Data Mining and Knowledge Discovery in Health

Ali Mohammad Nickfarjam

PHD of Artificial Intelligence

Outline

- What is DM and KD?
- Characteristics
- Applications
- Some Methods
- Software and Implementation
- Big Data

What is DM & KD?

What is DM & KD?

- “The process of identifying hidden patterns and relationships within data”

or

- “Data mining helps end users extract useful business information from large databases”

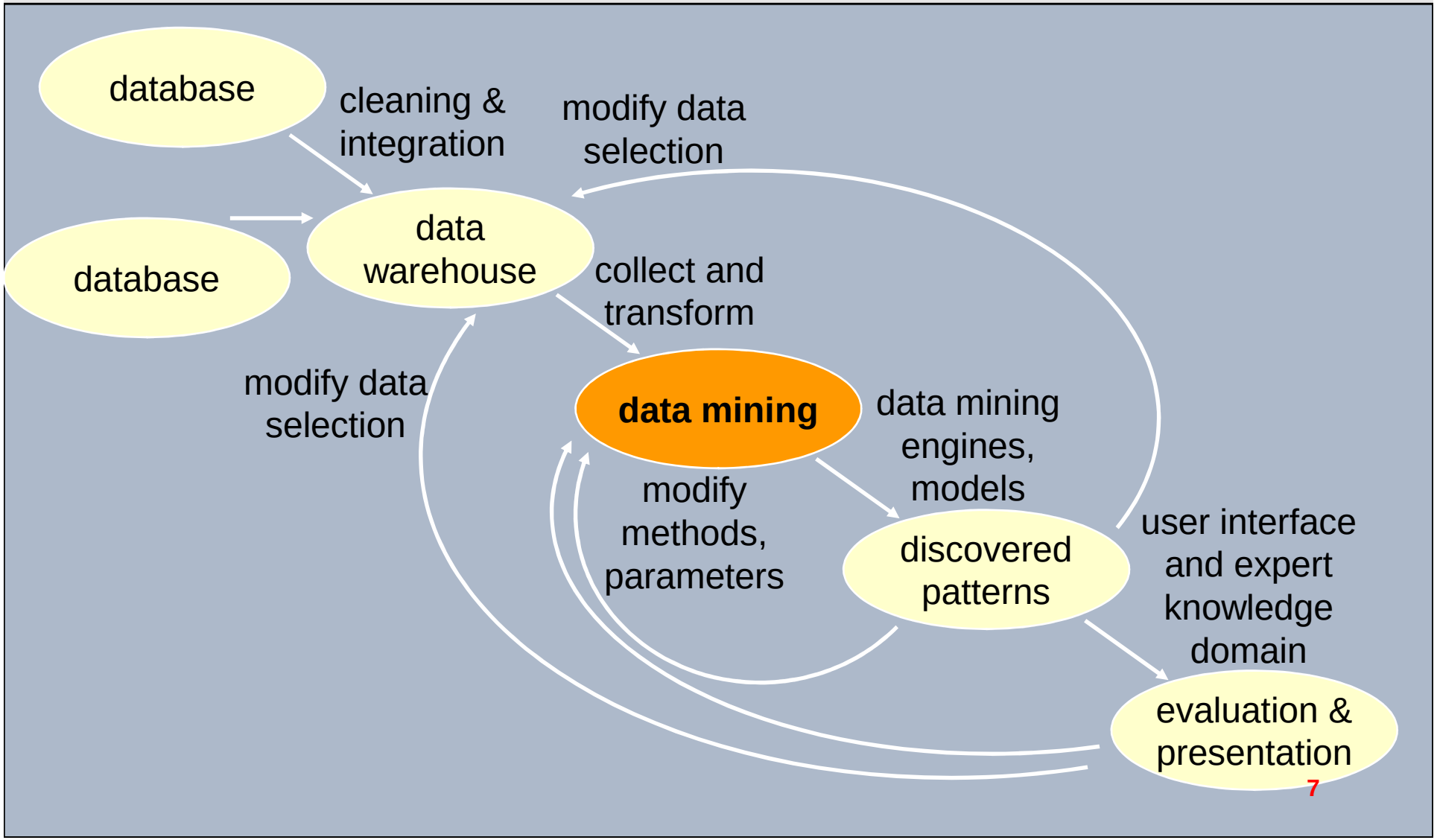
What's the Appeal?

- Hidden nuggets of valuable information buried deep within a mountain of otherwise unremarkable data
- Important data
- Seek competitive advantage

The Challenge

```
51020188905212001539458199000000001419881229448821996081
62100000010100010000000110000311111000000000100313020000
00000000002020010000000000000000000000000000004343888888888
42424342433301220202220000101001000000000441000000000110000
000000000000010000010000000000000000000000000000000000000
00000000000001998102751020189606012002126940968000000159
01998090337981199809173100100000100010000000011000032000 2
0000001000000012399000000000000002002222003131003120000000
0000000004243888888888842434242332121212220000001011000
000244100000000010020000000000000000000000100000000000000
00000000000000000000000000000000001998123051020189702032
00018626929200000047091998021356971199802273100000100100
01000000000110110000002000010000000000210110001000000000
01000000000000010001100000001110033888822223311323343330 0
00001100000111010011001020001000000000100000000100000000 0
00000000000000000000000000000000000000000000000000000001
99812215102018990930200520089867300000194101999011275981
19990126310010001010001000000000111110111112201010000011
12300100100000010210002200000000000020000000000000011133438
88843424242434242330000001111000001011001000024410000000
```

Process: KD In Databases



DM Might Mean…

- Statistics
- Visualization
- Artificial intelligence
- Machine learning
- Database technology
- Information retrieval
- High performance computing
- And so on...

What's needed?

- Suitable data
- Computing power
- Data mining implementation
- Skilled operator who knows both the nature of the data and the software tools
- Reason, theory, or hunch

Typical Applications of DM & KD

- Health Care
  - Epidemiological Analysis - incidence and prevalence of disease in large populations and detection of the source and cause of epidemics of infectious disease
  - Knowledge for funding
  - Policy, programs
  - …
  - …

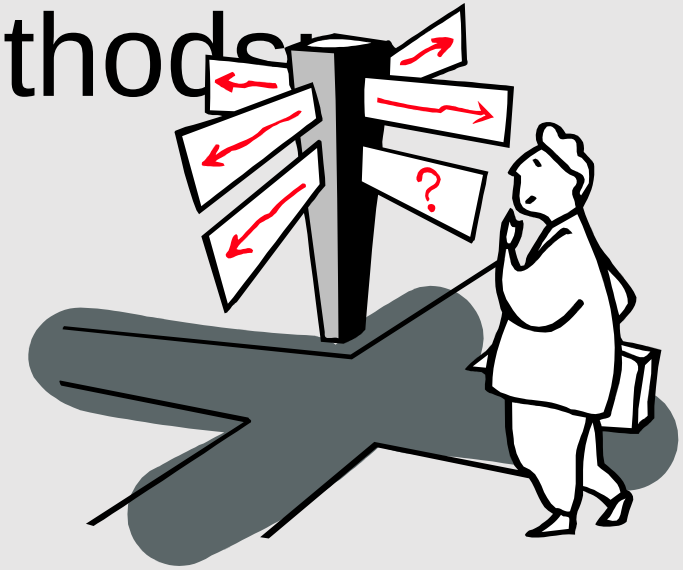
Two Basic Approaches

- Supervised
  - A dependent or target variable
- Unsupervised
  - “Pure Data Mining”
  - Fewer assumptions
  - Typically used for clustering techniques

Automation

- The ability to aim a tool at some data and push a button
- Some methods of KD/DM are more suitable for automation than others

Seven Basic Methods

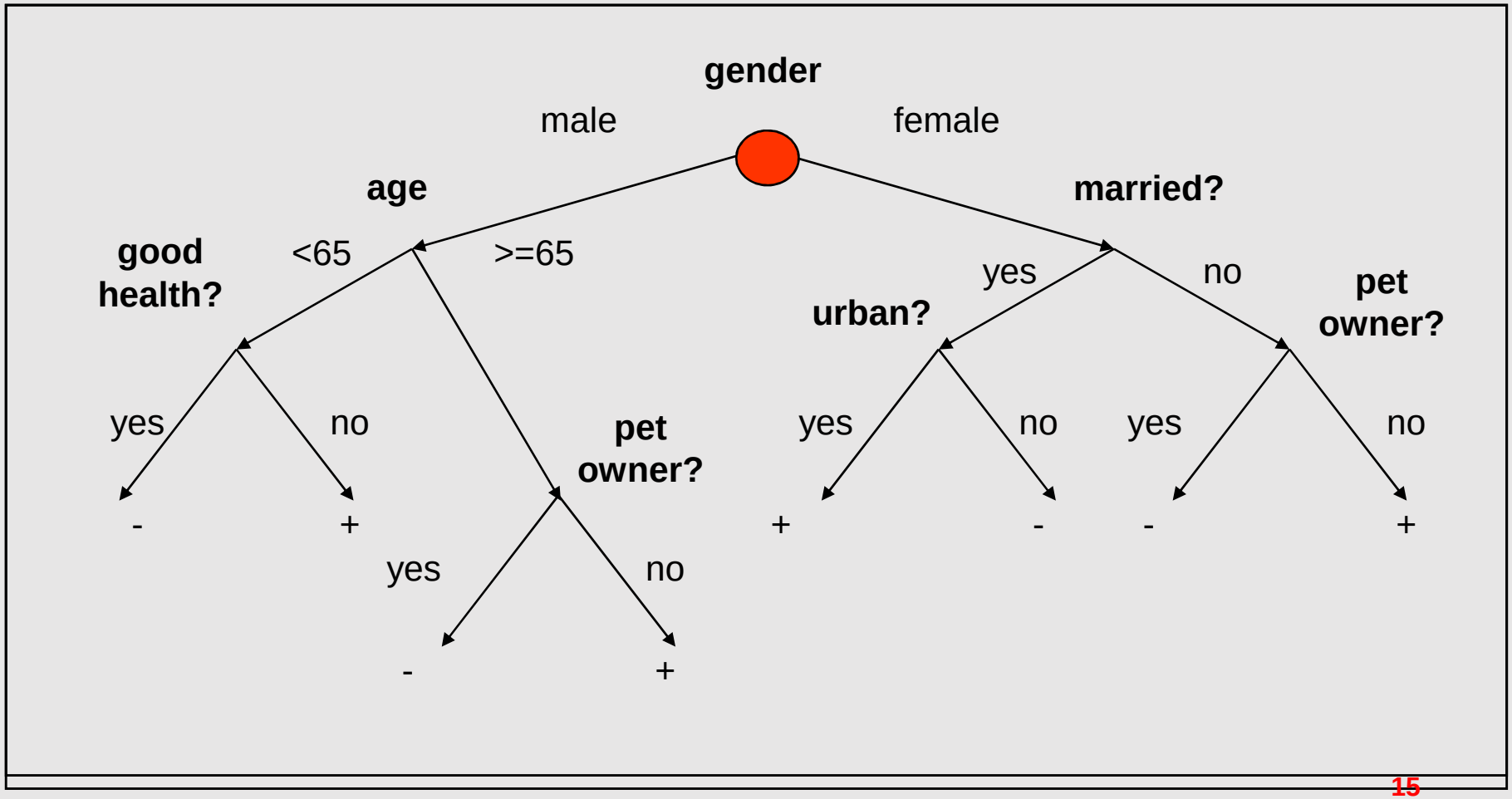


1. Decision Trees
2. (Artificial) Neural Networks
3. Nearest Neighbor
4. Genetic Algorithms/Evolutionary Computing
5. Hybrids
6. Deep Learning
7. …

Decision Trees

- Graphical representations of relationships with data
- Excel at Classification & Prediction Models

Sample of a Decision Tree



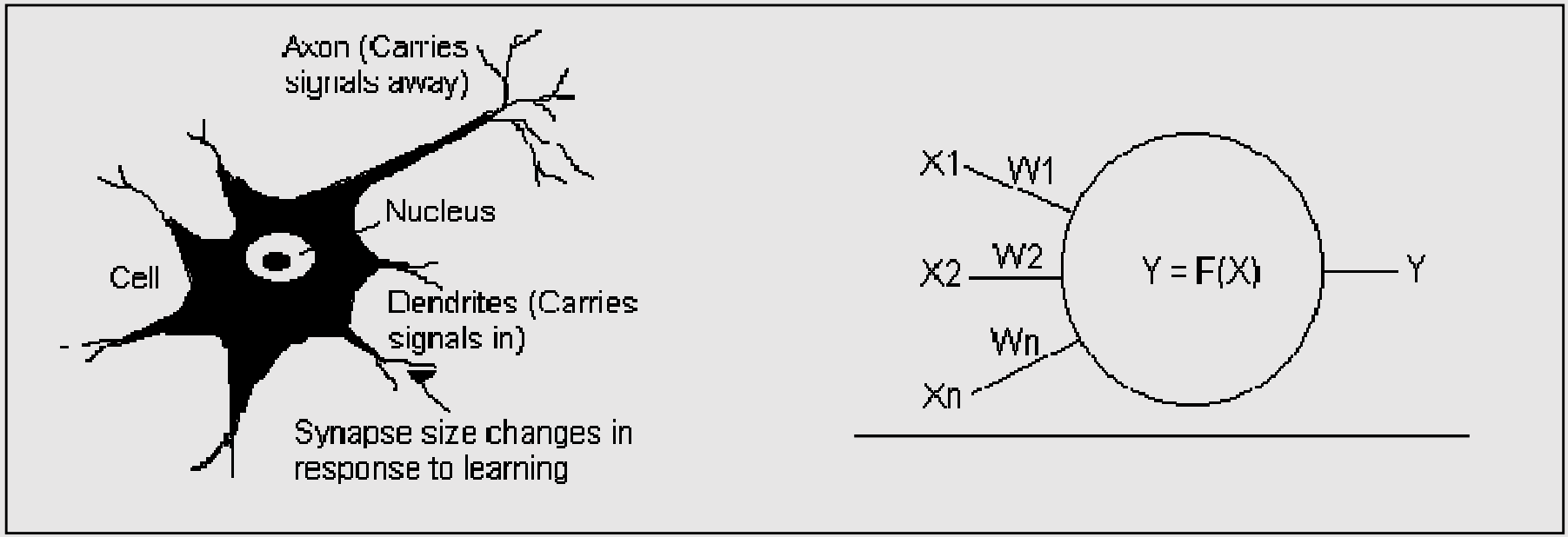
Decision Trees

- Strengths
  - Easily interpreted
  - Represent complexity in a compact form
  - Handle non-linear data well
  - Relatively well suited to automation.

- Weaknesses
  - Large trees with large numbers of variables become difficult to understand
  - Missing data must be appropriately managed in construction and use of the models

Neural Networks

- Derived from Artificial Intelligence Research
- Modelled on the Human Neuron



Neural Networks

- Strengths
  - Accuracy of prediction
  - Robust performance with a wide variety of data types

- Weaknesses
  - Overfitting
  - Poor clarity of model

Nearest Neighbor

- Aim to assign “like” records to a group
- Groups assigned according to some target variable or criteria
- Nearest neighbour used for prediction

Nearest Neighbor

- Applications:
  - Text processing: search engines
  - Image processing: radiology/image processing
  - Diseases detection

Nearest Neighbor

- Strengths
  - Easily understood and interpreted
  - Easily implemented in basic situations

- Weaknesses
  - complex data not well suited to automation (much preprocessing required)

Genetic Algorithms/ Evolutionary Computing

- Grounded in Darwin – applied using mathematics
- Require
  - a way to represent a solution to a problem
  - a way to test the “fitness” of the solution
- Solutions are mathematically “mutated”
- Fittest solutions survive
- Convergence

Genetic Algorithms/ Evolutionary Computing

- Strengths
  - Suited to novel problems that are poorly understood
  - Suitable where data is dirty or missing
  - May be useful where other methods cannot be applied

- Weaknesses
  - Not easily automated
  - Require creativity in their application

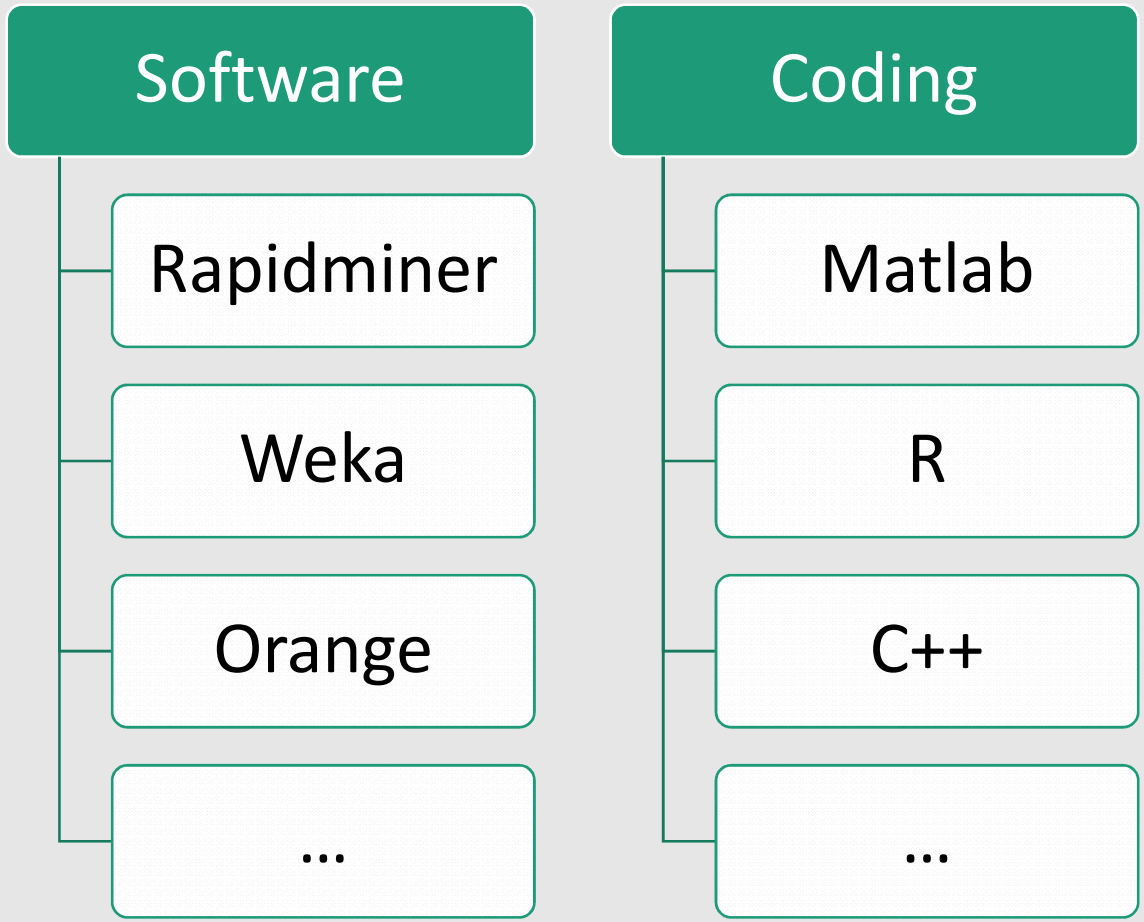
Hybrids

- Techniques used in combination
- Example: use of a genetic algorithm to identify target variables for inclusion in a neural network model

Deep Learning

- Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher level features from the raw output.

Software and Implementation



Big data


Pharmaceutical
Research Data
Health Care
Data
Clinical Data
Behavior and
Social Sentiment
Health Insurance
Claims Data

Big data

- Big data is a poorly defined marketing term comprised of a set of four things
  - Big Computer
  - Big Datasets (NoSQL data bases)
  - Big Models

Big Computer

- It's cheaper and more scalable to tie together large numbers of commodity computers than to make bigger single computers
  - That's what Hadoop and Spark are about
- This is important if you have a data set that's really really big
  - Won't fit onto single disk drives
  - Too big to be processed in a reasonable period of time

Big datasets (NoSQL data bases)

- Hadoop itself came from Google needing to manage its data
- Scanning the text in the Library of Congress is 10-20 terabytes, so a petabyte (1,000 terabytes) is VERY big
  - Library of Congress would be a tenth of that if converted to text
  - Texas Medicaid is about 40 terabytes

Big Models

- Text analysis
  - Such as looking at doctor's noted to suggest diagnosis codes
- Machine learning
  - Train a classifier on a sample and then score the entire population
  - Used for health risk scoring and fraud detection
- Network analysis (Analyze the links between entities)
  - Used for fraud detection and workflow analysis

Conclusion