Welcome to CS4437 / CS9637 Introduction to Data Science

Dr. Dan Lizotte (Comp. Sci., Epidemiology & Biostatistics)


Statistic
Visualisation
Pattern recognit.
DATA
Data Mining
Machine Learning
Science
Neurocomputing
Database & Data procss

“A data scientist is a statistician who lives in San Francisco”

“Data Science is Statistics on a Mac”

“A data scientist is better at statistics than any software engineer and better at software engineering than any statistician.”

Statistic

Visualisation

Pattern recognit.

DATA

Science

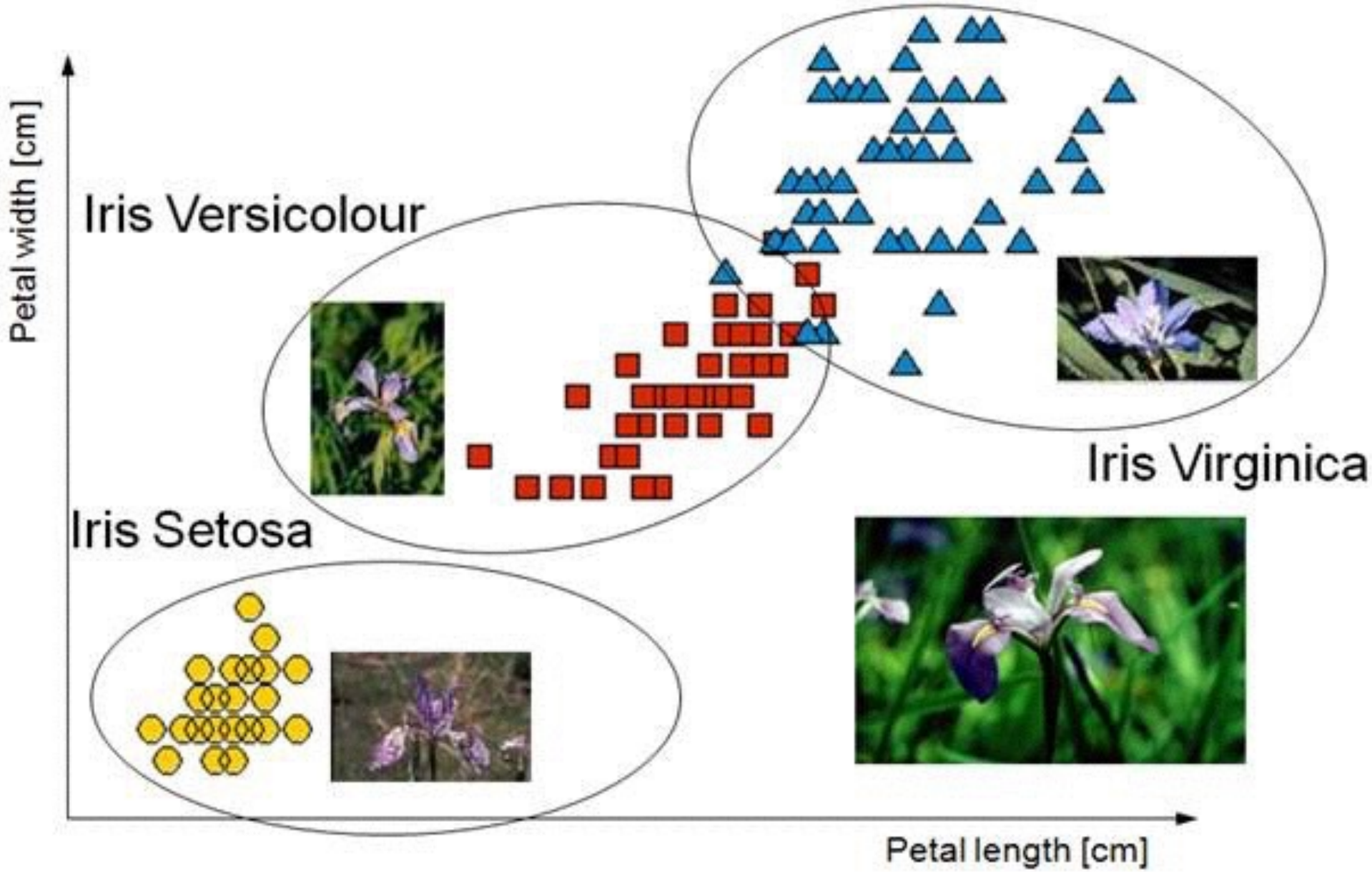
Data Mining

Machine Learning

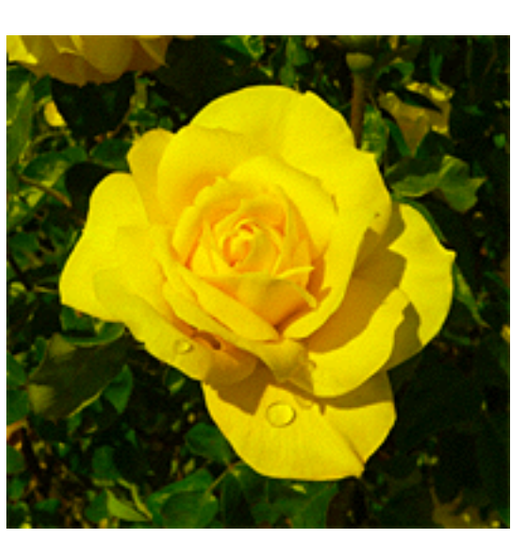
Neurocomputing

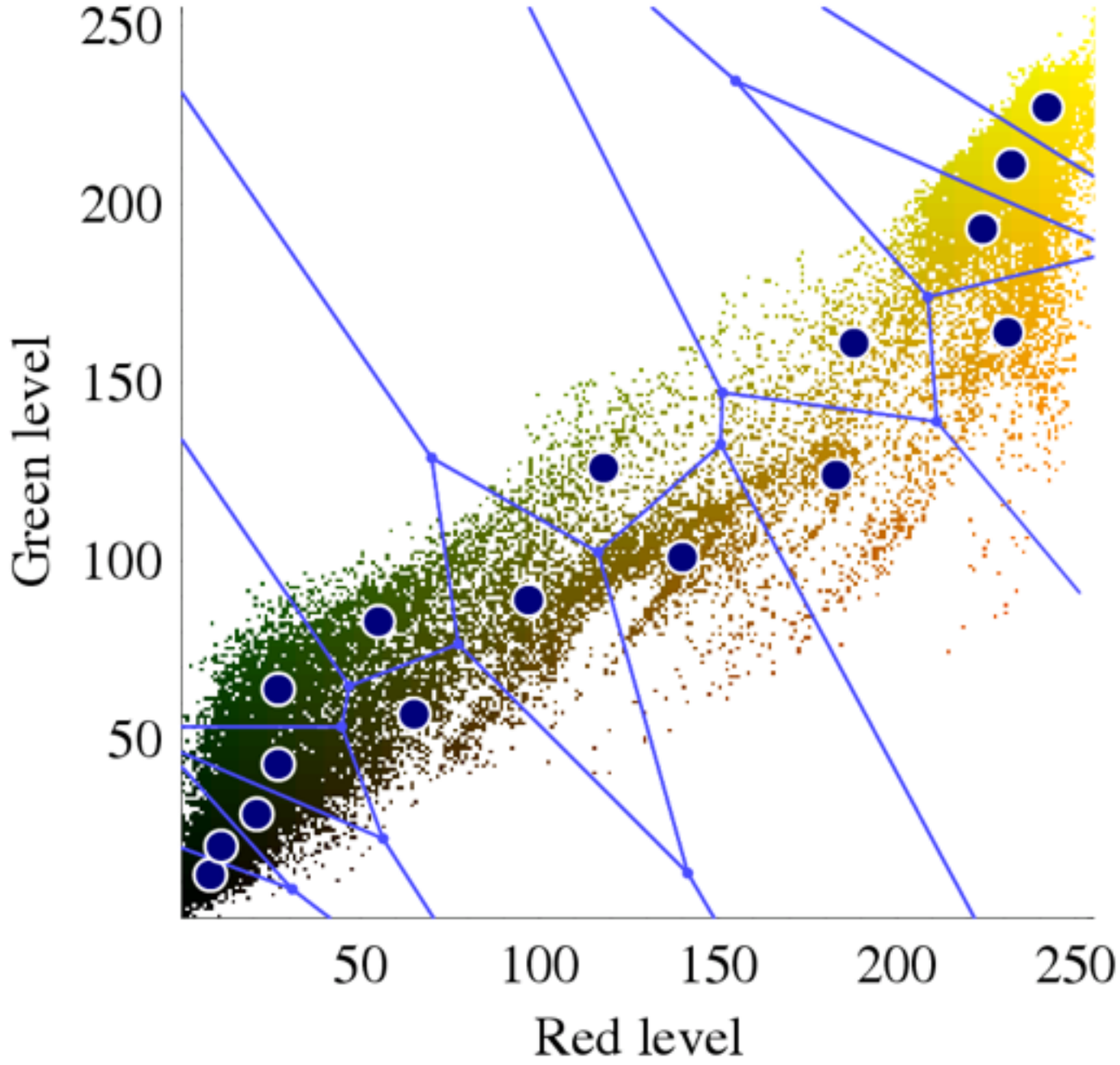
Database & Data procss

Supervised Learning



Unsupervised Learning





Reinforcement Learning

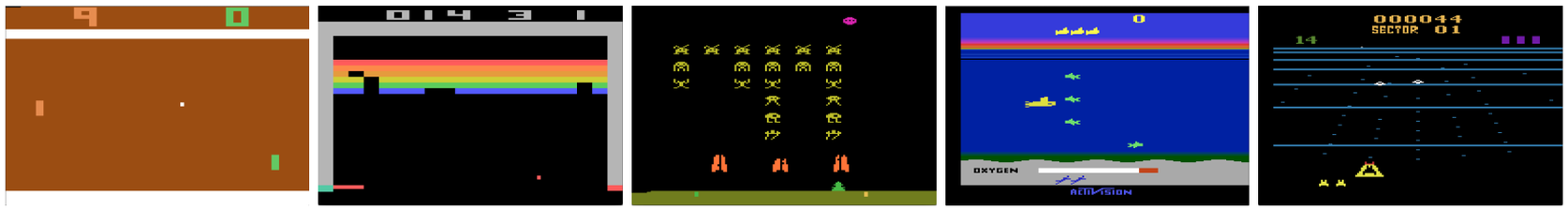
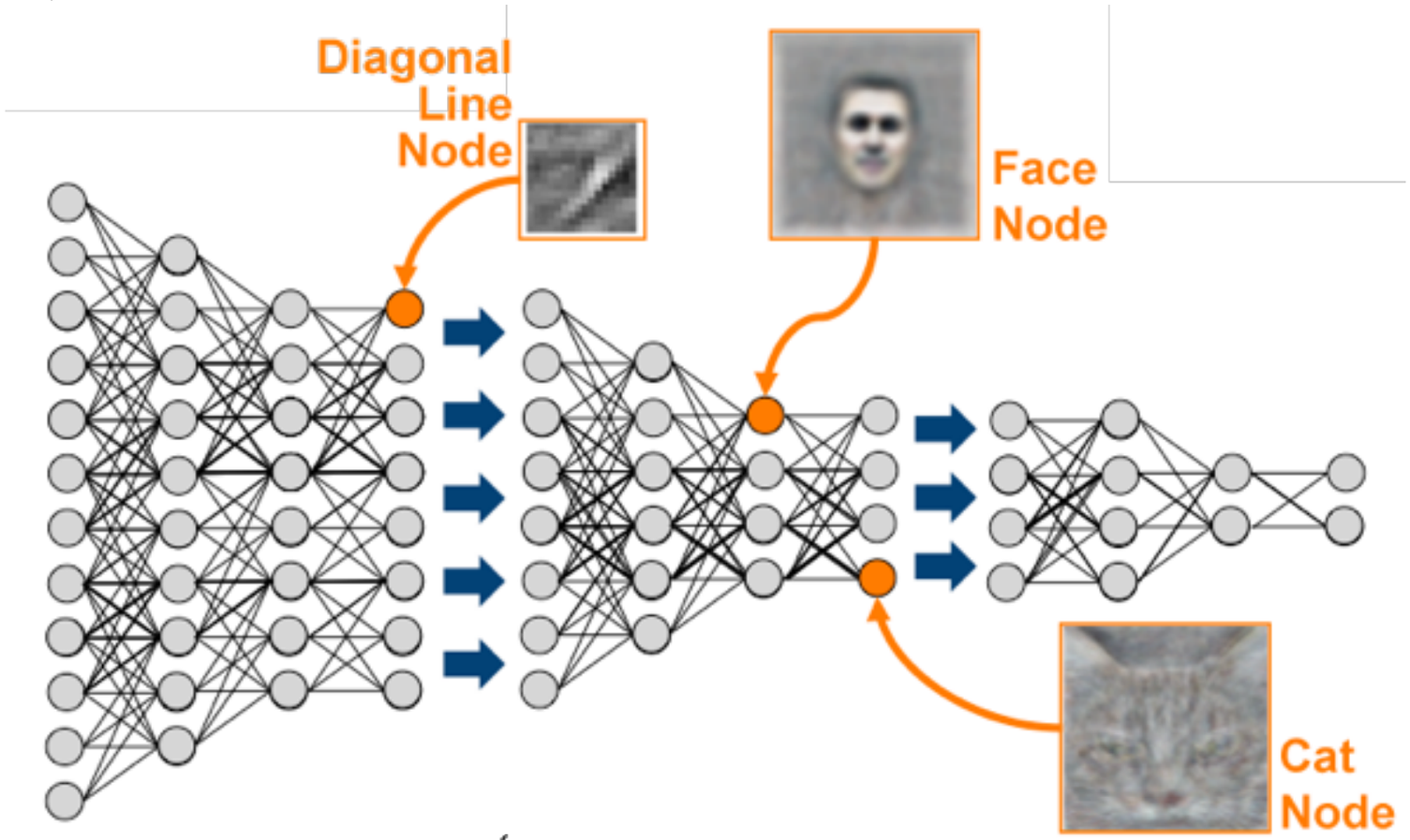
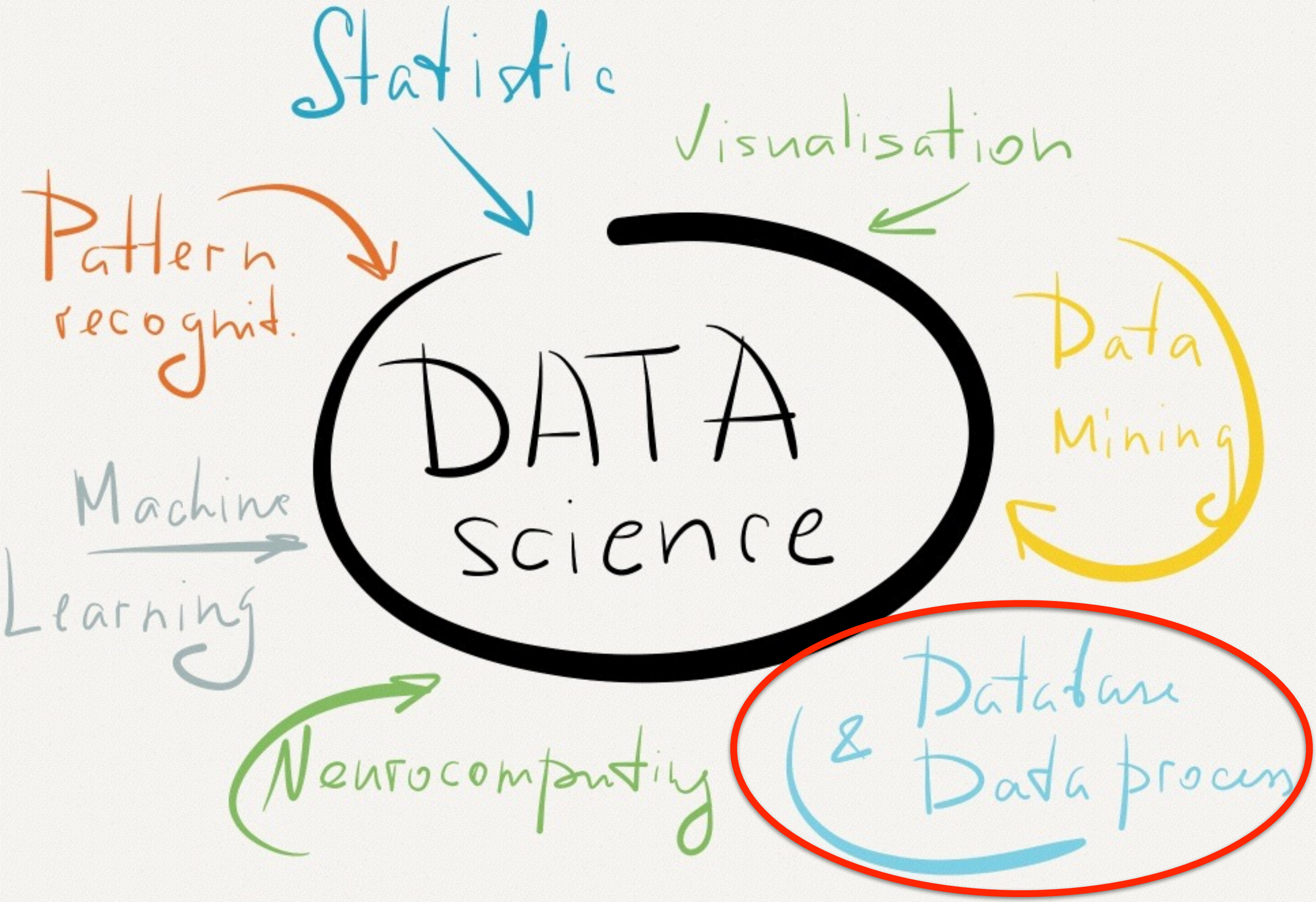


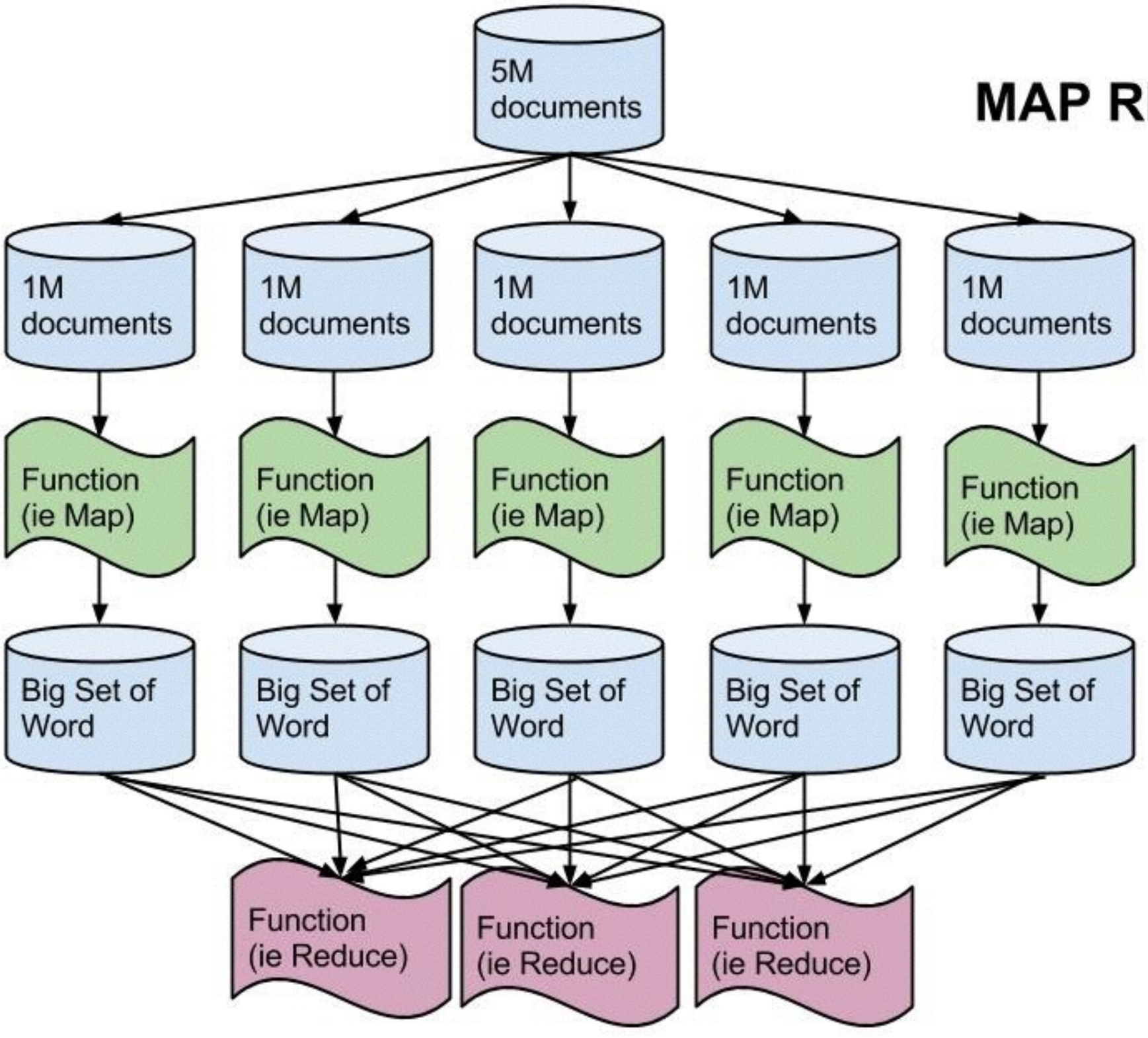
Figure 1: Screen shots from five Atari 2600 Games: (*Left-to-right*) Pong, Breakout, Space Invaders, Seaquest, Beam Rider



$$Q_{t+1}(s_t, a_t) = \underbrace{Q_t(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha_t(s_t, a_t)}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{R_{t+1}}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \underbrace{\max_a Q_t(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} - \underbrace{Q_t(s_t, a_t)}_{\text{old value}} \right)$$


Statistic
Visualisation
Pattern recognit.
DATA
Science
Data Mining
Machine Learning
Neurocomputing
Database & Data procesn

MAP REDUCE

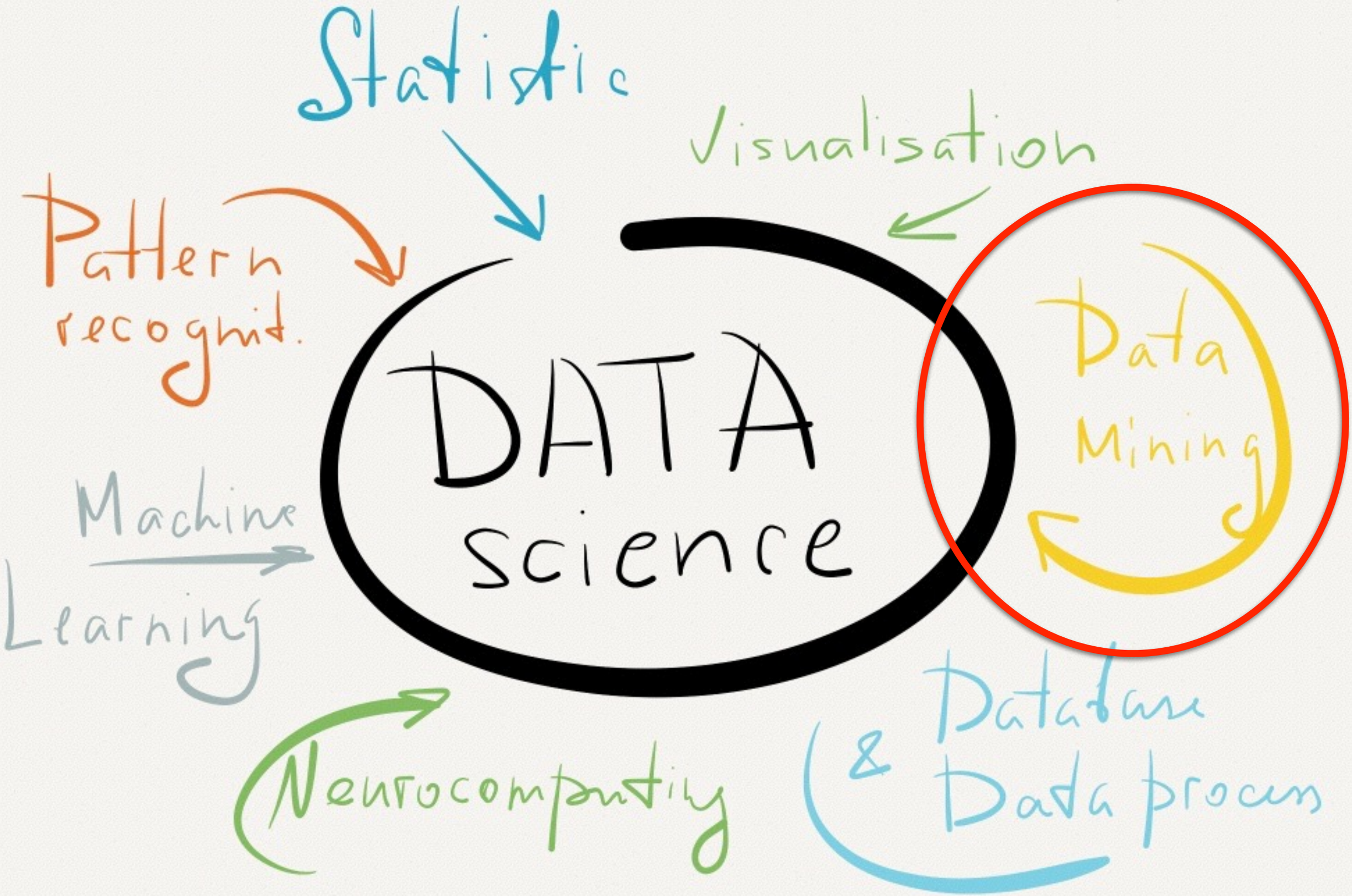


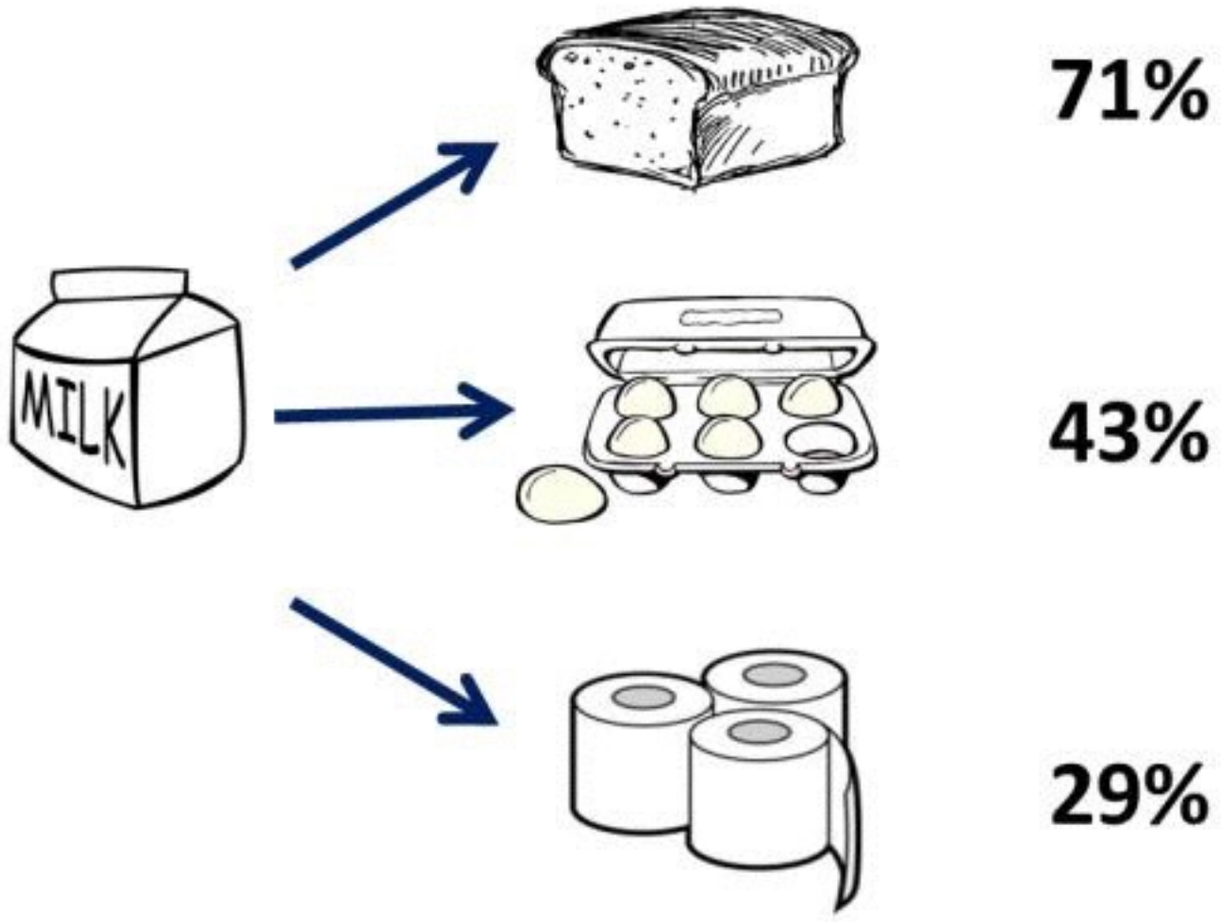
Distribute the documents across N computers

For each document, return a set of (word, frequency) pair

We get a big distributed list of sets of words frequency

Each Reduce function count the occurrences of one word.

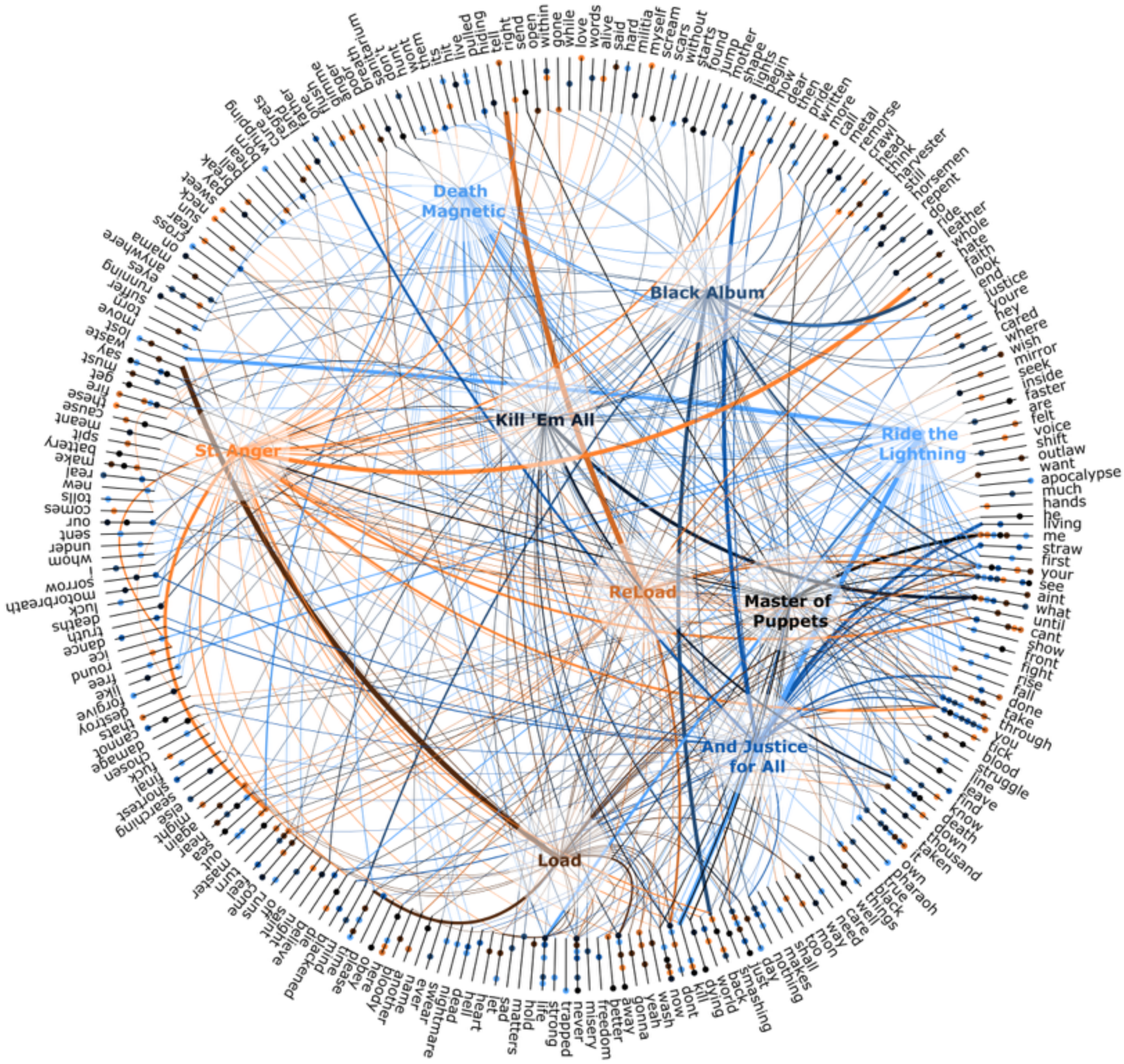

Statistic
Visualisation
Pattern recognit.
DATA
Science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss



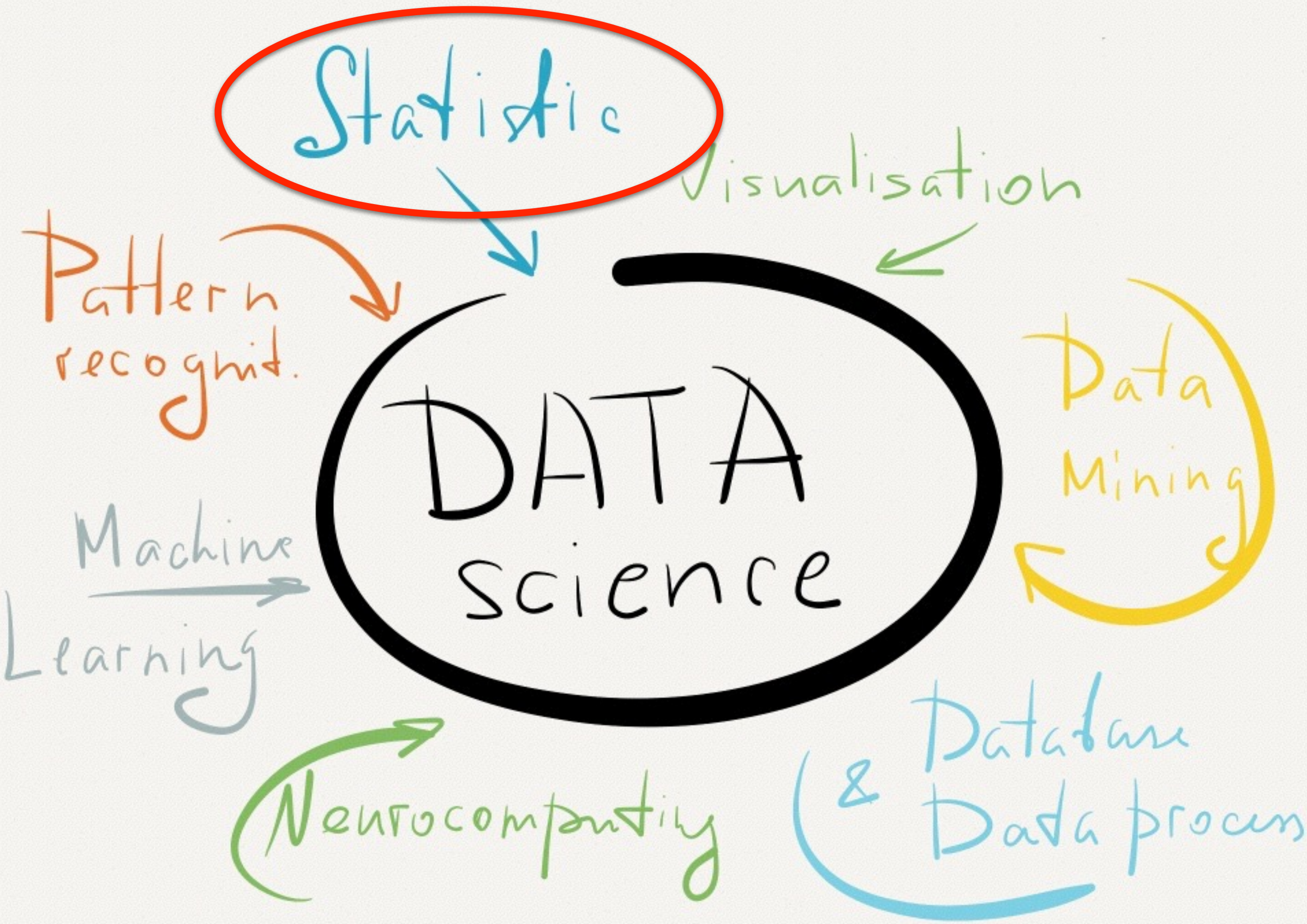
Of transactions that included milk:

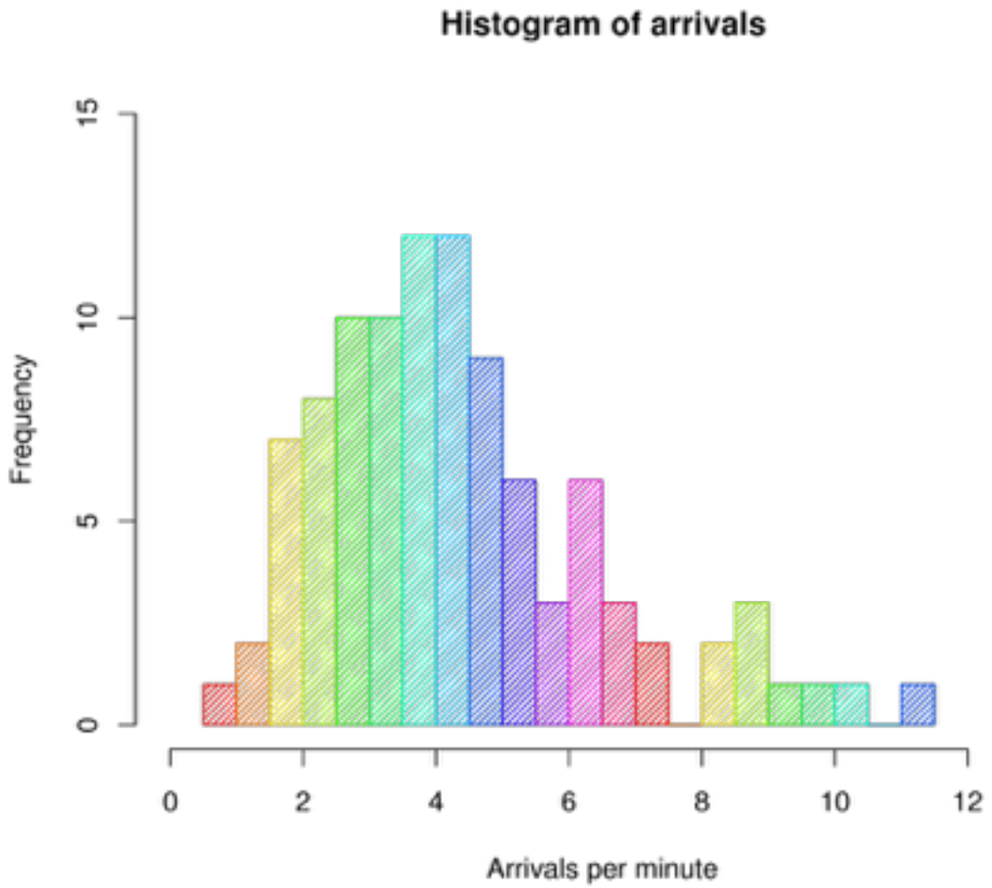
- 71% included bread
- 43% included eggs
- 29% included toilet paper

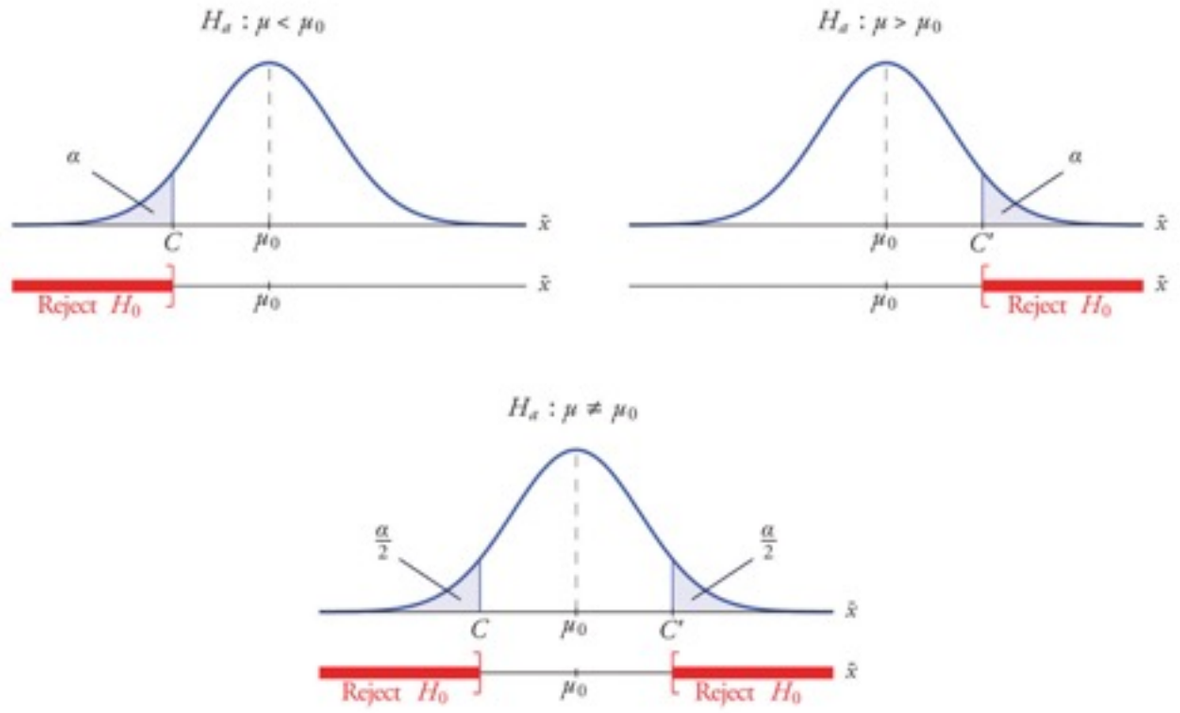

Statistics
Visualisation
Pattern recognit.
DATA
Data Mining
Machine Learning
Science
Neurocomputing
Database & Data procss

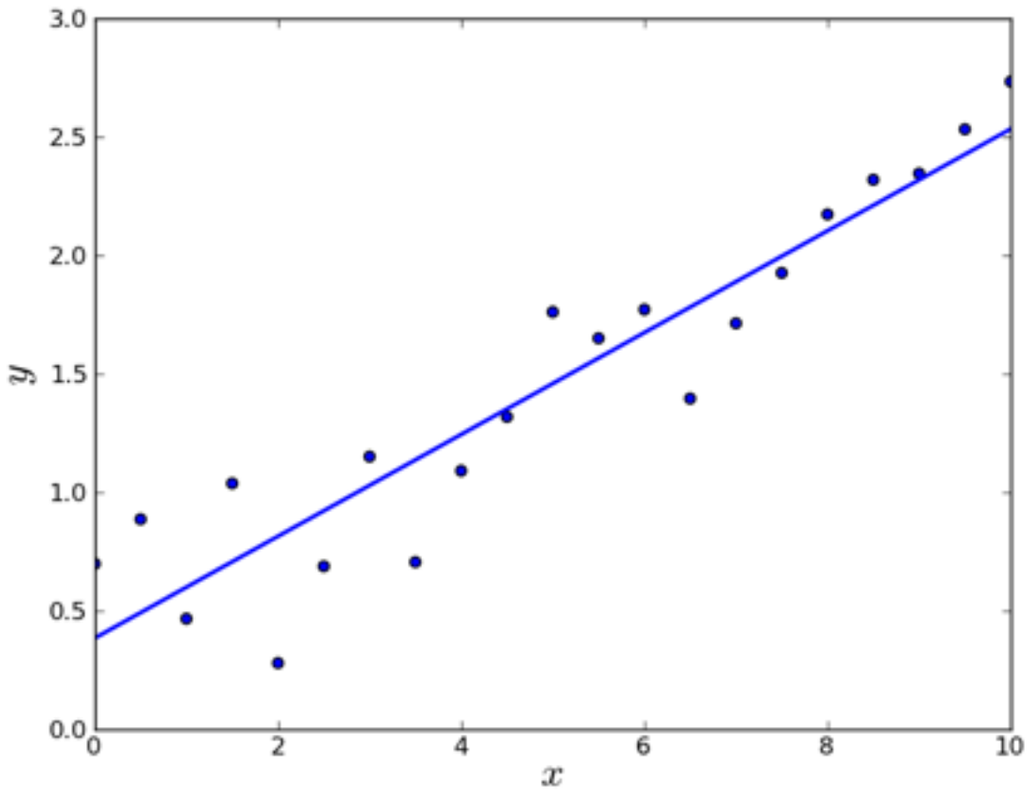


Death Magnetic
Black Album
Kill 'Em All
St. Anger
Ride the Lightning
ReLoad
Master of Puppets
And Justice for All
Load
within
gone
while
love
words
alive
said
hard
militia
myself
scream
scars
without
starts
found
jump
mother
shape
lights
begin
how
dear
then
pride
written
more
call
metal
remorse
crawl
head
think
harvester
still
horsemen
repent
do
ride
leather
whole
hate
faith
look
end
justice
youre
hey
cared
where
wish
mirror
seek
inside
faster
are
felt
voice
shift
outlaw
want
apocalypse
much
hands
he
living
me
straw
first
your
see
aint
what
until
cant
show
front
fight
rise
fall
done
take
through
you
tick
blood
struggle
line
leave
find
know
death
down
thousand
taken
it
own
pharaoh
true
track
things
well
care
need
way
mon
too
shall
makes
nothing
smashing
just
day
back
world
dying
kill
dont
now
wash
yeah
gonna
away
better
freedom
misery
never
trapped
strong
life
hold
matters
sad
let
heart
hell
dead
nightmare
swear
ever
name
another
bloody
here
obey
please
time
mind
blackened
die
believe
night
sight
off
runs
come
feel
turn
master
out
sea
hear
again
might
else
searching
shortest
final
fuck
chosen
damage
cannot
thats
destroy
forgive
like
free
found
ice
dance
truth
deaths
luck
motorbreath
sorrow
i
whom
under
sent
our
comes
tolls
new
real
make
battery
spit
meant
cause
these
fire
get
must
say
waste
lost
move
torn
suffer
running
eyes
anywhere
mama
on
fear
cross
sun
neck
sweet
open
send
right
tell
riding
pulled
lived
hit
its
them
wont
hunt
dont
sanitarium
breath
ginger
glimmer
father

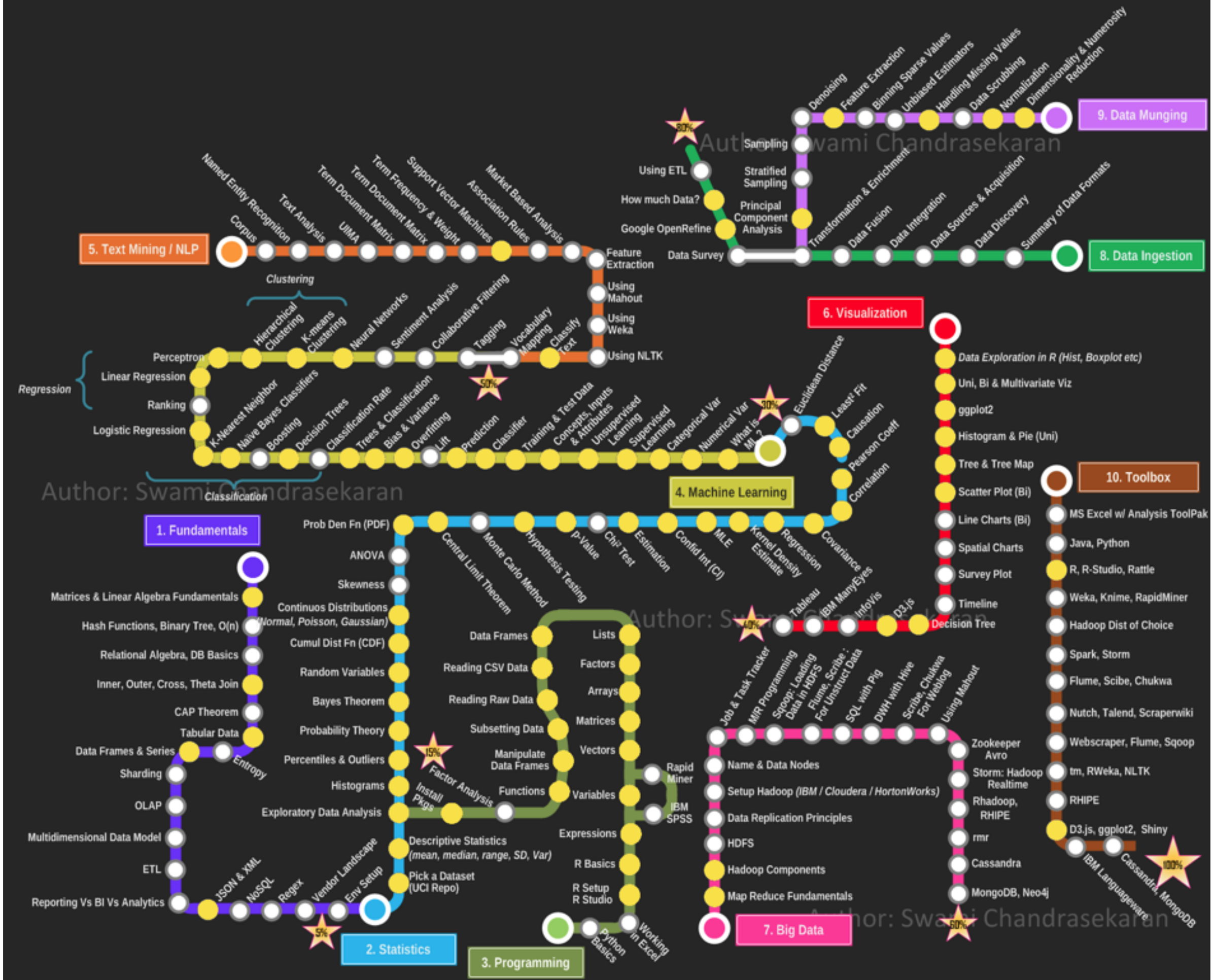

Statistic
Visualisation
Pattern recognit.
DATA
Data Mining
Machine Learning
Science
Neurocomputing
Database & Data procss


Histogram of arrivals
Frequency
0
5
10
15
0
2
4
6
8
10
12
Arrivals per minute


$H_a : \mu < \mu_0$
α
C
μ_0
$\bar{x}$
Reject H_0
$H_a : \mu > \mu_0$
α
C'
μ_0
$\bar{x}$
Reject H_0
$H_a : \mu \neq \mu_0$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
C
C'
μ_0
$\bar{x}$
Reject H_0
Reject H_0


3.0
2.5
2.0
1.5
1.0
0.5
0.0
y
0
2
4
6
8
10
x


Author: Swami Chandrasekaran
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Perceptron
Linear Regression
Ranking
Logistic Regression
Regression
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML ?
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
Classification
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
Tableau
IBM ManyEyes
InfoVis
D3.js
40%
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Sampling
Using ETL
How much Data?
Google OpenRefine
Data Survey
Stratified Sampling
Principal Component Analysis
80%
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%


Author: Swami Chandrasekaran
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
Sampling
Stratified Sampling
Principal Component Analysis
30%
Using ETL
How much Data?
Google OpenRefine
Data Survey
8. Data Ingestion
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
5. Text Mining / NLP
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Corpus
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML ?
30%
4. Machine Learning
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
Covariance
Regression
Kernel Density Estimate
MLE
Confid Int (CI)
Estimation
Chi² Test
p-Value
Hypothesis Testing
Monte Carlo Method
Central Limit Theorem
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
15%
Factor Analysis
Install Pkgs
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Rapid Miner
IBM SPSS
Python Basics
Working in Excel
3. Programming
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
7. Big Data

Course Objective

- Introduce students to data science (DS) techniques, with a focus on application to substantive (i.e. "applied") scientific problems.

- Through individual projects, students will gain experience in identifying which problems can be tackled by DS methods, and learn to identify which specific DS methods are applicable to a problem at hand.

- This course **requires students to show substantial initiative in investigating methods that are applicable for their project**. The lectures give an overview of important methods, but the **lecture content alone is not sufficient to produce a high quality course project**.

Logistics

- **READ. THE. WIKI.**
  http://www.csd.uwo.ca/~dlizotte/teaching/cs4437/

- Instructor: Dan Lizotte – dlizotte at uwo dot ca – MC363
  TA: Brent Davis — bdavis56 at uwo dot ca

- Time: Tuesday from 11:30AM – 1:30PM, and on Thursday from 3:30PM – 4:30PM

- Place: Talbot College TC342

- Communication: We will be using OWL for electronic communication.

- Question & Collaboration Hour: Middlesex College MC320, Thursday 4:30PM to 5:30pm

Materials

- **READ. THE. WIKI.**

Anticipated Topics and Schedule

- **READ. THE. WIKI.**

Evaluation

- Daily Quizzes – **5%**
- Midterm - **35%**
- Brainstorming Session – **5%**
- Project Proposal – *4437*: **15%** *9637*: **10%**
- Report Draft – **5%**
- Project Report – **35%**
- Peer Review – *9637 only*: **5%**

Daily Quizzes

- Very short quiz at the beginning of class covering the previous day's materials
- The final quiz will be on 2 Mar.
- The lowest quiz mark will be dropped.
- Quiz marks will only be excused for medical reasons.

Individual Project

- Project Proposal – 4437: 15% 9637: 10%
    - Document detailing the plan for the project. See **Project Guidelines on the wiki** for detailed requirements.
- Report Draft – 5%
    - The purpose of the draft is to allow the instructor to provide feedback on the quality of the writing and the direction of the project.
- Project Report – 35%
    - Each student will prepare a research paper detailing a substantive problem, the data available, the applicable DS methods, and empirical results obtained on the problem.

Brainstorming - 5%

- Each student will prepare a presentation explaining an **applied problem**, as well as some **potential data science methods** that could be applied to the problem.

- The presentation should be **no more than 10 minutes**.

- We will then **discuss the problem as a class**, along with possible approaches for solving the problem using ML methods.

- Student is expected to **be prepared to answer deep questions about the nature of their problem** to ensure that they receive high quality feedback from the brainstorming session.

- See **Project Guidelines on the Wiki** for detailed requirements.

Brainstorming

- You must pick a brainstorming slot.

1. Create account on the Wiki
2. Edit the schedule at the bottom of the main page, replacing “SlotX” with your name.

- **Pick one before Friday, 3 February at 5pm or I will pick one for you and you won’t like it.**

Peer Review

- Each **graduate** (9637) student will be assigned three project reports to review
- Primary Purpose: Provide feedback to authors that they can make use of in their future careers, which gives them a better return on the investment they have made in their course project.
- Secondary Purpose: Give students a view of the variety of work that has been done in the course, and further develop reviewing skills.
- **Reviews from other students will not affect the grade of the author in any way.**
- **See the wiki for more details.**

Accessibility and Support, Missed Course Components

- **Check the wiki.**

Questions and Chat:

Why are you here?