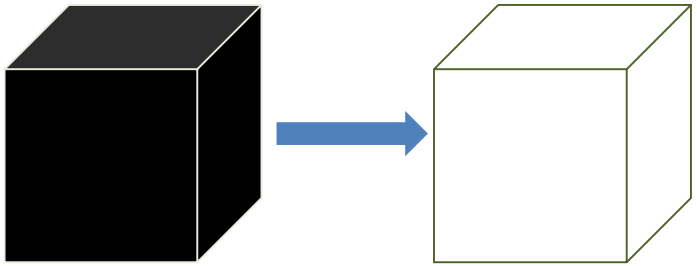
At the end of this session:

Everyone has had a good time and will have better familiarity with Data Science

Most will have a decent understanding of Data Science, its underlying principles and significance

Some Will be inspired and be able to identify practical application of data science in their own work



BEYOND

BINGO

This presentation contains gross oversimplifications and generalisations, in general

interdisciplinary field

extract knowledge and insights from :

- Structured data
- Unstructured data

Data Science

Scientific methods:

- Processes
- Algorithms
- Systems

"concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.

"fourth paradigm" of science:

1. Empirical
2. Theoretical
3. Computational
4. Data-driven

"everything about science is changing because of the impact of information technology" and the **data deluge**

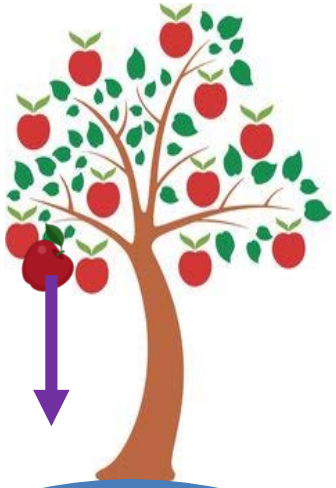
Pre-Science Paradigm : natural phenomena explained by deity

Empirical Paradigm:

Observed natural phenomena confirmed by theory

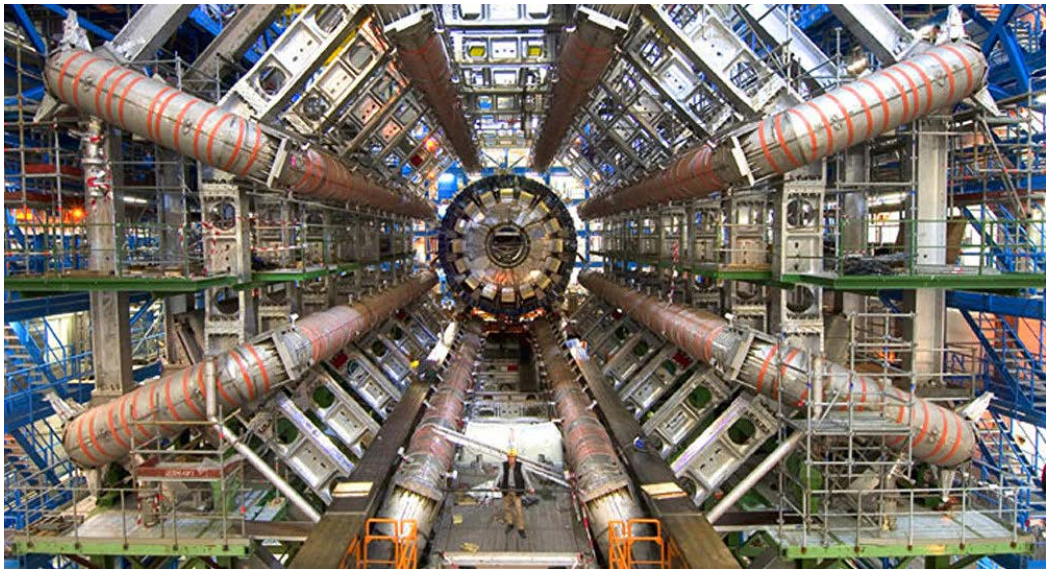
‘Discovery driven by observation’

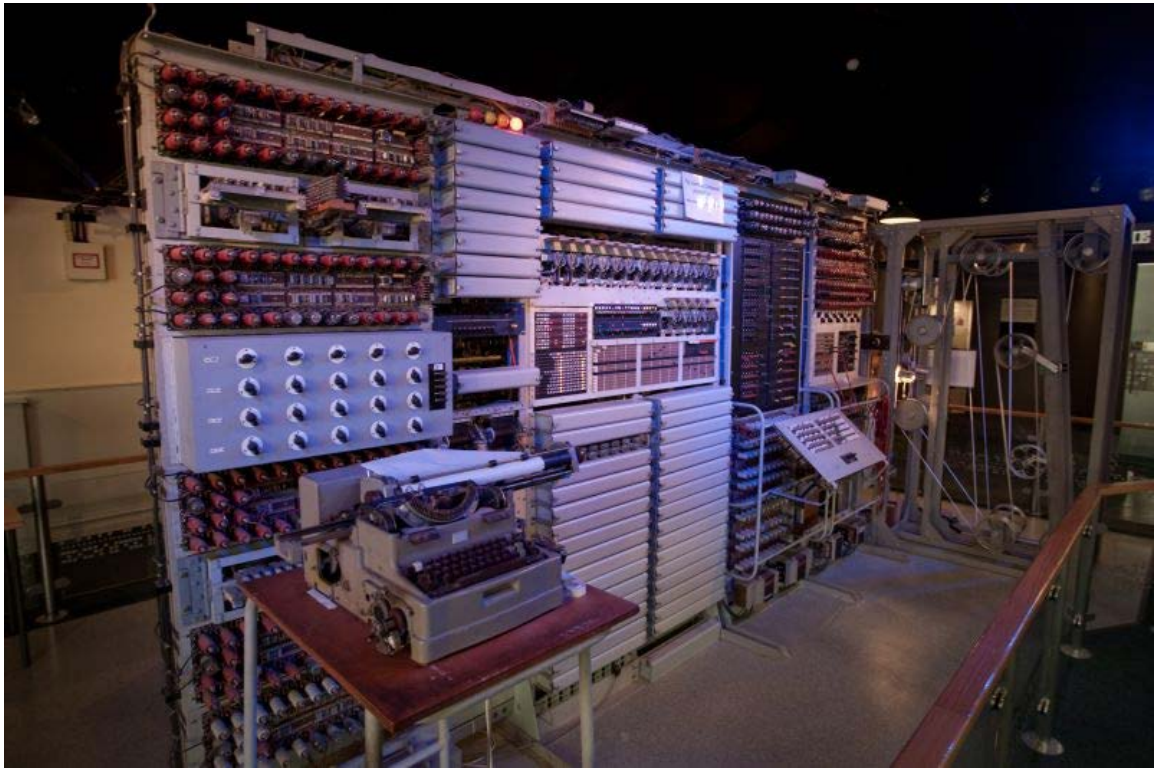
?



Theoretical Paradigm: theory confirmed by observed natural phenomena

‘Discovery driven by theory ’



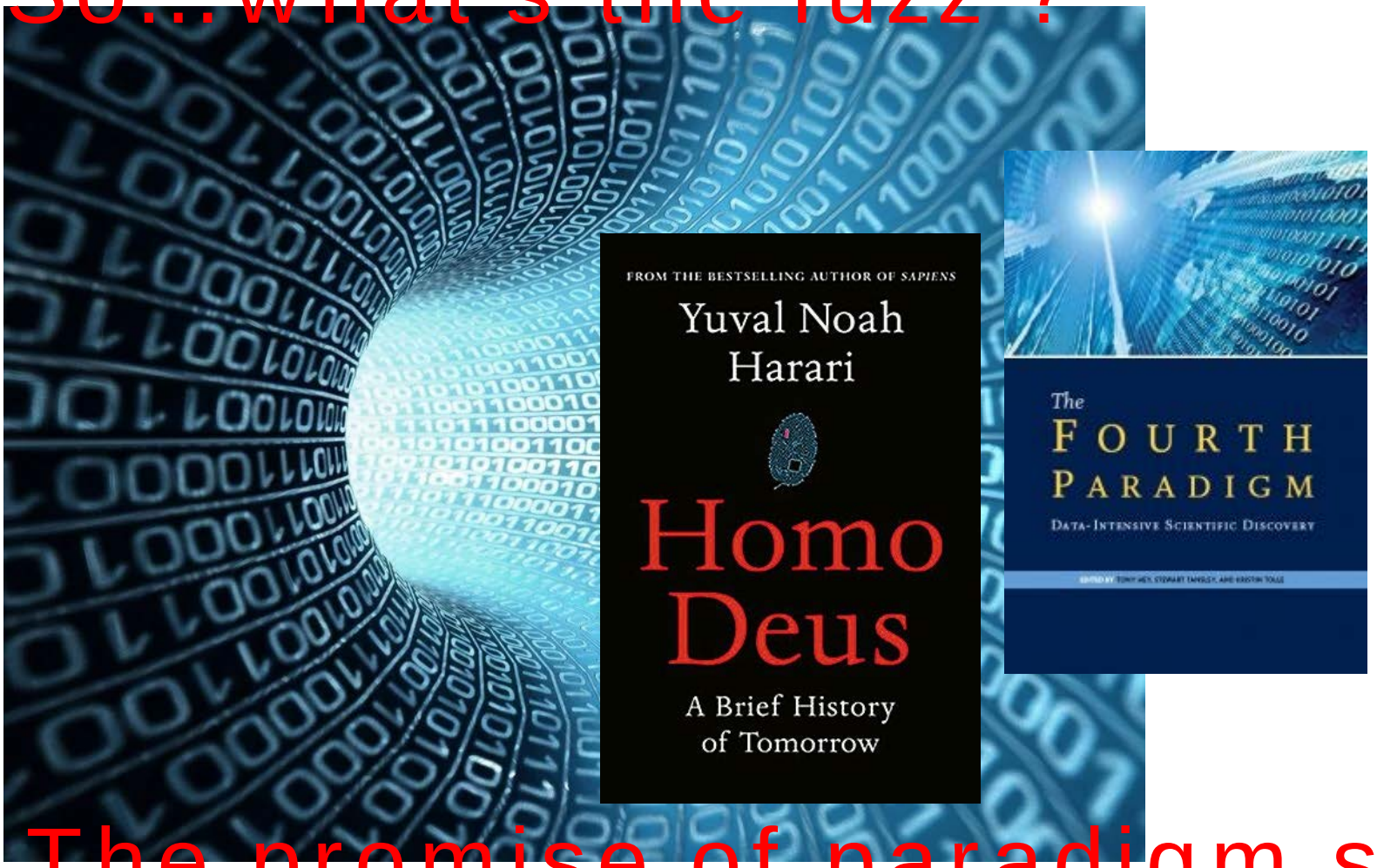


Computational Paradigm:

observed natural phenomena explained by computational simulation of reality

‘Discovery driven by computation’

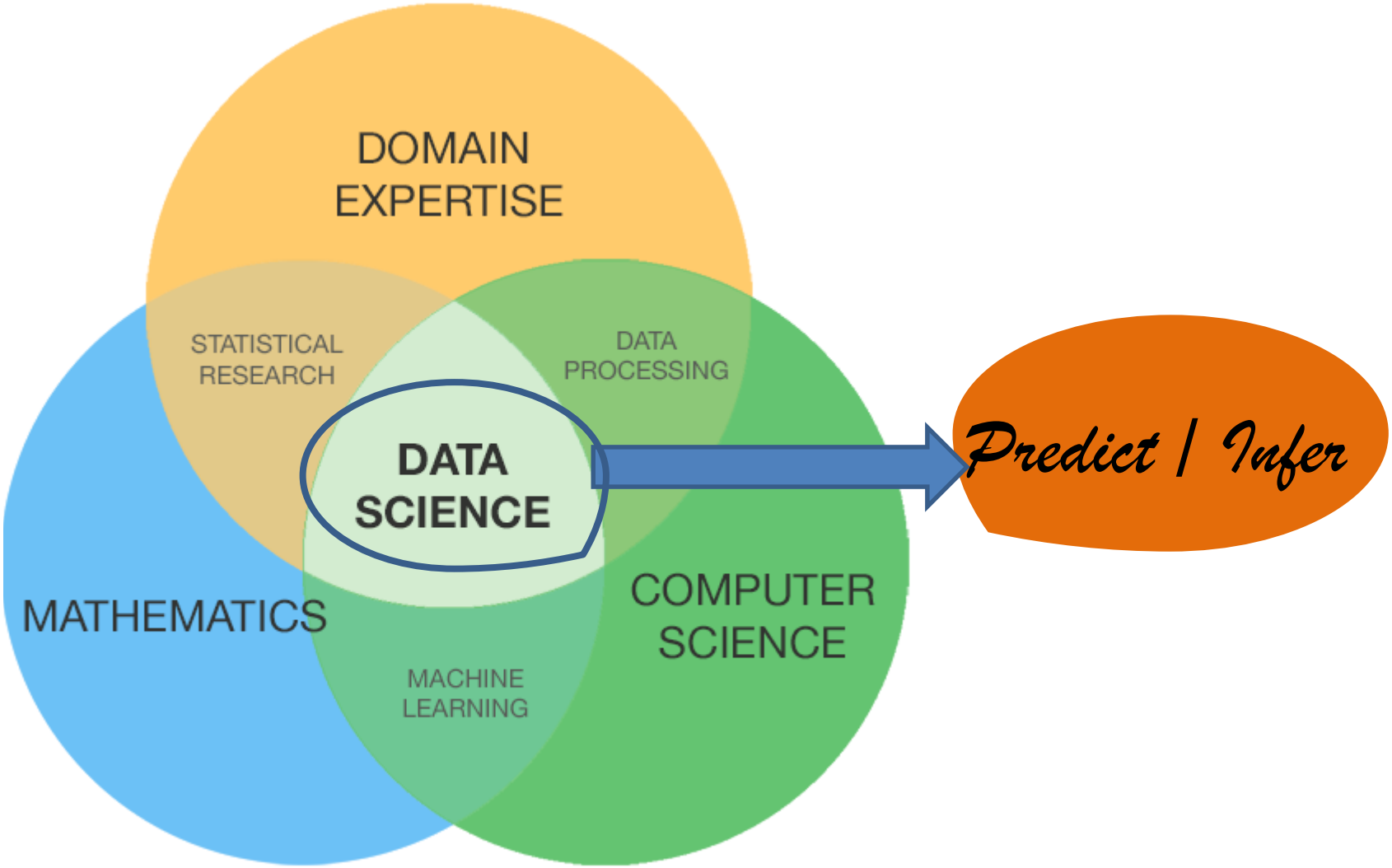
So…what's the fuzz ?



The promise of paradigm shift

Data Paradigm: *unobserved* complex natural phenomena *discovered* from computational simulation of reality

'Discovery driven by data '


DOMAIN
EXPERTISE
STATISTICAL
RESEARCH
DATA
PROCESSING
DATA
SCIENCE
Predict | Infer
MATHEMATICS
COMPUTER
SCIENCE
MACHINE
LEARNING

autocomplete

yahoo autocomp

disable autocomp

internet explorer a

BEREKEN ZELF - WAT IS MIJN HUIS WAARD?

De waarde van een huis fluctueert continu. Wegwijs geeft antwoord op de vraag *Wat is mijn huis waard* door cijfers van het CBS en Kadaster slim te combineren. Vul je gegevens in en bekijk wat je huis nu waard is en hoe de waarde zich heeft ontwikkeld.

Elke maand een update over de waarde van je huis? (1) Doe de berekening. (2) Vul daarna je e-mailadres in.

Je gegevens

Koopprijs huis € 300.000

Wanneer heb jij je huis gekocht mei 2008

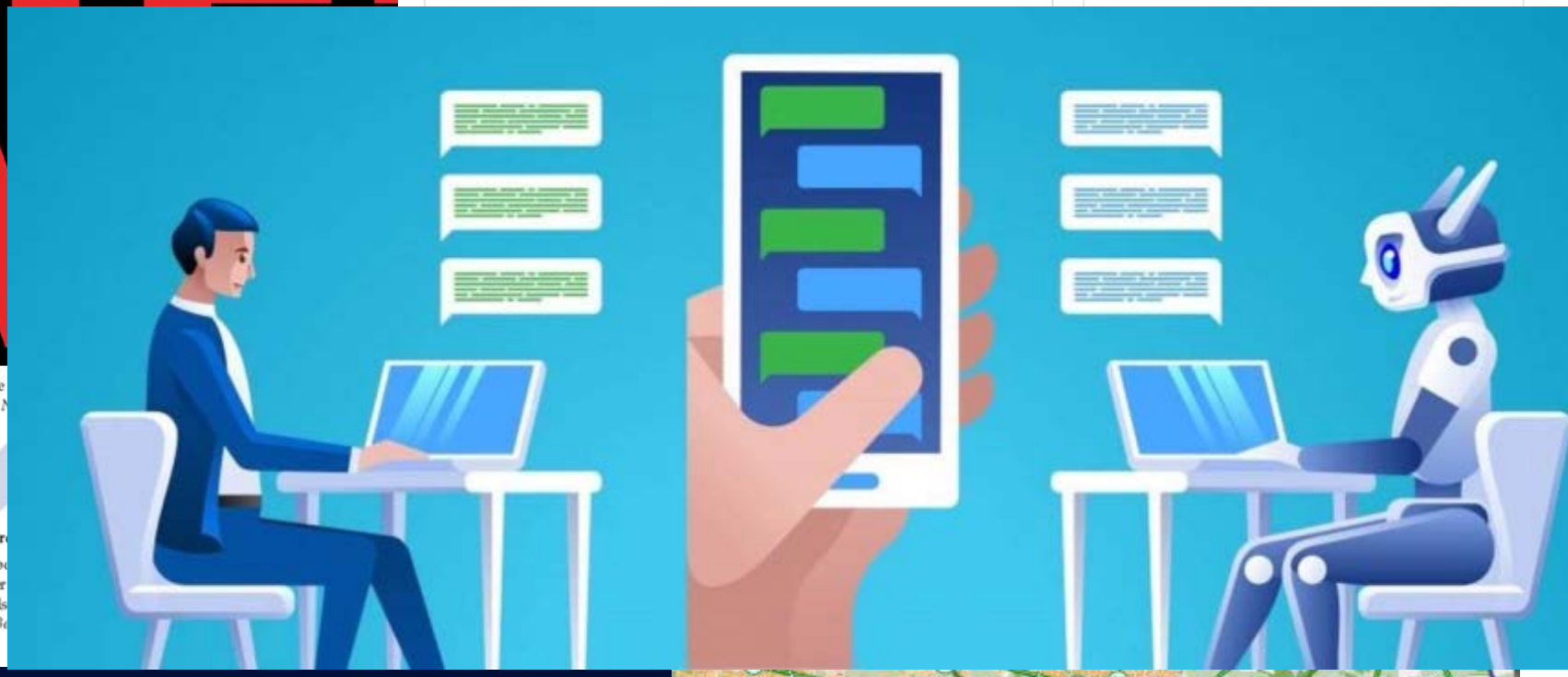
In wat voor huis woon je? Tussenwoning

CHECK >

Waarde van jouw huis

€313.081

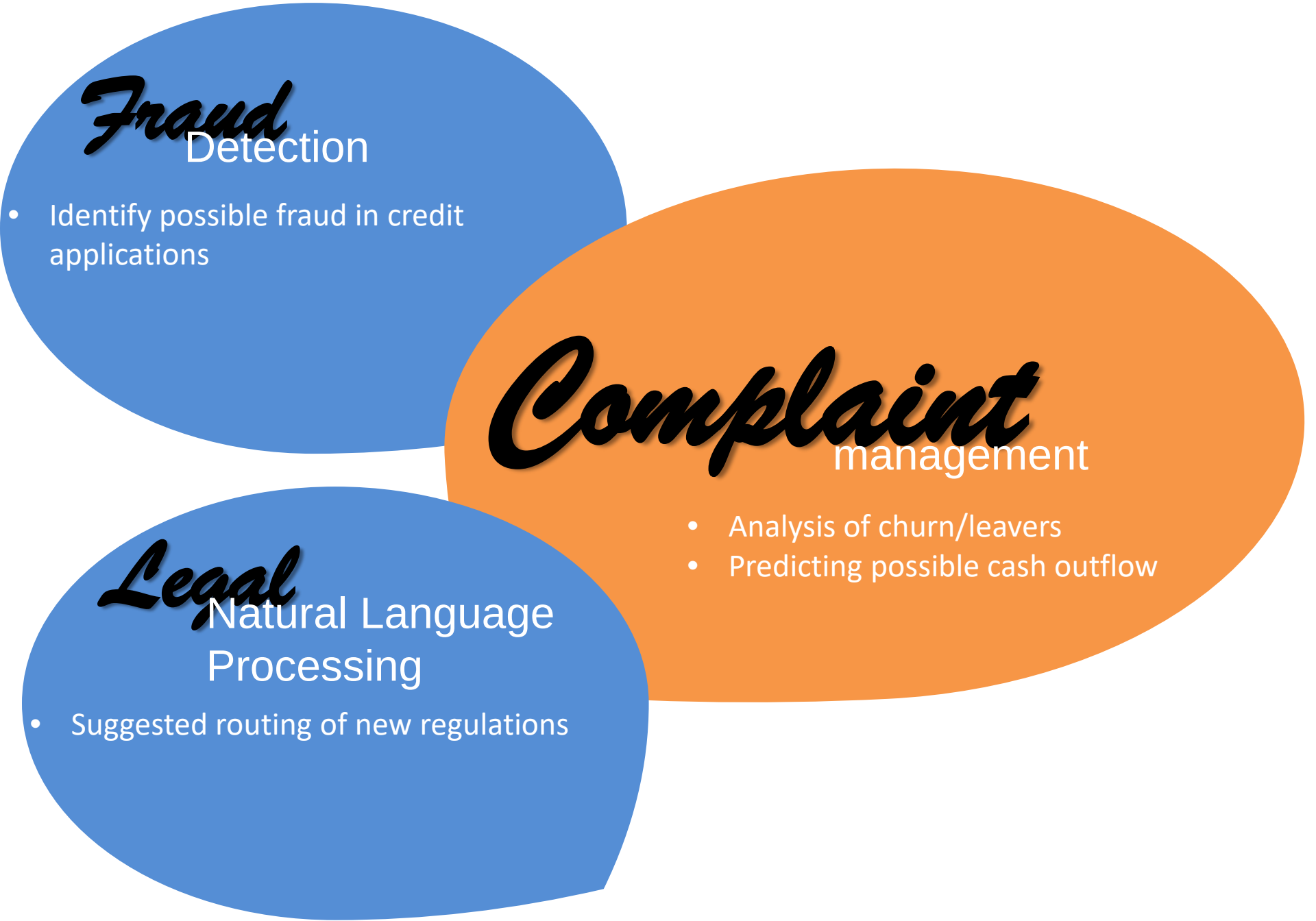
Dit is een indicatie



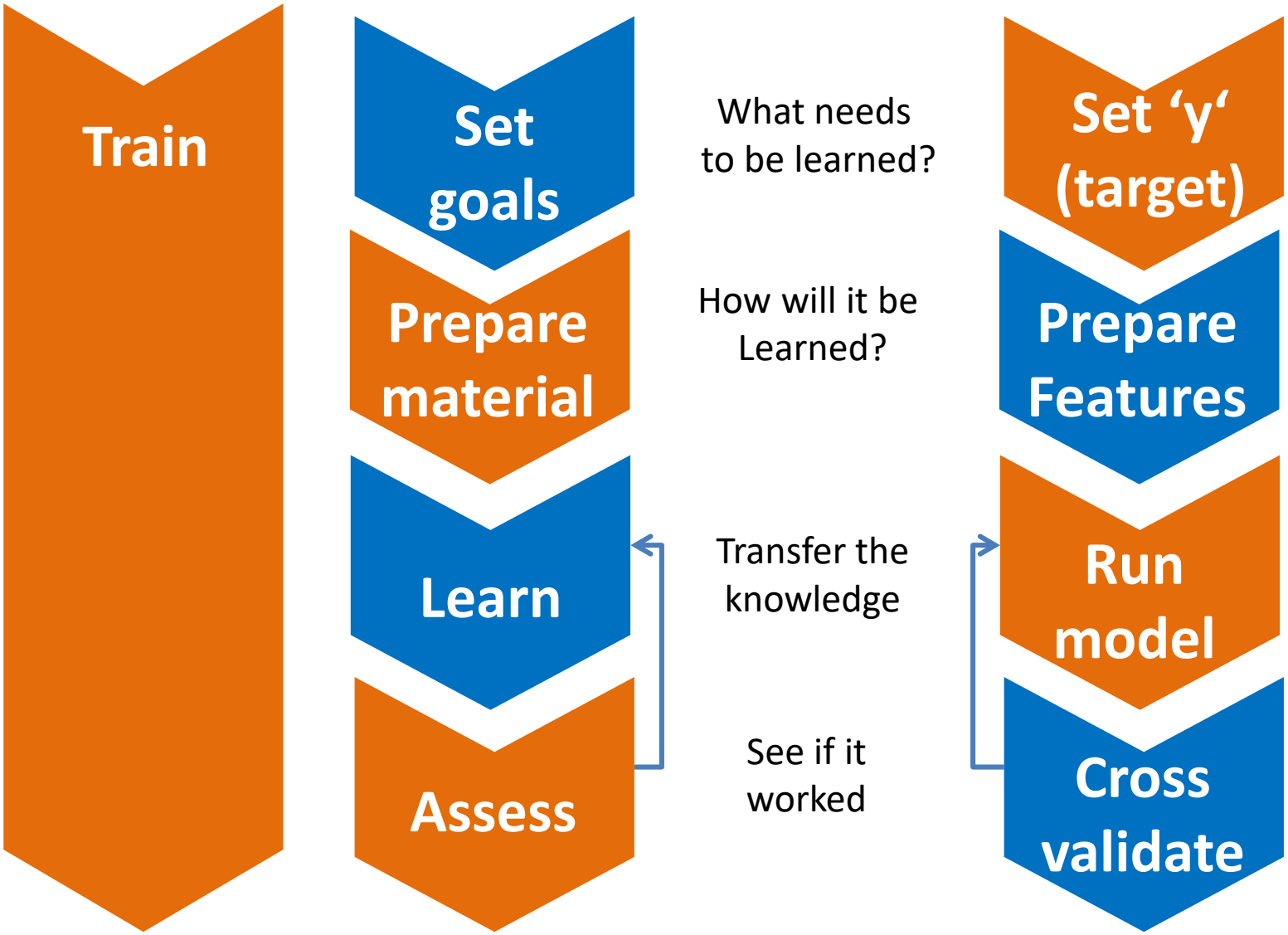
Item
Hierar
(You b
Printer
will als
ink - B

naar de Vuursche

4. Ga na 396 meter bij de zijstraat


Fraud
Detection
• Identify possible fraud in credit applications
Complaint
management
• Analysis of churn/leavers
• Predicting possible cash outflow
Legal
Natural Language Processing
• Suggested routing of new regulations

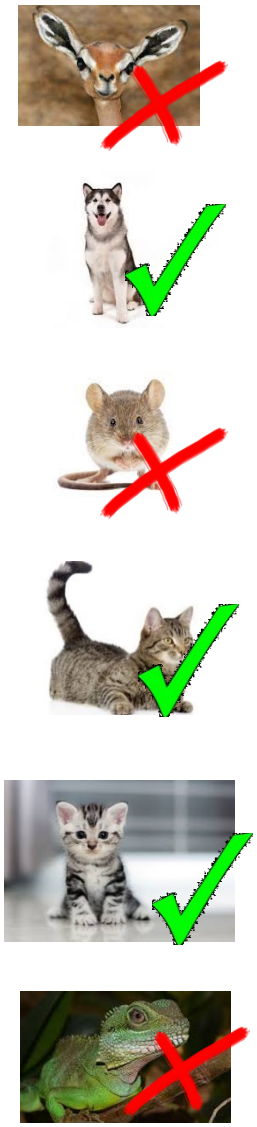

FIRST DAY WORK
Hire
Onboard
Train
Test
Apply
HELLO
I AM...
NEW
Get data
Select model
Train
Test
Apply


Train
Set goals
Prepare material
Learn
Assess
What needs to be learned?
How will it be Learned?
Transfer the knowledge
See if it worked
Set 'y' (target)
Prepare Features
Run model
Cross validate

‘y’ : Is this a cat?

‘X’
Train

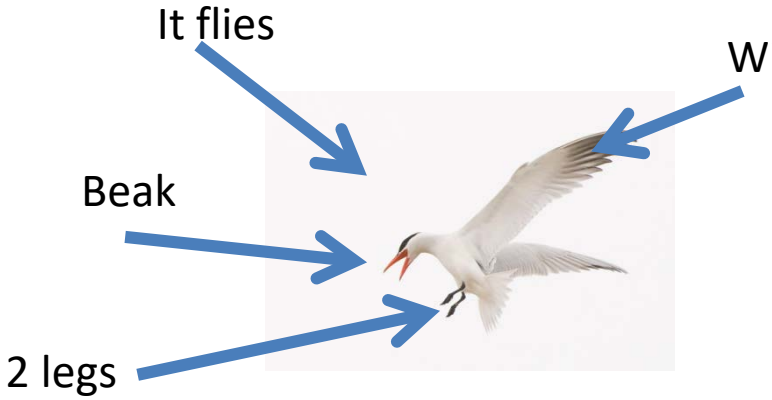
‘X’
Test



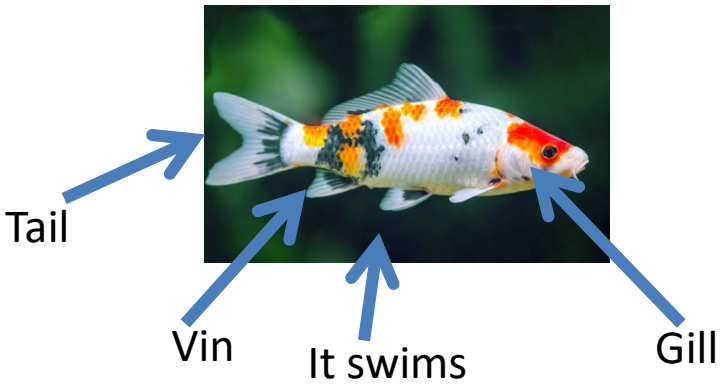
Unsupervised

‘X’ Train

How do we know ? learn

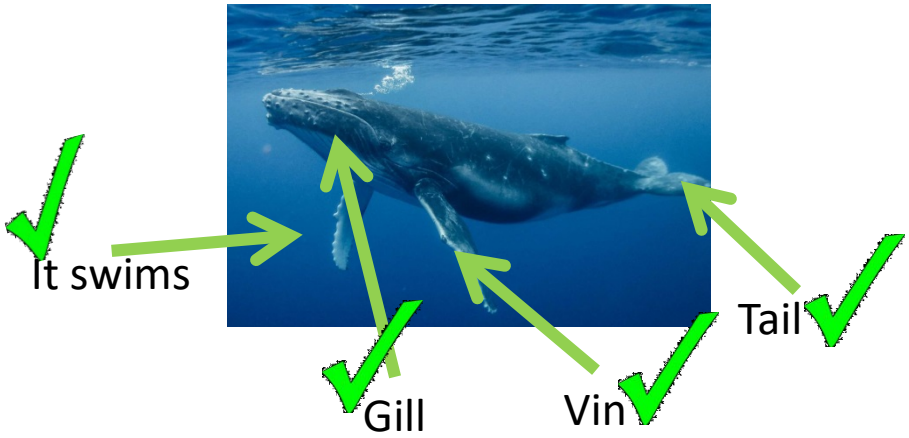


Bird

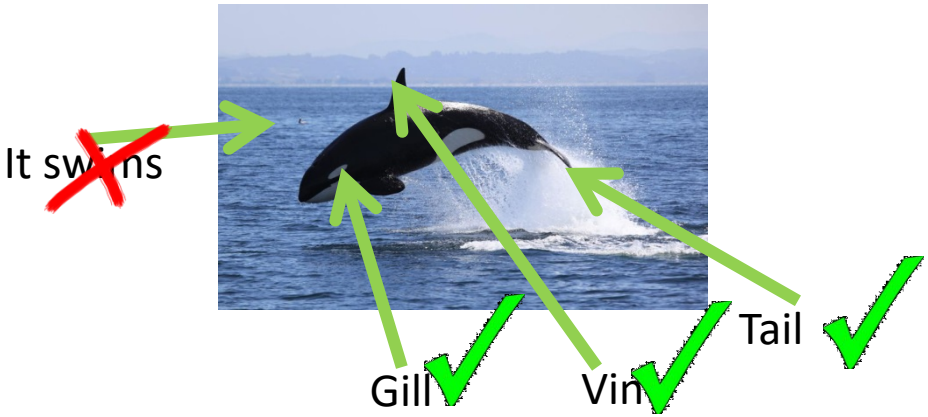


Fish

Pattern Recognition



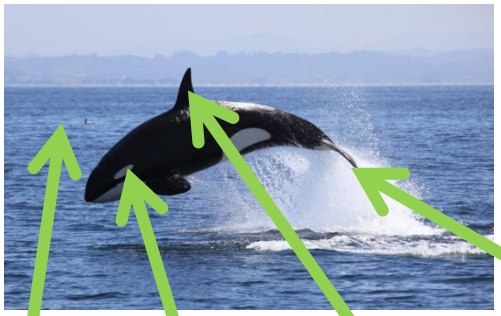
Fish ✓



Fish ?

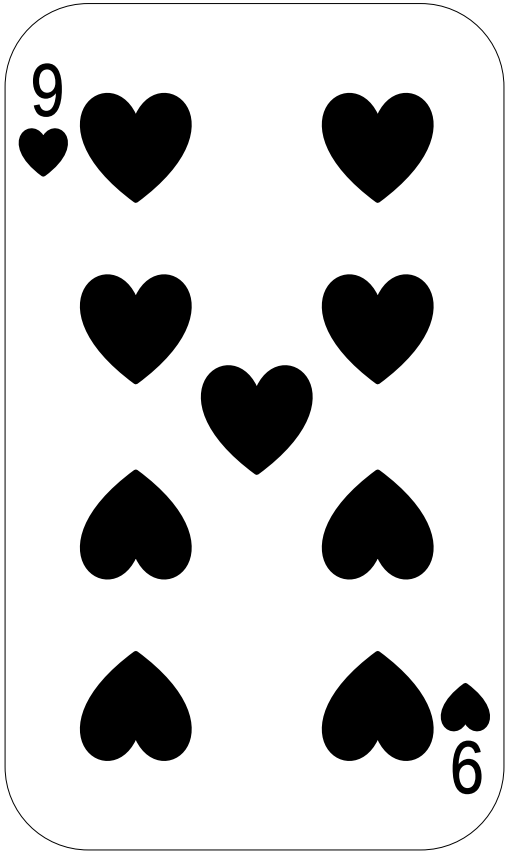
K N N
K-Next Neighbour

We look a ‘k’ number of neighbours and take the closest



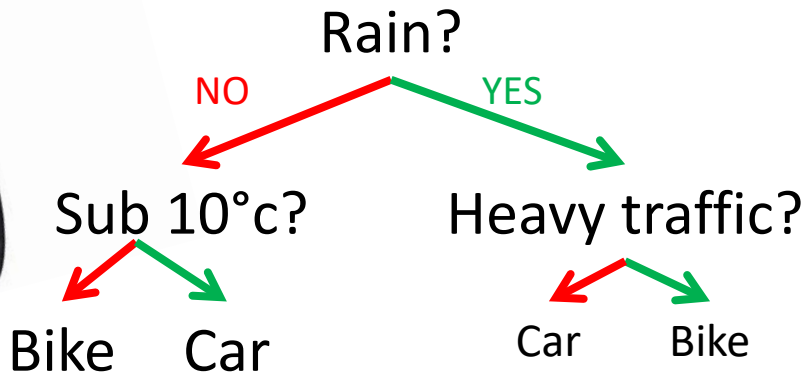
It swims	Gill	Vin	Tail	Fish
It flies	Beak	2 legs	Wings	Bird
It walks	Snout	4 legs	Tail	Dog

‘fish’ for a 6 year old or ‘mammal’ for 12 years and up


9
6

How do we Decide? Decision Tree

How to commute: bike or car?

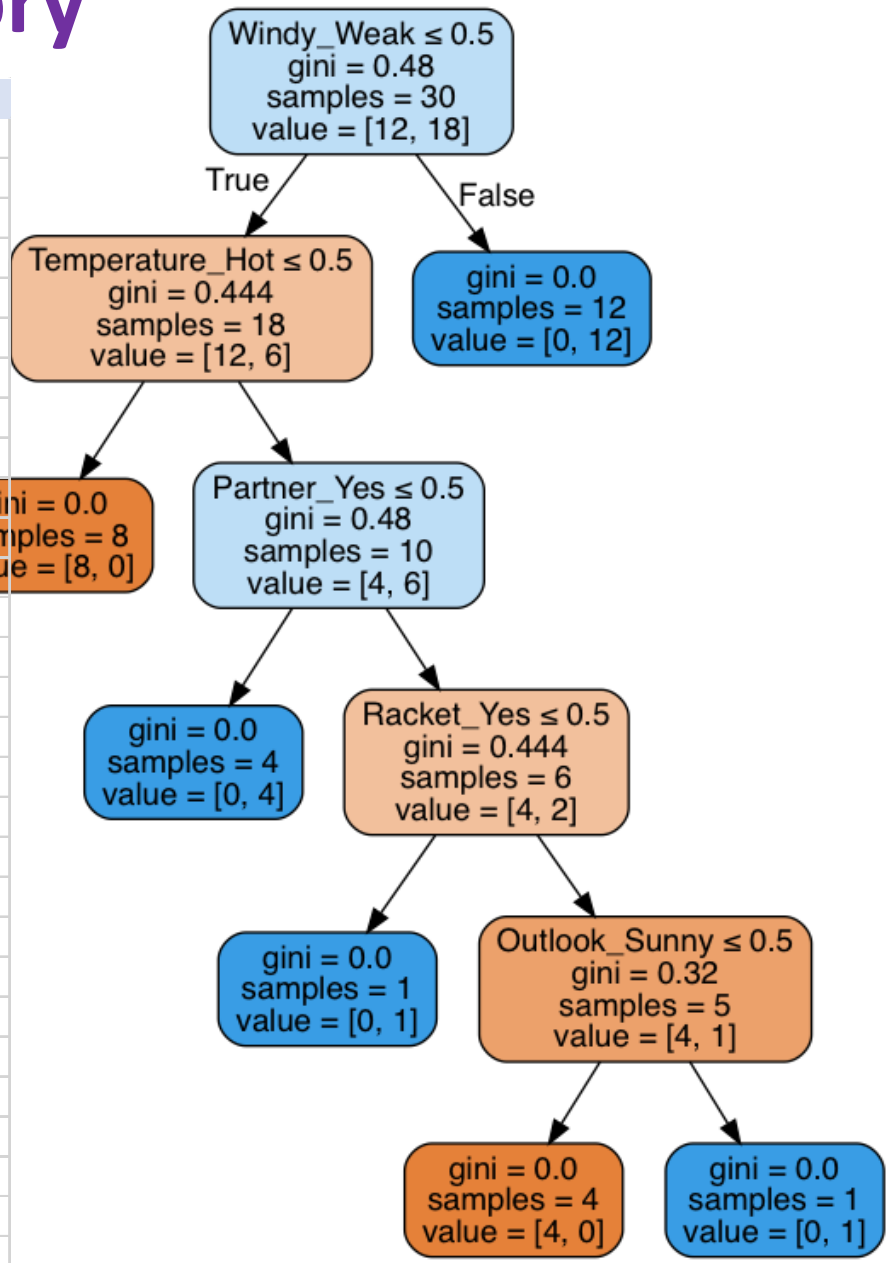


Lets's predict:

- No Rain
- 19 Celsius
- Light traffic

My tennis playin' history

Day	Racket	Partner	Windy	Temperat	Outlook	Humidity	Tennis
1	Yes	Yes	Strong	Hot	Rain	High	NO
2	Yes	Yes	Strong	Hot	Rain	High	NO
3	Yes	Yes	Weak	Cool	Sunny	Normal	YES
4	Yes	Yes	Strong	Hot	Sunny	High	YES
5	maybe	Yes	Weak	Cool	Rain	Normal	YES
6	Yes	No	Strong	Hot	Rain	Normal	YES
7	Yes	No	Weak	Cool	Rain	High	YES
8	Yes	No	Strong	Cool	Sunny	Normal	NO
9	Yes	No	Weak	Hot	Sunny	Normal	YES
10	Yes	No	Strong	Cool	Rain	High	NO
11	Yes	No	Strong	Hot	Rain	Normal	YES
12	No	Yes	Weak	Cool	Rain	High	YES
13	Yes	Yes	Strong	Cool	Sunny	Normal	NO
14	Yes	Yes	Weak	Hot	Sunny	Normal	YES
15	Yes	Yes	Strong	Cool	Rain	High	NO
16	Yes	Yes	Strong	Hot	Rain	High	NO
17	Yes	Yes	Strong	Hot	Rain	High	NO
18	Yes	Yes	Weak	Cool	Sunny	Normal	YES
19	Yes	No	Strong	Hot	Sunny	High	YES
20	Yes	No	Weak	Cool	Rain	Normal	YES
21	Yes	No	Strong	Hot	Rain	Normal	YES
22	Yes	No	Weak	Cool	Rain	High	YES
23	maybe	No	Strong	Cool	Sunny	Normal	NO
24	Yes	Yes	Weak	Hot	Sunny	Normal	YES
25	No	Yes	Strong	Cool	Rain	High	NO
26	No	Yes	Strong	Hot	Rain	Normal	YES
27	No	Yes	Weak	Cool	Rain	High	YES
28	Yes	Yes	Strong	Cool	Sunny	Normal	NO
29	Yes	Yes	Weak	Hot	Sunny	Normal	YES
30	Yes	Yes	Strong	Cool	Rain	High	NO



Decision trees get very complex, 'Overfitting'
Very fast

Create a Random Forest

Lots of small trees from random subsets of data and take a majority vote

Lets do a Random Forest

Day	Racket	Partner	Windy	Temperat	Outlook	Humidity	Tennis
1	Yes	Yes	Strong	Hot	Rain	High	NO
2	Yes	Yes	Strong	Hot	Rain	High	NO
3	Yes	Yes	Weak	Cool	Sunny	Normal	YES
4	Yes	Yes	Strong	Hot	Sunny	High	YES
5	maybe	Yes	Weak	Cool	Rain	Normal	YES
6	Yes	No	Strong	Hot	Rain	Normal	YES
7	Yes	No	Weak	Cool	Rain	High	YES
8	Yes	No	Strong	Cool	Sunny	Normal	NO
9	Yes	No	Weak	Hot	Sunny	Normal	YES
10	Yes	No	Strong	Cool	Rain	High	NO
11	Yes	No	Strong	Hot	Rain	Normal	YES
12	No	Yes	Weak	Cool	Rain	High	YES
13	Yes	Yes	Strong	Cool	Sunny	Normal	NO
14	Yes	Yes	Weak	Hot	Sunny	Normal	YES
15	Yes	Yes	Strong	Cool	Rain	High	NO
16	Yes	Yes	Strong	Hot	Rain	High	NO
17	Yes	Yes	Strong	Hot	Rain	High	NO
18	Yes	Yes	Weak	Cool	Sunny	Normal	YES
19	Yes	No	Strong	Hot	Sunny	High	YES
20	Yes	No	Weak	Cool	Rain	Normal	YES
21	Yes	No	Strong	Hot	Rain	Normal	YES
22	Yes	No	Weak	Cool	Rain	High	YES
23	maybe	No	Strong	Cool	Sunny	Normal	NO
24	Yes	Yes	Weak	Hot	Sunny	Normal	YES
25	No	Yes	Strong	Cool	Rain	High	NO
26	No	Yes	Strong	Hot	Rain	Normal	YES
27	No	Yes	Weak	Cool	Rain	High	YES
28	Yes	Yes	Strong	Cool	Sunny	Normal	NO
29	Yes	Yes	Weak	Hot	Sunny	Normal	YES
30	Yes	Yes	Strong	Cool	Rain	High	NO

- You all get a small part of a dataset
- Make your own decision tree based on the data
- The decision trees together form a forest
- The forest is used to predict if we play or not

How to make a Decision Tree?

- Which ingredient is most useful to distinguish Pancakes from Omelettes?

Milk	Flour	Cheese	Pancake/Omelette
Yes	Yes	Yes	Pancake
Yes	Yes	No	Pancake
No	No	Yes	Omelette
Yes	No	No	Omelette

Milk	Pancake	Omelette
No	0	1
Yes	2	1

Flour	Pancake	Omelette
No	0	2
Yes	2	0

Cheese	Pancake	Omelette
No	1	1
Yes	1	1

Count !

Will I play?

Temperature	Outlook	Humidity	Windy	Play tennis?
Cool	Sunny	High	Weak	
Hot	Rain	High	Strong	
Hot	Rain	Normal	Strong	

The Democratic Forest:
Majority rules

Random Forest
Decision tree

Supervised Learning

- Training set contains the ‘solution’
- Predictions made based on similar observations

- **Regression**
- **Support Vector Machines**
- **Naïve Bayes**
- **Stochastic Gradient Descent**

Unsupervised Learning

- ‘Solution’ unknown
- Discovery based on patterns

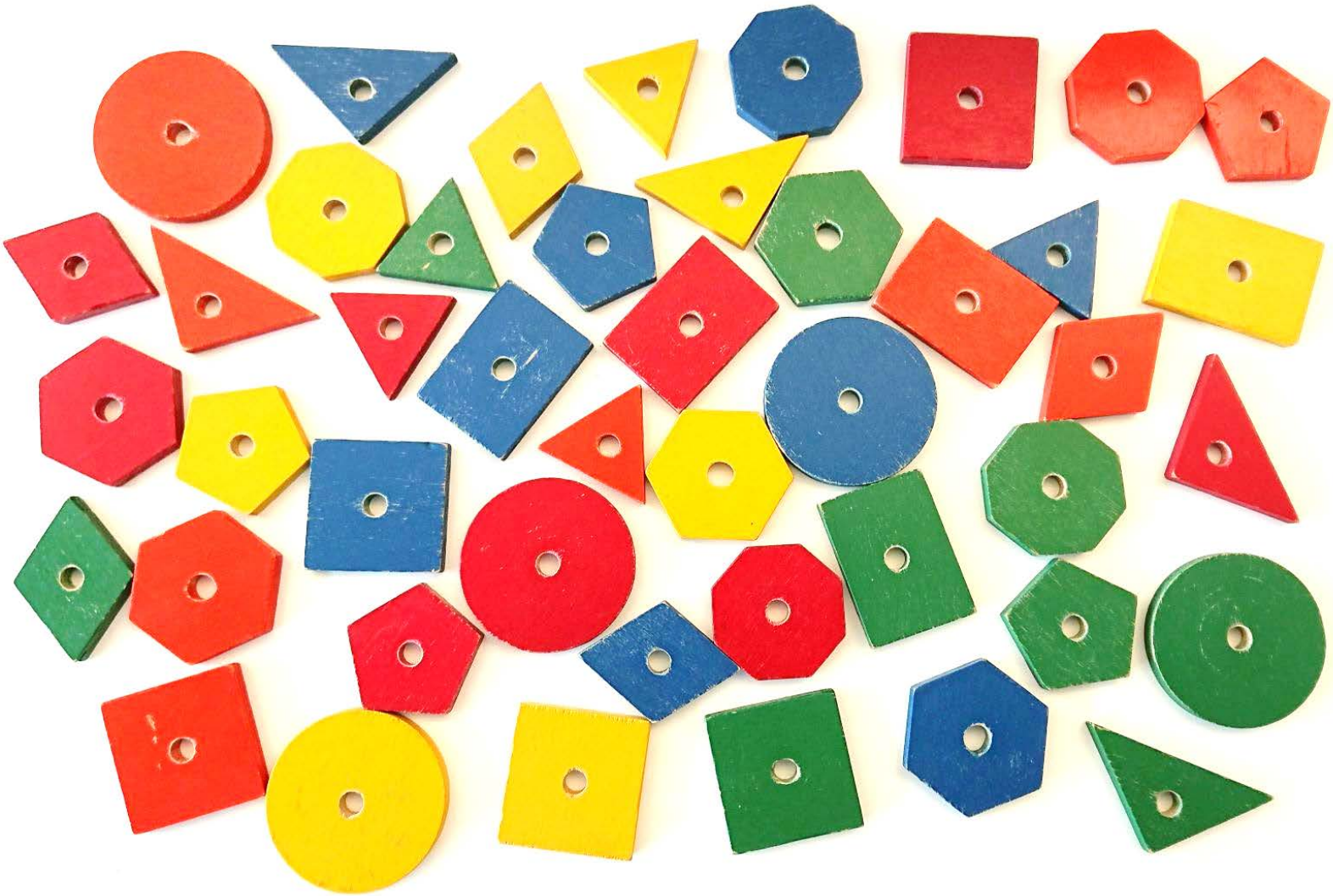
Unsupervised Learning intuition

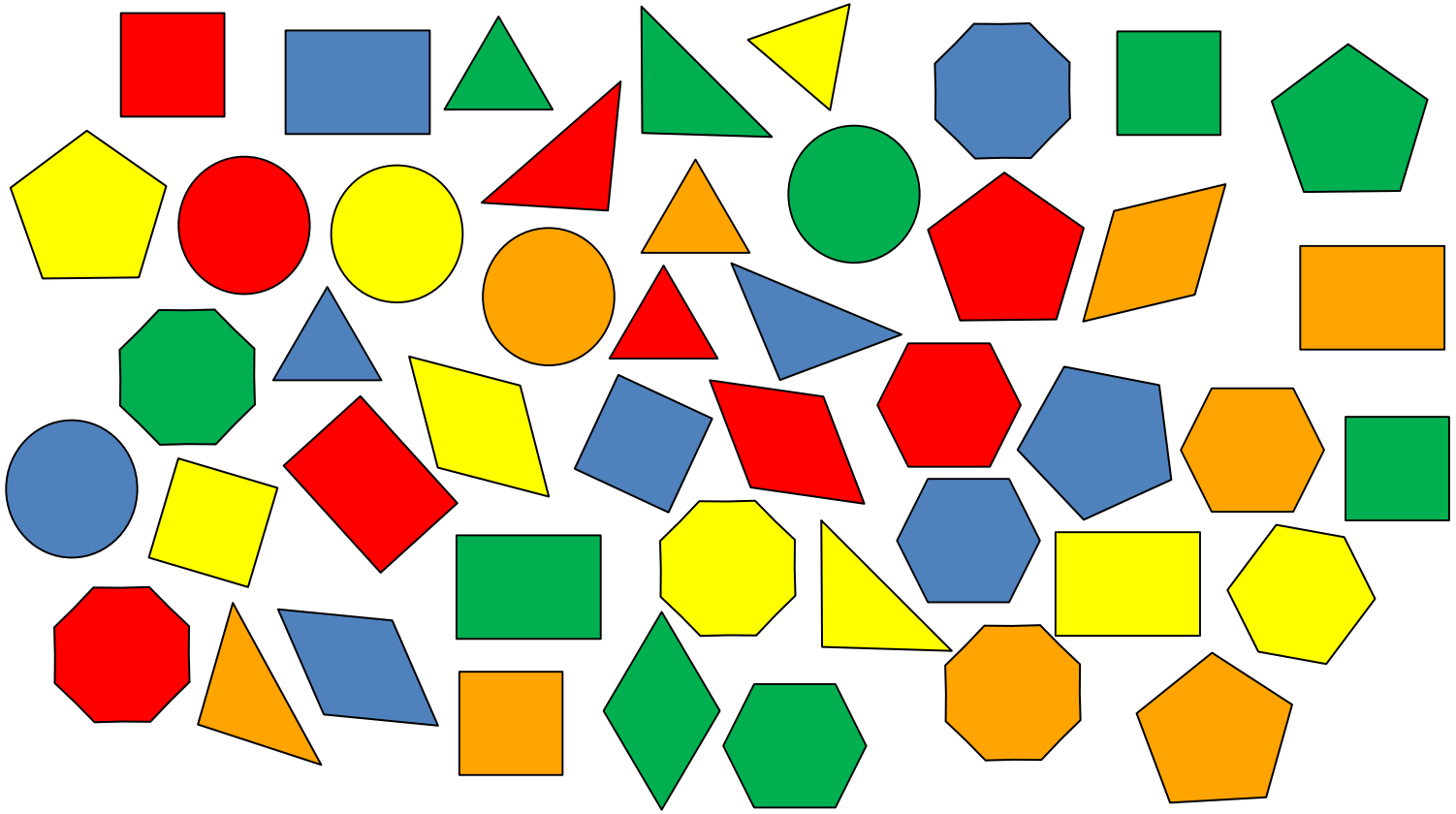
K-Means

1. Run through the entire data set
2. Look for most similarity for k-number of groups

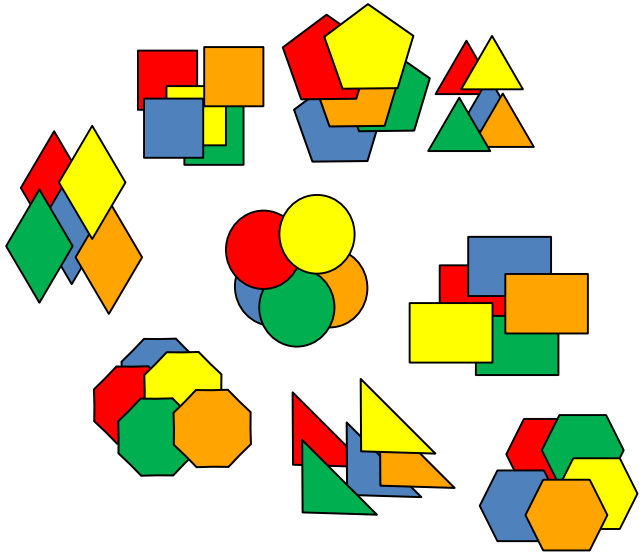
Dbscan (Density-Based)

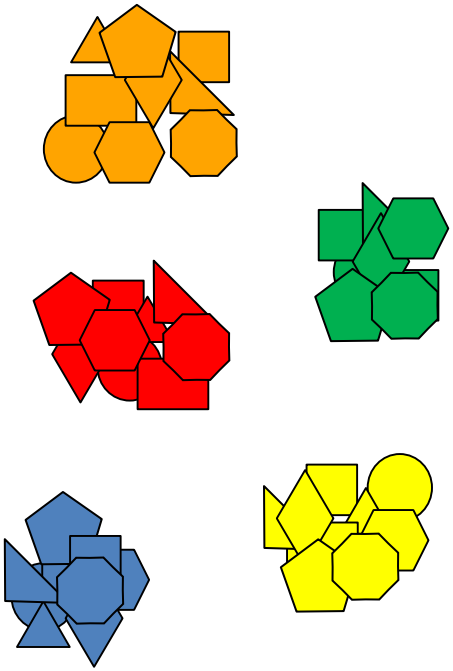
1. Take an item
2. Take the next one and see if it is similar ‘enough’
    1. Yes, add to group, go to the next
    2. No, try on the next one for ‘x’ number of times
        1. If none found, start a new group





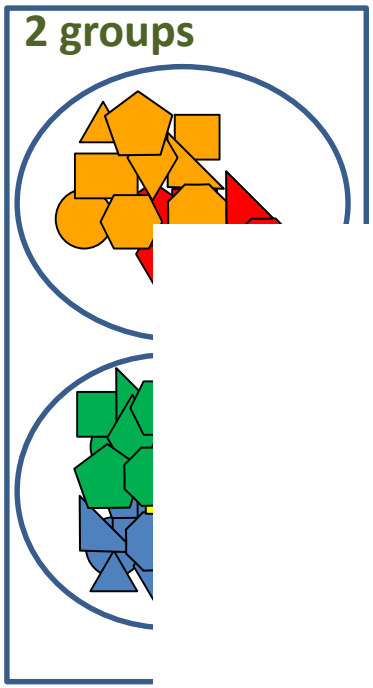
What did we learn in school?

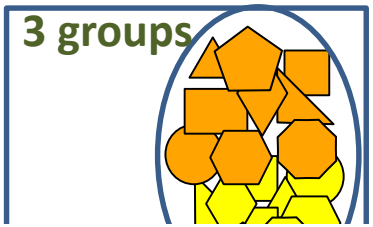


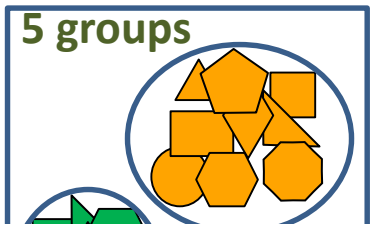


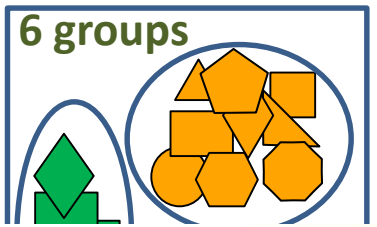
Code on https://github.com/MaartenKool/

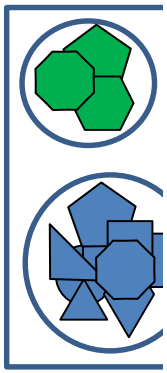
What did de computer do?



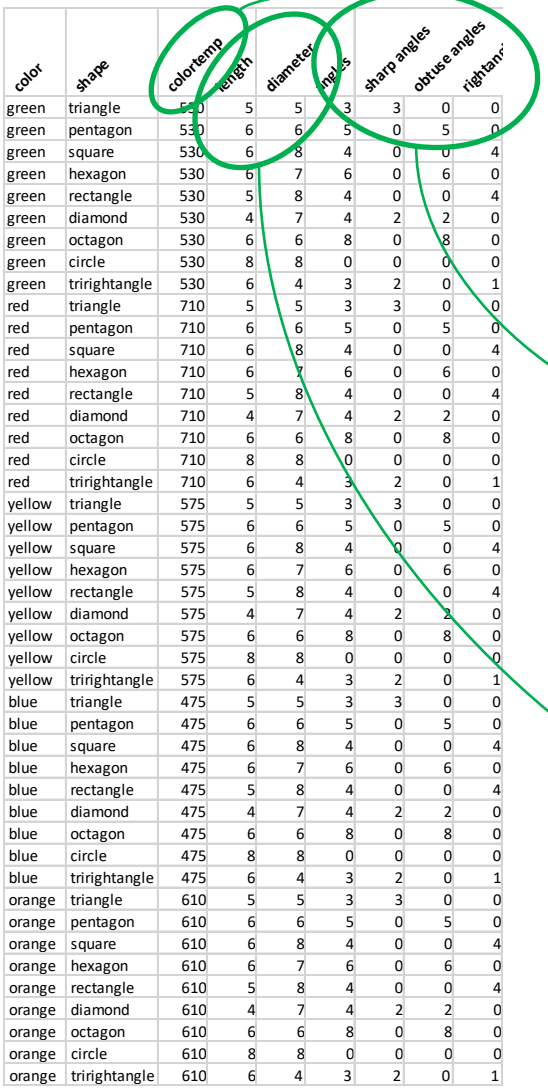








How does this magic work?



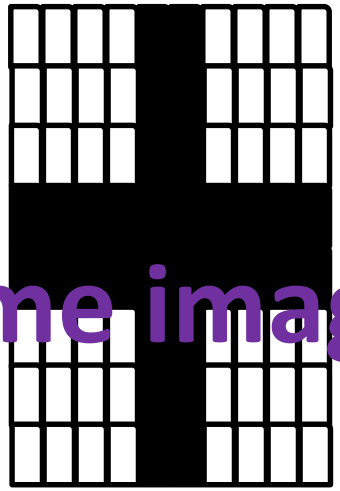
color	shape	colortemp	length	diameter	angles	sharp angles	obtuse angles	rightangle
green	triangle	530	5	5	3	3	0	0
green	pentagon	530	6	6	5	0	5	0
green	square	530	6	8	4	0	0	4
green	hexagon	530	6	7	6	0	6	0
green	rectangle	530	5	8	4	0	0	4
green	diamond	530	4	7	4	2	2	0
green	octagon	530	6	6	8	0	8	0
green	circle	530	8	8	0	0	0	0
green	trirightangle	530	6	4	3	2	0	1
red	triangle	710	5	5	3	3	0	0
red	pentagon	710	6	6	5	0	5	0
red	square	710	6	8	4	0	0	4
red	hexagon	710	6	7	6	0	6	0
red	rectangle	710	5	8	4	0	0	4
red	diamond	710	4	7	4	2	2	0
red	octagon	710	6	6	8	0	8	0
red	circle	710	8	8	0	0	0	0
red	trirightangle	710	6	4	3	2	0	1
yellow	triangle	575	5	5	3	3	0	0
yellow	pentagon	575	6	6	5	0	5	0
yellow	square	575	6	8	4	0	0	4
yellow	hexagon	575	6	7	6	0	6	0
yellow	rectangle	575	5	8	4	0	0	4
yellow	diamond	575	4	7	4	2	2	0
yellow	octagon	575	6	6	8	0	8	0
yellow	circle	575	8	8	0	0	0	0
yellow	trirightangle	575	6	4	3	2	0	1
blue	triangle	475	5	5	3	3	0	0
blue	pentagon	475	6	6	5	0	5	0
blue	square	475	6	8	4	0	0	4
blue	hexagon	475	6	7	6	0	6	0
blue	rectangle	475	5	8	4	0	0	4
blue	diamond	475	4	7	4	2	2	0
blue	octagon	475	6	6	8	0	8	0
blue	circle	475	8	8	0	0	0	0
blue	trirightangle	475	6	4	3	2	0	1
orange	triangle	610	5	5	3	3	0	0
orange	pentagon	610	6	6	5	0	5	0
orange	square	610	6	8	4	0	0	4
orange	hexagon	610	6	7	6	0	6	0
orange	rectangle	610	5	8	4	0	0	4
orange	diamond	610	4	7	4	2	2	0
orange	octagon	610	6	6	8	0	8	0
orange	circle	610	8	8	0	0	0	0
orange	trirightangle	610	6	4	3	2	0	1

Color: Wavelength in Nanometers

Convert to numbers

Shape: Counts of angles

Dimensions: in centimeters

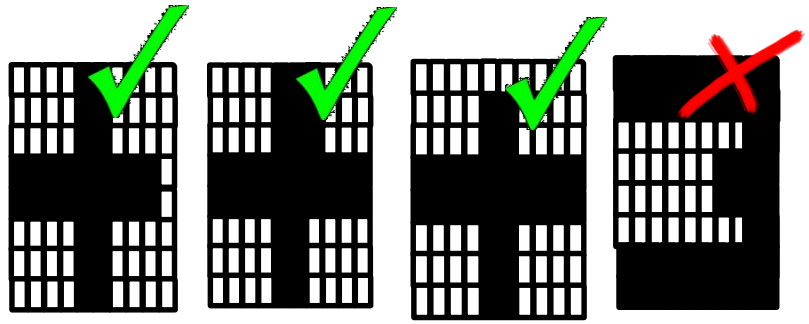


0000110000
0000110000
0000110000
1111111111
1111111111
0000110000
0000110000
0000110000

Let’s do some image recognition

000011000000001100000000110000111111111111111111111000011000000001100000000110000

pixel10	pixel11	pixel12	pixel13	pixel14	pixel15	pixel16	pixel17	pixel18	pixel19	pixel20	pixel21	pixel22	pixel23	pixel24	pixel25	pixel26	pixel27	pixel28	pixel29	pixel30	pixel31	pixel32	pixel33	pixel34	pixel35	pixel36	pixel37	pixel38	pixel39	pixel40	pixel41	pixel42	pixel43	pixel44	pixel45	pixel46
0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1



0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1
0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1
0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

‘y’ : what is the subject of the newspaper article

Politiek leeft niet bij Facebookers

Van een onzer verslaggeefsters

AMSTERDAM, dinsdag

Bijna de helft van alle Amsterdammers heeft een Facebook-account, maar slechts acht procent vindt de pagina van zijn of haar stadsdeel leuk. Meer dan negentig procent volgt de medelingen die het stadsdeel via social media bekend maakt dus helemaal niet.

Het onderzoek vond plaats in het verlengde van een eerder onderzoek naar de manier waarop Amsterdamse diensten en stadsdelen hun mediabeleid hebben vormgegeven. Uit dat onder-

Negentig procent negeert stadsdeel

Amsterdamse jongeren op Facebook te vinden en ruim de helft van de 39-50 jarigen. Van de 50- tot 64-jarigen is het een derde. In de jongste groep worden LinkedIn en Hyves door respectievelijk 39%

woensdag 11 juli 2012

Uithet AD - RD

STAD & REGIO

Frisbee en knotshockey

Gemeente: 45-plussers moeten meer sporten

ERDAM • Te veel 45-plussers in dam zitten liever op de bank dan ichzelf in het zweet werken bij tvereniging. Om meer volwas- eweging te krijgen, introdu- ds meer clubs speciale pro- et frisbee en korfbal.

ELLEN VAN GAALEN

Met een speciaal programma, waaronder spelletjes knotshockey, moeten 45-plussers naar de sportverenigingen worden gerokken.

ningen sporten r Rotterdam- inette Laan n de bewo- 'n 2014 in ralsnog rocent. ek en atie In 'n

CHIEFFOTO JAN RODEN

eciaal programma opgestart olussers naar de vereniging gen.

ssenen sporten niet, aan hun wensen vol- woordvoerder van De clubs moeten veranderen. Dat zie je steeds meer gebeuren." Korfbalvereniging OZC in Vreewijk loopt als één van de verenigingen voorop. Rotterdammers die 45 jaar of ouder zijn mogen voor 17 euro per kwartaal wekelijks een uur sporten.

Elke week staat een andere sport op het programma, van een potje jeu-de-boules tot korfbal. „Bij volwassenen moet je niet met korfbal aankomen," zegt coördinator Bobbie van der Meer. „Dat vinden ze veel te intensief."

Het gevarieerde aanbod valt wél in goede aarde. OZC kreeg veertien bewoners in beweging. „De jongste deelnemer is 44, de oudste 83. Ze vinden het hartstikke leuk om actief bezig te zijn."

	room	female	law	pixel44	pixel45
text1	1	1	1	1	1
text2	1	0	0	1	
text3	0	0	1	1	

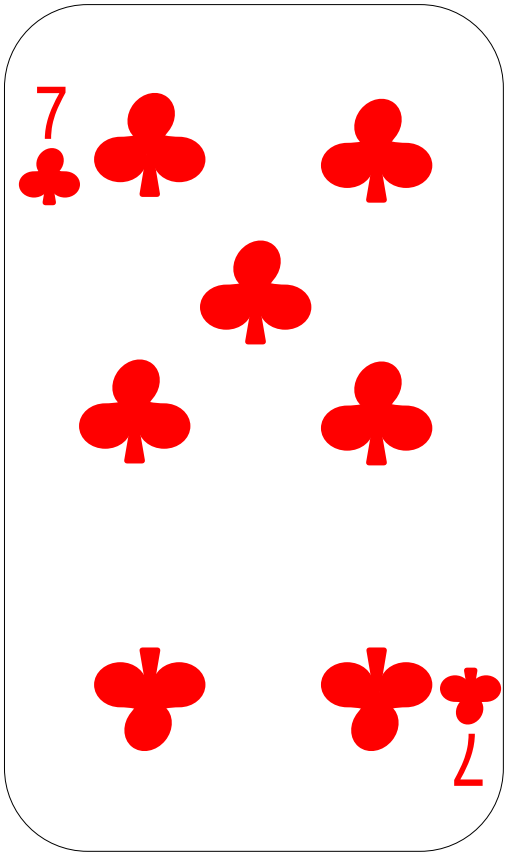
Polit

Convert text to numbers: Count words

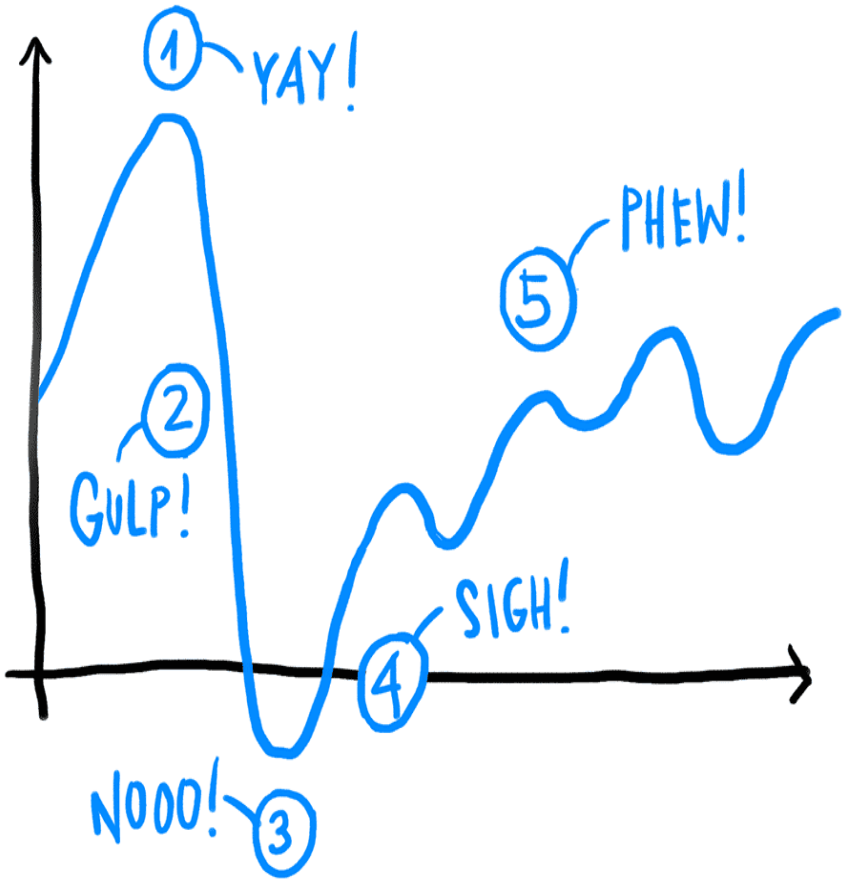
	act	acts	article	association	attacks	basis	beings	belief	committed	common	community	compelled	conscience	contrary	country	cultural	declarati
0	3	3	30	3	2	3	2	3	2	3	3	2	3	2	5	2	

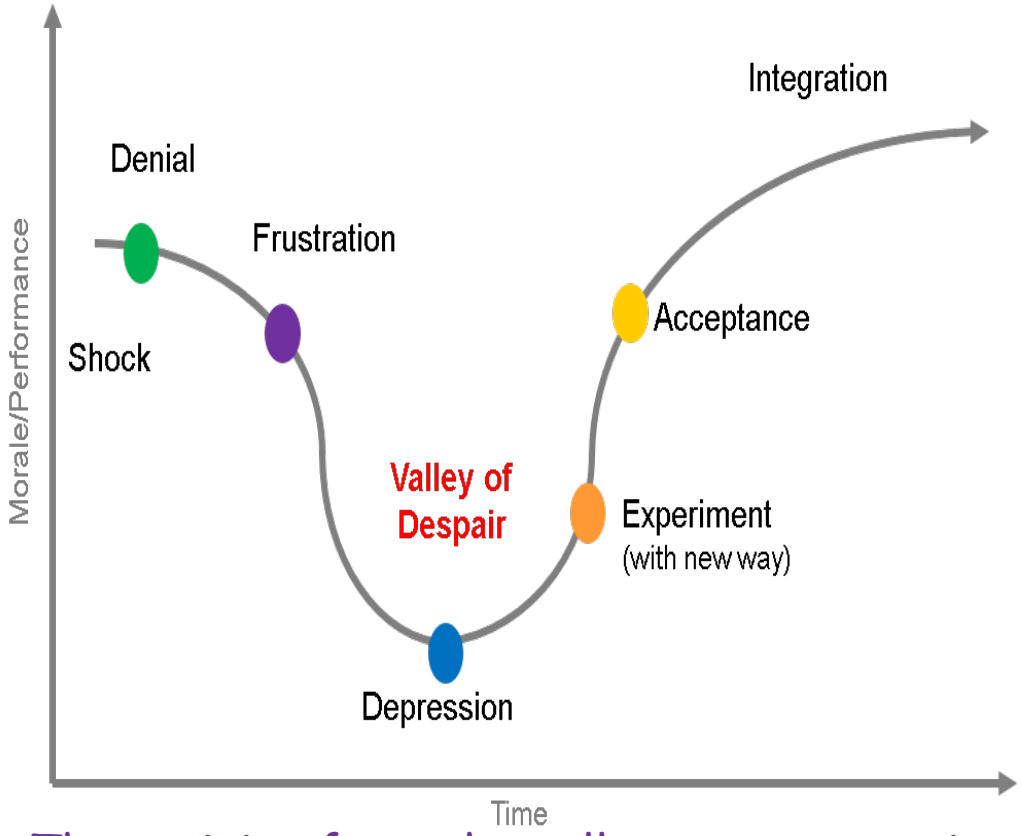
	french	english	german	finnish
0	droit	right	recht	oikeus
1	article	article	artikel	artikla
2	toute	shall	darf	jokaisella
3	droits	rights	rechte	seka
4	etre	human	anspruch	hanen
5	libertes	equal	nationen	saa
6	homme	freedom	freiheiten	kaikkien
7	liberte	freedoms	erklarung	yhteiskunnan
8	pays	entitled	schutz	etta
9	nations	protection	freiheit	kansakuntien
10	nul	law	sowie	vastaan
11	protection	education	niemand	suojaan
12	considerant	social	fur	ketaan
13	contre	nations	menschen	tassa
14	declaration	free	menschenrechte	opetuksen

What simple wordcounting can do


7
7

Fallible pattern recognition? by a human?





The training from the collegues perspective

Sooo…

I thought it was all about Big Data

Small Data

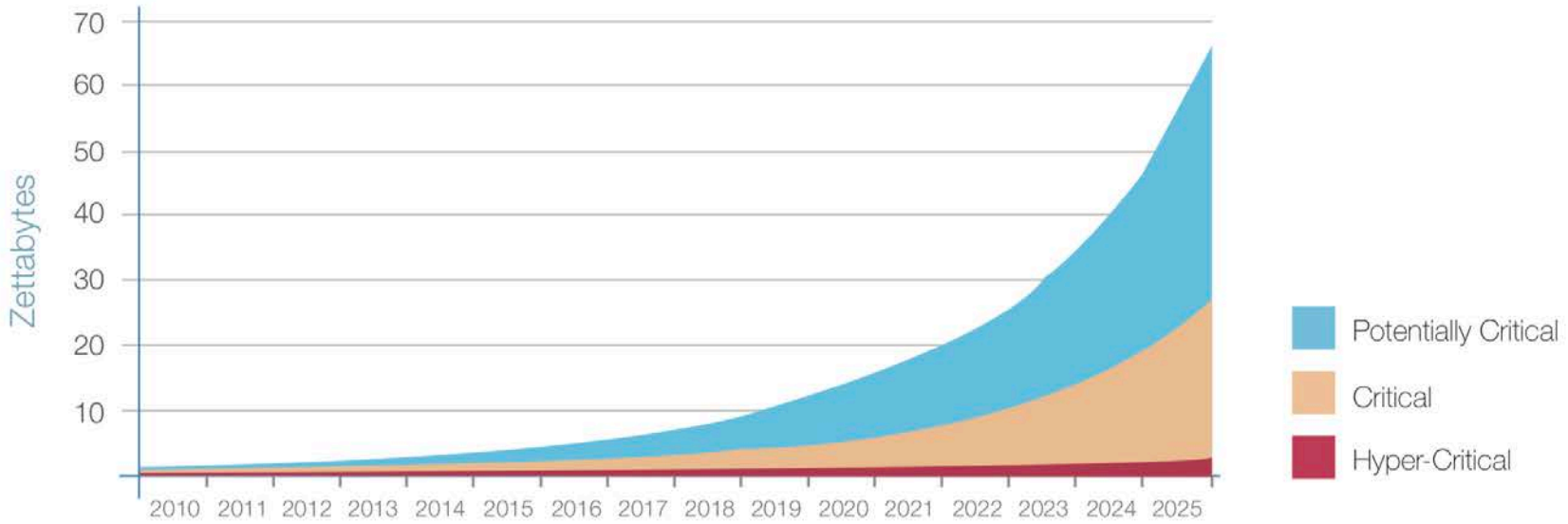
- Purposely collected
- Structured, well defined
- Typically a subset of reality
- Traditional statistics are applicable

Big Data

- Data collection often a by-product
- 4/5/6/7 V's:
  1. Volume
  2. Variety
  3. Velocity
  4. Veracity
  5. Viability
  6. Value
  7. Visualisation
- May be the complete reality

V for Viability
V for Volume

Figure 5. | Data Criticality Over Time



Idc whitepaper 2017

V for Variability

V for Velocity

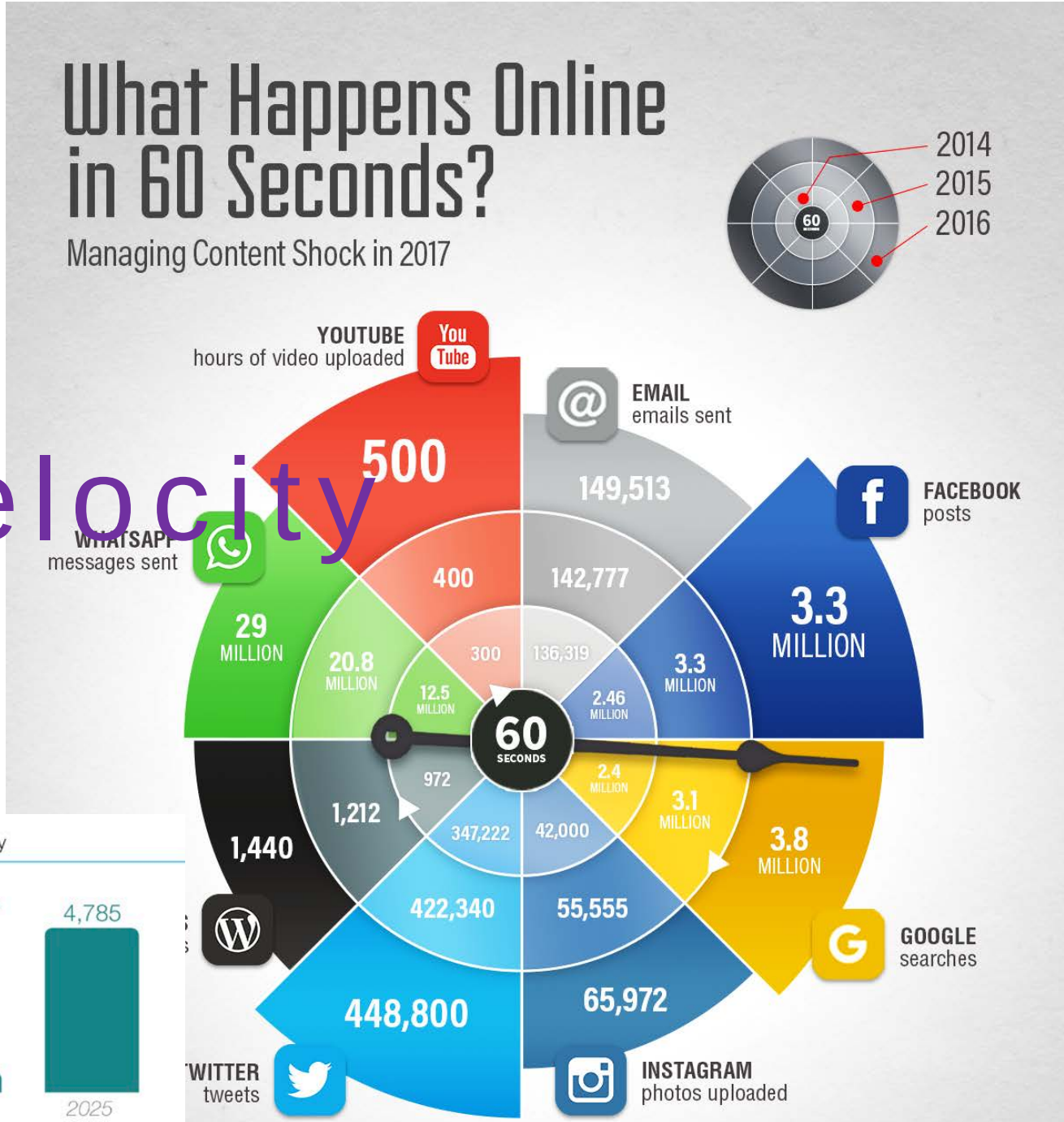
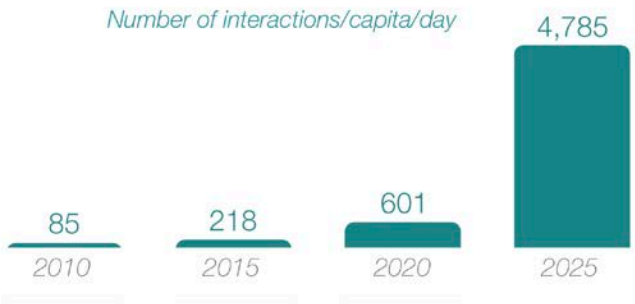
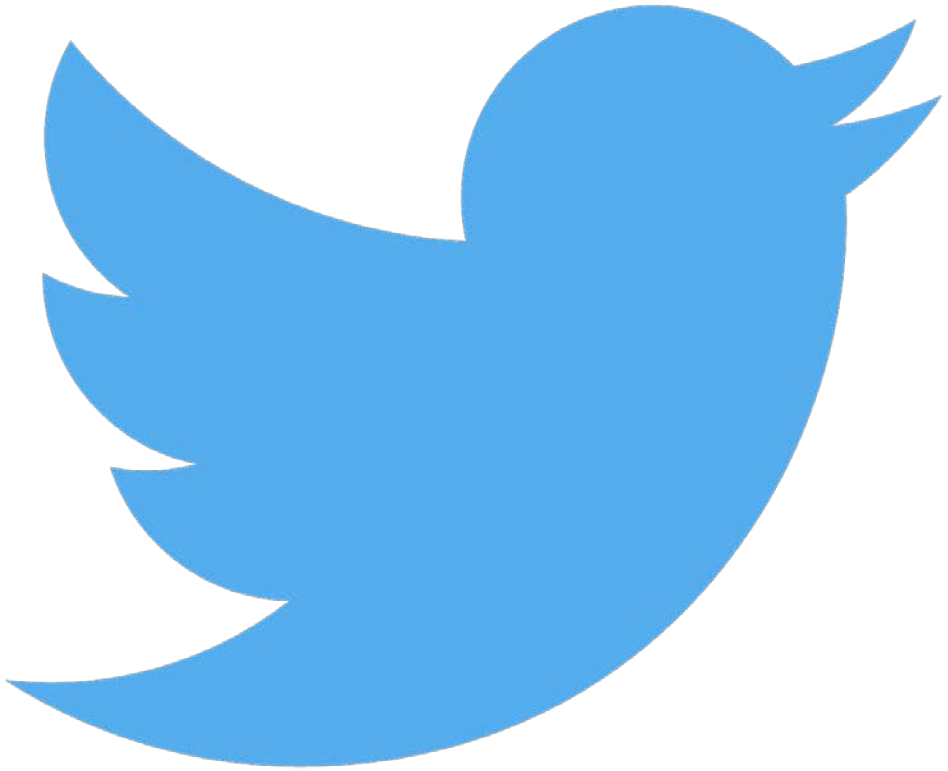


Figure 8. | Interactions per Connected Person per Day



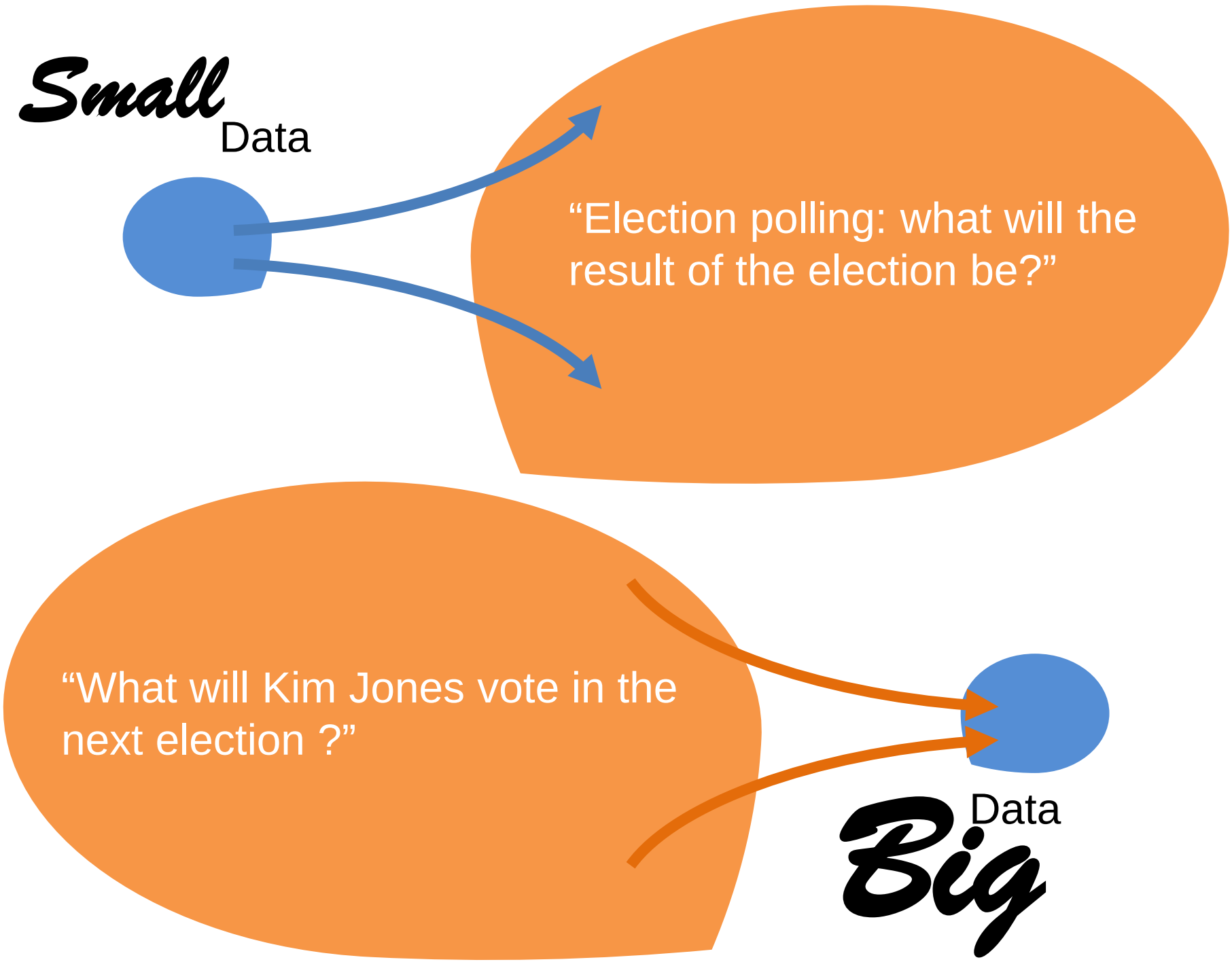
Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017



V for Veracity

V for Visibility

https://hackercombat.com/the-best-10-deep-web-search-engines-of-2017/


Small Data
“Election polling: what will the result of the election be?”
“What will Kim Jones vote in the next election ?”
Big Data

What's In It For Me? W.I.I.F.M

1. Data
2. Suspected/likely pattern
3. Existing training set, or willingness to build one
4. practical implementation/application
5. In NFR Grid domain

Come and talk to me:

Maarten.kool@nl.abnamro.com

Played tennis? (dataset 1)

Temperature	Outlook	Humidity	Played tennis?
Hot	Rain	High	NO
Hot	Rain	High	NO
Cool	Sunny	Normal	YES
Hot	Sunny	High	YES
Cool	Rain	Normal	YES

1) Which feature is the most informative?
2) Which follow-up feature is most informative?

Played tennis? (dataset 2)

Temperature	Humidity	Windy	Played tennis?
Hot	Normal	Strong	**YES**
Cool	High	Weak	**YES**
Cool	Normal	Strong	**NO**
Hot	Normal	Weak	**YES**
Cool	High	Strong	**NO**

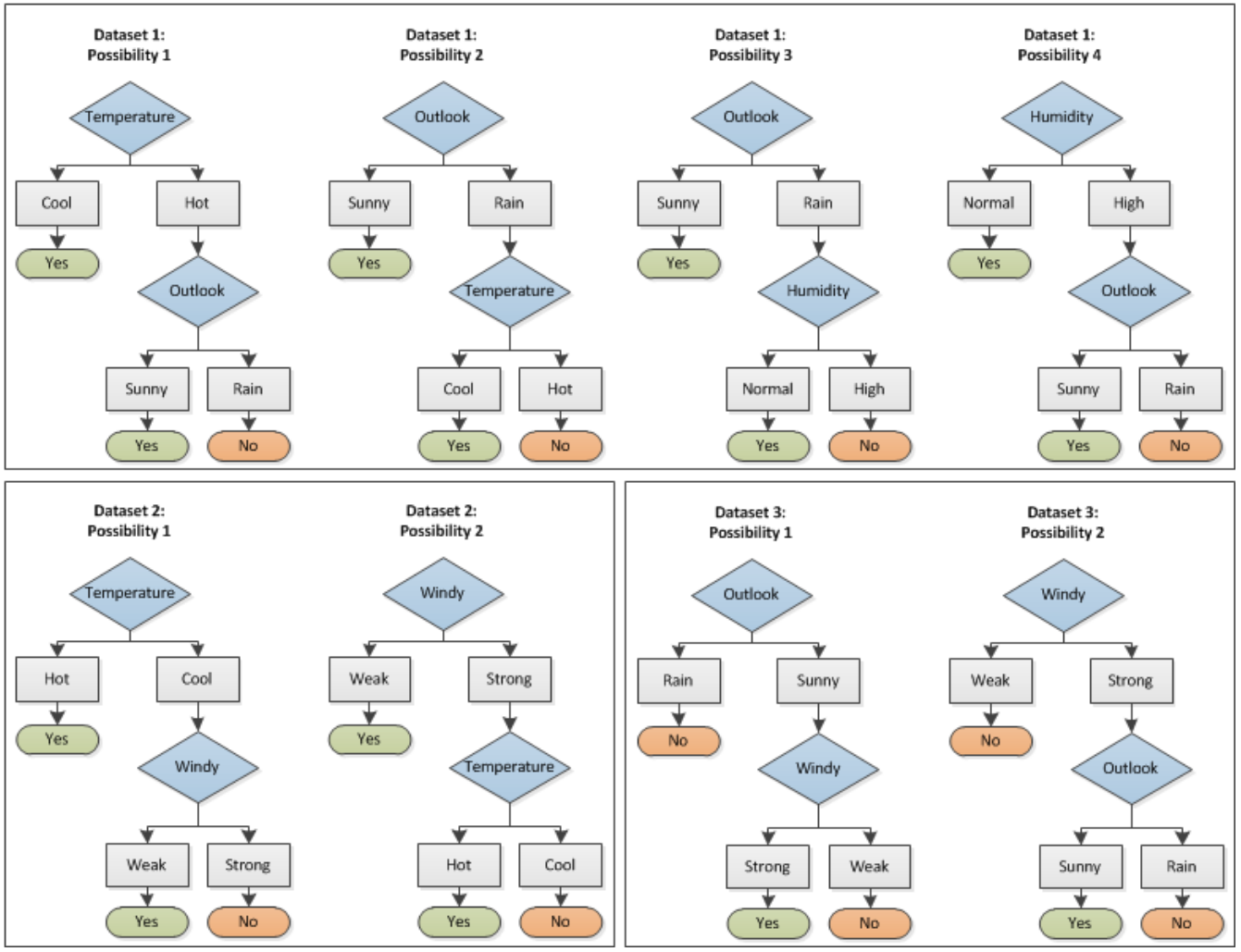
1) Which feature is the most informative?
2) Which follow-up feature is most informative?

Played Tennis? (dataset 3)

Temperature	Outlook	Windy	Played tennis?
Hot	Rain	Strong	NO
Hot	Sunny	Strong	YES
Cool	Sunny	Weak	NO
Cool	Sunny	Strong	YES
Cool	Sunny	Strong	YES
Cool	Rain	Weak	NO

1) Which feature is the most informative?
2) Which follow-up feature is most informative?

Answers


Dataset 1:
Possibility 1
Temperature
Cool
Hot
Yes
Outlook
Sunny
Rain
Yes
No
Dataset 1:
Possibility 2
Outlook
Sunny
Rain
Yes
Temperature
Cool
Hot
Yes
No
Dataset 1:
Possibility 3
Outlook
Sunny
Rain
Yes
Humidity
Normal
High
Yes
No
Dataset 1:
Possibility 4
Humidity
Normal
High
Yes
Outlook
Sunny
Rain
Yes
No
Dataset 2:
Possibility 1
Temperature
Hot
Cool
Yes
Windy
Weak
Strong
Yes
No
Dataset 2:
Possibility 2
Windy
Weak
Strong
Yes
Temperature
Hot
Cool
Yes
No
Dataset 3:
Possibility 1
Outlook
Rain
Sunny
No
Windy
Strong
Weak
Yes
No
Dataset 3:
Possibility 2
Windy
Weak
Strong
No
Outlook
Sunny
Rain
Yes
No