Statistical Computing with R
Masters in Data Science 503 (S11)
Fourth Batch, SMS, TU, 2025

Shital Bhandary
Associate Professor
Statistics/Bio-statistics, Demography and Public Health Informatics
Patan Academy of Health Sciences, Lalitpur, Nepal
Faculty, Masters in Medical Research, NHRC/Kathmandu University
Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Review Preview (Unit 2, Session 5)

- **Data Mining**
  - **What did you find in this reading provided to you?**
  - **https://aws.amazon.com/what-is/data-mining/**

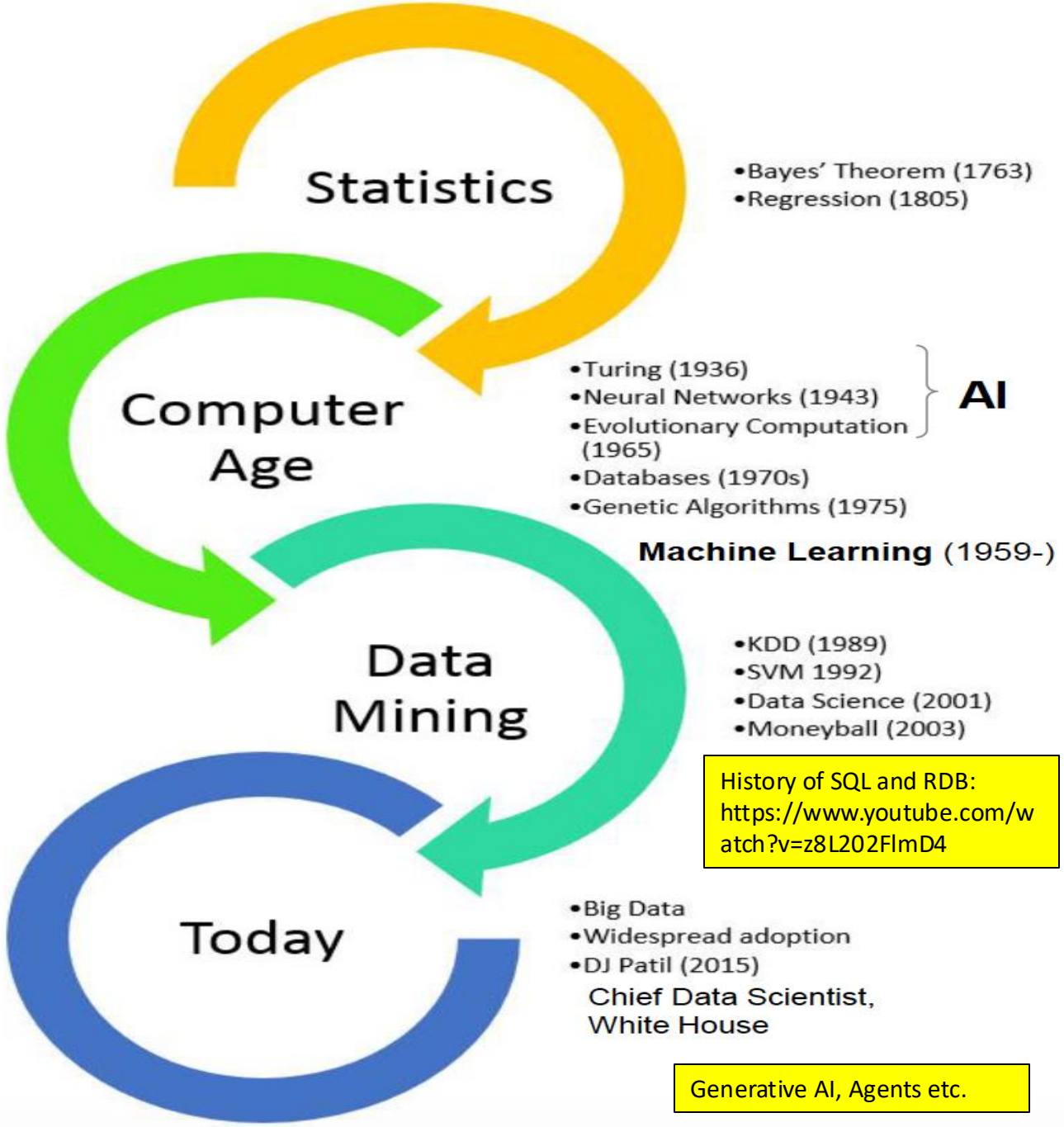
- **Text Mining**
  - **What did you find in this reading provided to you?**
  - **https://rpubs.com/vipero7/introduction-to-text-mining-with-r**

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.

- **Alan Mathison Turing** (/ˈtjʊərɪŋ/; 23 June 1912 – 7 June 1954) was an English mathematician, computer scientist, logician, cryptanalyst, philosopher and theoretical biologist.
- He was highly influential in the development of theoretical computer science, providing a formalisation of the concepts of algorithm and computation with the Turing machine, which can be considered a model of a general-purpose computer.
- Turing is widely considered to be the father of theoretical computer science. (Wikipedia)
- The ACM A.M. Turing Award, "Nobel Prize of Computing"!



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

Data Mining (What is):

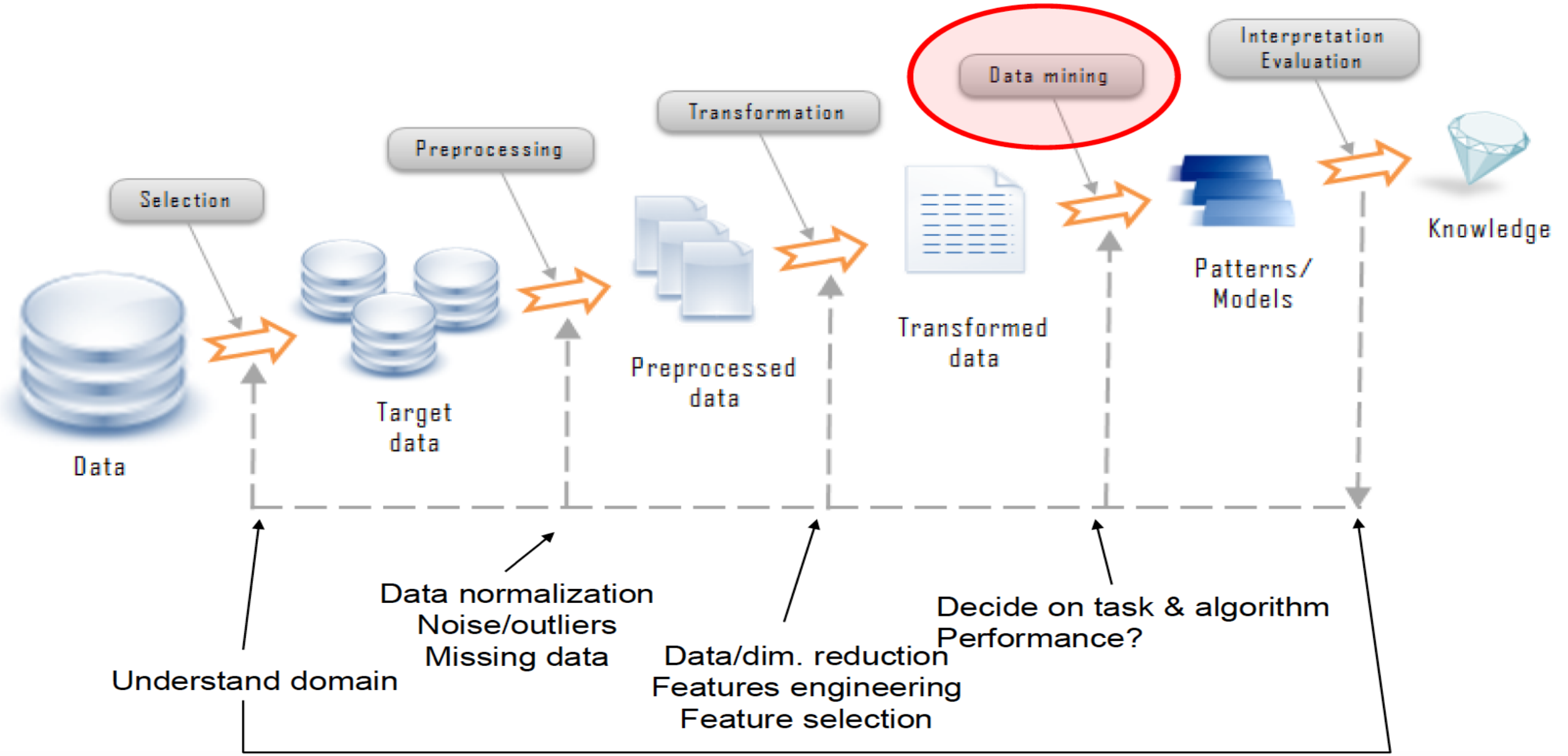
- Data Mining refers to a set of methods applicable to large and complex databases to eliminate the randomness and discover the hidden pattern. (https://online.stat.psu.edu/stat857/node/142/)

- 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.

(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

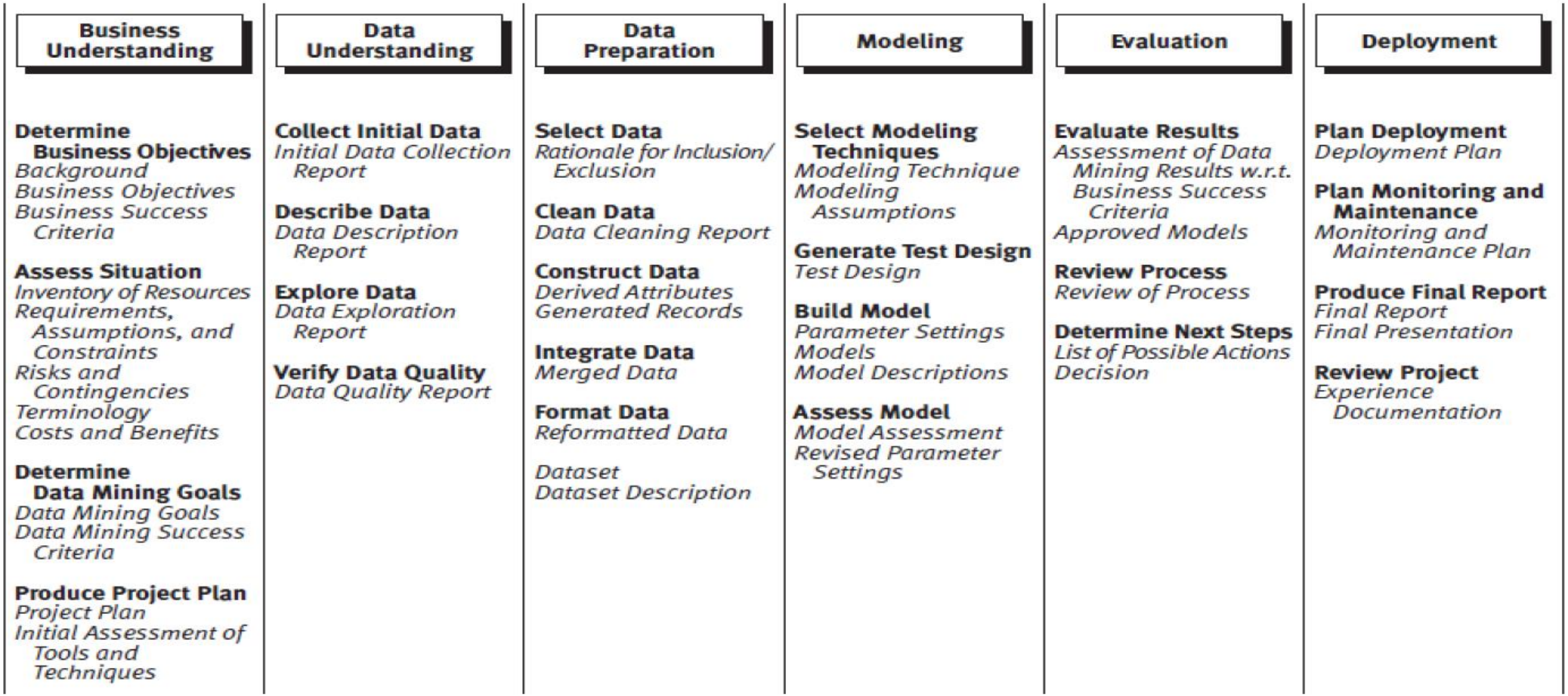
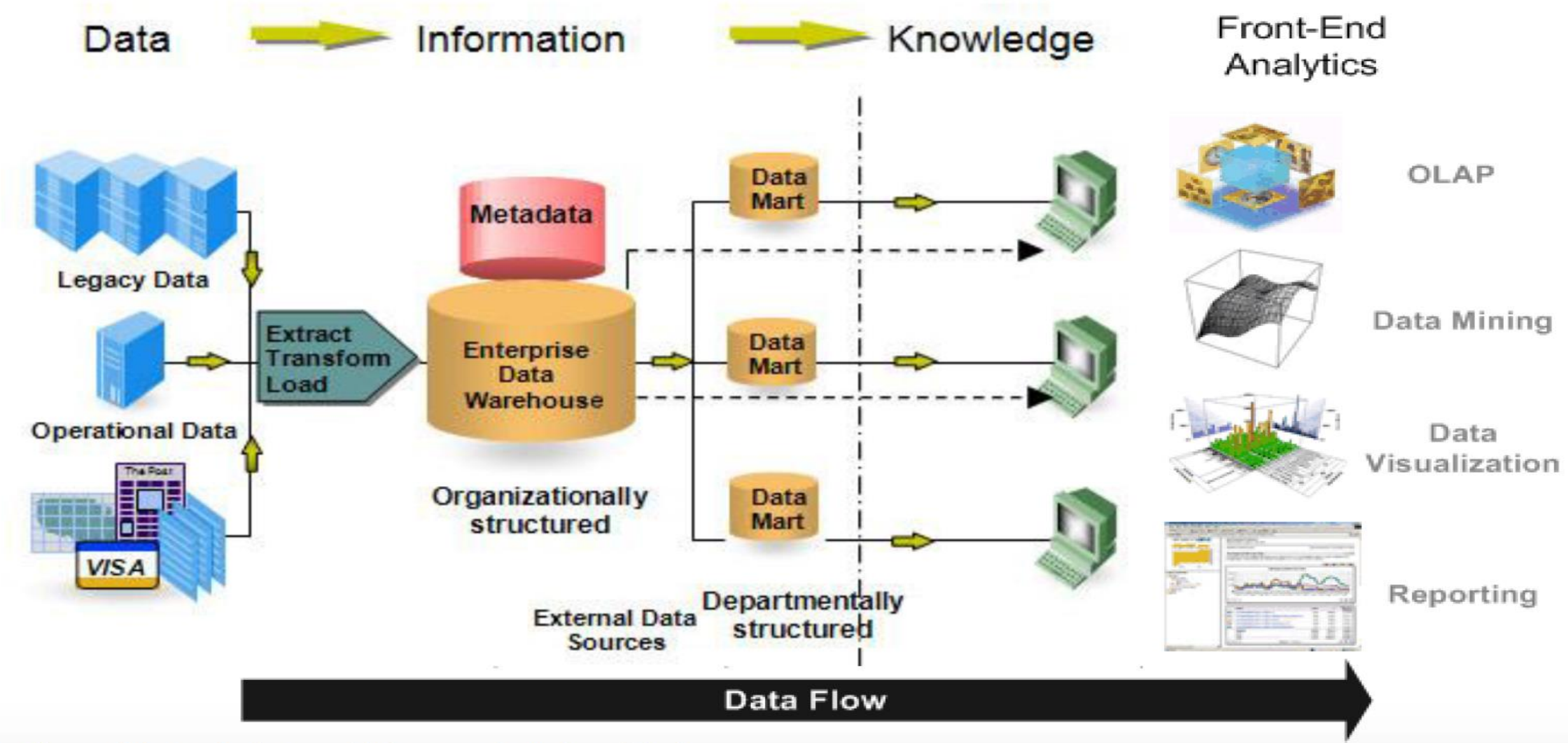


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Data Warehouse



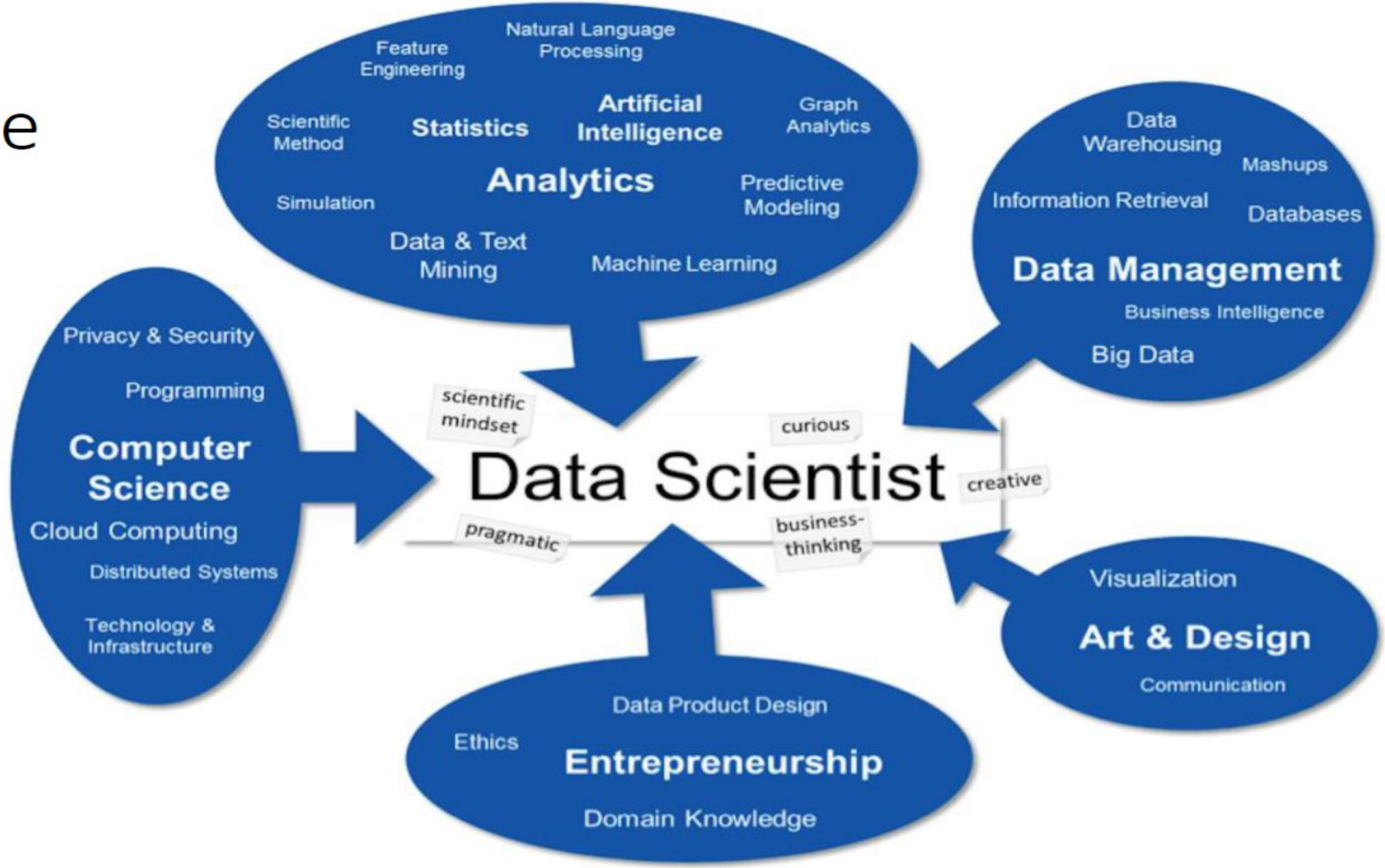
http://www.fulcrumlogic.com/data_warehousing.shtml

What is?

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

- Data:
  - Old, obsolete but retained data
  - Highly volatile, real time analysis
- Extract, transform, load
- Information:
  - Directory
  - Integrated, static data, analytics
  - Simple form of data warehouse
- Knowledge
  - Online analytical processing
  - Descriptive, predictive, prescriptive

Data Science

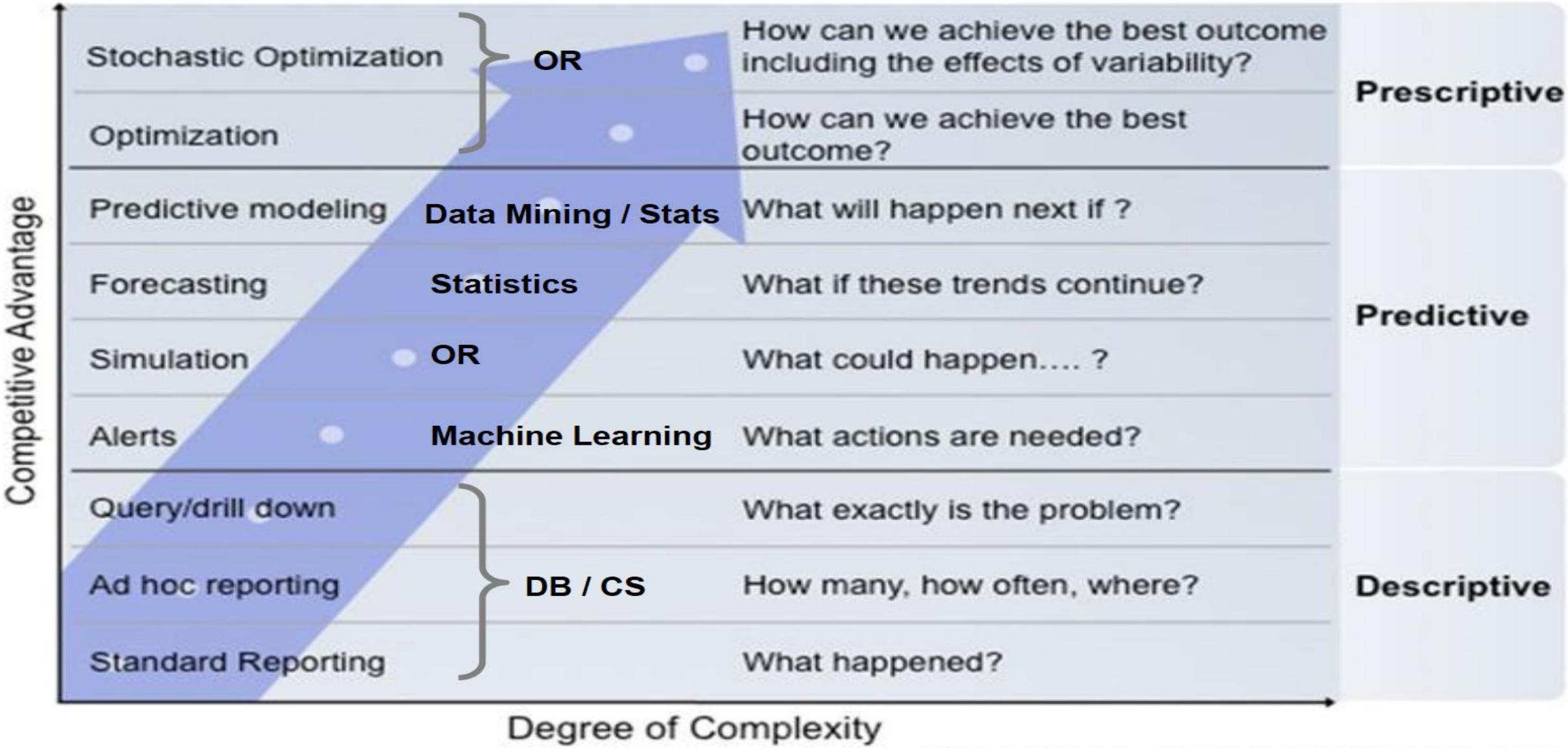


Source: T. Stadelmann, et al., Applied Data Science in Europe

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.)

Data Mining & Analytics



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

Data Mining Tasks:

- **Text Mining – Unit 2**
- **Exploratory/Descriptive Data Analysis – Unit 3**
- **Predictive Modelling (Regression and classification) – Unit 4**
- **Dimensionality Reduction, Cluster Analysis – Unit 5**
- **Association Analysis – Unit 5**

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) **aka text parameters**
- Find frequent terms and associations of key term with other terms
- Use network graph/word cloud to visualize the DTM
- Perform “topic modelling” and compare it with network graph result!
- Perform cluster analysis to find clusters of similar words

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

Option 1: retrieve tweets from Twitter (**must have API keys!**)

- 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 = "./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 = "./rdmTweets.RData") #Option 3 used, data in working folder!
- tweets <- rdmTweets #Assign tweets as rdmTweets
- str(tweets)

- (n.tweet <- length(tweets)) #If rmdTweets is assigned as “tweets”
- [1] 154 #Option 3 used, 154 tweets only!

- strwrap(tweets[[154]]$text, width = 55) #Text variable of tweet 154
- [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."

Checking the content of the last tweet

#With string wrap and line break at 55th and 62nd positions

- strwrap(tweets[[154]]$text, width = 55)
- strwrap(tweets[[154]]$text, width = 62)

#What happens if a single square bracket is used?

strwrap(tweets[154]$text, width = 55)

Output: ?

Why?

Checking the content of first three tweets

- **tweets[1:3]**
- [[1]]
- [1] "RDataMining: Postdoc/Research Scientist Position on Big Data at MIT http://t.co/hZ1ojAW2"

- [[2]]
- [1] "RDataMining: Research scientist position for privacy-preserving data publishing, Singapore http://t.co/GPA0TyG5"

- [[3]]
- [1] "RDataMining: Easier Parallel Computing in R with snowfall and sfCluster http://t.co/BPcinvzK"

Text cleaning in R: **Pre-processing I** (tweets to data frame and text corpus formation)

- **library(twitteR)**

convert tweets to a data frame

- df <- twListToDF(tweets)
- str(df)

library(tm)

build a corpus

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

#Inspect first 3 elements

inspect(myCorpus[1:3])

Text cleaning in R: **Pre-processing I** (Corpus to lower case, remove punctuation/numbers)

convert to lower case

- myCorpus <- tm_map(myCorpus, tolower)
- inspect(myCorpus[1:3])

remove punctuations and numbers

- myCorpus <- tm_map(myCorpus, removePunctuation)
- inspect(myCorpus[1:3])
- myCorpus <- tm_map(myCorpus, removeNumbers)
- inspect(myCorpus[1:3])

Text cleaning in R: **Pre-processing II** (Remove URL)

remove URLs, http followed by non-space characters

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

- myCorpus <- tm_map(myCorpus, removeURL)

- inspect(myCorpus[1:3])

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

remove r and big from the list of stopwords

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

remove stopwords

- myCorpus <- tm_map(myCorpus, removeWords, myStopwords)
- inspect(myCorpus[1:3])

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)
- inspect(myCorpus[1:3])

We must use SnowballC package for proper stemming if this does not work!

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

replace ”posit" with ”position", because ”position" was first stemmed to ”posit" and then completed to ”posit"

- myCorpus <- tm_map(myCorpus, gsub, pattern=”posit", replacement=”position")
- strwrap(myCorpus[154], width=55) #check the corpus again (iteratively)!

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

[2] "functions packag data mining applic"

Create Term Document Matrix and Check “Frequent terms”:

#Create Term Document Matrix and check its structure via tm package

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

inspect frequent words on TDM of non-stemmed myCorpus

(freq.terms <- findFreqTerms(myTdm, lowfreq=10))

- [1] "data" "research" "r" "package" "tutorial"
- [6] "using" "slides" "mining" "analysis" "network"
- [11] "social" "introduction" "examples"

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

- # which words are associated with r?

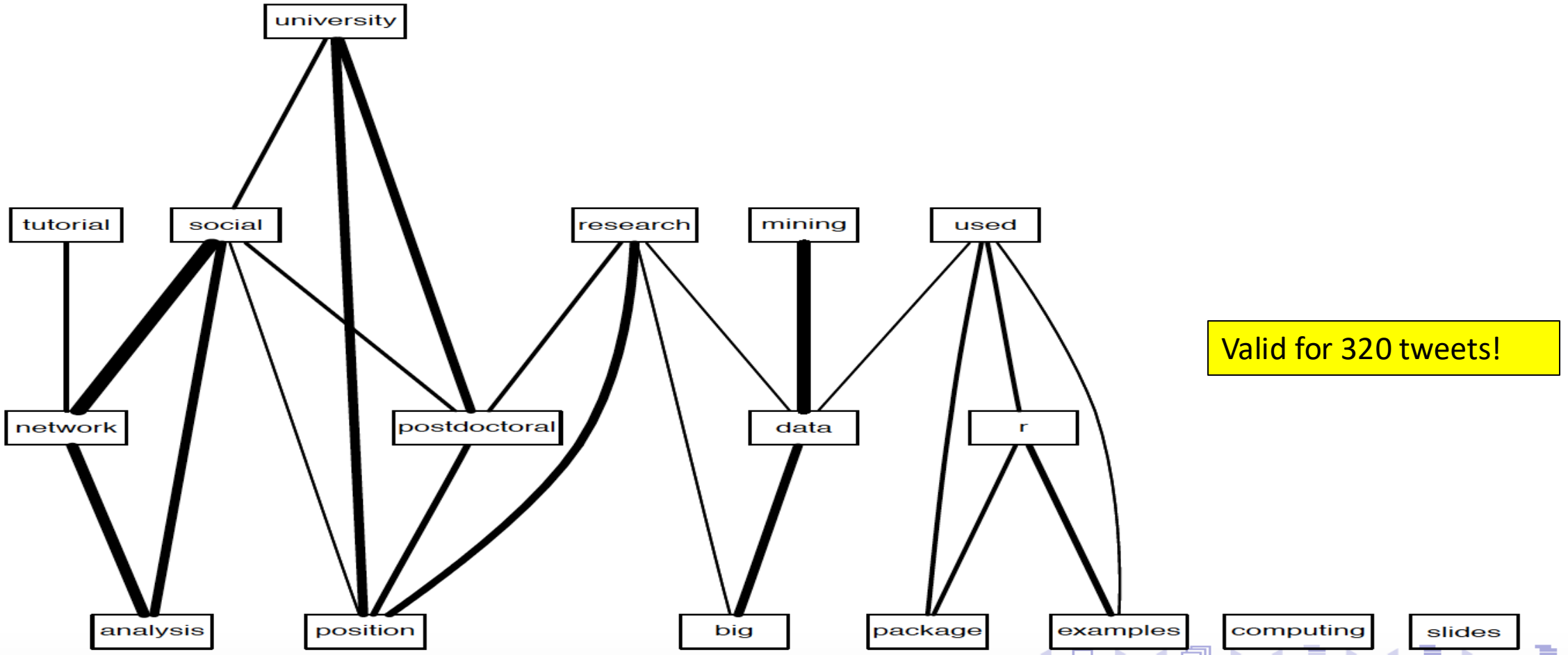
```
findAssocs(myTdm, "r", 0.2)
```

```
$r
 packages     users      many  canberra      cran      card         functions reference
     0.35      0.30      0.26      0.26      0.26      0.24              0.24      0.24
      see  examples
     0.24      0.23
```

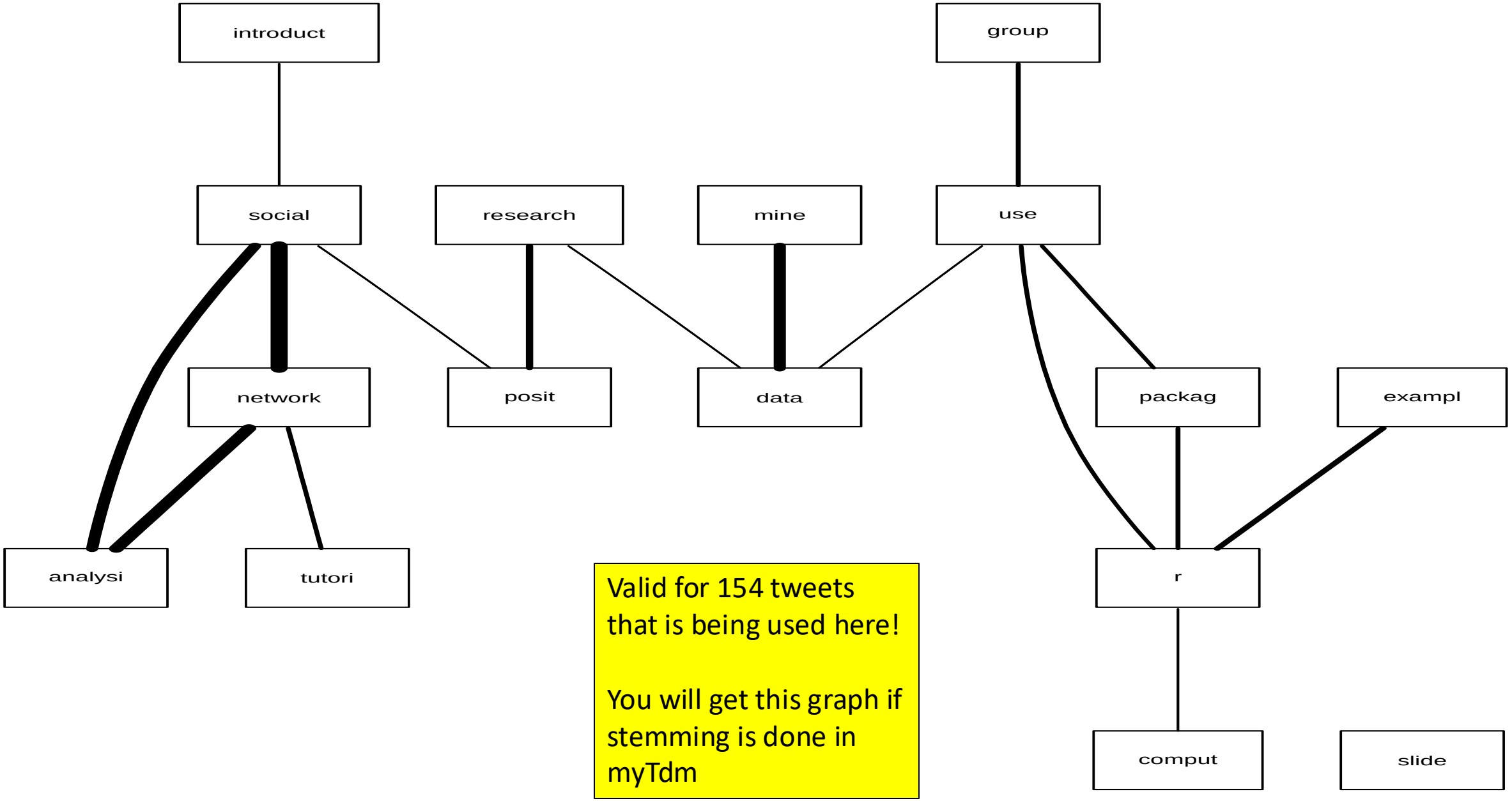
Network of Terms

install.packages("BiocManager")
BiocManager::install("Rgraphviz")

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

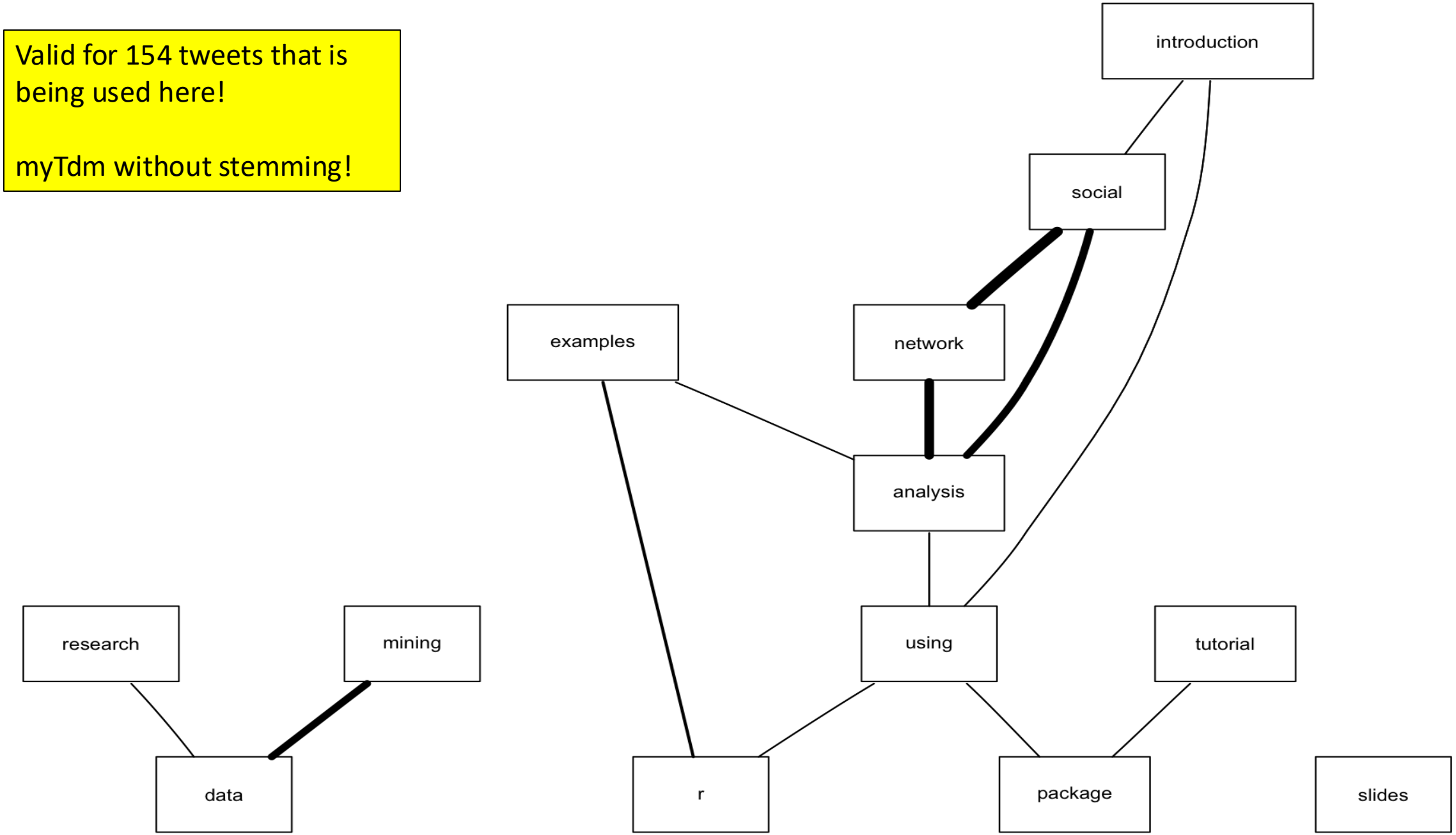


Valid for 320 tweets!


introduct
group
social
research
mine
use
network
posit
data
packag
exampl
analysi
tutori
r
comput
slide
Valid for 154 tweets that is being used here!
You will get this graph if stemming is done in myTdm

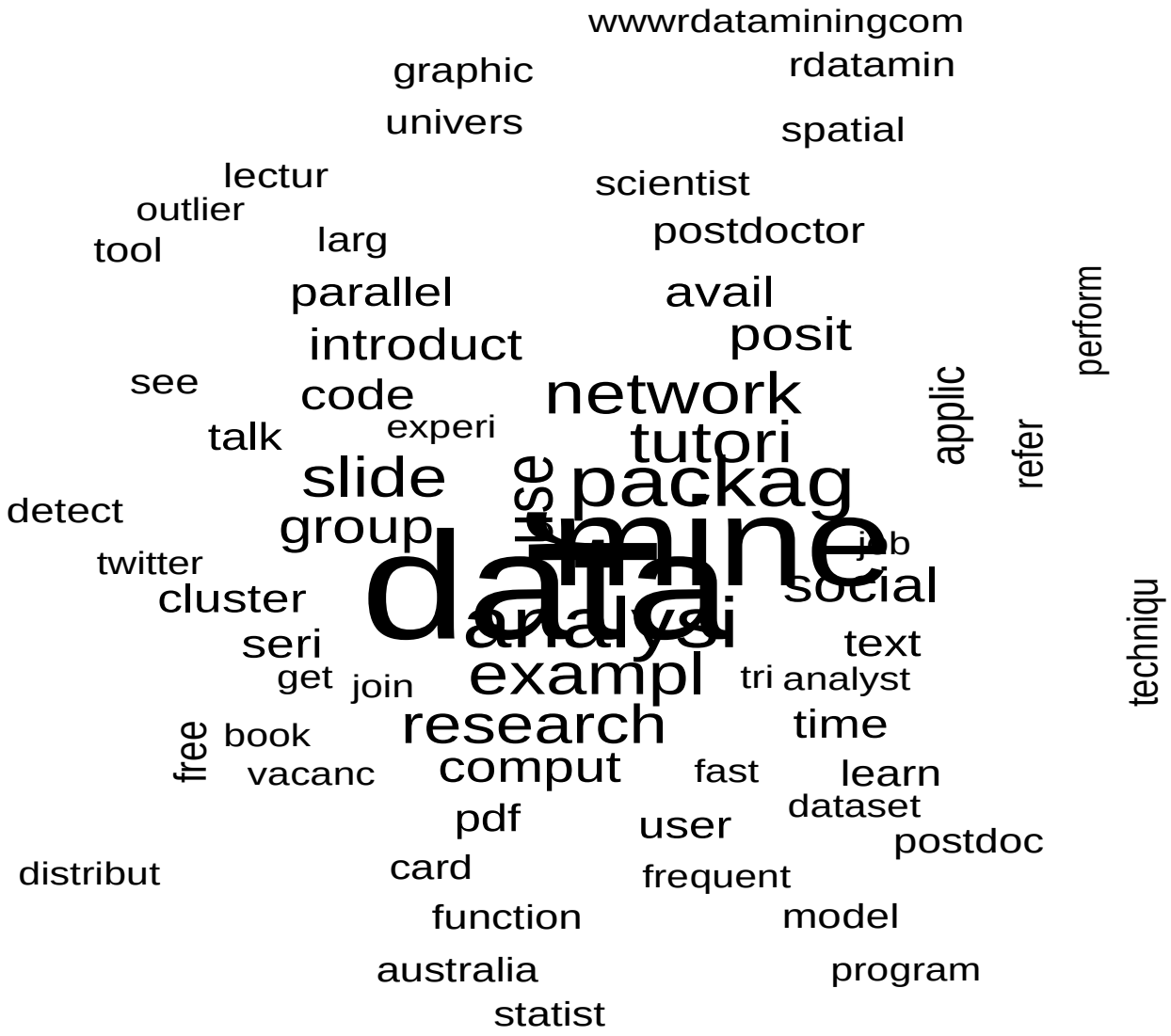
Valid for 154 tweets that is being used here!

myTdm without stemming!



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)



Try it with minimum frequency of 1 and 10

If you want colorful word cloud then use the RColorBrewer package!

More here:

https://www.r-bloggers.com/2011/07/word-cloud-in-r/

Topic Modelling: “topicmodels” package

- library(topicmodels)
- set.seed(123)
- myLda <- LDA(as.DocumentTermMatrix(myTdm), k=5) #5 topics
- terms(myLda, 3) #Three 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	Topic 5
[1,]	"data"	"r"	"analysi"	"r"	"r"
[2,]	"mine"	"exampl"	"network"	"packag"	"data"
[3,]	"r"	"code"	"social"	"comput"	"mine”

- Compare this result with the Rgraphviz results obtained above
- Are you happy?
- Do you want to change LDA parameters?
- Change LDA parameters!

Refined model: Four topics with 3 terms each (Cluster analysis or Thematic Analysis?)

- set.seed(123)
- > myLda <- LDA(as.DocumentTermMatrix(myTdm), k=4)#5 topics
- > terms(myLda, 3) #Three terms in each topic (can be changed)

	Topic 1	Topic 2	Topic 3	Topic 4
[1,]	"data"	"data"	"r"	"r"
[2,]	"mine"	"research"	"tutori"	"group"
[3,]	"r"	"r"	"exampl"	"data"

Are you happy with these four topics?

Did it miss something important?

Question/Queries?

Project 2, Part 1: Replicate outputs from other software using base R or dplyr package

- You will be provided with excel and SPSS datafiles and sample table outputs
- You need to replicate them in R with Scripts using R Studio
- You need to submit your script file, markdown file and output knitted as PDF file

Project 2, Part 2: Text mining with Ten pdf files of "Text Mining" journal articles!

- You must search and download first 10 free pdf files on this topic using **Google Scholar (https://scholar.google.com/)** (DO NOT USE GOOGLE.COM)
- You must put all the 10 pdf files in a folder called "MDS503P2" i.e. setwd()
- Use the "pdftools" package to read these ten pdf files in R from MDS503P2
- Once you read them in R, create a "corpus" and perform text mining
- Submit the R Script file, R markdown file and knitted pdf report file of the Project work (Project 2:Unit 2) in Google classroom

Thank you!

@shitalbhandary