Databases in Finance

by Angel Marchev, Jr.

Data science overview

by Angel Marchev, Jr.

Shift happens

- https://www.youtube.com/watch?v=fbcMPGyPr8k

THE 5 V‘s OF DATA

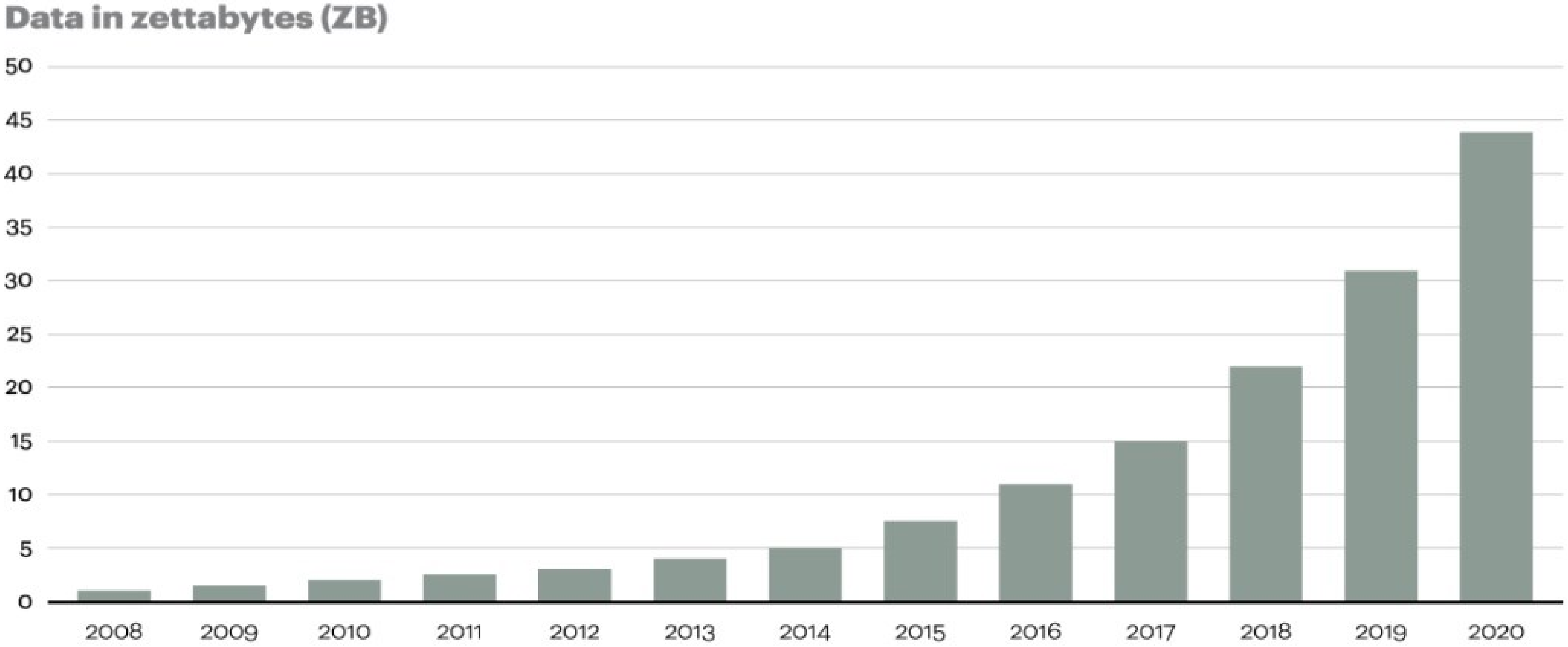
Big Data does a pretty good job of telling us what happened, but not why it happened or what to do about it. The **5 V‘s represent specific characteristics and properties** that can help us understand both the challenges and advantages of big data initiatives.

Volume	**The magnitude** of the data being generated.	90% of the data in the world today has been created in the last 2 years alone.
Velocity	**The speed** at which data is being generated and aggregated.	Literally the speed of light! Data doubles every 40 months.
Variety	The different **types** of data.	Structured, semi-structured and unstructured data.
Veracity	**The trustworthiness** of the data in terms of accuracy in quality.	Because of the anonimity of the Internet or possibly false identities, the reliability of data is often in question.
Value	**The economic value** of the data.	Having access to big data is no good unless we can turn it into value.

How big ?

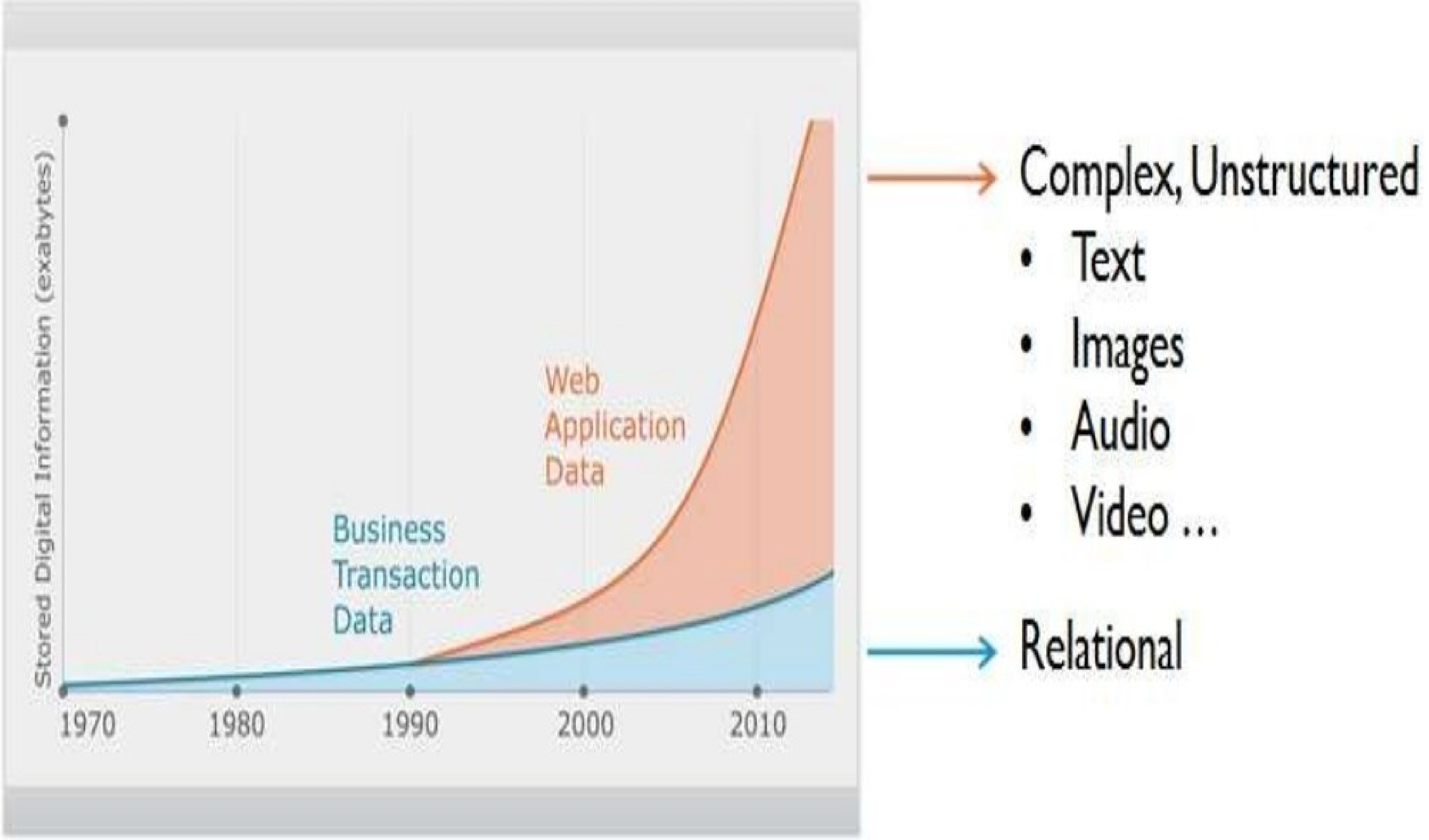
Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

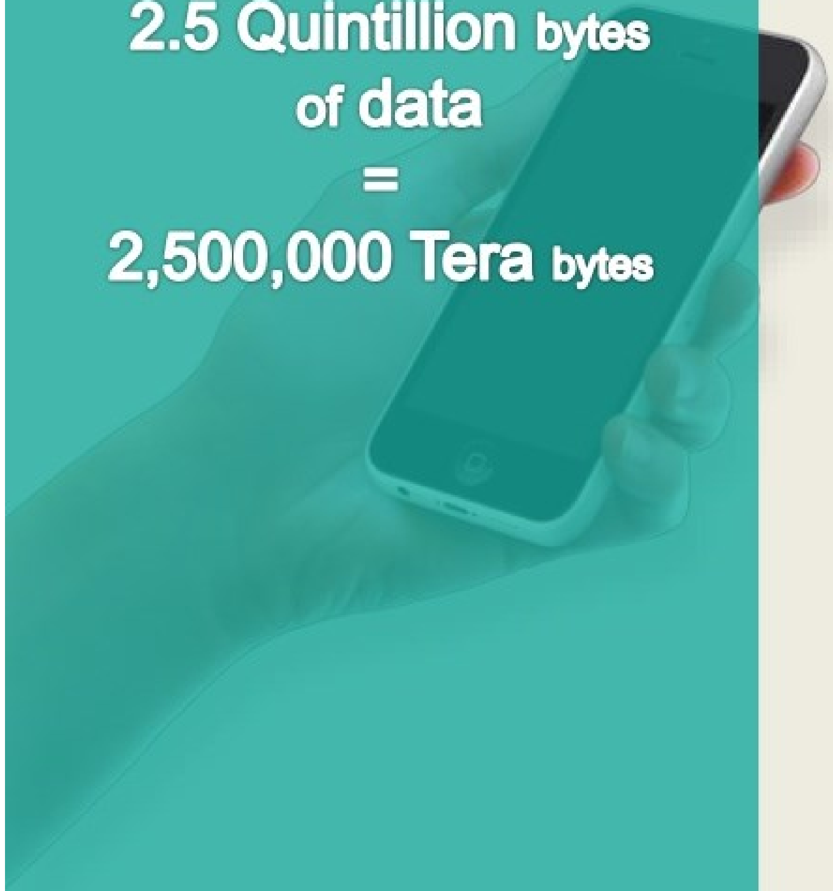


Source: Oracle, 2012

Structured data vs Unstructured data



How Much Are Data That You Create ?



we create
2.5 Quintillion bytes
of data
=
2,500,000 Tera bytes

Daily routine

▪ What Happens In An Internet Minute ?

In an internet Minute

FACEBOOK

701,389
Facebook
Logins

TWITTER

347,222
New Tweets

LINKED IN

120+
NEWLINKEDIN
ACCOUNTS

INSTAGRAM

38,194
Posts To
Instagram

Google

2.4MILLION
Search Queries

WHATSAPP

20.8MILLION+
Messages

Email

150MILLION
Emails sent

Just one min

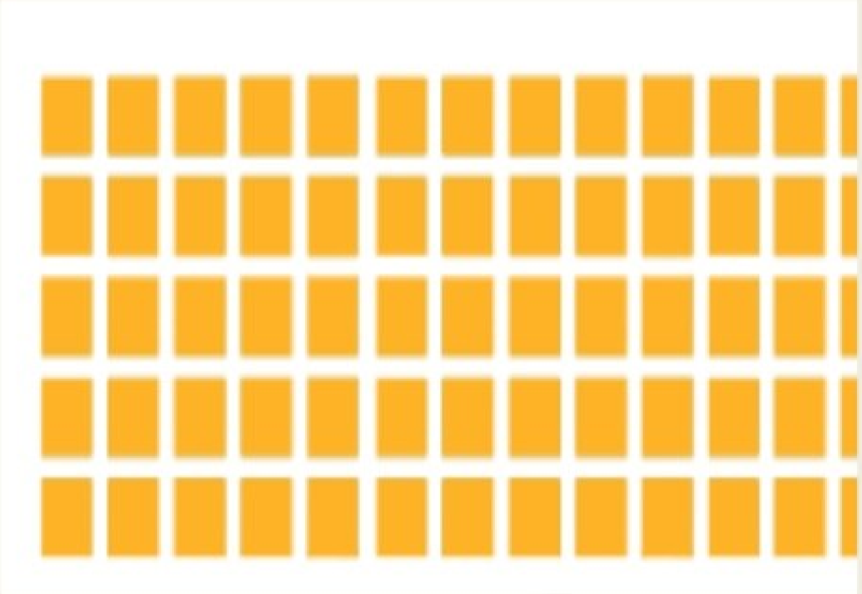
YouTube

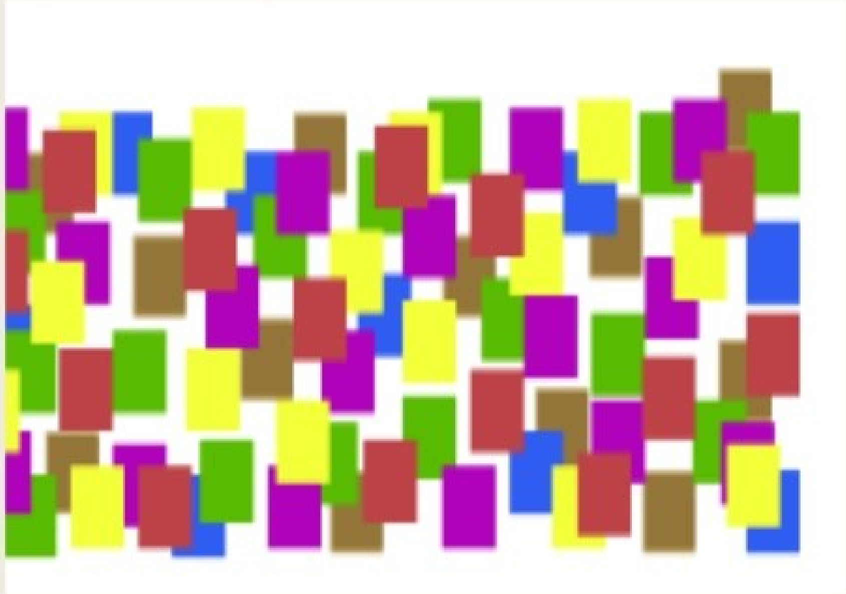
2.78MILLION
Video Views

App Store

51,000 app
downloads from
apples

- Structured and Unstructured Data: What is It?





What is Structured Data?

What is Unstructured Data?

Need For Data Science

So Data Science is mainly needed for:

Better Decision Making

Whether A or B?

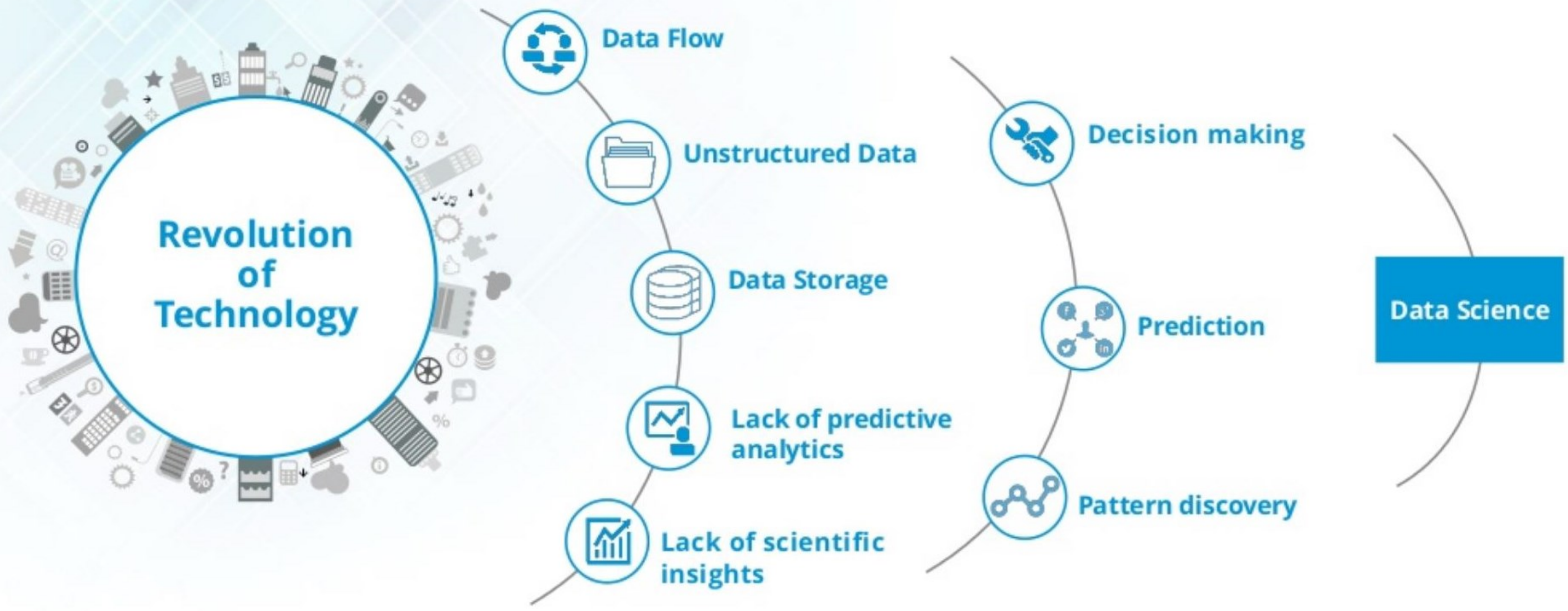
Predictive Analysis

What will happen next?

Pattern Discovery

Is there any hidden information in the data?

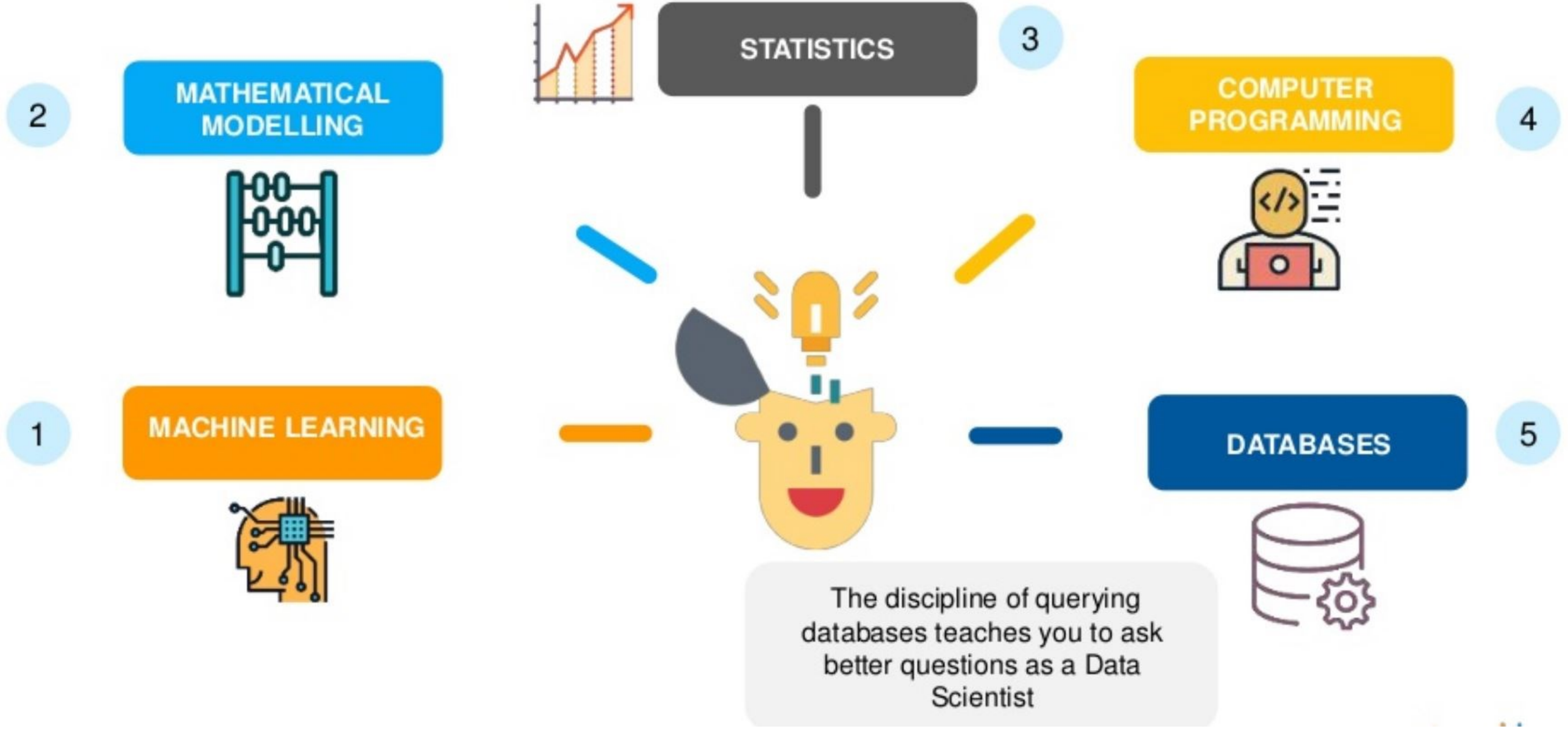
Need Of Data Science



BI Vs. Data Science

Characteristics	Business Intelligence	Data Science
Perspective	Looking Backward	Looking Forward
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Prerequisites for Data Science



Tools/Skills used in Data Science

Data Warehousing

Skills: ETL, SQL,Hadoop, Apache Spark,

Tools: Informatica/ Talend, AWS Redshift

Data Analysis

Skills: R, Python, Statistics

Tools: SAS, Jupyter, R studio, MATLAB, Excel, RapidMiner

Data Visualization

Skills: R, Python libraries

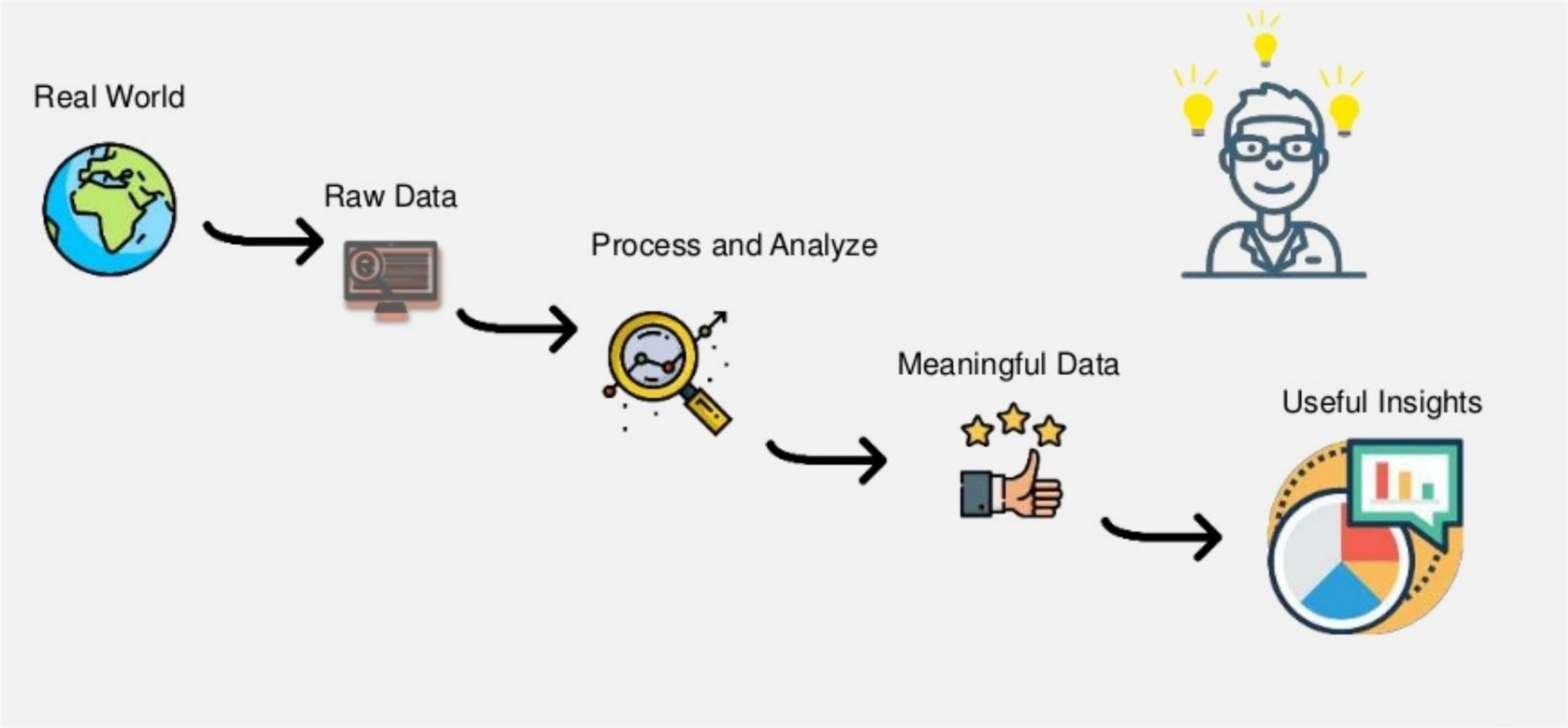
Tools: Jupyter, Tableau, Cognos, RAW

Machine Learning

Skills: Python, Algebra, ML Algorithms, Statistics

Tools: Spark MLib, Mahout, Azure ML studio

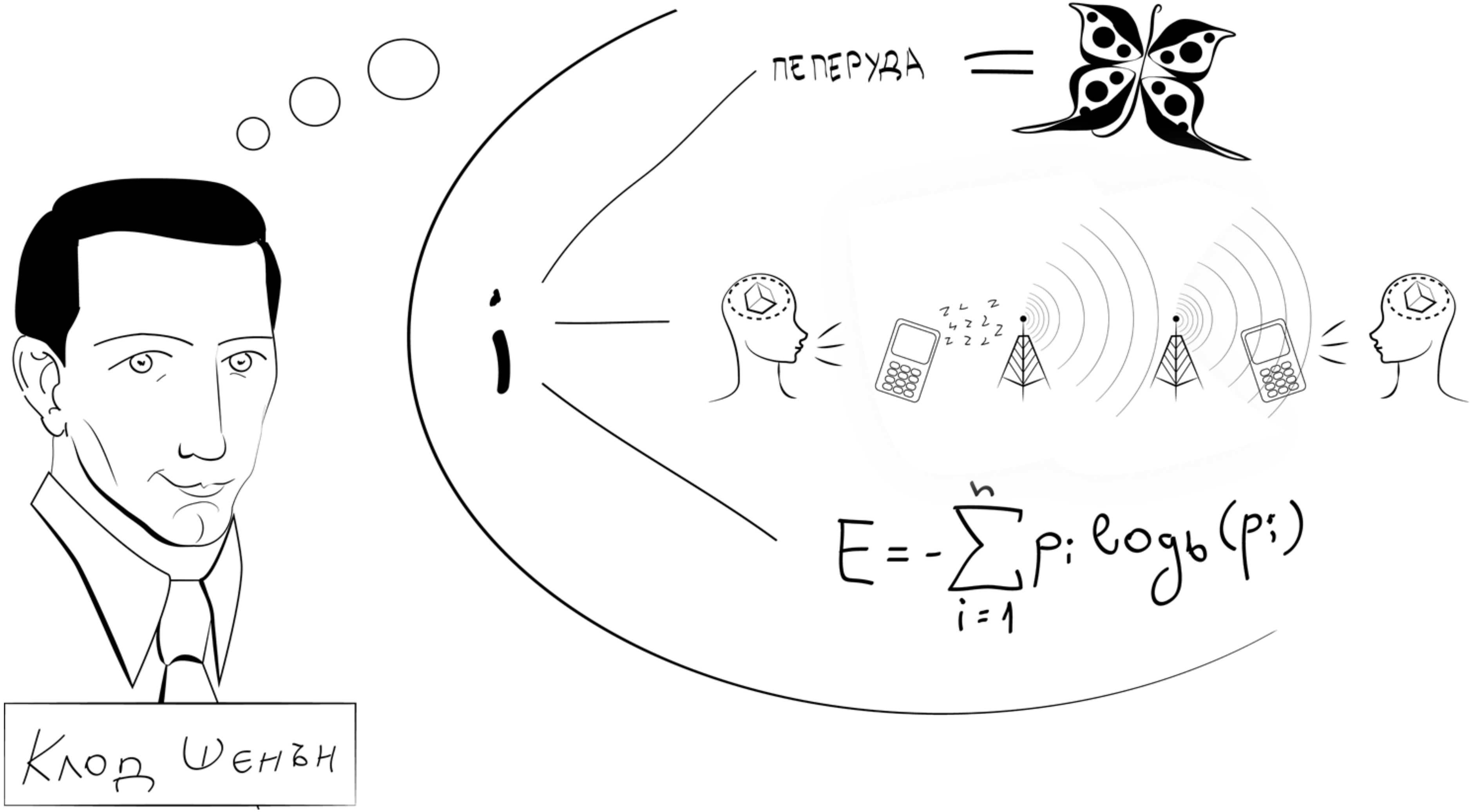
What does a Data Scientist do?



Theory of Information

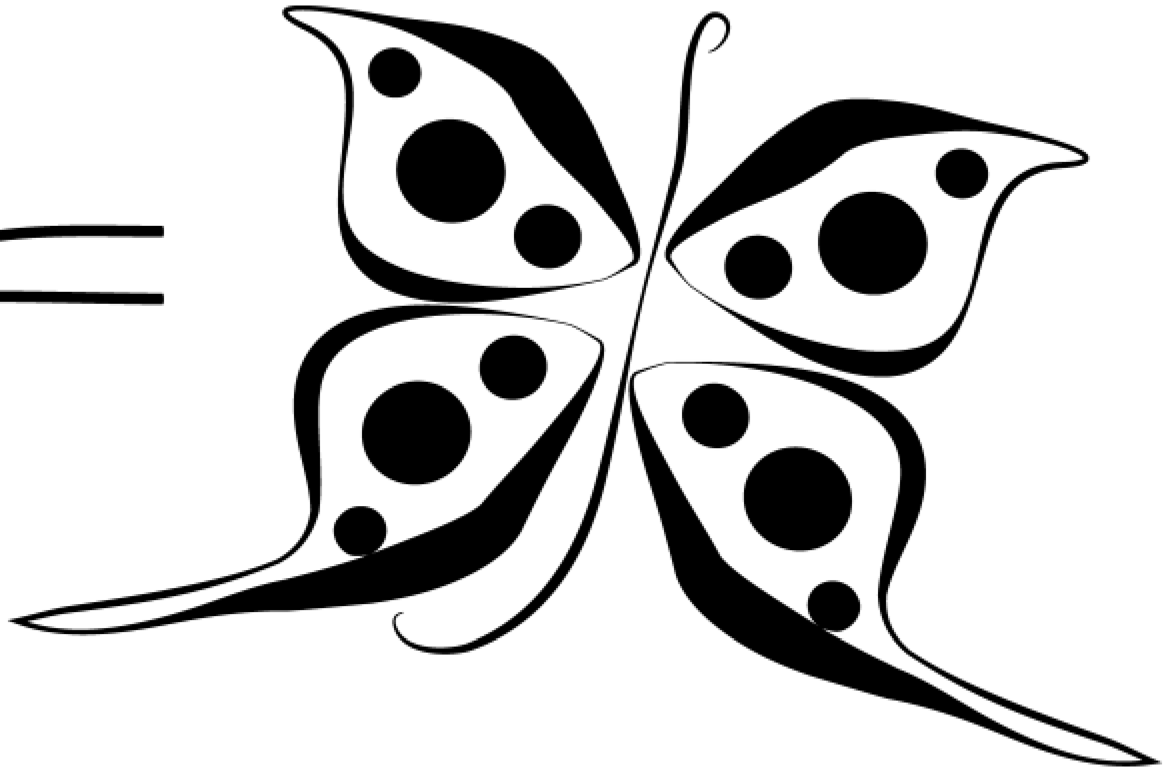
by Angel Marchev, Jr.

Theory of Information

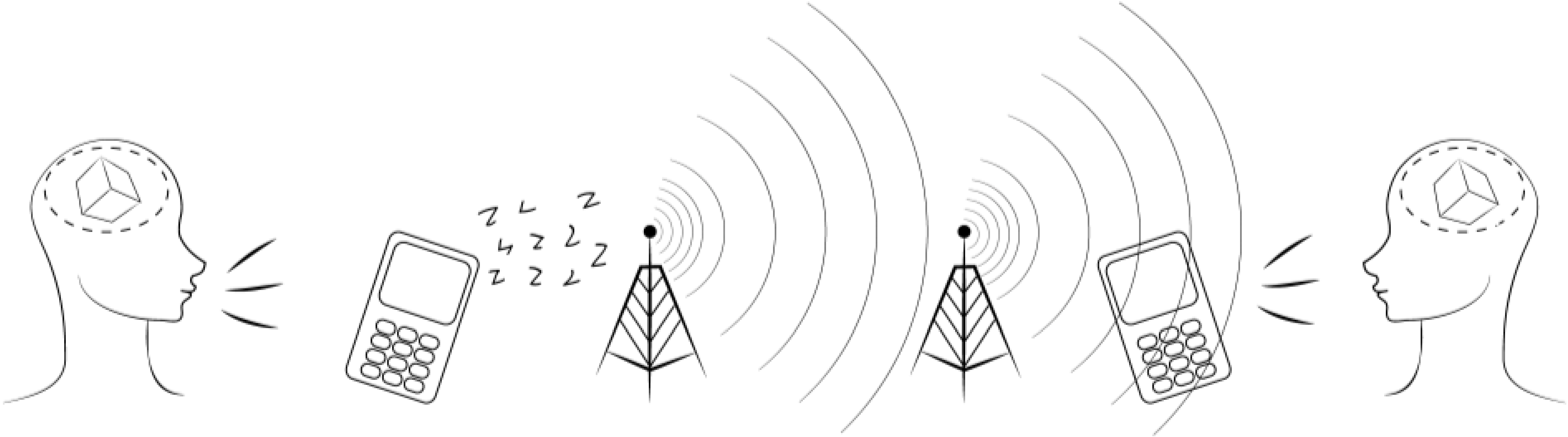


Semantics

ПЕПЕРУДА =



Communication



Entropy

$$E = -\sum_{i=1}^{n} p_i \log_b(p_i)$$

Informational entropy

- Let the system S have n possible states Ai with corresponding probabilities Pi :
- $A_1 \quad A_2 \quad \ldots \quad A_i \quad \ldots \quad A_n$
- $p_1 \quad p_2 \quad \ldots \quad p_i \quad \ldots \quad p_n$

$$\sum_{i=1}^{n} p_i = 1$$

- The degree of uncertainty of the system is estimated by the value ENTROPY

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

Properties of entropy

- 0) Entropy is determined ONLY by :
  - n (the number of possible system states / number of possible outputs)
  - P_n (corresponding probabilities)
- 1) $H \geq 0,\ 0 \leq p_i \leq 1 \Rightarrow log_2 p_i > 0$
- 2) $H = max,\ p_1 = p_2 = p_3 = \ldots = p_n$
- 3) $H = 0,\ p_k = 1\ p_{j \neq k} = 0$
- 4) $0 \leq H \leq log_2 n$

A measure of entropy

- Unit of measure for entropy: the uncertainty of an event with two equally probable outcomes.
- $H = \log_2 n$
- $N=2 \Rightarrow H=1$
- "BIT"

Problem

- System with 5 possible states with probabilities:
- $$p_1 = \frac{1}{4}, p_2 = \frac{1}{8}, p_3 = \frac{1}{2}, p_4 = 0, p_5 = \frac{1}{8}$$
  $$H = ?$$

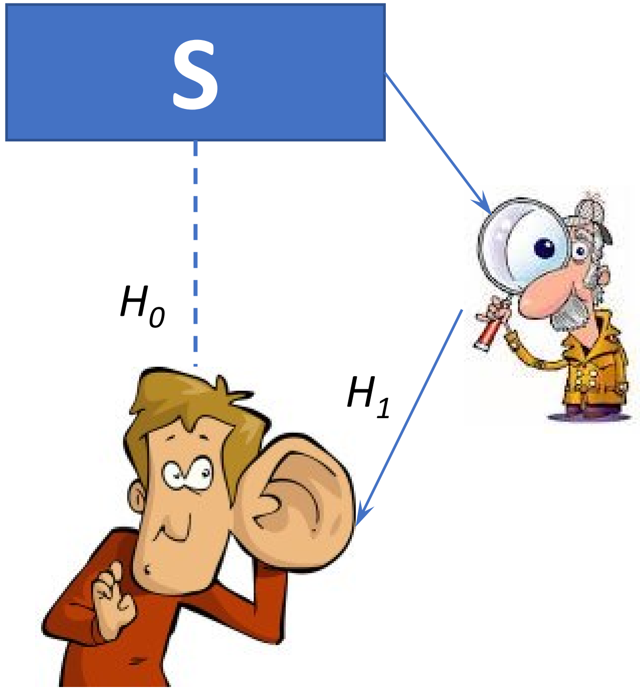
$$H = -\sum_{i=1}^{n} p_i . \log_2 p_i$$

$$H = -\frac{1}{4}.\log_2 \frac{1}{4} - \frac{2}{8}.\log_2 \frac{1}{8} - \frac{1}{2}.\log_2 \frac{1}{2} - 0 - \log_2 \frac{1}{4}$$

$$\log_2 4 = 2, (2^2 = 4)$$

$$H = \frac{1}{4}.2 + \frac{2}{8}.3 + \frac{1}{2}.1 = \frac{14}{8} = 1.75$$

Amount of information



- A measure to reduce uncertainty, ie. Measure the novelty received as a result of the message

$$H_1 < H_0$$

$$I = H_0 - H_1$$

$$H_1 = 0, I = H_0$$

Problems

$I=?$	H_0	H_1
p_1	1/4	0
p_2	1/8	1/4
p_3	1/2	1/2
p_4	0	0
p_5	1/8	1/4

How many bits of information is contained in the statement: "My wife gave birth to a girl"?

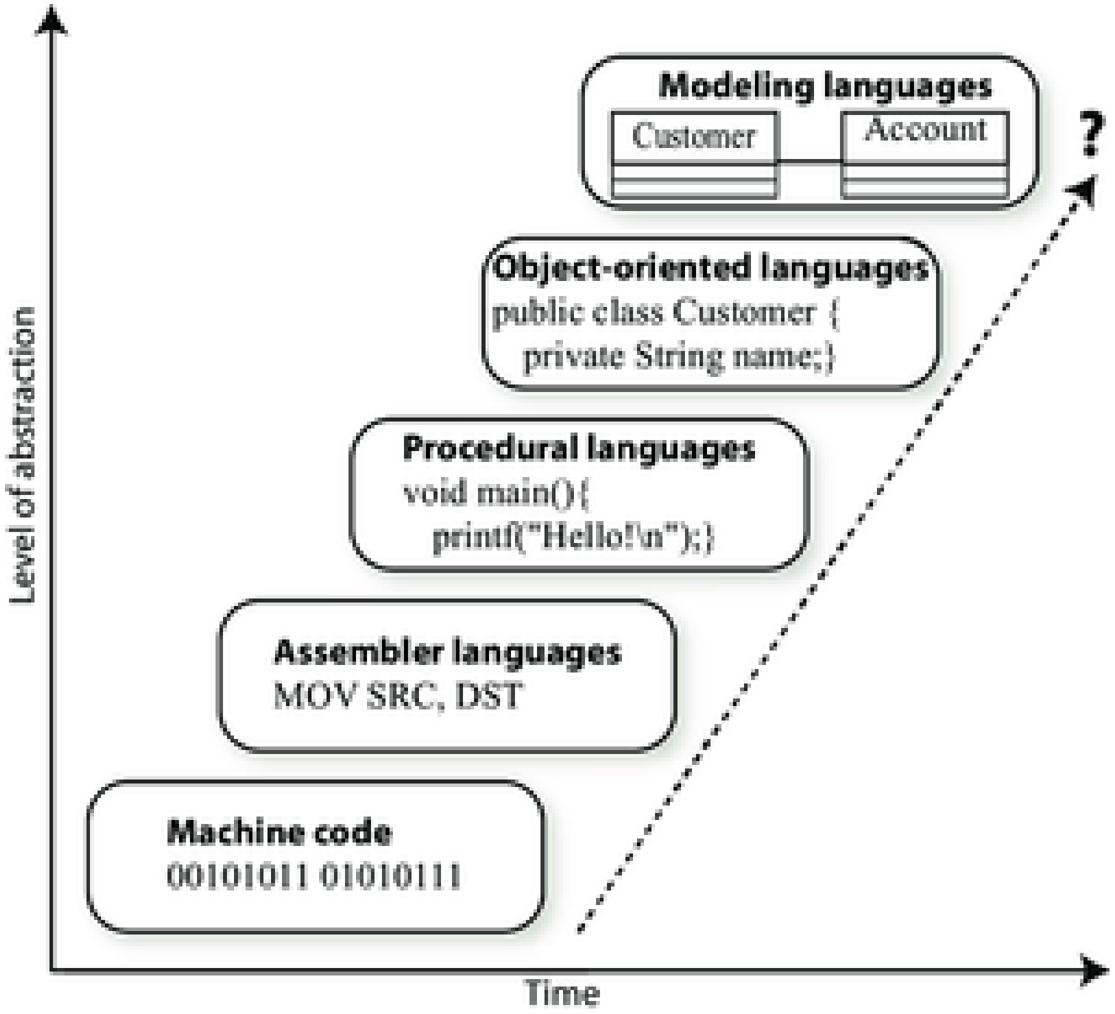
How many bits of information are contained in the statement: "Of the 5 possible answers to question 20 of the test, I know for sure that the answer is either A or B"?

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

$$I = H_0 - H_1$$

Data as information

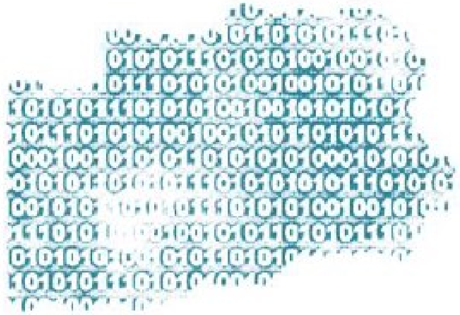
by Angel Marchev, Jr.



Level of abstraction
Time
Machine code
00101011 01010111
Assembler languages
MOV SRC, DST
Procedural languages
void main(){
printf("Hello!\n");}
Object-oriented languages
public class Customer {
private String name;}
Modeling languages
Customer
Account
?

Introduction / Data, Information & Knowledge

- Data = Information + Uncertainty



- Information = Meaningful Component in Data

$$y = f(x)$$

- Knowledge = Comprehended Information

Data:

A set of values recorded on one or more observational units i.e. Object, person etc

Types of data:

(D) Qualitative/ Quantitative data

(E) Discrete/ Continuous data

(F) Primary/ Secondary data

(G) Nominal/ Ordinal data

- **Qualitative data**:
  - also called as *enumeration data* .
  - Represents a particular quality or attribute.
  - There is no notion of magnitude or size of the characteristic, as they can't be measured.
  - Expressed as numbers without unit of measurements . Eg: religion, Sex, Blood group etc.
- **Quantitative data:**
  - Also called as *measurement* data.
  - These data have a magnitude.
  - Can be expressed as number with or without unit of measurement. Eg: Height in cm, Hb in gm%, BP in mm of Hg, Weight in kg.

Quantitative data	Qualitative data
Hb level in gm%	Anemic or non anemic
Ht in cms	Tall or short
BP in mm of Hg	Hypo, normo or hypertensive
IQ scores	Idiot, genius or normal

- **Discrete / Continuous data**:

  Discrete data: Here we always get a whole number. Eg. Number of beds in hospital, Malaria cases .

  Continuous data : it can take any value possible to measure or possibility of getting fractions. Eg. Hb level, Ht, Wt.

- **Primary/ Secondary data**:

  Primary data : Obtained directly from an individual , it gives precise information .

  Secondary data : Obtained from outside source ,Eg: Data obtained from hospital records, Census.

- **Nominal/ Ordinal data**:

  Nominal data: the information or data fits into one of the categories, but the categories cannot be ordered one above another . E.g. Colour of eyes, Race, Sex.

  Ordinal data: here the categories can be ordered, but the space or class interval between two categories may not be the same. E.g.. Ranking in the class or exam

Types of measurement scales and their properties

Stevens, S. S. (1946). "On the Theory of Scales of Measurement". *Science* **103** (2684): 677–680.

		Category (Nominal) Characterizes the measured objects and / or phenomena according to the presence or absence of a certain feature.	**Ordinal (Rank)** Characterizes the measured objects and / or phenomena according to the degree of manifestation of a certain relative property in an interrupted magnitude.	**Interval** Characterizes the measured objects and / or phenomena according to the degree of manifestation of a certain absolute property in an interrupted magnitude.	**Absolute** Characterizes the measured objects and / or phenomena according to the degree of manifestation of a certain absolute property in a continuous magnitude.	**Relative** Characterizes the measured objects and / or phenomena according to the degree of change of a certain relative property in a continuous magnitude.
Logical / Mathematical operations	× ÷	✗	✗	✗	✓	✓
	+ −	✗	✗	✓	✓	✓
	< >	✗	✓	✓	✓	✓
	= ≠	✓	✓	✓	✓	✓
Examples: Dichotomous and non-dichotomous Variable name (possible values)		*Dichotomous:* Gender (male or female) *Non-dichotomous:* Nationality (Bulgaria / Romania / others)	*Dichotomous:* Health status (healthy or sick), Truth (True or false), Beauty (beautiful or ugly) *Non-dichotomous:* Opinion ('completely agree'/ 'rather agree' / 'rather disagree' / 'completely disagree')	Date ('From 1878 to 1945' / 'From 1945 to 1989' / 'After 1989') Age ('Under 18' / 'from 18 to 25' / '25 to 35 years old' / 'over 35 years old') Temperature ('Below 0 °' / 'from 0 ° C to 20 ° C') 'From 20 ° C to 40 ° C' / 'above 40 ° C')	Age (0, 1, 2, 3, ... 100 y.) Temperature (10°, 11°, 12°, 13°, ... 40°)	Temperature change (- 6°, +2°) Annual return (- 31%, +12%)
Measure for central tendency		Mode	Median Mode	Arithmetic mean Median Mode	Geometric mean Arithmetic mean Median Mode	Geometric mean Arithmetic mean Median Mode
Qualitative or quantitative		Qualitative	Qualitative	Quantitative	Quantitative	Quantitative

Data types

1) Solve the quiz:
https://www.med.soton.ac.uk/stats_eLearning/typesofdataquiz/index.html

2) Make a print screen with your final score

3) Submit it here:
https://forms.gle/7YRmC4CehbdGBBby7

Data types by storage (programming)

type	*set of values*	*common operators*	*sample literal values*
`int`	integers	`+ - * / %`	`99 12 2147483647`
`double`	floating-point numbers	`+ - * /`	`3.14 2.5 6.022e23`
`boolean`	boolean values	`&& || !`	`true false`
`char`	characters		`'A' '1' '%' '\n'`
`String`	sequences of characters	`+`	`"AB" "Hello" "2.5"`

Binary Systems

https://www.youtube.com/watch?v=LpuPe81bc2w

https://www.youtube.com/watch?v=b7pOcU1xMks

Binary representation of integers

	128	64	32	16	8	4	2	1
8 bit binary digit	1	0	1	1	0	0	0	1
	128 + 32 + 16 + 1 = 177							

Signed byte (8 bit) integer

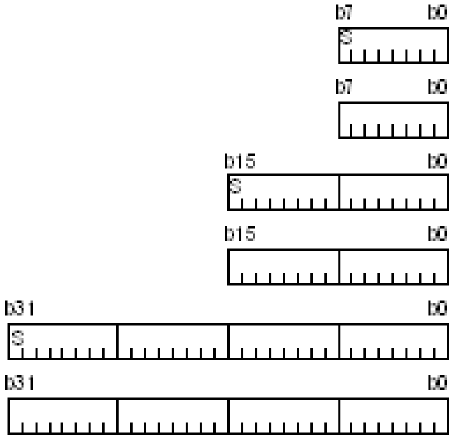
Unsigned byte (8 bit) integer

Signed word (16 bit) integer

Unsigned word (16 bit) integer

Signed long word (32 bit) integer

Unsigned long word (32 bit) integer



S: Sign bit

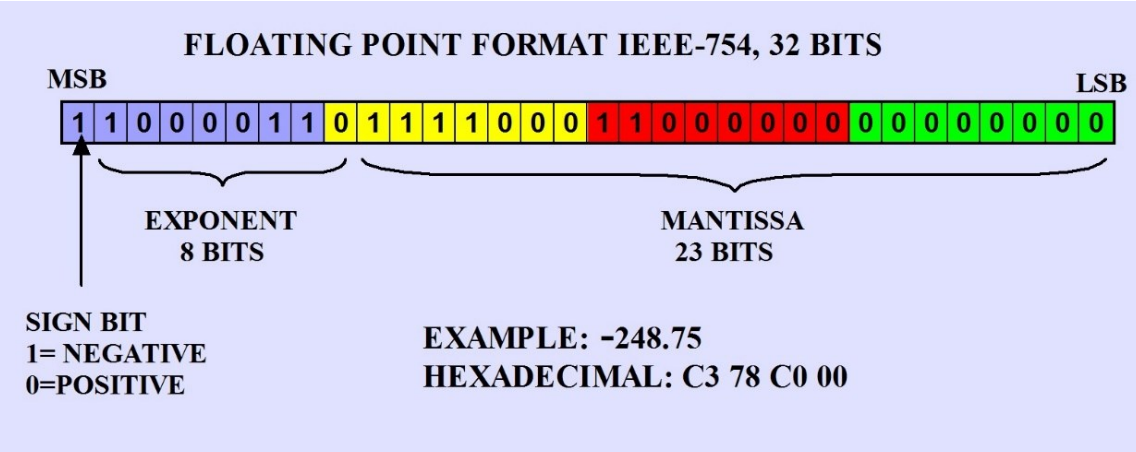
3.4 Data Type Definitions

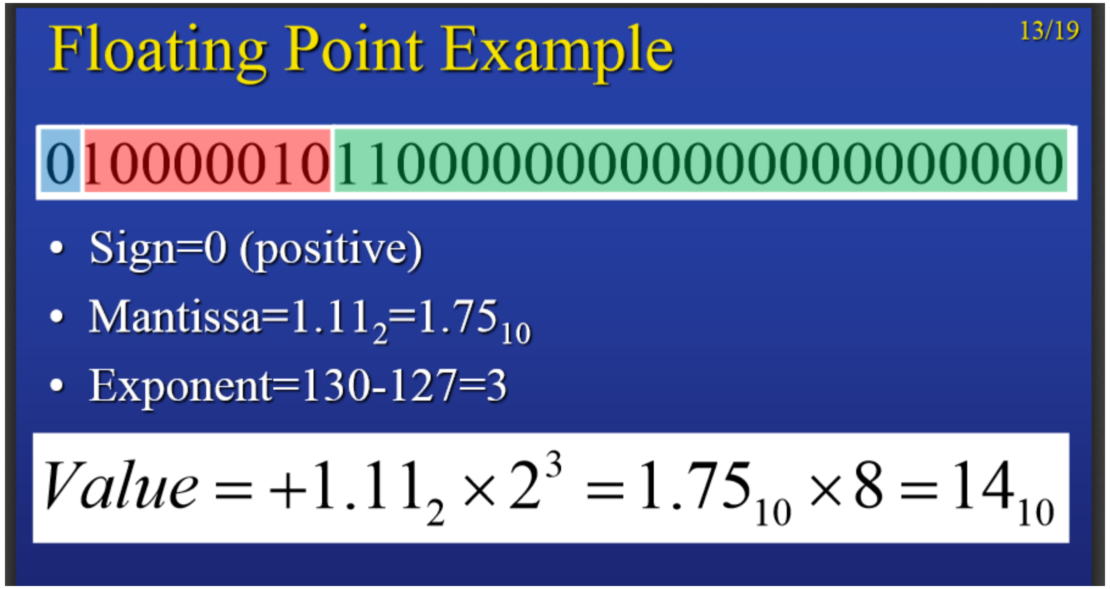
Below is a table of all used data types.

Name	Data type	Size bits	Size bytes	Range
char, __int8	signed integer	8	1	– 128 ... 127
BYTE	unsigned integer	8	1	0 ... 256
short	signed integer	16	2	– 32'768 ... 32'767
WORD	unsigned integer	16	2	0 ... 65'535
long	signed integer	32	4	– 2'147'483'648 ... 2'147'483'647
DWORD	unsigned integer	32	4	0 ... 4'294'967'295
BOOL	signed integer	32	4	TRUE = 1 FALSE = 0
HANDLE	pointer to an object	32	4	0 ... 4'294'967'295

Table 2: Data type definitions

Binary representation of floating-point numbers





Example IEEE-decimal conversion

- Let's find the decimal value of the following IEEE number.

  1 01111100 11000000000000000000000

- First convert each individual field to decimal.
  - The sign bit s is 1.
  - The e field contains 01111100 = 124_{10}.
  - The mantissa is 0.11000… = 0.75_{10}.
- Then just plug these decimal values of s, e and f into our formula.

$$(1 - 2s) * (1 + f) * 2^{e\text{-bias}}$$

- This gives us $(1 - 2) * (1 + 0.75) * 2^{124-127} = (-1.75 * 2^{-3}) = -0.21875$.

Binary Game

1) Play as long as possible
https://basaga.org/basaga_files/binary_game/binary_game.html

2) Make a print screen with your final score

3) Submit it here (incl. your final score) :
https://forms.gle/7YRmC4CehbdGBBby7

	Sepal.Length	Sepal.Width
1	5.1	3.5
2	4.9	3.0
3	4.7	3.2

Tabular Data

- ↳Name: Robin
    - ↳ Species: Hedgehog
    - ↳ Owner: Justice Smith
        - ↳ Address: 1234 Main St.
        - ↳ Phone #: 123-4567
- ↳Name: Bunny
    - ↳ Species: Rabbit
        - ↳ Breed: Holland Lop
    - ↳ Color: Brown and white

Hierarchical Data

石室诗士施氏，嗜狮，誓食十狮。氏时时适市视狮。十时，适十狮适市。是时，适施氏适市。氏视是十狮，恃矢势，使是十狮逝世。氏拾是十狮尸，适石室。石室湿，氏使侍拭石室。石室拭，氏始试食是十狮尸。食时，始识是十狮，实十石狮尸。试释是事。

Raw Text

<table>
<tr><th>Statistics</th><th>Programming / Storage</th><th>Complexity</th></tr>
<tr><td>interval scale</td><td rowspan="2">Floating-point</td><td rowspan="11">Basic</td></tr>
<tr><td>ratio scale</td></tr>
<tr><td>count data</td><td rowspan="3">Integer</td></tr>
<tr><td>Ranking</td></tr>
<tr><td>Rating data</td></tr>
<tr><td>binary data</td><td>Boolean</td></tr>
<tr><td rowspan="4">categorical data</td><td>String</td></tr>
<tr><td>Integer (enumerated)</td></tr>
<tr><td>Boolean (dummied)</td></tr>
<tr><td>Character (abbrivieted)</td></tr>
<tr><td>Text</td><td>String</td></tr>
<tr><td>vector</td><td rowspan="2">List or Array</td><td rowspan="10">Arrays</td></tr>
<tr><td>Sequence data</td></tr>
<tr><td>matrix</td><td>two-dimensional array</td></tr>
<tr><td>Tensor</td><td>n-dimensional array</td></tr>
<tr><td rowspan="2">Image</td><td>2-dimensional array (compressed)</td></tr>
<tr><td>3-dimensional array (raw)</td></tr>
<tr><td rowspan="2">Audio</td><td>1-dimensional array (compressed)</td></tr>
<tr><td>2-dimensional array (raw)</td></tr>
<tr><td rowspan="2">Video</td><td>n-dimensional array (visual stream)</td></tr>
<tr><td>2-dimensional array (audio stream)</td></tr>
<tr><td>tree</td><td>tree (data structure)</td><td>Hierarchical</td></tr>
</table>

Data processing

by Angel Marchev, Jr.

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Reasons for inaccurate data

- Data collection instruments may be faulty
- Human or computer errors occurring at data entry
- Users may purposely submit incorrect data for mandatory fields when they don't want to share personal information
- Technology limitations such as buffer size
- Incorrect data may also result from inconsistencies in naming conventions or inconsistent formats
- Duplicate tuples also require cleaning

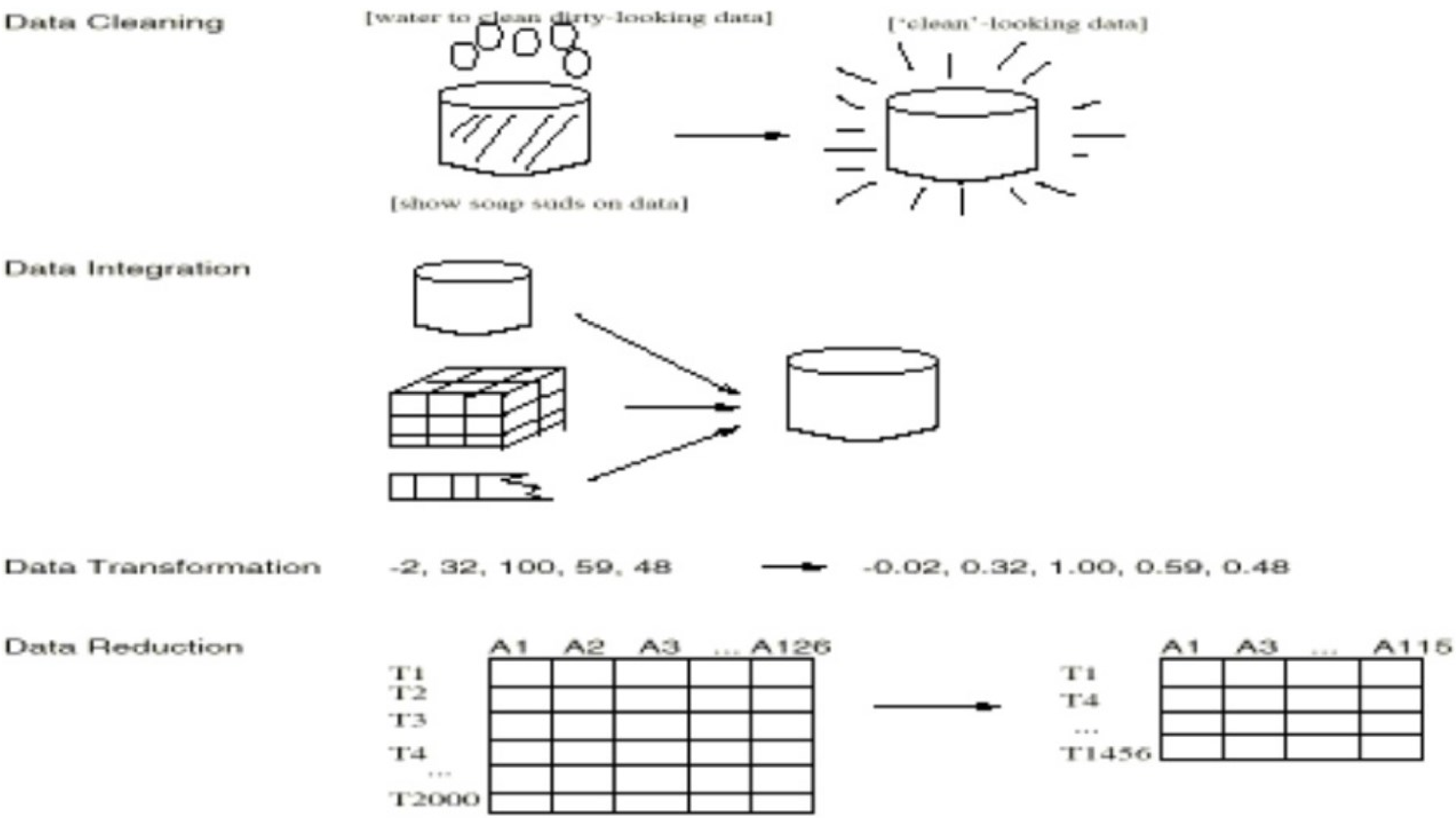
Reasons for incomplete data

- Attributes of interest may not be available
- Other data may not be included as it was not considered imp at the time of entry
- Relevant data may not be recorded due to misunderstanding or equipment malfunctions
- Inconsistent data may be deleted
- Data history or modifications may be overlooked
- Missing data

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Forms of Data Preprocessing



Why Is Data Dirty?

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Data Cleaning

- Data in the Real World Is Dirty :- Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data
  - **noisy**: containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - **inconsistent**: containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - **Intentional**(e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Data Cleaning

- Importance
  - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
  - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

Noisy Data

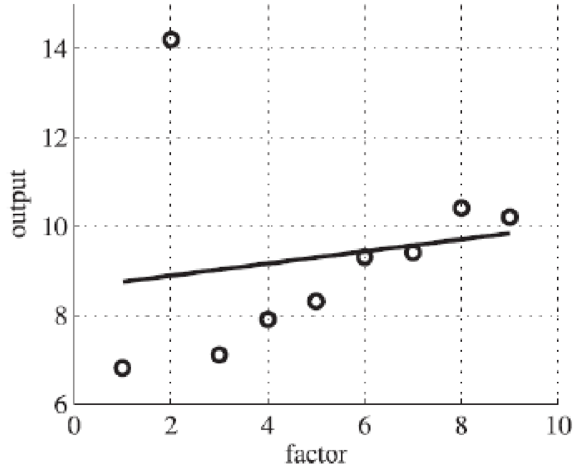
- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

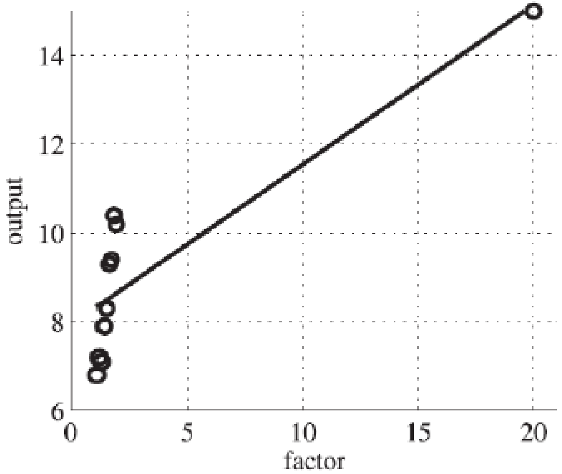
How to Handle Noisy Data?

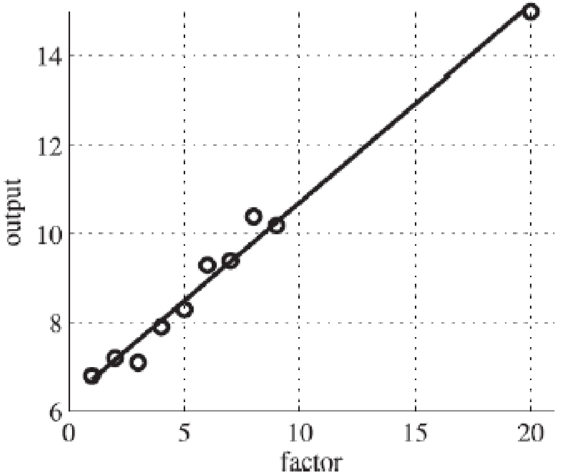
- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Outlier Analysis by Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Preparation / Data Cleaning

- **Outliers**
  - Effect on the model
  - Wrong conclusions



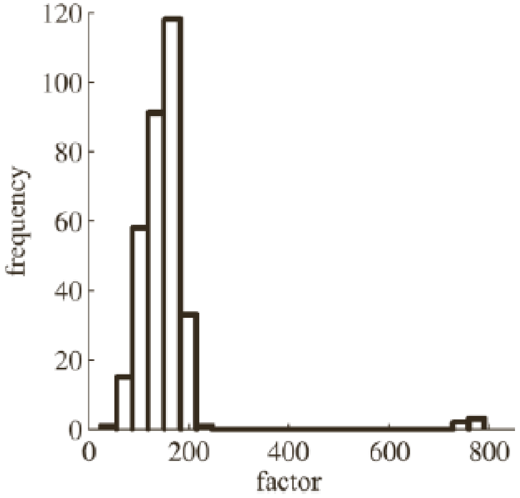




- **Outliers Detection & Manipulation**
  - Capping

replace x ≥ p_{95}
with p_{95}

p_{95} – 95-th percentile



- Time-series

Low frequency component:

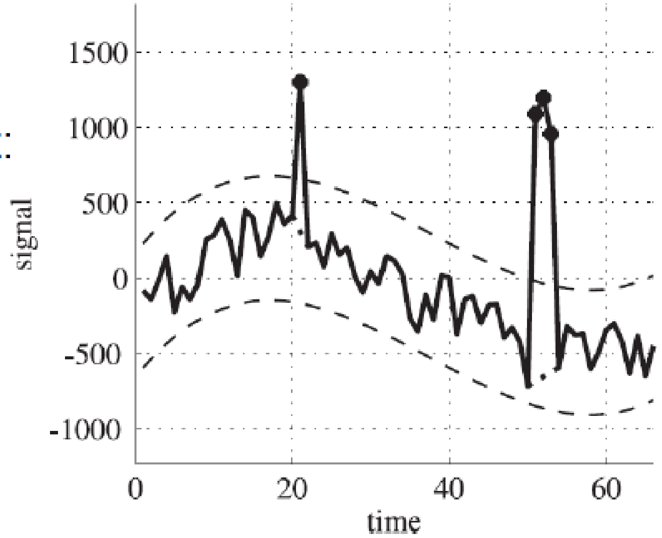
$x_{t,k}$ = low-pas-filter(x_k)

$\tilde{x}_k = x_k - x_{t,k}$

- Sleeve

$x_{l,k} = x_{t,k} - n\sigma_{\tilde{x}}$

$x_{u,k} = x_{t,k} + n\sigma_{\tilde{x}}$



Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Data Preparation / Pre-processing

– Encoding
  - categorical → numeric
  - Dummy variables
  - Dependent mean (y – numeric)

    $\tilde{\varphi}_i = \bar{y}_i = 1/N_i \sum_{k, xk = xi} y_k$

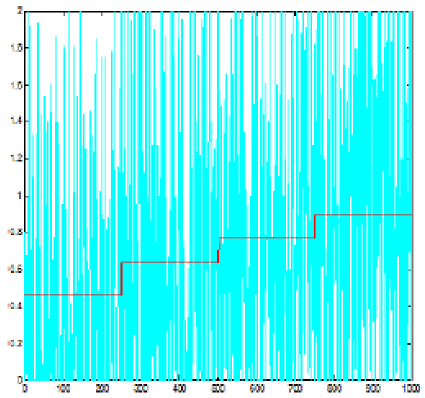
    x_i – i-th unique value
  - Weight of evidence (y – binary)

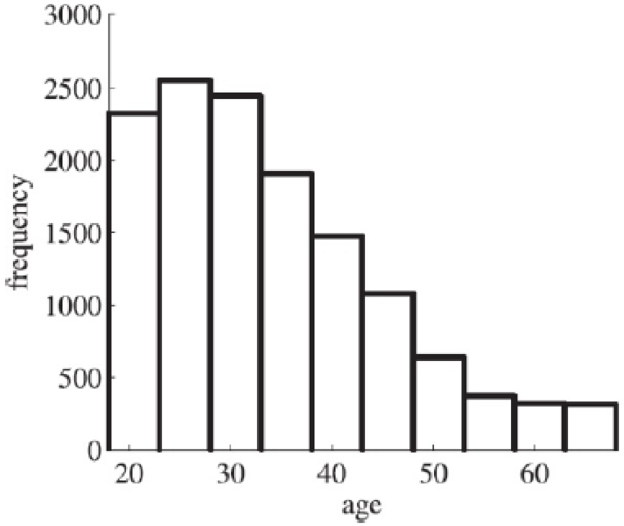
    $\tilde{\varphi}_i = \mathrm{WoE}_i = \log((n_{i,1}/N_1) / (n_{i,2}/N_2))$

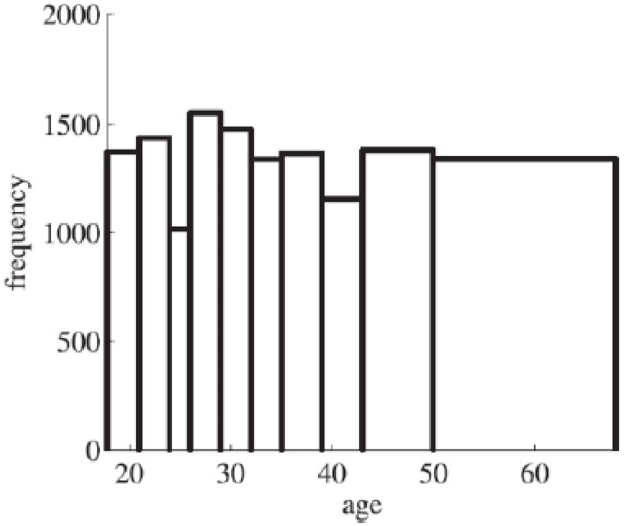
week	promotion type	dv1	dv2	dv3
1	promo 2	0	1	0
2	promo 1	1	0	0
3	promo 1	1	0	0
4	promo 3	0	0	1
5	promo 3	0	0	1
6	promo 1	1	0	0
7	promo 1	1	0	0
8	promo 2	0	1	0
9	promo 2	0	1	0
10	promo 3	0	0	1

Data Preparation / Pre-processing

– Binning
  - numeric / categorical → categorical
  - Applications:
    - uncertainty reduction
    - finding на relations
    - account for business logic
    - avoid outliers effect
  - Approaches
    - unsupervised binning
      - equal number of records
      - equal ranges
    - supervised binning
      - Chi Square, Entropy Gain, Gini…







Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id $\equiv$ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

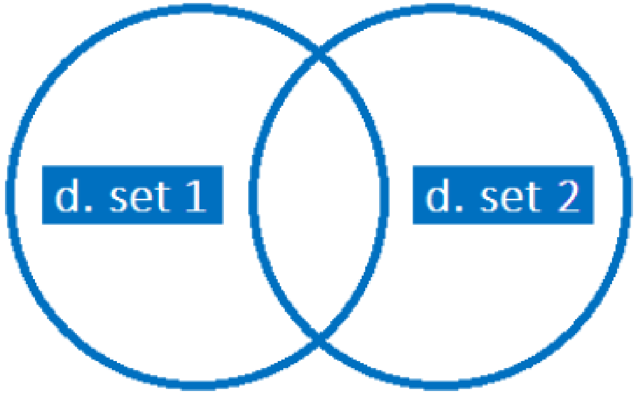
Data Preparation / Data Manipulation

- Load Data

```
df1, df2 ← load('data_DManip/data.Rdata')
```

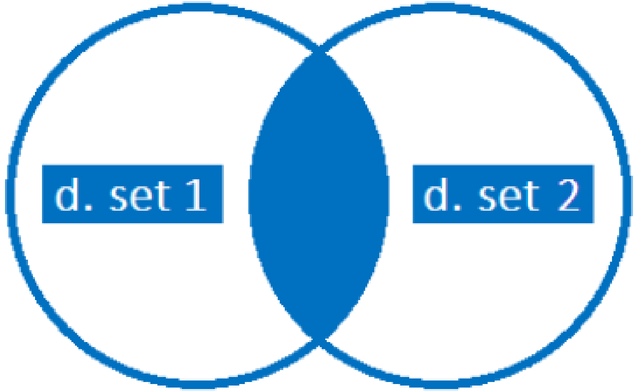
- data.table

```
dt1 <- data.table(df1, key = 'id')
dt2 <- data.table(df2, key = 'ucc')
```



- Inner Join

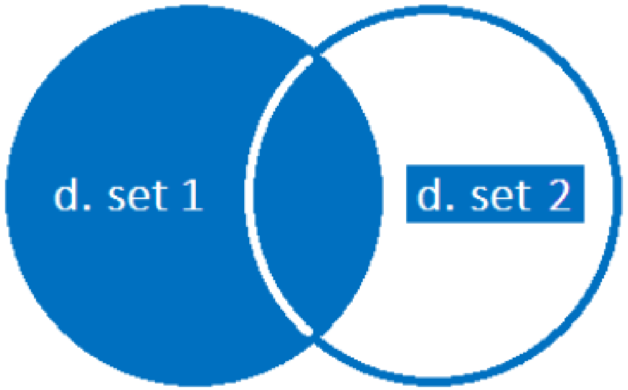
```
dt <- merge(dt1, dt2, by.x = 'id', by.y = 'ucc')
```



Data Preparation / Data Manipulation

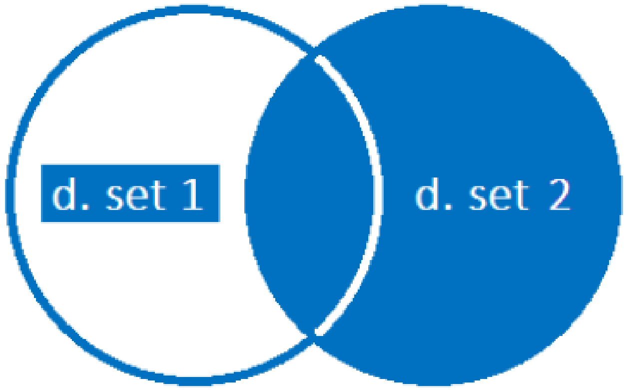
- Left Outer Join

```
dt <- merge(dt1, dt2, by.x = 'id', by.y = 'ucc', all.x = T)
# alternative
dt <- dt2[dt1]
```



- Right Outer Join

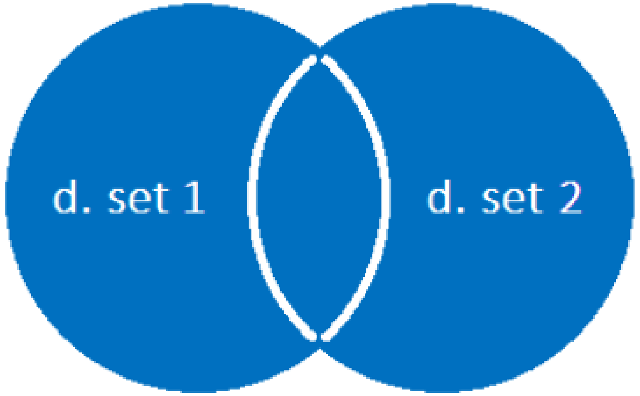
```
dt <- merge(dt1, dt2, by.x = 'id', by.y = 'ucc', all.y = T)
# alternative
dt <- dt1[dt2]
```



Data Preparation / Data Manipulation

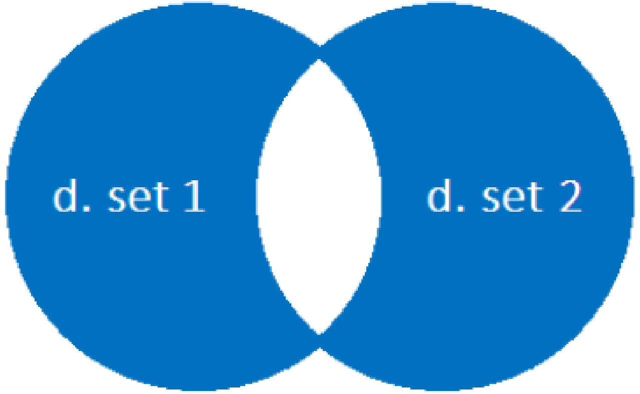
– Full Outer Join

```
dt <- merge(dt1, dt2, by.x = 'id', by.y = 'ucc', all = T)
```



– NOT Inner Join

```
dt <- merge(dt1, dt2, by.x = 'id', by.y = 'ucc', all = T)
dt[is.na(name) | is.na(ctype)]
```



– Mapping

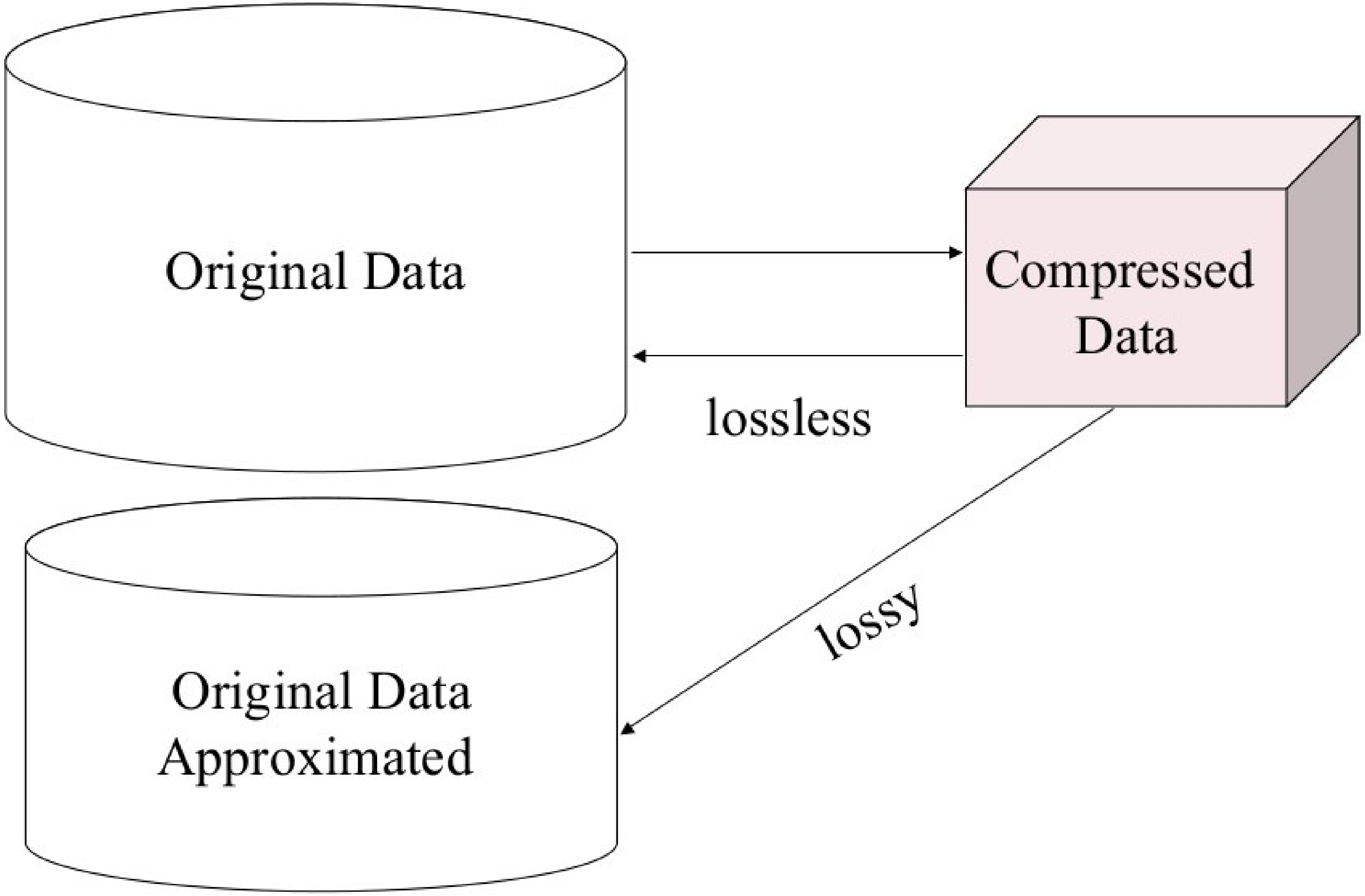
Data Reduction Strategies

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation:
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation

Data Compression

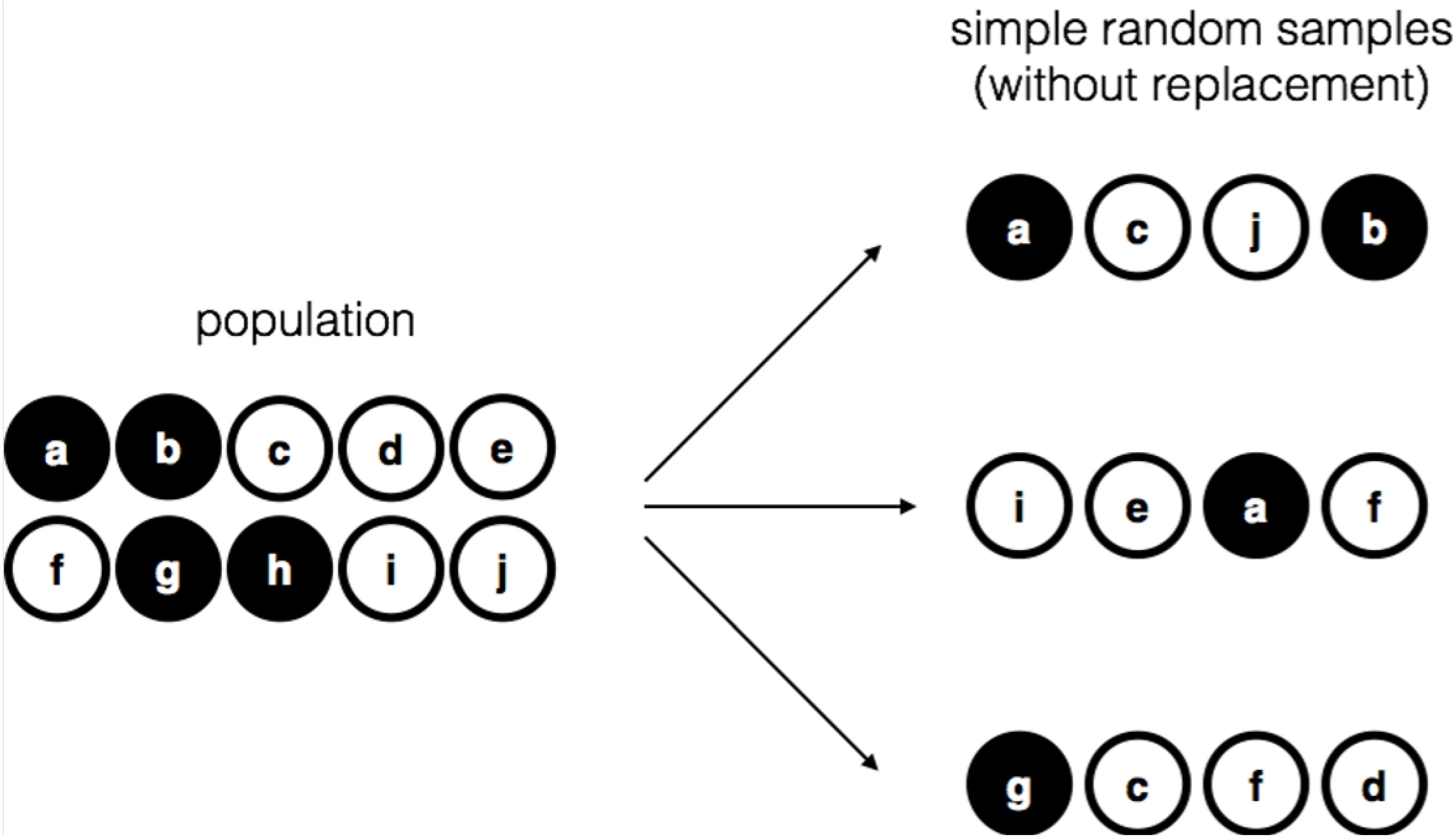
- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time

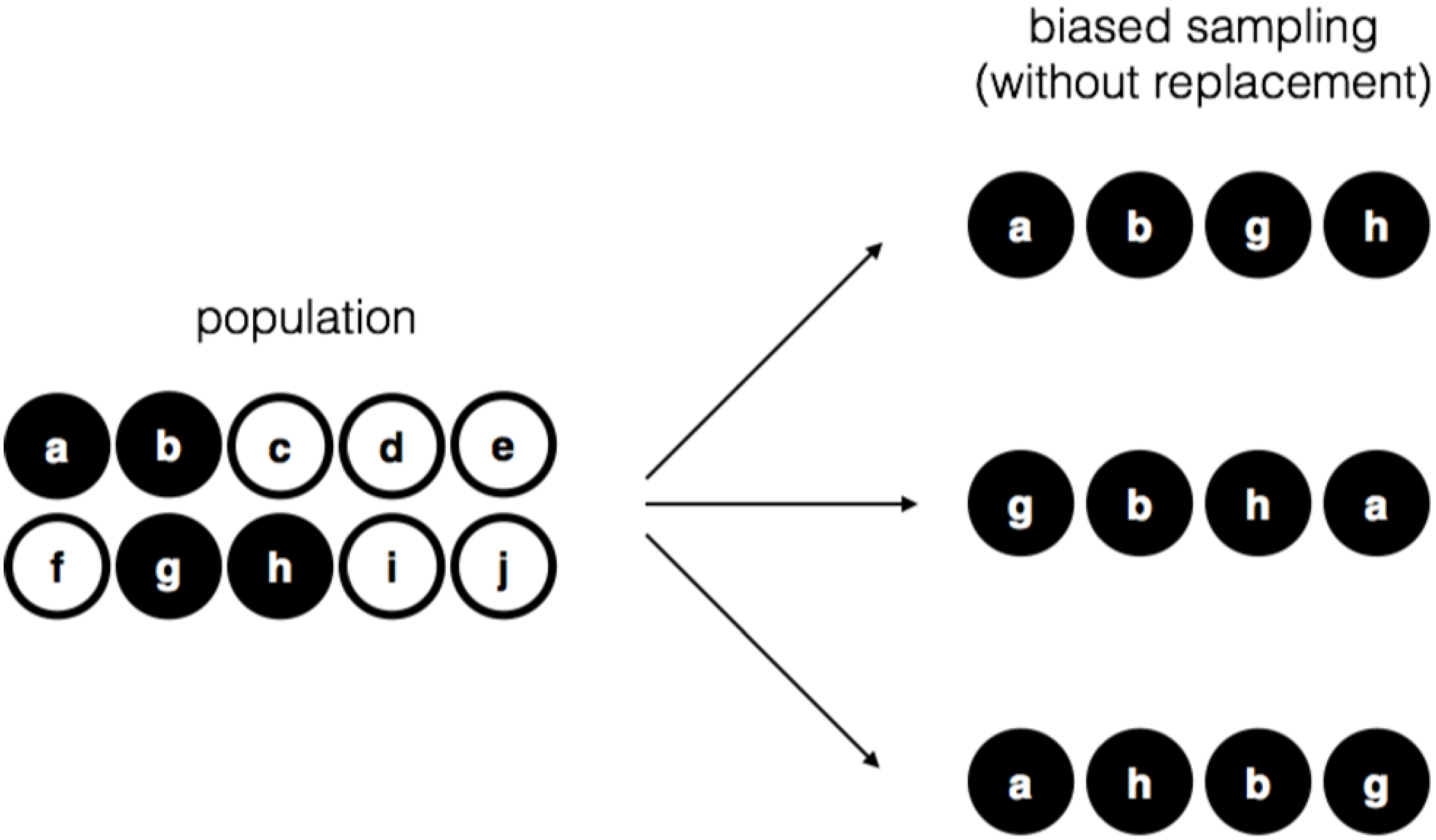
Data Compression

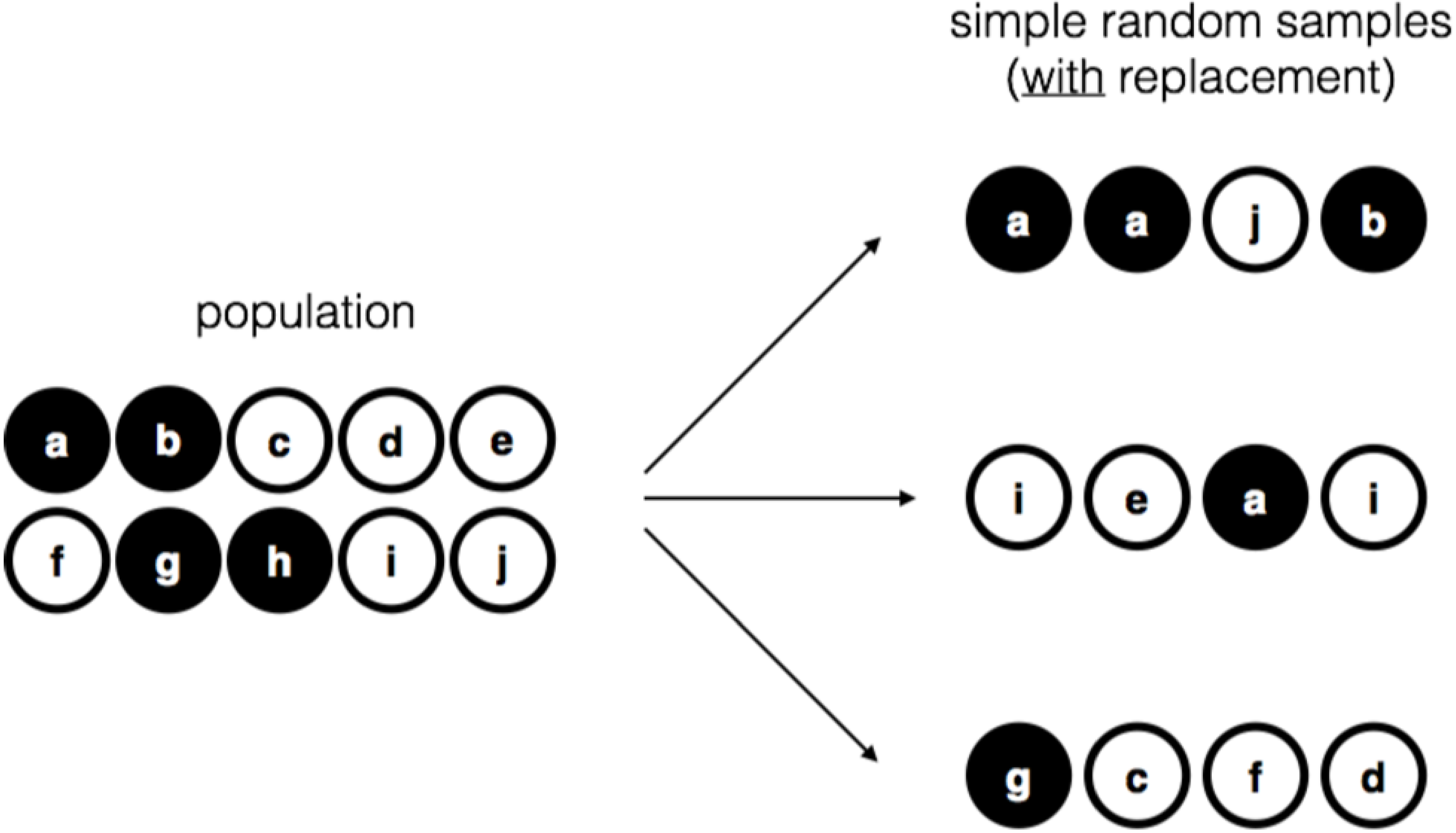


Data Reduction Method (4): Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)


simple random samples
(without replacement)
population
a
b
c
d
e
f
g
h
i
j
a
c
j
b
i
e
a
f
g
c
f
d

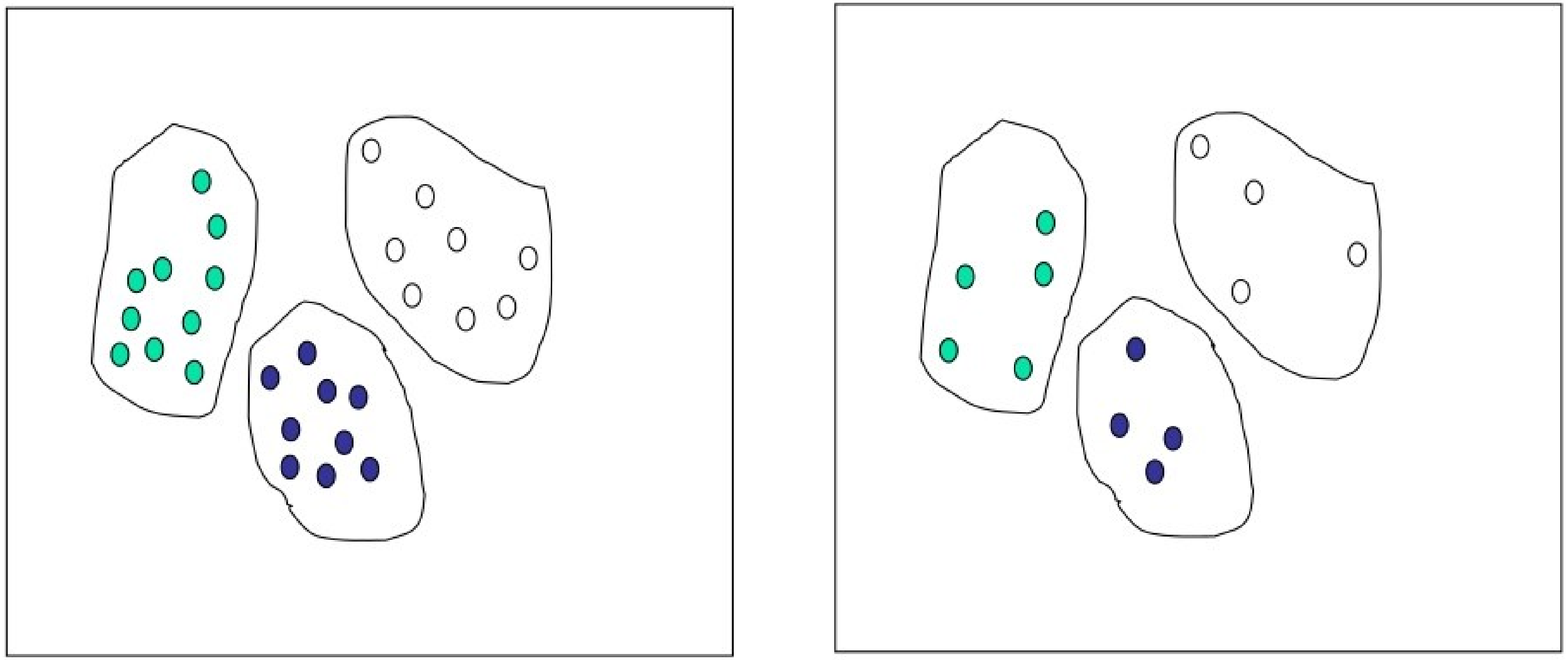

biased sampling
(without replacement)
population
a
b
c
d
e
f
g
h
i
j
a
b
g
h
g
b
h
a
a
h
b
g


simple random samples
(with replacement)
population
a
b
c
d
e
f
g
h
i
j
a a j b
i e a i
g c f d

Sampling: Cluster or Stratified Sampling

Raw Data

Cluster/Stratified Sample



Sampling size

$$S = \frac{Z^2.p.(1-p).N}{N.c^2 + Z^2.p.(1-p)}$$

S - *размер на извадката*

Z - *стойност Z (1.96 при допускане за 95% доверителност)*

p - *вероятност за съвкупността (допуснете че е 0.5)*

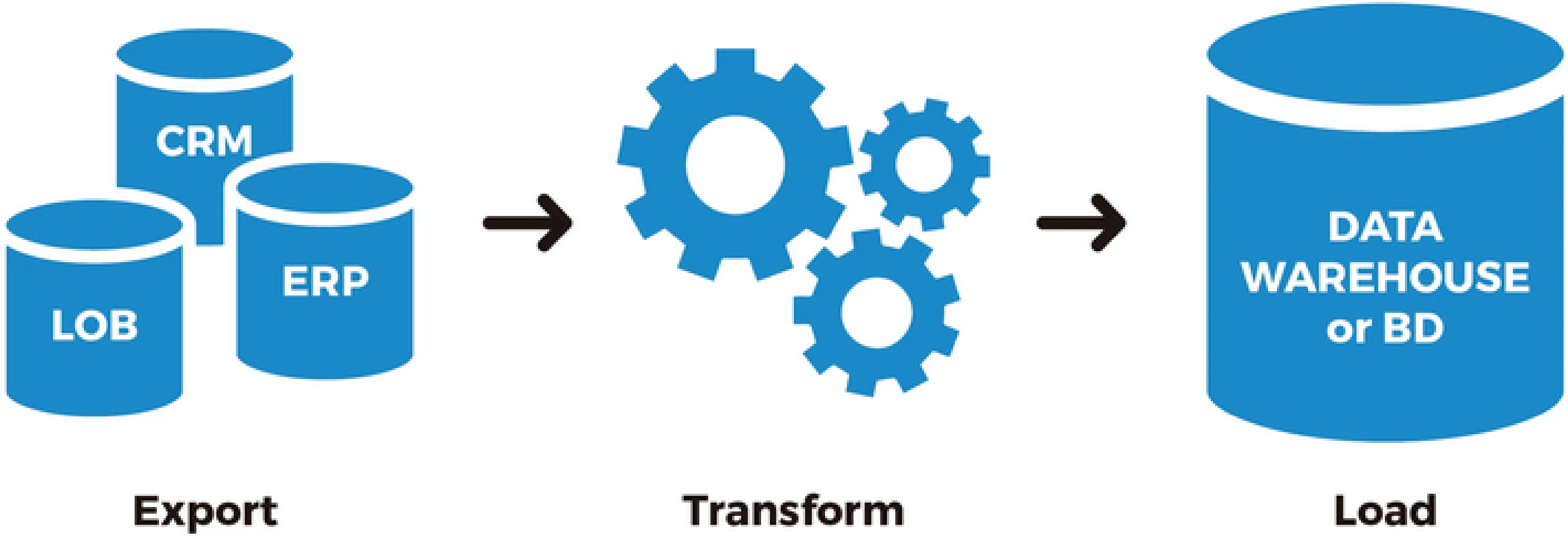
c - *допустима грешка (нормално 5%)*

N - *размер на генералната съвкупност*

Required Sample Size†

Population Size	Confidence = 95%				Confidence = 99%			
	Margin of Error				Margin of Error			
	5.0%	3.5%	2.5%	1.0%	5.0%	3.5%	2.5%	1.0%
10	10	10	10	10	10	10	10	10
20	19	20	20	20	19	20	20	20
30	28	29	29	30	29	29	30	30
50	44	47	48	50	47	48	49	50
75	63	69	72	74	67	71	73	75
100	80	89	94	99	87	93	96	99
150	108	126	137	148	122	135	142	149
200	132	160	177	196	154	174	186	198
250	152	190	215	244	182	211	229	246
300	169	217	251	291	207	246	270	295
400	196	265	318	384	250	309	348	391
500	217	306	377	475	285	365	421	485
600	234	340	432	565	315	416	490	579
700	248	370	481	653	341	462	554	672
800	260	396	526	739	363	503	615	763
1,000	278	440	606	906	399	575	727	943
1,200	291	474	674	1067	427	636	827	1119
1,500	306	515	759	1297	460	712	959	1376
2,000	322	563	869	1655	498	808	1141	1785
2,500	333	597	952	1984	524	879	1288	2173
3,500	346	641	1068	2565	558	977	1510	2890
5,000	357	678	1176	3288	586	1066	1734	3842
7,500	365	710	1275	4211	610	1147	1960	5165
10,000	370	727	1332	4899	622	1193	2098	6239
25,000	378	760	1448	6939	646	1285	2399	9972
50,000	381	772	1491	8056	655	1318	2520	12455
75,000	382	776	1506	8514	658	1330	2563	13583
100,000	383	778	1513	8762	659	1336	2585	14227
250,000	384	782	1527	9248	662	1347	2626	15555
500,000	384	783	1532	9423	663	1350	2640	16055
1,000,000	384	783	1534	9512	663	1352	2647	16317
2,500,000	384	784	1536	9567	663	1353	2651	16478
10,000,000	384	784	1536	9594	663	1354	2653	16560
100,000,000	384	784	1537	9603	663	1354	2654	16584
300,000,000	384	784	1537	9603	663	1354	2654	16586

What Is an ETL Process?



Briefly explained, an ETL process (Extract, Transform, Load) is a system that allows organizations to **move data from multiple sources** (ERP, CRM, Excel, Open Data, Internet Of Things, Social Networks ...) to integrate them into a single place, which could be a database, a data warehouse, and so on.

Data cleaning task

1) Do the task

2) Save the file

3) Submit it here (incl. your answers) :
https://forms.gle/7YRmC4CehbdGBBby7