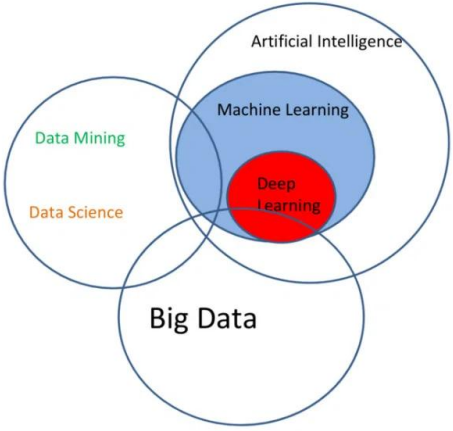
Session 1: Introduction to Machine Learning

Javier Serrano
Applied Machine Learning
Master in Data Science and Analytics

Strathmore University
@iLabAfrica Centre

Differences between AI, Machine learning and Data Analytics



Why are we interested in Mining Data?

- Explosive Growth of Data: From Terabytes to Petabytes
  - Availability of Data Collections:
    - Clinical records
    - Web sites
    - Remote monitoring IoT
    - Genomics and proteomics
    - Network traffic
    - And many others sources...
- We are drowning in data but short on knowledge
- Need for automated analysis of massive datasets

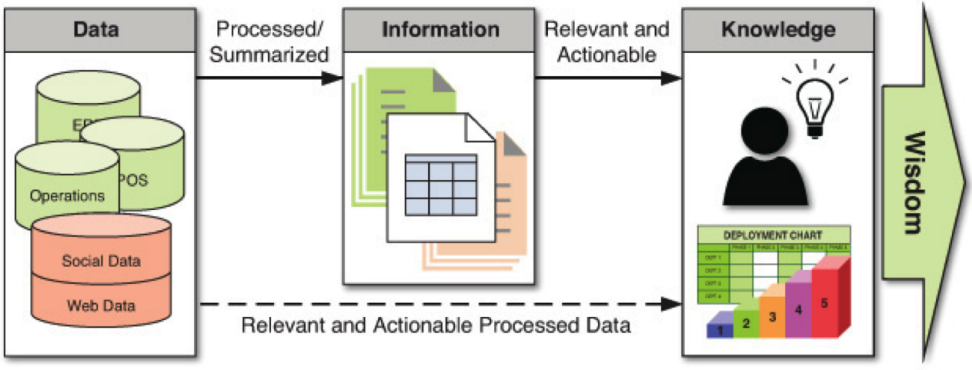
Sumary

What's Data Mining

- **Extraction** of interesting (non-trivial, implicit, previously unknown and potentially useful) **patterns or knowledge** from huge amount of data
- **Alternative names**: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis/analytics, business intelligence, etc.
- **Watch out**: Is everything "data mining"?
  - Look up phone number in phone directory (Not)
  - Query a Web search engine for information about "Cancer" (Not)
  - Group together similar patients returned from some medical DB (Yes)

What's Data Mining?

From Data to Knowledge

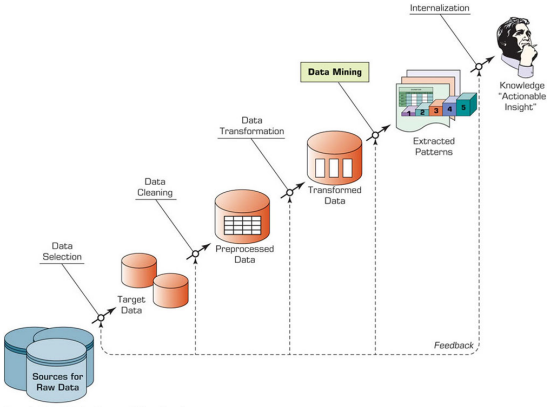


What's Data Mining?

A Multidiciplinary Approach

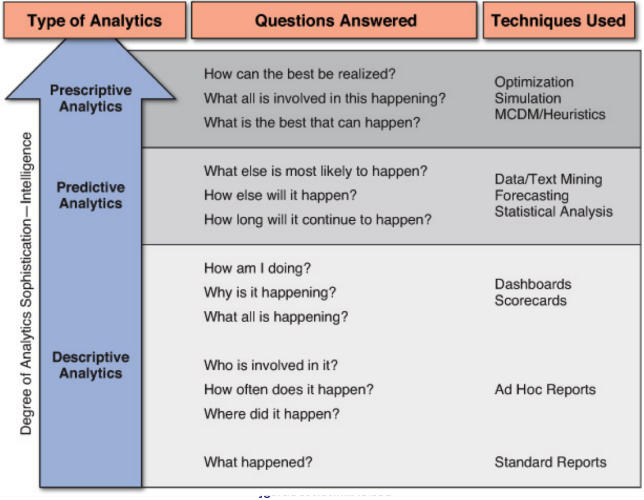
What's Data Mining?

Process of Knowledge Discovering



What's Data Mining?

Types of Data Analysis



What's Data Mining?

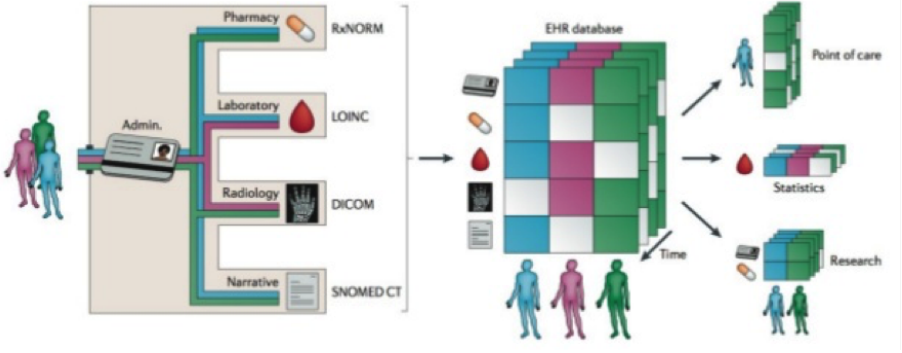
A Multi-Dimensional View of Data Mining

- **Data to be mined**: Database data, data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined**: Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
- **Techniques utilized**: Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**: EHR mining, bio-data (gen) mining, clinical reports mining, medical signals, network analytics

Sumary

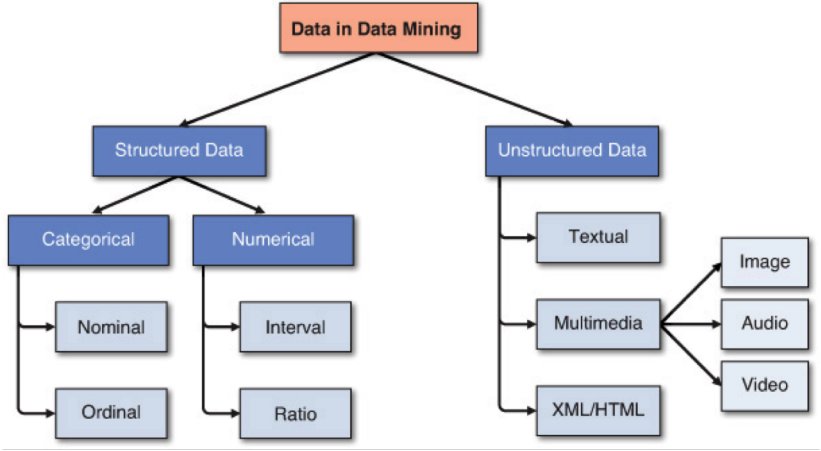
What Kind of Data Can Be Mined?

Medical Data: Electronic Health Record (EHR)



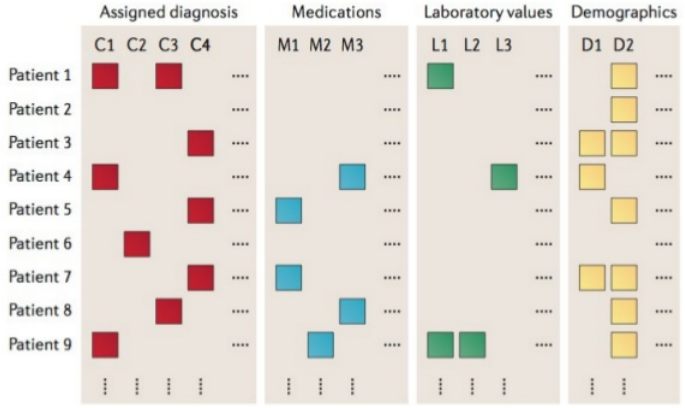
What Kind of Data Can Be Mined?

Data Classification



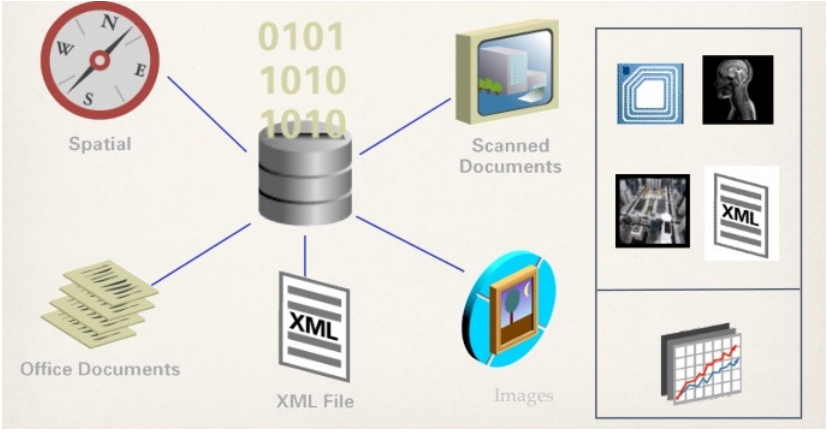
What Kind of Data Can Be Mined?

Medical Structured Data



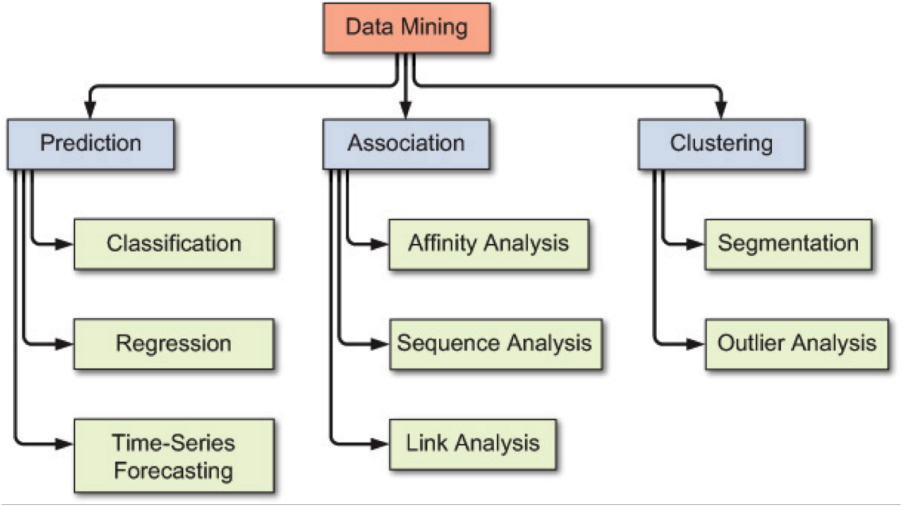
What Kind of Data Can Be Mined?

Unstructured Data



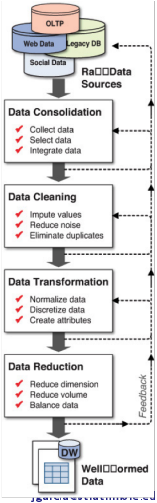
Sumary

Simple Taxonomy for DM



Data Mining Functions

1. Preprocessing



Data Mining Functions

2. Association Analysis

- **Frequent patterns** (or frequent itemsets): What features are frequently occurred together for some disease?
- **Association, correlation vs. causality**
  - A typical association rule

    $$\text{Fever, sore throat} -> \text{flu } [0.5\%, 75\%] \text{ (support, confidence)}$$

  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in **large datasets**?
- **How to use such patterns** for classification, clustering, and other applications?

Data Mining Functions

3. Classification and Prediction

- Classification and **label prediction**
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
  - E.g., classify AF from ECG
  - Predict some unknown class labels
- **Typical methods**: Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- **Typical applications**: classify potential risk patients, gens, diseases

Data Mining Functions

4. Cluster Analysis

- **Unsupervised learning** (i.e., Class label is unknown)
- **Group data** to form new categories (i.e., clusters), e.g., cluster gens to find distribution patterns
- **Principle**: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Functions

5. Outlier Analysis

- **Outlier**: A data object that does not comply with the general behavior of the data
- **Noise or exception?** - One person's garbage could be another person's treasure
- **Methods**: by product of clustering or regression analysis, ...
- Useful in **fraud detections** or **rare disease analysis**

Data Mining Functions

6. Sequence Patterns and Data Stream Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining: Motifs and biological sequence analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

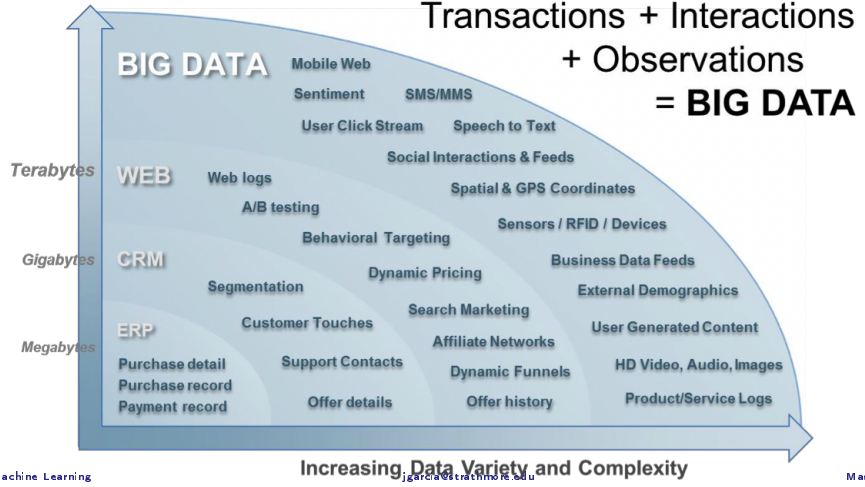
Sumary

What's Big Data?

- A **collection of data sets** so large and complex that it becomes difficult to process using traditional data processing applications
- Big Data Analytics is the same as 'Small Data' Analytics, only with the added challenges of large datasets (50M records or 50GB size, or more)
- **Challenges** :
  - Data storage and management
  - De-centralized/multi-server architectures
  - Performance bottlenecks, poor responsiveness
  - Increasing hardware requirements

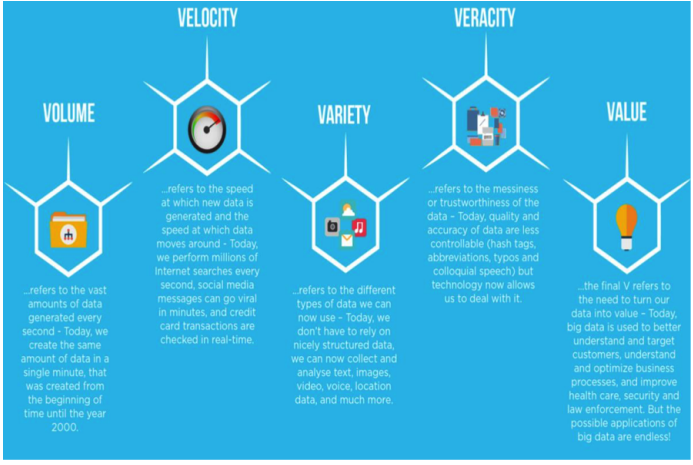
What's Big Data?

Small vs Big Data



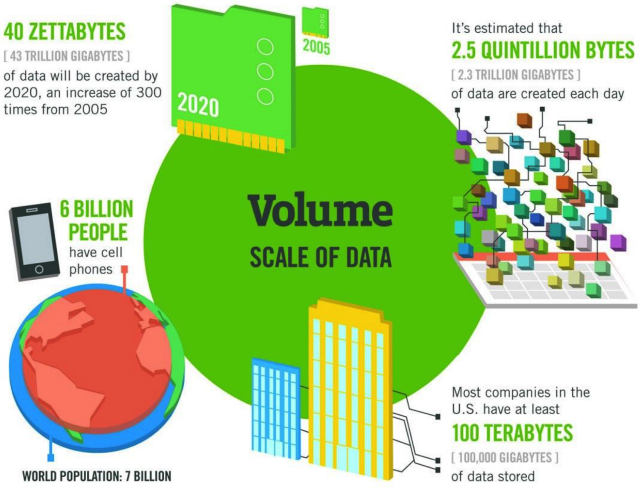
What's Big Data?

The 5 V's



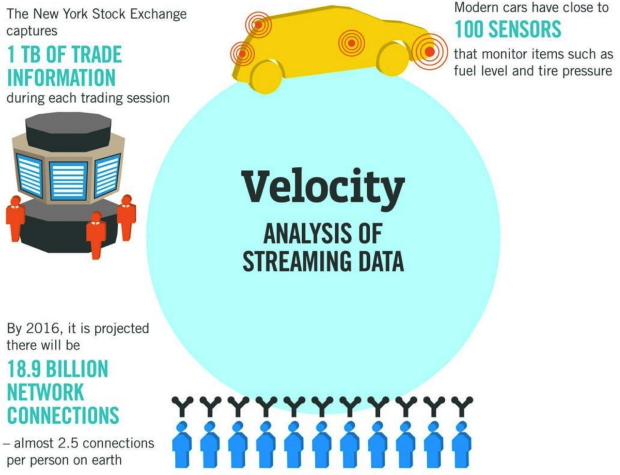
What's Big Data?

First Feature: Volumen



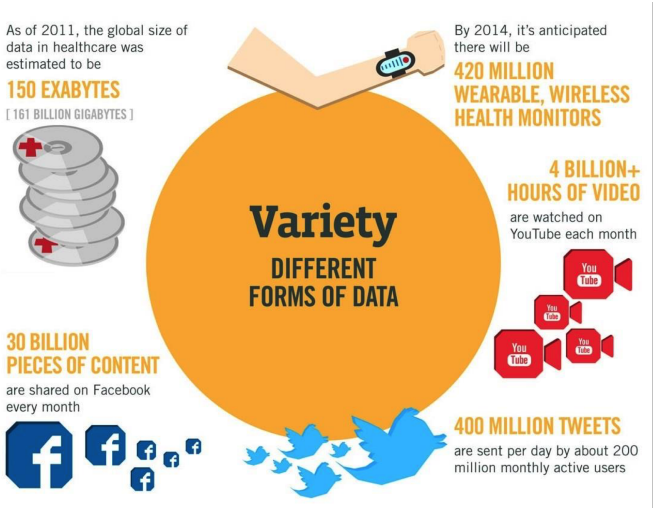
What's Big Data?

Second Feature: Velocity



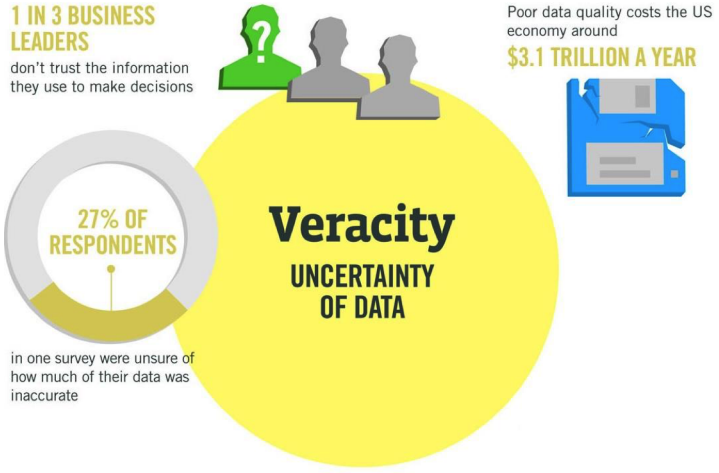
What's Big Data?

Third Feature: Variety

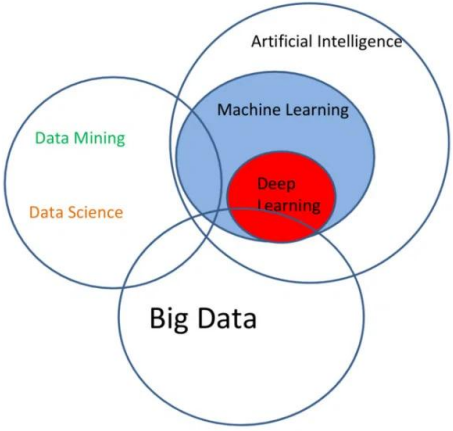


What's Big Data?

Fourth Feature: Veracity



Differences between AI, Machine learning and Data Analytics



Sumary

Life Science Applications

- Medical diagnostics tools
- Medical image analysis
- Biological sequence analysis
- Micro-array data analysis
- New drugs development
- Disease surveillance
- Environmental health impacts
- ....

Business: Marketing and Sales

- **Customer Segmentation**: Grouping customers based on their behavior, demographics, or purchase history to create targeted marketing campaigns.
- **Recommendation Engines**: Suggesting products or services to users, like on Netflix or Amazon, to increase sales and engagement.
- **Churn Prediction**: Identifying customers who are likely to stop using a service or product, allowing the business to take proactive steps to retain them.
- **Dynamic Pricing**: Automatically adjusting prices for goods or services in real-time based on demand, competition, and other market factors.
- **Lead Scoring**: Prioritizing sales leads by assigning a score that represents their likelihood to convert into a customer.

Business: Comment Other applications in these areas

- Human Resources
- Finance and Accounting
- Operations and Supply Chain
- Customer Service
-

Business: Finance and Accounting

- **Fraud Detection:** Identifying and preventing fraudulent transactions by analyzing patterns in financial data. This is widely used for credit card transactions and insurance claims.
- **Algorithmic Trading:** Using ML models to execute trades at high speeds based on market data and predictive models.
- **Credit Scoring & Risk Assessment:** Evaluating a loan applicant's creditworthiness or assessing the financial risk of investments.
- **Automated Data Entry:** Extracting and processing information from invoices, receipts, and other financial documents to reduce manual work.

Business: Operations and Supply Chain

- **Demand Forecasting:** Predicting future product demand to optimize inventory levels and production schedules.
- **Predictive Maintenance:** Forecasting when machinery or equipment is likely to fail, enabling maintenance to be scheduled proactively to prevent costly downtime.
- **Inventory Management:** Optimizing stock levels to avoid overstocking or stockouts, balancing supply and demand efficiently.
- **Route Optimization:** Calculating the most efficient routes for delivery vehicles, considering factors like traffic, weather, and delivery windows to save time and fuel.

Business: Human Resources

- **Resume Screening:** Automatically scanning and shortlisting job applications based on keywords, skills, and experience to speed up the hiring process.
- **Employee Turnover Prediction:** Identifying employees at risk of leaving the company, helping HR to address issues and improve retention.
- **Performance Analysis:** Analyzing employee performance data to identify high-potential individuals and areas for training and development.

Business: Customer Service

- **Chatbots and Virtual Assistants:** Providing 24/7 automated customer support by answering common questions and resolving simple issues.
- **Sentiment Analysis:** Analyzing customer feedback from reviews, social media, or surveys to gauge public opinion and identify areas for improvement.

Summary

What's Data Mining

What Kind of Data Can Be Mined?

What Kinds of Patterns Can Be Mined

What's Big Data?

Applications