Lesson A-1

Introduction to Data Mining

Topics

- Introduction to Data Mining/Big Data
- Data mining project development process

POPULAR SCIENCE

THE FUTURE NOW

THE CONTROL CENTERS

Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate P.52

OFFICER ALGORITHM

Can a Crime Be Prevented Before It Begins? P.38

NEW WAYS OF SEEING

A Gallery of Extraordinary Infographics P.69

PLUS

Juan Enriquez Reprograms Life P.31

James Gleick Unsplits the Bit P.58

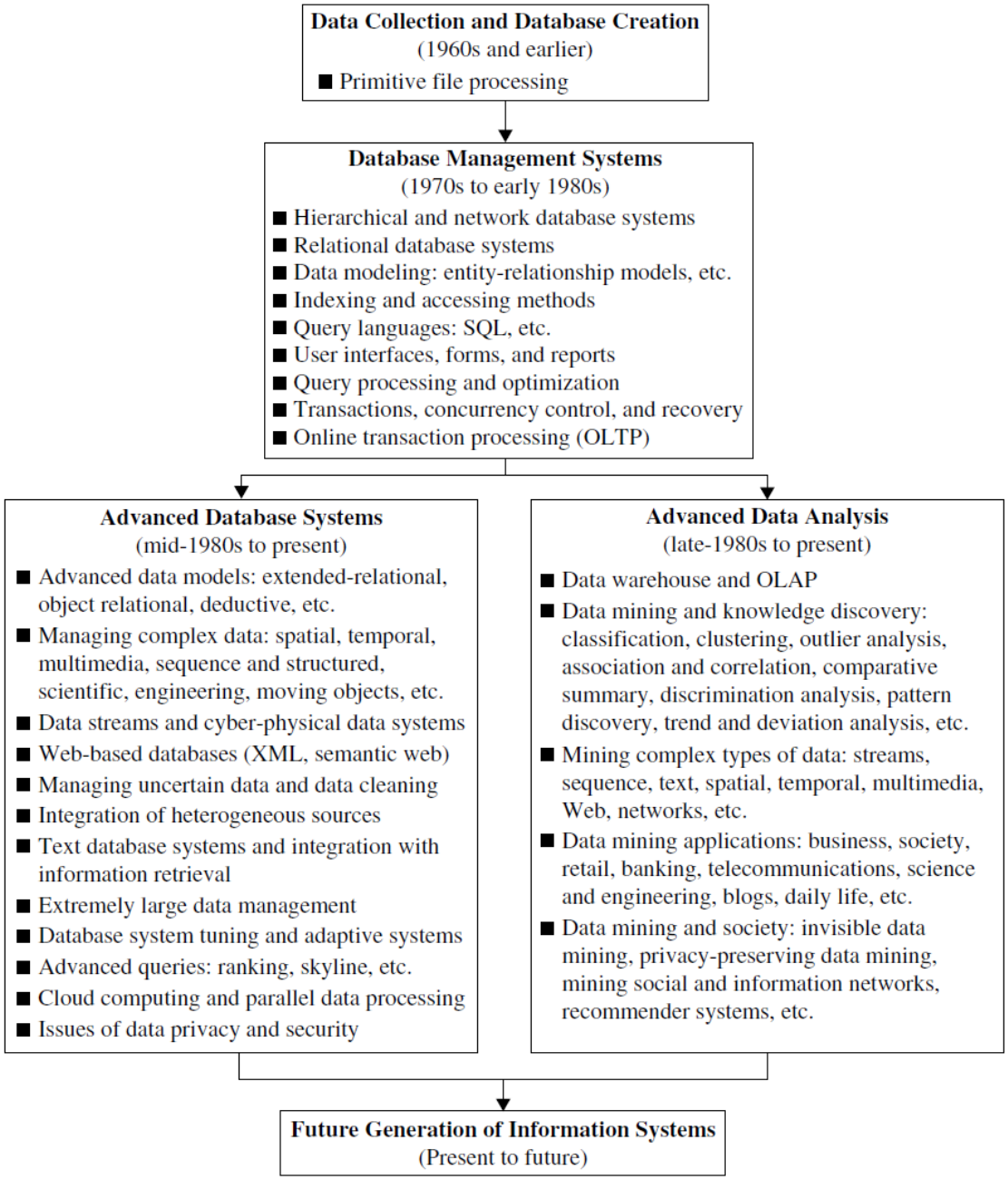
AND **Lawrence Weschler** Questions the Cloud P.76

SPECIAL ISSUE

DATA IS POWER

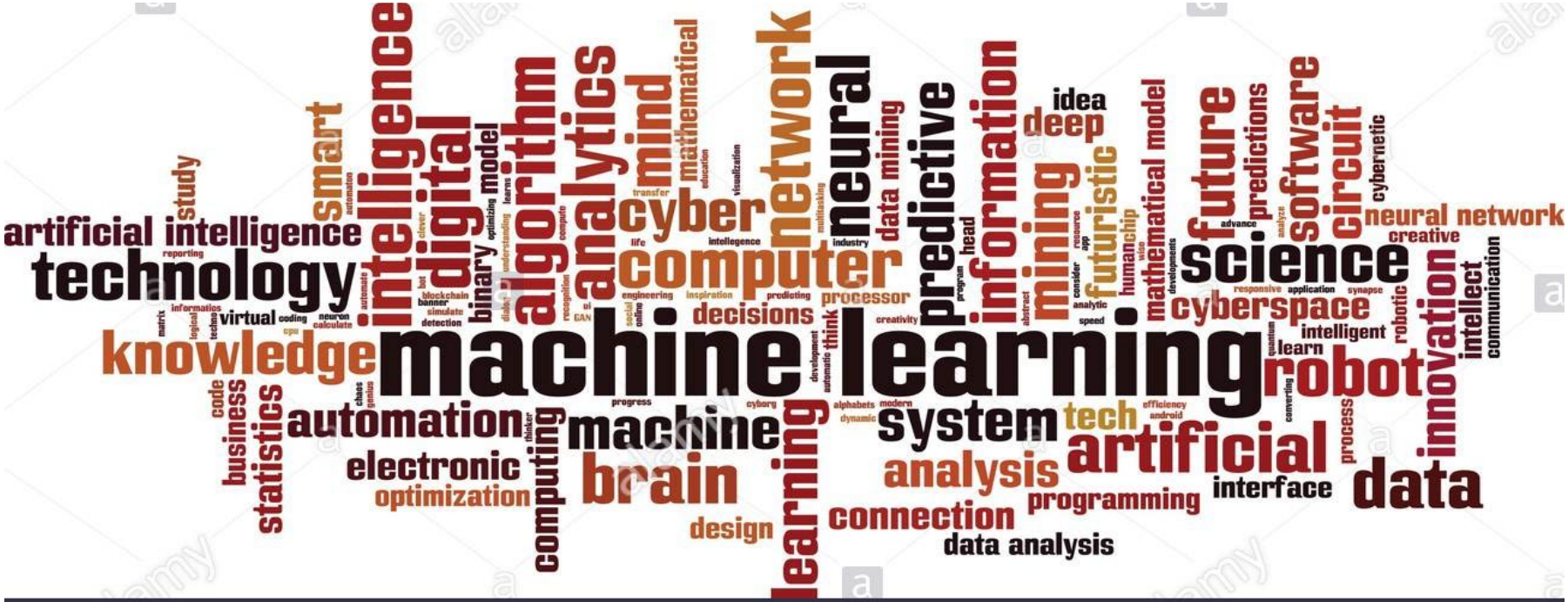
HOW INFORMATION IS DRIVING THE FUTURE

NOVEMBER 2011 US $4.99

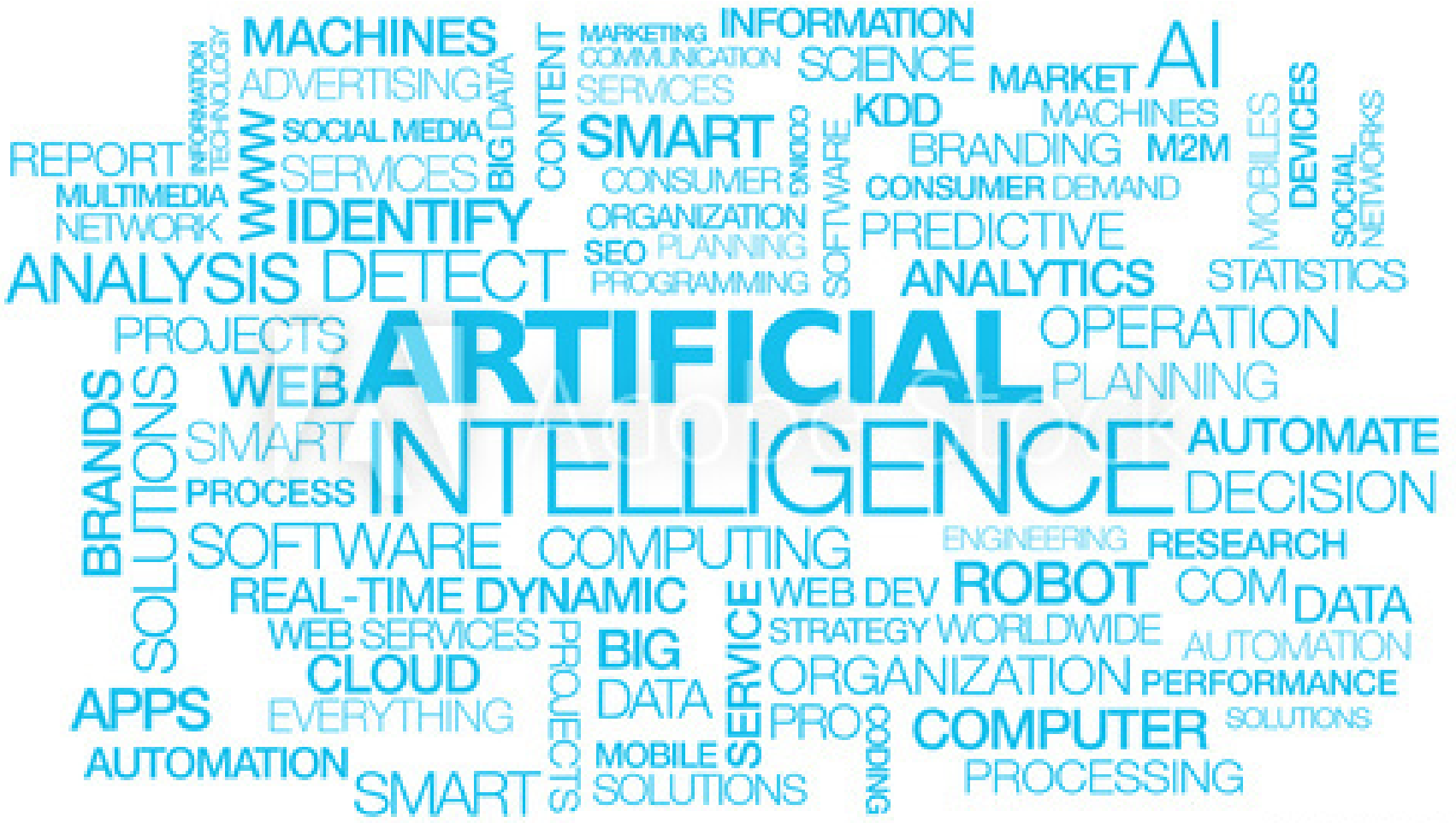


The evolution of database system technology.

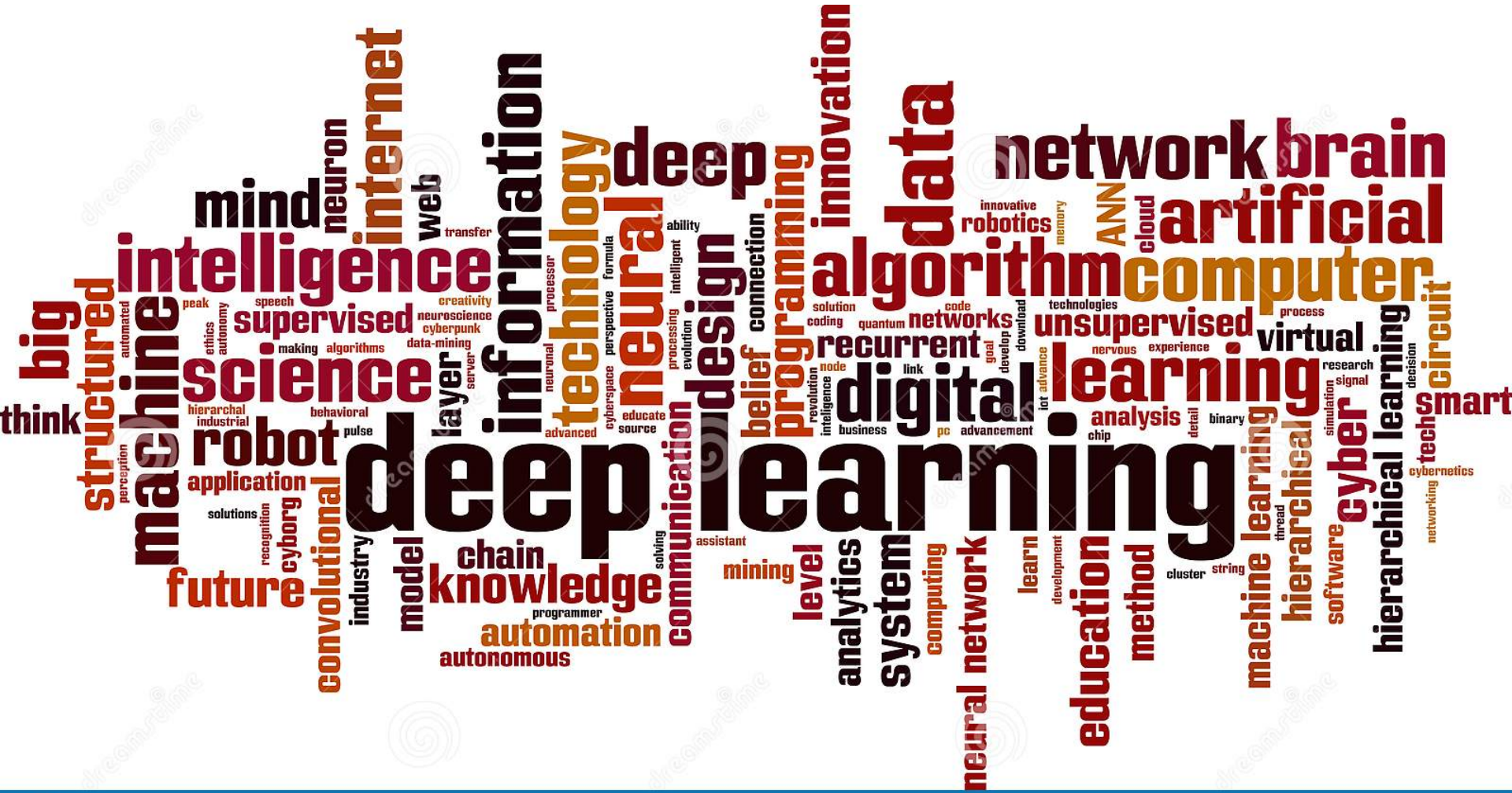

DATA
MINING
PATTERNS
LEARNING
ANALYSIS
SET
LIKELY
PRESENT
GROUPS
MARKETING
GENERAL
SENDING
BUSINESS
KNOWN
MANAGEMENT
METHODS
STRUCTURES
LEARNED
PROCESS
NEW
DATABASE
DECISION
PREDICT
MIGHT
PRODUCT
BASKET
APPLICATIONS
ALGORITHMS
VISUALIZATION
TEST
TERM
POTENTIAL
MUST
USING
WITHOUT
E-MAILS
WITHIN
MARKET
FINAL
DESIRED
OFFERS
DISCOVERY
ASSOCIATION
OFFER
SEND
USE
TASK
MACHINE
TARGET
CUSTOMERS
INVOLVES
EXAMPLE
STEP
ANOTHER
DETERMINE
PEOPLE
RULE
LARGER
DISCOVERED
INTELLIGENCE
CUSTOMER
SUPPORT
CITATION
KNOWLEDGE
MODELING
TRAINED
INFORMATION
ARTIFICIAL
NEEDED
NUMBER
PRE-PROCESSING
IDENTIFY
OFTEN
HELP
SYSTEM
SOPHISTICATED
RECORDS
DETECTION
ALGORITHM
MANY
E-MAIL
MODEL
MAKING
MODELS
COMMON
FINALLY
LARGE




DETECTION
BIG DATA
COMPUTING
INFORMATION TECHNOLOGY
WWW
SOCIAL MEDIA
PROMOTION
E-MARKETING
COMMUNICATION
MODELS
SERVICES
BIG
CONTENT
PROCESSING
BRANDING
COMPUTER
MULTIMEDIA
NETWORK
PROJECTS
CONSUMER
CONSUMER DEMAND MARKETS
SOFTWARE
ORGANIZATION
CODING
WEB MARKETING
PREDICTIVE
PLANNING
PROGRAM
ANALYTICS
EVENTS
PROGRAMMING
DATA MINING
DATA SCIENCE
INFORMATION
MACHINE LEARNING
VISION
ENGINEERING
RESEARCH
PROBABILITY
KDD
WEB DEV
COMPUTING
WEB SERVICES
BIG
SERVICE
STRATEGY
WORLDWIDE
MATHS PATTERN
DATA
VISUALIZATION
BIG DATA
ENGINEERING
PROJECTS
SOLUTIONS
PRO
PRICING
PLANNING
MEDIA
MOBILE
SEGMENTATION
SOCIAL NETWORK
INFORMATION
CODING
SOCIAL NETWORKS
STATISTICS
DIGITAL
TARGET


MACHINES
ADVERTISING
INFORMATION TECHNOLOGY
WWW
SOCIAL MEDIA
SERVICES
BIG DATA
CONTENT
MARKETING
COMMUNICATION
SERVICES
INFORMATION
SCIENCE
SMART
CONSUMER
ORGANIZATION
SEO
PLANNING
PROGRAMMING
SOFTWARE
KDD
MARKET
AI
MACHINES
BRANDING
M2M
CONSUMER DEMAND
PREDICTIVE
ANALYTICS
MOBILES
DEVICES
SOCIAL NETWORKS
STATISTICS
REPORT
MULTIMEDIA
NETWORK
IDENTIFY
ANALYSIS
DETECT
PROJECTS
ARTIFICIAL
INTELLIGENCE
OPERATION
PLANNING
AUTOMATE
DECISION
BRANDS
SOLUTIONS
WEB
SMART
PROCESS
SOFTWARE
COMPUTING
ENGINEERING
RESEARCH
ROBOT
COM
DATA
REAL-TIME
DYNAMIC
WEB SERVICES
CLOUD
APPS
EVERYTHING
AUTOMATION
PROJECTS
BIG
DATA
MOBILE
SOLUTIONS
SERVICE
WEB DEV
STRATEGY
WORLDWIDE
ORGANIZATION
PROG
CODING
COMPUTER
PROCESSING
AUTOMATION
PERFORMANCE
SOLUTIONS
SMART

#87886806

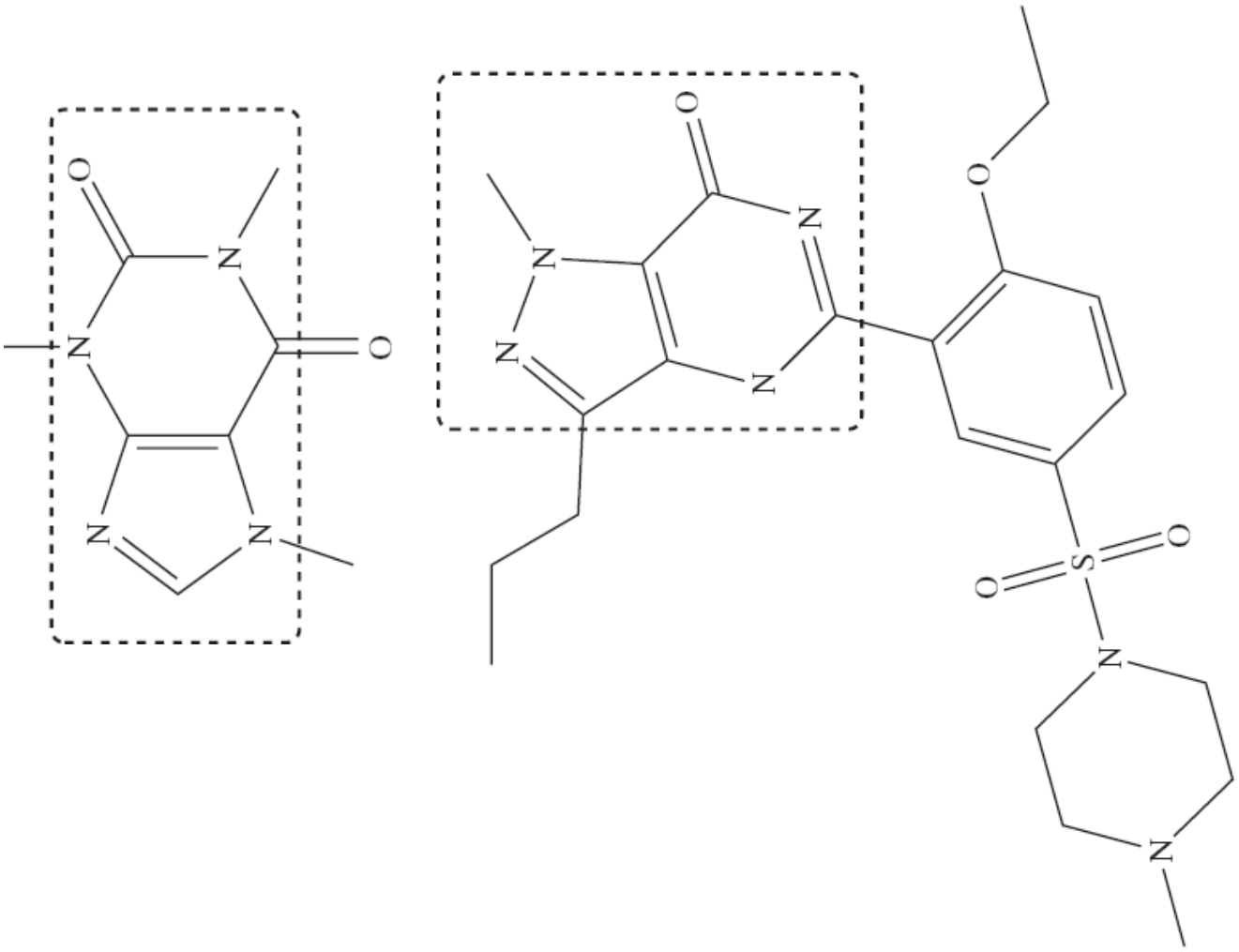



BIG
DATA
ALSO
TOOLS
DISK
TARGET
APPLIED
SENSOR
MAY
EXAMPLES
SHARED
DEFINITION
CURRENT
MOVING
WIRELESS
MANAGE
RECONSIDER
WITHIN
ARCHIVES
BIOGEOCHEMICAL
ZETTABYTES
PRACTITIONERS
CAPTURE
THOUGHT
TECHNOLOGIES
USING
INTERNET
BUSINESS
DISTRIBUTED
SETS
GENOMICS
DESCRIBING
RADIO-FREQUENCY
COMPLEXITY
NEEDED
REQUIRING
TIME
ONE
LARGE
MANAGEMENT
TERABYTES
CASE
ABILITY
SOFTWARE
PROCESS
PETABYTES
INCLUDE
ELAPSED
FG
FORMS
PERFORMANCE
STORE
SYSTEMS
TOLERABLE
SIZE
RESEARCH
INFORMATION
PRESENTATIONS
COMBAT
STORAGE
MPP
LOGS
COST
SIGNIFICANT
DEFINING
RELATED
COMPUTING
PARALLEL
SAN
QUALITIES
SINCE
MASSIVELY
GROW
SEARCH
DIFFICULTY
ANALYTICS
COMPLEX
TENS
LARGER
CONTINUES
SOCIAL
CITATION
INDEXING
SET
USE
RELATIONAL
CONNECTOMICS
ORGANIZATIONS
USED
CREATED
NOW
BIOLOGICAL
PROCESSING
UBIQUITOUS
CAPACITY
BURIED
WORLD'S
HUNDREDS
DESKTOP
RECORDS
NETWORKS
DATABASES
SOLID
TYPES
CURRENTLY
GARTNER
WORKING
AMOUNT
OPPORTUNITIES

What is Data Mining?

- *Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems*.
- *Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.*

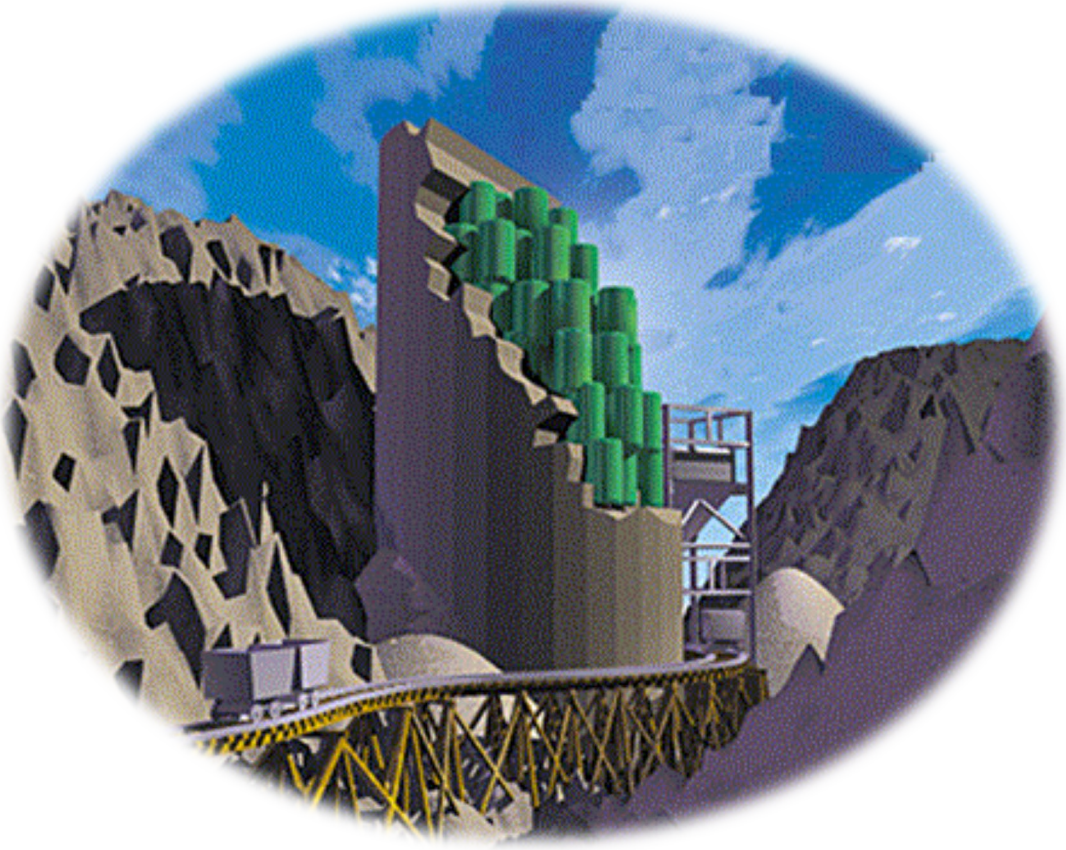
Market Basket Analysis



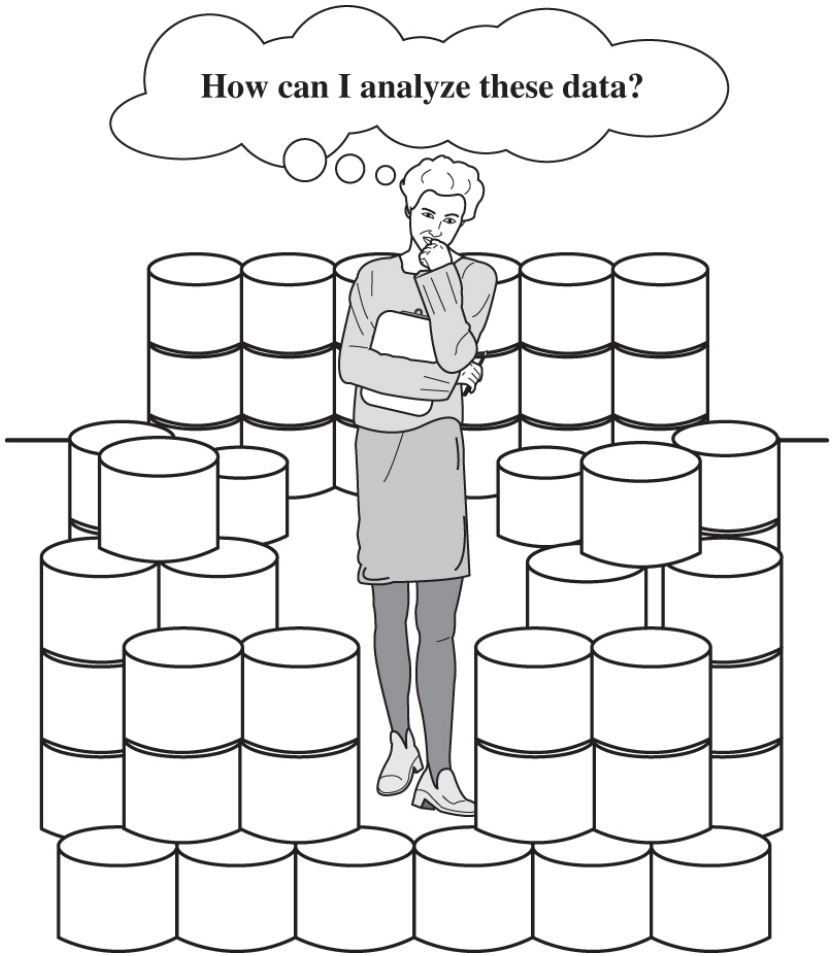
A Chemical database.

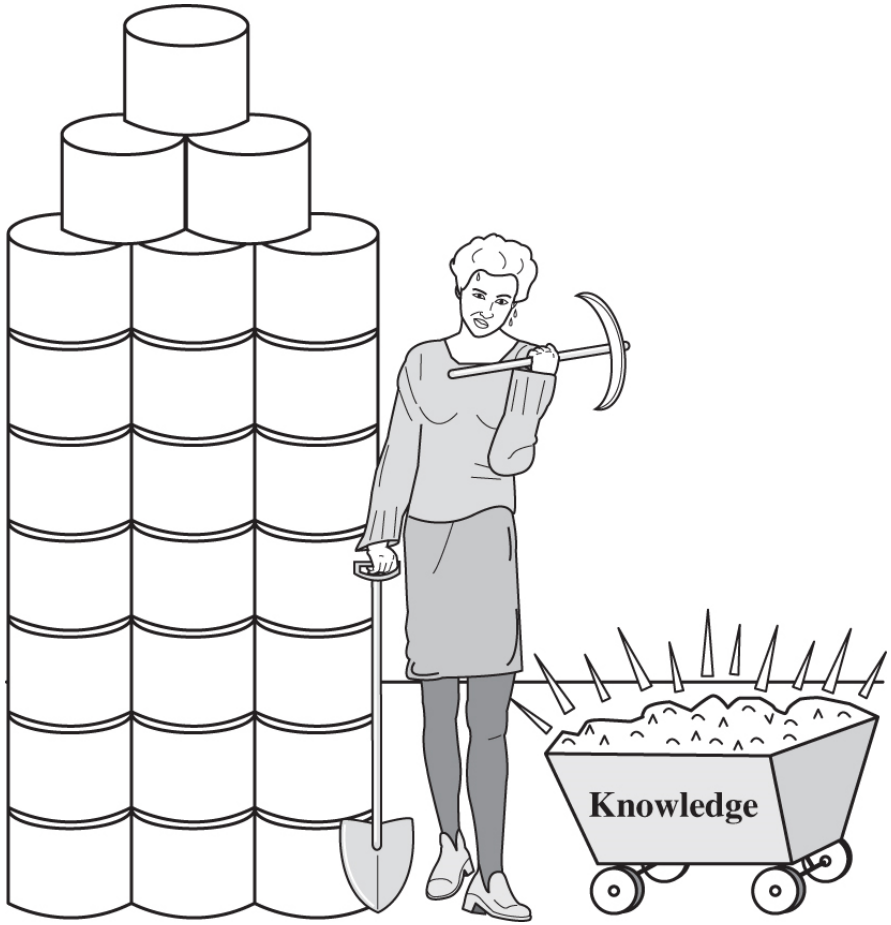
Chemistry Informatics

What is Data Mining?



Source: Cover page of *Advanced in Knowledge Discovery and Data Mining ,* edited by U. Fayyad, G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, MIT Press

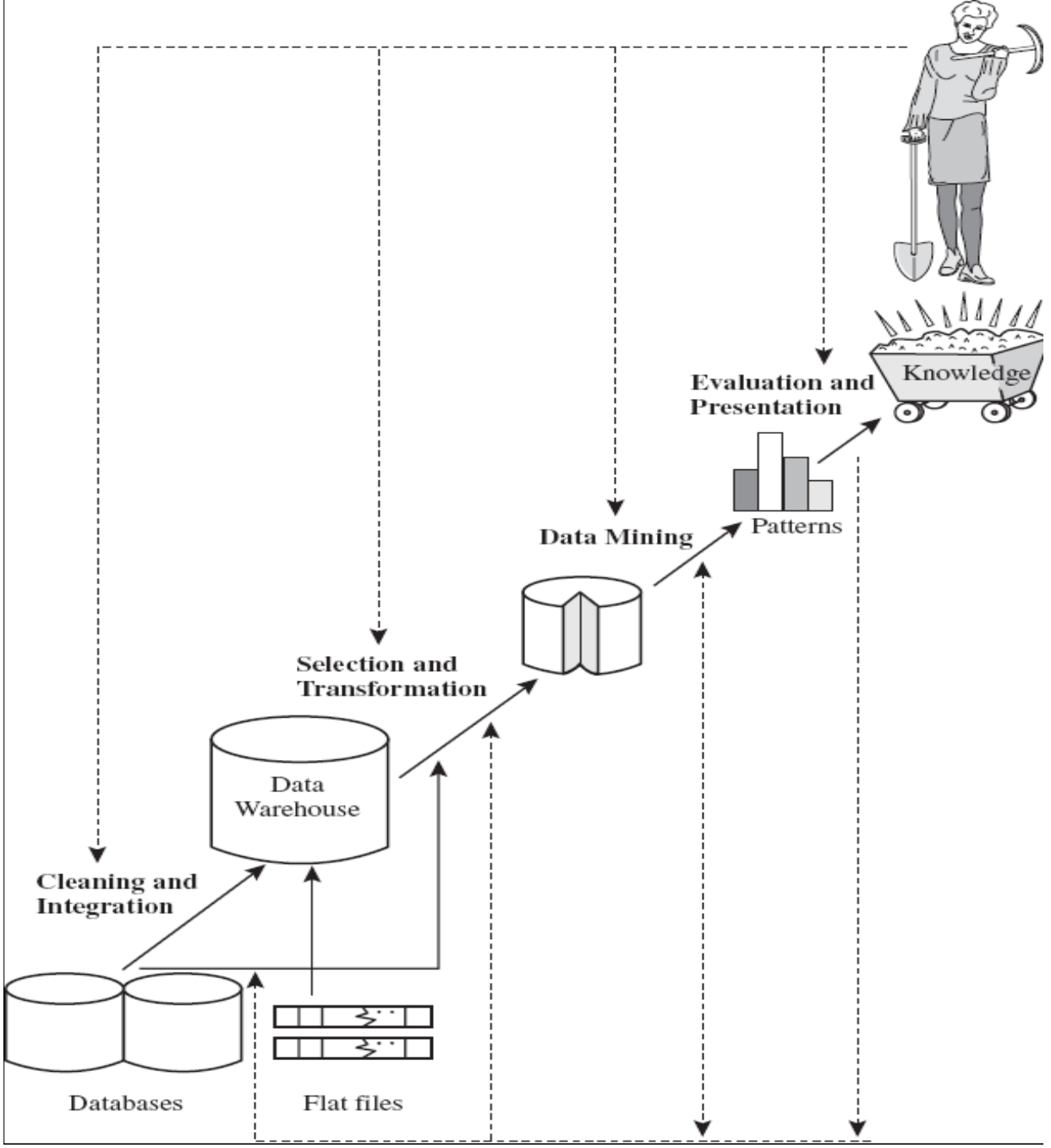

How can I analyze these data?


Knowledge

What Is Data Mining?

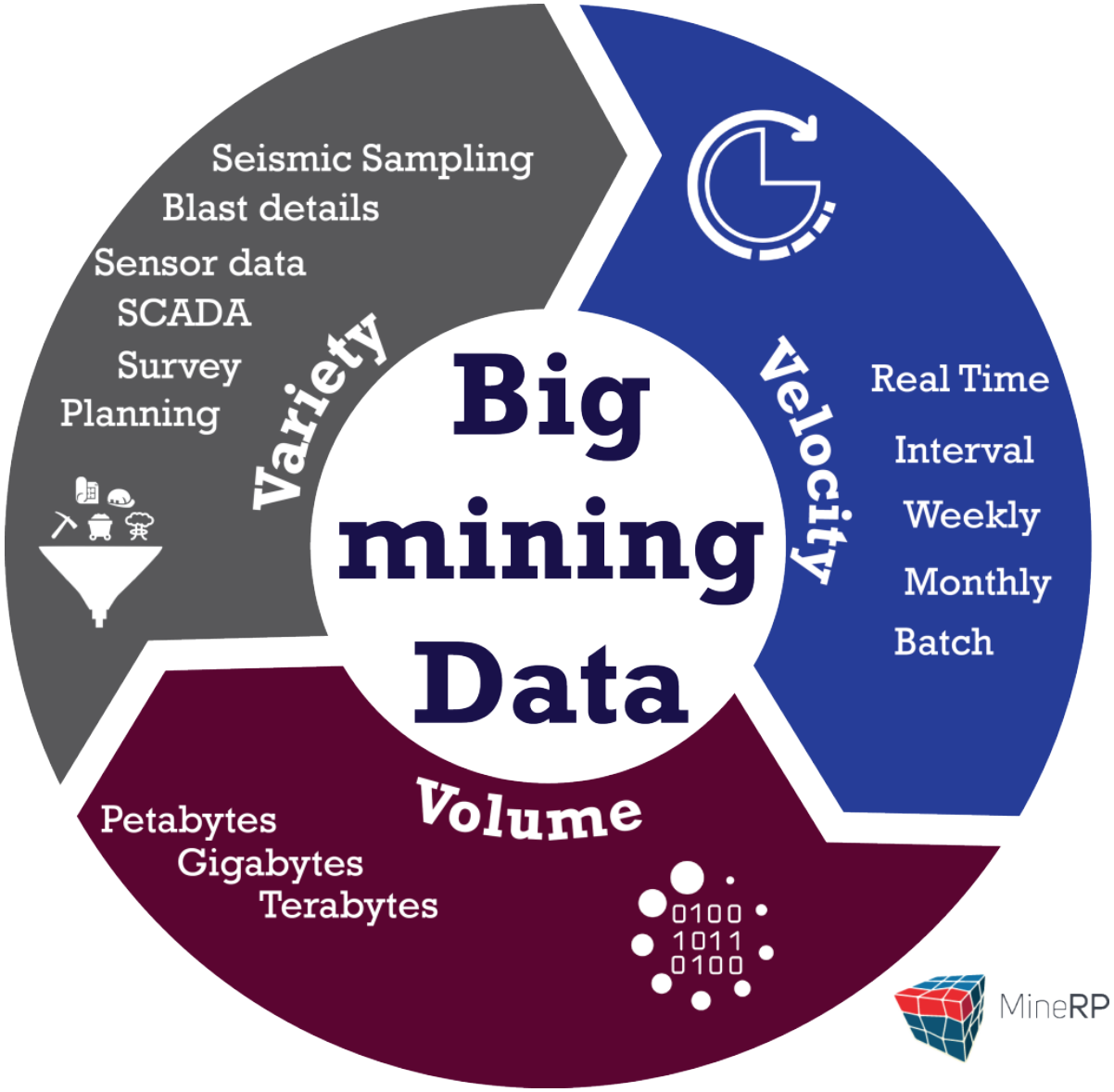
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

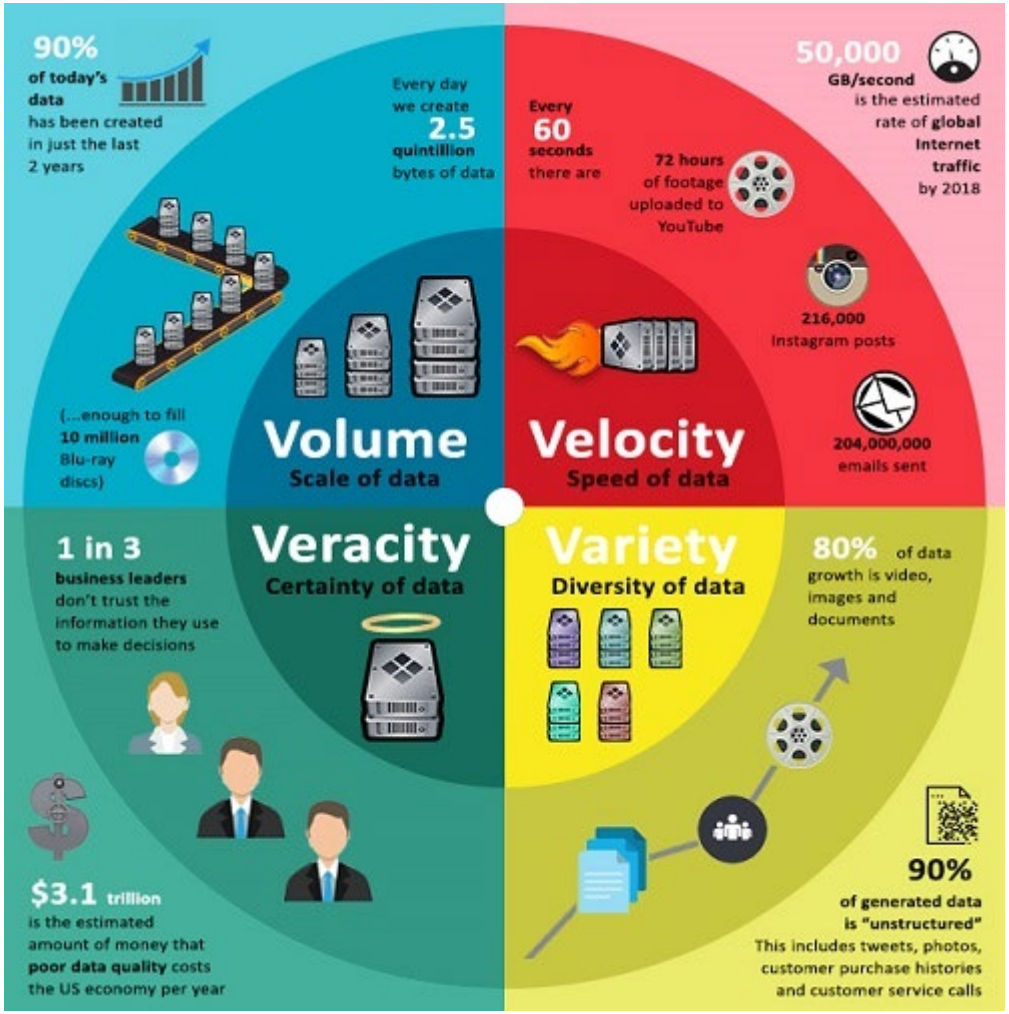
Data Mining is a Process of knowledge discovery



Data mining as a step in the process of knowledge discovery


Google
https://youtu.be/zDAYZU4A3w0


Big mining Data
Variety
Seismic Sampling
Blast details
Sensor data
SCADA
Survey
Planning
Velocity
Real Time
Interval
Weekly
Monthly
Batch
Volume
Petabytes
Gigabytes
Terabytes
0100
1011
0100
MineRP


90%
of today's data
has been created
in just the last
2 years
Every day
we create
2.5
quintillion
bytes of data
(...enough to fill
10 million
Blu-ray
discs)
Volume
Scale of data
Every
60
seconds
there are
72 hours
of footage
uploaded to
YouTube
50,000
GB/second
is the estimated
rate of global
Internet
traffic
by 2018
216,000
Instagram posts
204,000,000
emails sent
Velocity
Speed of data
1 in 3
business leaders
don't trust the
information they use
to make decisions
Veracity
Certainty of data
$3.1 trillion
is the estimated
amount of money that
poor data quality costs
the US economy per year
Variety
Diversity of data
80% of data
growth is video,
images and
documents
90%
of generated data
is "unstructured"
This includes tweets, photos,
customer purchase histories
and customer service calls

The Five V's of Big Data

Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.

The New York Stock Exchange capture **1 TB of Trade Information**

Analysis of Streaming Data

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.

By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**

4 Billion+ hours of video are watched on You Tube each month

30 Billion pieces of content are shared on facebook every month

400 Million tweets are sent per day by about 200 million monthly active users

Diffrent forms of data

As more and more data is being digitized.

1 in 3 Business leaders don't trust the information they use to make decisions

Uncertainty Of Data

This refers to the discrepancies found in the data.

Poor data quality costa the US economy around **$ 3.1 Trillion a year**

Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.

Exhibit 1

Big data can generate significant financial value across sectors

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

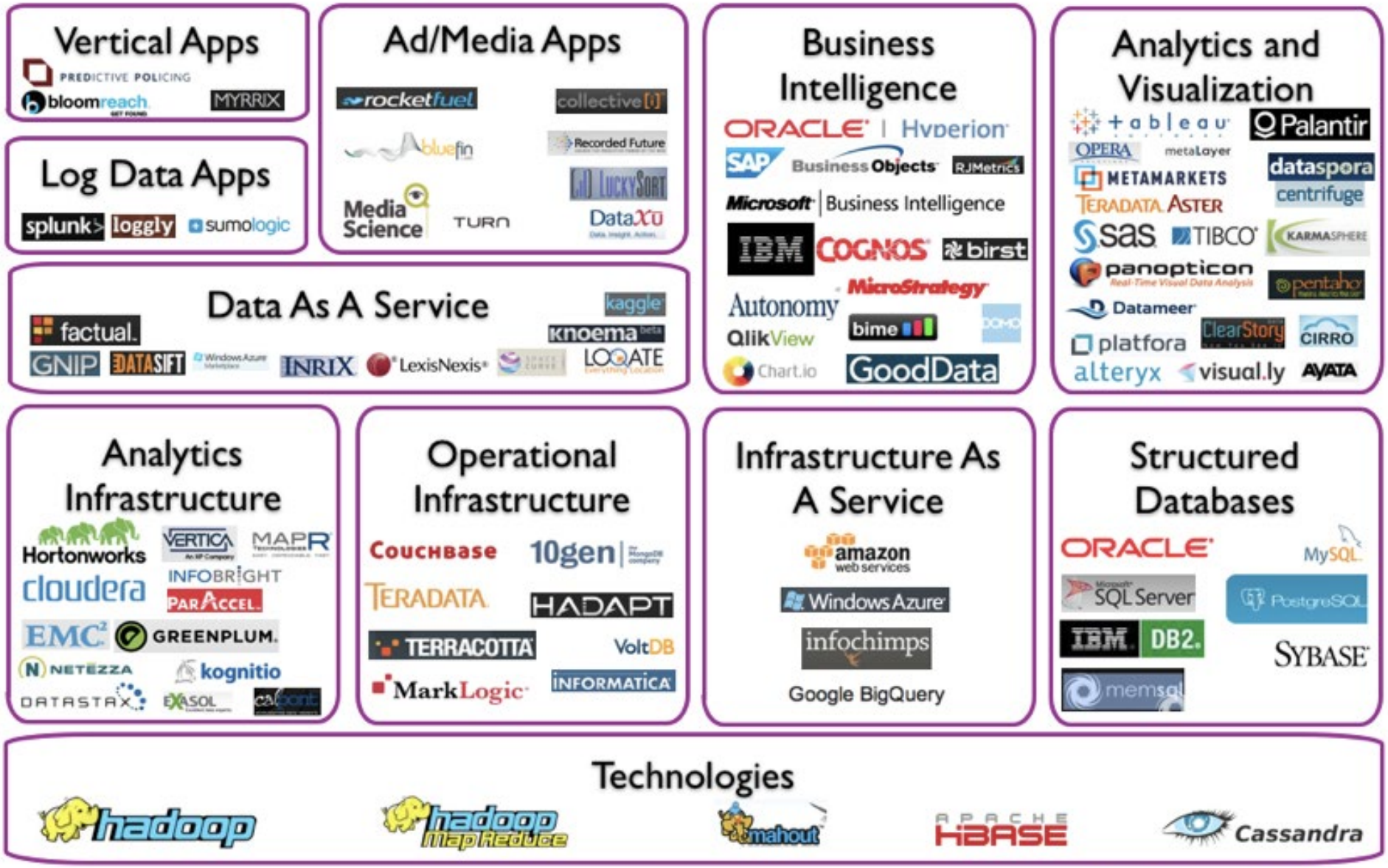
Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

SOURCE: McKinsey Global Institute analysis

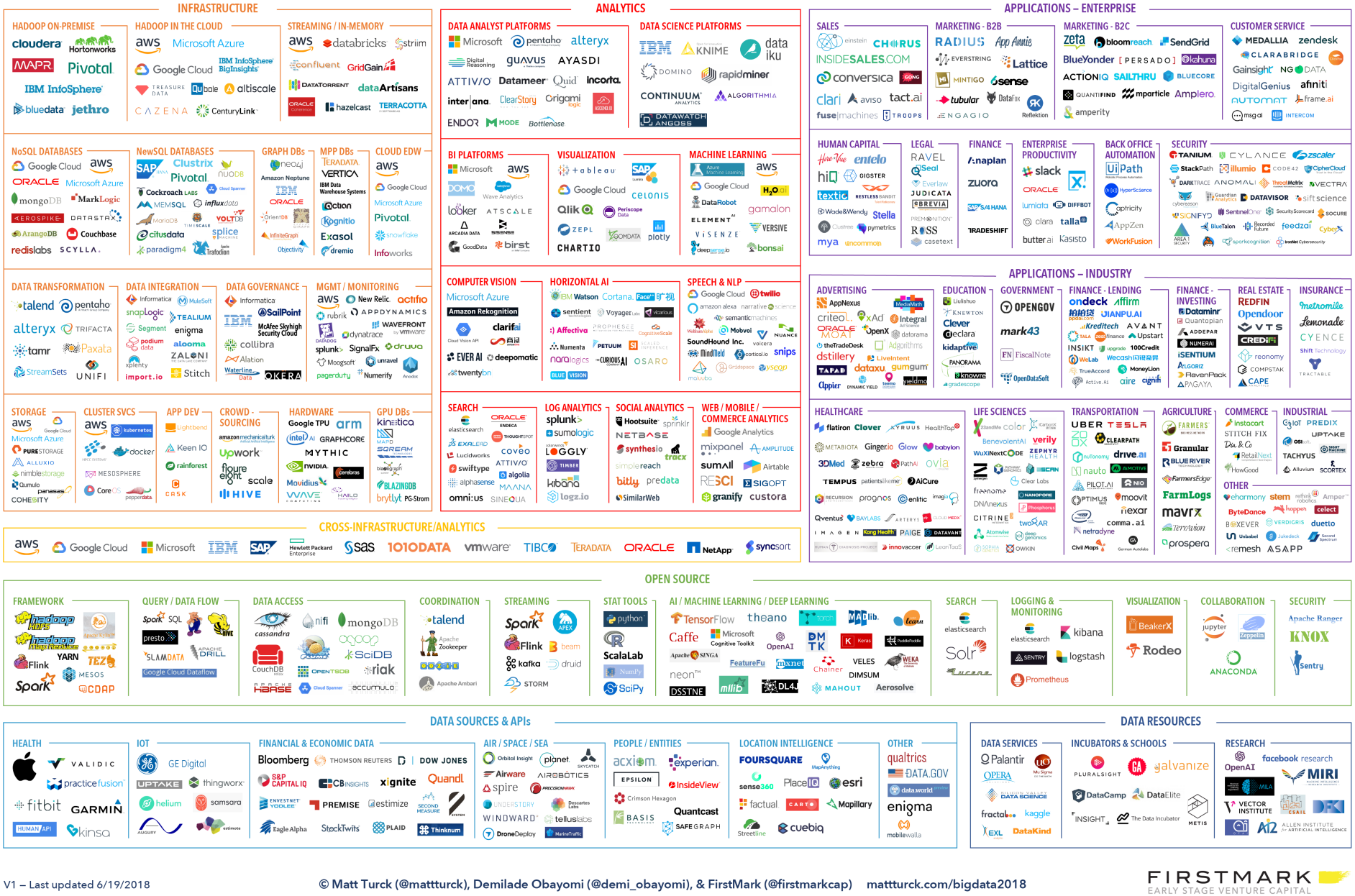
Big Data Landscape



dave@vcdave.com

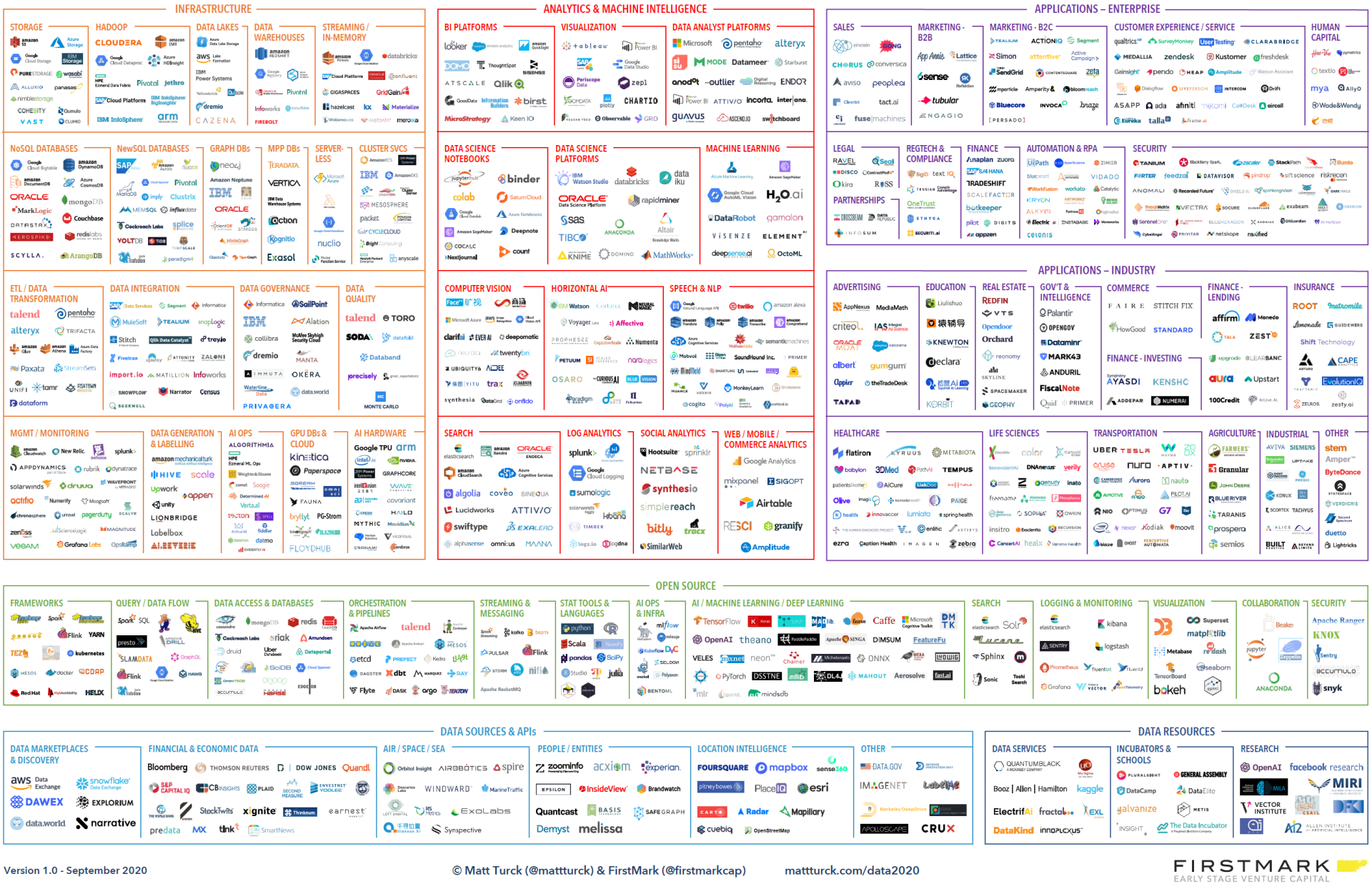
blogs.forbes.com/davefeinleib

BIG DATA & AI LANDSCAPE 2018



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

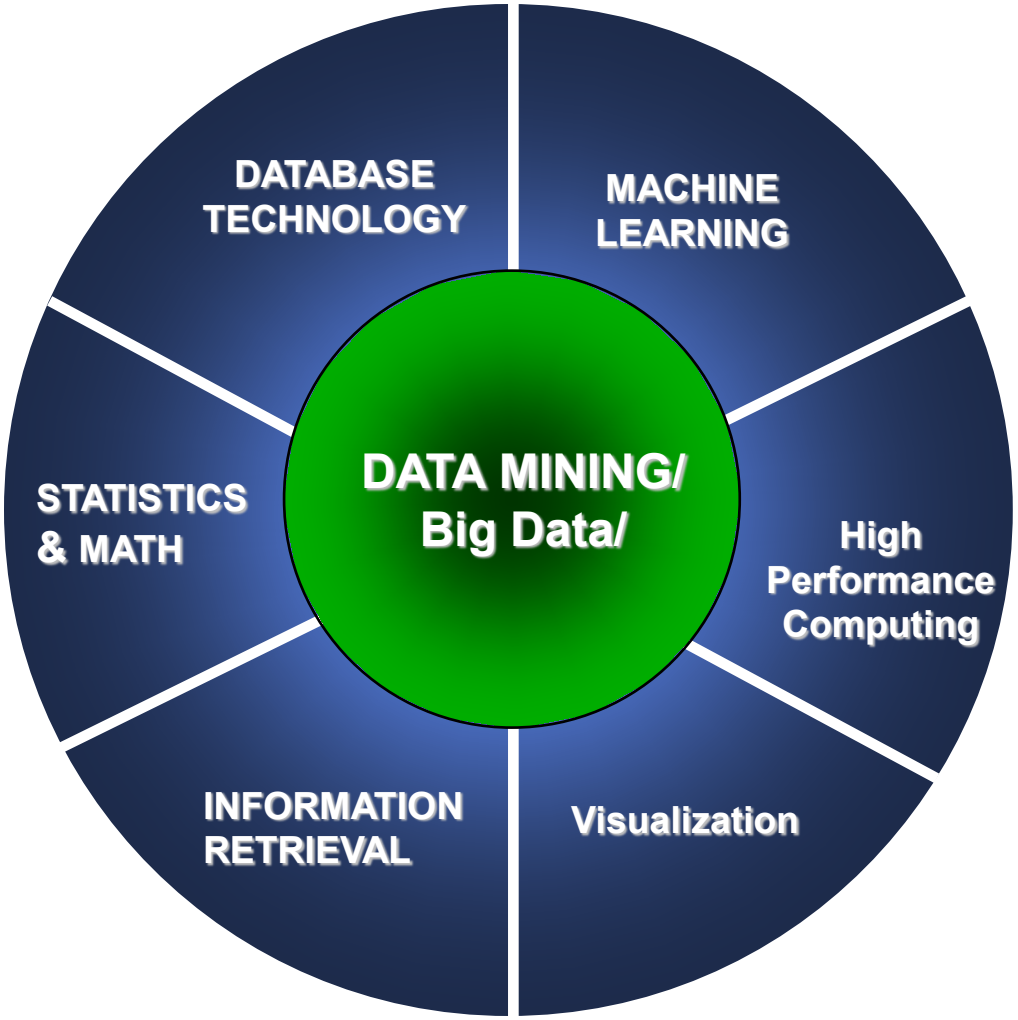
DATA & AI LANDSCAPE 2020



mattturck.com/data2020

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Data Mining/Big Data/Data Science is an Interdisciplinary and Multidisciplinary Field



Why Data Mining/Big Data is Important?

- *McKinsey predicts that data--driven technologies will bring an additional $300 billion of value to the U.S. health care sector alone, by 2020, 1.5 million more "data--savvy managers" will be needed to capitalize on the potential of data*

  – *May 1, 2011*

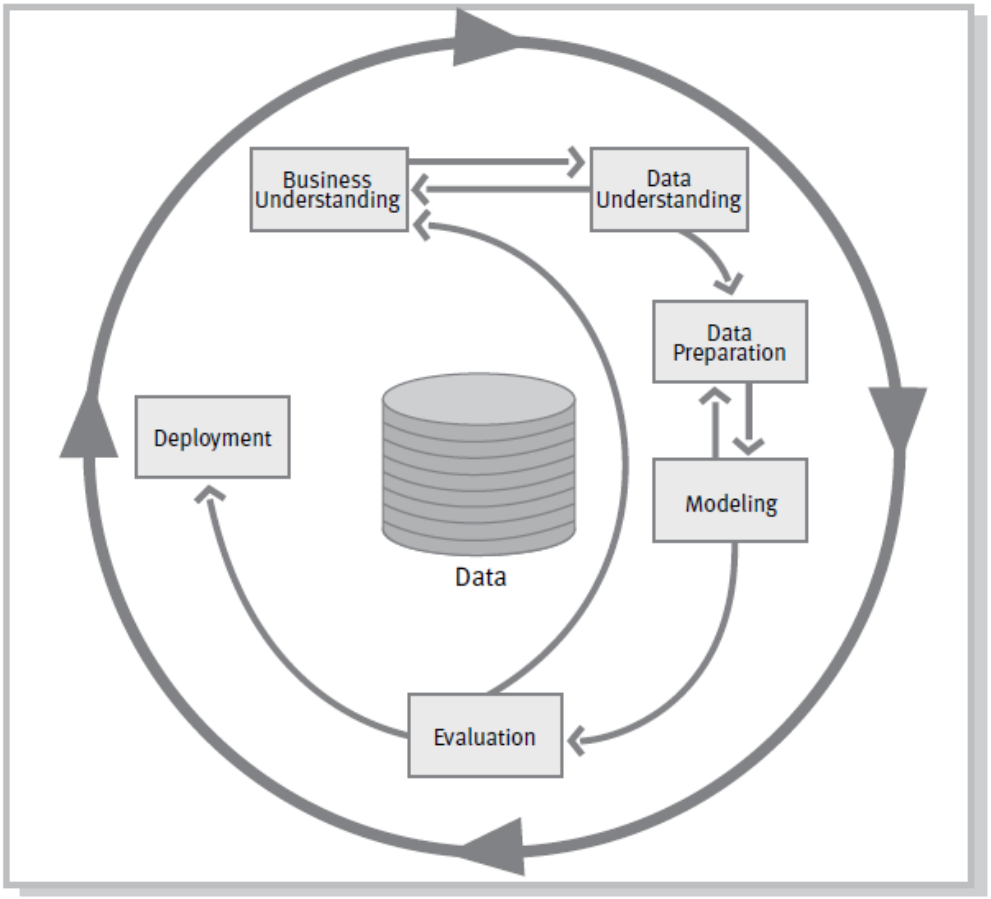
Multi-Dimensional View of Data Mining/Big Data

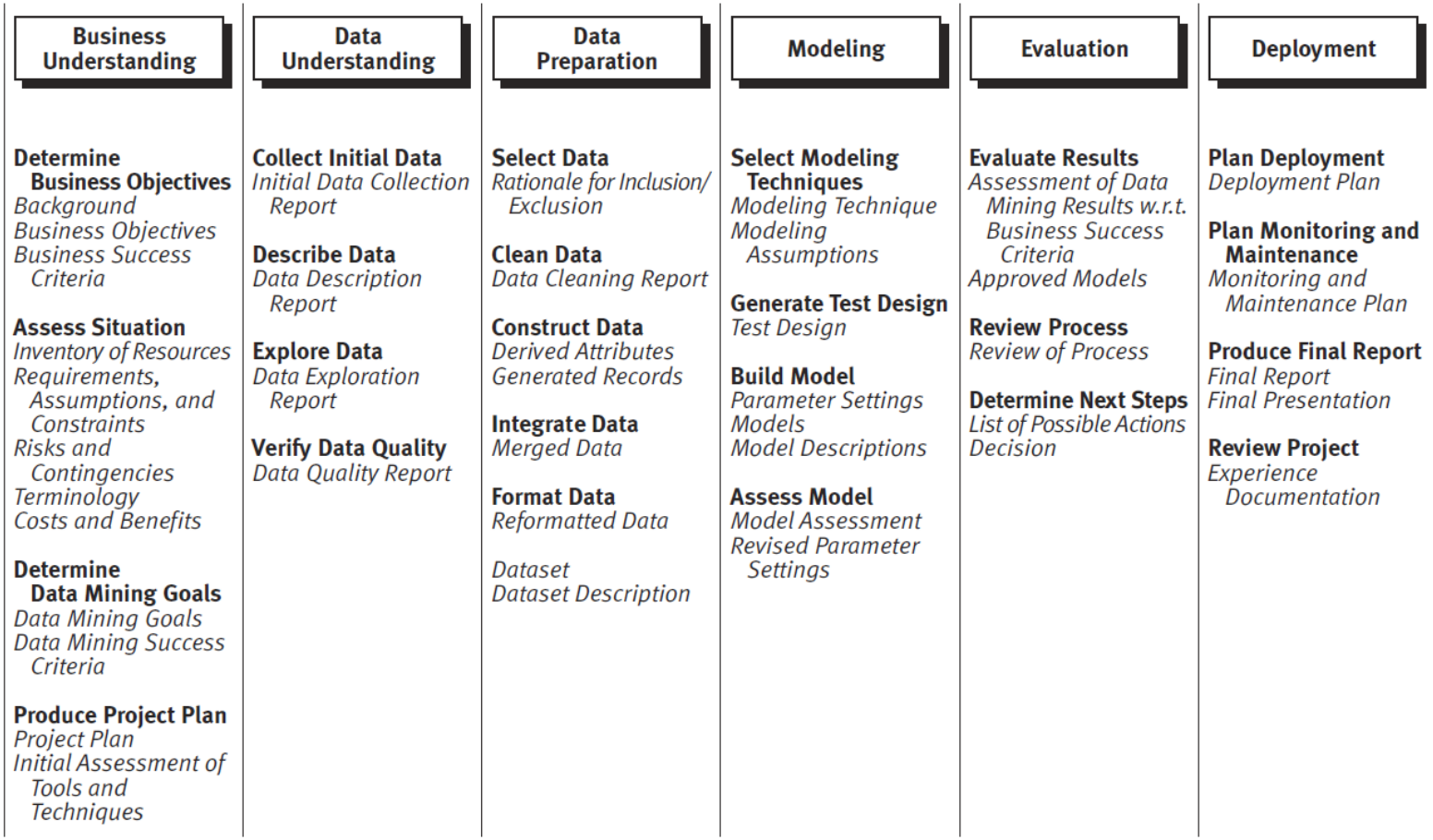
- **Data to be mined**
  - Structured data, semi-structured data, unstructured data, relational data, data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media data, graphs data, social network, etc.
- **Data preprocessing and feature engineering**
  - Cleaning, integration, reduction, transformation, generalization, feature engineering, etc.
- **Knowledge to be mined (Data mining functions and models)**
  - generalization, association, classification, clustering, etc.
- **Techniques utilized**
  - data warehouse (OLAP), frequent pattern mining, machine learning, statistics, pattern recognition, visualization, high-performance computing, parallel and distributed computing, cloud computing, etc.
- **Applications adapted**
  - Market basket analysis, healthcare, manufacturing engineering, education, CRM, fraud detection, customer segmentation, intrusion detection, financial banking, criminal investigation, bio Informatics, direct marketing, classifying stars, medical diagnosis, computer vision, drug discovery, speech recognition, handwriting recognition, credit scoring, human genetic clustering, medical imaging, market segmentation, product positioning, new product development, social network analysis, image segmentation, recommender systems, anomaly detection, crime analysis, climatology, market research, customer trend analysis, market movement tracking, customer buying patterns, analyze tax records, decision making, on line operations and report, etc.

CRISP-DM

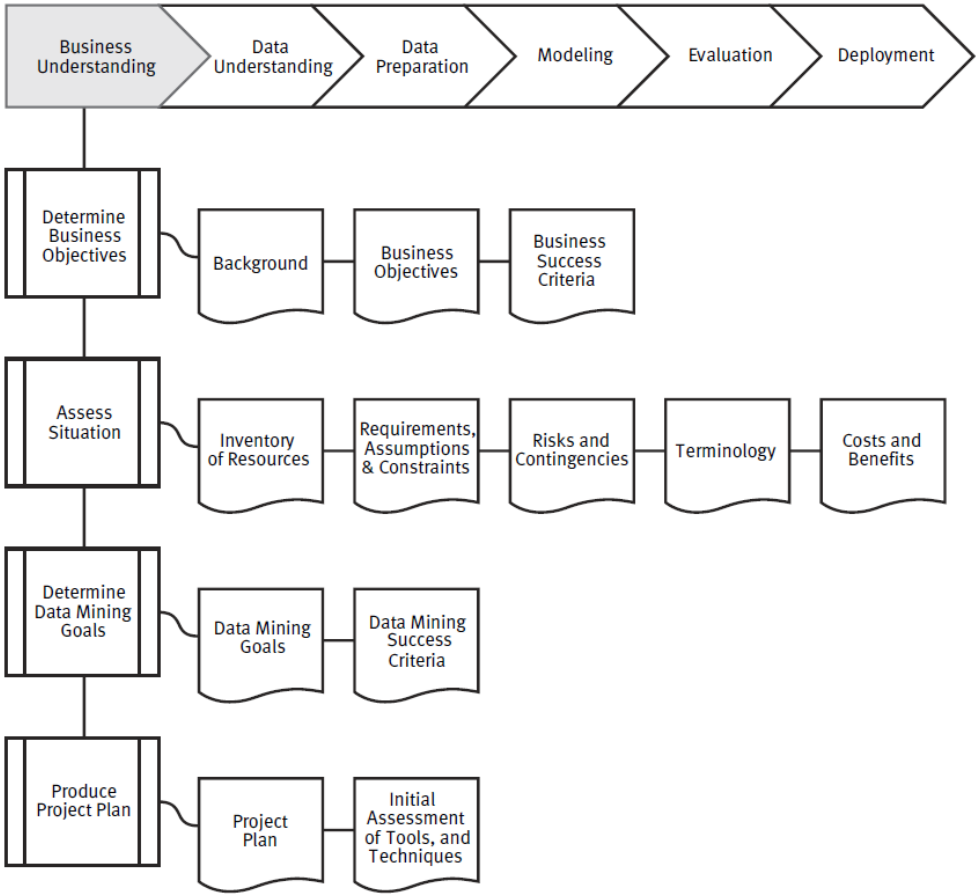
- CRoss-Industry Standard Process for Data Mining
  - is described in terms of a hierarchical process model, consisting of sets of tasks described at four levels of abstraction (from general to specific): phase, generic task, specialized task, and process instance

The life cycle of a data mining project

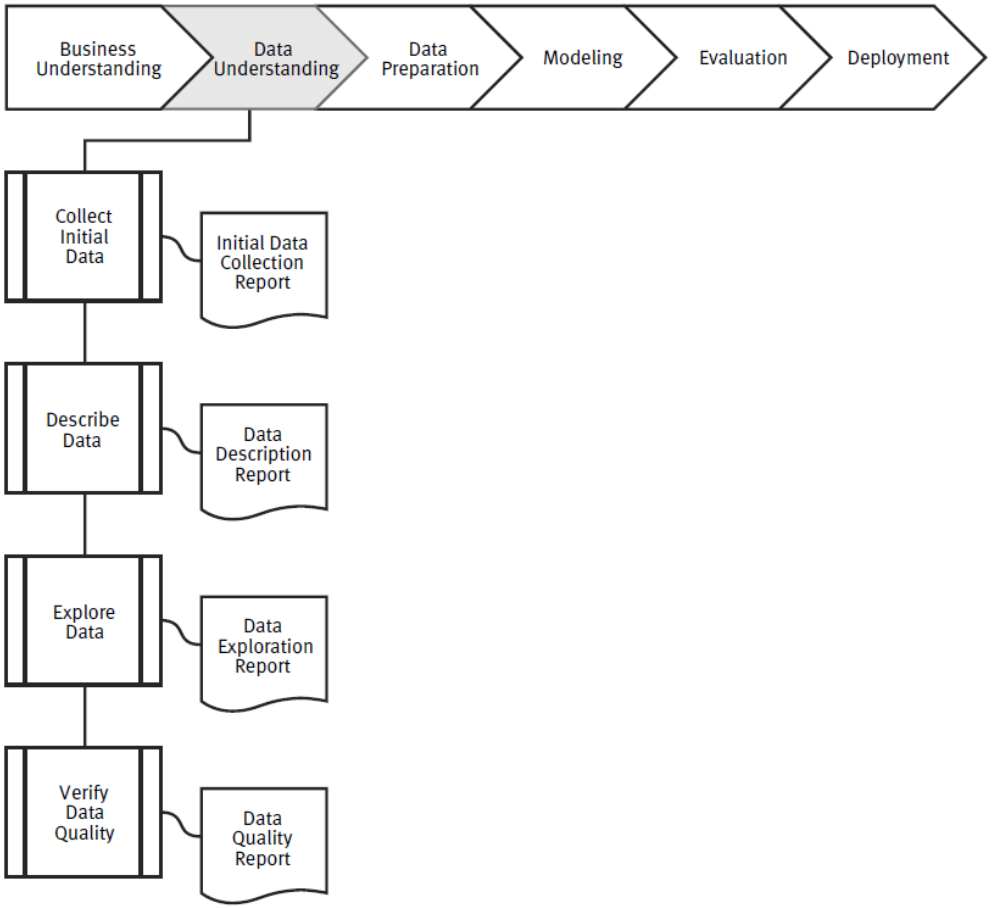



Business Understanding
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
Data Understanding
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report
Data Preparation
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data
Dataset
Dataset Description
Modeling
Select Modeling Techniques
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Descriptions
Assess Model
Model Assessment
Revised Parameter Settings
Evaluation
Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

Business understanding



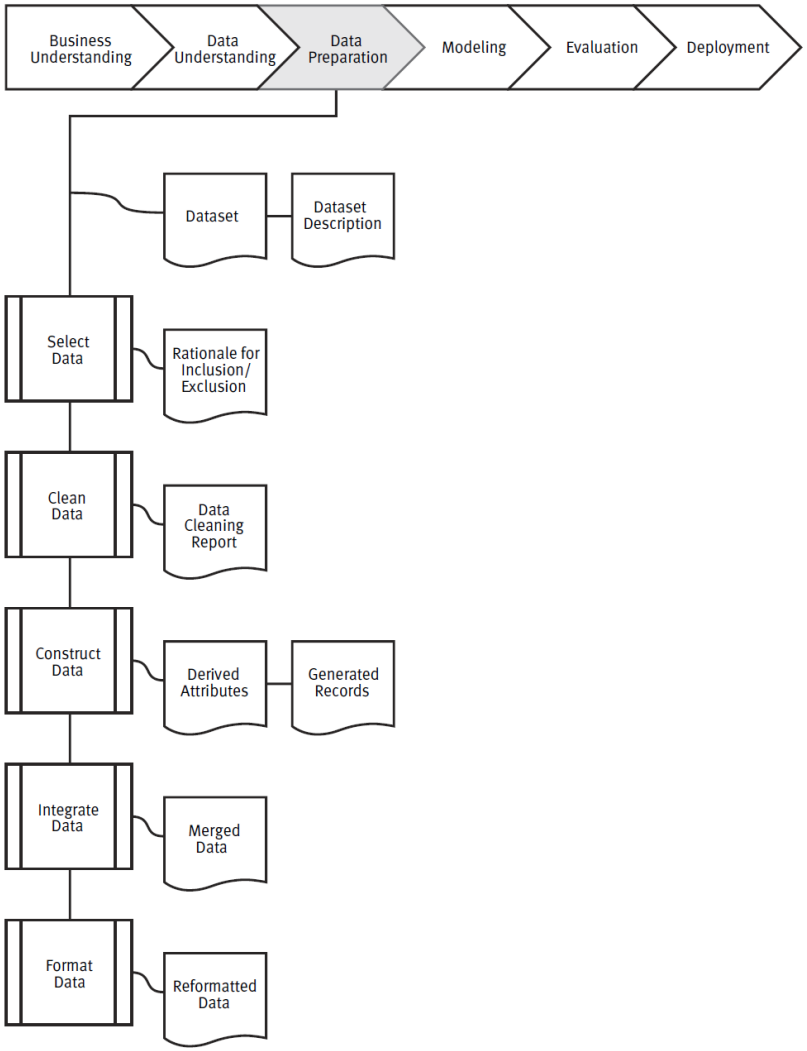
Data understanding



- Structured
- Semi-structured
- Unstructured
- Categorical
  - Nominal
    - Binary
  - Ordinal
- Numeric
  - Interval
  - Ratio
- Discrete
- Continuous

- Descriptive statistics
  - Count, mean, mode, standard deviation, quantiles, etc.
- Class distribution
- Feature skewness
- Correlations between features
  - Chi-square
  - Correlation
- Data visualization

Data preparation



Data preprocessing

Data cleaning

- **Missing values**
  - Use the most probable value to fill in the missing value (and five other methods)
- **Noisy data**
  - Binning; regression; clustering

Data integration

- **Entity ID problem**
  - Metadata
- **Redundancy**
  - Correlation analysis (Correlation coefficient, chi-square test)

Data trasnformation

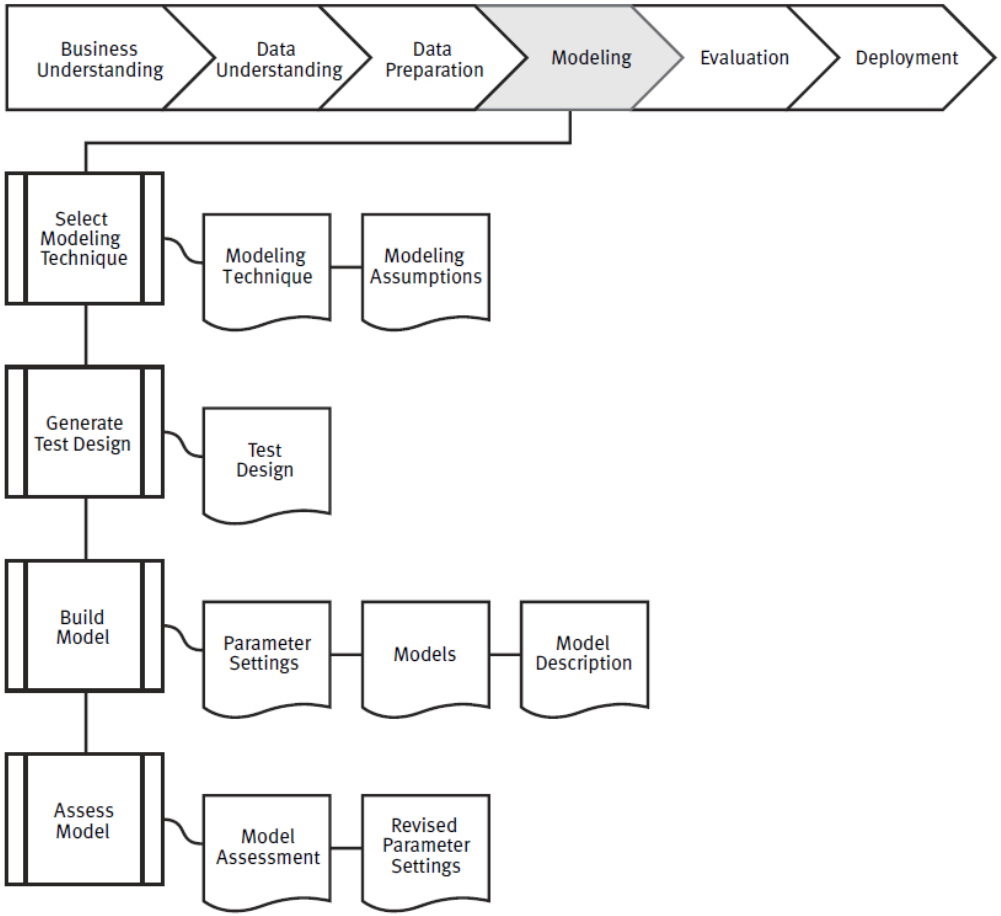
- **Smoothing**
  - **Data cleaning**
- **Aggregation**
  - **Data reduction**
- **Generailization**
  - **Data reduction**
- **Normalization**
  - Min-max; z-score; decimal scaling
- **Attribute construction**
  - New attributes are constructed and added from the given set of attributes to help mining process

Data reduction

- **Data cube aggregation**
  - Data cube store multidimensional aggregated information
- **Attribute selection**
  - Stepwise forward selection; stepwise backward selection; combination; decision tree induction
- **Dimensionality reduction**
  - Discrete wavelet transforms (DWT); Principle components analysis (PCA)
- **Numerosity reduction**
  - Parametric data reduction; clustering; data cube aggregation;
- **Data discretization**
  - Binning; historgram analysis; entropy-based discretization;
    Interval merging by chi-square analysis; cluster analysis; intuitive partitioning
- **Concept hierarchy**
  - Concept hierarchy generation

- Feature Engineering
  - Curse of dimensionality
  - Feature selection
    - Filter methods
    - Wrapper methods
    - Embedded methods
    - Hybrid methods
  - Feature construction
  - Feature creation

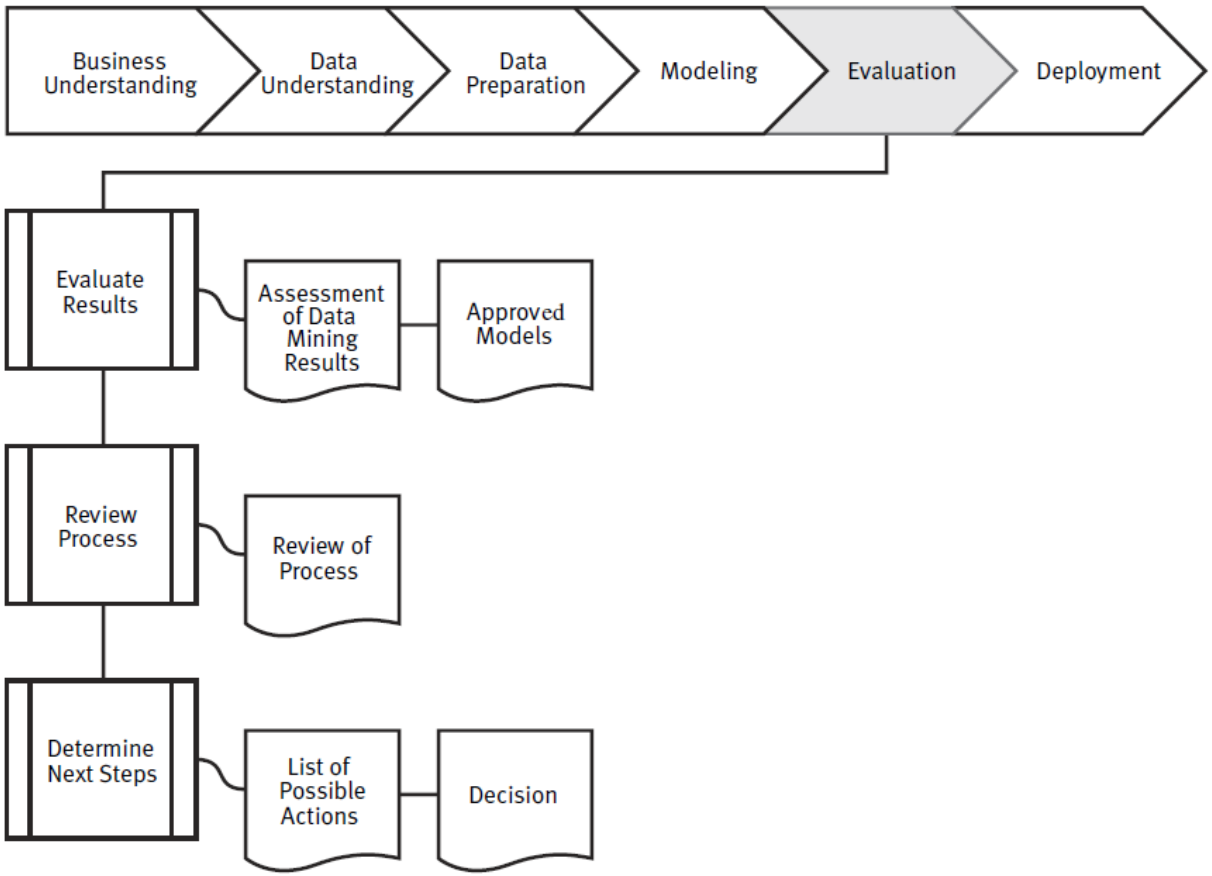
Modeling


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Settings

- Decision tree
- Bayesian classifier
- Regression
  - Simple linear regression
  - Multiple linear regression
  - Polynomial regression
  - Logistic regression
- k-Nearest Neighbor classifier (kNN)
- Perception
- Gradient Descent
- Neural Network
  - Feed-Forward Neural Network (FNN)
  - Backpropagation
  - Recurrent Neural Network (RNN)
  - Convolutional Neural Network (CNN)
  - Probabilistic Neural Network (PNN)
- Support Vector Machine (SVM)
- Binary Classification
- Multiclass Classification
- Rule-Based Classification
- Bayesian Belief Networks
- Genetic Algorithms
- Rough Set Approach
- Fuzzy Set Approaches
- Etc.

- Multidimensional Data Modeling
  - Data Cube: A Multidimensional Data Model
  - Stars, Snowflakes, and Fact Constellations
  - Dimensions: The Role of Concept Hierarchies
  - OLAP Operations
- Mining Frequent Patterns, Associations, and Correlations
  - Apriori Algorithm
  - Frequent Itemset
  - Generating Association Rules from Frequent Itemsets
- Clustering
  - Partitional methods
    - k-means
    - k-medoids
  - Hierarchical methods
    - Agglomerative hierarchical clustering method
    - Divisive hierarchical clustering method
  - Density-Based Methods
    - DBSCAN: Density-Based Clustering Based on Connected Regions with High Density
    - OPTICS: Ordering Points to Identify the Clustering Structure
    - DENCLUE: Clustering Based on Density Distribution Functions
  - Grid-Based Methods
    - STING: STatistical INformation Grid
    - CLIQUE: An Apriori-like Subspace Clustering Method

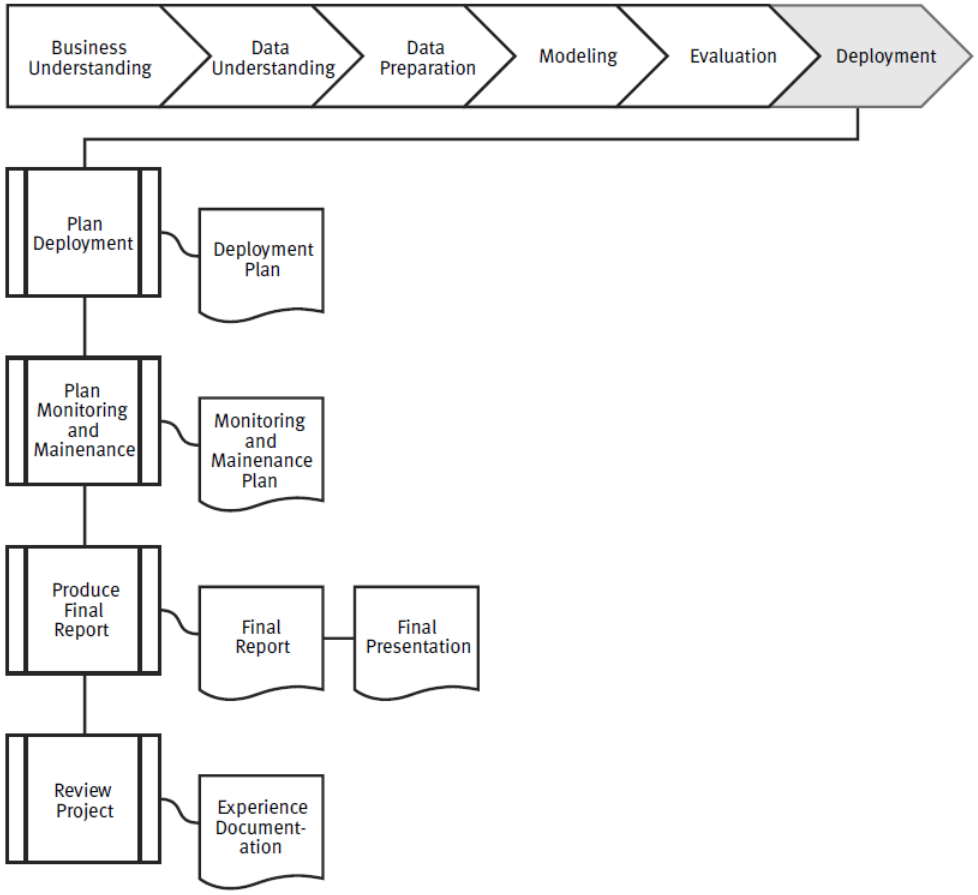
Evaluation


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment of Data Mining Results
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision

confusion matrix
accuracy, recognition rate
error rate, misclassification rate
sensitivity, true positive rate, recall
specificity, true negative rate, precision
F1-score, harmonic mean of precision and recall
F-beta , where is a non-negative real number
k-fold cross-validation
bootstrap
.632 bootstrap
model selection using statistical tests of significance
comparing classifiers based on cost–benefit and ROC curves

Ensemble Methods
- Random Forests
- Bagging
- Boosting
- AdaBoost

Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Mainenance
Monitoring and Mainenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Document-ation