# Introduction

Statistical Data Mining I

Rachael Hageman Blair

# What is Data Mining?

**Data mining:** Tools, methodologies and theories for revealing patterns in the data – a critical step in knowledge discovery.

**Driving forces:**

- Explosive growth of data in a variety of fields
  - Cheaper storage devices with high capacity
  - Faster Communications
  - Better database management systems
- Better Computing Power
- More emphasis on ‘making the data work for us’.

# Also known as … Multivariate Analysis

**Multivariate Analysis:** the simultaneous statistical analysis of a collection of random variables.

**Origins:** social, behavioral sciences, agriculture, biology, astrology.

- Factor Analysis -> explain psychological theories of human behavior.
- PCA -> analyze student scores over a battery of tests concerning psychological measurement.
- Discriminant analysis -> classification based on botanical measurements.
- Regression and correlation -> heredity and the orbits of the planets.

# Two flavors of data mining

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.

- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy or those models and procedures when applied to fresh data.

# Two flavors of data mining

- **Descriptive data mining:** Search massive data sets and discover the locations of unexpected structures or relationships, patterns, trends, clusters, and outliers in the data.

- **Predictive data mining:** Build models and procedures for regression, classification, pattern recognition, or machine learning tasks, and assess the predictive accuracy or those models and procedures when applied to fresh data.

# Research Fields

- Statistics
- Engineering
- Bioinformatics
- Signal Processing
- Machine Learning
- Pattern Recognition
- Computer Science
- Databases

# A Few Applications …..

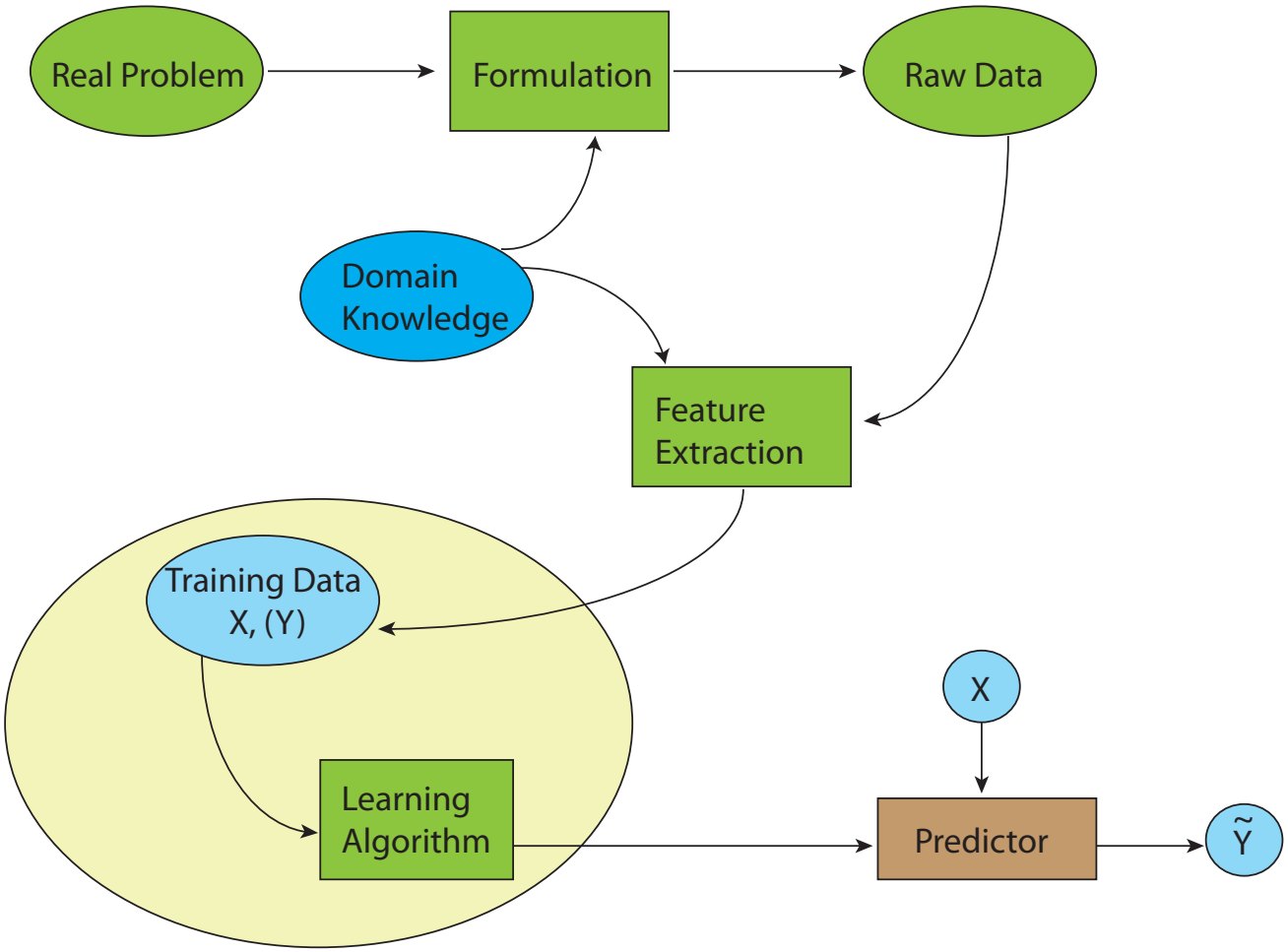
- Business
  - Insurance Companies
  - Internet Marketing ~ suggested buys (e.g., Amazon).

- Life Sciences
  - High-throughput Analysis
  - Medical Imaging

- Communication Systems
  - Speech recognition
  - Image Analysis

- Social Networks
  - Friend and group recommendations.
  - Terrorist Links

# Birds Eye View: Prediction



# Terminology

**Notation:**

- **Input X:** X is often multi-dimensional and in matrix form. Each dimension is denoted by $X_j$ and is referred to as a feature, predictor, or independent variable.

- **Output Y:** response, dependent variable.

**Input and Output may be of different variable type:**

- **Quantitative variable:** cholesterol levels, height.
- **Qualitative variable:** flower species, low/high risk.
- **Ordered categorical:** small, medium, and large.

**Why does it matter?**

- Supervised Learning vs. Unsupervised learning

  *Is Y available?*
- Regression vs. Classification

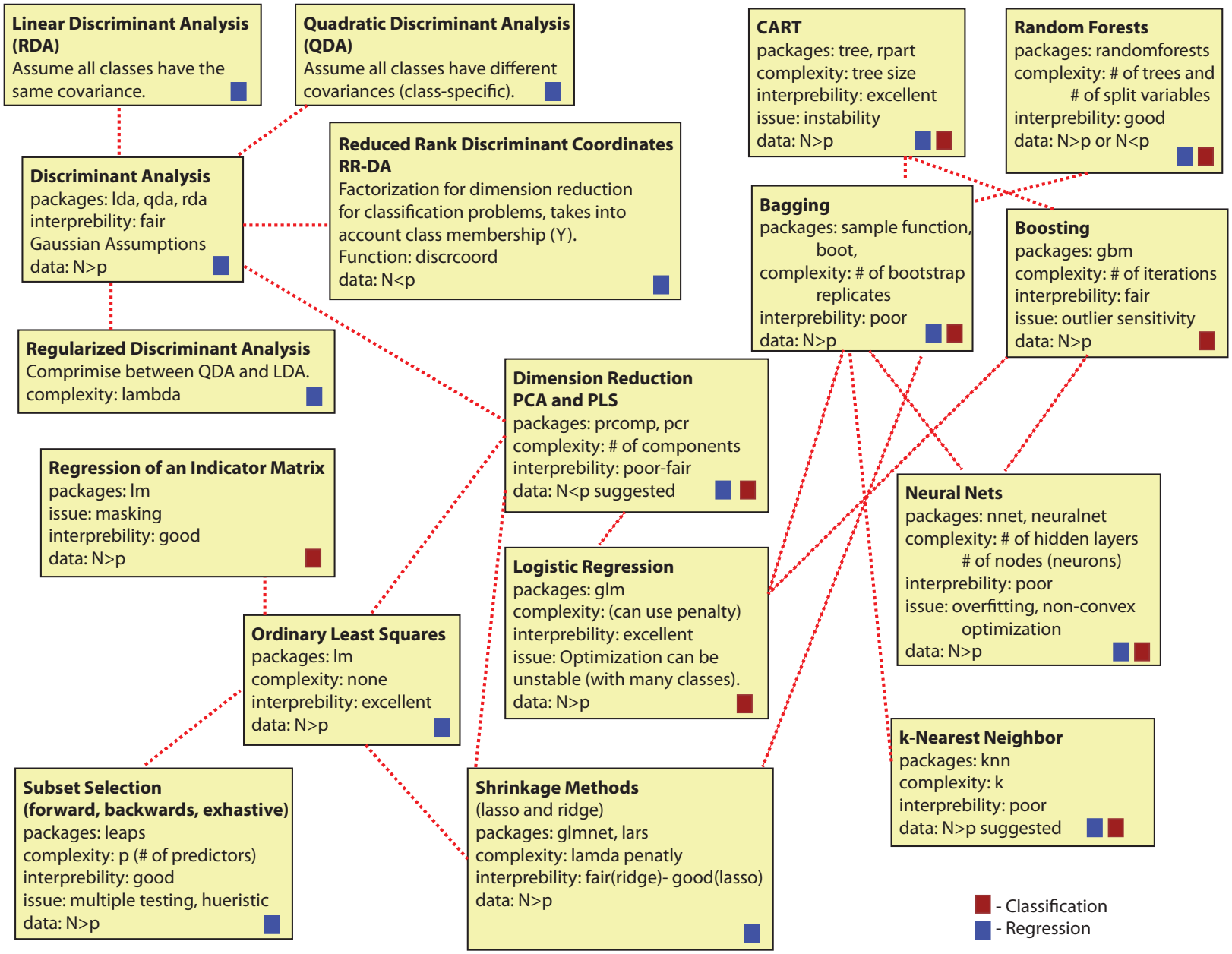
  *Is Y qualitative or quantitative?*

# Where we are going?

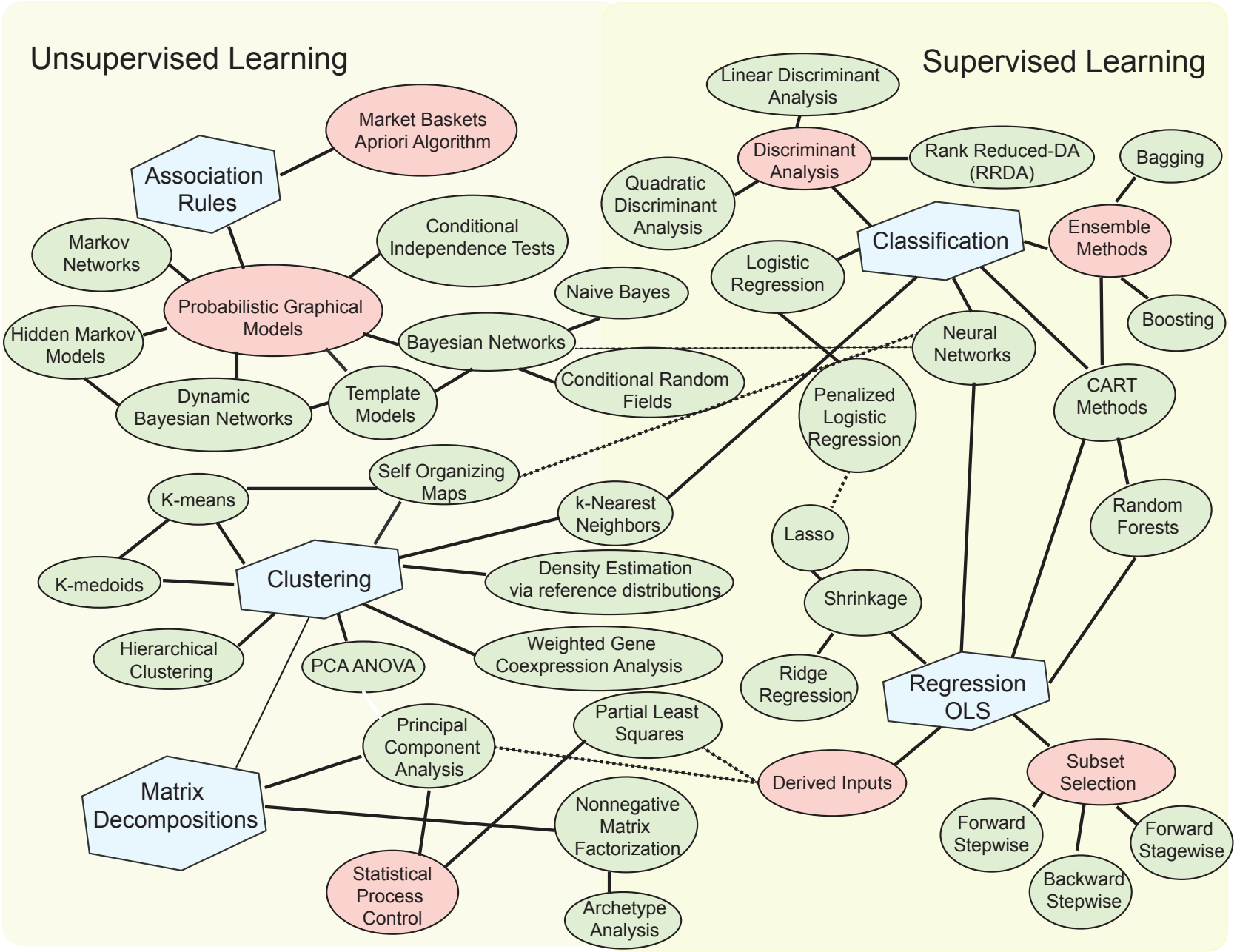
Supervised Learning

Unsupervised Learning

# Data Mining I



# Data Mining I & II



# Our Focus: Supervised Learning

**Task:** Use inputs to predict outputs

- Analogous to a human learning from past experience.
- A computer does not have “experience”.
- A computer system learns from “data” which represents some form of “past experience” from our application domain.

# Example: Emergency Room

- An emergency room measures 15 variables (blood pressure, temperature, age, … ) of newly admitted patients.
- Must decide on whether to place a new patient in the Intensive Care Unit (ICU).
- Due to high cost and limited space, those patients that may survive less than a month are given priority.
- Problem: to predict high-risk and low-risk patients.

# Example: Credit Card Application

- A credit card-company receives thousands of applications for new cards.  Each application contains the standard information (e.g., age, marital status, annual salary, outstanding debts, credit rating….).
- Problem: to decide whether an application should be “approved” or “not approved”

# Example: Handwriting Recognition

- Raw data: Images that are scaled segments from five digit zip codes.
  - 16 x 16 eight-bit grayscale maps
  - Pixel intensities range from 0 (black) to 255 (white)
- Input data: a 256 dimension vector, or feature vectors with lower dimensions.
- Problem: identify single digits 0~9 based on images. Or assign an "I don't know". Keep the error rate down.

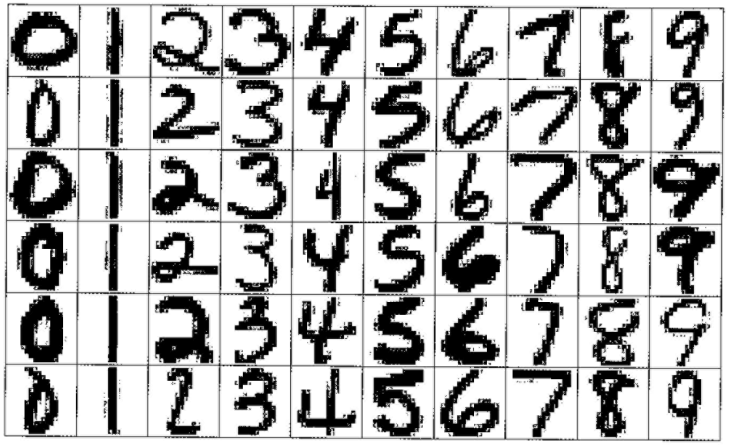


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

# Example: Image Segmentation

- Raw data: Picture
- Problem: Reconstruct image in gray-scale image for classification.

# Example: Email Spam

- Raw data: 456 Email messages
- Input data: Each of the 456 email messages is classified as "spam" or "important". Relative frequencies of key words and punctuation marks in the message.
- Problem: Try to figure out if an incoming message is "spam" or "important".

Possible Decision Rule:

If (%rachael < .6) & (%you >1.5) then spam, else important

Or,

If (.2 %you - .3 % rachael) > 0 then spam, else important

# Example: Prostate Cancer

- Raw data: clinical measures from 97 men.
- Problem: Predict prostate specific antigen (PSA) from a number of clinical measurements.

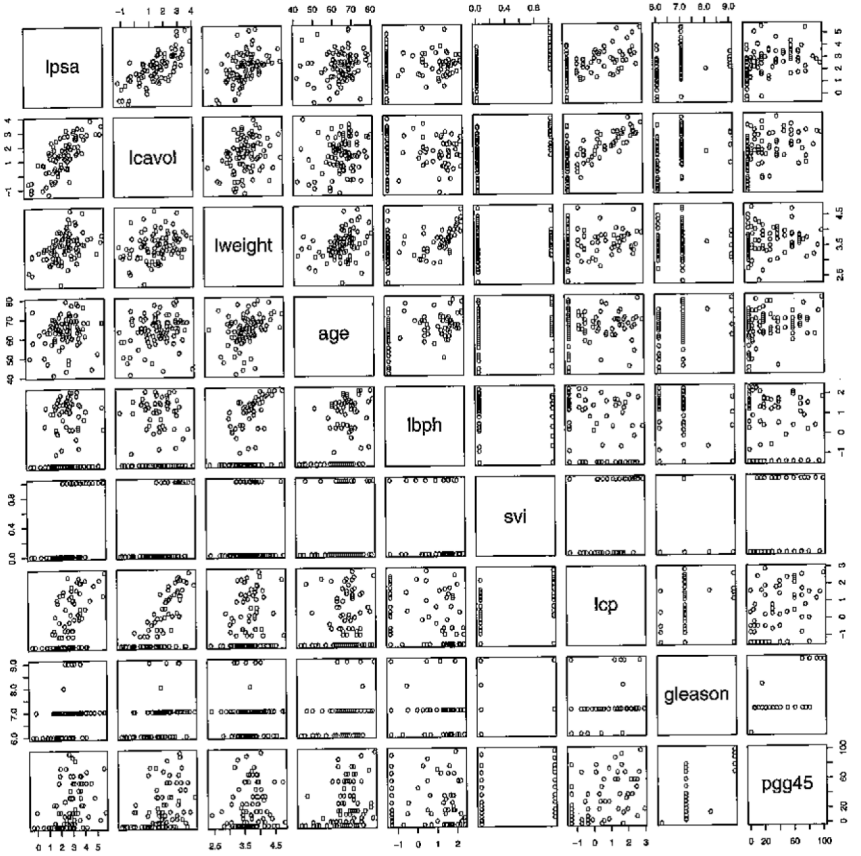



**FIGURE 1.1.** *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

# What is a Data Scientist?

The data scientist role has been described as “part analyst, part artist.”

“A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization.”

- Anjul Bhambhri, vice president of big data products at IBM

# What is a Data Scientist?

The data scientist role has b

“A data scientist is somebod
It's almost like a Renaissanc
an organization.”

- Anjul Bhambhri, vice presi

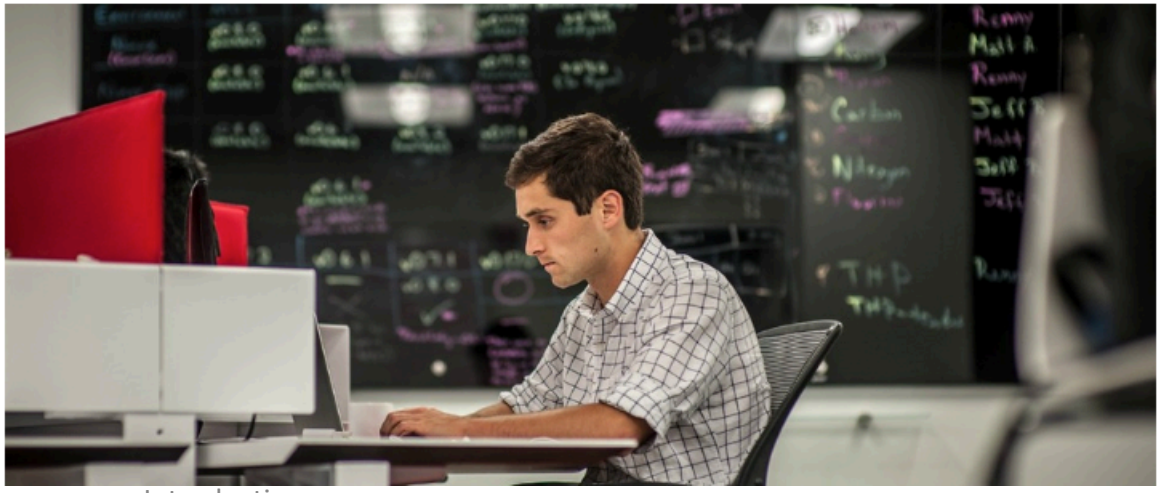
Data mining tops LinkedIn's list of the 'hottest skills of 2014'

5.1k SHARES

Share on Facebook | Share on Twitter | +

WHAT'S THIS?





# What is “Big Data”…..

- “Big Data" is a popular term inspired by the explosive growth of digital data in various forms.
- More concretely, the term “Big Data" refers to data sets of extreme size (tall, fat, or both) and complexity (heterogeneous, dynamic, sparse, etc.)

Consequently…. limited applicability of conventional methods of data processing and analysis.

# The four V's …..

**Volume -** The huge decrease in the cost of DNA sequencing has had a transformative accumulation of genomic data, such that large genomic datasets on the order of 2 petabytes are presently being gathered every year. Electronic health records. A single hospital may generate up to 665 terabytes of data in medical records every year.

**Variety -** refers to the structural and functional heterogeneity of data received from multiple sources. Healthcare data is extremely heterogeneous including structured lab reports, narrative reports, images, physiological measures, genomic (and other -omic) features, social-media, billing and utilization data, to name just a few.

**Velocity -** indicates the speed with which data is captured, parsed, and processed in real time.

**Veracity -** invokes the inconsistent quality of data, along with ambiguities that are characteristic of complex datasets