COSC 757 Data Mining

About Me – Contact Info.

Dr. Weixian Liao
Associate Professor,
Computer and Information Sciences Department,
Towson University

- Office: YR-470
- Phone: 410-704-2774
- Email: wliao@towson.edu
- Office Hour: Wednesday 3:00 PM - 4:00 PM and by appointment
- YR 470 or Online: https://towson.webex.com/meet/wliao

Teaching Assistant

- Name: Papa Pene
- Email: ppene1@students.towson.edu
- Office: YR 224
- Office Hour: Tuesday 3:00 PM – 4:00 PM and by appointment

About Me - Education

- Ph.D. in Computer Engineering, Case Western Reserve University, 2018.
  - Dissertation Title: “Security and Privacy in Cyber-physical Systems.”
- M.S. in Computer Engineering, Mississippi Sate University, MS, 2015.
- B.S. in Information Engineering, Xidian University, China, 2012.

About Me - Research

- Data science, machine learning and deep learning
- Cybersecurity in Cyber-physical systems and IoT

https://scholar.google.com/citations?user=cqwOuaEAAAAJ&hl=en&oi=ao

About You

- Name
- Background/Experience?
- Data mining experience?
- Programming experience?
- **Research interests**?
- **What you hope to get out of this course**?
- Interesting story?

General Course Description

- **Understand** the field of data mining and knowledge discovery in data (KDD)
- **Become** familiar with the foundations of data mining from a number of perspectives
- **Explore** cutting-edge research in data mining
- **Gain** hands-on experience with data mining tools
- **Perform** data mining experiments with real-world datasets

Course Topics

- Introduction to data mining & Applications of data mining
- Exploratory data analysis
- Data preprocessing
- Frequent itemset mining
- Pattern evaluation methods and visualization of frequent patterns
- Advanced pattern mining techniques and applications
- Classification methods such as decision trees, Bayes methods, rule-based methods,
- Classification model evaluation, and visualization of classification results
- Advanced classification methods and applications
- Cluster analysis including partitioning-based methods, hierarchical methods, density-based methods, grid-based methods.
- Cluster evaluation techniques
- Applications of clustering, such as Outlier detection
- Spatial and temporal data mining and visualization techniques
- Cutting edge data mining trends and research directions

Detailed Course Objectives

- Describe the **key components** of the **knowledge discovery process**.
- Understand the key **theoretical underpinnings** of data mining and data analytics
- Understand the **feasibility, usefulness, effectiveness, and scalability** of various approaches to data mining
- Perform **exploratory analysis** on datasets through visualization and descriptive statistics
- **Pre-process and clean data** for various data mining approaches
- **Formulate data mining problems** and **choose** the proper **algorithms** and **evaluation methods** for knowledge discovery.

Course Requirements

- Class Preparation
  - Attendance is mandatory.
  - Please complete the assigned readings before class.
  - Ask questions and participate!
- Article Summary and Presentation
  - Each Student will be required to select a scholarly data mining research article from a data mining conference or journal.
  - Students must write a summary of the article and present it to the class (20-min presentation and 5-min discussion).
- Labs
  - Labs are practice for assignments
  - We will be using the R software and/or Python

Course Requirements

- Assignments
  - Complete **4** data mining experiments during the semester
  - Write **1** conference-style paper detailing their experiment
- Final Project
  - **Individual** project
  - An **original** research paper on a topic covered in class or something new not covered in class.
  - Paper should be formatted for a conference or workshop (IEEE or ACM)
  - The projects will be discussed throughout the semester.
  - I have some ideas/datasets if you need inspiration

Course Grading

- **4** Assignments – 40%
  - Each is 10%
- **1** Paper Summary and Presentation – 25%
  - Written paper summary report: 12.5%
  - Paper presentation: 12.5%
- **1** Final Project – 32%
  - Proposal, midterm progress report, final report, and presentation
  - Each is 8%
- Attendance and participation – 3%

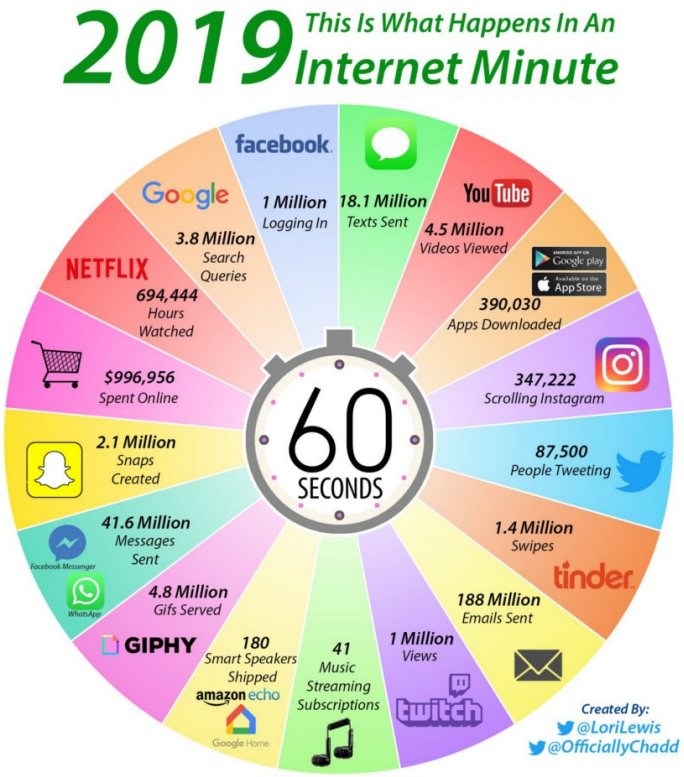
  - **No midterm exam!!!**
  - **No final exam!!!**

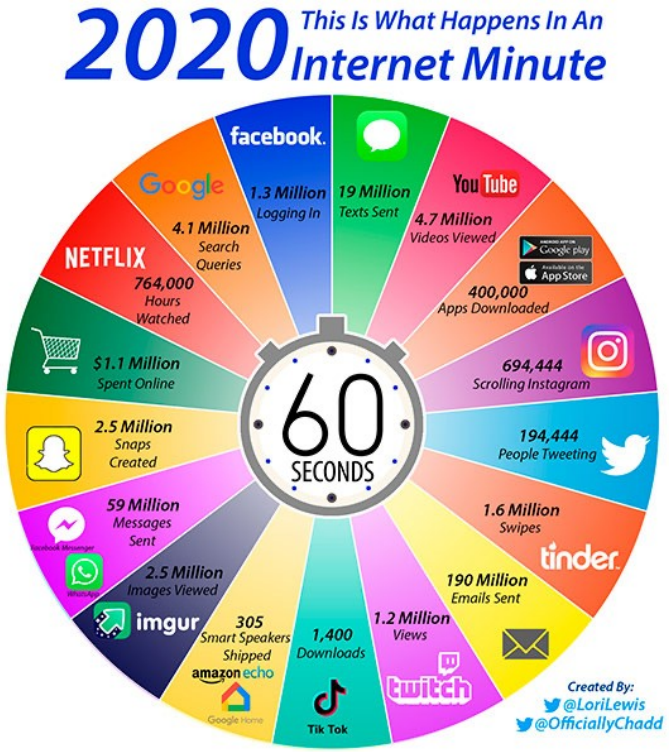
Course Information

- Course Website:
  - Blackboard
  - If you do not have access to this site, please see me!

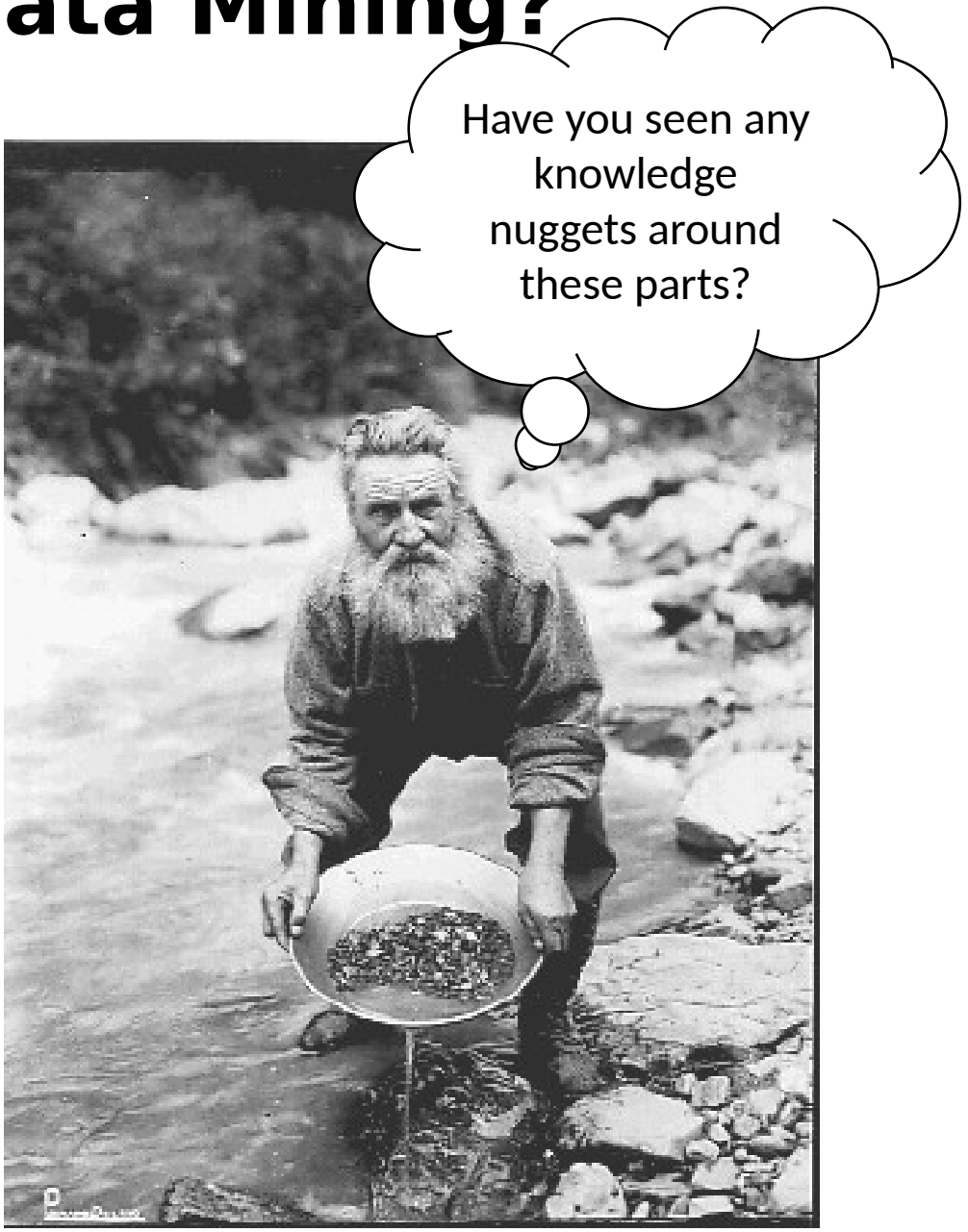
Introduction to Data Mining

“Necessity is the Mother of Invention”





What is Data Mining?



What is Data Mining?

- What is old?
  - Economists, statisticians, forecasters, and communication engineers have long worked with the idea that there exists **patterns** in data

- What is new?
  - Increase for opportunities in finding **patterns** in data

Why?

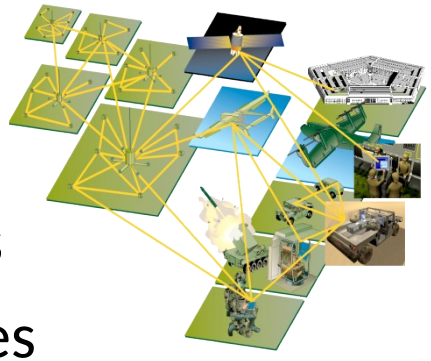
Why?-- Commercial Viewpoint

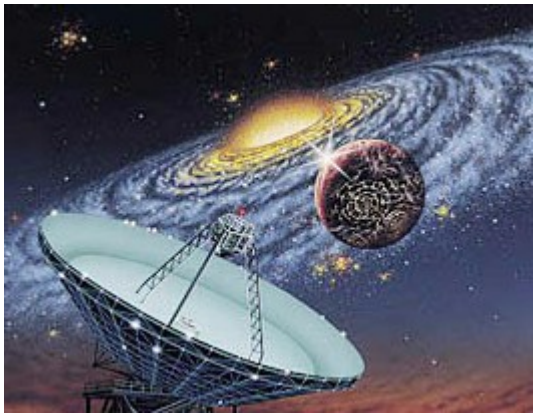
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

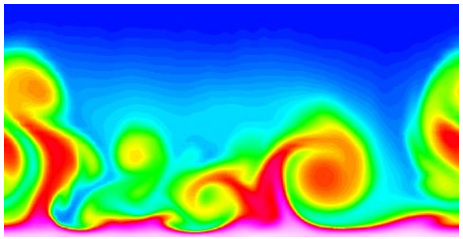
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

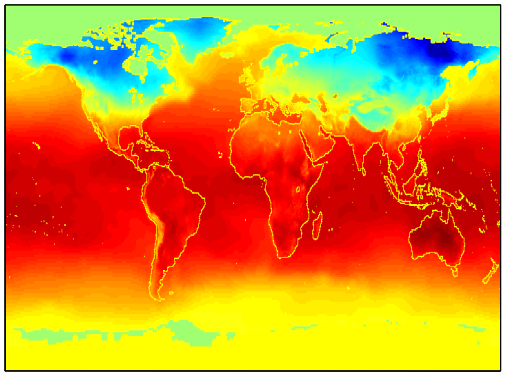
Why?-- Scientific Viewpoint

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









Buzz Words

- Big Data
- Data Mining
- Data Analytics
- Data Warehousing
- Data Science
- Can you relate any current events to these terms?

Data Analysis, Mining, Analytics, and Science: What is it?

- These are all terms that you will hear that generally refer to the same thing.

  - ✓ knowledge discovery from data (KDD)
  - ✓ knowledge discovery in databases
  - ✓ knowledge extraction
  - ✓ data/pattern analysis
  - ✓ data archeology
  - ✓ data dredging
  - ✓ business intelligence

Evolution of Sciences

- Before 1600, **empirical** science
- 1600-1950s, **theoretical** science
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational** science
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data** science
  - The *flood* of data from new scientific instruments and simulations
  - The *ability* to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and 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
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

More Formally What is Data Mining?

- According to the text:
  - **"The process of discovering useful patterns and trends in large datasets."**

- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Before You Ask… Here is where I stand.

- Quote from Witten et al.
- "What is the difference between **[data mining]** and **statistics**? Cynics looking wryly at the explosion of commercial interest (and hype) in this area, equate data mining to statistics plus marketing. In truth, you should not look for a dividing line between [data mining] and statistics because there is a **continuum** – and a multidimensional one at that – of data analysis techniques. Some derive from the skills taught in standard statistics courses, and others are more closely associated with the kind of machine learning that has arisen out of computer science."

Fallacies of Data Mining

- Fallacy 1: There are data mining tools that we can turn loose on our data repositories, and find answers to our problems.

- Reality: There are no automatic data mining tools which will mechanically solve your problems while you wait. As you will see, data mining is a **process**.

(Jen Que Louie, President of Nautilus Systems, Testimony Before US House of Representatives, March, 25, 2003))

Fallacies of Data Mining

- Fallacy 2: The data mining process is autonomous, requiring little or no human oversight.

- Reality: Data mining is not magic. Without skilled supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. The wrong analysis can be worse than no analysis. Why?

(Jen Que Louie, President of Nautilus Systems, Testimony Before US House of Representatives, March, 25, 2003))

Fallacies of Data Mining

- Fallacy 3: Data mining pays for itself quite quickly.

- Reality: The return rates vary depending on start-up costs, analysis personnel costs, data warehousing preparation costs, and so on.

(Jen Que Louie, President of Nautilus Systems, Testimony Before US House of Representatives, March, 25, 2003))

Fallacies of Data Mining

- Fallacy 4: Data mining software packages are intuitive and easy to use.

- Reality: As is the case with fallacy 1, you can't just purchase a data mining suite, sit back, and let it solve all of your problems.

(Jen Que Louie, President of Nautilus Systems, Testimony Before US House of Representatives, March, 25, 2003))

Fallacies of Data Mining

- Fallacy 5: Data mining will identify the causes of our business or research problems.

- Reality: Data mining is not magic. The knowledge discovery process will help you uncover patterns of behavior. Humans identify the causes.

(Larose and Larose, Discovering Knowledge in Data: An Introduction to Data Mining 2014)

Fallacies of Data Mining

- Fallacy 6: Data mining will automatically clean up our messy database.

- Reality: Not automatically.
  - As a preliminary phase in the data mining process, data preparation often deals with data that has not been examined or used.
  - Organizations beginning a new data mining operation will often be confronted with the problem of data that has been lying around for years, is stale, and needs considerable updating.

(Larose and Larose, Discovering Knowledge in Data: An Introduction to Data Mining 2014)

Fallacies of Data Mining

- Fallacy 7: Data mining always provides positive results.

- Reality: There is **no guarantee** of positive results when mining data for actionable knowledge. Data mining is not a panacea for solving business problems. But when used properly by people who understand the models involved, the data requirements and the overall project objectives, data mining can provide actionable results.

(Larose and Larose, Discovering Knowledge in Data: An Introduction to Data Mining 2014)

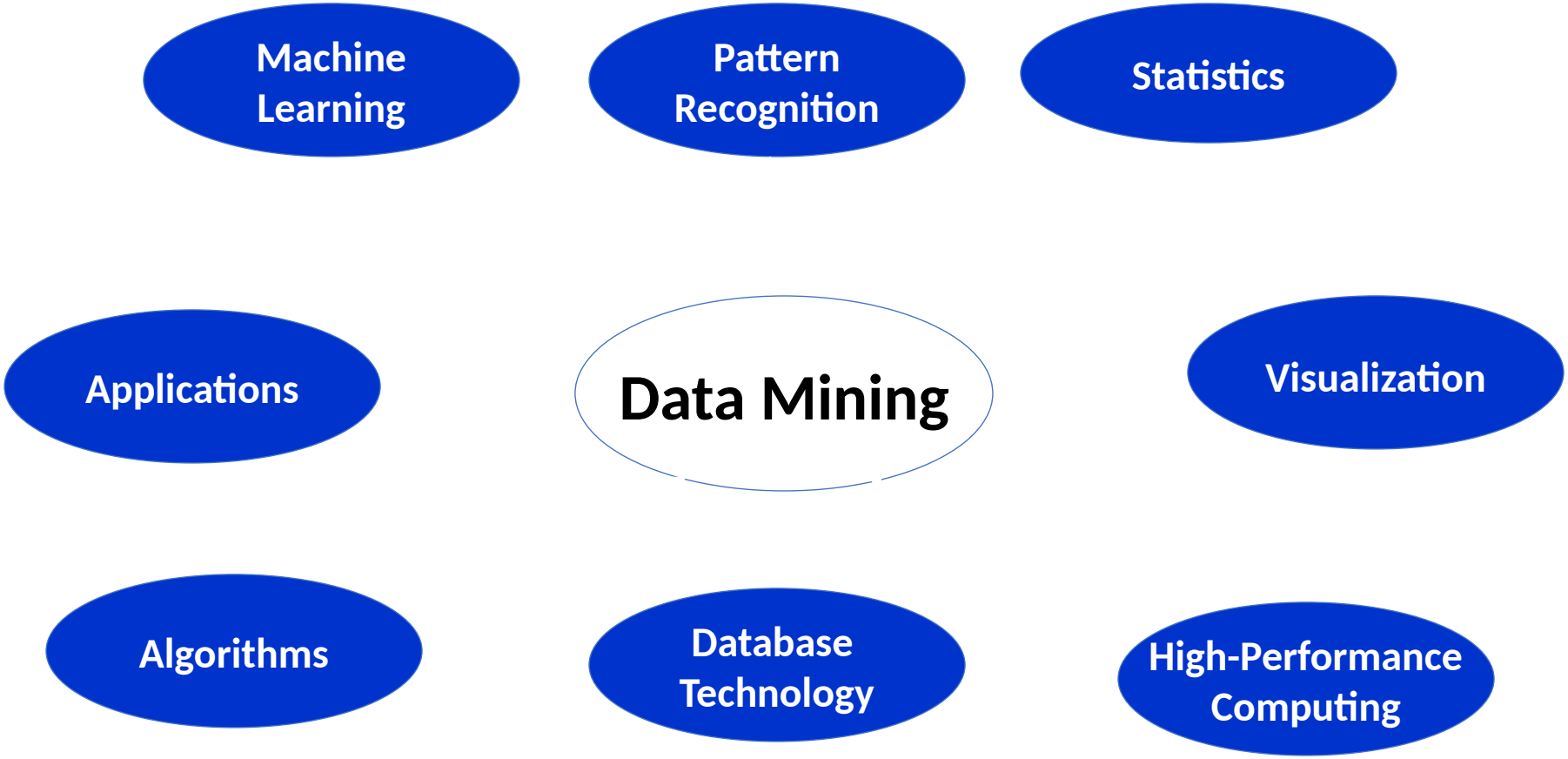
Data Mining as a Discipline

- The need for human direction of data mining:
- Berry and Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support (1997):

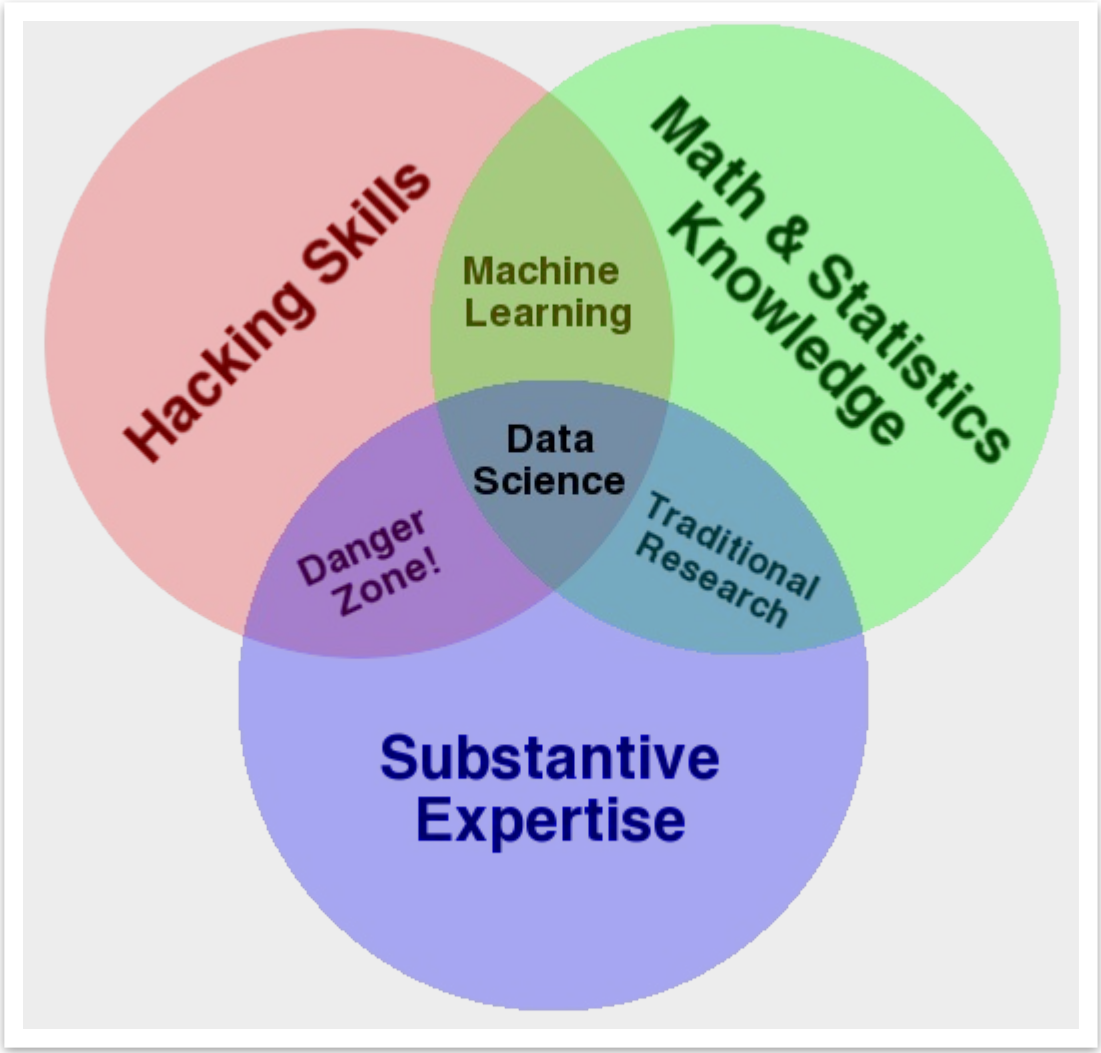
"Data mining is the process of exploration and analysis, by ***automatic or semi-automatic means***, of large quantities of data in order to discover meaningful patterns and rules."

- Berry and Linoff, Mastering Data Mining (2000):
  - "If there is anything that we regret, it is the phrase "by **automatic or semi-automatic means**" …because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered."

Actually, Data Mining is the Confluence of Multiple Disciplines

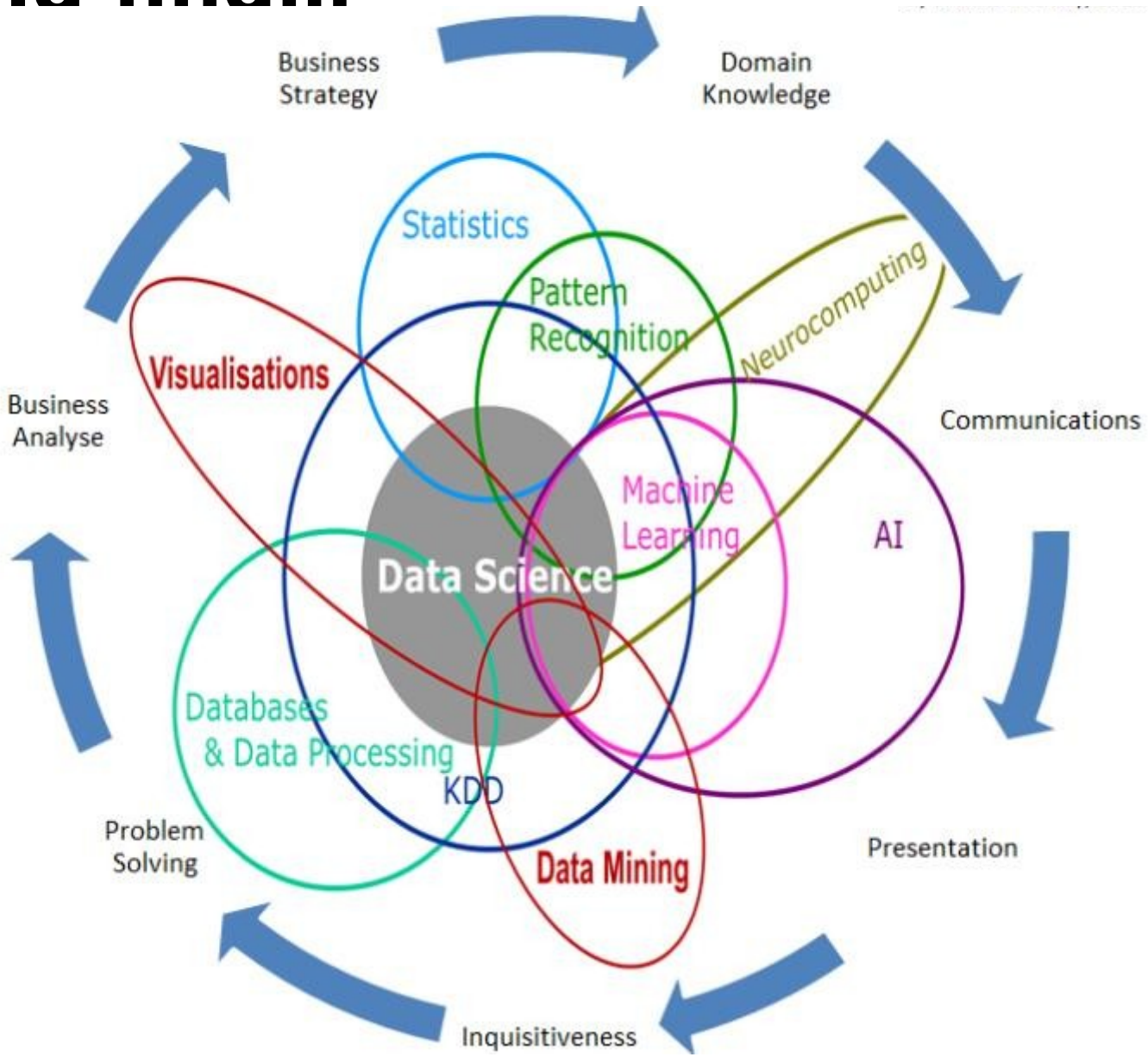


Data Science Venn Diagram

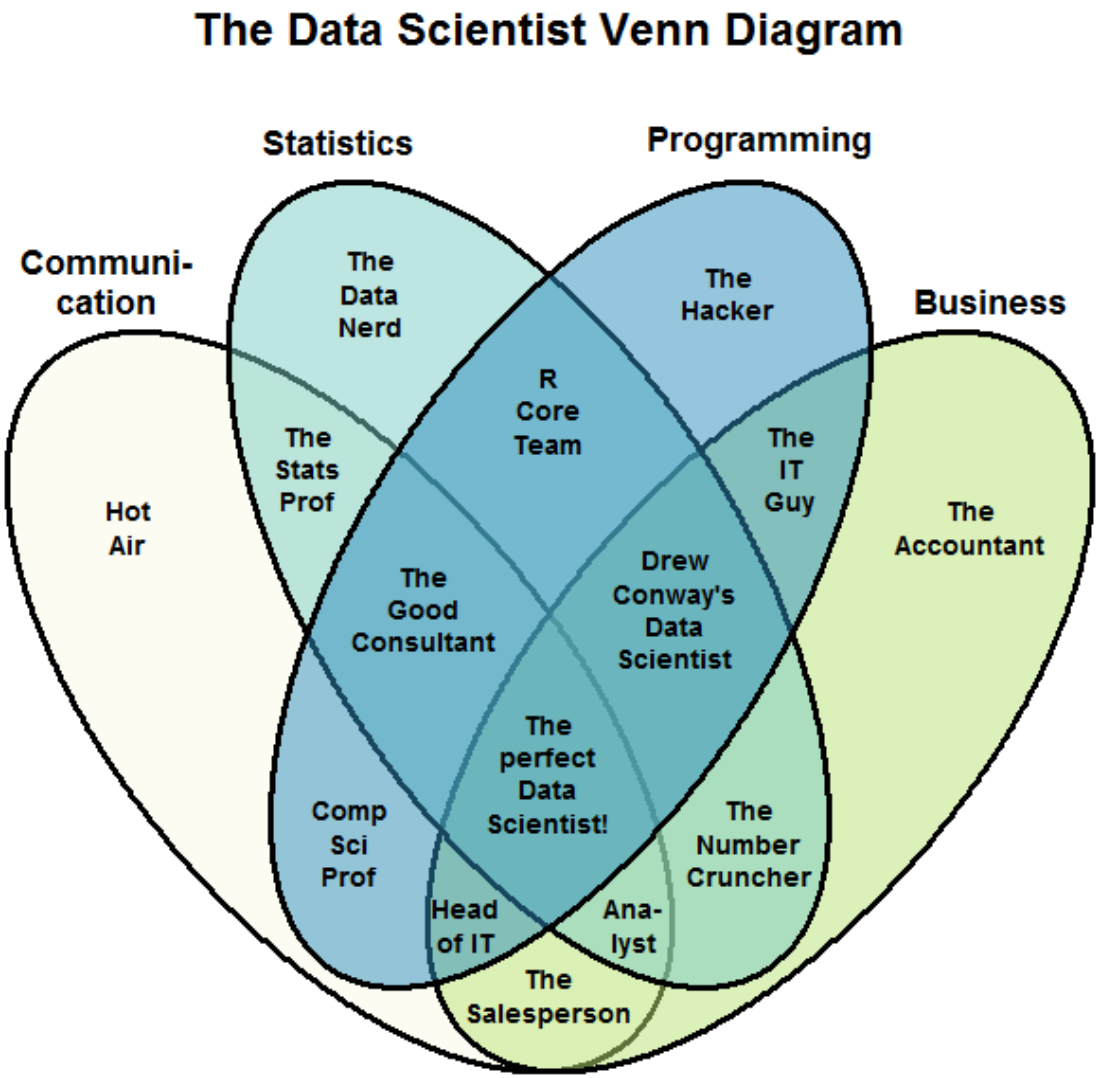


http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

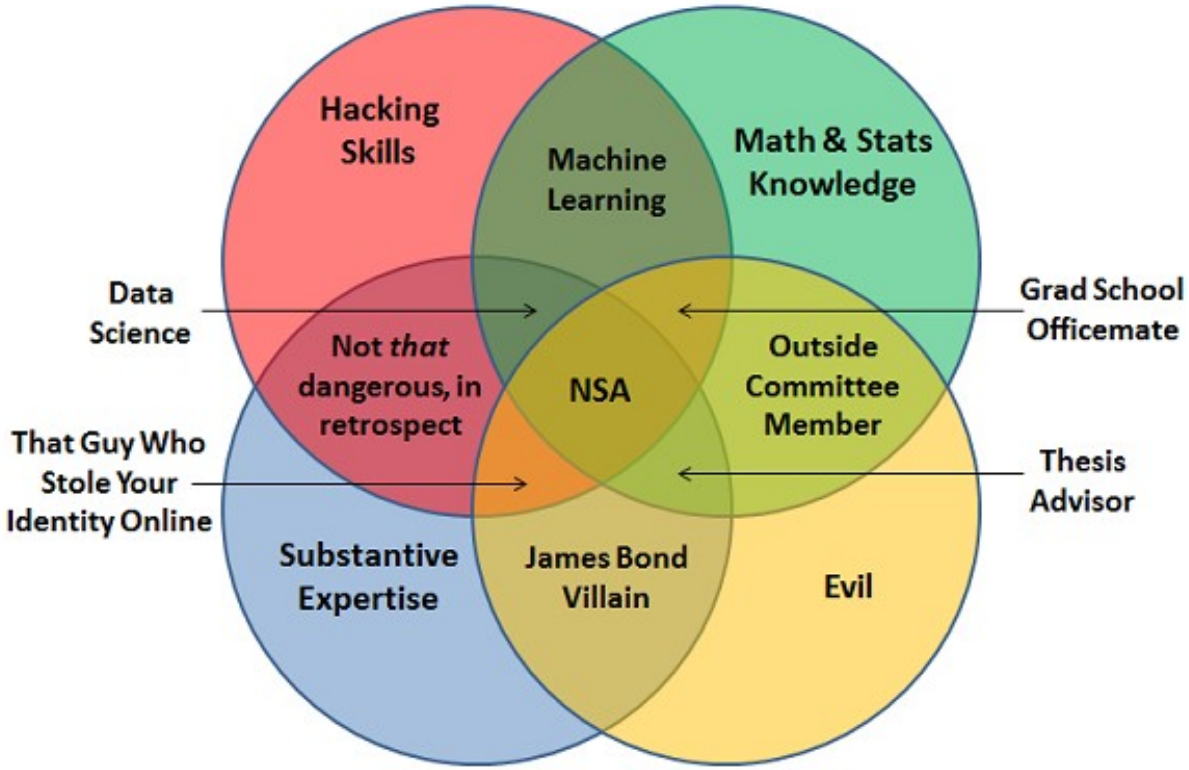
The most complex one that I could find…



In people terms…

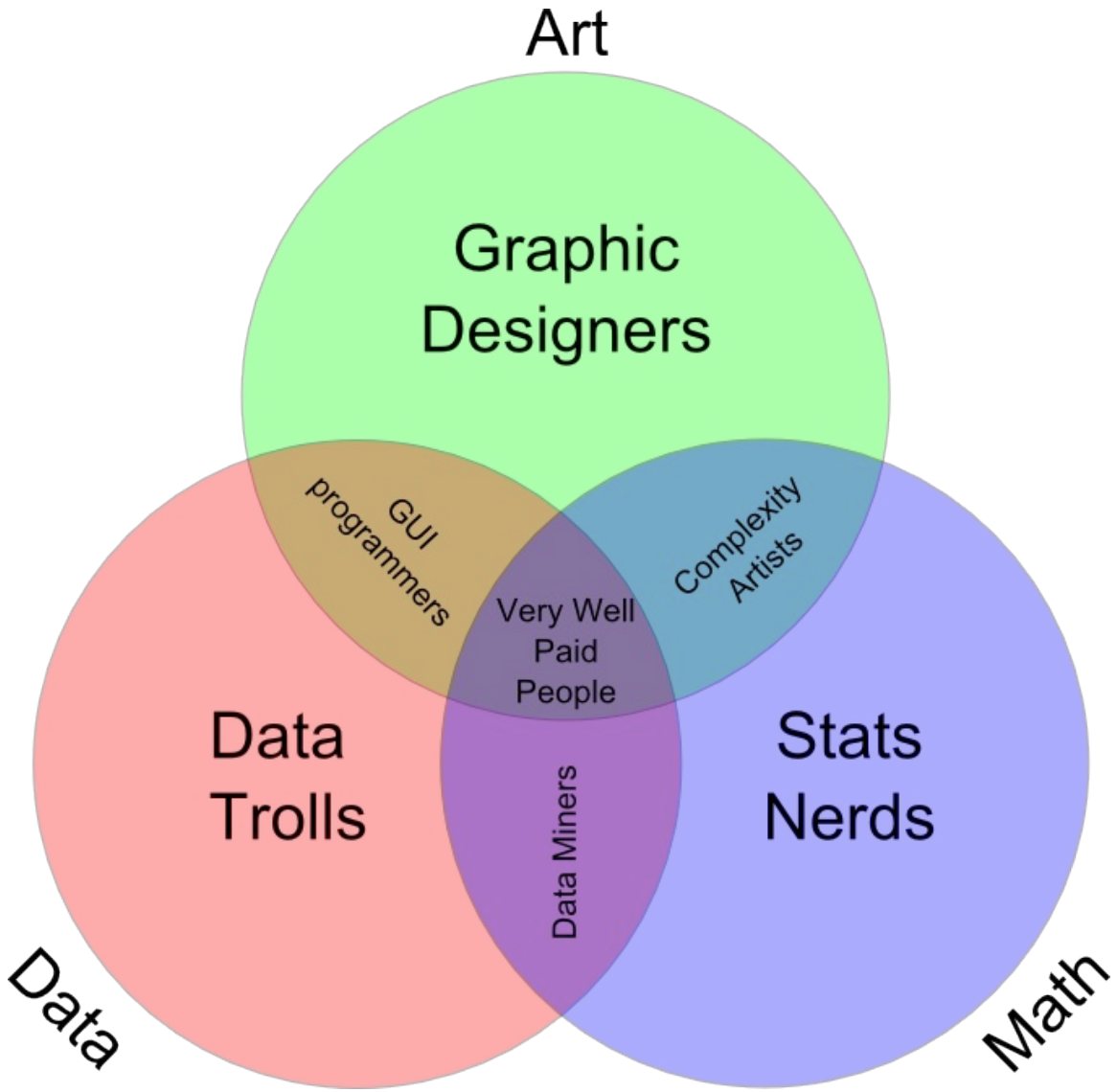


Post Data Science



- http://joelgrus.com/2013/06/09/post-prism-data-science-venn-diagram/

Just for Fun



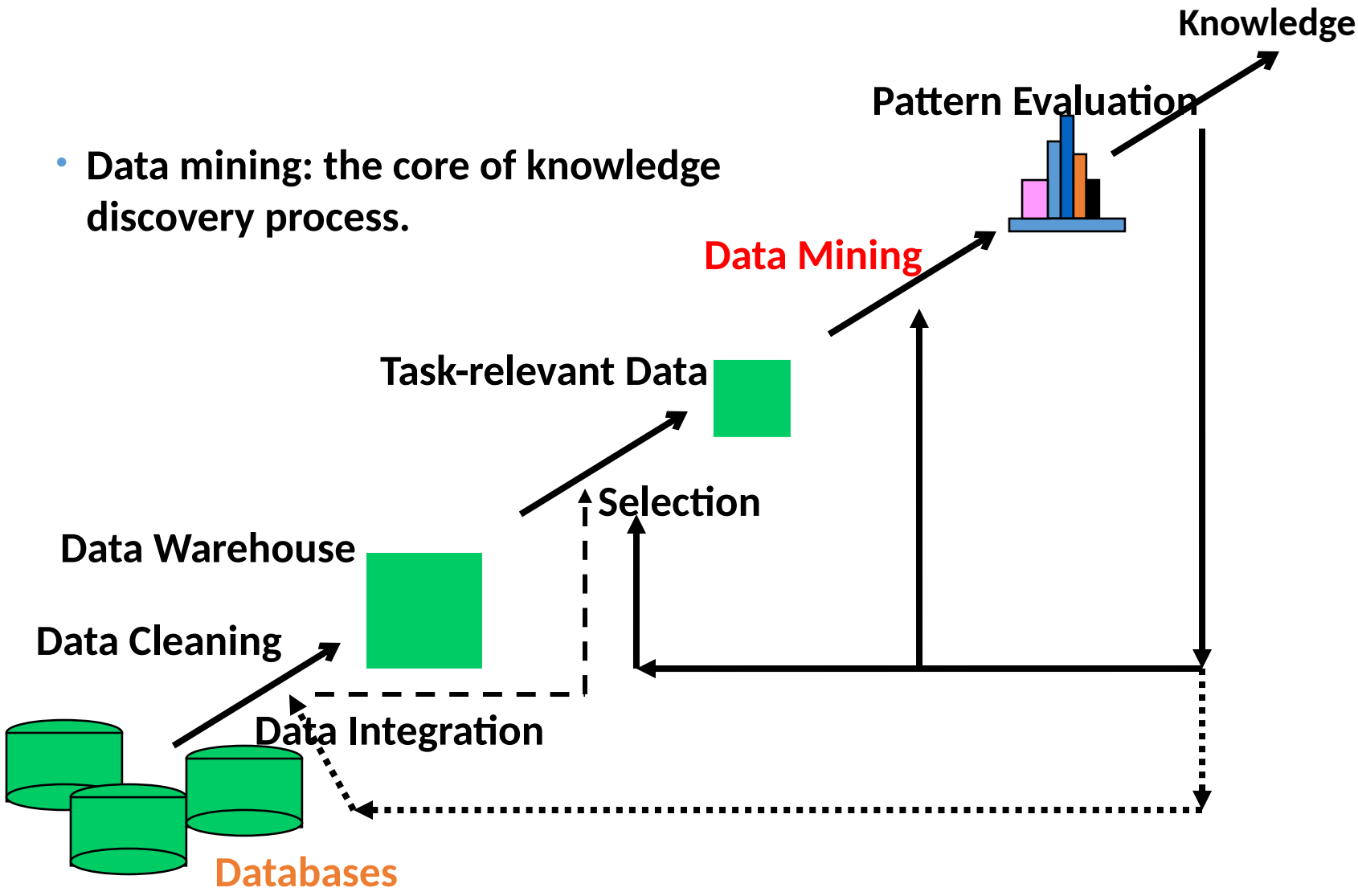
- http://ericksondata.com/wp/2012/venn-diagram-of-data-science/

Why Multiple Disciplines

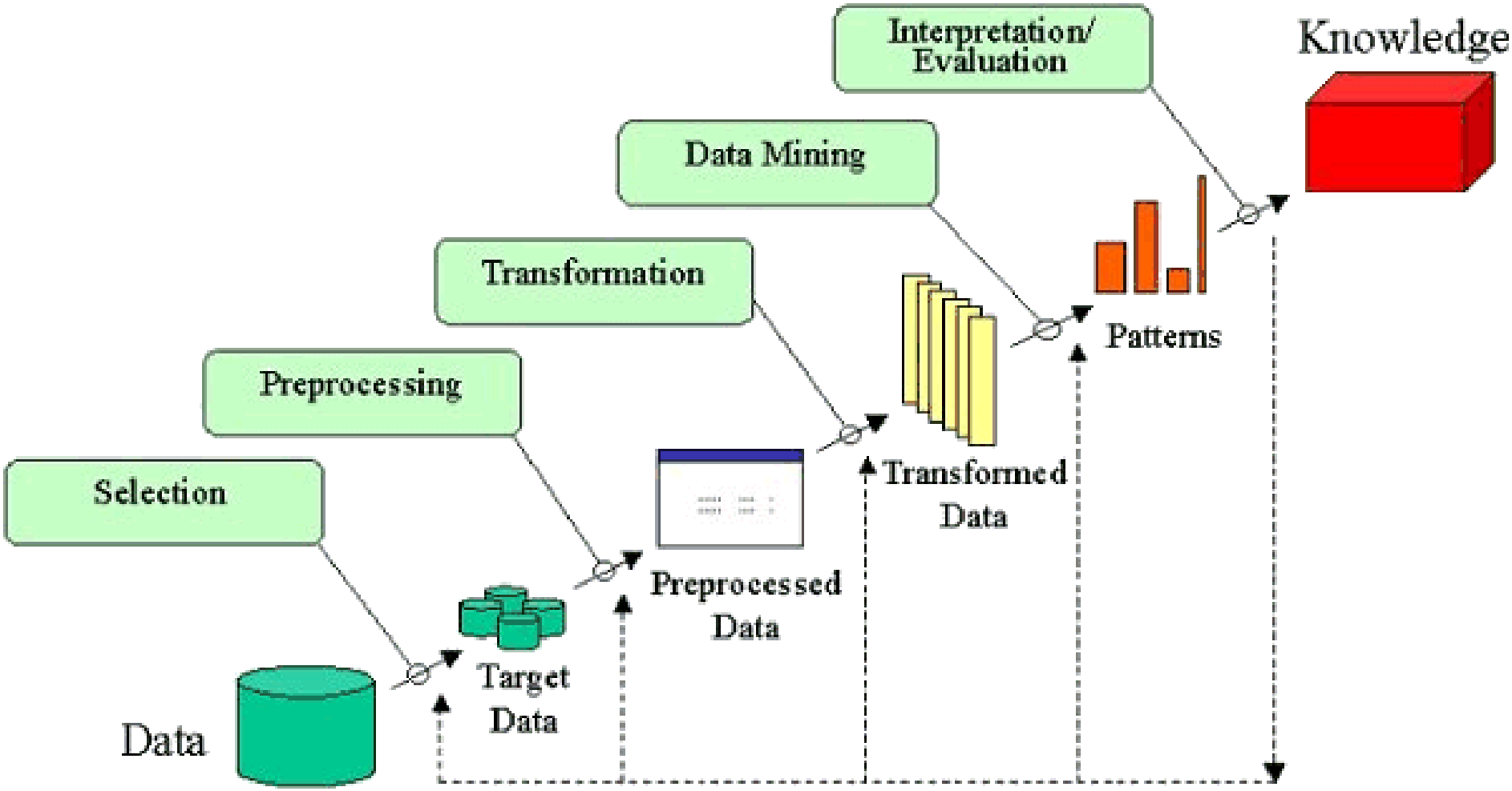
- Tremendous amount of data
  - Algorithms must be highly scalable to handle tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure 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

KDD Process

- Data mining: the core of knowledge discovery process.



KDD Process


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Patterns
Preprocessing
Selection
Transformed
Data
Preprocessed
Data
Target
Data
Data

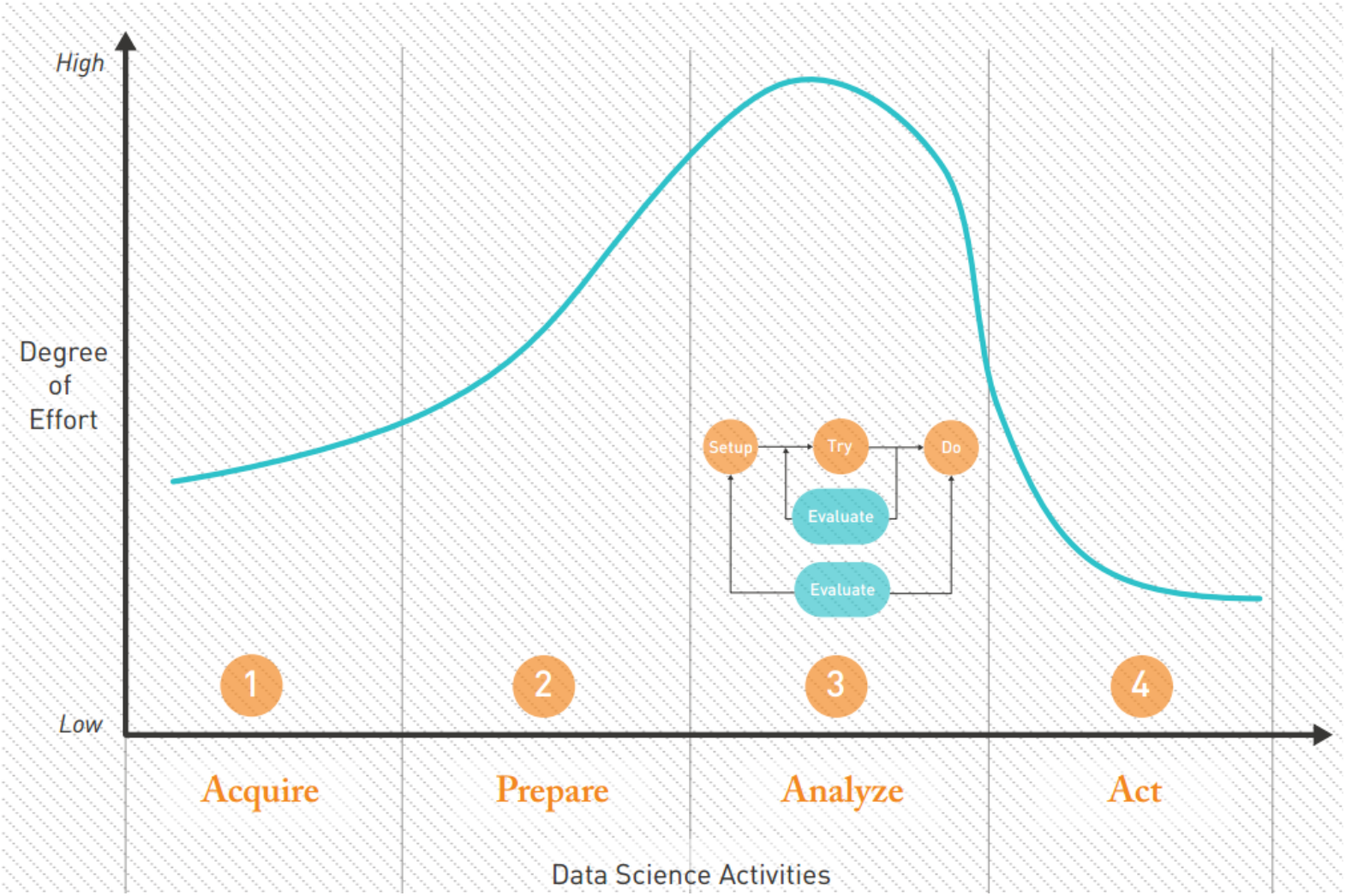
Steps in the KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering, prediction.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - Validation, visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Crisp-DM: 6 Phases

- Business/Research Understanding Phase
- Data Understanding Phase
- Data Preparation Phase
- Modeling Phase
- Evaluation Phase
- Deployment Phase

Data Science Activities



Dimensions of Data Mining

- Data to be mined
- Knowledge to be mined (data mining functions)
- Techniques utilized
- Applications adapted

Data to be Mined

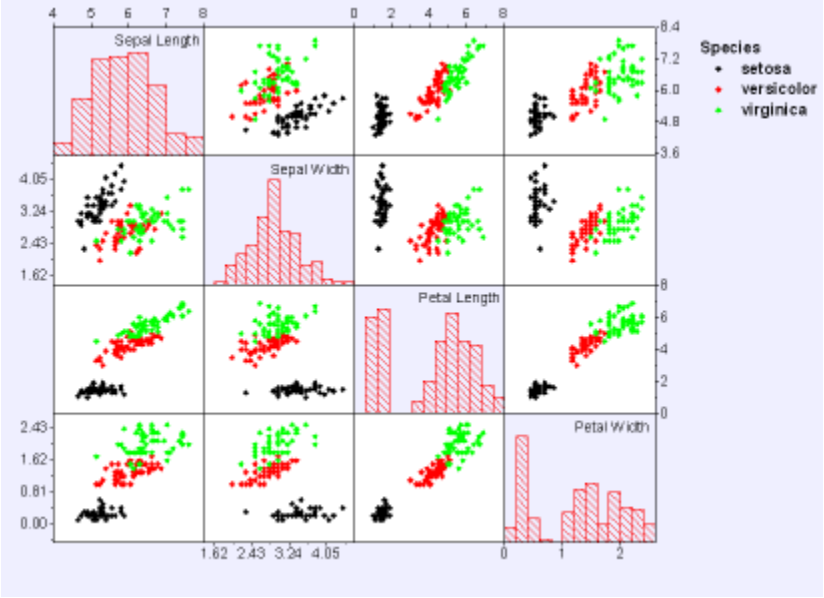
- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW
- Big Data
  - Hadoop
  - MongoDB
  - Neo4j
  - CouchDB….

Knowledge to be Mined (Data Mining Function)

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Description

- Describes patterns or trends in data
  - Descriptions of patterns often suggest possible explanations
  - Example: People who are laid off of work support a certain political candidate/movement
- Methods
  - Descriptive statistics
  - Correlation analysis
  - Regression analysis
  - Clustering
  - Association
  - Exploratory Data Analysis

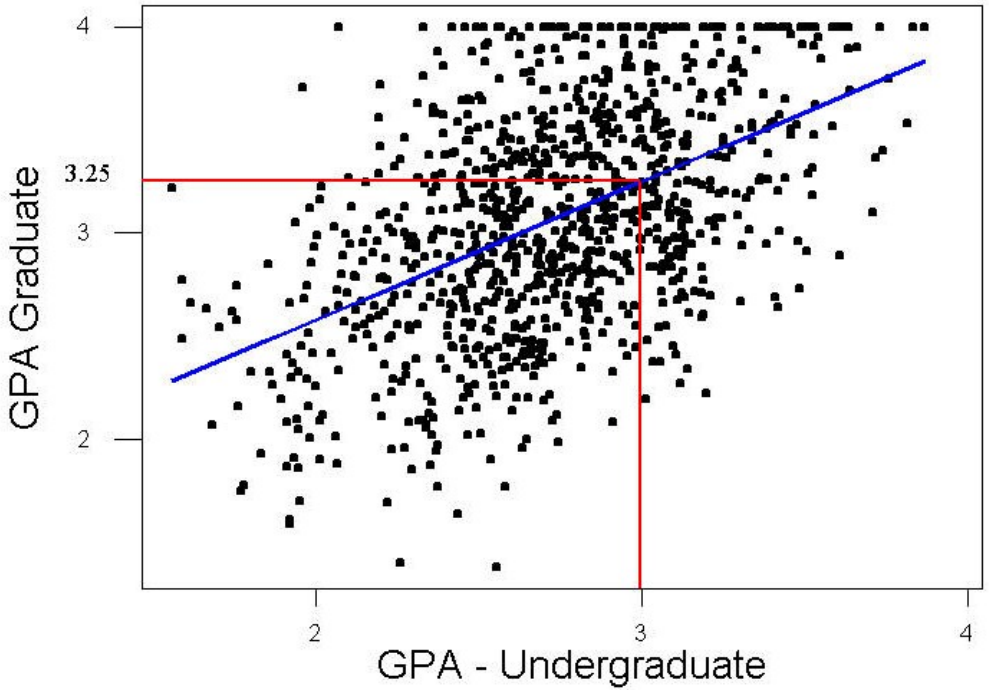


Estimation

- Similar to classification but target variable is numeric
- Models built from complete data records
- For new observations, estimate the target variable
- Examples
  - Estimate a patients blood pressure based on the patient's age, gender, BMI, and sodium levels
  - Estimate GPA of graduate student based on undergraduate GPA
- Methods
  - Point estimation
  - Confidence interval estimation
  - Linear regression
  - Correlation
  - Multiple regression

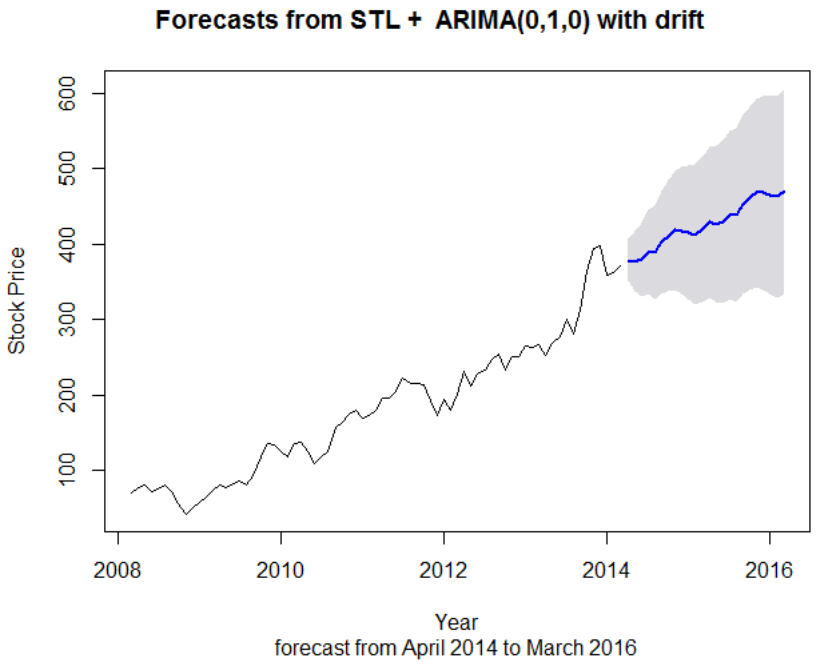
Estimation Example

- Regression line estimate student's ***graduate*** GPA based on their ***undergraduate*** GPA resulting in the following model:
  - $\hat{y} = 1.24 + 0.67x$
- Suppose a student's undergraduate GPA = 3.0
  - GPA = 1.24 + 0.67(3.0) = 3.25



Prediction

- Similar to classification and estimation, except results lie in the future
- Methods used for classification and estimation applicable to prediction
- Forecasting models are also used



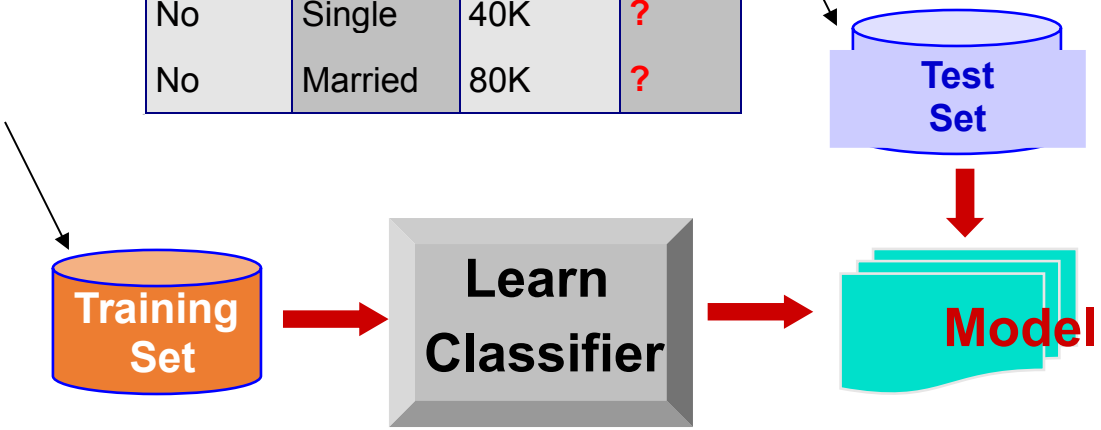
Classification

- Supervised learning – discrete classes are known
- Classification and label prediction
  - Construct models (functions) based on some training examples
- Describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - 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, …

Classification: Experimental Setup

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification Application

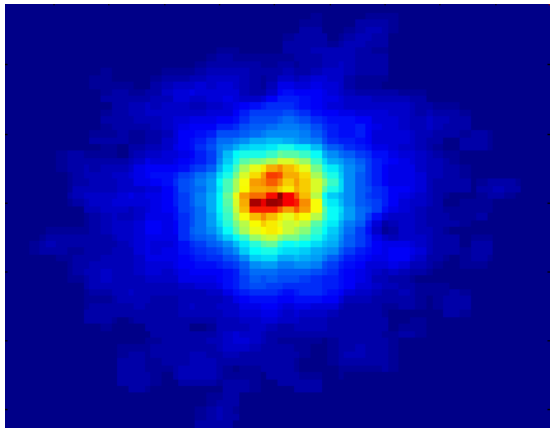
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc?
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification Application

Courtesy: http://aps.umn.edu

http://www.galaxyzoo.org/

Early



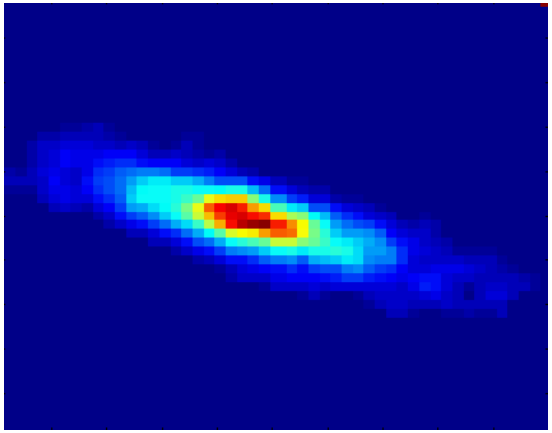
Class:

- **Stages of Formation**

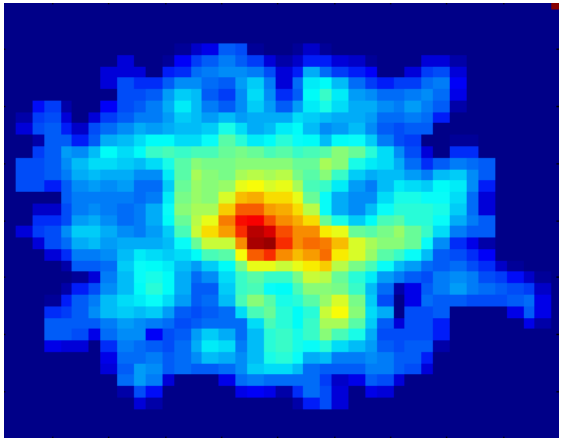
Attributes:

- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:

- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

Clustering

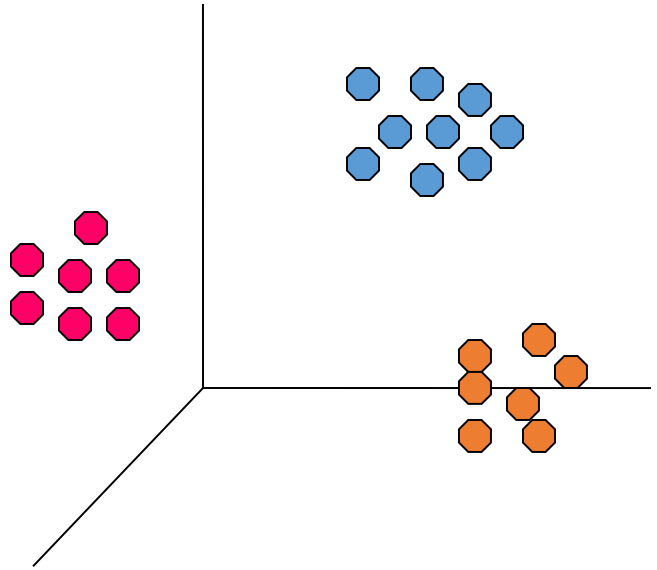
- Unsupervised learning (i.e., Class label is unknown)
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Clustering Illustration

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized

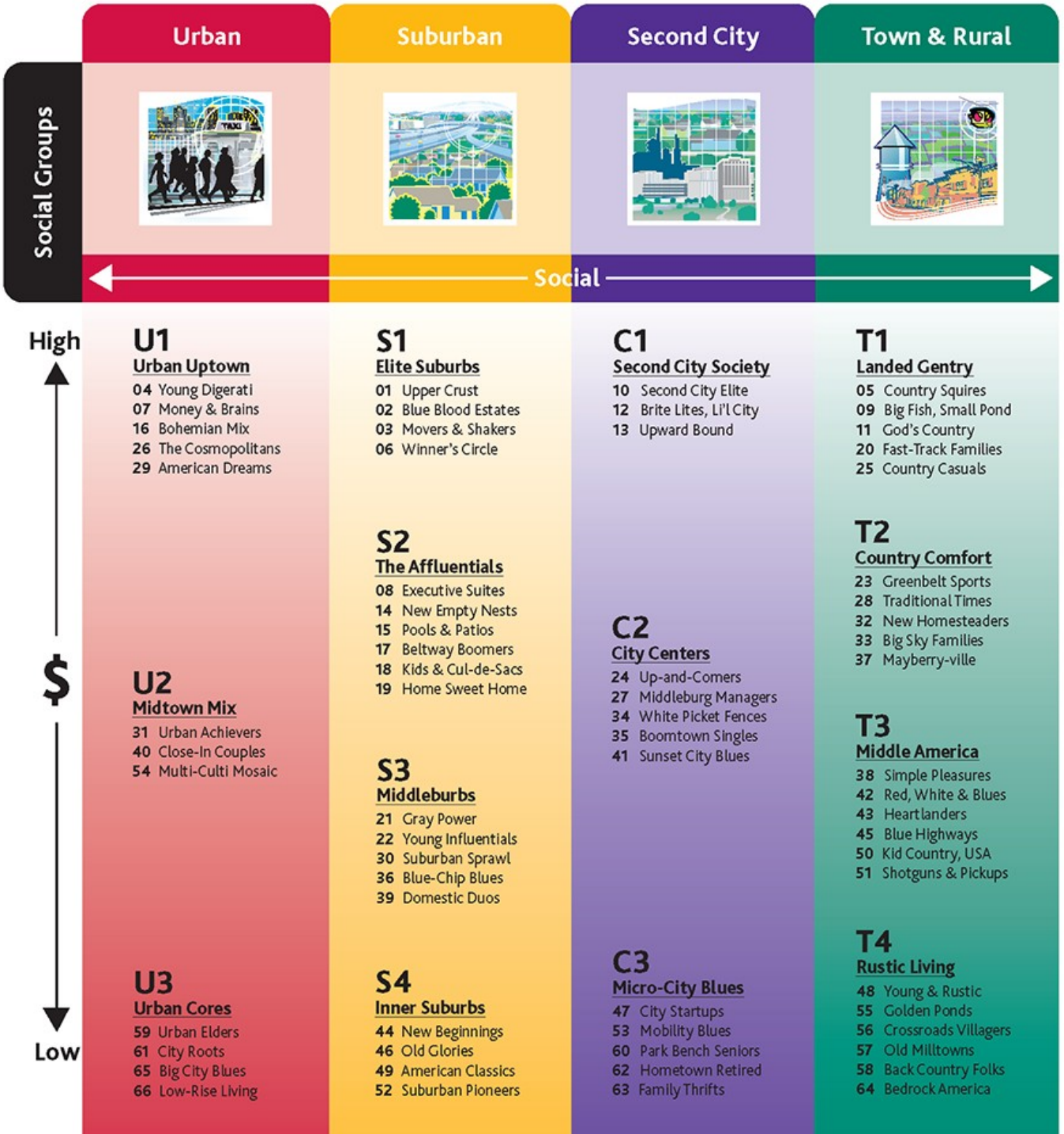


Clustering Applications

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering Applications PRISM Social Groups



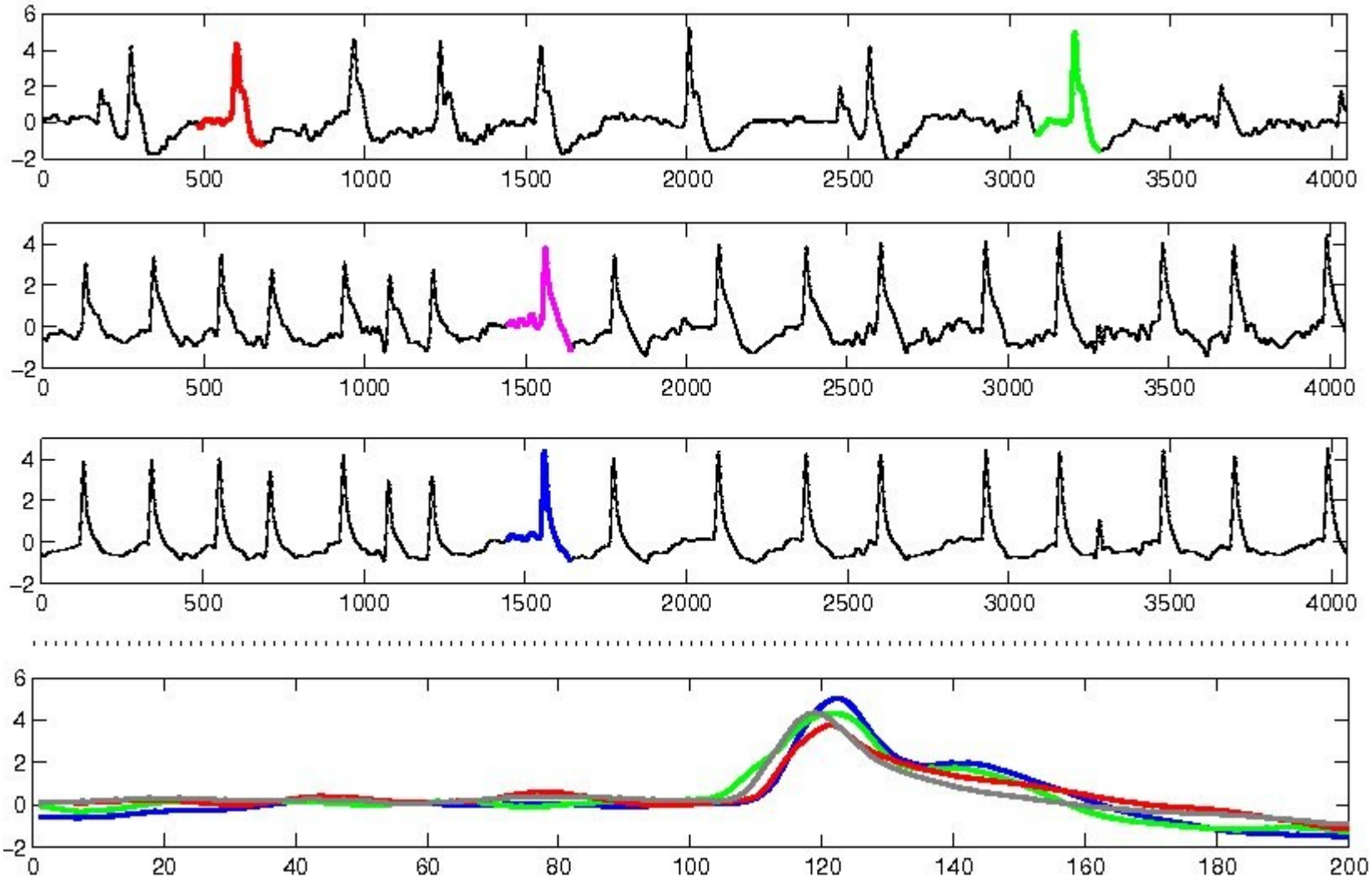
Association: Application

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Time and Ordering

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

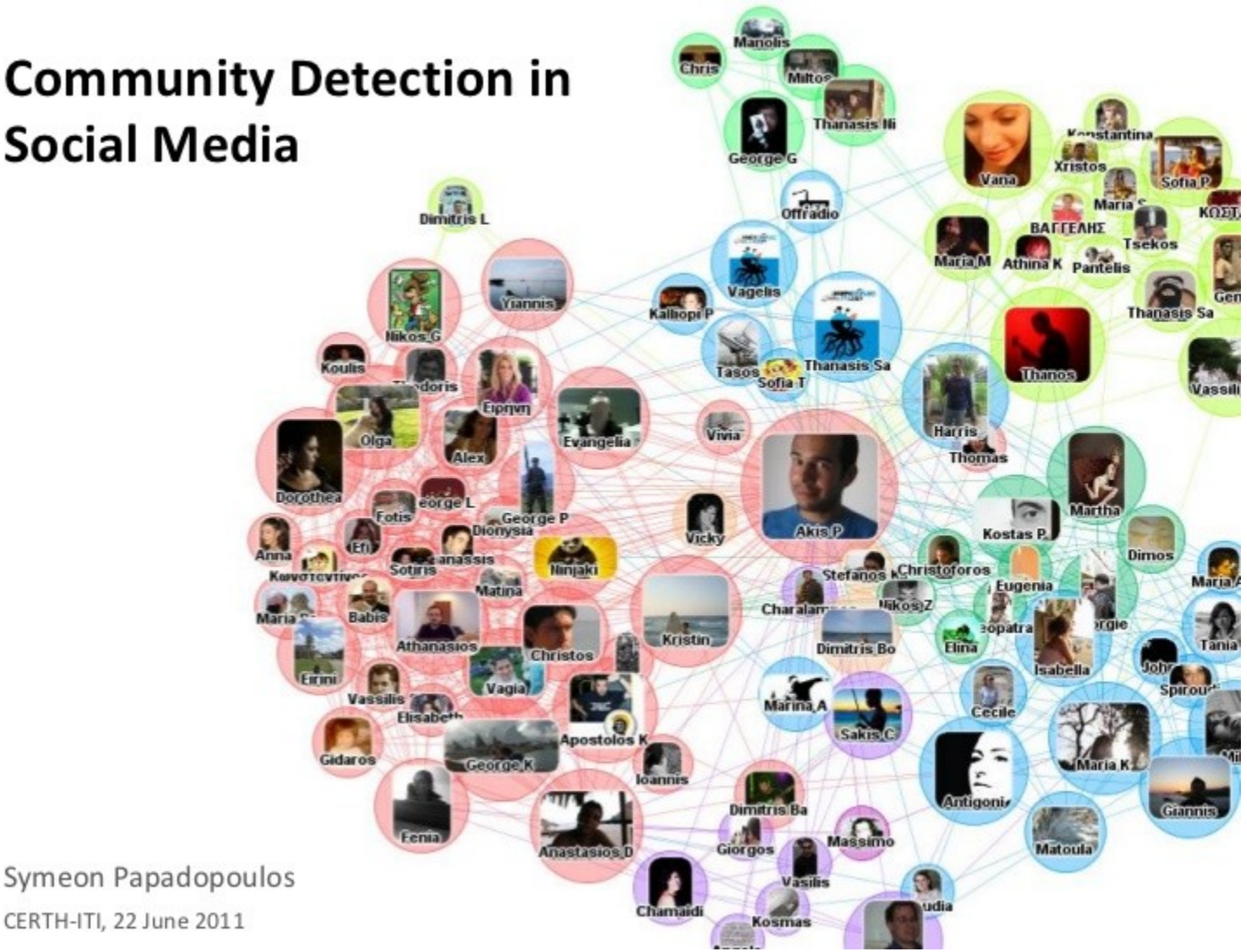
Sequence Example: Time Series Motifs



Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Network Analysis Example: Community Detection



Symeon Papadopoulos
CERTH-ITI, 22 June 2011

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine a tremendous amount of “patterns” and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Mundane vs. novel
  - Accuracy
  - Timeliness
  - …

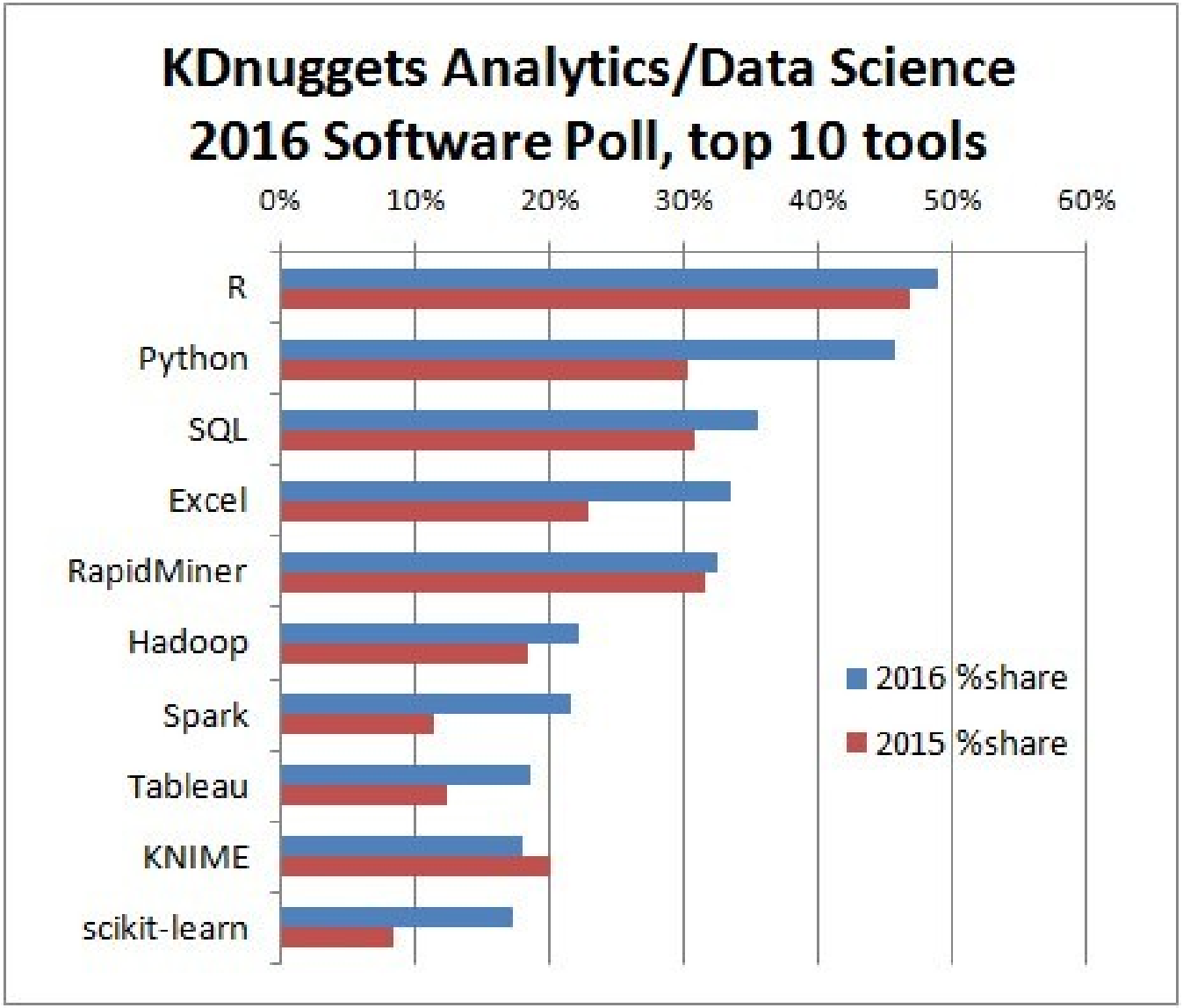
Major Issues and Challenges!

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

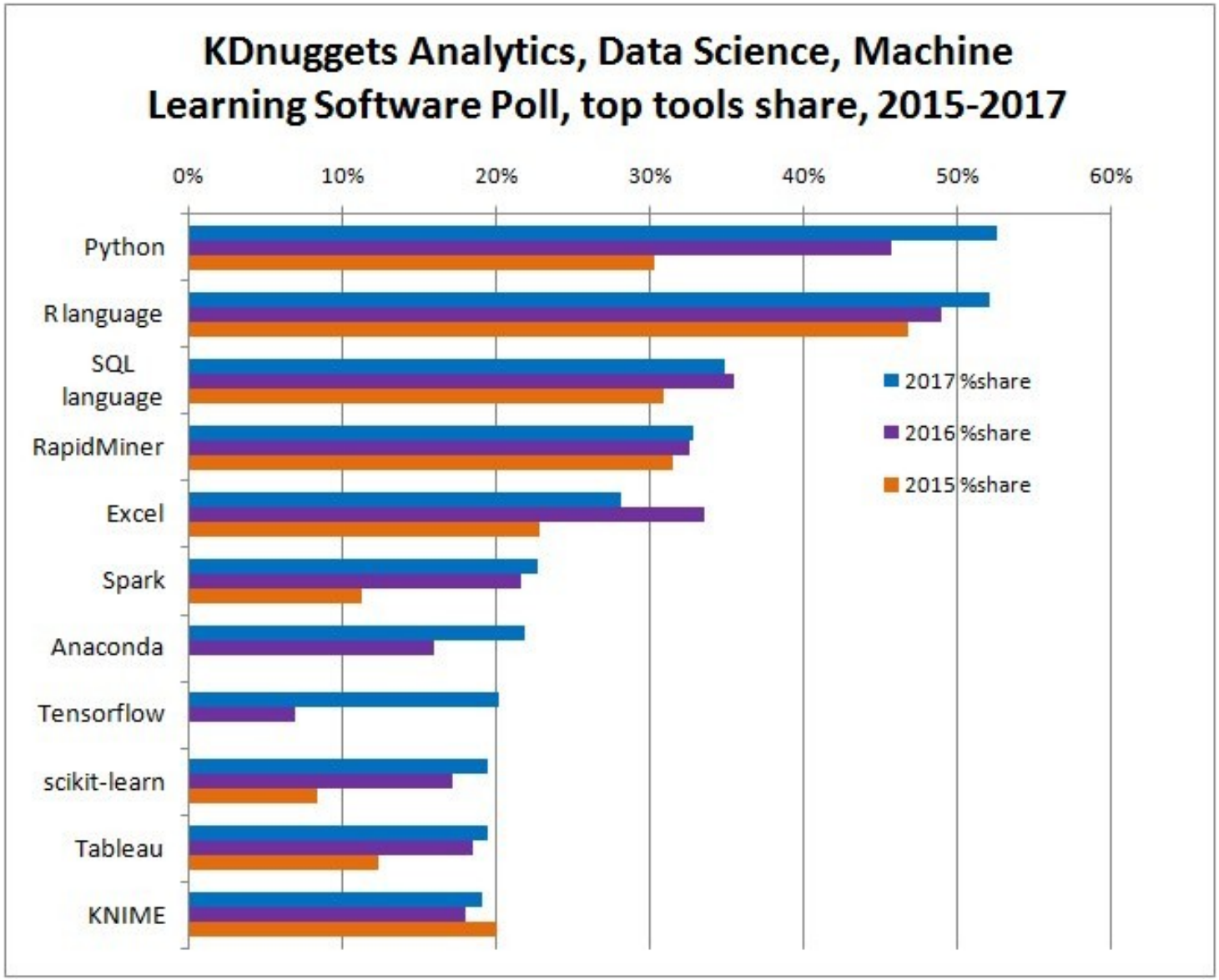
Major Issues and Challenges!

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

Data Mining Software



Data Mining Software



Publications

- **KDD Conferences**
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)

Other related conferences

DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …

Web and IR conferences: WWW, SIGIR, WSDM

ML conferences: ICML, NIPS

PR conferences: CVPR,

Journals

Data Mining and Knowledge Discovery (DAMI or DMKD)

IEEE Trans. On Knowledge and Data Eng. (TKDE)

KDD Explorations

ACM Trans. on KDD

Where to Find Articles

- DBLP, CiteSeer, Google Scholar
- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.