Statistical Computing with R
Masters in Data Science 503 (S8)
First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Review Preview (Unit 2, Session 3)

- **Data Mining**
  - **Resources used/recommended:**

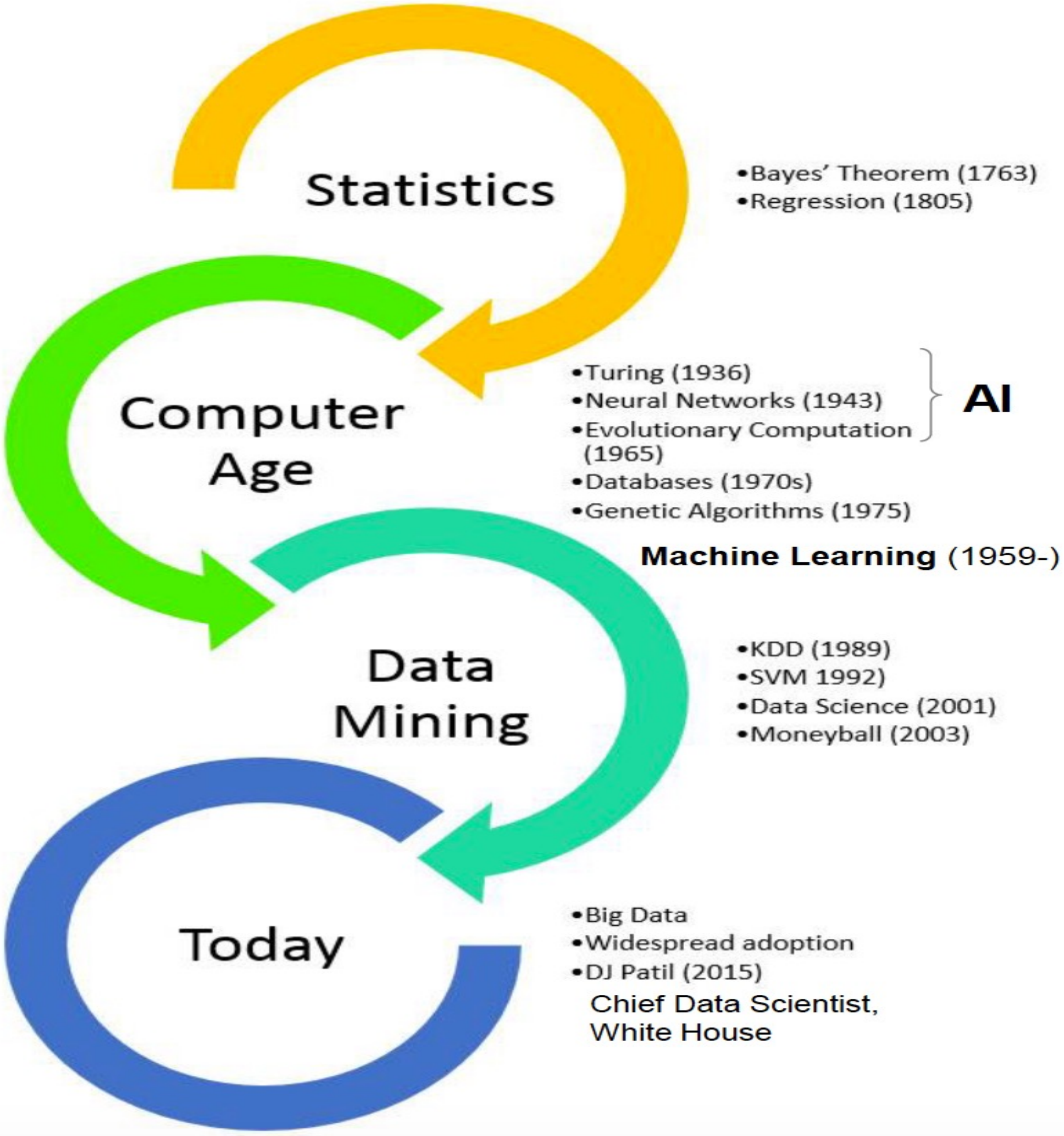
1. https://online.stat.psu.edu/stat857/intro/
2. https://michael.hahsler.net/research/misc/Intro_Data_Mining.mini.pdf
3. https://www.rdatamining.com/

- **Text Mining**
  - **Resource used/recommended:**

1. **RDataMining-slides-text-mining.pdf, GD link from rdatamining.com:** https://drive.google.com/file/d/1JSlWQLPrAUrtdLrGFuS8kckxhqHp885f/view
2. **R and Data Mining: Examples and Cases Studies. Text Mining (Chapter 10), 2015:** https://drive.google.com/file/d/1gn7cMdpMkDwHVTfDldAkn5i3_pRtoH-H/view

Origins of Data Mining

- Draws ideas from AI, machine learning, pattern recognition, statistics, and database systems.
- There are differences in terms of
  - used data and
  - the goals.



https://rayli.net/blog/data/history-of-data-mining/

Data Mining (What is):

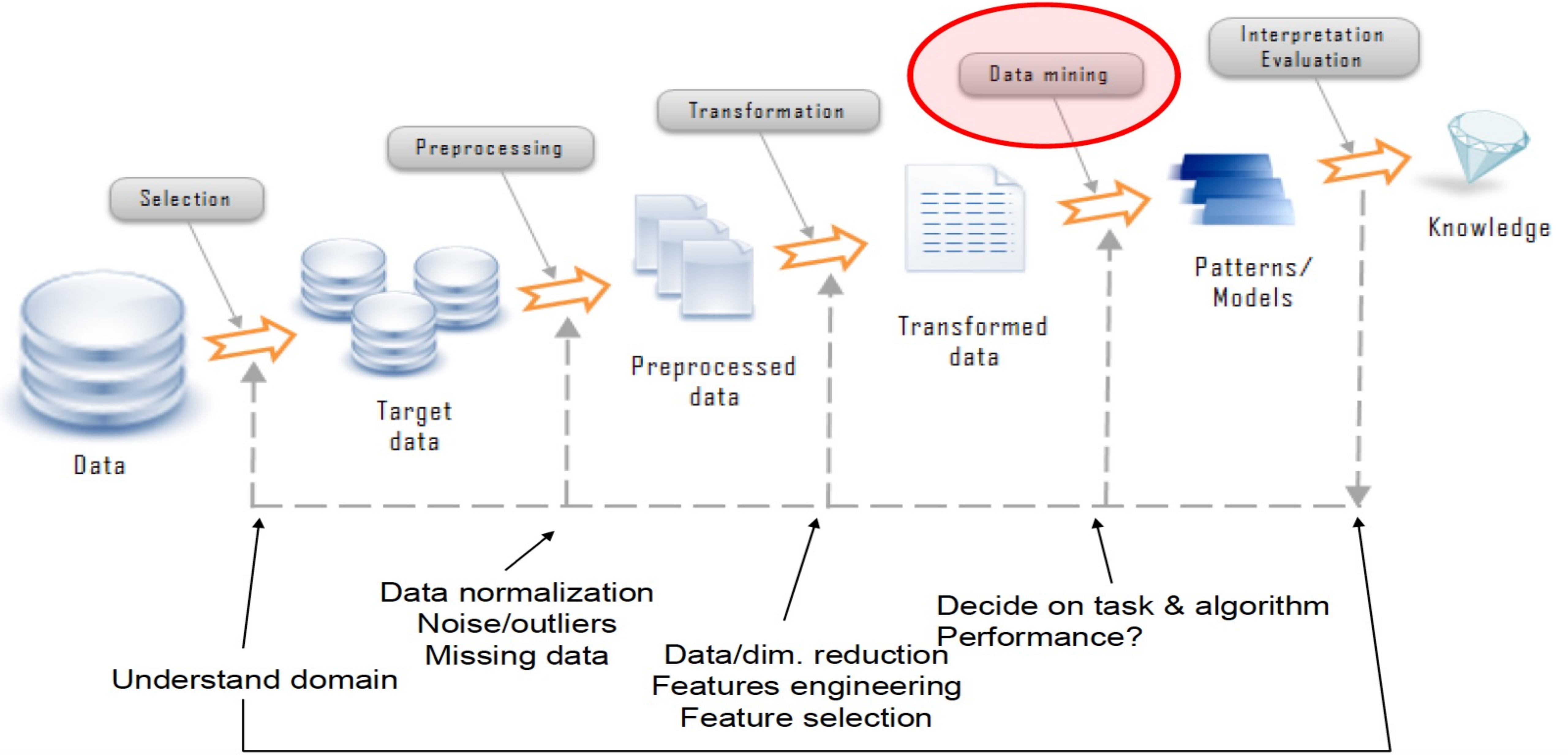
- Data Mining refers to a set of methods applicable to large and complex databases to eliminate the randomness and discover the hidden pattern. (https://online.stat.psu.edu/stat857/node/142/)

- Data Mining is the science of **extracting useful information** from huge data repositories/warehouse (http://www.kdd.org/curriculum)

- Data Mining helps to:
  - identify patterns and relationships
  - classify and segment data
  - formulate hypothesis

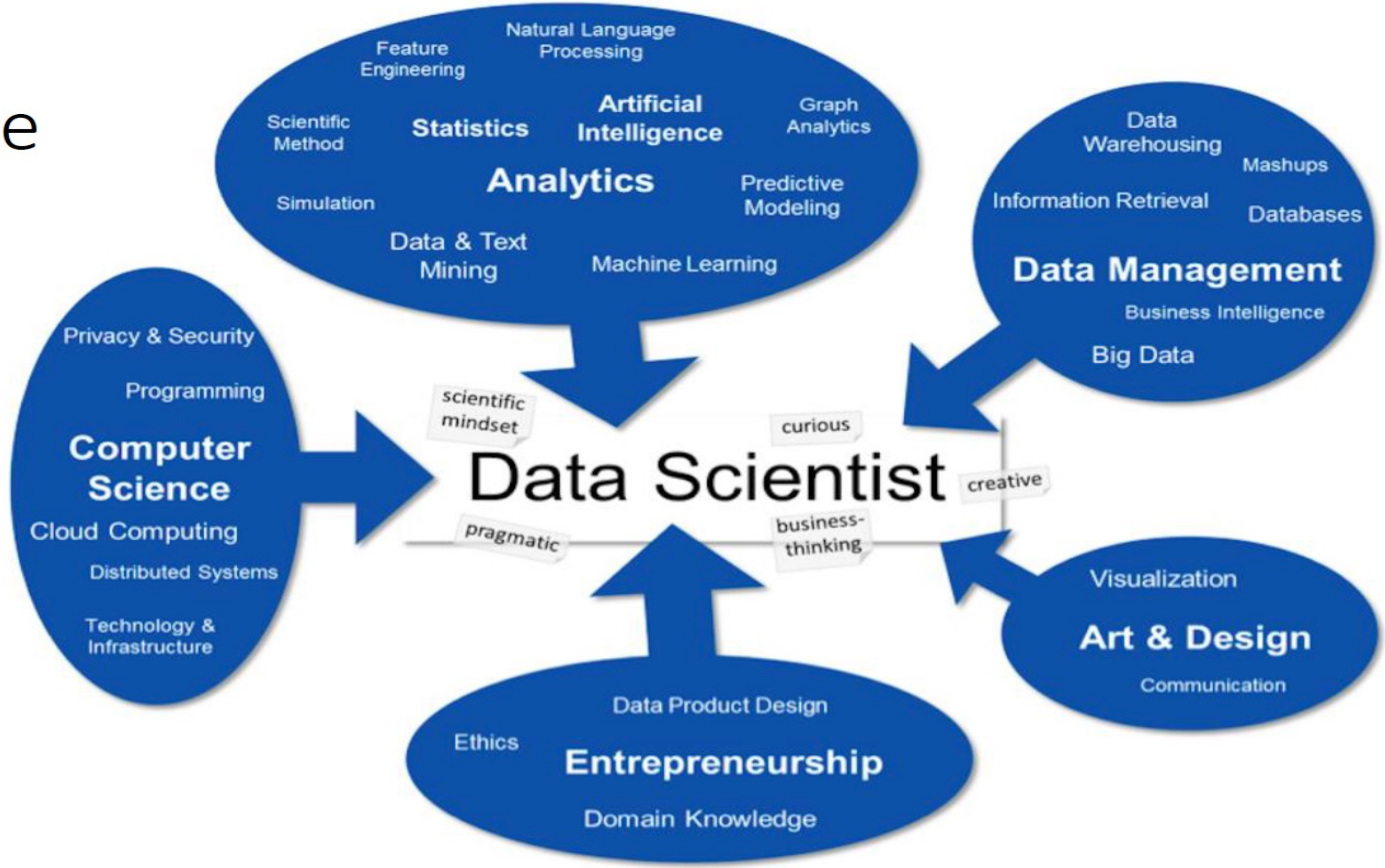
KDD = Knowledge Discovery in/from Database

Knowledge Discovery in Databases (KDD) Process



Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery: an overview.

Data Science



Source: T. Stadelmann, et al., Applied Data Science in Europe

For Data Science, Data Mining is:

- interdisciplinary and overlaps significantly with many fields such as
  - Statistics
  - Computer Science (Machine Learning, AI, Databases)
  - Optimization
- requires a team effort with members who have expertise in several areas such as
  - Data management
  - Statistics
  - Programming
  - Communication
  - + application domain (health, business, physics, biology etc.)

(IBM) CRISP-DM Reference Model:

- Cross Industry Standard Process for Data Mining (CRISP-DM):

- Business Understanding
- Data understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

Tasks in the CRISP-DM Model

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

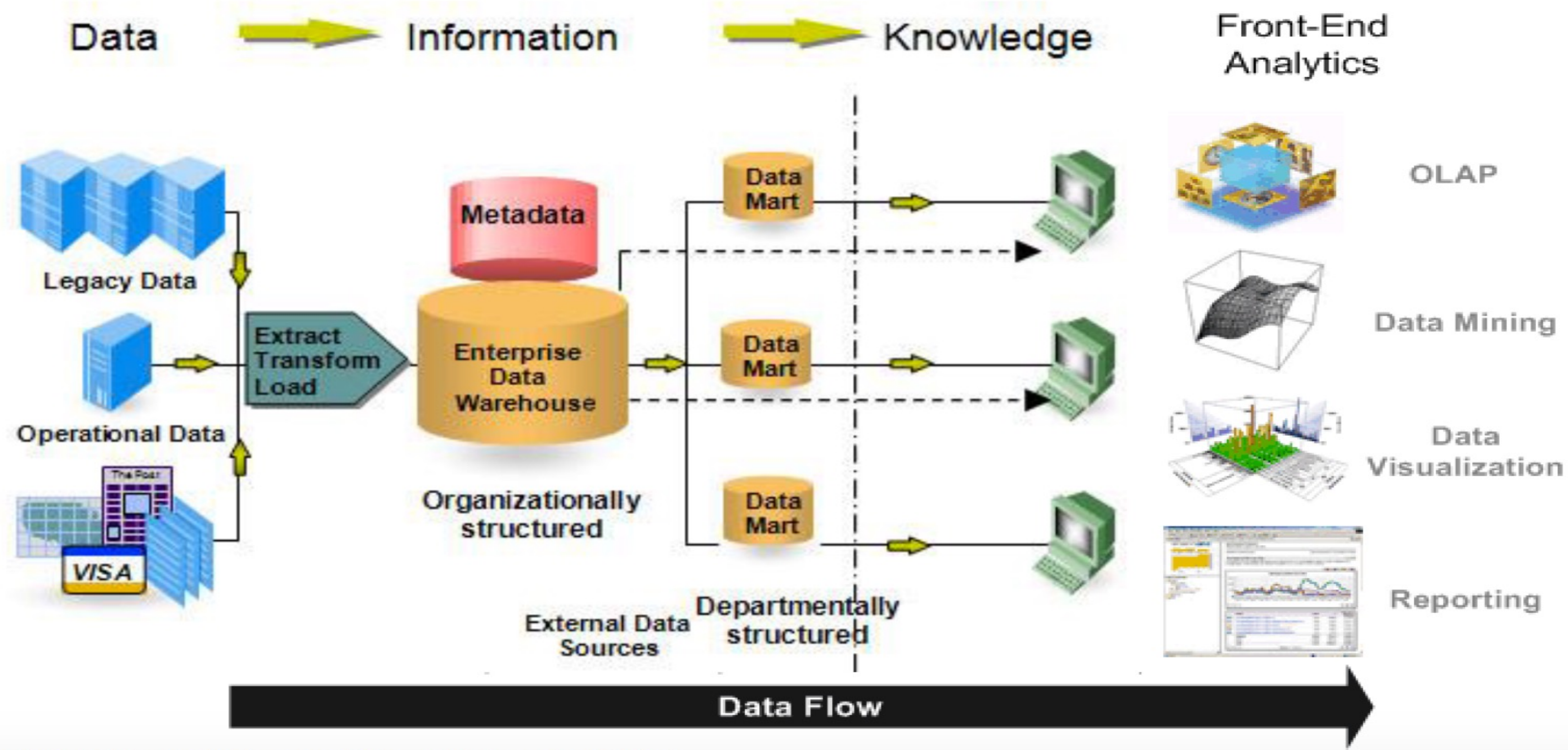
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Data Warehouse



http://www.fulcrumlogic.com/data_warehousing.shtml

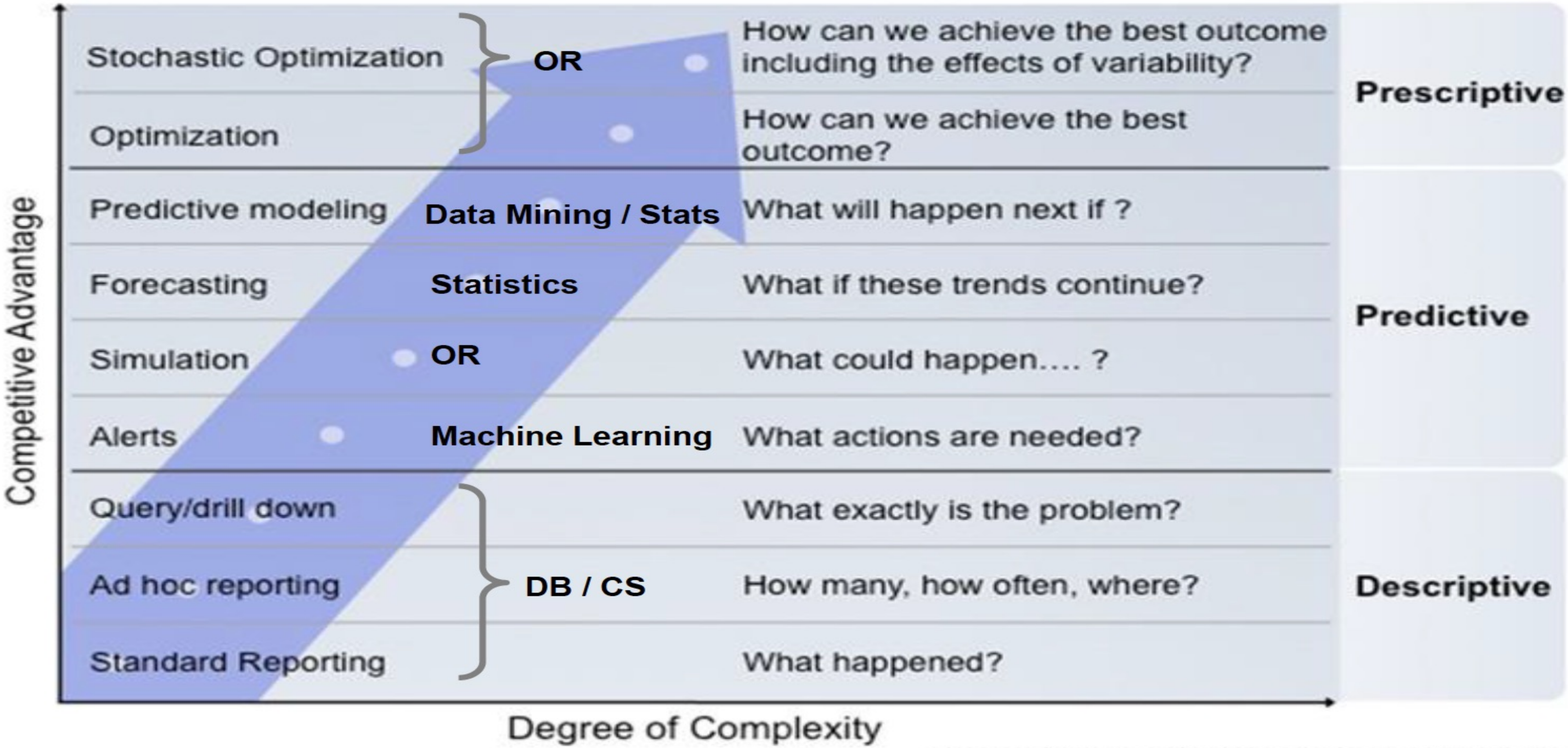
What is?

- Data:
  - Legacy data?
  - Operational data?
- ETL process?
- Information:
  - Metadata?
  - Enterprise Data Warehouse?
  - Data Mart?
- Knowledge:
  - OLAP?
  - Data Mining?

Data Mining Tasks:

- Descriptive Methods:
  - Find human interpretable patterns that describes the data (Unit 1, 2 and 3 of this course)
- Predicting Methods:
  - Use some features (variables) to predict unknown or future value of other variable (Unit 4 and 5 of this course)
- Prescriptive Methods:
  - Optimization
  - Stochastic optimization

Data Mining & Analytics



Based on: Competing on Analytics, Davenport and Harris, 2007

Data Mining Tasks:

- **Predictive Modelling (Regression and classification) – Unit 4**
- **Dimensionality Reduction, Cluster Analysis – Unit 5**
- Association Analysis – Not covered in this course
- Anomaly detection – Not covered in this course

Predictive modelling:

- **Supervised Learning**

- Regression
  - Linear regression (simple and multiple)
  - Logistic regression (bi-variate and multivariate)

- Classification
  - Decision trees, Random forests, Neural networks
  - Support Vector Machines, Naïve Bayes

- **We will discuss more on these topics in Unit 4**

Dimensionality Reduction: Column/Variable Based Data Reduction Methods

- **Unsupervised Learning**

- Principal Component Analysis (PCA)

- Principal Axis Factoring (PAF)

- Multidimensional scaling (MDS)
  - Classical (Principal coordinate analysis)
  - Metric MDS, Non-metric MDS
  - Generalized MDS

Cluster Analysis: Row/Case Based Data Reduction Methods (HC, k-means etc.)

- **Unsupervised Learning**
- Data points in one cluster are more similar to one another
- Data points in separate clusters are less similar to one another
- **We will discuss more on dimension reductions/cluster analysis in Unit 5**

Data Mining Tools:

- Simple Graphical User Interface (GUI) based on R
  - Weka
  - Rattle
- Process oriented
  - Rapid Miner
  - IBM SPSS Modeler
  - SAS Enterprise Miner etc.
- Programming oriented
  - R, Rattle, R Studio (shiny), Microsoft R (reticulate package to run python in R)
  - Python, Numpy, Scipy scikit-learn, pandas, Jupyter notebook (rpy2 to run R in python)

Other Data Mining Tasks:

- **Text Mining (we will discuss it today with an example from web)**
- **Graph Mining (Unit 3)**
- Data stream mining – not covered in this course
- Mining spatiotemporal data (e.g. moving objects) – not covered
- Distributed data mining etc. – not covered in this course

Question/queries so far?

Text Mining:

- Import texts (Interviews, Twits, Facebook posts, Comments, Reviews etc.) in R
- Transform the texts to data frame and define the “Corpus”
- Perform pre-processing of the “Corpus” using standard methods
- Build document-term matrix (DTM)
- Find frequent terms and associations of key term with other terms
- Use network graph/word cloud to visualize the DTM
- Perform cluster analysis to find clusters of similar words
- Perform “topic modelling”, if required!

Packages required for Text Mining:

- Text mining: *tm*

  (Details: https://cran.r-project.org/web/packages/tm/tm.pdf)

- Topic modelling: *topicmodels, lda*

- Word cloud: *wordcloud*

- Twitter data access: *twitteR* (Optional)

Example of tweet mining: rdatamining.com

(Alternative solution: https://rstudio-pubs-static.s3.amazonaws.com/66739_c4422a1761bd4ee0b0bb8821d7780e12.html)

Option 1: retrieve tweets from Twitter

- library(twitteR)
- tweets <- userTimeline("RDataMining", n = 3200)

Option 2: download @RDataMining tweets from RDataMining.com

- url <- "http://www.rdatamining.com/datasets/rdmTweets.RData"

  download.file(url, destfile = "./data/rdmTweets.RData")

Option 3: Download @RDataMining tweeks from RDataMining.com manually: http://www.rdatamining.com/datasets/rdmTweets.RData and save it to the folder you want to use e.g. Downloads!

Load tweets in R, check length and its structure (The "twitteR" package must be installed *a priori*):

- load(file = "./data/rdmTweets.RData") #Option 2 used!
- choose.file() #Locate rmdTweets located at "Downloads" folder

- (n.tweet <- length(tweets)) #If rmdTweets is assigned as "tweets"
- [1] 320 #Option 2 used, 320 tweets only!

- strwrap(tweets[[320]]$text, width = 55) #Text variable of tweet 320
- [1] "An R Reference Card for Data Mining is now available"
- [2] "on CRAN. It lists many useful R functions and packages"
- [3] "for data mining applications."

Text cleaning in R: Pre-processing I (data frame, corpus, lower case, punctuation)

- library(tm)

convert tweets to a data frame

- df <- twListToDF(tweets)

build a corpus

- myCorpus <- Corpus(VectorSource(df$text))

convert to lower case

- myCorpus <- tm_map(myCorpus, tolower)

remove punctuations and numbers

- myCorpus <- tm_map(myCorpus, removePunctuation)
- myCorpus <- tm_map(myCorpus, removeNumbers)

Text cleaning in R: Pre-processing II (Remove URL and Stop Words)

remove URLs, http followed by non-space characters

- removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
- myCorpus <- tm_map(myCorpus, removeURL)

remove r and big from stopwords

- myStopwords <- setdiff(stopwords("english"), c("r", "big"))

remove stopwords

- myCorpus <- tm_map(myCorpus, removeWords, myStopwords)

Text cleaning in R: Pre-processing III (Stemming, be careful with this process!)

keep a copy of corpus

- myCorpusCopy <- myCorpus

stem words

- myCorpus <- tm_map(myCorpus, stemDocument)

stem completion

- myCorpus <- tm_map(myCorpus, stemCompletion, dictionary = myCorpusCopy)

replace "miners" with "mining", because "mining" was first stemmed to "mine" and then completed to "miners"

- myCorpus <- tm_map(myCorpus, gsub, pattern="miners", replacement="mining")
- strwrap(myCorpus[320], width=55) #check the corpus again (iteratively)!

[1] "r reference card data mining now available cran list"

[2] "used r functions package data mining applications"

Check “Frequent terms”:

- myTdm <- TermDocumentMatrix(myCorpus, control=list(wordLengths=c(1,Inf)))

inspect frequent words

(freq.terms <- findFreqTerms(myTdm, lowfreq=20))

- [1] "analysis" "big" "computing" "data" ...
- [5] "examples" "mining" "network" "package"...
- [9] "position" "postdoctoral" "r" "research...
- [13] "slides" "social" "tutorial" "universi...
- [17] "used"

Check “Associations” with word “r”: Association >=0.2 of r with other words!

- # which words are associated with r?

findAssocs(myTdm, "r", 0.2)

- ## r
- ## examples 0.32
- ## code 0.29
- ## package 0.20

What is done here?
(This will not work if stemming is not corrected!)

which words are associated with

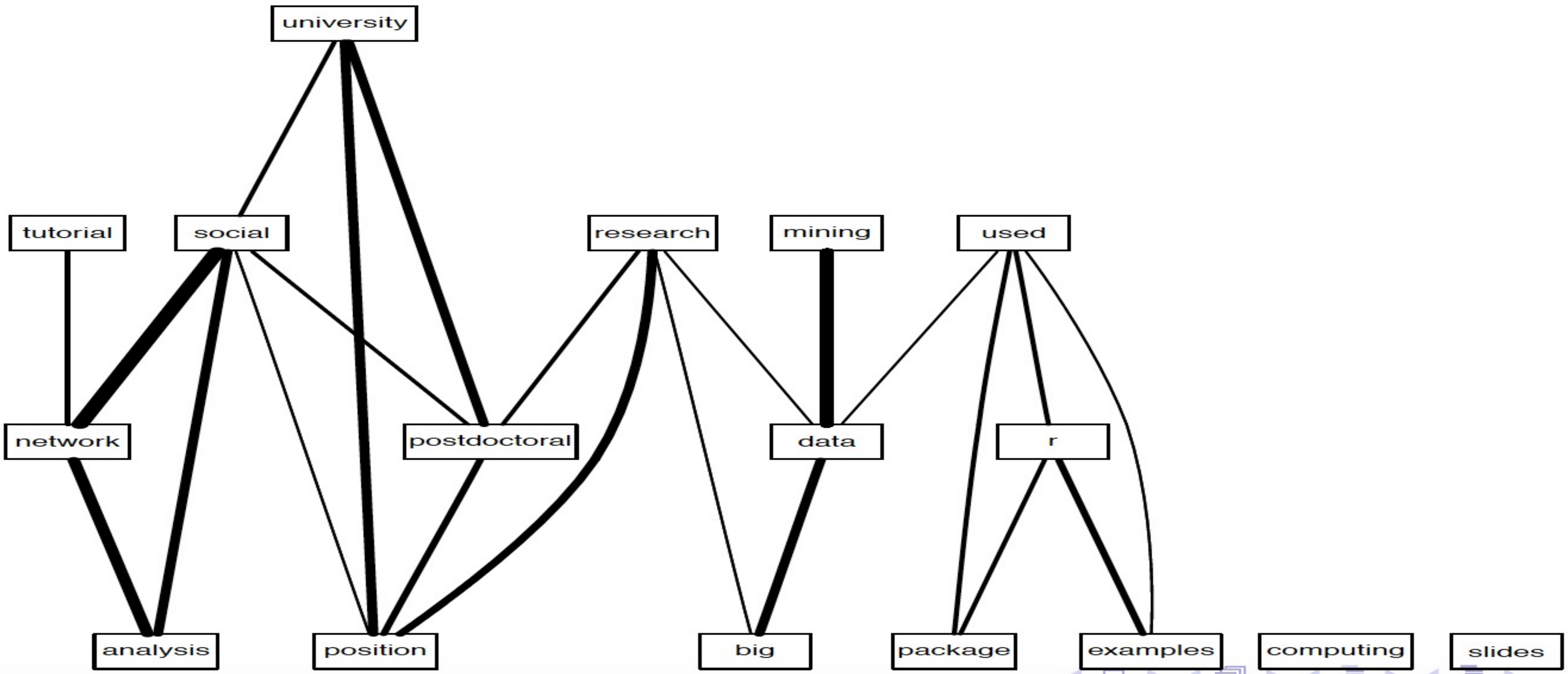
- findAssocs(myTdm, "mining", 0.25)

mining

- data 0.47
- mahout 0.30
- recommendation 0.30
- sets 0.30
- supports 0.30
- frequent 0.26
- itemset 0.26

Network of Terms

```
library(graph)
library(Rgraphviz)
plot(myTdm, term=freq.terms, corThreshold=0.1, weighting=T)
```



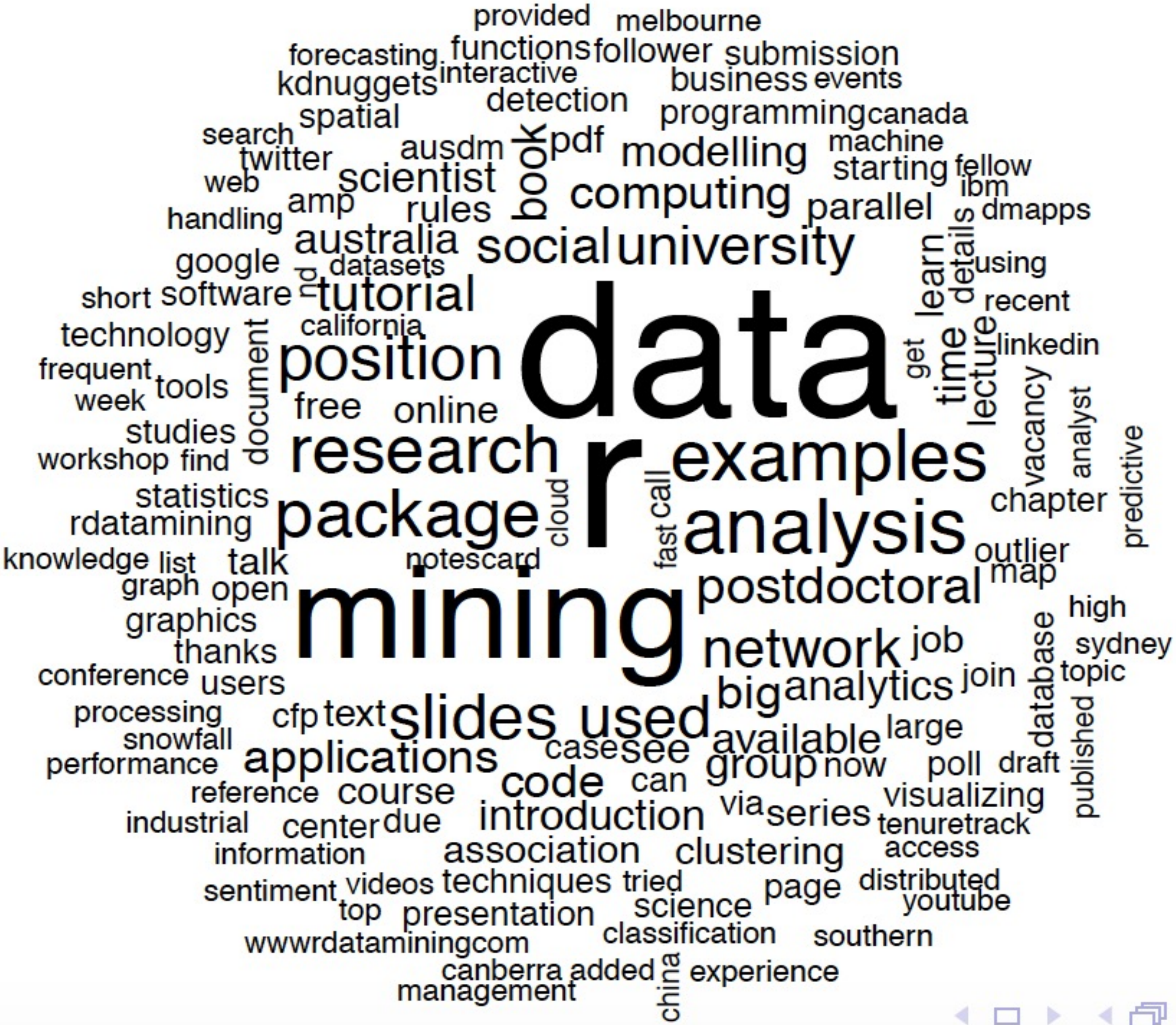
Word cloud:

- library(wordcloud)

- m <- as.matrix(myTdm)

- freq <- sort(rowSums(m), decreasing=T)

- wordcloud(words=names(freq), freq=freq, min.freq=4, random.order=F)


provided
melbourne
forecasting
functions
follower
submission
kdnuggets
interactive
business
events
detection
programming
canada
spatial
search
ausdm
book
pdf
modelling
machine
twitter
scientist
starting
fellow
web
computing
ibm
amp
rules
parallel
dmapps
handling
australia
social
university
details
learn
using
google
datasets
nd
short
software
tutorial
recent
technology
california
data
linkedin
frequent
position
get
time
lecture
tools
week
free
online
document
vacancy
analyst
studies
research
r
examples
predictive
workshop
find
call
chapter
statistics
package
cloud
analysis
rdatamining
fast
outlier
knowledge
list
talk
notescard
map
graph
open
postdoctoral
mining
high
graphics
thanks
network
job
sydney
database
conference
users
big
analytics
join
topic
processing
cfp
text
slides
used
published
snowfall
large
applications
case
see
available
performance
group
now
poll
draft
reference
course
code
can
visualizing
industrial
center
due
introduction
via
series
tenuretrack
information
association
clustering
access
sentiment
videos
techniques
tried
page
distributed
top
presentation
science
youtube
wwwrdataminingcom
classification
southern
canberra
added
china
experience
management

Clustering words: Hierarchical Clustering (HC) R Data Mining Book: Chapter 10, Page 109

\# remove sparse terms

- myTdm2 <- removeSparseTerms(myTdm, sparse=0.95)
- > m2 <- as.matrix(myTdm2)

\# cluster terms

- > distMatrix <- dist(scale(m2))
- > fit <- hclust(distMatrix, method="ward.D")
- > plot(fit)

\# cut tree into 10 clusters

- rect.hclust(fit, k=10)
- (groups <- cutree(fit, k=10))

Clustering tweets: K-means clustering R Data Mining Book: Chapter 10, page 111

transpose the matrix to cluster documents (tweets)

- m3 <- t(m2)

set a fixed random seed

- set.seed(122)

k-means clustering of tweets

- k <- 8
- > kmeansResult <- kmeans(m3, k)

cluster centers

- > round(kmeansResult$centers, digits=3)

Function to show top three words in every cluster of tweets: Check with topic modelling!

```
for (i in 1:k) {
+ cat(paste("cluster ", i, ": ", sep=""))
+ s <- sort(kmeansResult$centers[i,], decreasing=T)
+ cat(names(s)[1:3], "\n")
+ # print the tweets of every cluster
+ # print(rdmTweets[which(kmeansResult$cluster==i)])
+ }
```

Topic Modelling: “topicmodels” package

- library(topicmodels)
- set.seed(123)
- myLda <- LDA(as.DocumentTermMatrix(myTdm), k=8)   #8 topics
- terms(myLda, 5)   #Five terms in each topic (can be changed)

Note: LDA = Latent Dirichlet Allocation: NLP->ML->AI (Self-learning)

- **## Topic 1 Topic 2 Topic 3 Topic 4**
- ## [1,] "mining" "data" "r" "position"
- ## [2,] "data" "free" "examples" "research"
- ## [3,] "analysis" "course" "code" "university"
- ## [4,] "network" "online" "book" "data"
- ## [5,] "social" "ausdm" "mining" "postdoctoral"
- **## Topic 5 Topic 6 Topic 7 Topic 8**
- ## [1,] "data" "data" "r" "r"
- ## [2,] "r" "scientist" "package" "data"
- ## [3,] "mining" "research" "computing" "clustering"
- ## [4,] "applications" "r" "slides" "mining"
- ## [5,] "series" "package" "parallel" "detection"

Question/Queries?

Thank you!

@shitalbhandary