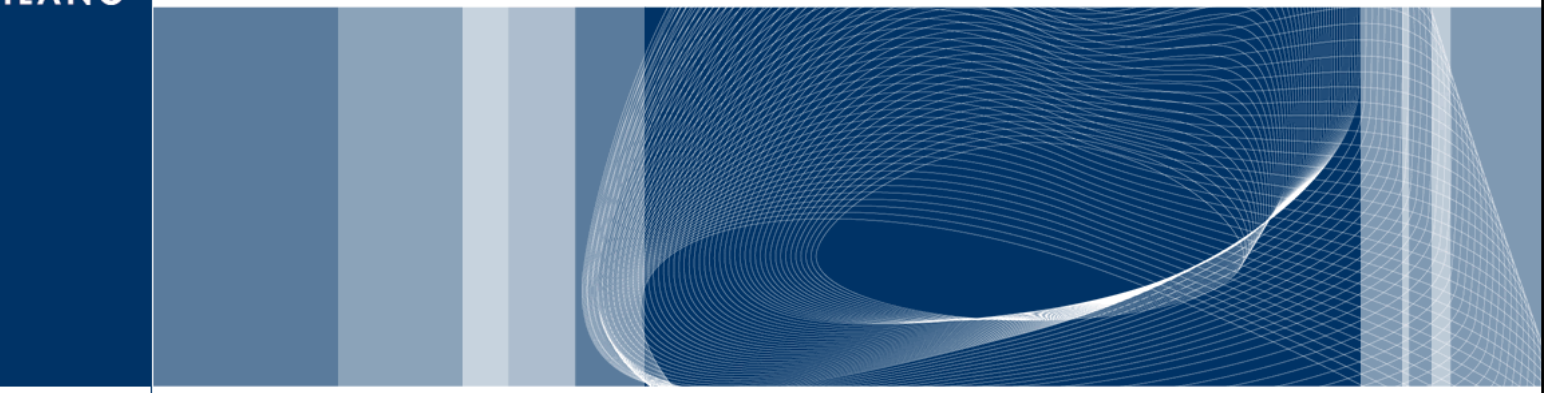
POLITECNICO DI MILANO



Data Mining

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

Prof. Pier Luca Lanzi

Motivations

Why Data Mining?

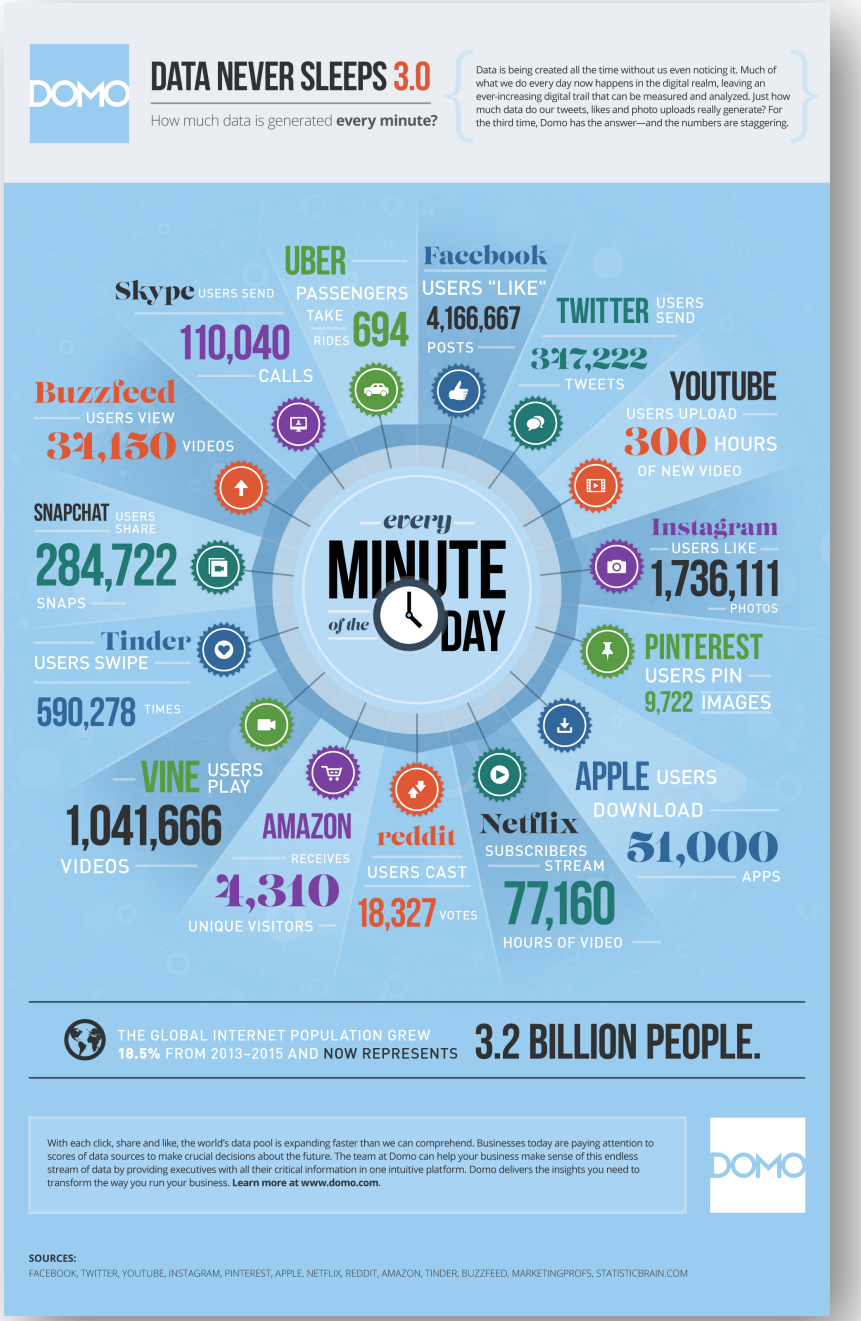
"Necessity is the mother of invention"

Explosive Growth of Data

Pressing need for the automated analysis of massive data

Emerged in the late 1980s

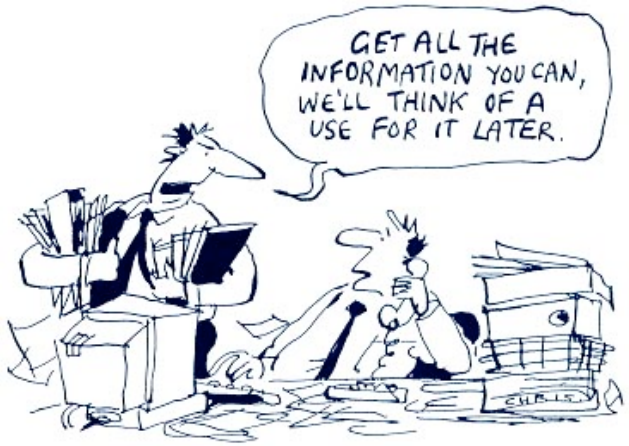
Major developments in the mid 1990s



http://www.iflscience.com/technology/amount-data-internet-generates-every-minute-crazy/

Evolution of Technology

- 1960s: data collection, database creation, & network DBMS
- 1970s: relational data model, relational DBMS implementation
- 1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.); application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s: data mining, data warehousing, multimedia databases, and Web databases
- 2000s: stream data management and mining, web technology (XML, data integration), global information systems
- 2010s: social networks, NoSQL, unstructured data, etc.



http://launchhack.com/content/25-cartoons-give-current-big-data-hype-perspective/

What is the Commercial Viewpoint?

- Huge amounts of data is being collected and warehoused everyday
  - Web data, e-commerce
  - Purchases at department stores
  - Bank/Credit Card transactions

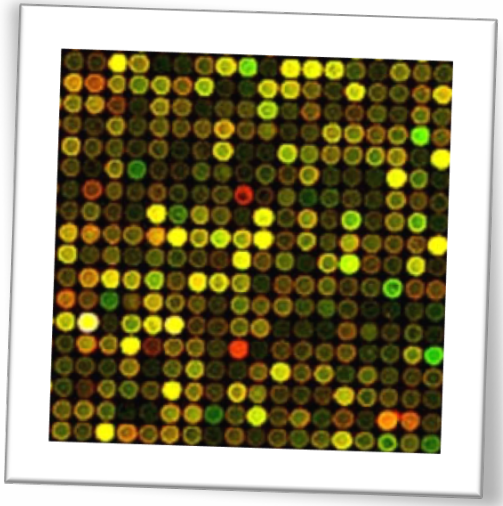
- Computers have become cheaper and more powerful

- Competitive pressure is strong to provide better, customized services (e.g., CRM or Customer Relationship Management)

- Poor data across businesses and the government costs huge amount of money

What is the Scientific Viewpoint?

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



Examples

- Customer attrition
  - Given customer information for the past months
  - Predict who is likely to attrite next month, or estimate customer value
- Credit assessment
  - Given a loan application
  - Predict whether the bank should approve the loan
- Customer segmentation
  - Given several information about the customers
  - Identify interesting groups among them
- Community detection
  - Given a social network of users
  - Identify community based on their connections (friendship relation, discussions, etc.)

Big Data?

FINAL FINAL

POLICYFORUM

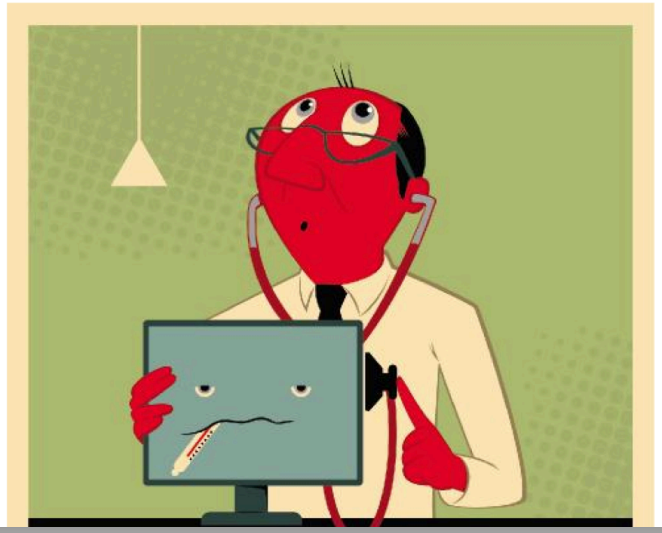
BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw



the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the

http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/

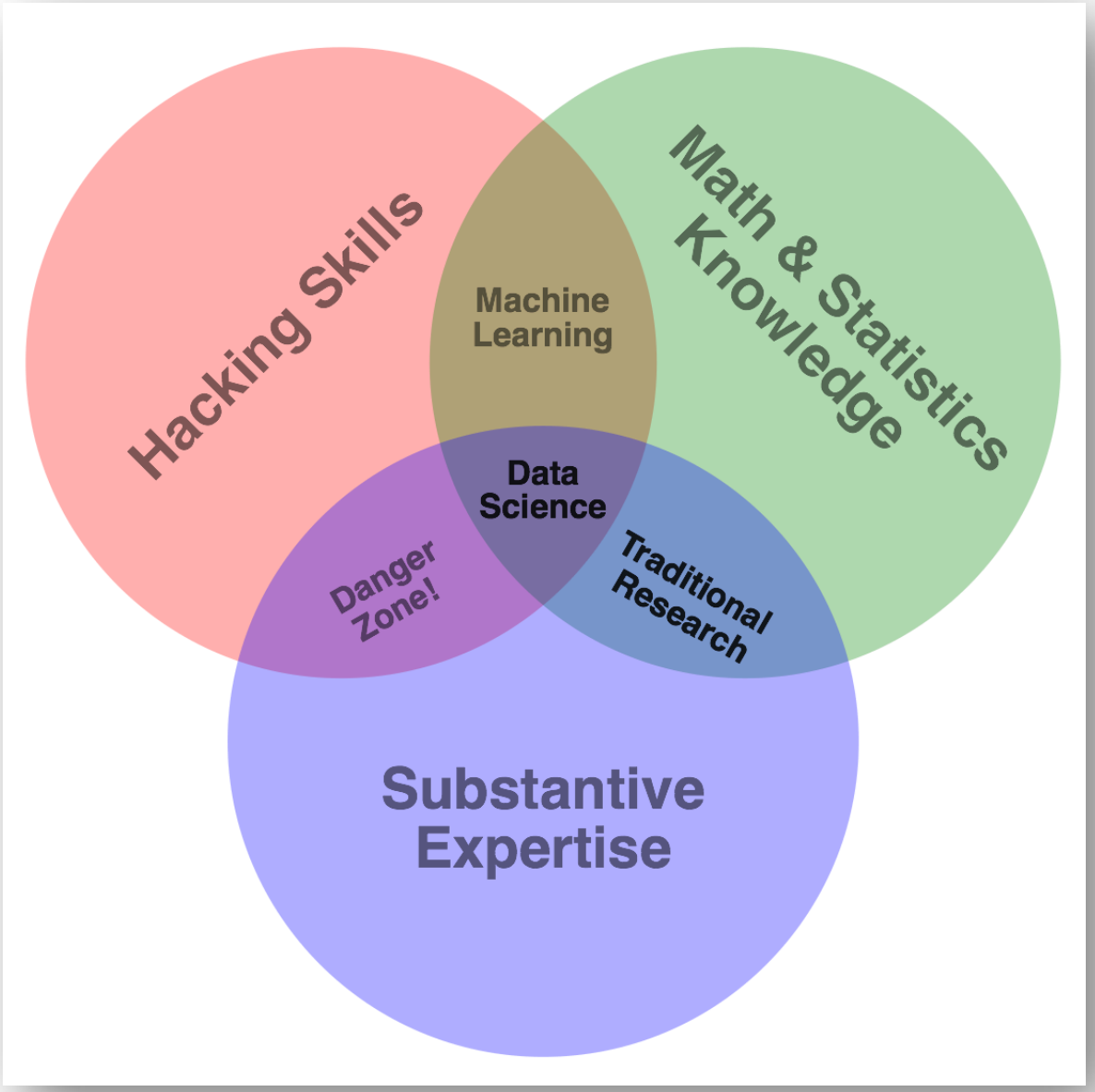
http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/?_r=0

http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf


Dan Ariely
Yesterday
Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...
Unlike · Comment · Share
359

Data Science?

http://wikibon.org/blog/role-of-the-data-scientist/



https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html

Data Mining?

Data Mining

The non-trivial process of identifying
(1) valid, (2) novel, (3) potentially useful,
and (4) understandable patterns in data.

An Example Using Contact Lens Data

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

An example of possible pattern

if astigmatism = yes
and tear production rate = normal
and spectacle prescription = myope
then recommendation = hard

is it a “good” pattern?

How Can We Evaluate a Pattern?

- Is it valid?
  - The pattern has to be valid with respect to a certainty level (rule true for the 86%)
- Is it novel?
  - Is the relation between astigmatism and hard contact lenses already well-known?
- Is it useful? Is it actionable?
  - The pattern should provide information useful to the bank for assessing credit risk
- Is it understandable?

but there is another important question …

was it "worth" finding it? (I mean $ worth)

how much did the search cost?

how much value did it bring

Example of Cost-Based Model Evaluation

- A bank has a predictive model that can identify risky loans with an accuracy of 72%
- Your company develops a model that can improve their performance by 3% reaching an accuracy of 75%
- Is this a good result?
- We might simply evaluate the 3% of improvement but giving out loans has a cost that depends on the type of error we make

Example of Cost-Based Model Evaluation

- Predictive accuracy in this case is defined as

$$accuracy = \frac{number\ of\ correct\ predictions}{number\ of\ applications}$$

- True Positives, true negatives
  - Safe and risky loans predicted as safe and risky respectively
- False Positive Errors
  - We accept a risky loan which we predicted was safe
  - We are likely not to get the money back (let's say on average 30000 euros)
- False Negative Errors
  - We don't give a safe loan since we predicted it was risky
  - We will loose the interest money (let's say on average 10000 euros

Example of Cost-Based Model Evaluation

- Original Model
  - 1576 false positives and 1224 false negatives
  - Total cost is 59525443
- Our Model
  - 1407 false positives and 1093 false negatives
  - Total cost is 53147717 (more than 6 millions saved)
- What if we can change the way our model makes mistakes?
  - 1093 false positives and 1407 false negatives
  - Total cost becomes 46852283 (more than 12 millions saved)

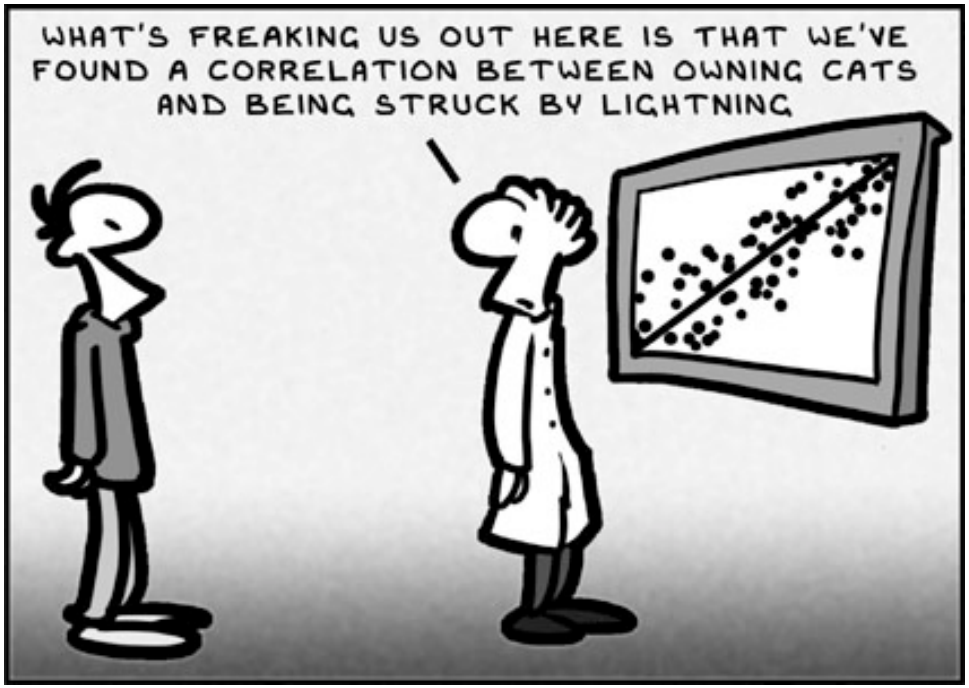
What is the General Idea?

- Build computer programs that navigate through databases automatically, seeking regularities or patterns

- There will be problems
  - Most patterns are banal and uninteresting
  - Most patterns are spurious, inexact, or contingent on accidental coincidences in the particular dataset used
  - Real data is imperfect: Some parts will be garbled, and some will be missing

- Algorithms need to be robust enough to cope with imperfect data and to extract regularities that are inexact but useful

Pitfalls

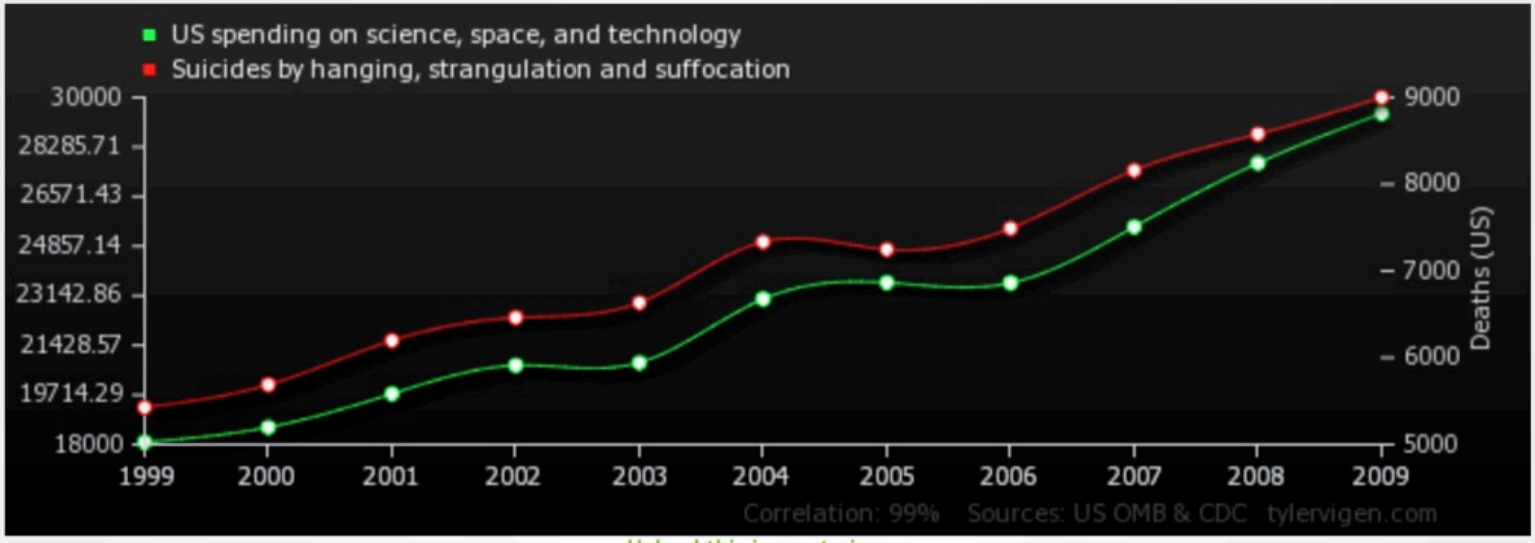


http://launchhack.com/content/25-cartoons-give-current-big-data-hype-perspective/

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



Upload this image to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

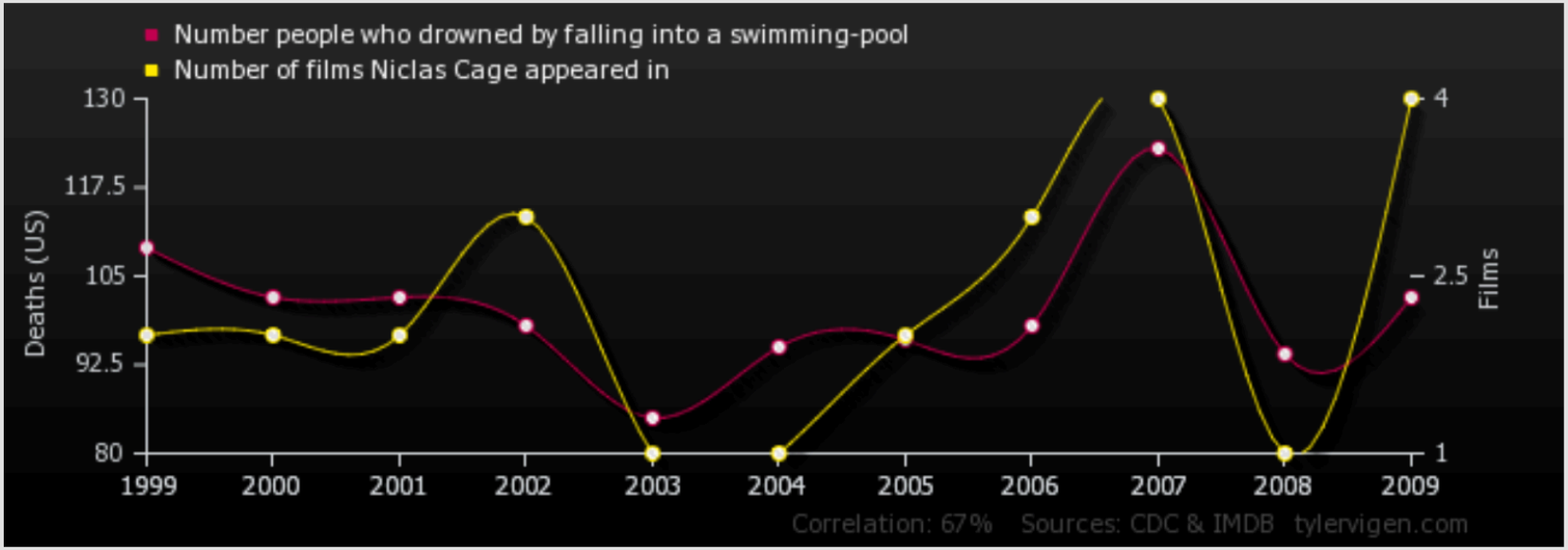
Correlation: 0.992082

http://www.tylervigen.com

Number people who drowned by falling into a swimming-pool

correlates with

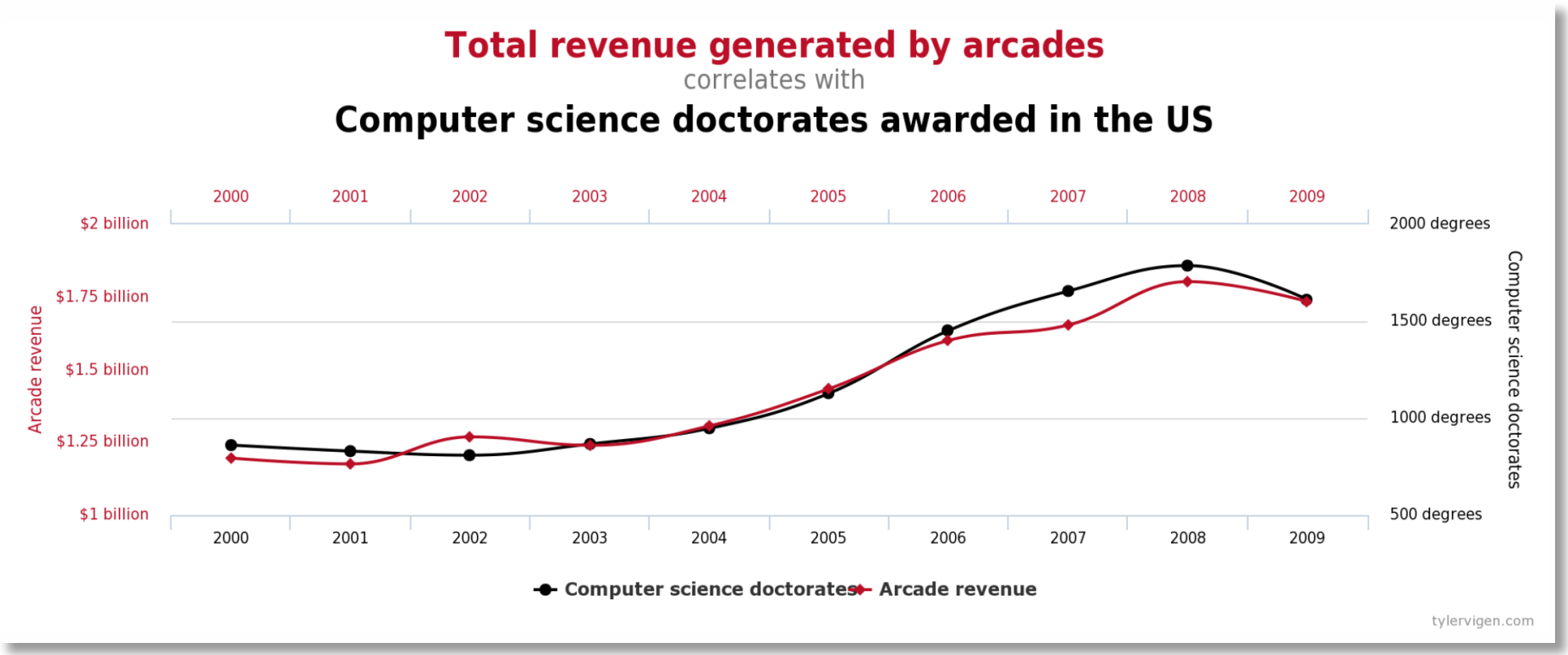
Number of films Nicolas Cage appeared in



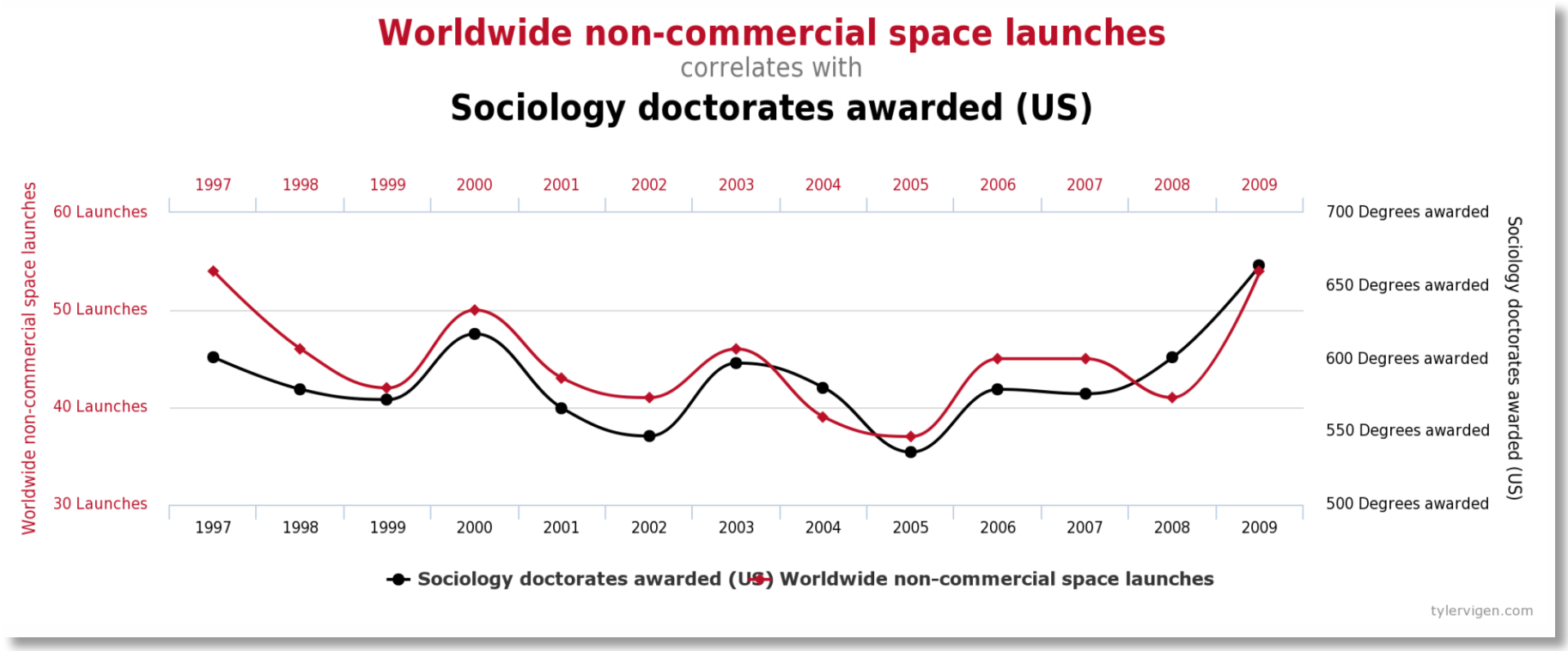
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool *Deaths (US) (CDC)*	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in *Films (IMDB)*	2	2	2	3	1	1	2	3	4	1	4

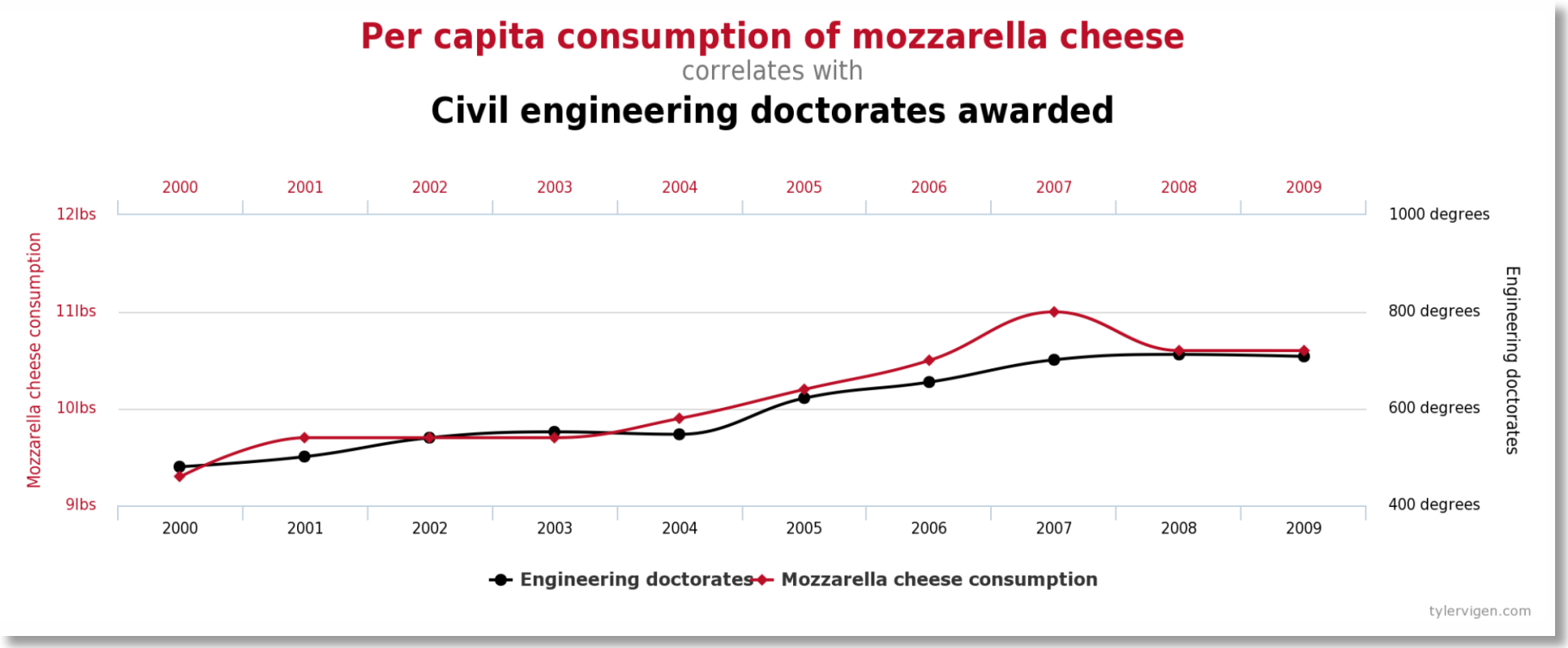
Correlation: 0.666004

http://www.tylervigen.com

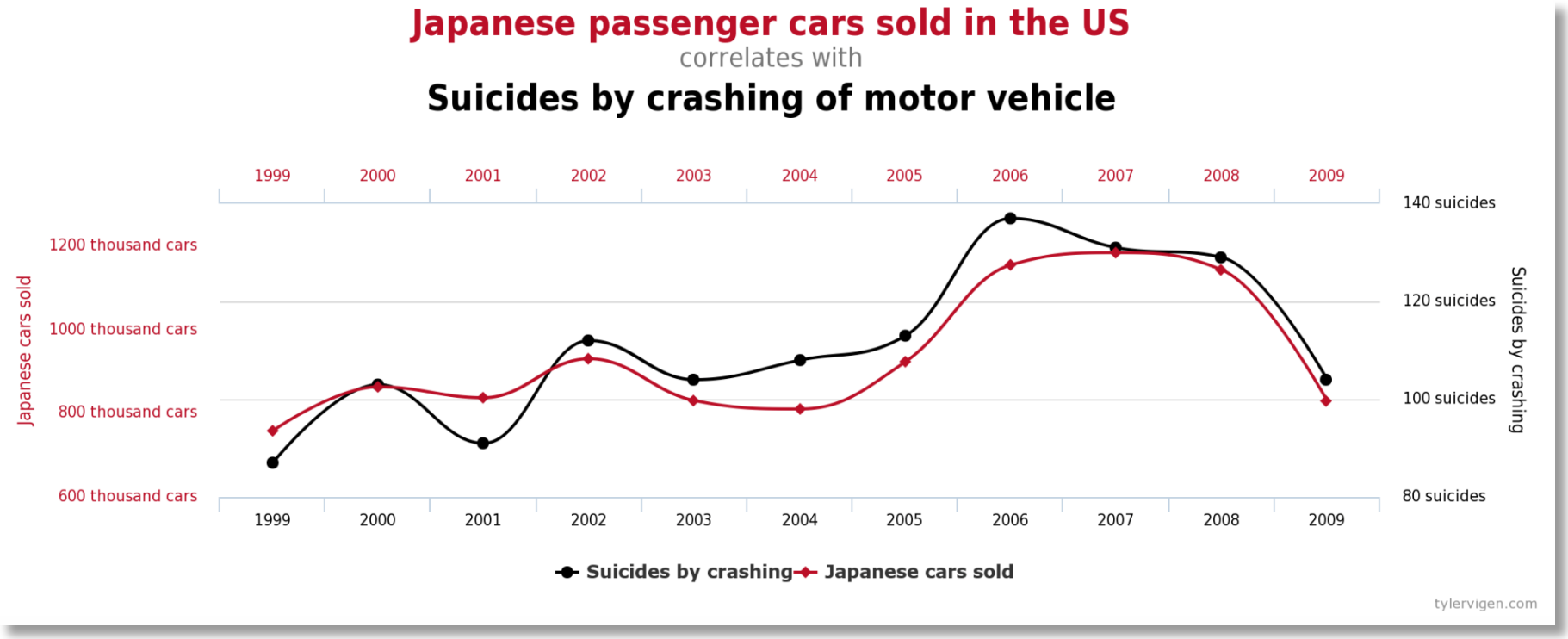


http://www.tylervigen.com


Worldwide non-commercial space launches
correlates with
Sociology doctorates awarded (US)
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
60 Launches
50 Launches
40 Launches
30 Launches
Worldwide non-commercial space launches
700 Degrees awarded
650 Degrees awarded
600 Degrees awarded
550 Degrees awarded
500 Degrees awarded
Sociology doctorates awarded (US)
Sociology doctorates awarded (US)
Worldwide non-commercial space launches
tylervigen.com



http://www.tylervigen.com

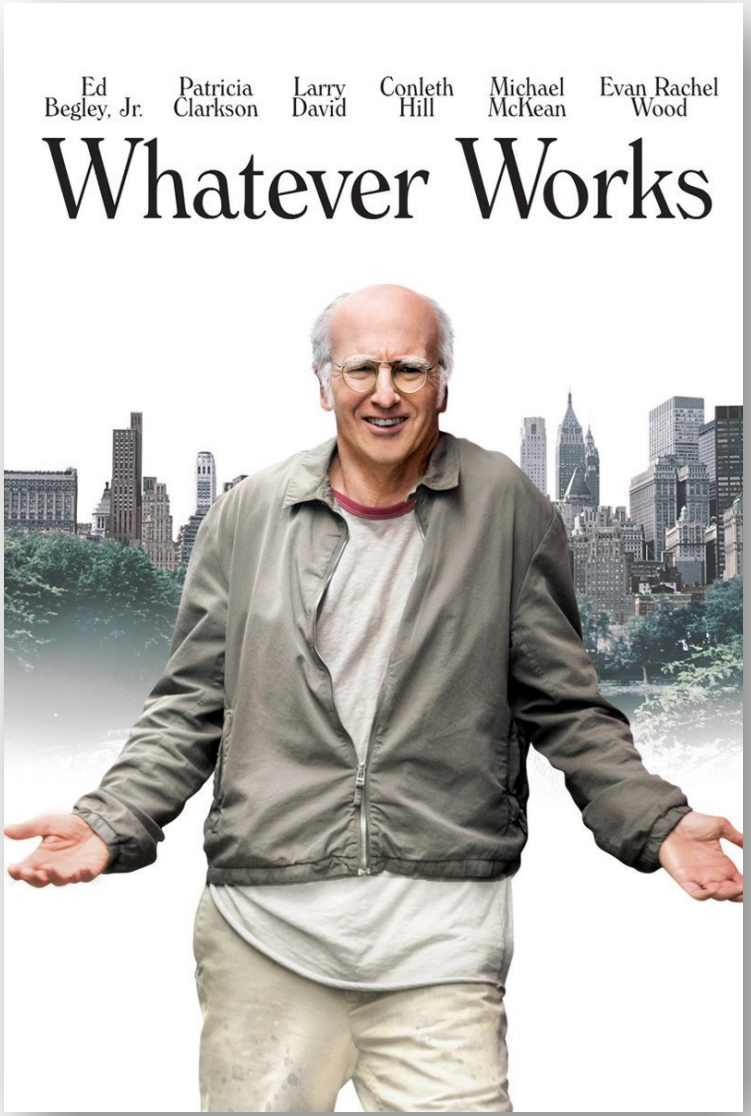


http://www.tylervigen.com

Statistics, Machine Learning, and Data Mining

- Statistics is more theory-based, focuses on testing hypotheses
- Machine learning is more based on heuristic, focuses on building program that learns, more general than Data Mining
- Data Mining
  - Integrates theory and heuristics
  - Focus on the entire process of discovery, including data cleaning, learning, integration and visualization

Distinctions are blurred!

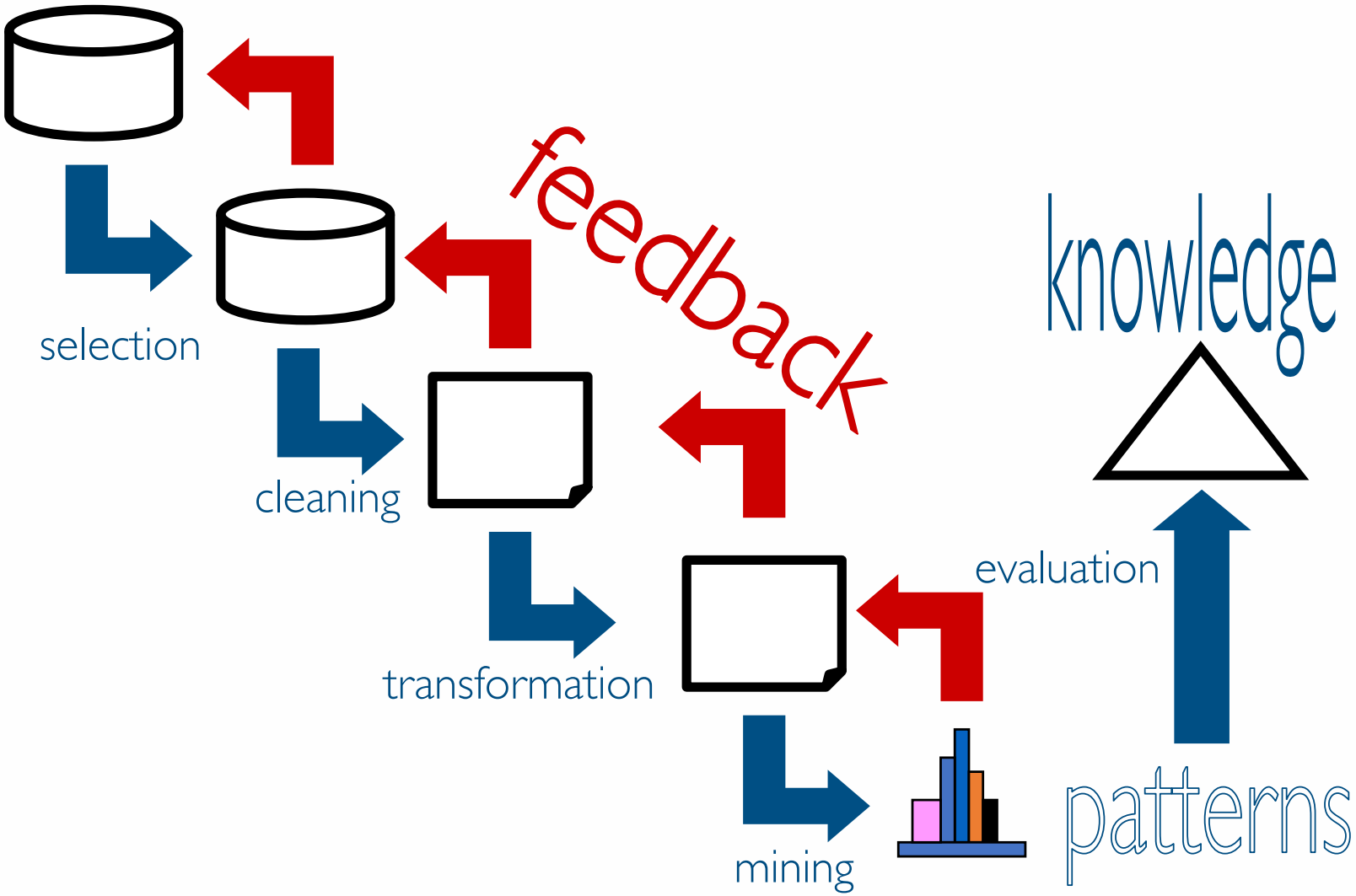


http://www.imdb.com/title/tt1178663/

Why Is It Different?

- Tremendous amount of data
  - High scalability to handle terabytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Knowledge Discovery Process



Knowledge Discovery Process
What are the main steps?

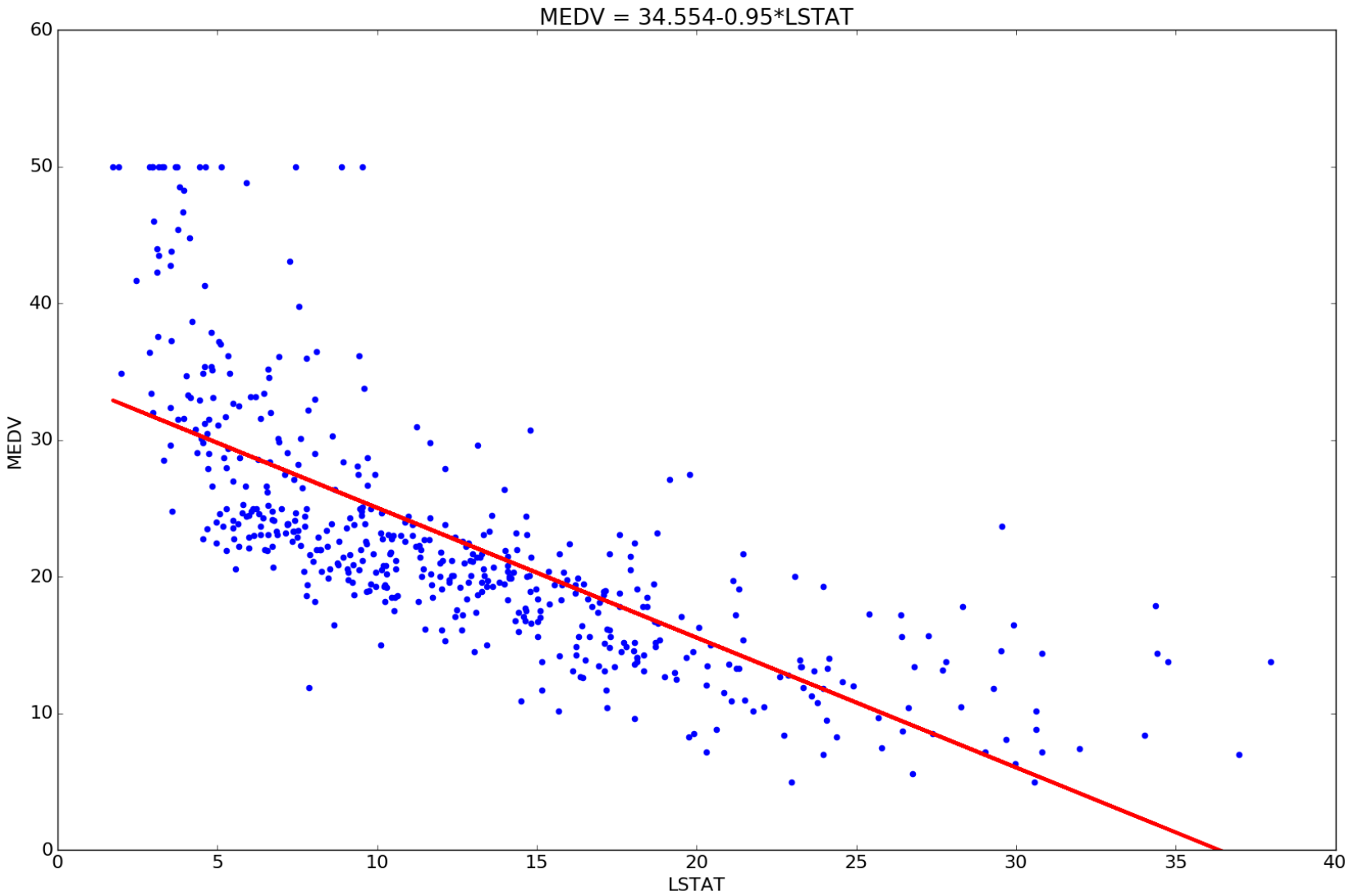
- Learning the application domain to extract relevant prior knowledge and goals
- Data selection
- Data cleaning
- Data reduction and transformation
- Mining
  - Select the mining approach: classification, regression, association, clustering, etc.
  - Choosing the mining algorithm(s)
  - Perform mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

What are the typical Data Mining tasks?

What are the Major Data Mining Tasks?

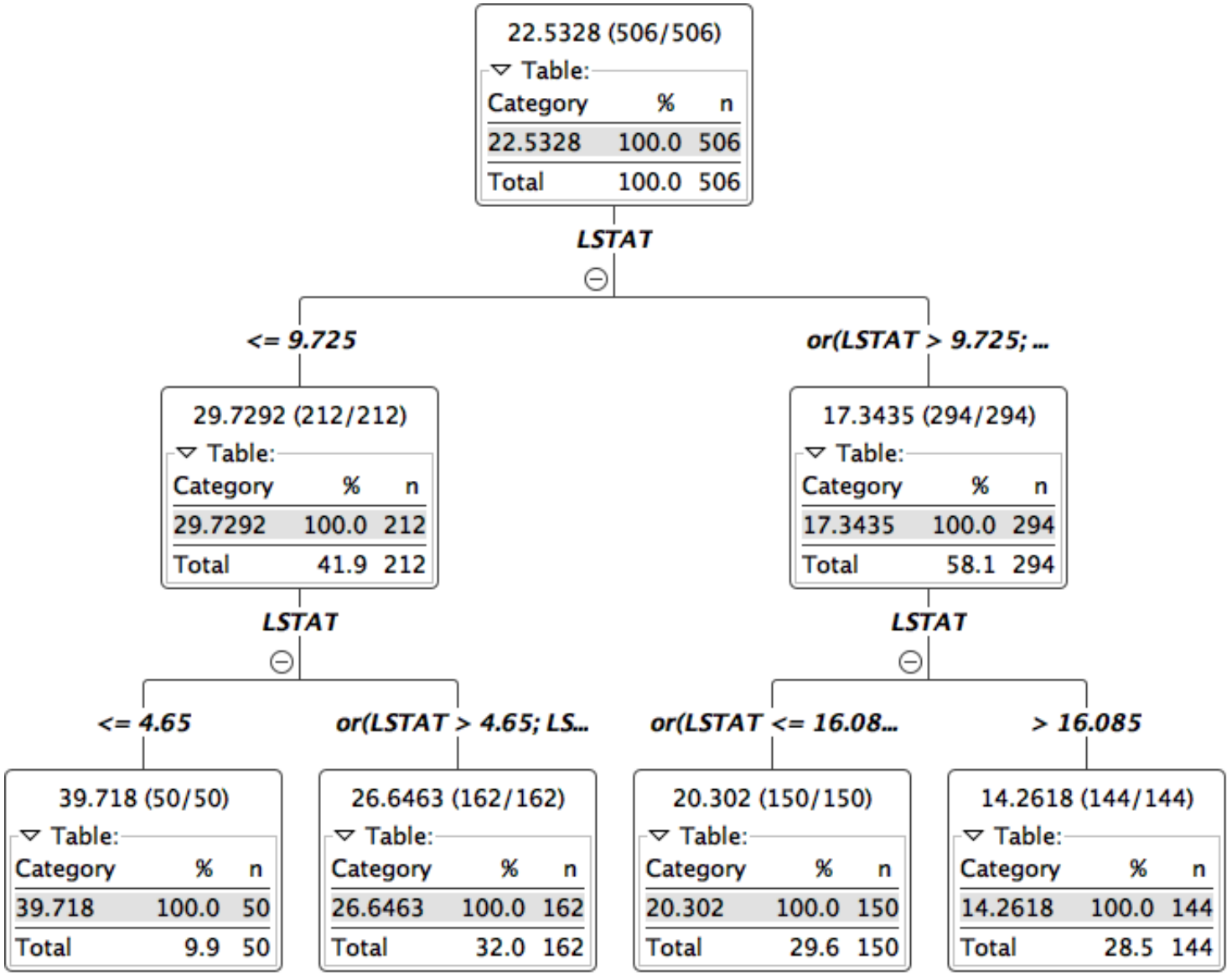
- Classification: predicting an item class
- Clustering: finding clusters in data
- Associations: frequent occurring events…
- Visualization: to facilitate human discovery
- Summarization: describing a group
- Deviation Detection: finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationship
- "Sentiment" Analysis, "Opinion" Mining
- But many appears as time goes by, opinion mining, sentiment mining

Prediction & Regression



Input variable: LSTAT - % lower status of the population

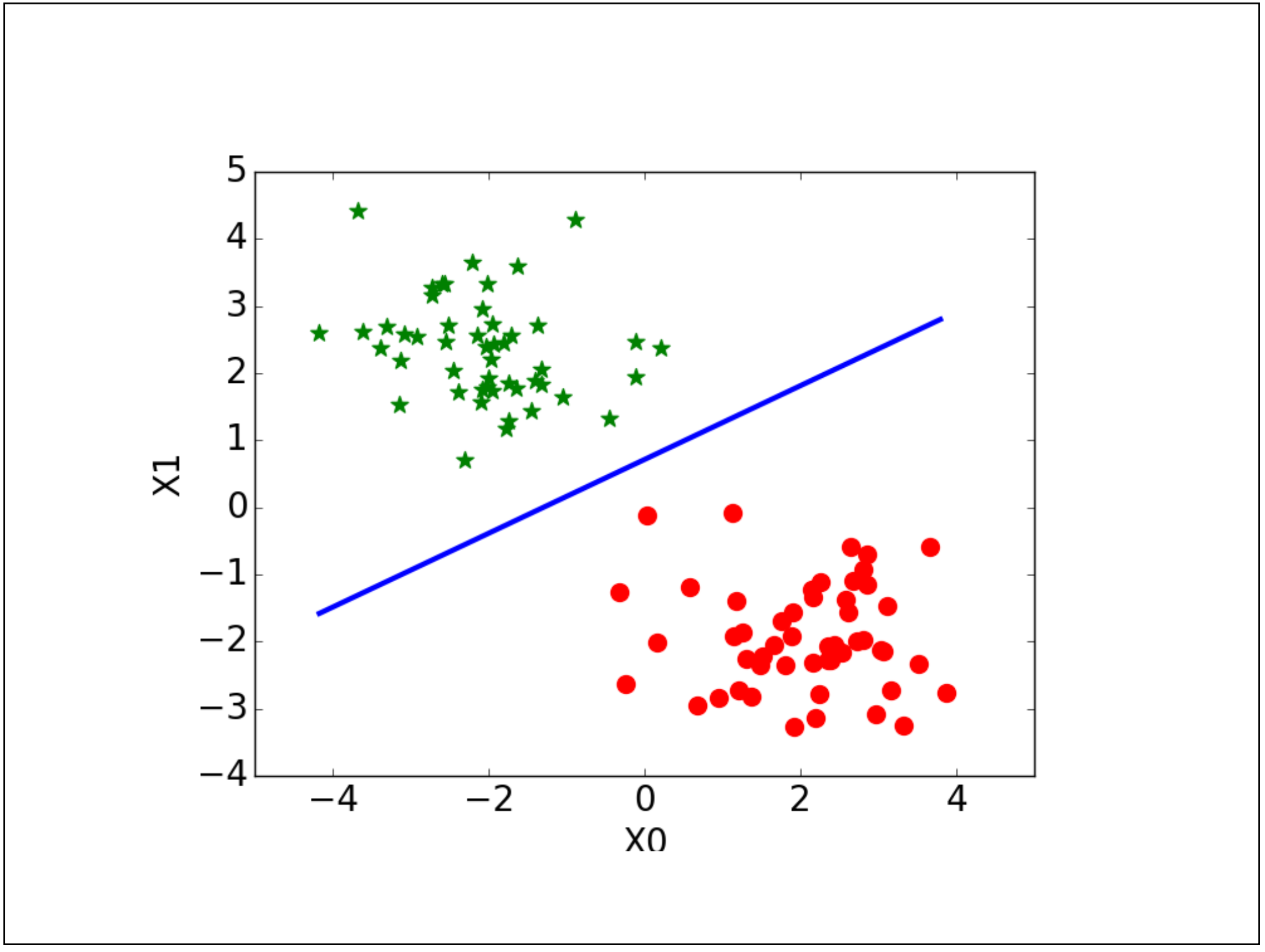
Output variable: MEDV - Median value of owner-occupied homes in $1000's



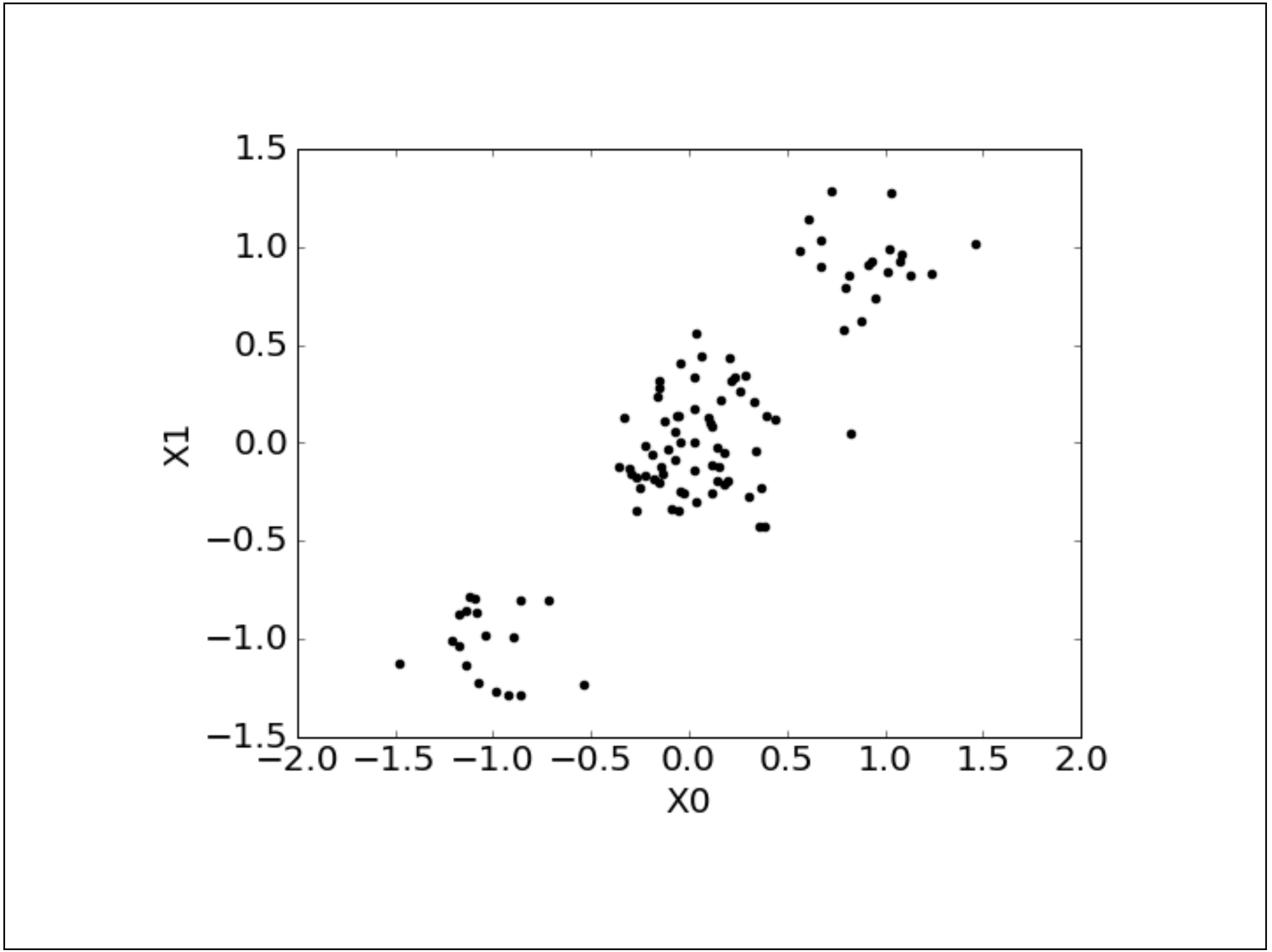
Input variable: LSTAT - % lower status of the population

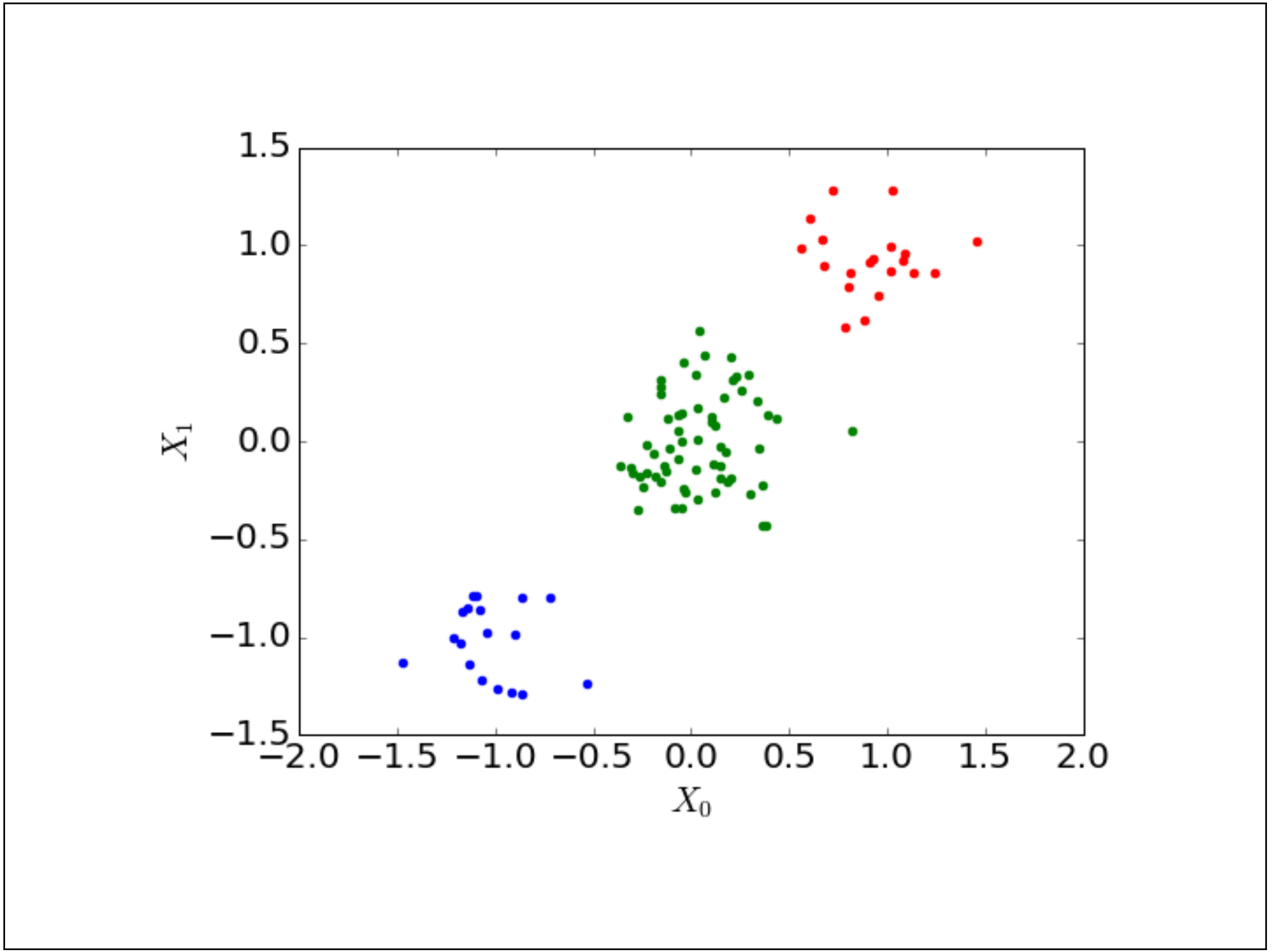
Output variable: MEDV - Median value of owner-occupied homes in $1000's

Classification


5
4
3
2
1
0
−1
−2
−3
−4
X1
−4
−2
0
2
4
X0

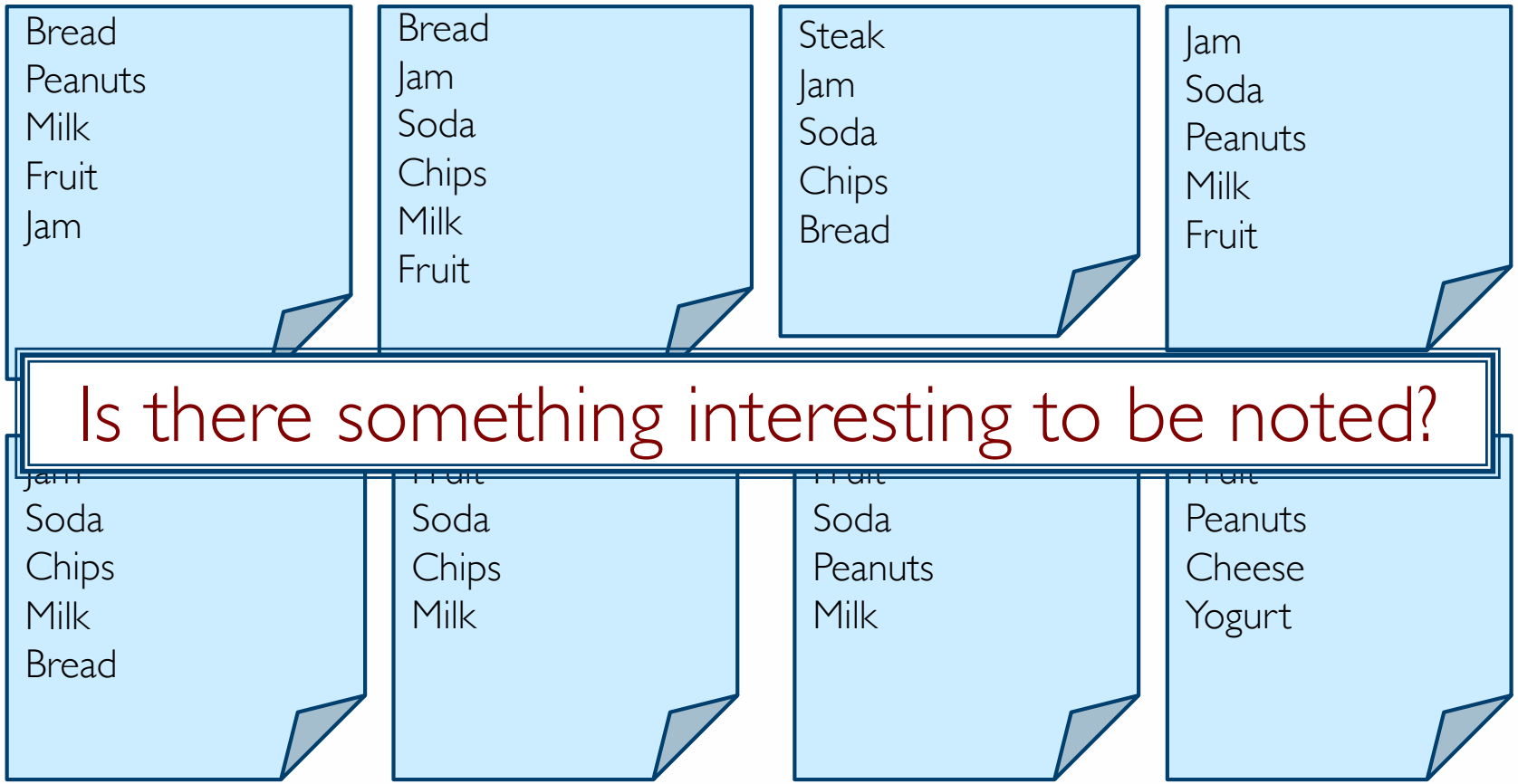
Clustering


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
X1
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X0


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
X_1
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X_0

Associations

Data Mining Tasks: Associations



Data Mining Tasks: Associations

- Finds interesting associations and/or correlation relationships among large set of data items.

- E.g., 98% of people who purchase tires and auto accessories also get automotive services done

Data Mining Tasks: Other Tasks

- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Text Mining, Topic Modeling, Graph Mining, Data Streams
- Sentiment Analysis, Opinion Mining, etc.
- Other pattern-directed or statistical analyses

Relevant Issues

Are all the "Discovered" Patterns Interesting?

- Data Mining may generate thousands of patterns, not all of them are interesting.
- Interestingness measures: a pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- Objective vs. subjective interestingness measures:
- Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
- Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, etc.

Can We Find All and Only Interesting Patterns?

- Completeness
  - Find all the interesting patterns
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering

- Optimization
  - Search for only interesting patterns:
  - Can a data mining system find only the interesting patterns?
  - Two approaches: (1) first general all the patterns and then filter out the uninteresting ones; (2) generate only the interesting patterns—mining query optimization

What About Background Knowledge?

- A typical kind of background knowledge: Concept hierarchies
- Schema hierarchy
  - E.g., Street < City < ProvinceOrState < Country
- Set-grouping hierarchy
  - E.g., {20-39} = young, {40-59} = middle_aged
- Operation-derived hierarchy
  - email address: hagonzal@cs.uiuc.edu
  - login-name < department < university < country
- Rule-based hierarchy
  - LowProfitMargin (X) <= Price(X, P1) and Cost (X, P2) and (P1 - P2) < $50

https://www.youtube.com/watch?v=CO2mGny6fFs

Assignments

- Mining of Massive Datasets (Chapter 1)