MAS Data Science and Engineering Machine Learning

Natasha Balac, Ph.D.

Predictive Analytics Center of Excellence, Director

San Diego Supercomputer Center

University of California, San Diego

WELCOME

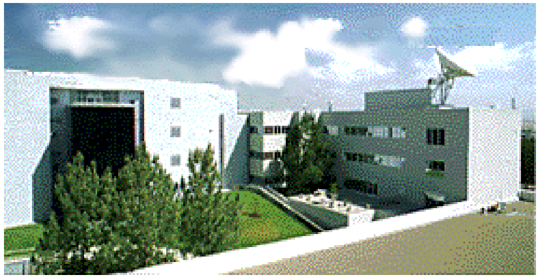
- **Logistics**
  - Check-in from 8:15am – 9:00am
  - Parking
  - Restrooms
  - Lunch
- **Agenda**
  - Technical Sessions
  - Hands-on Sessions
  - Interactive format
  - Assignments & Final
  - TA Hours; Communication; Piazza

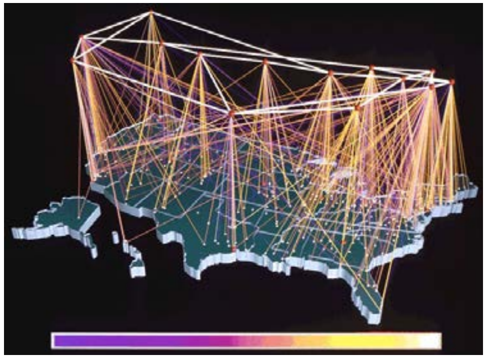
Team

- Anwaya Aras - aaras@eng.ucsd.edu
- Natasha Balac – nbalac@ucsd.edu

Brief History of SDSC

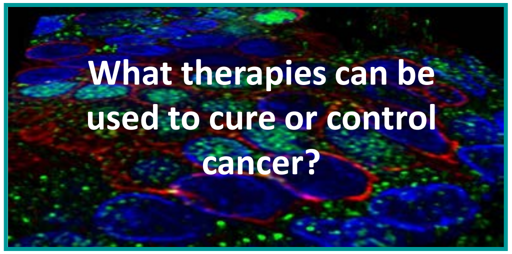
- **1985-1997: NSF national supercomputer center; managed by General Atomics**
- **1997-2007: NSF PACI program leadership center; managed by UCSD**
  - PACI: Partnerships for Advanced Computational Infrastructure
- **2007-2009: Internal transition to support more diversified research computing**
  - still NSF national “resource provider”
- **2009-future: Multi-constituency cyberinfrastructure (CI) center**
  - provide data-intensive CI resources, services, and expertise for campus, state, and nation
- **Approaching $1B in lifetime contract and grant activity**

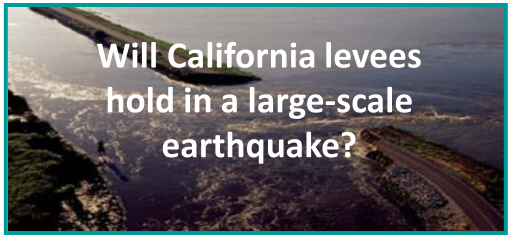




Today's Information Technologies Drive 21st Century Solutions

- **Labs and Centers**
  - Advanced Query Processing
  - Spatial Data Systems
  - Data-Driven High Performance Computing
  - Scientific Workflow Automation Technologies
  - Large-scale Data Systems
  - Predictive Analytics
  - Data Visualization
- **We teach and mentor students**
  - Graph Data Management
  - Semantic Web
  - Machine Learning and Analytics
  - GIS and Spatial Technologies
  - Information Integration
  - Text Analytics
  - Big Data technologies





GORDON

- **A “data-intensive” supercomputer based on SSD flash memory and virtual shared memory**
  - *Emphasizes MEM and IO over FLOPS*
- **A custom-designed system based on COTS to accelerate access to massive data bases being generated in all fields of science, engineering, medicine, and social science**
- **Designed specifically for data-intensive HPC applications**

Gordon Speeds and Feeds

Intel Sandy Bridge Compute node	
Sockets & Cores	2 & 16
Clock speed	2.6 GHz
DRAM capacity and speed	64 GB, 1,333 MHz
Intel710 eMLC Flash I/O Node	
NAND flash SSD drives	16
SSD capacity per drive & per node	16 * 300 GB = 4.8 TB
SMP Super-Node (via vSMP)	
Compute nodes / I/O Nodes	32 / 2
Addressable DRAM	2 TB
Addressable memory including flash	11.6 TB
Gordon (Aggregate)	
Compute Nodes	1,024
Compute cores	16,384
Peak performance	341 TF
DRAM/SSD memory	64 TB DRAM; 300 TB SSD
InfiniBand Interconnect	
Architecture	Dual-Rail, 3D torus
Link Bandwidth	QDR
Vendor	Mellanox
Lustre-based Disk I/O Subsystem (Shared)	
Total storage: current/planned	4 PB/6 PB (raw)
Total bandwidth	100 GB/s

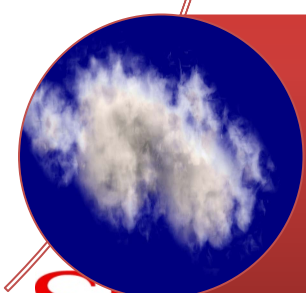
SDSC Repertoire of Storage Systems

Data Oasis (PFS)

- High-Performance Parallel File System for HPC Systems; Partitioned for Scratch and Medium-Term Parking Space
- Access: Lustre on HPC Systems (Gordon, Trestles, Triton)

Project Storage

- Purpose: Typical Project / Home Directory / User File Server Storage Needs
- Access: NFS/CIFS, iSCI



SDSC Cloud

- Storage of Digital Data for Ubiquitous Access and High-Durability
- Access: Multi-platform web interface, S3-type interfaces, backup SW

PACE
Predictive Analytics Center of Excellence

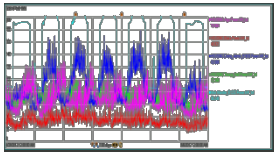
Closing the gap between
Government, Industry and Academia

PACE: Closing the gap between Government, Industry and Academia

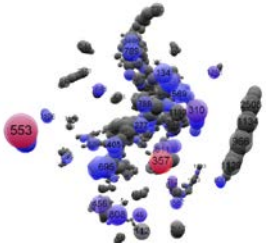
PACE is a non-profit, public educational organization

- To promote, educate and innovate in the area of Predictive Analytics
- To leverage predictive analytics to improve the education and well being of the global population and economy
- To develop and promote a new, multi-level curriculum to broaden participation in the field of predictive analytics

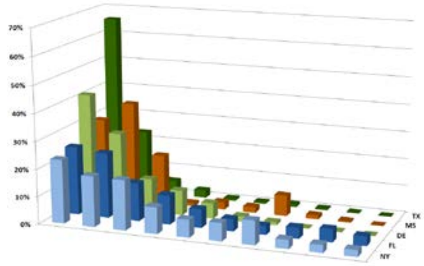
Foster Research and Collaboration



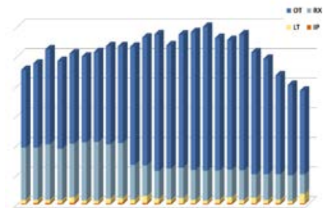
- Fraud Detection
- Modeling user behaviors
- Smart Grid Analytics
- Distributed Energy Generation
- Microgrid anomaly detection
- Battery Storage Analytics
- Sport Analytics
- Genomics

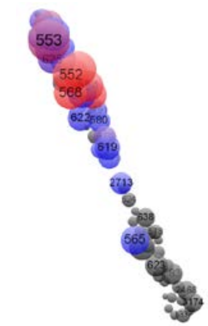


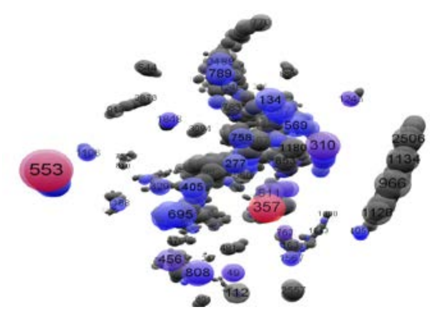
CMS Fraud, Waste and Abuse Detection and Prediction



- **Descriptive Statistics**
  - Claims summary information
  - History and trends
  - Distributions across periods, transactions, etc.
- **Exploratory Analysis**
  - Profiles of provider transactions
  - Provider similarity according to profiles
  - Visual summaries of large amounts of data
  - Eligibility data link to provider billing
- **Predictive analytics**
  - Adjustments
  - Equipment, Service Codes
  - Long term vs. short term hospital stay
  - Provider profiles



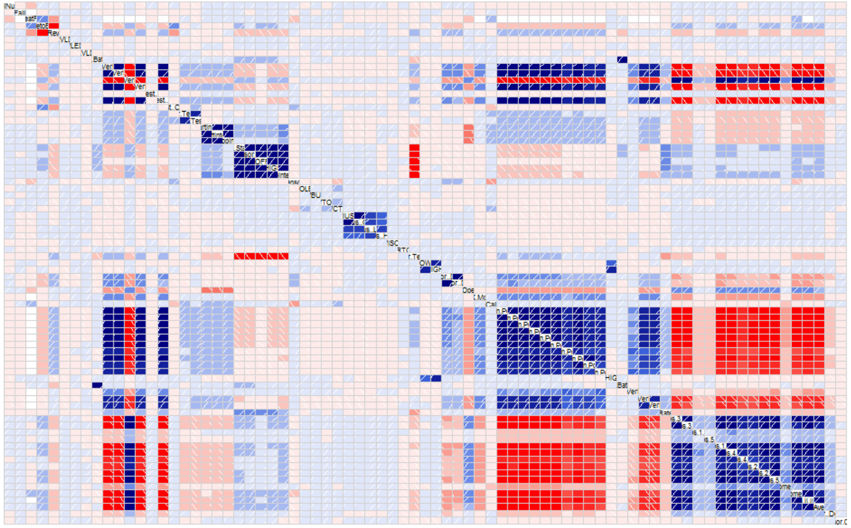


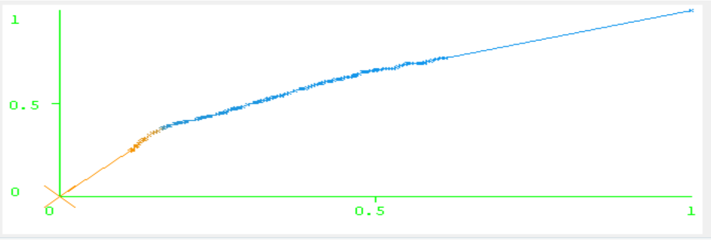


Predictive Analytics In Action

Sports Analytics

Manufacturing



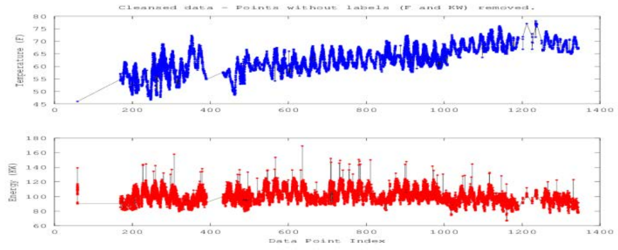


UCSD Smart Grid

- **UCSD Smart Grid sensor network data set**
  - **45MW peak micro grid; daily population of over 54,000 people**
  - **Self-generate 92% of its own annual electricity load**
- **Smart Grid data – over 100,000 measurements/sec**
  - **Sensor and environmental/weather data**
    - Large amount of multivariate and heterogeneous data streaming from complex sensor networks
  - **Predictive Analytics throughout the Microgrid**

Predictive Analytics for Discovering Energy Consumption Patterns

- **The utility and the consumer both benefit from consumption analytics**
- **Forecasting the energy consumption patterns in the UCSD campus microgrid**
- **Different spatial and temporal granularities**
- **Novel Feature Engineering**
- **Machine learning for demand response optimization**



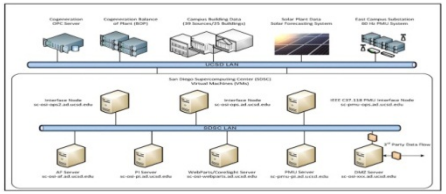
Sustainable San Diego Partnership

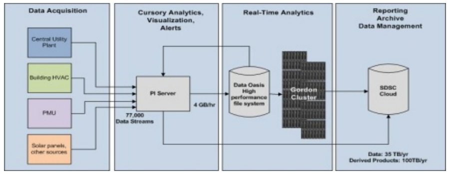
- **Clean Tech San Diego, OSIsoft, SDG&E and UC San Diego Common data infrastructure connects physical assets: electrical, gas, water, waste, buildings, transportation &traffic**
- **Platform to securely transfer high volumes of Big Data from multiple, distributed measurement units**
- **Crowd-sourced Big Data in a cyber-secure, private cloud**
- **Predictive analytics on real-time time-series data**

White House Big Data Event: "Data to Knowledge to Action" – Launch Partners Award

Big Data

- **Complexities introduced by the large amount of multivariate and heterogeneous data streaming from complex sensor networks**
- **Extremely large, complex sensor networks, enabling a novel feature reduction method that scales well**



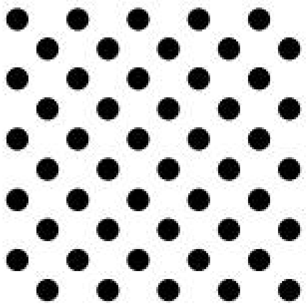



BIG
DATA
ANALYTICS
STORAGE
TECHNOLOGIES
INFORMATION
SOFTWARE
USE
ONE
SAN
TOLERABLE
SIZE
NEEDED
PARALLEL
PETABYTES
INTERNET
PROCESSING
LOGS
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
EVERY
CASE
LARGER
CURRENT
EXAMPLES
APPLIED
SETS
ABILITY
TARGET
SENSOR
ARCHIVES
DIFFICULTY
COMPLEX
DATABASES
RESEARCH
TERABYTES
CONTINUES
COST
CITATION
RECORDS
SEARCH
NETWORKS
INCLUDE
SYSTEMS
SHARED
TIME
DISK
HUNDREDS
BIOLOGICAL
GARTNER
MASSIVELY
USING
WORKING
PERFORMANCE
ALSO
RELATIONAL
TOOLS
WITHIN
PROCESS
DEFINITION
GENOMICS
TENS
CAPACITY
PRACTITIONERS
BUSINESS
MOVING

IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WordPress DOWNLOADS
125+ PLUGIN DOWNLOADS
695,000+ facebook STATUS UPDATES
79,364 WALL POSTS
510,040 COMMENTS
1,700+ Firefox DOWNLOADS
694,445 SEARCH QUERIES
Google
Google Search
168 MILLION EMAILS ARE SENT
1,500+ BLOG POSTS
60+ NEW BLOGS
70+ DOMAINS REGISTERED
25+ HOURS TOTAL DURATION
600+ NEW VIDEOS
You Tube
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
GO-Globe.com web technologies

4 V's of Big Data

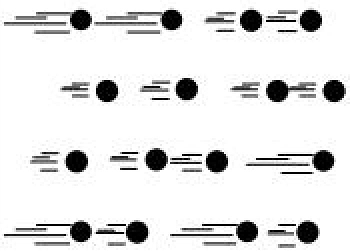
Volume



Data at Rest

Terabytes to exabytes of existing data to process

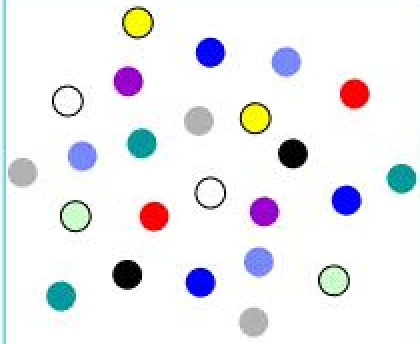
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

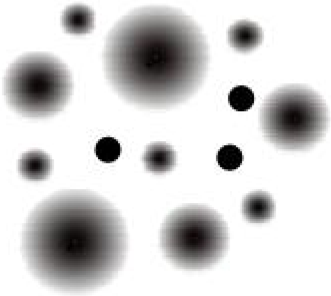
Variety



Data in Many Forms

Structured, unstructured, text, multimedia

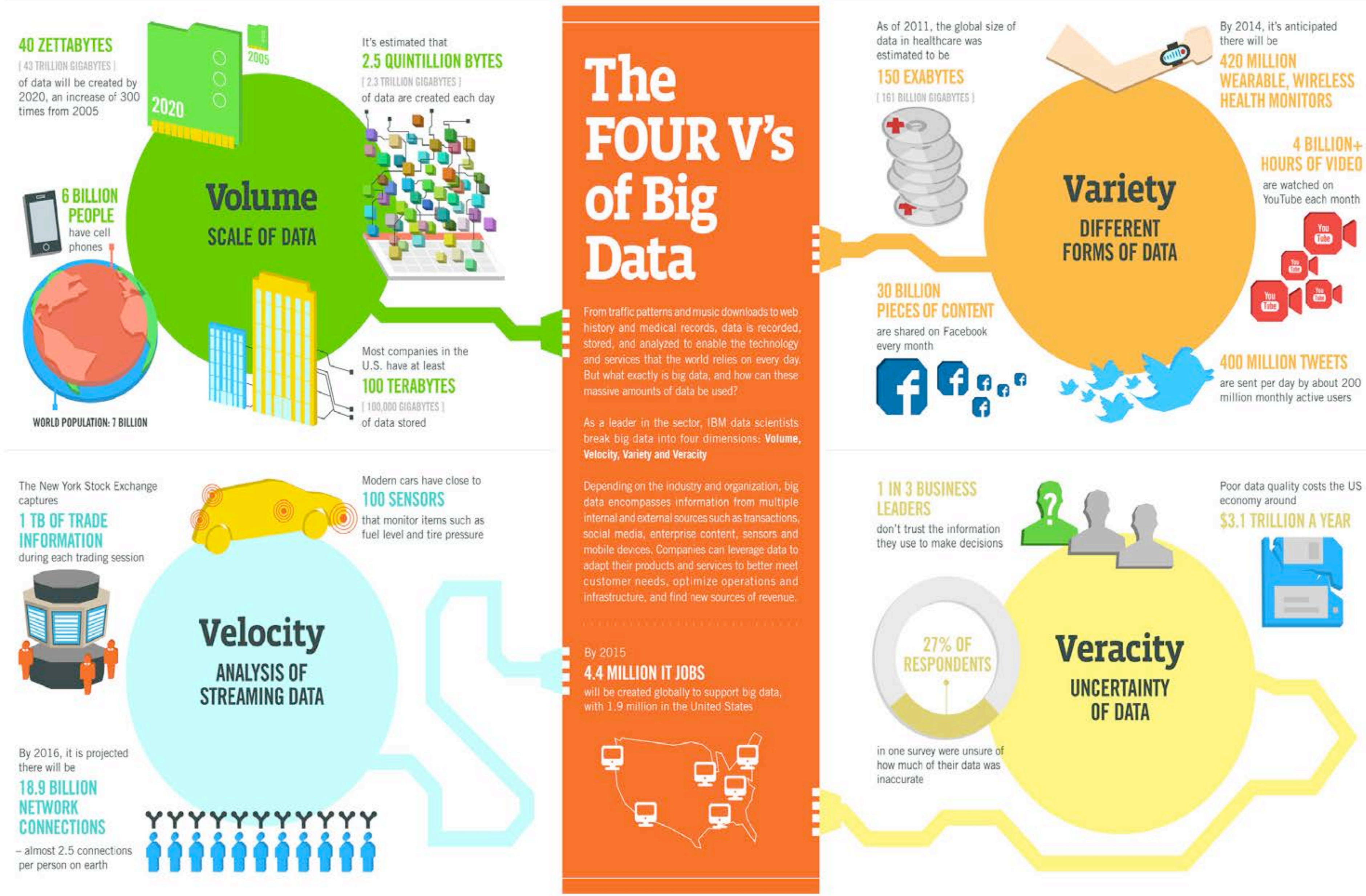
Veracity*



Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

IBM, 2012



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Kmeans Results from 10 million NYTimes articles

cluster means shown with coordinates determining fontsize

7 viable clusters found

Big Data – Big Training

- **"Data Scientist"**
  - **The "Hot new gig in town"**
    - O'Reilly report
  - **Data Scientist: The Sexiest Job of the 21st Century**
    - Harvard Business Review, October 2012
    - The next sexy job in next 10 years will be statistician" – Hal Varian, Google Chief Economist
    - Geek Chic – Wall Street Journal – new cool kids on campus
  - The future belongs to the companies and people that turn data into products
- *"The human expertise to capture and analyze big data is both the most expensive and the most constraining factor for most organizations pursuing big data initiatives" –* **Thomas Davenport**
- **New curriculum – Boot camps, Certificates, Data Science Institute, '14 MAS**

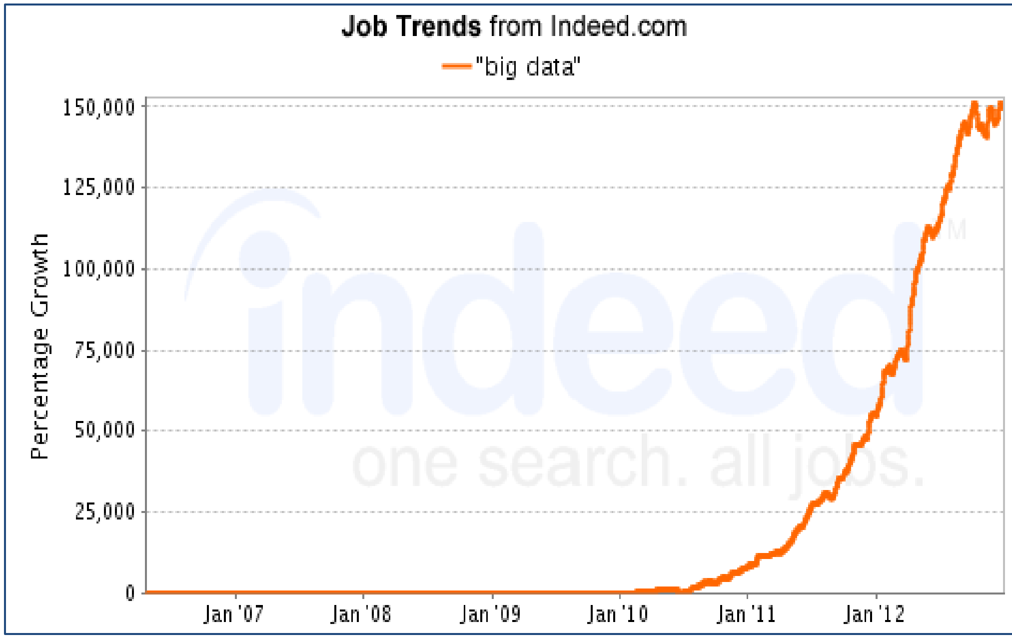
Big Data – Big Data Science

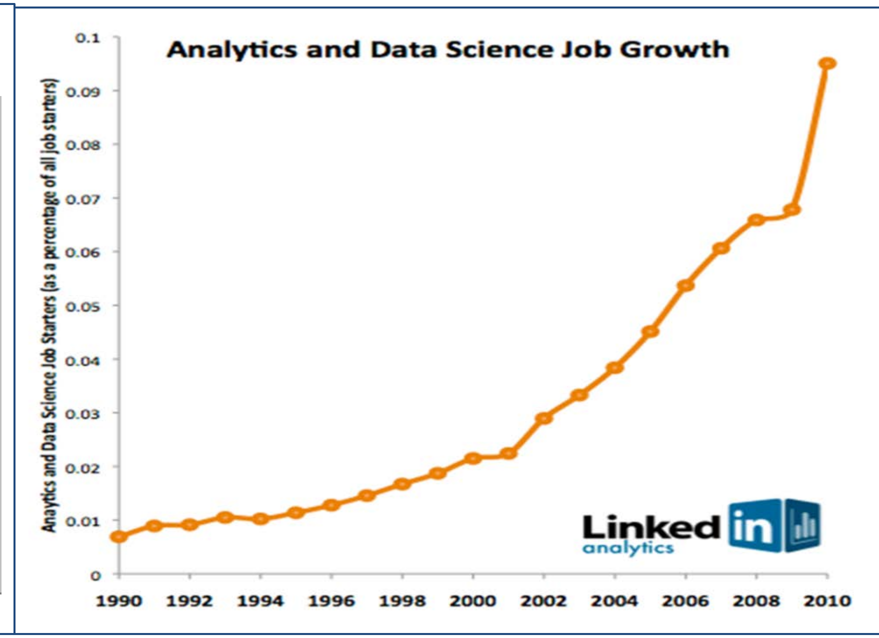
- **"Data Scientist"**
  - **The "Hot new gig in town"**
    - O'Reilly report
  - **Data Scientist: The Sexiest Job of the 21st Century**
    - Harvard Business Review, October 2012
    - The next sexy job in next 10 years will be statistician" – Hal Varian, Google Chief Economist
    - Geek Chic – Wall Street Journal – new cool kids on campus
  - The future belongs to the companies and people that turn data into products
- *"The human expertise to capture and analyze big data is both the most expensive and the most constraining factor for most organizations pursuing big data initiatives"* **– Thomas Davenport**

Data scientist: The hot new gig in tech

- **Article in *Fortune***
  - *"The unemployment rate in the U.S. continues to be abysmal (9.1% in July), but the tech world has spawned a new kind of highly skilled, nerdy-cool job that companies are scrambling to fill: data scientist"*
- ***McKinsey Global Institute* "Big data Report"**
  - By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions

Data Science Job Growth



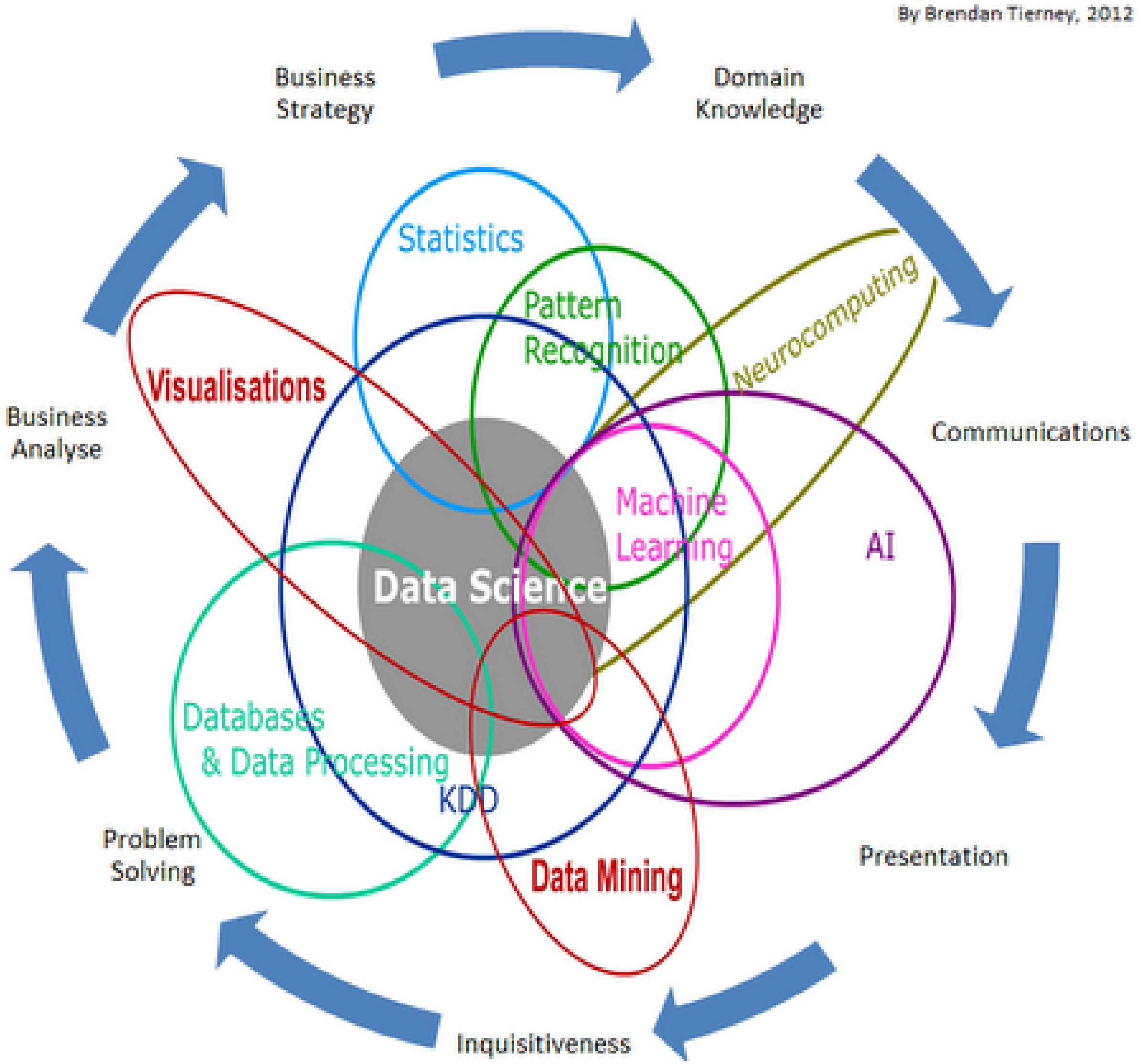


By 2018 shortage of 140-190,000 predictive analysts and 1.5M managers / analysts in the US

Data Miners: Past and Present

- **Traditional approaches have been for DM experts: "White-coat PhD statisticians"**
  - DM tools also fairly expensive
- **Today: approach is designed for those with *some* Database/Analytics skills**
  - DM built into DB, easy to use GUI, Workflows
  - Many jobs available from Statistical analyst to Data Scientist!
- **Data Science: The Art of mathematically sophisticated data engineers delivering insights from data into business decisions and systems**

Data Science Is Multidisciplinary



Successful Data Scientist Characteristics

- **Intellectual curiosity, Intuition**
  - Find needle in a haystack
  - Ask the right questions – value to the business
- **Communication and engagements**
- **Presentation skills**
  - Let the data speak but tell a story
  - Story teller – drive business value not just data insights
- **Creativity**
  - Guide further investigation
- **Business Savvy**
  - Discovering patterns that identify risks and opportunities
  - Measure

To Ph.D or NOT Ph.D?
That is the Question!

- **LinkedIn Poll:**

  Do You Need a PhD to Analyze Big Data?

YES	NO
301 (12%)	*2476* votes (87%)

Data Scientist Self-ID

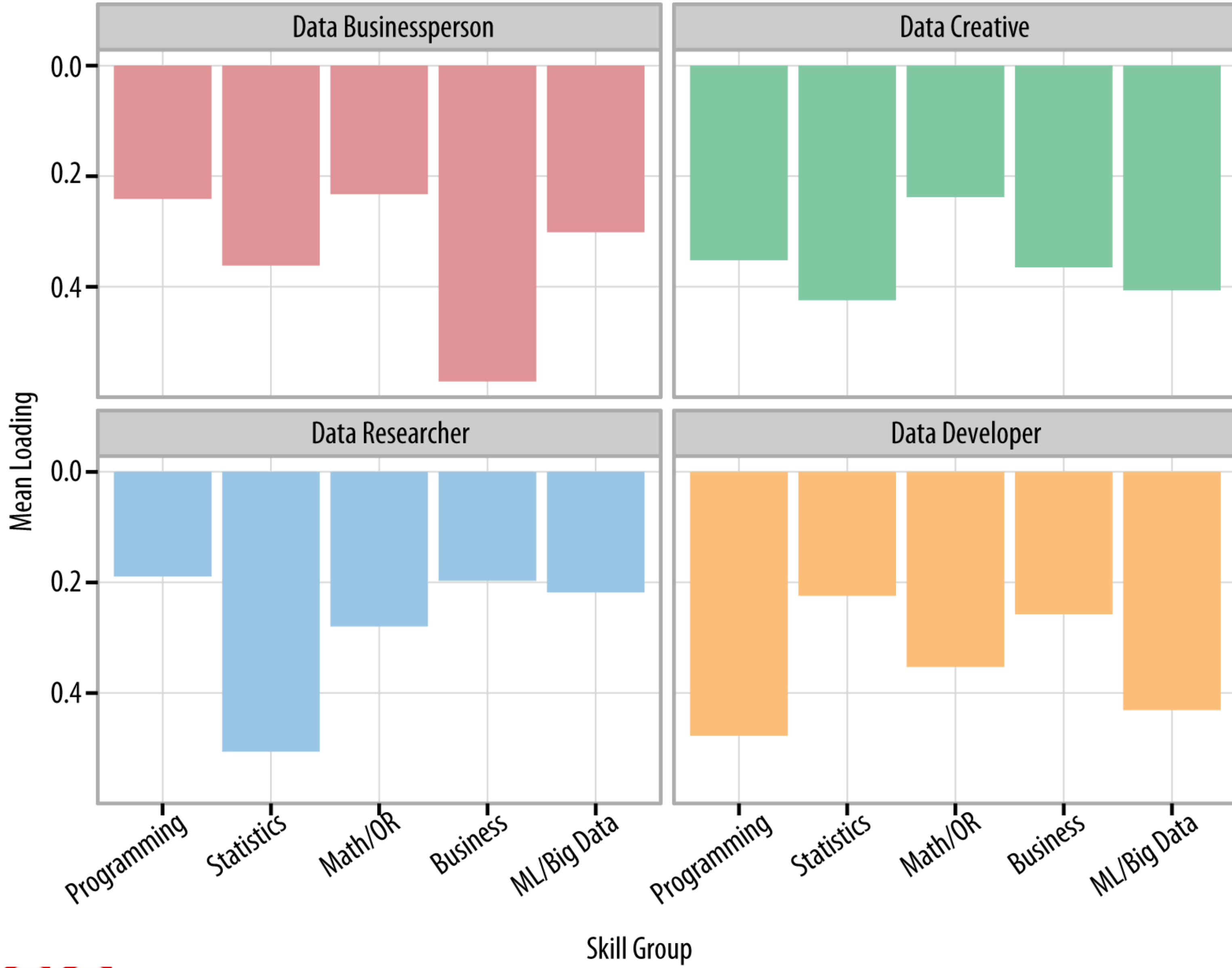
Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

O'Reilly Strata Survey suggested Self-ID Group, along with the self-ID categories most strongly associated with each Group

Strata Survey Skills

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

Strata survey, 2013



Strata survey, 2013

Learning and Training Opportunities

- **Many MS, MAS, Courses, Training, Workshops, Certificates, Boot camps, etc.**
- **Introduction to Data Science Example**
  - Part 1: Data Manipulation at scale
    - Databases and the relational algebra
    - Parallel databases, parallel query processing, in-database analytics, MapReduce, Hadoop, relationship to databases, algorithms, extensions, languages
    - Key-value stores and NoSQL; Entity resolution, record linkage
  - Part 2: Analytics, Predictive Analytics, Text mining
  - Part 3: Communicating Results
    - Visualization, data products, visual data analytics
    - Provenance, privacy, ethics, governance

How long does it take for a beginner to become a good data scientist per Region?

Region (Count)	Avg Years to become a good data scientist
AU/NZ (9)	6.9 years
E. Europe (19)	5.9 years
US/Canada (143)	4.9 years
W. Europe (60)	4.9 years
Asia (25)	4.9 years
Africa/Middle East (9)	4.4 years
Latin America (12)	3.9 years

INTRO TO MACHINE LEARNING
DATA MINING
PREDICTIVE ANALYTICS
DATA SCIENCE

Necessity is the Mother of Invention

Data explosion

Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories

- "*We are drowning in data, but starving for knowledge!*" (John Naisbitt, 1982)

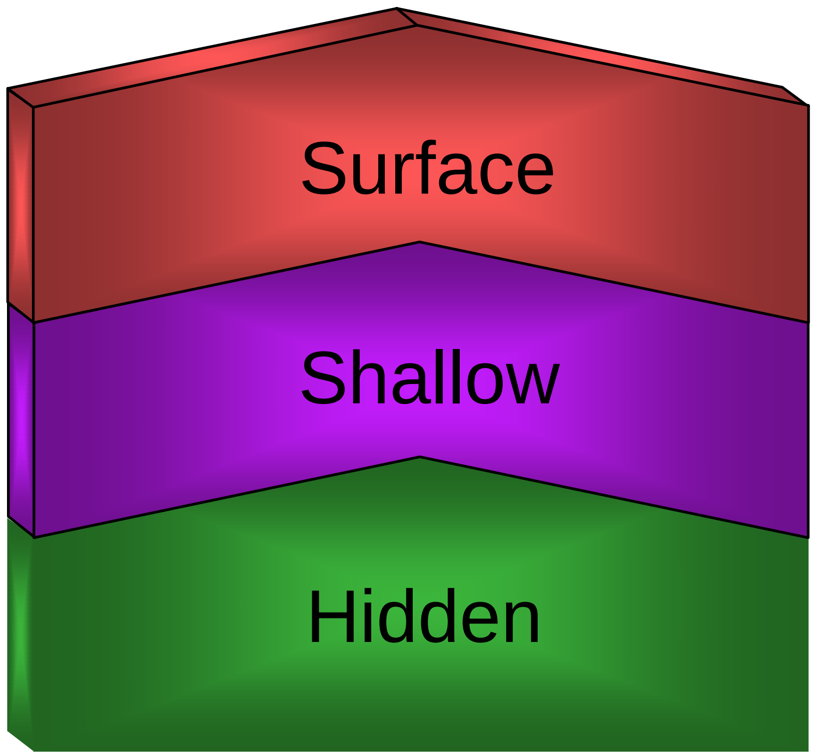
Necessity is the Mother of Invention

- **Solution**
  - **Predictive Analytics or Data Mining**
    - Extraction or “mining” of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
    - Data -driven discovery and modeling of hidden patterns in large volumes of data
    - Extraction of implicit, previously unknown and unexpected, potentially extremely useful information from data

Predictive Analytics

Top-Down
Methodology

Bottom-Up
Methodology



<u>Analytical Tools</u>

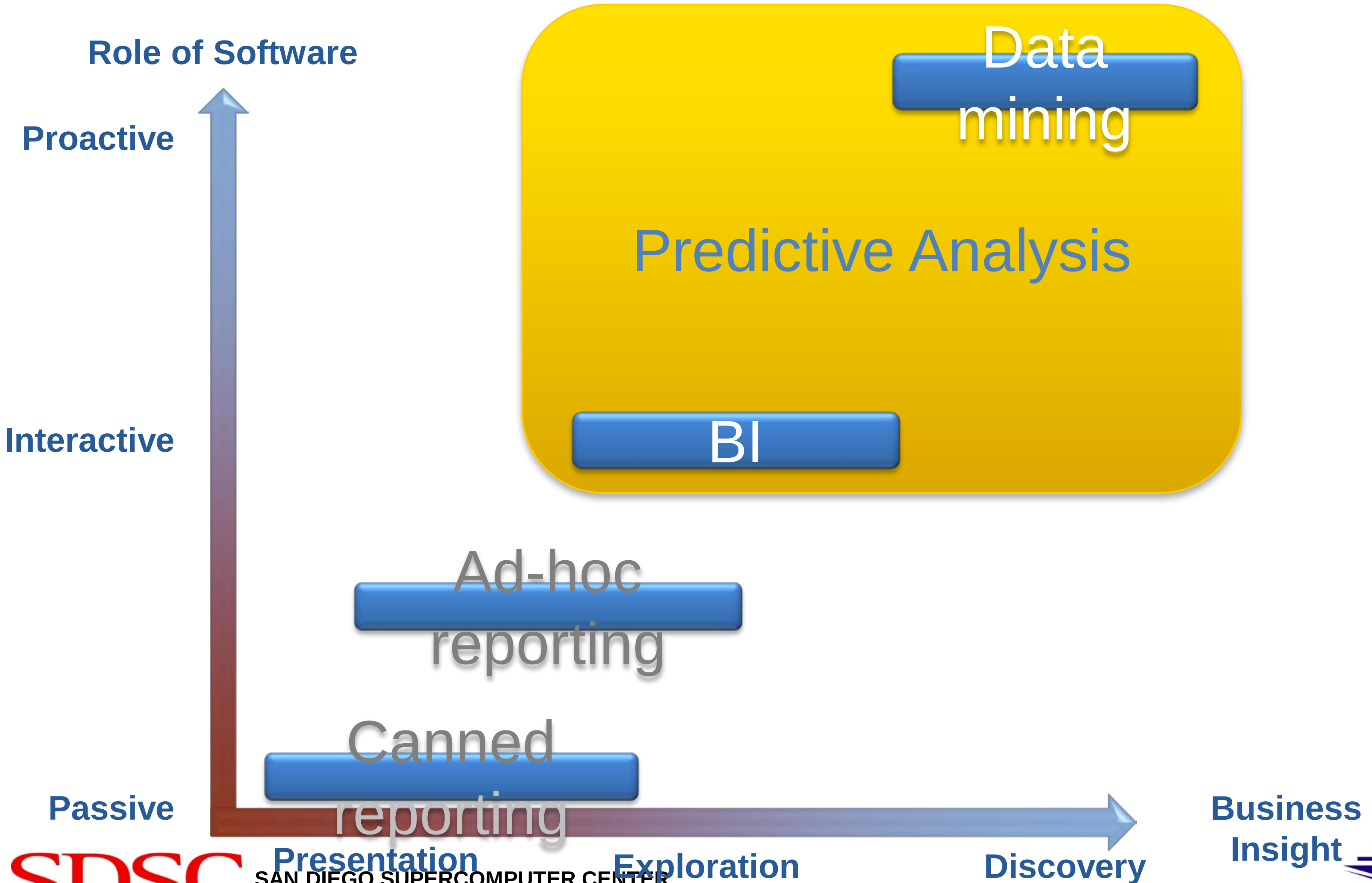
SQL tools for simple queries and reporting

Statistical & BI tools for summaries and analysis

Data Mining methods for knowledge discovery

Query Reporting	BI	Data Mining
Extraction of data; detailed and/or summarized	Analysis, summaries, Trends	Discovery of hidden patterns, information, predicting future trends
Information	Analysis	Insight knowledge and prediction
Who purchased the product in the last 2 quarters?	What is an average income of the buyers per quarter by district?	Which customers are likely to buy a similar product in the future and why?

DM Enables Predictive Analytics



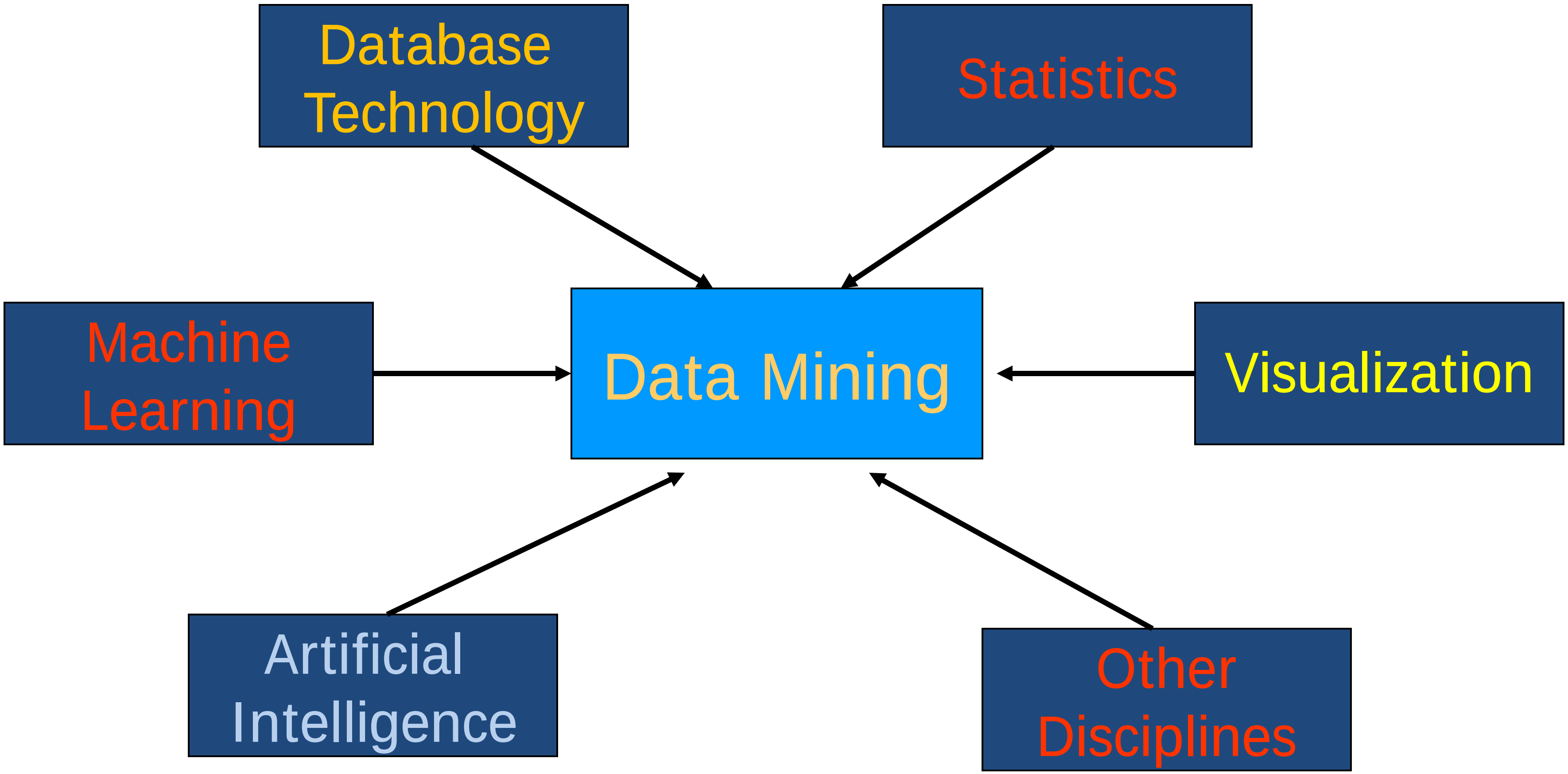
What Is Data Mining?

- **Combination of AI and statistical analysis to discover information that is “hidden” in the data**
  - associations (e.g. linking purchase of pizza with beer)
  - sequences (e.g. tying events together: marriage and purchase of furniture)
  - classifications (e.g. recognizing patterns such as the attributes of employees that are most likely to quit)
  - forecasting (e.g. predicting buying habits of customers based on past patterns)

Data Mining is NOT…

- **Data Warehousing**
- **(Deductive) query processing**
    - SQL/ Reporting
- **Software Agents**
- **Expert Systems**
- **Online Analytical Processing (OLAP)**
- **Statistical Analysis Tool**
- **Data visualization**
- **BI – Business Intelligence**

Multidisciplinary Field



Data Mining is…

- **Multidisciplinary Field**
  - Database technology
  - Artificial Intelligence
    - Machine Learning including Neural Networks
  - Statistics
  - Pattern recognition
  - Knowledge-based systems/acquisition
  - High-performance computing
  - Data visualization
  - Other Disciplines

History of Data Mining

History

- **Emerged late 1980s**
- **Flourished –1990s**
- **Roots traced back along three family lines**
  - Classical Statistics
  - Artificial Intelligence
  - Machine Learning

Statistics

- **Foundation of most DM technologies**
  - Regression analysis, standard distribution/deviation/variance, cluster analysis, confidence intervals
- **Building blocks**
- **Significant role in today's data mining – but alone is not powerful enough**

Artificial Intelligence

- **Heuristics vs. Statistics**
- **Human-thought-like processing**
- **Requires vast computer processing power**
- **Supercomputers**

Machine Learning

- **Union of statistics and AI**
  - Blends AI heuristics with advanced statistical analysis
- **Machine Learning – let computer programs**
  - learn about data they study - make different decisions based on the quality of studied data
  - using statistics for fundamental concepts and adding more advanced AI heuristics and algorithms

Terminology

- Gold Mining
- Knowledge mining from databases
- Knowledge extraction
- Data/pattern analysis
- Knowledge Discovery Databases or KDD
- Information harvesting
- Business intelligence
- Predictive Analytics
- Data Science

TAXONOMY

- **Predictive Methods**
  - *Use some variables to predict some unknown or future values of other variables*
- **Descriptive Methods**
  - *Find human –interpretable patterns that describe the data*

- **Supervised vs. Unsupervised**

What does Data Mining Do?

Explores Your Data

Finds Patterns

Performs Predictions

What can we do with Data Mining?

- **Exploratory Data Analysis**
- **Predictive Modeling: Classification and Regression**
- **Descriptive Modeling**
  - Cluster analysis/segmentation
- **Discovering Patterns and Rules**
  - Association/Dependency rules
  - Sequential patterns
  - Temporal sequences
- **Deviation detection**

Data Mining Applications

- **Science: Chemistry, Physics, Medicine, Energy**

  Biochemical analysis, remote sensors on a satellite, medical image analysis
- **Bioscience**

  Sequence-based analysis, protein structure and function prediction, protein family classification, microarray gene expression
- **Pharmaceutical, Insurance, Health care, Medicine**

  Drug development, medical therapies, claims analysis, fraudulent behavior, medical diagnostics
- **Financial Industry, Banks, Businesses, E-commerce**

  Stock and investment analysis, identify loyal customers vs. risky customer, predict customer spending, risk management, sales forecasting
- **Market analysis and management**

  Target marketing, CRM, market basket analysis, cross selling, market segmentation
- **Risk analysis and management**

  Forecasting, customer retention, improved underwriting, quality control, competitive analysis
- **Sports and Entertainment**

  IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

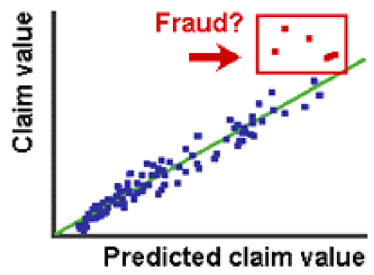
Data Mining Tasks

- Concept/Class description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions; “normal” vs. fraudulent behavior
- Association (correlation and causality)
  - Multi-dimensional interactions and associations

age(X, “20-29”) ^ income(X, “60-90K”) à buys(X, “TV”)

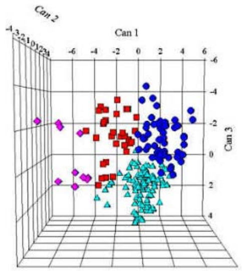
Hospital(area code) ^ procedure(X) ->claim (type) ^ claim(cost)

Data Mining Tasks



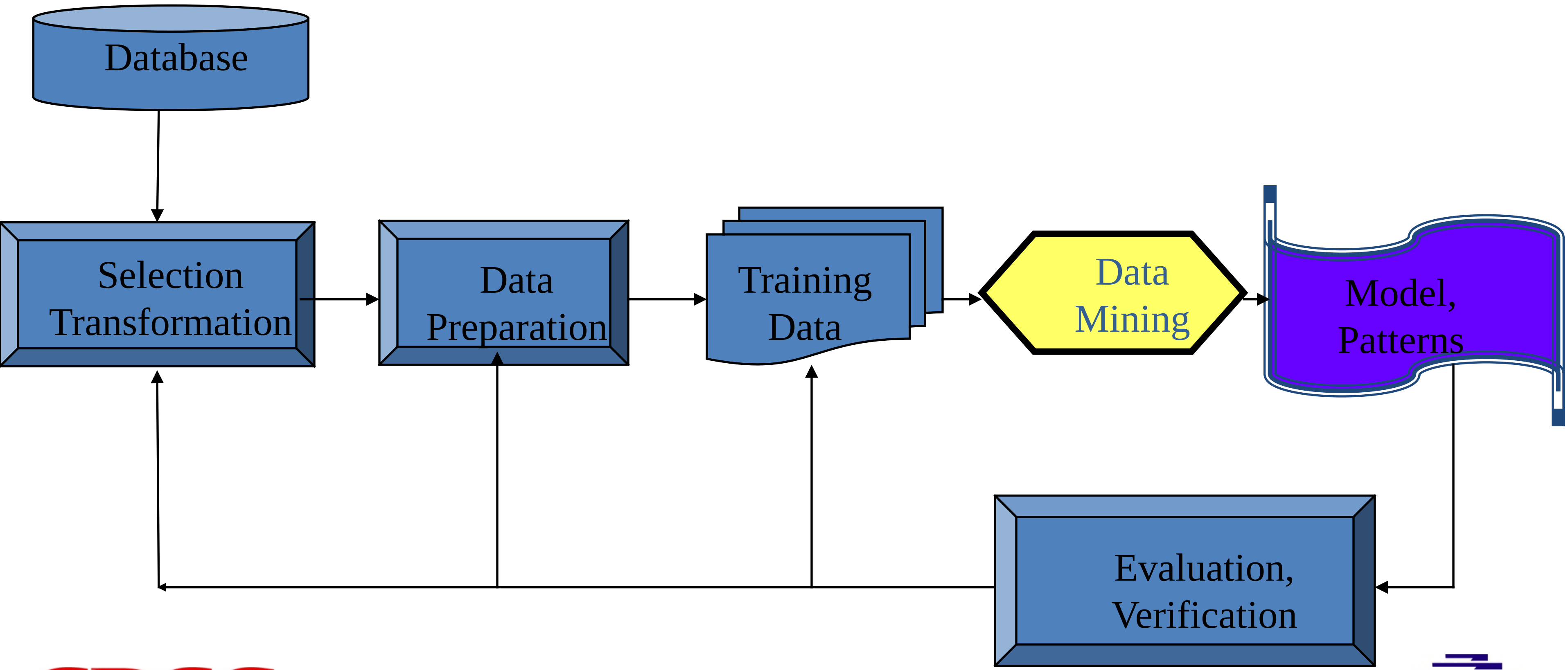
- Classification and Prediction
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - Example: classify countries based on climate, or classify cars based on gas mileage, fraud based on claims information, energy usage based on sensor data
  - Presentation:
    - If-THEN rules, decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values

Data Mining Tasks



- **Cluster analysis**
  - Class label is unknown: Group data to form new classes
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity
- **Outlier analysis**
  - Data object that does not comply with the general behavior of the data
  - Mostly 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

KDD Process


Database
Selection
Transformation
Data
Preparation
Training
Data
Data
Mining
Model,
Patterns
Evaluation,
Verification

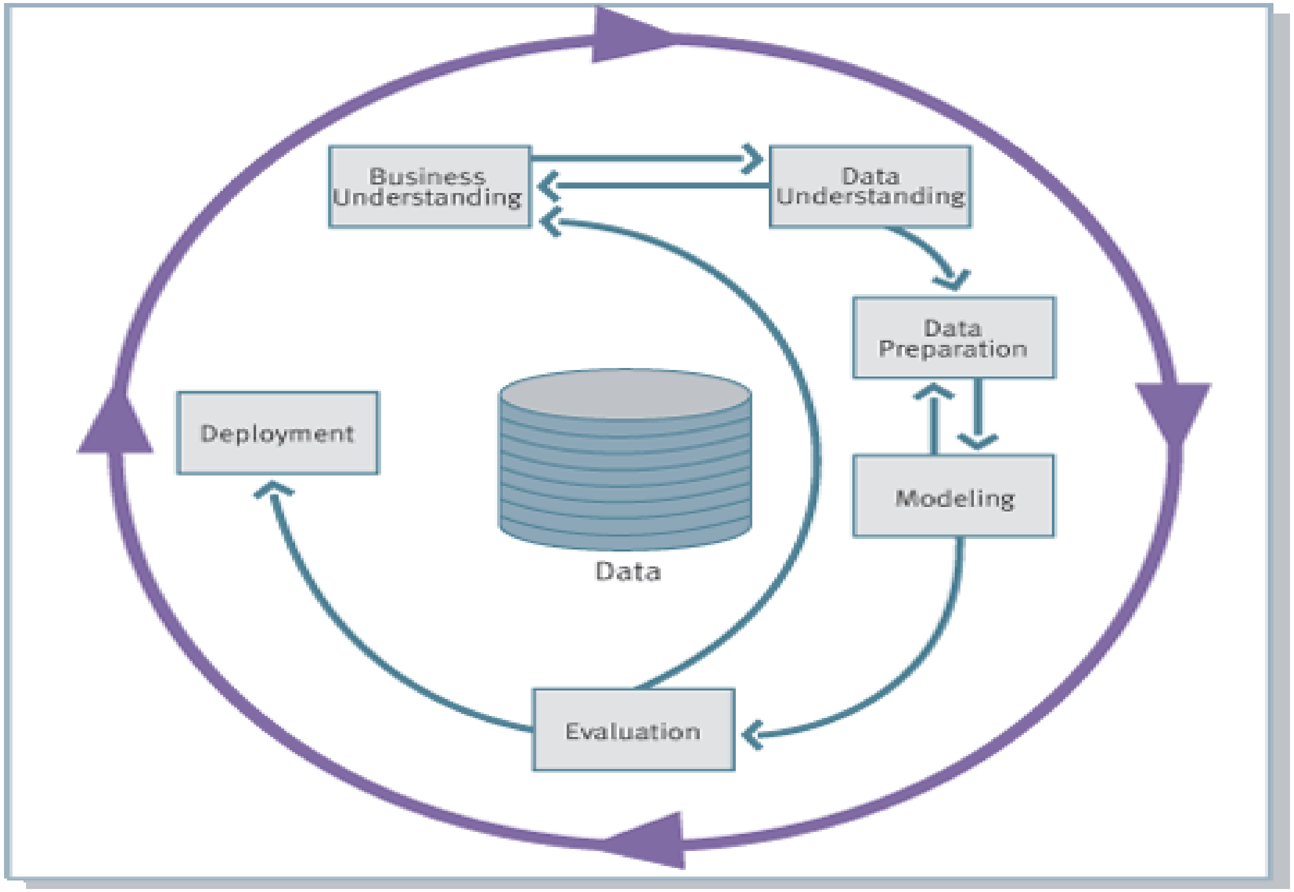
KDD Process Steps

- 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, representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering

KDD Process Steps (2)

- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use and integration of discovered knowledge

CRISP-DM - Cross Industry Standard Process for Data Mining



CRISP-DM Process Model

Learning and Modeling Methods

- **Decision Tree Induction (C4.5, J48)**
- **Regression Tree Induction (CART, MP5)**
- **Multivariate Regression Tree (MARS)**
- **Clustering (K-means, EM, Cobweb)**

- **Artificial Neural Networks (Backpropagation, Recurrent)**
- **Support Vector Machines (SVM)**
- **Various other models**

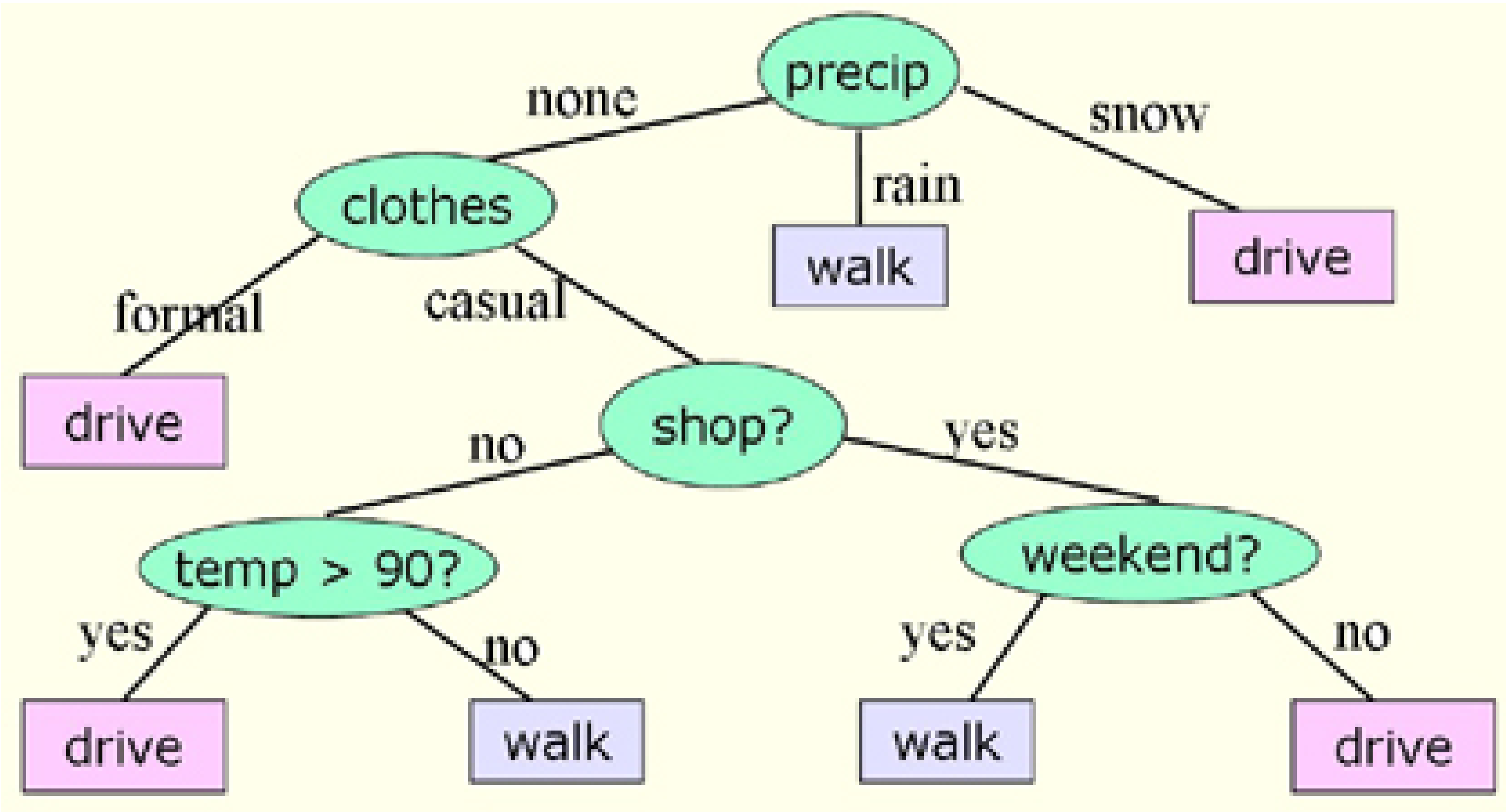
Decision Tree Induction

- **Method for approximating discrete-valued functions**
  - robust to noisy/missing data
  - can learn non-linear relationships
  - inductive bias towards shorter trees

Decision Tree Induction

- **Applications:**
  - medical diagnosis – ex. heart disease
  - analysis of complex chemical compounds
  - classifying equipment malfunction
  - risk of loan applicants
  - Boston housing project – price prediction
  - fraud detection

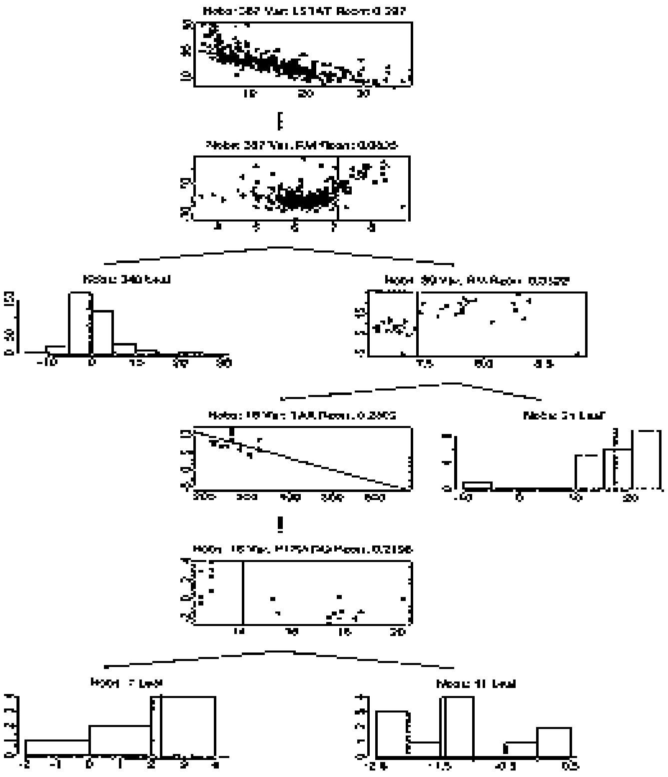
Decision Tree Example



Regression Tree Induction

- **Why Regression tree?**
  - Ability to:
    - Predict continuous variable
    - Model conditional effects
    - Model uncertainty

Regression Trees



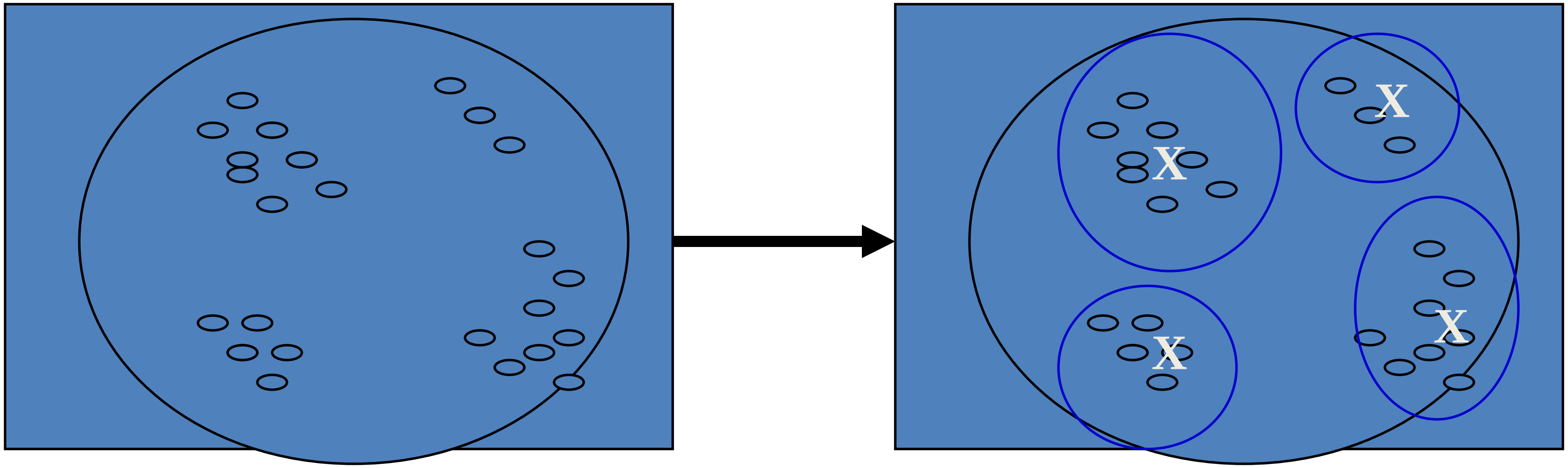
- **Continuous goal variables**
- **Induction by means of an efficient recursive partitioning algorithm**
- **Uses linear regression to select internal nodes**

Quinlan, 1992

Clustering

- Basic idea: Group similar things together
- Unsupervised Learning – Useful when no other info is available
- K-means
  - Partitioning instances into *k* disjoint clusters
  - Measure of similarity

Clustering

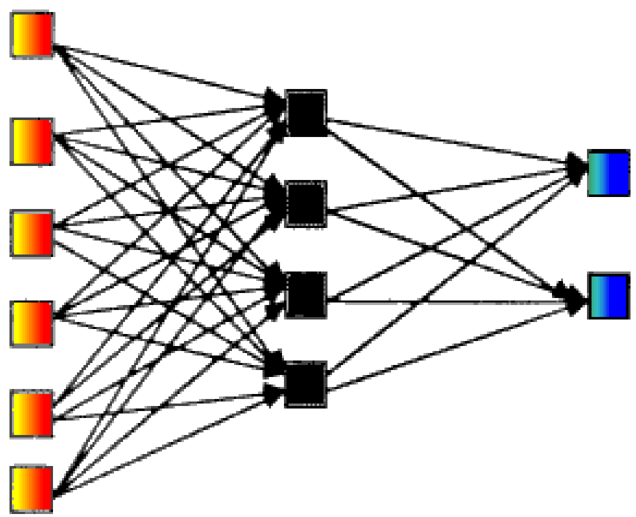


Kmeans Results from 10 million NYTimes articles

cluster means shown with coordinates determining fontsize

7 viable clusters found

Artificial Neural Networks (ANNs)



- Network of many simple units
- Main Components
  - Inputs
  - Hidden layers
  - Outputs
- Adjusting weights of connections
- Backpropagation

Evaluation

- **Error on the training data vs. performance on future/unseen data**
- **Simple solution**
  - Split data into training and test set
  - Re-substitution error
    - error rate obtained from the training data
- **Three sets**
  - training data, validation data, and test data

Training and Testing

- **Test set**
  - set of independent instances that have not been used in formation of classifier in any way
  - Assumption
    - data contains representative samples of the underlying problem
- **Example: classifiers built using customer data from two different towns A and B**
  - To estimate performance of classifier from town in completely new town, test it on data from B

Error Estimation Methods

- **Holdout**
  - ½ training and ½ testing (2/3&1/3)
- **Repeated Holdout Method**
  - Random sampling – repeated holdout
- **Cross-validation**
  - Partition in K disjoint clusters
  - Train k-1, test on remaining
- **Leave-one-out Method**
- **Bootstrap**
  - Sampling with replacement

Data Mining Challenges

- **Computationally expensive to investigate all possibilities**
- **Dealing with noise/missing information and errors in data**
- **Mining methodology and user interaction**
  - Mining different kinds of knowledge in databases
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
  - Expression and visualization of data mining results

Data Mining Heuristics and Guide

- **Choosing appropriate attributes/input representation**
- **Finding the minimal attribute space**
- **Finding adequate evaluation function(s)**
- **Extracting meaningful information**
- **Not overfitting**

Available Data Mining Tools

COTs:

- IBM Intelligent Miner
- SAS Enterprise Miner
- Oracle ODM
- Microstrategy
- Microsoft DBMiner
- Pentaho
- Matlab
- Teradata

Open Source:

- Python
- R
- WEKA
- KNIME
- Orange
- RapidMiner
- Rattle
- Mahout
- MlLib

Data mining applications at SDSC

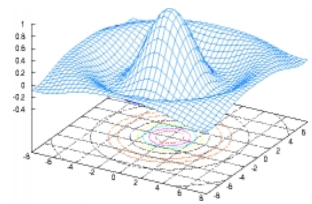
DM Suites

WEKA
The University of Waikato

RAPID|MINER

MathWorks

python



Octave

Computational Packages with DM tools

Others as Requested

FASTlib
A library of Fundamental Algorithmic and Statistical Tools

Libraries for building tools

The Phoenix System for MapReduce Programming

Summary

- **Discovering interesting patterns from large amounts of data**
- **CRISP-DM Industry standard**
- **Learn from the past**
  - High quality, evidence based decisions
- **Predict for the future**
  - Prevent future instances of fraud, waste & abuse
- **React to changing circumstances**
  - Current models, continuous learning

Thank you!


KEEP
CALM
AND
ANALYZE
BIG DATA

Questions?

- www.sdsc.edu
- For further information, contact Natasha Balac (nbalac@ucsd.edu)