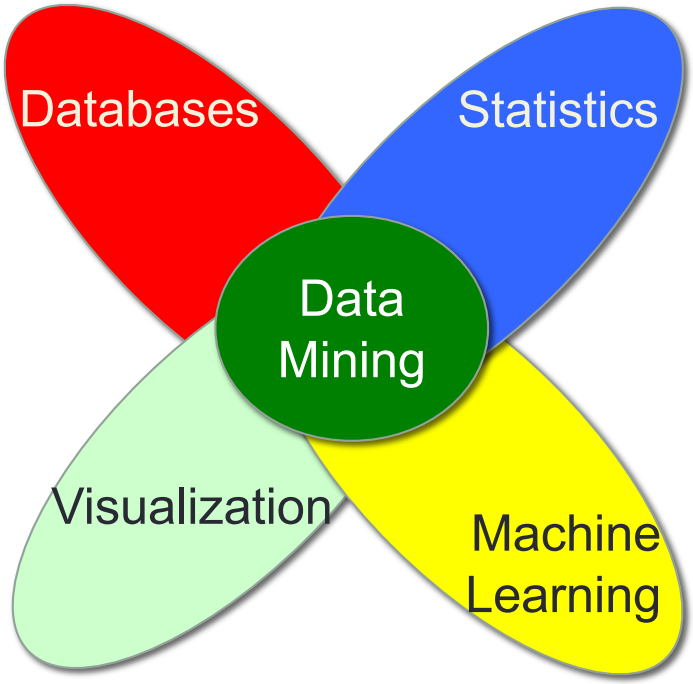
CS 1656

Introduction to Data Science

Alexandros Labrinidis – http://labrinidis.cs.pitt.edu
University of Pittsburgh

Data Mining definition

- What is data mining?
  - Computational process to discover patterns in large data sets



Data Mining definition (cont)

Must produce novel and interesting patterns!

[Source: http://dilbert.com/strips/comic/1996-04-17/]

A bit of history

- Let us go back 160 years ago (1854) in London, England
  - On 31 August 1854, after several other outbreaks had occurred elsewhere in the city, a major outbreak of cholera struck Soho.
  - Over the next three days, 127 people on or near Broad Street died. In the next week, three quarters of the residents had fled the area.
  - By 10 September, 500 people had died and the mortality rate was 12.8 percent in some parts of the city. By the end of the outbreak, 616 people had died.

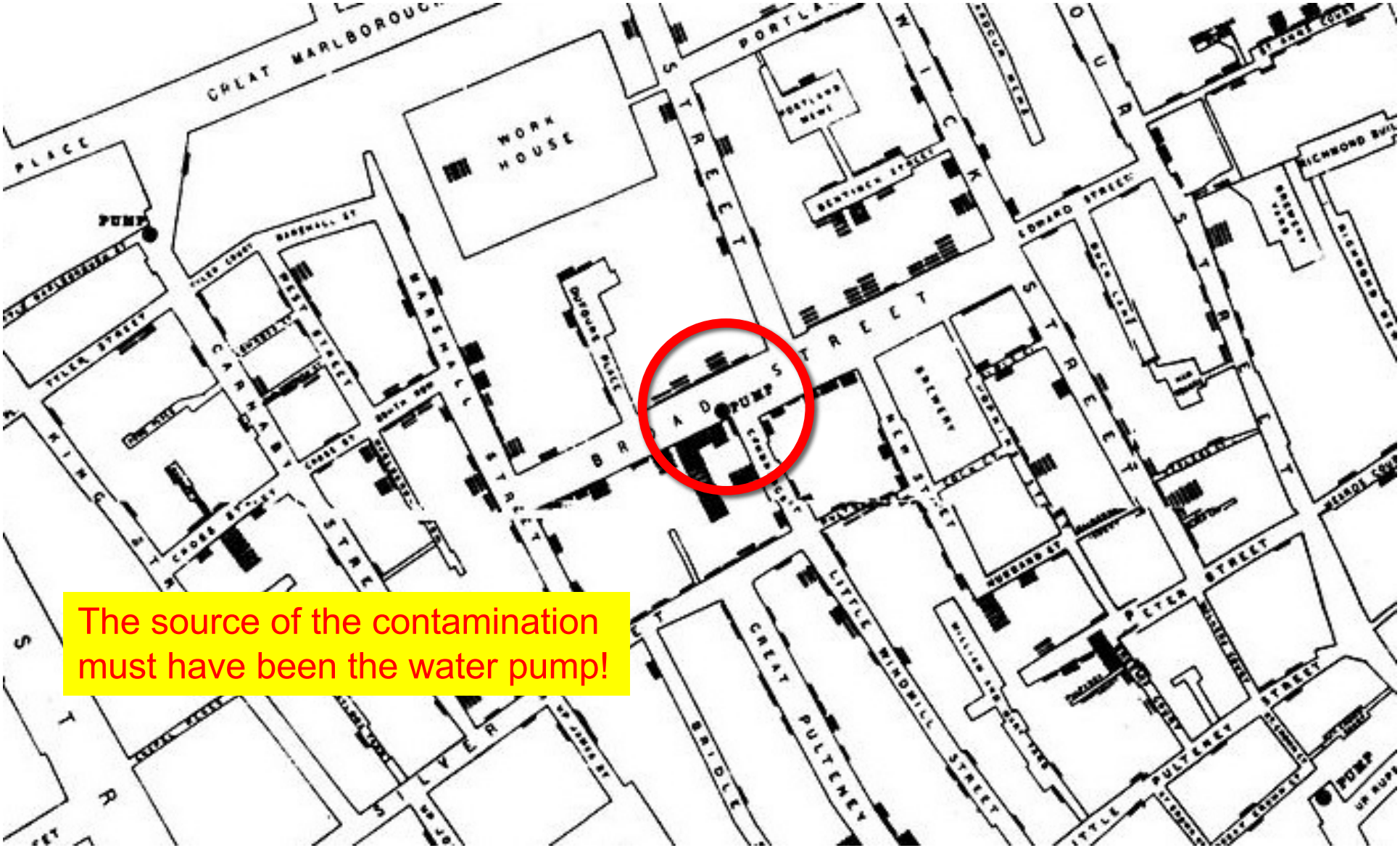
[Source: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak]

1854 Cholera Outbreak

- John Snow (a physician) wanted to investigate cause
  - Was skeptic of the *Miasma theory*, of "bad air"
  - Note that the germ theory was not created until 1861 by Louis Pasteur
  - Snow studied spread of disease
  - Made a map of fatalities

Map of Soho with fatalities

Map of Soho with fatalities (zoom)



In Snow's own words

On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street...

With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump water from Broad Street, either constantly or occasionally...

The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-mentioned pump well.

I had an interview with the Board of Guardians of St James's parish, on the evening of the 7th inst [September 7], and represented the above circumstances to them. In consequence of what I said, the handle of the pump was removed on the following day.

—John Snow, *letter to the editor of the Medical Times and Gazette*

DATA MINING TASKS

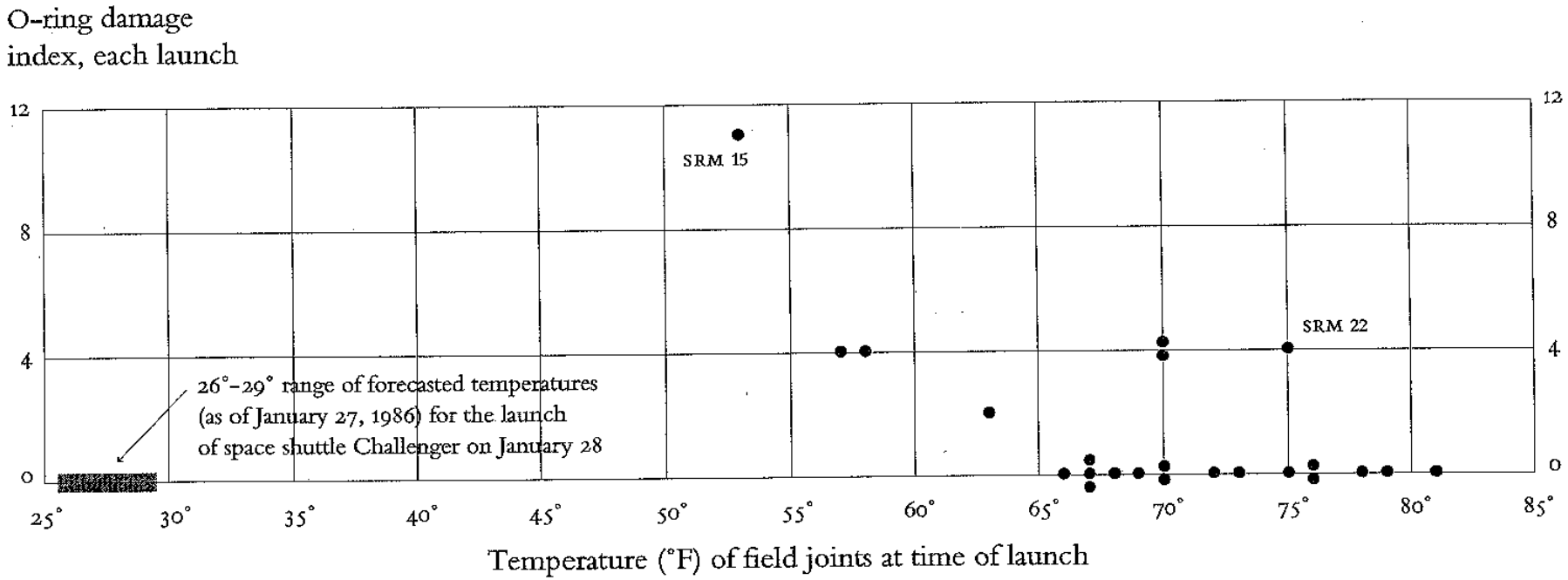
Typical Data Mining Tasks

- Anomaly Detection
- Association Rule Learning
- Clustering
- Classification
- Regression
- Summarization

Source: http://en.wikipedia.org/wiki/Data_mining

Anomaly Detection

- Anomaly Detection (Outlier/change/deviation detection)
  The identification of unusual data records, that might be interesting or data errors that require further investigation.



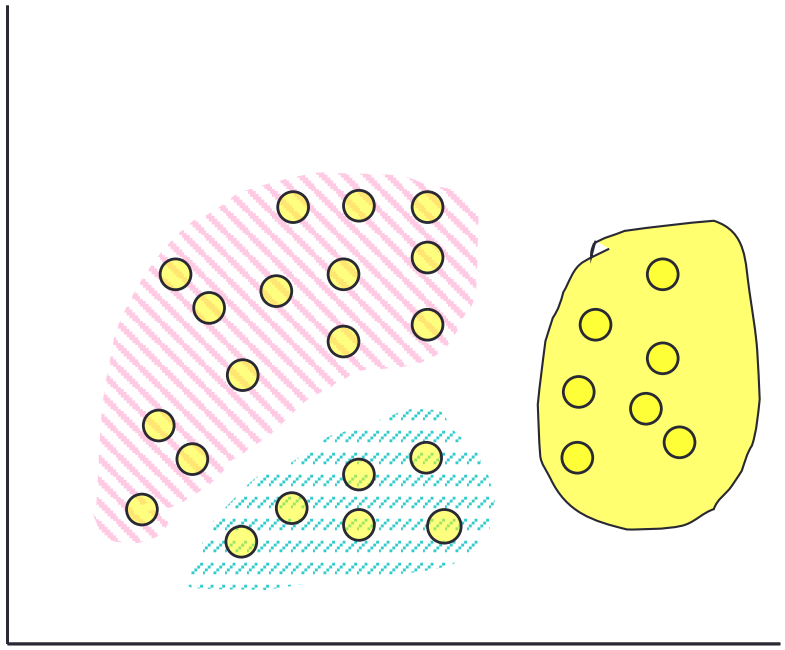
Source: Visual and Statistical Thinking: Displays of Evidence for Making Decisions, Edward Tufte, 1997

Association Rule Learning

- Association rule learning (Dependency modeling)
  Searches for relationships between variables.
  - For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes.
  - This is sometimes referred to as market basket analysis.

Clustering

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
  - Members of a cluster should be more “alike” among each other, than to members of other clusters
  - Clustering is one type of **unsupervised learning**



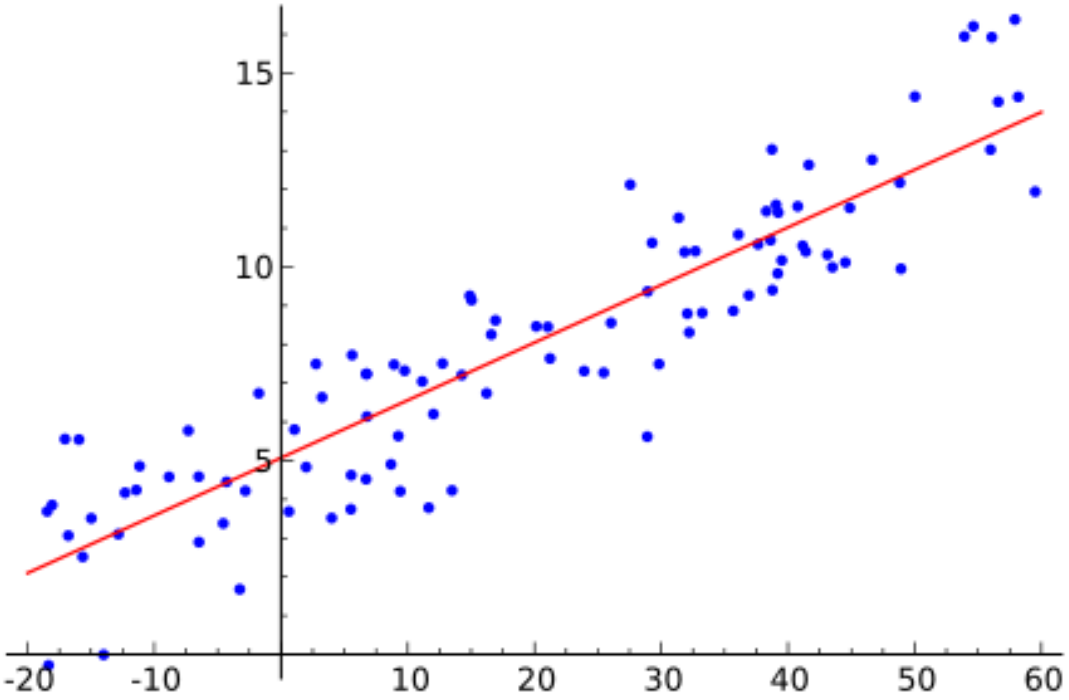
Classification

- Classification – is the task of generalizing known structure to apply to new data.
  - In other words: learn a method to predict a label for new data from pre-labeled (classified) data
  - Classification is one type of **supervised learning**

- Examples:
  - an e-mail program classifies an e-mail as "legitimate" or as "spam".
  - a lender classifies its customers as credit-worthy or credit-risky.
  - a credit card company identifies fraudulent transactions.
  - a phone company identifies which customer would abandon contract for another carrier.
  - a security agency identifies potential *evil-doers*.
  - personalized medicine – will drug work for specific patient?

Regression

- Regression – attempts to find a function which models the data with the least error.



Source: http://en.wikipedia.org/wiki/Regression_analysis

Summarization

- Summarization –
  providing a more compact representation of the data set, including **visualization** and report generation.

- Examples:
  - Document summarization (e.g., snippets in Gmail)
  - Choosing one representative member from each cluster
  - Choosing a few representative data points through *sampling*
- Question:
  - What would be a trivial (but simple) summarization of a set of numbers?
- Answer:
  - computing the average or the median