Welcome to CS4414 / CS9637 / CS9114 Introduction to Data Science I

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

“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.”

Data

Data

1. As a count noun: an item of information; a datum; a set of data. Also *fig.*

2. As a mass noun.

a. Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation. Cf. datum *n.* 1a.

b. *Computing*. Quantities, characters, or symbols on which operations are performed by a computer, considered collectively. Also (in non-technical contexts): information in digital form. Cf. datum *n.* 1b.

[OED]

Science

Science

A branch of study that deals with a connected body of demonstrated truths or with observed facts systematically classified and more or less comprehended by general laws, and incorporating trustworthy methods (now esp. those involving the **scientific method and which incorporate falsifiable hypotheses**) for the discovery of new truth in its own domain. [OED]

Science

1. Generate a hypothesis
2. Generate data through observation and/or experiment
3. Assess whether the data are consistent with the hypothesis or not

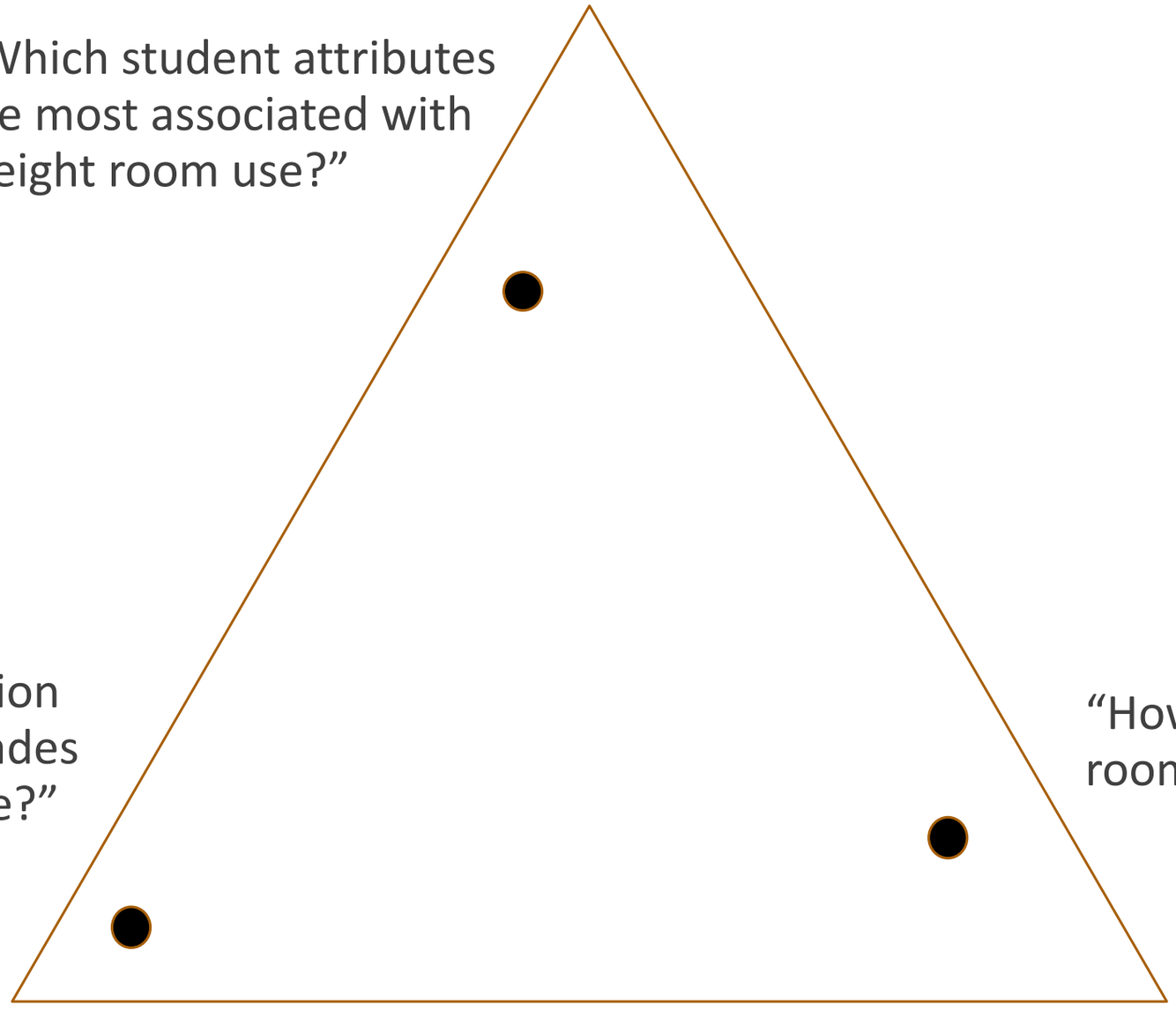
Data Science

1. Get some data
2. Try some methods
3. …

- https://twitter.com/WesternGymBot

Exploratory

"Which student attributes are most associated with weight room use?"



"Is there an association between student grades and weight room use?"

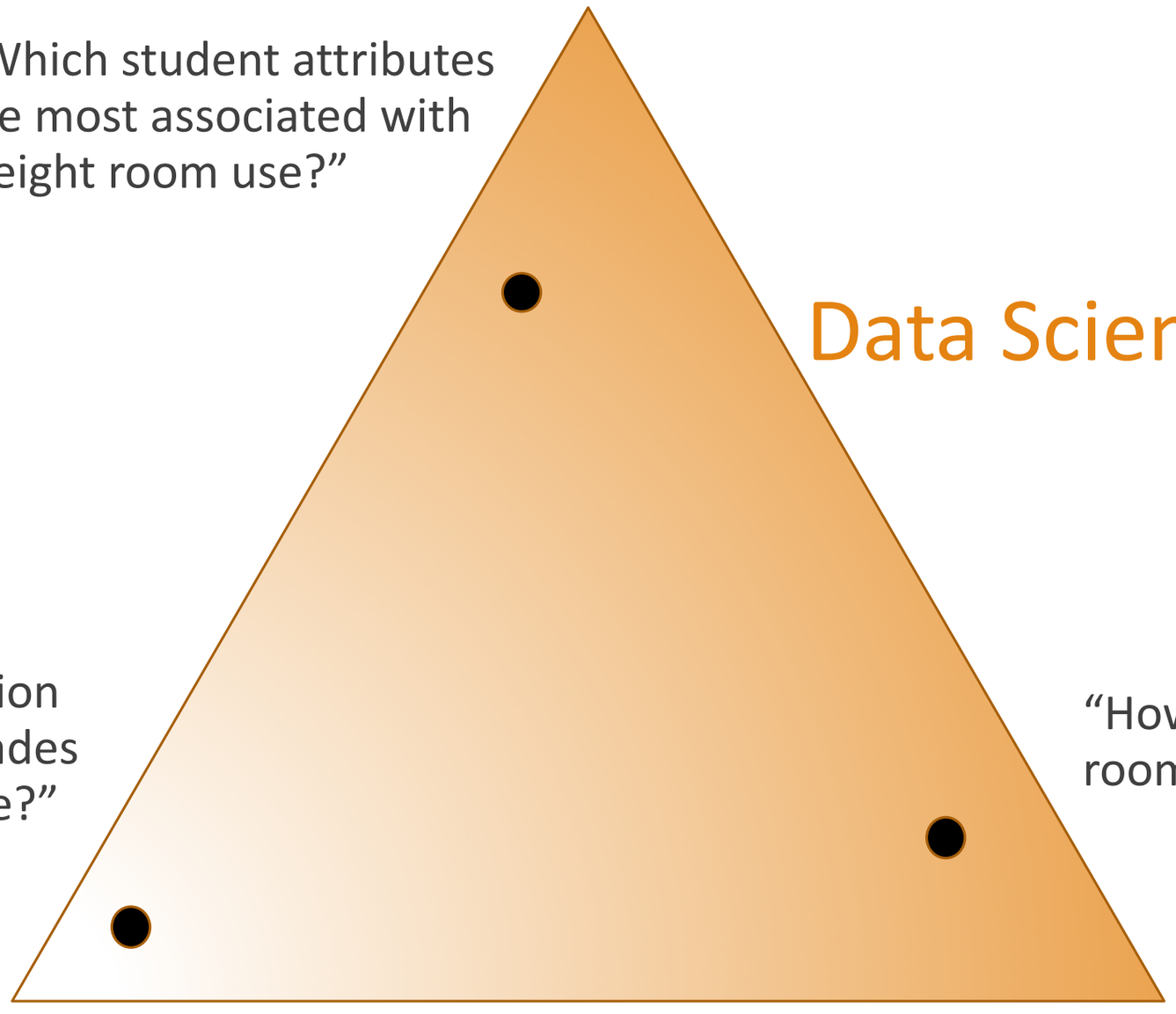
"How busy will the weight room be in the next hour?"

Confirmatory

Predictive

Exploratory

"Which student attributes are most associated with weight room use?"



Data Science

"Is there an association between student grades and weight room use?"

"How busy will the weight room be in the next hour?"

Confirmatory

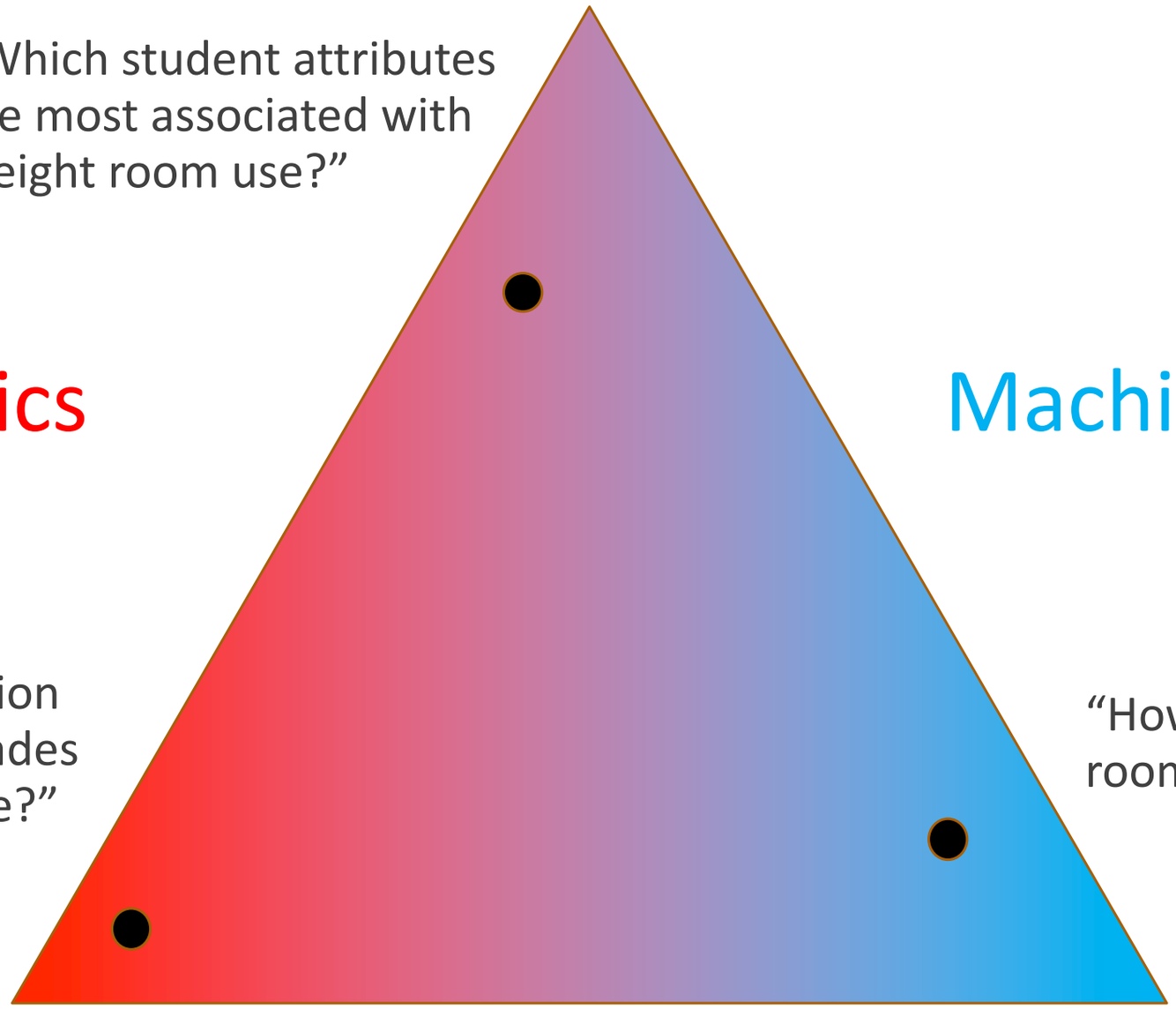
Predictive

Exploratory

"Which student attributes are most associated with weight room use?"

Statistics

Machine Learning



"Is there an association between student grades and weight room use?"

"How busy will the weight room be in the next hour?"

Confirmatory

Predictive

Any dangers with starting from data?

Any dangers with starting from data?

- Spurious results because of random noise (variance)

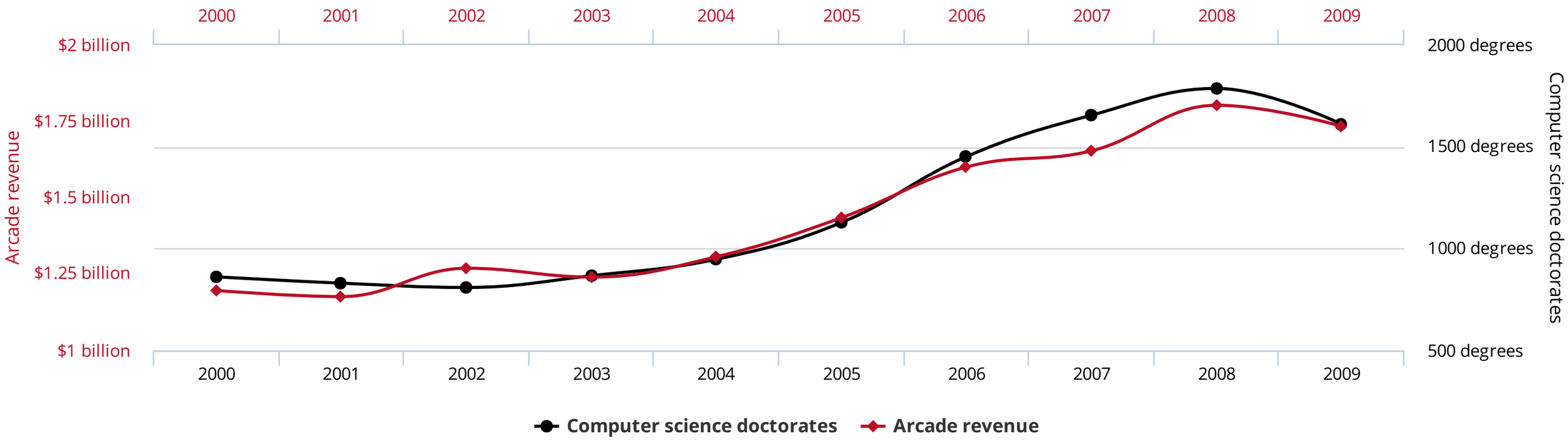
- Spurious results because of missing information (bias)

Variance

Total revenue generated by arcades

correlates with

Computer science doctorates awarded in the US



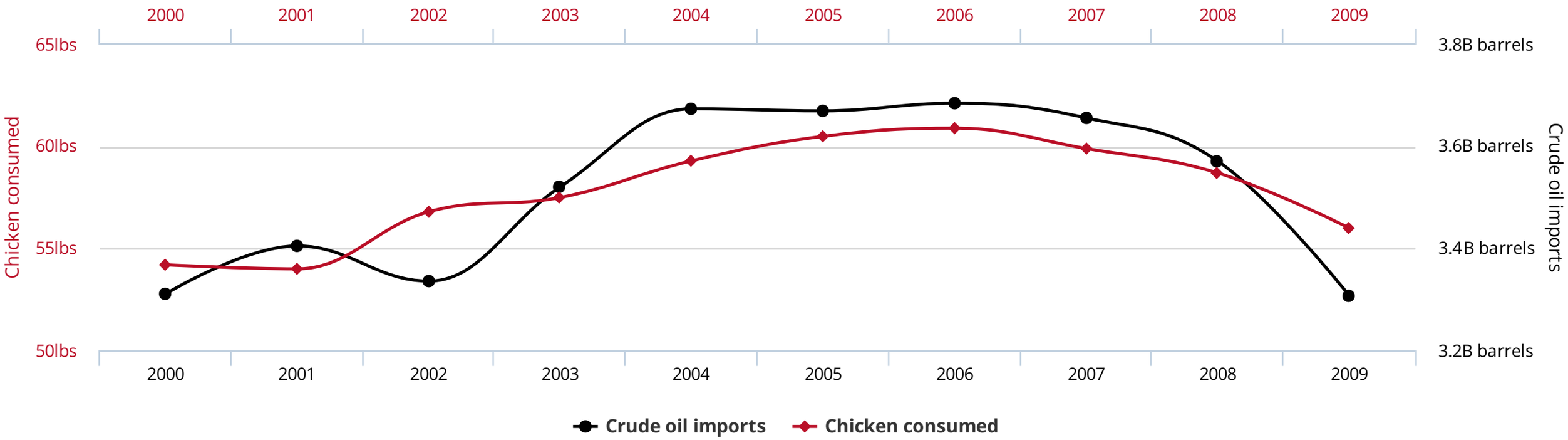
tylervigen.com

Correlation: 0.98

Per capita consumption of chicken

correlates with

Total US crude oil imports



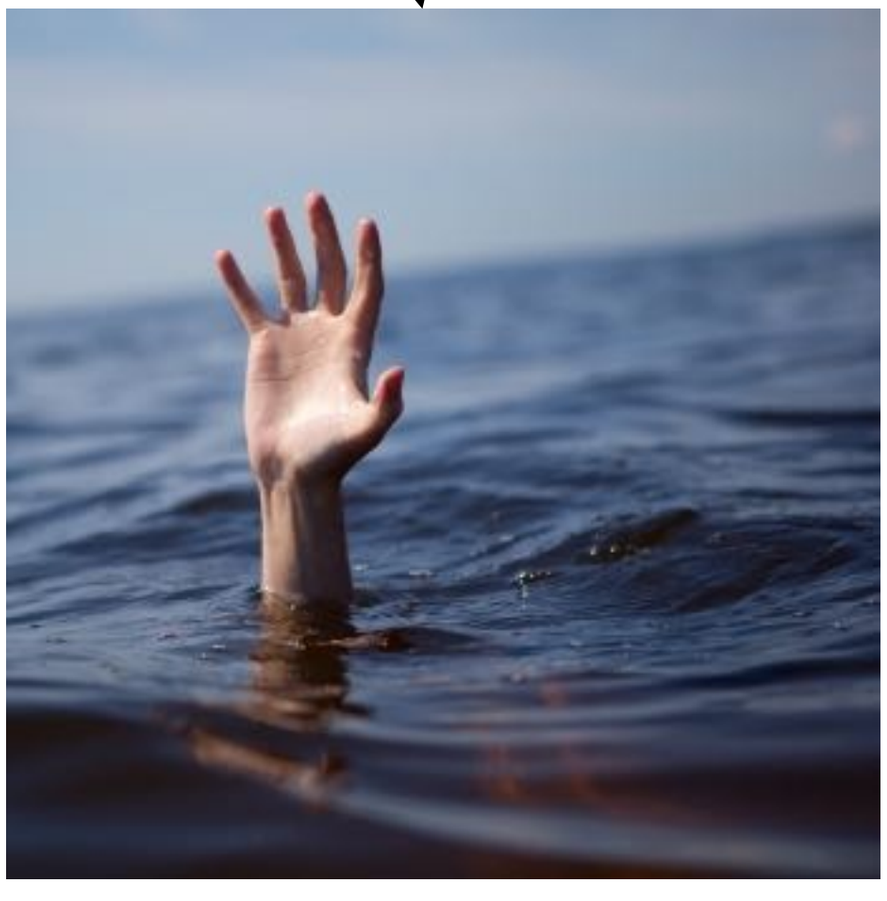
tylervigen.com

Correlation: 0.89

Bias

?

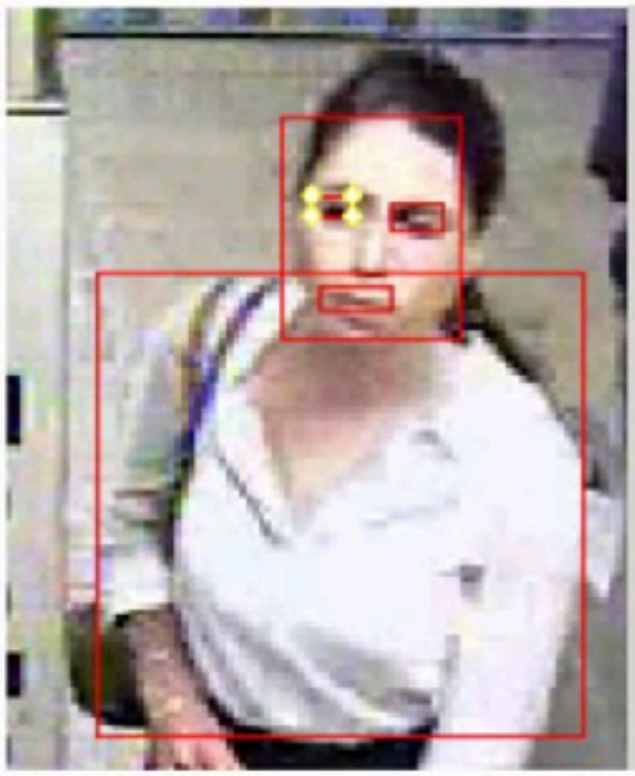
?



George Joseph @georgejoseph94 · 3h

SCOOP: With secret access to NYPD CCTV @IBM created software which tags people based on their skin tone + hair/clothing color. **IBM** gave NYPD access, then pitched them on a new AI product which identifies people on camera as "Black," "White," and "Asian":

theintercept.com/2018/09/06/nyp…



Event Attribute | Local Attribute

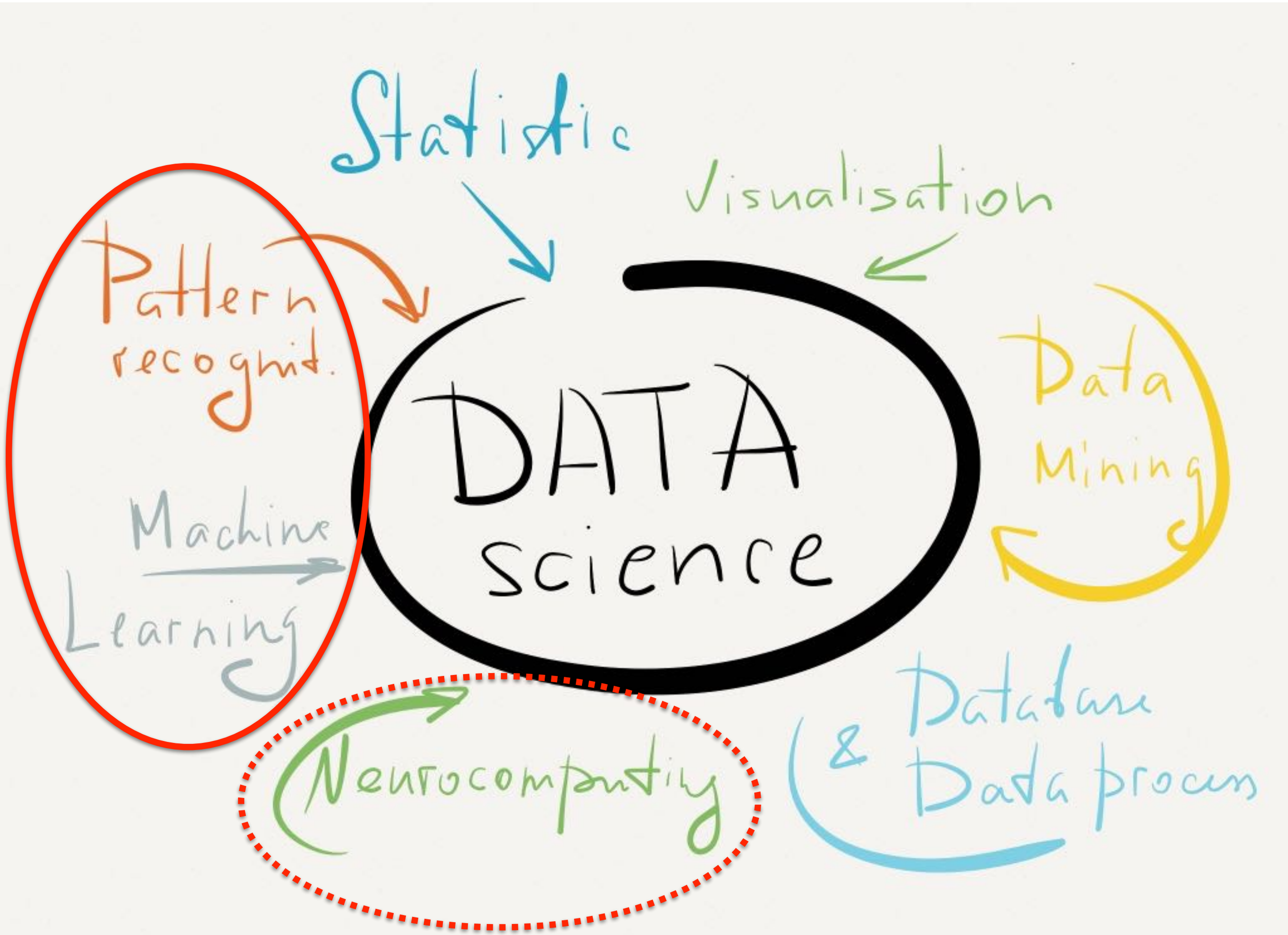
Add Attribute

ID	Name	Value
1	Bald	No
2	Sunglasses	No
3	Eyeglasses	No
4	Hat	No
5	Moustache	No
6	Beard	No
7	Skin Tone	Light
8	Gender	Female
9	Age	Adult (26-35)
10	Head Color	Brown
11	Torso Pattern	Solid
12	Torso Color	White

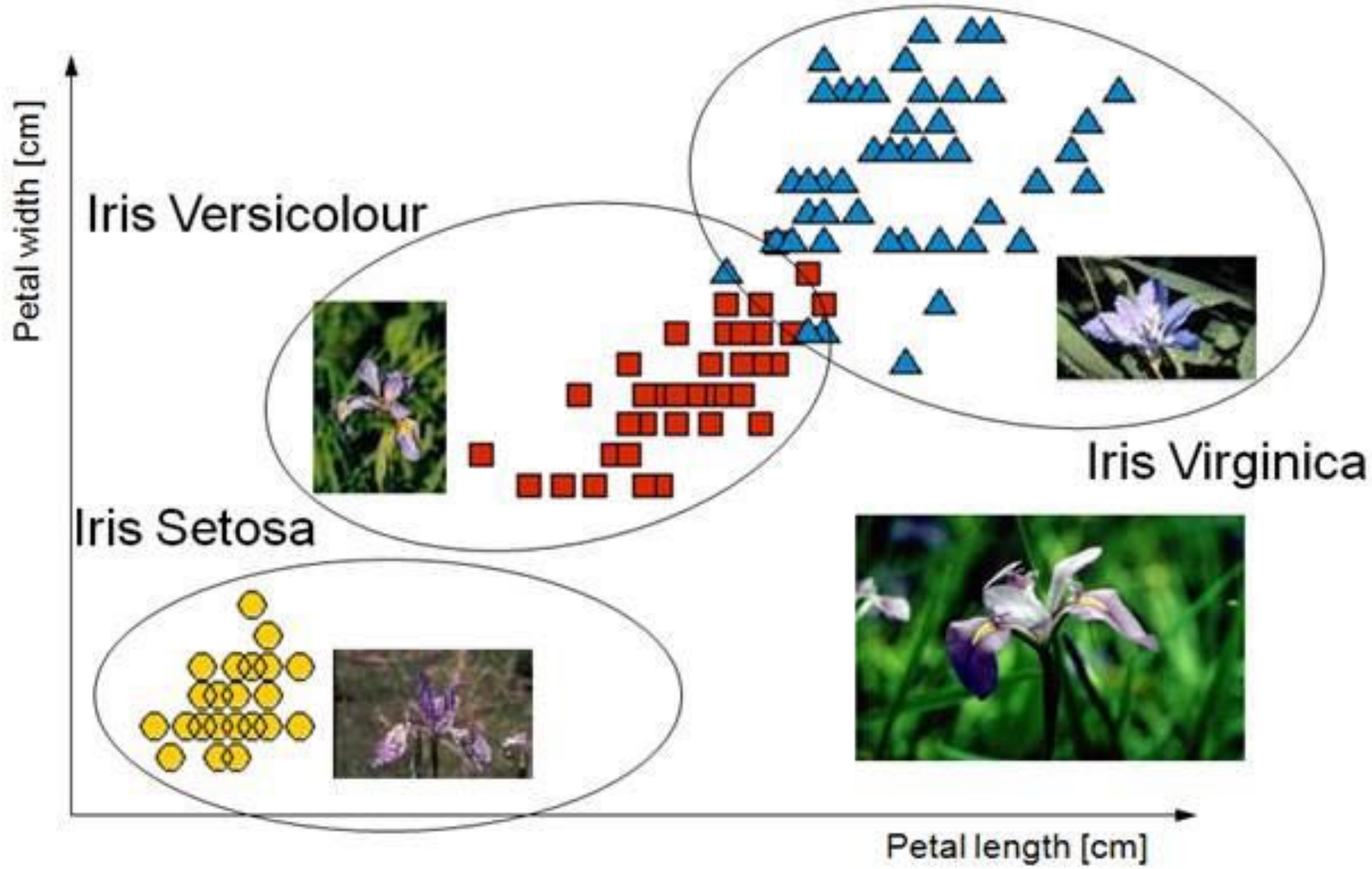
Event Attribute | Local Attribute

Add Attribute

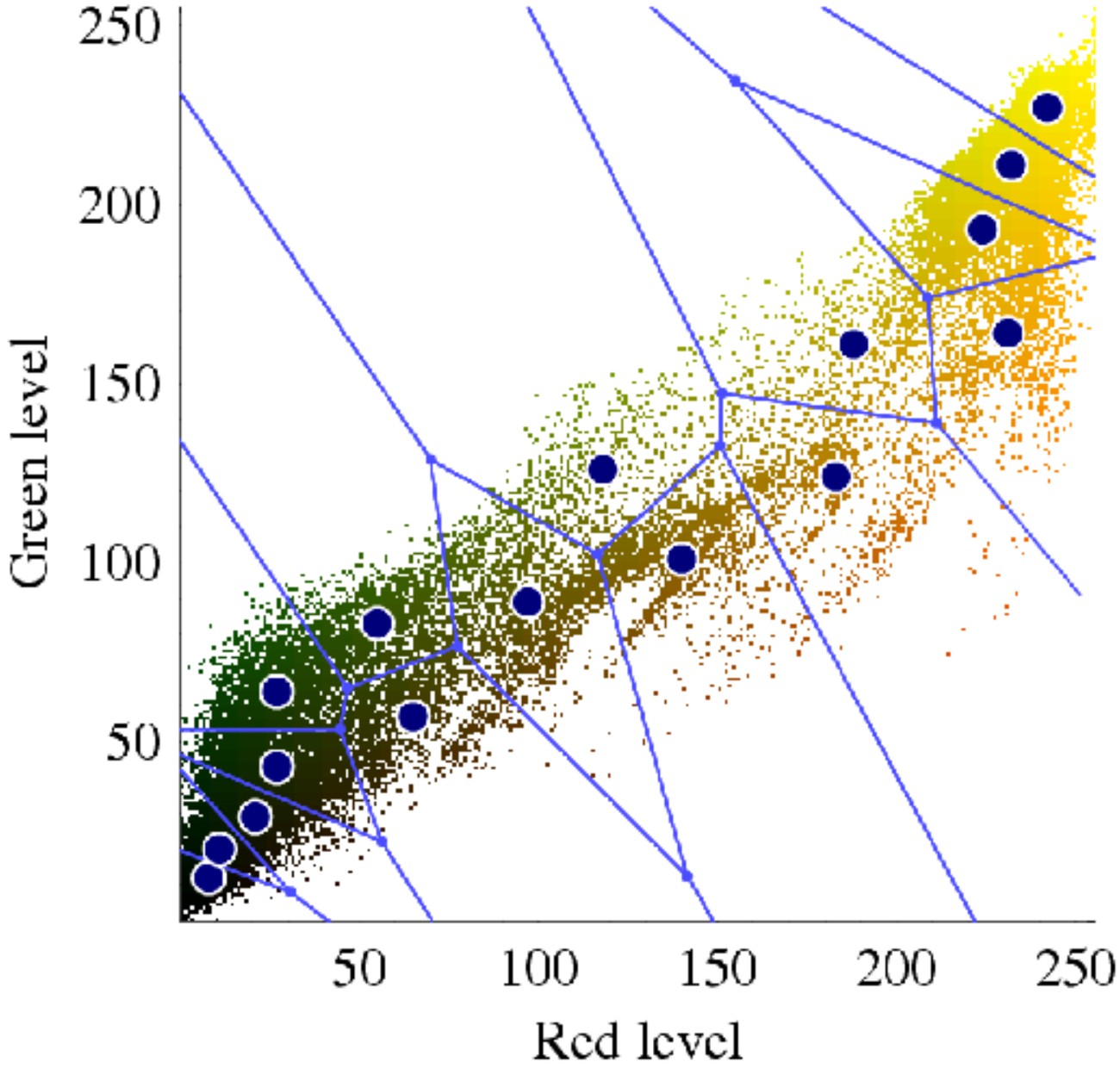
ID	Name	Value
1	Torso Visibility	Visible

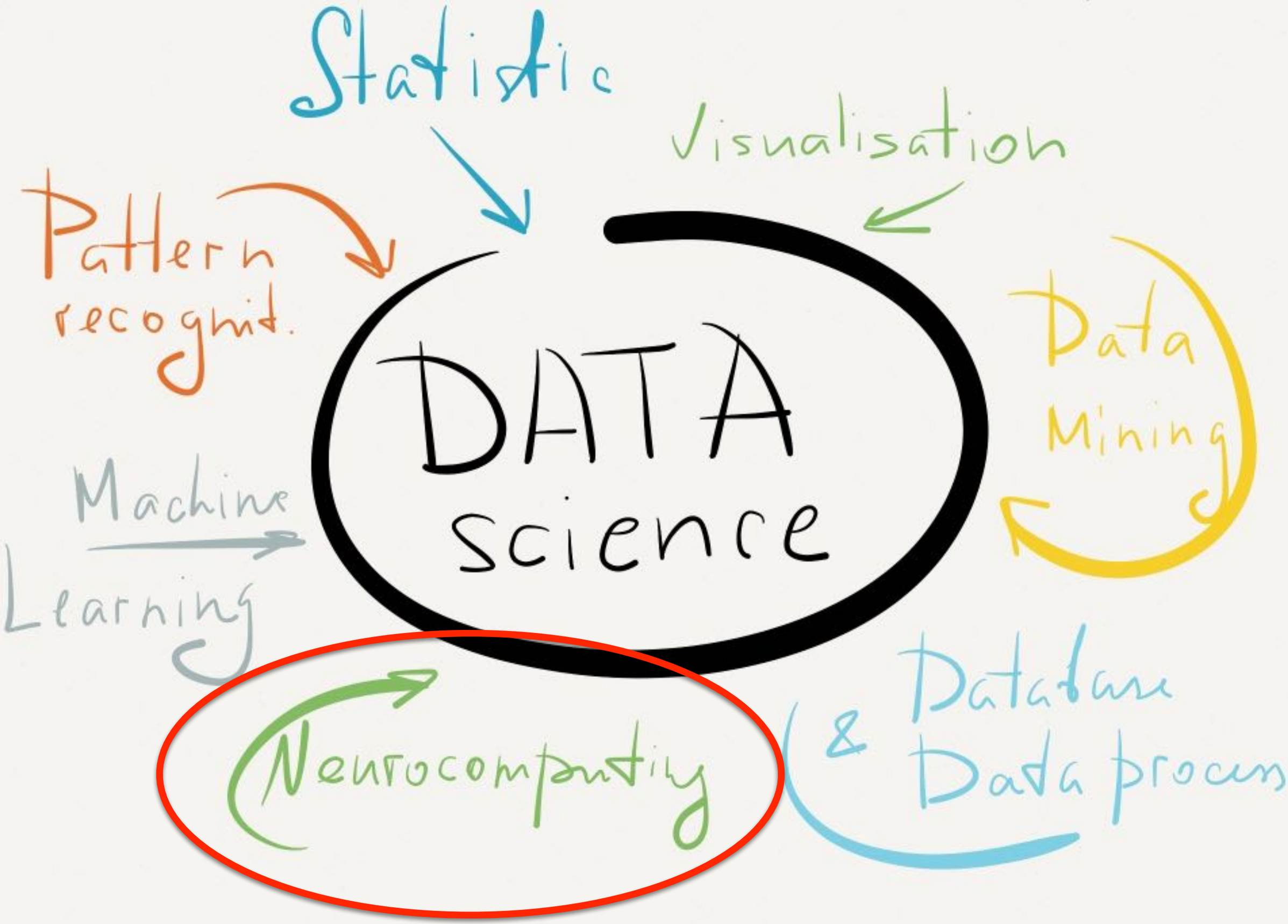

Statistic
Visualisation
Pattern recognit.
DATA
science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss

Supervised Learning

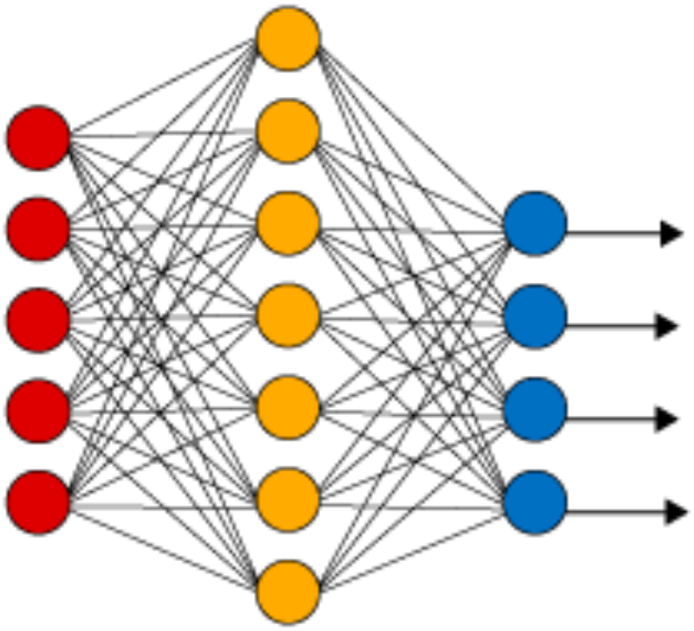


Unsupervised Learning

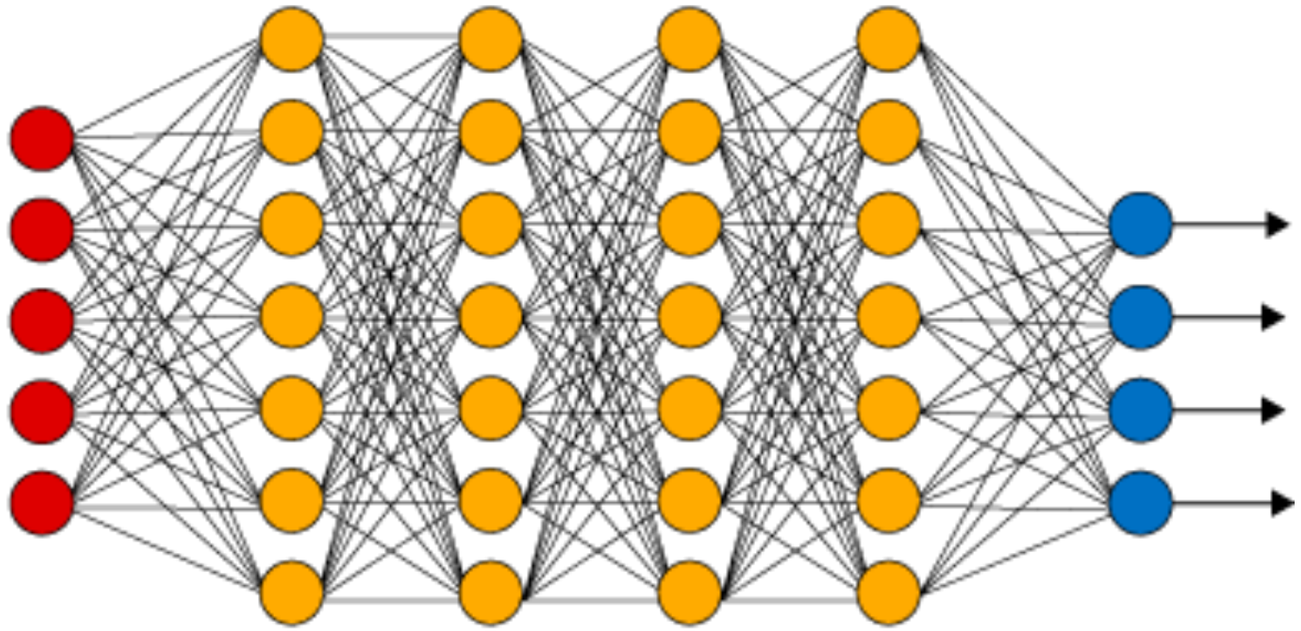



Statistic
Visualisation
Pattern recognit.
DATA
science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss

Simple Neural Network



Deep Learning Neural Network

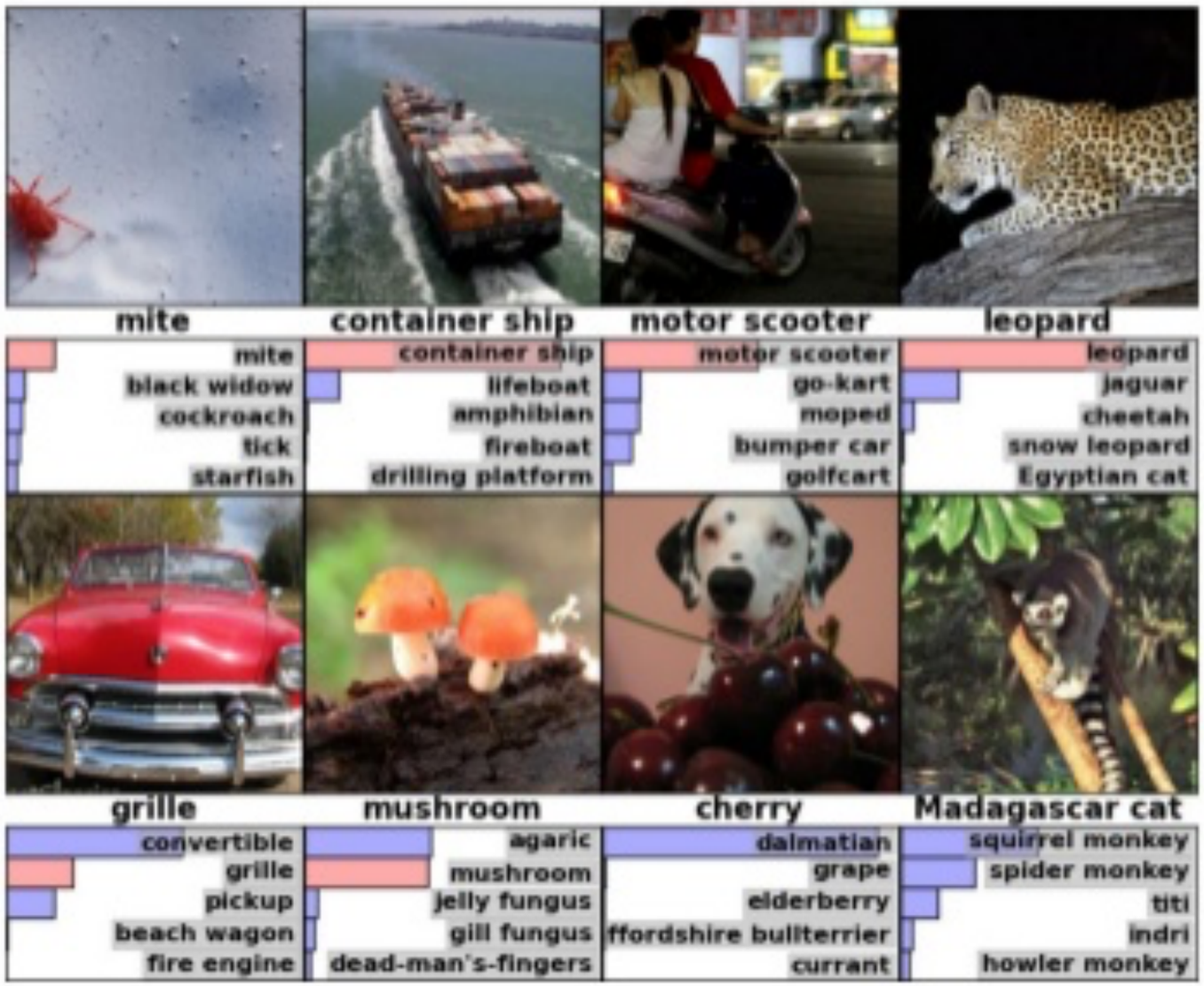


Input Layer Hidden Layer Output Layer

ImageNet Challenge

IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

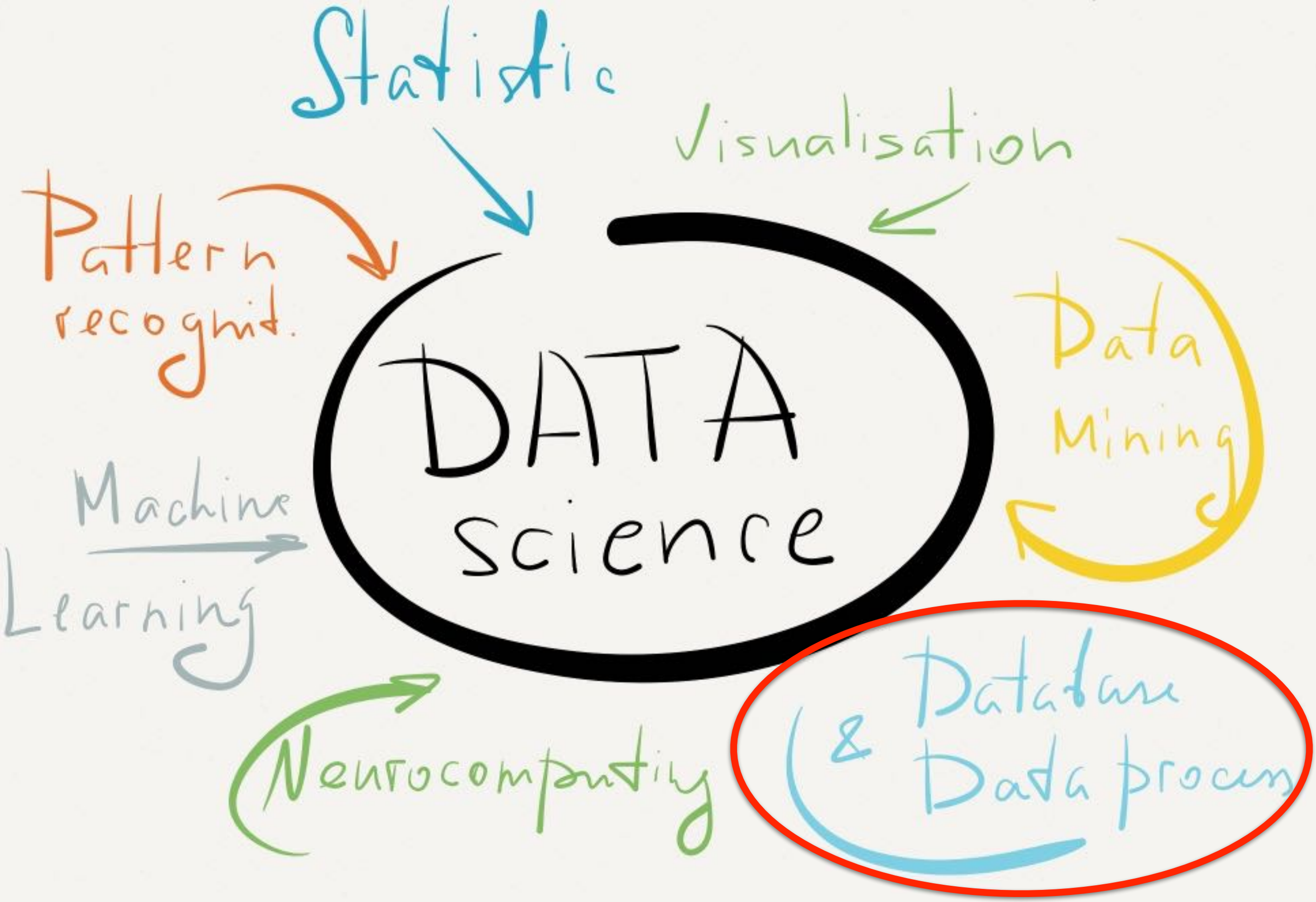


ImageNet Challenge: 2014

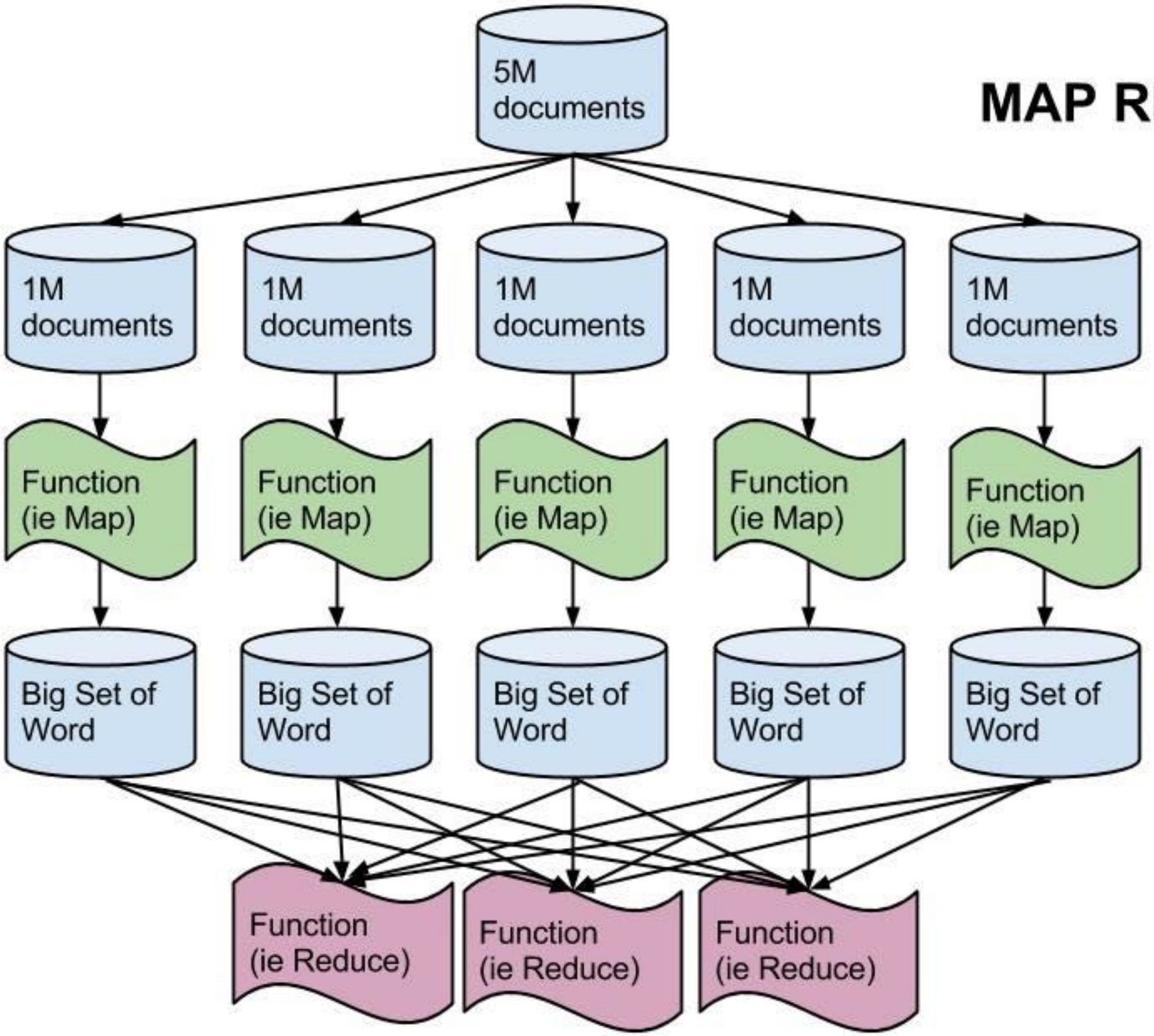
AlexNet

- image
- conv-64
- conv-192
- conv-384
- conv-256
- conv-256
- FC-4096
- FC-4096
- FC-1000

- image
- conv-64
- conv-64
- maxpool
- conv-128
- conv-128
- maxpool
- conv-256
- conv-256
- conv-256
- conv-256
- maxpool
- conv-512
- conv-512
- conv-512
- conv-512
- maxpool
- conv-512
- conv-512
- conv-512
- conv-512
- maxpool
- FC-4096
- FC-4096
- FC-1000
- softmax


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

MAP REDUCE

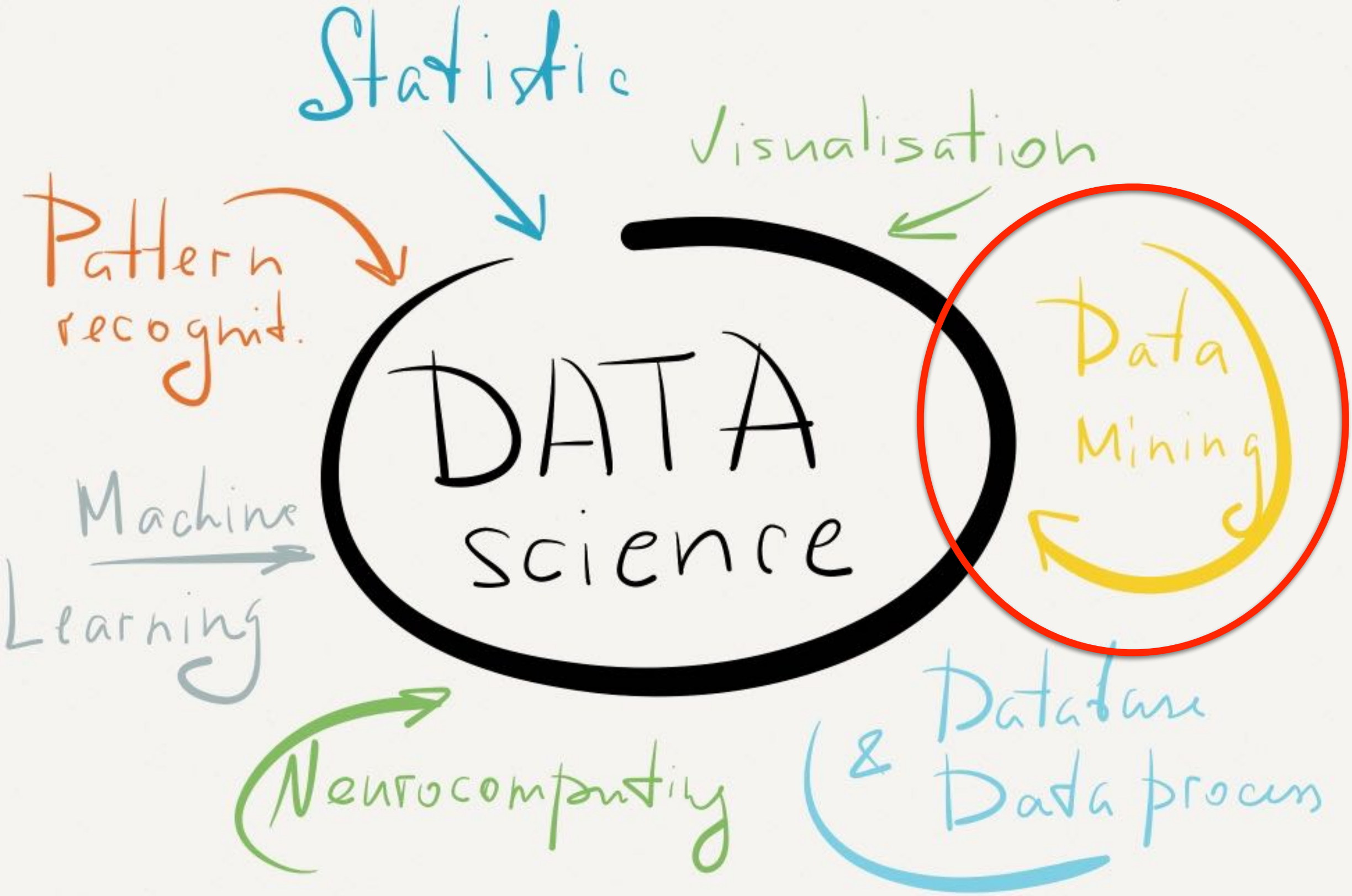


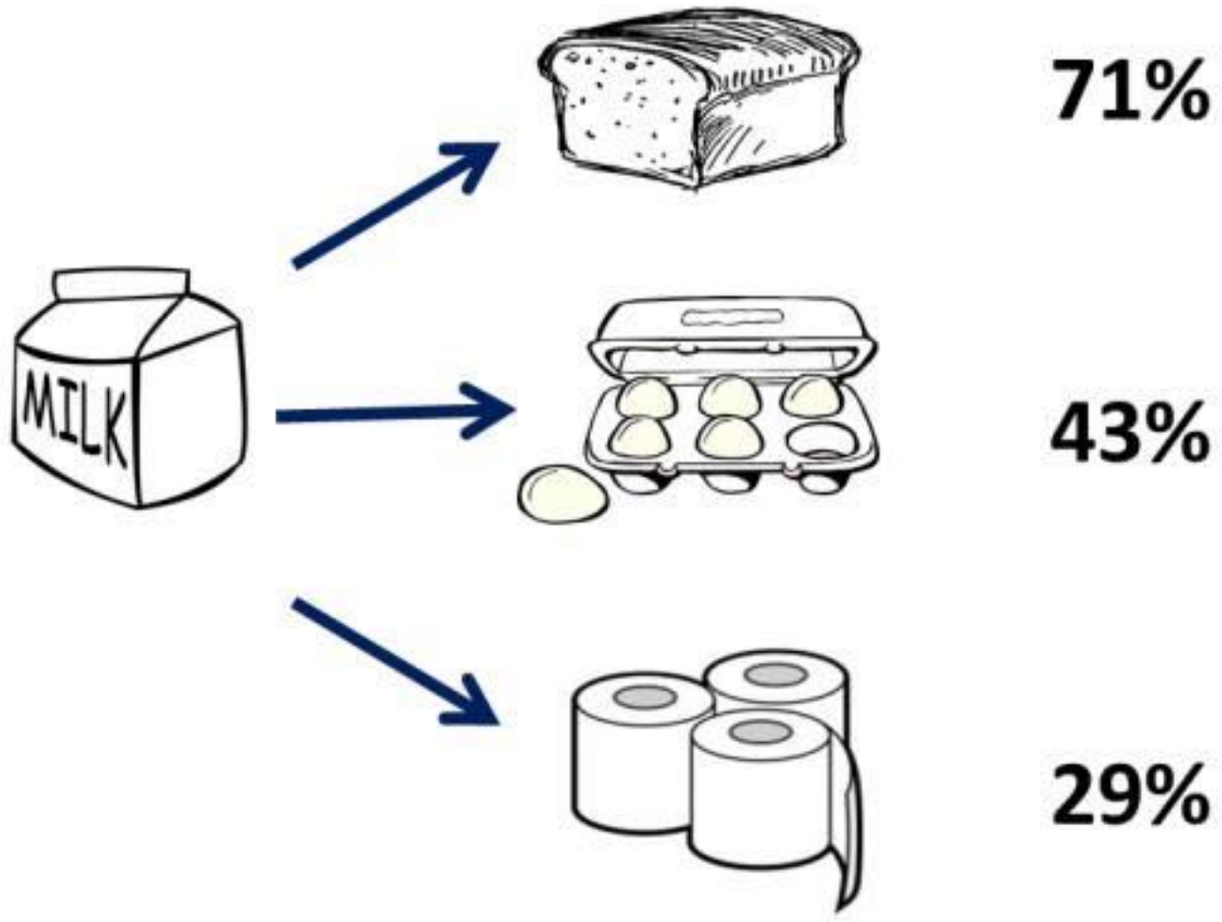
Distribute the documents across N computers

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

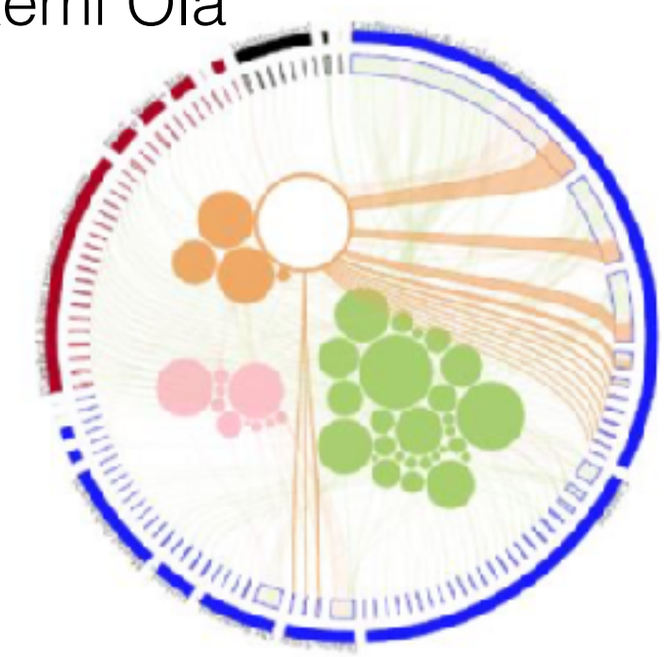


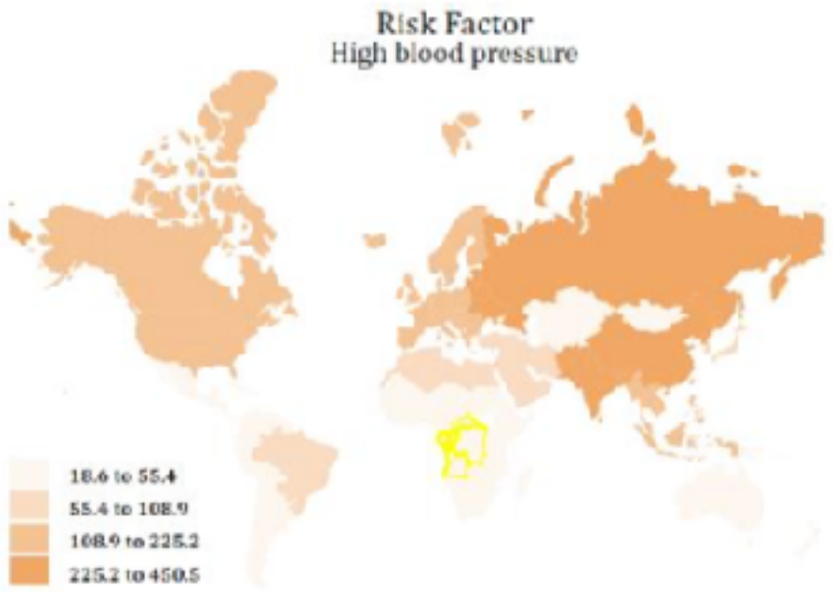
Of transactions that included milk:

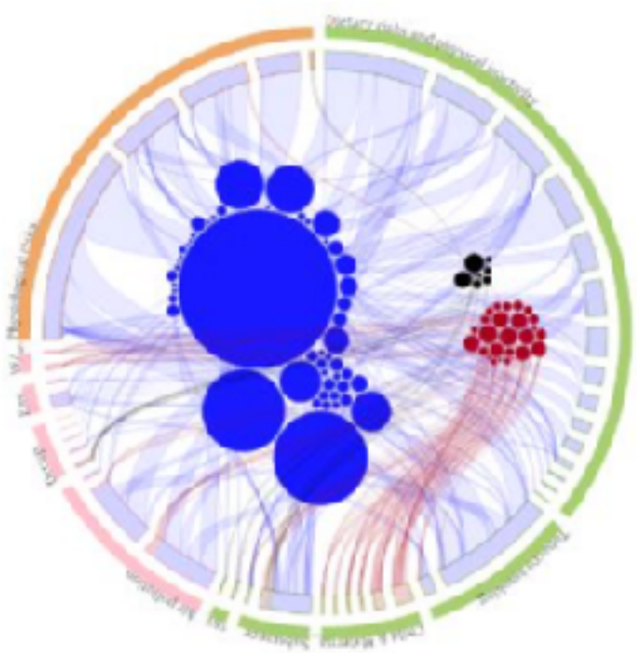
- 71% included bread
- 43% included eggs
- 29% included toilet paper


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

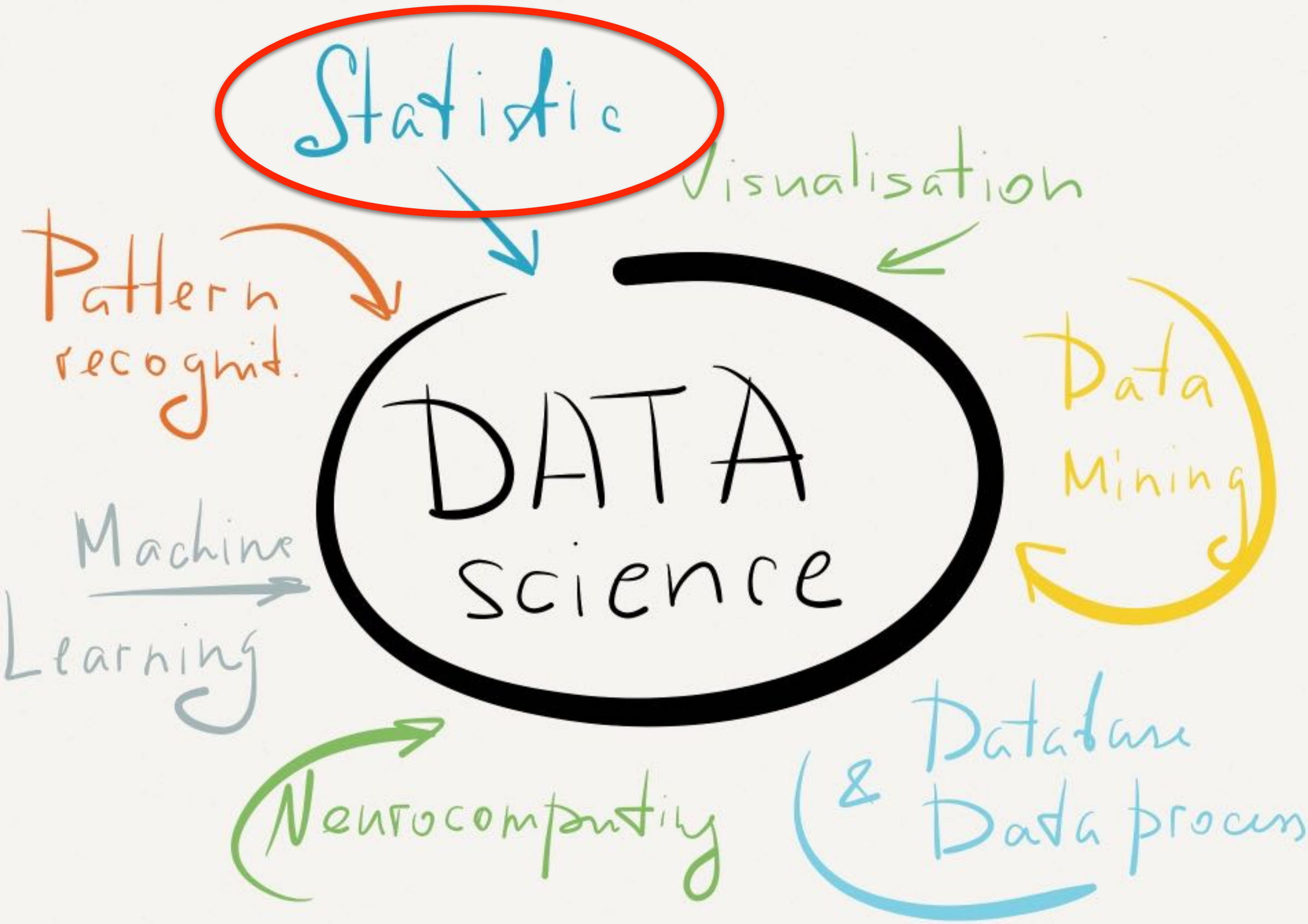
Dr. Kemi Ola

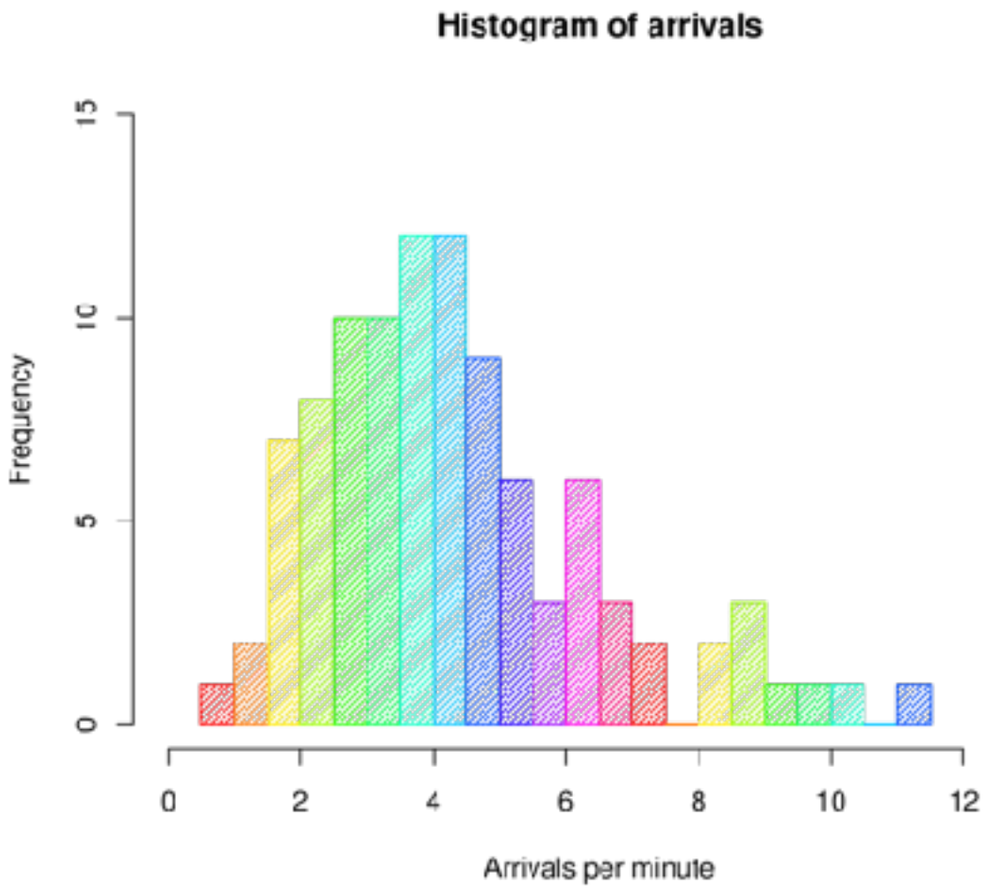


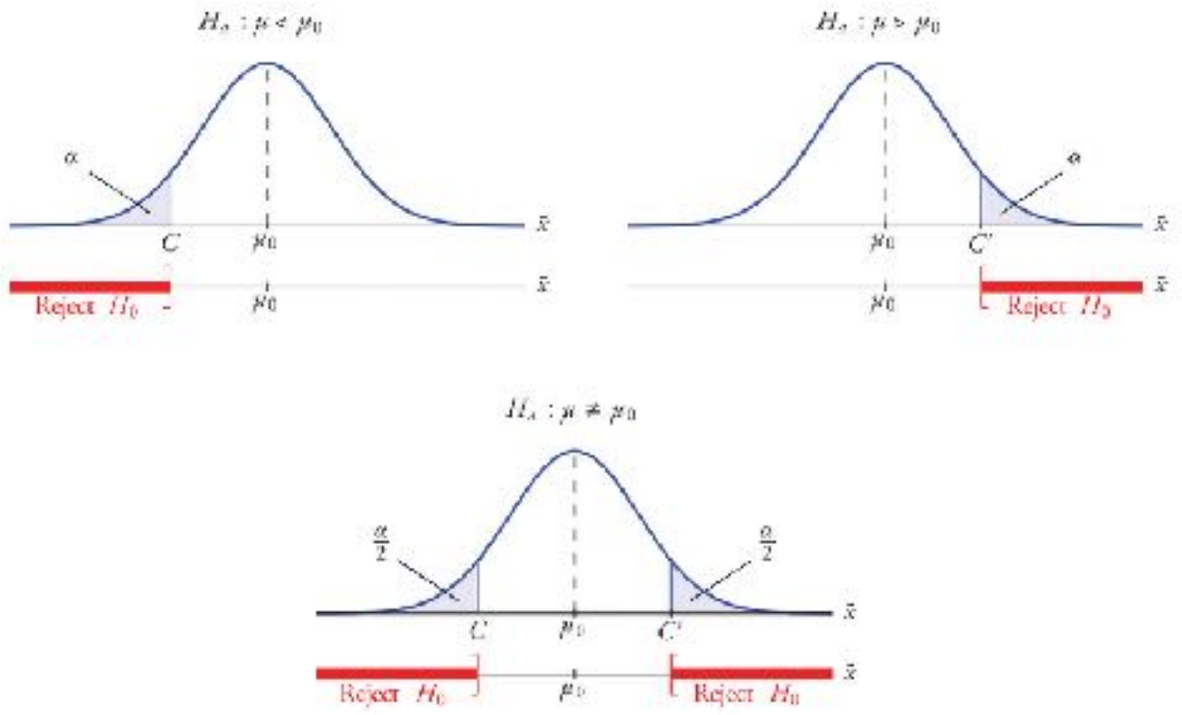

Risk Factor
High blood pressure
18.6 to 55.4
55.4 to 108.9
108.9 to 225.2
225.2 to 450.5

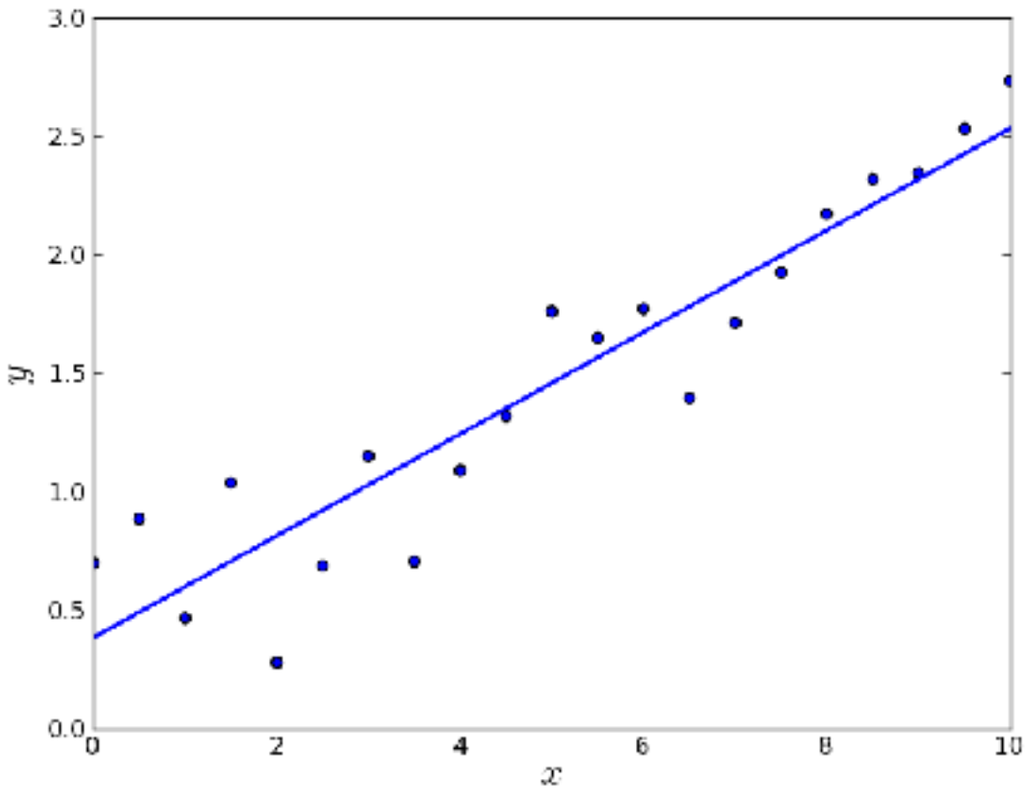


GLOBAL
Caribbean
Central Latin America
High-income North America
Western Europe
Central Europe
Eastern Europe
Andean Latin America
Central Asia
Tropical Latin America
East Asia
Southern Latin America
South Asia
North Africa and Middle East
Southeast Asia
Western Sub-Saharan Africa
Central Sub-Saharan Africa
Eastern Sub-Saharan Africa
Southern Sub-Saharan Africa
Australasia
Oceania
High-income Asia Pacific
Angola
Central African Republic
Congo
DR Congo
Equatorial Guinea
Gabon
High blood pressure
High body-mass index
High fasting plasma glucose
High total cholesterol
Low bone mineral density
0.0 to 4.0
4.0 to 8.0
8.0 to 57.3
57.3 to 106.5
Air pollution
Dietary risks & physical inactivity
Physiological risks
Smoking
Undernutrition
Water & sanitation
Cardiovascular & circulatory diseases
Chronic Respiratory diseases
Diabetes and blood/endocrine
Diarrheal & LR. diseases
Aortic aneurysm
Atrial fibrillation
Cardiomyopathy
Endocarditis
Hemorrhagic stroke
Hypertensive heart disease
Ischemic heart disease
Ischemic stroke
Other cardio & circulatory
Peripheral vascular disease
Rheumatic heart disease
0.0 to 4.0
4.0 to 8.0
8.0 to 33.9
33.9 to 59.8


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
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
Continues 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
Entropy
19%
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
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
5. Text Mining / NLP
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
Regression
Perceptron
Linear Regression
Ranking
Logistic 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?
Classification
30%
Euclidean Distance
Least Fit
Causation
Pearson Coeff
Correlation
80%
Sampling
Using ETL
How much Data?
Google OpenRefine
Data Survey
Stratified Sampling
Principal Component Analysis
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
9. Data Munging
8. Data Ingestion
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%
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scribe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
lm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
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%

Course Objective

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

- Through group 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/IDS/

- Instructor: Dan Lizotte – dlizotte at uwo dot ca – MC363
  TA: Nathan Phelps — nphelps3 at uwo dot ca

- Time: Tuesday from 2:30AM – 4:30PM, and on Thursday from 2:30PM – 3:30PM

- Place: Talbot College TC 205

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

- Question & Collaboration Hour: TBA

Materials

- **READ. THE. WIKI.**
- “Required” materials are materials that I expect you to consult if you have questions. Not required reading cover-to-cover.

Anticipated Topics and Schedule

- **Introduction to Data Science:** Definitions, Components, Relationships to Other Fields
- **Data Cleaning:** Working with structured data: selecting, filtering, joining, aggregating, Simple visualizations, “Face validity”
- **Supervised Machine Learning:** Regression, Classification. Linear Regression, SVMs, Trees, (Maybe also Reinforcement Learning and Sequential Decision Making)
- **(Re)-introduction to Statistics:** Data Summaries, Randomness, Sample Spaces and Events, Probability, Random Variables, Inference: Hypothesis testing, P-values, confidence Intervals Multivariate Statistics: conditional probability, correlation, independence
- **Evaluation:** Test set, cross-validation, bootstrap, confounding, causal inference
- **Unsupervised Machine Learning, Representations, and Feature Construction:** Clustering, Dimensionality reduction, Domain-specific Feature Development, Deep Learning, Images, Sounds, Text
- **Visualization**
- **Your picks?**

Evaluation

- Midterm - *4414/9114*: **35%** *9637*: **30%**
- Brainstorming Session – **10%**
- Project Proposal – *4414/9114*: **15%** *9637*: **10%**
- Report Draft – **5%**
- Project Report – **35%**
- Peer Review – *9637 only*: **10%**

Group Project (2 or 3)

- Project Proposal – 4414/9114: 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 - 10%

- Each group 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.

- Students are 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. Find the wiki on OWL
2. Edit the schedule a replacing “SlotX” with your names.

- **Pick one before Friday, 5 October at 5pm or we will pick one for you.**

Peer Review

- Each **research 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?