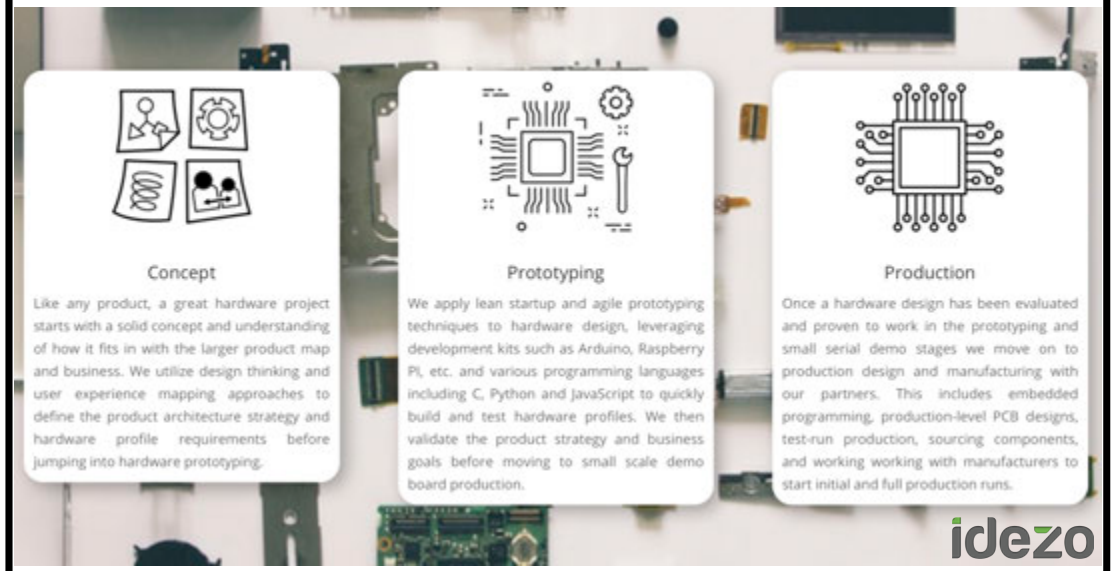
HANDS ON APPROACH TO

Data Science

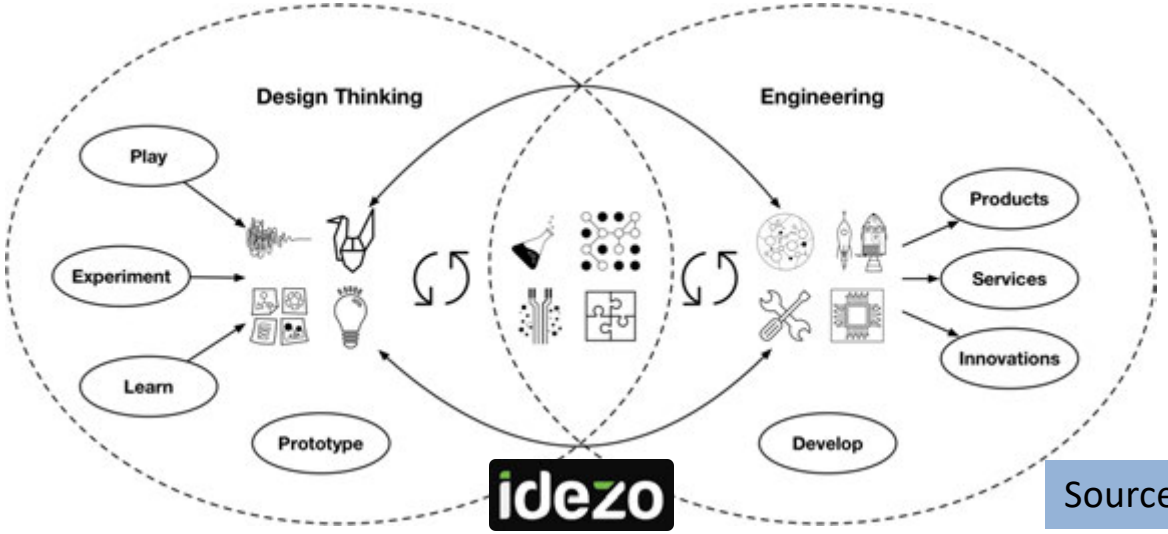
for (the)

IoT



Rob van der Willigen

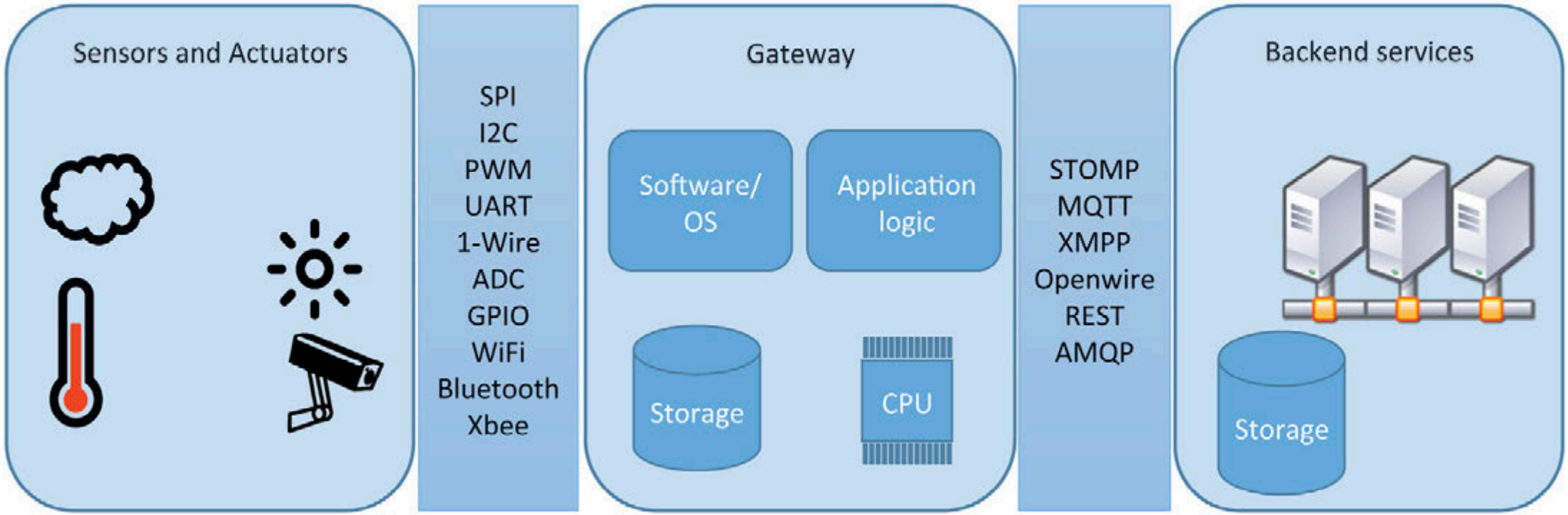
HANDS ON APPROACH TO DATA SCIENCE for (the) IoT



Source: https://idezo.ch

This Data Science Course was developed for keuzevak- program of the School of Communication, Media and Information Technology (CMI, at the Hogeschool Rotterdam **(Rotterdam University of Applied Sciences, RUAS)**.
If you find errors or omissions, please contact the author, Rob van der Willigen, at r.f.van.der.willigen@hr.nl. Materials of this course and code examples used will become available at:
https://github.com/robvdw/CMIDAT01K-DATA-SCIENCE-for-IOT

IoT Concepts: key Components & Protocols



SENSOR TO GATEWAY COMMUNICATION

Course Setup

Lesson 01:	**Discovering the IoT Data Science Domain**
Lesson 02:	**Defining project requirements**
	+ Cost calculation/estimate
Lesson 03	**Learn to write code**
Lesson 04	**Data Science: How to start your own IoT Project**

Week 09/10: FEEDBACK + GRADING

Lesson Four

HOGESCHOOL ROTTERDAM

LES: 04

DATA Science: How to start your own IoT-project

Preview Les 05

Write your first line of IoT code (Python)

for Rapberry Pi + Things Speak

Data Science

HOGESCHOOL ROTTERDAM

THE [Data Science] HYPE

There's huge and growing demand especially in business

Predicting the future take a lot of effort & expertise

Because of the hype, everyone wants to "own" data science!
- many of them are just selling their stuff with a new label

Here, I will ignore the hype and talk about Data
and how to use it for IoT-projects

Data Science is about:

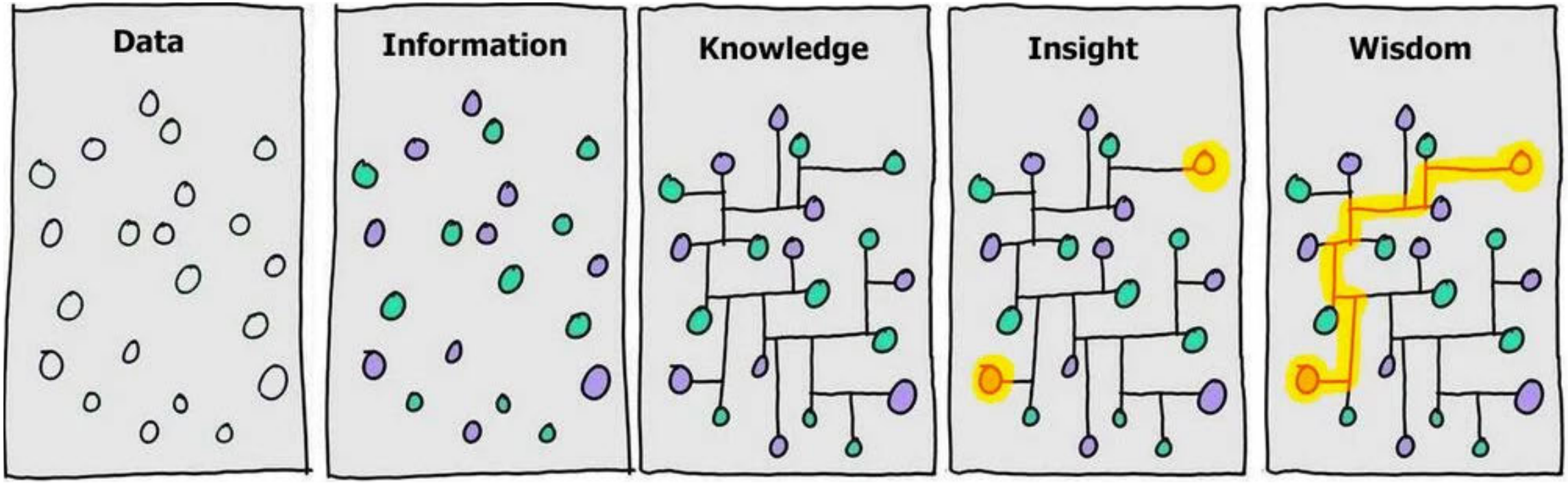
A Data Scientist can:

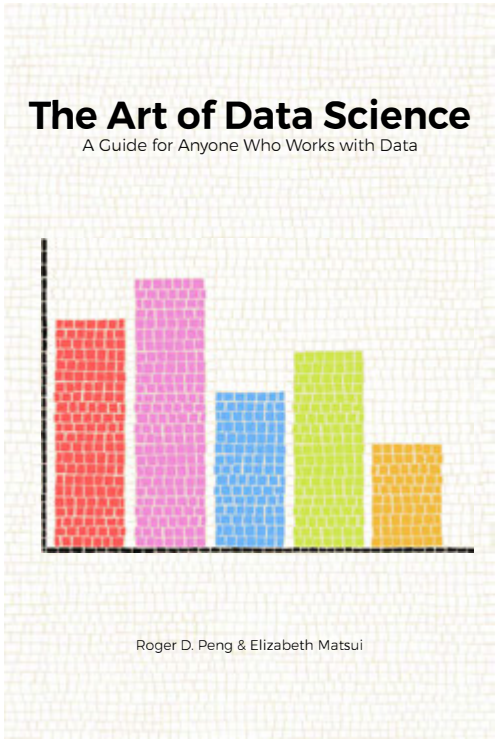
- *understand* the background domain
- *design* solutions that produce added value to the organization
- *implement* the solutions efficiently
- *communicate* the findings clearly (important!)

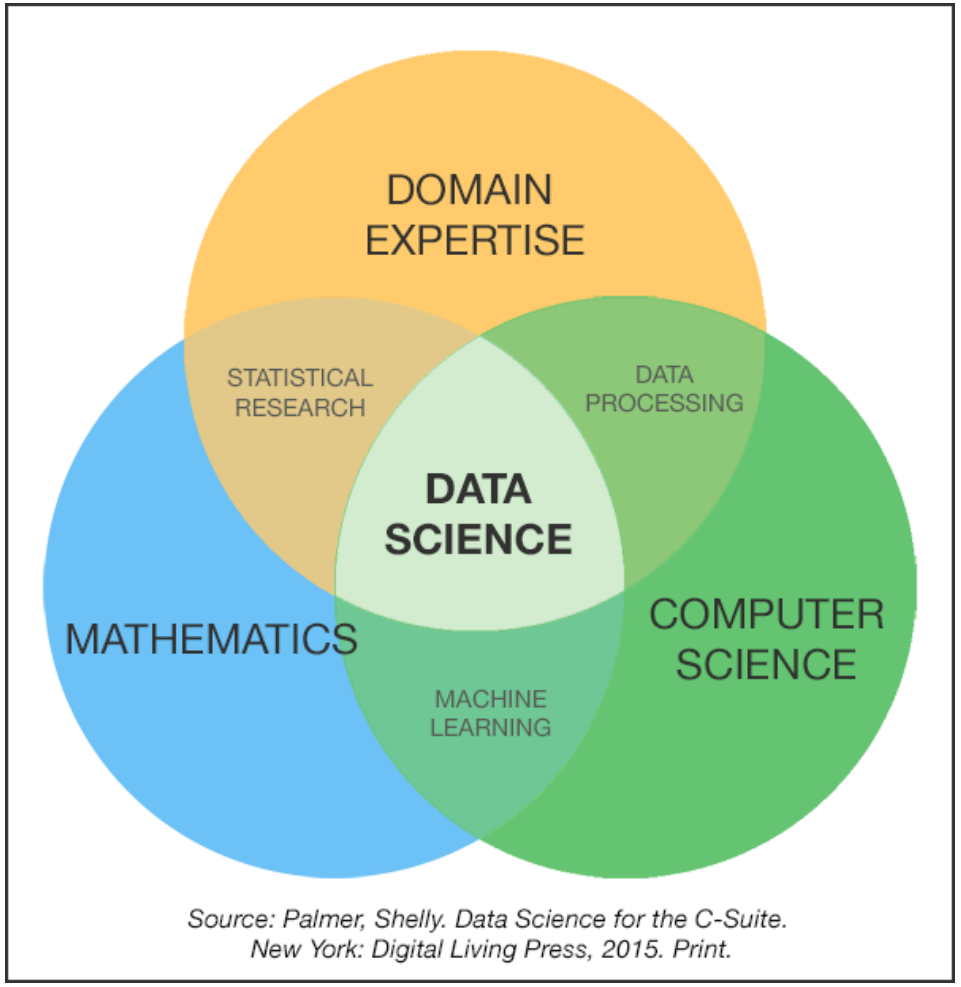
Data Scientist is a *practitioner* with sufficient expertise in software engineering, statistics/machine learning, **and** the application domain.

Making sense of Data (Types)

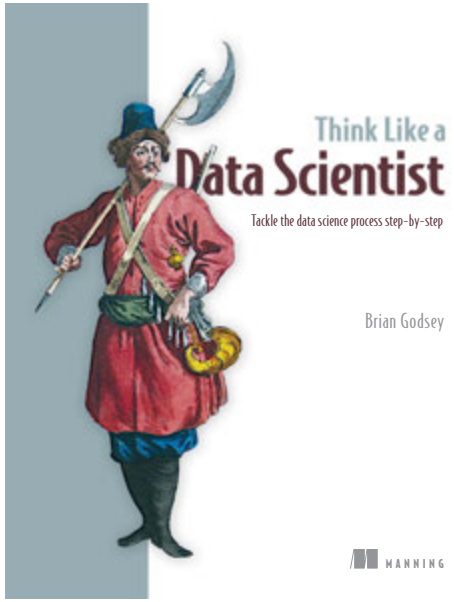
How to make sense of IoT Data (Science)

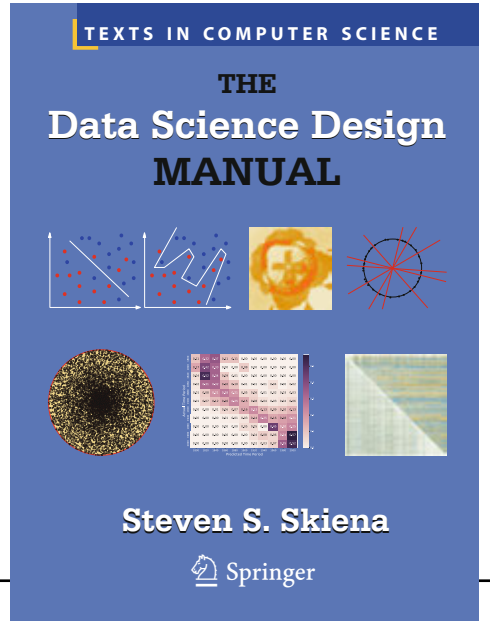




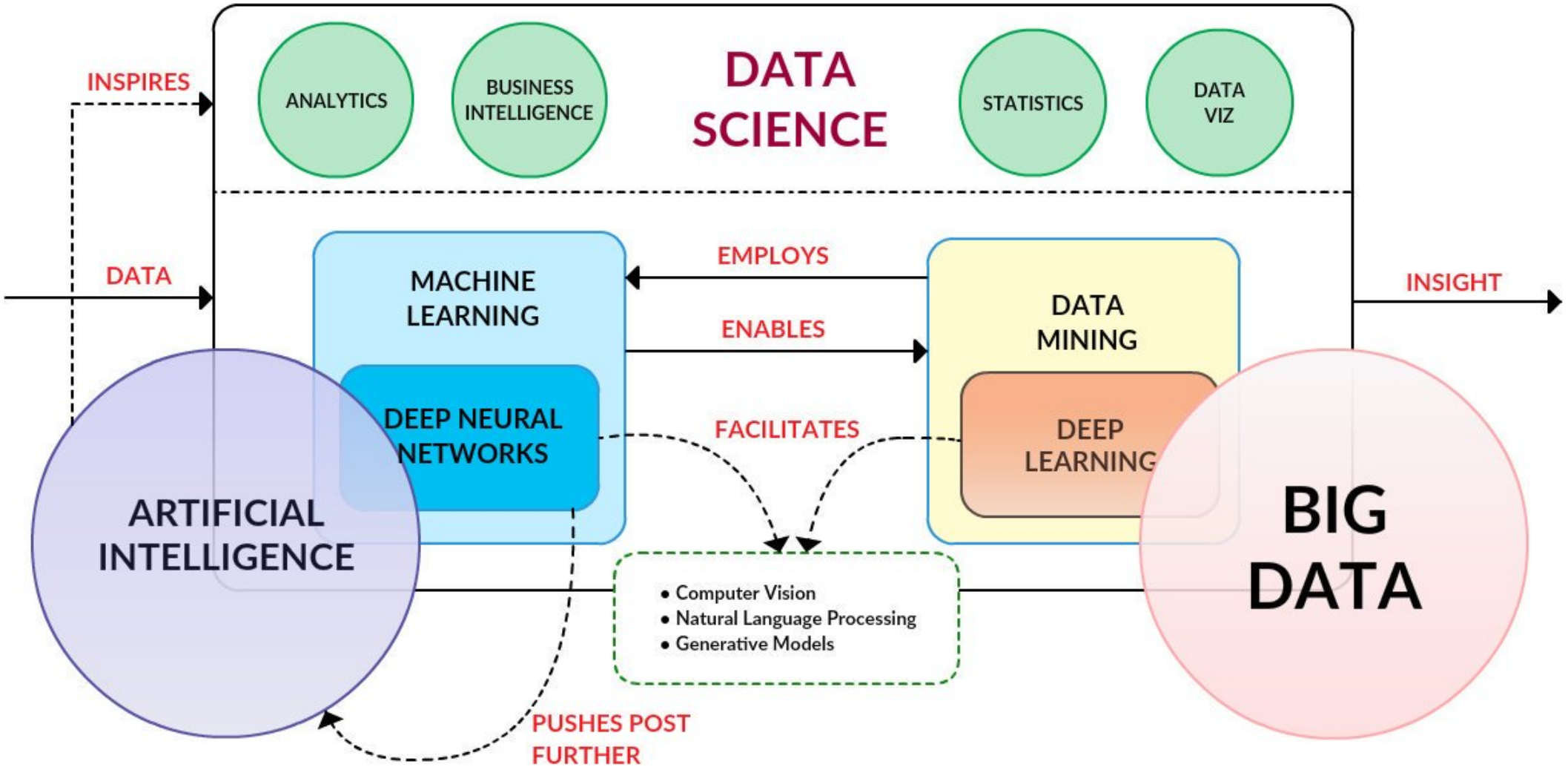


Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

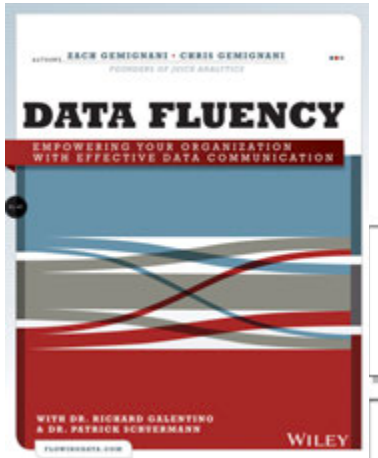


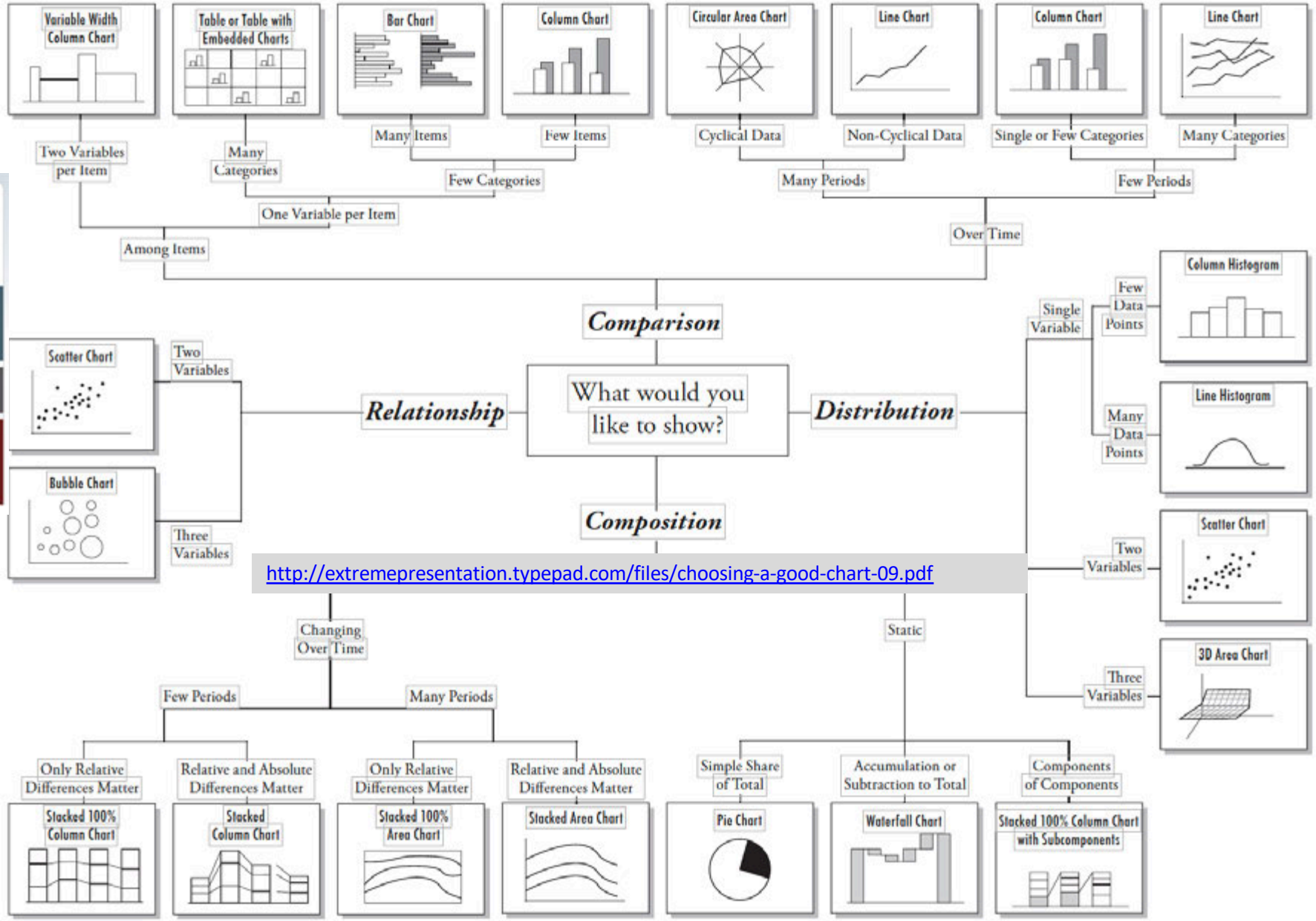


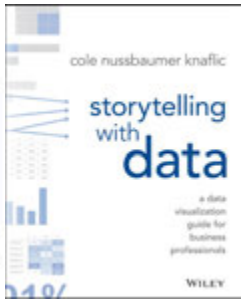
Wat geeft grip op DATA?



http://labs.juiceanalytics.com/chartchooser/index.html



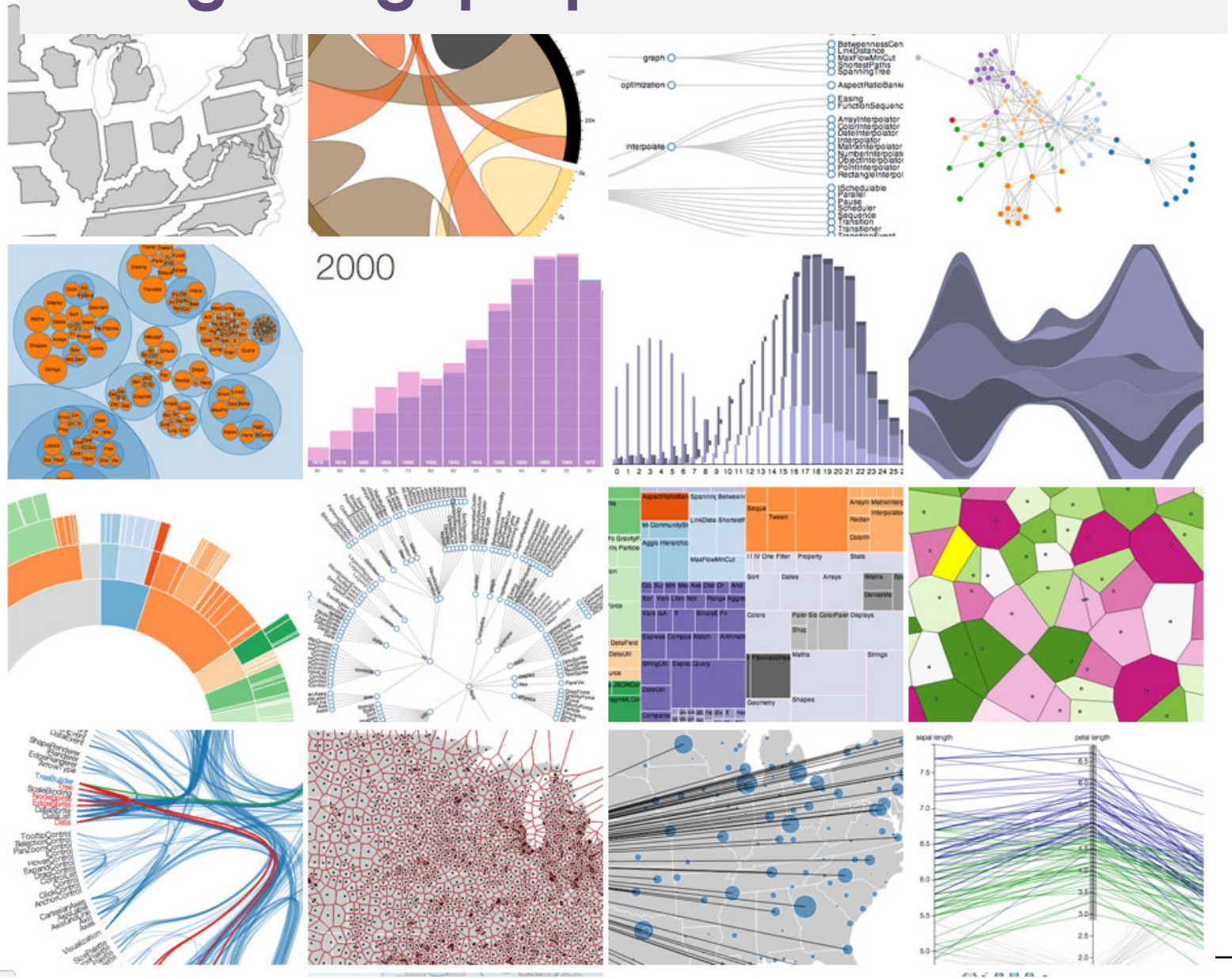


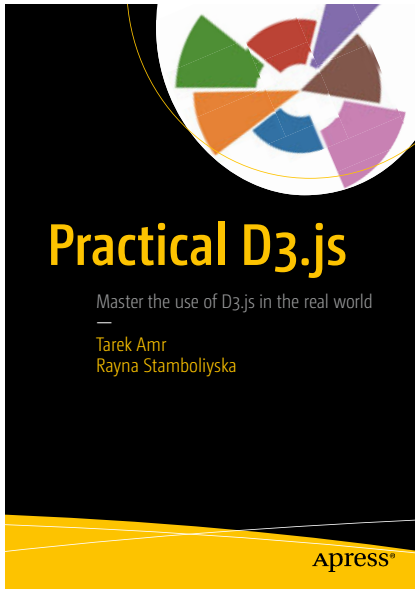


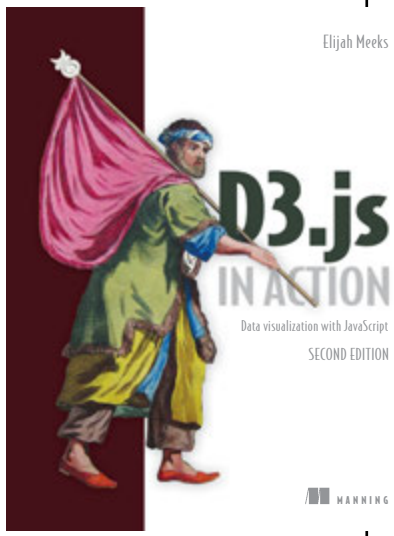
Wat geeft grip op DATA?

https://101innovations.wordpress.com/workflows/

Wat geeft grip op DATA?



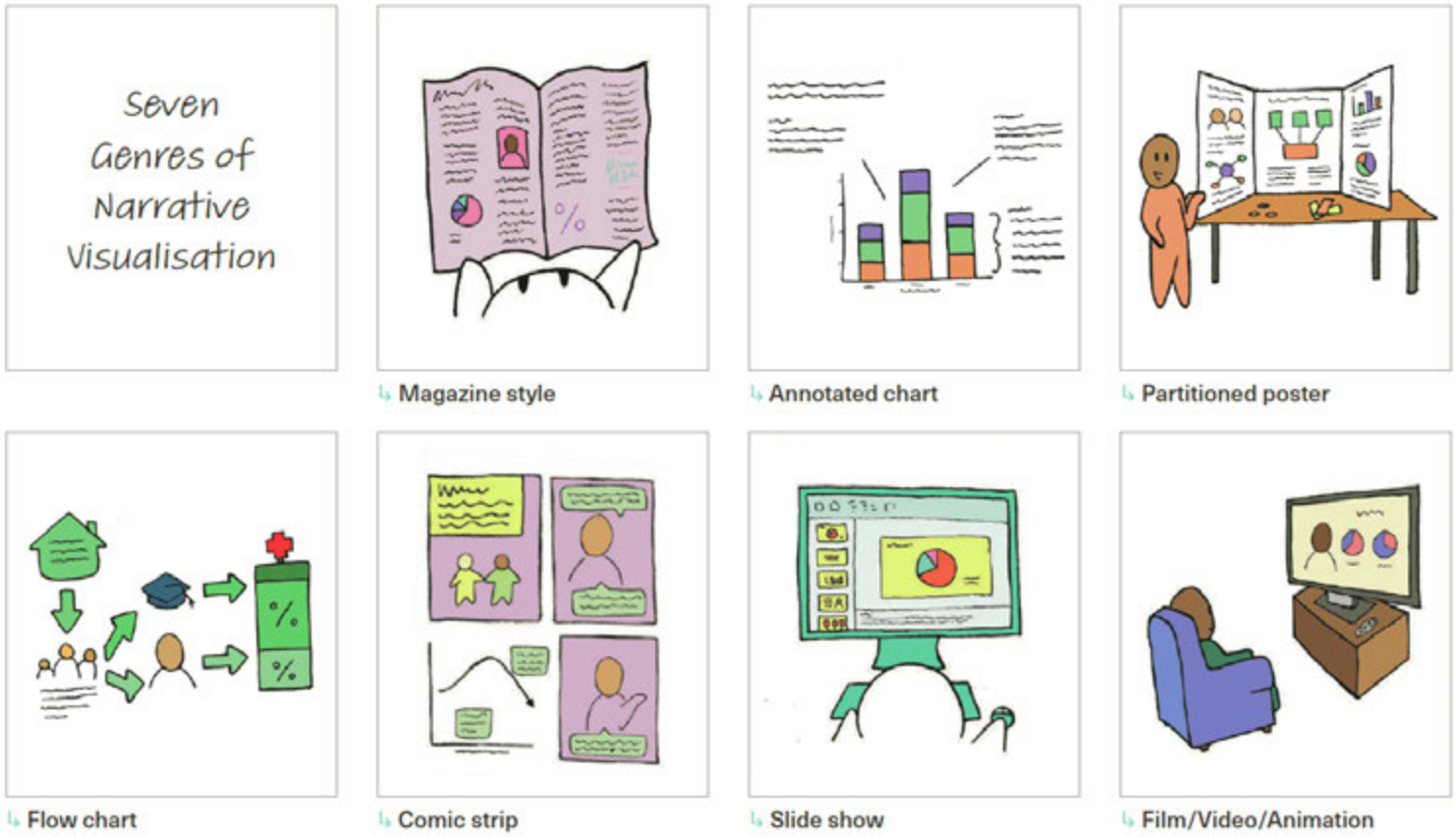




Data Types

HOGESCHOOL ROTTERDAM

Data VIZ ➔ Top-down **Cognition**
requires a Narrative (Story Telling)



DATA-DRIVEN IoT: WHAT IS DATA?

HOGESCHOOL ROTTERDAM

Data [gegevens]

Raw Facts
No Context
Numbers
Symbols

Data comes from the Latin word, "datum," meaning a "thing given."

Although the term "data" has been used since as early as the 1500s, modern usage started in the 1940s and 1950s as practical electronic computers began to input, process, and output data.

98734975471894614398734578
20875980542158009258202908
12349823094823048002343423

98734975471894614398734578
20875980542158009258202908
12349823094823048002343423

HOGESCHOOL ROTTERDAM

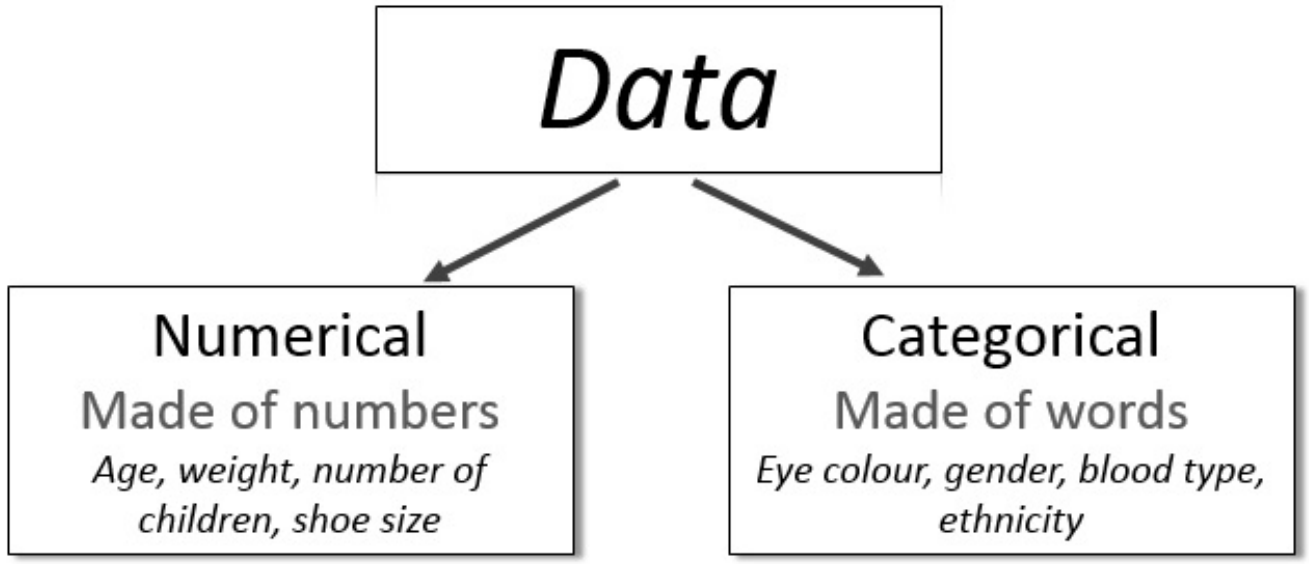
Can you find the the **mistake**?

1 2 3 4 5 6 7 8 9

TYPES OF DATA: Quantitative versus Qualitative [numerical vs categorical]

Data Quantification

Quantitative [Numerical] data: This data can be described using numbers, and basic mathematical procedures, including addition, are possible on the set. It can be discrete (countable numbers) or continuous (infinitely large or small)
Qualitative [Categorical] data: This data are categories. It cannot be described using numbers and basic mathematics. Is generally thought of as being described using"natural" categories and language.



- Quantitative values
  - **Measure** things
  - *Revenue, Units, Marketshare, Duration, Customer Satisfaction, Visits, Price, etc.*
- Categorical values
  - Subdivide things into **groups**
  - *Region, product, category, employee, etc.*

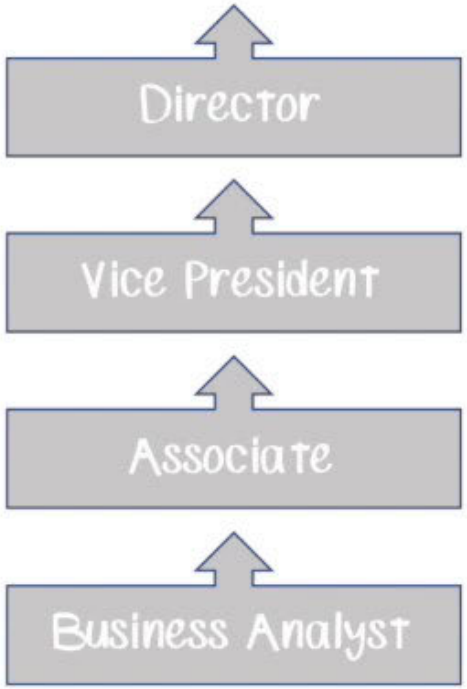

Data
Qualitative
Descriptive information
"I drink coffee every day"
Quantitative
Numerical information
Discrete
(Counted)
4
"I drink 4 coffees every day"
Continuous
(Measured)
80g
"I drink 80grs of coffee every day"

CATEGORICAL DATA

NOMINAL DATA

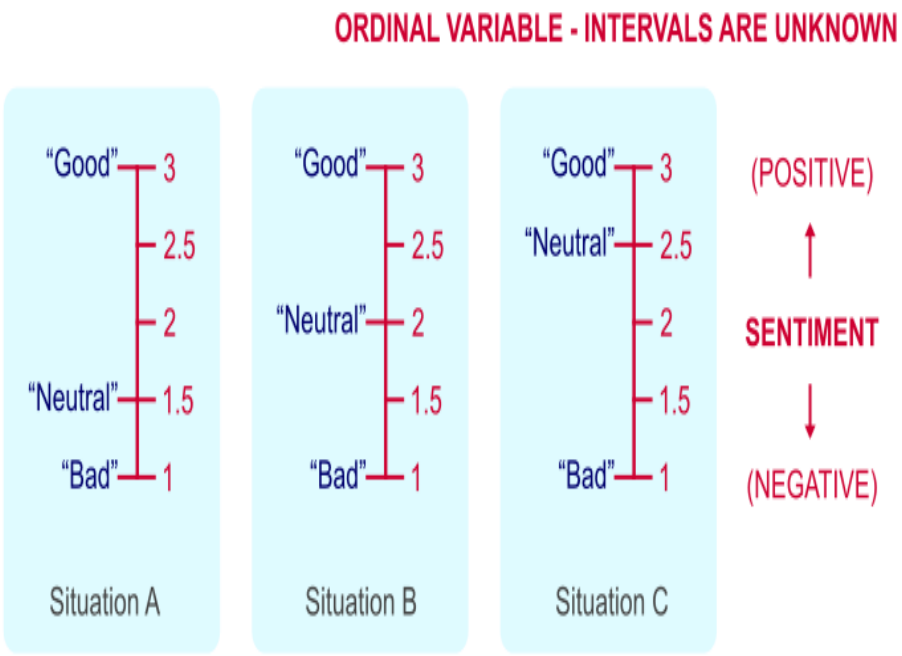
hair color

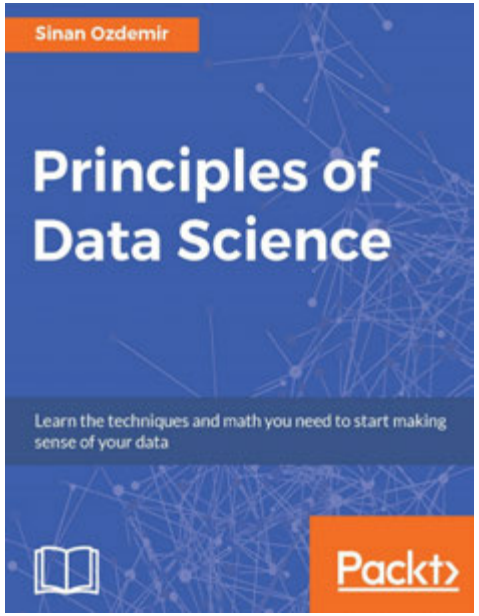
ORDINAL DATA



job hierarchy

Likert-scale data





Data

Numerical
Made of numbers

Age, weight, number of children, shoe size

Continuous
Infinite options

Age, weight, blood pressure

Discrete
Finite options

Shoe size, number of children

Categorical
Made of words

Eye colour, gender, blood type, ethnicity

Ordinal
Data has a hierarchy

Pain severity, satisfaction rating, mood

Nominal
Data has no hierarchy

Eye colour, dog breed, blood type

LEVELS OF DATA: LEVELS OF MEASUREMENTS/OBSERVATIONS

We onderscheiden 4 meetniveaus:

nominaal + ordinaal [discrete data]

interval + ratio [continue data]

Meetniveaus / Meetschalen:

Wanneer je onderzoek doet heb je vaak variabelen die je hierin moet verwerken.

Variabelen zijn elementen uit een onderzoek die verschillende waarden kunnen aannemen. Deze waarden kunnen worden gecategoriseerd in verschillende meetniveaus.

Meetniveaus kunnen iets vertellen over welke data-analyse geschikt is voor structurering.

LEVELS OF DATA: LEVELS OF MEASUREMENTS/OBSERVATIONS

Meetniveau	Wat je kunt berekenen met behulp van waarden op het meetniveau
Nominaal	Tellen, percentages berekenen
Ordinaal	Tellen, percentages berekenen en hoger/lager aangeven
Interval	Tellen, hoger/lager aangeven, verschillen in eenheden aangeven, gemiddelde, spreiding
Ratio	Tellen, hoger/lager aangeven, verschillen in eenheden aangeven, gemiddelde, spreiding en het berekenen van verhoudingen

LEVELS OF DATA: LEVELS OF MEASUREMENTS/OBSERVATIONS

HOGESCHOOL ROTTERDAM

Meetniveaus [level] /Meetschalen [scale]:

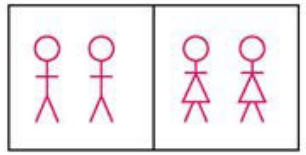
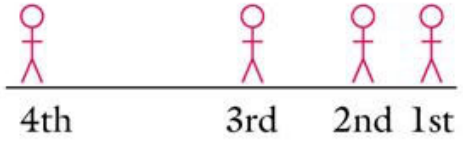
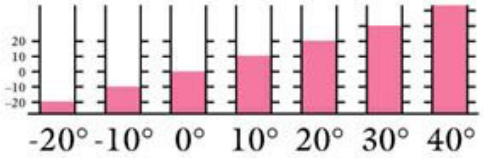
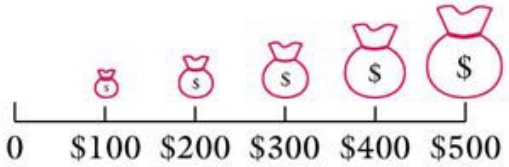
De hoogte van het meetniveau is bepalend voor:

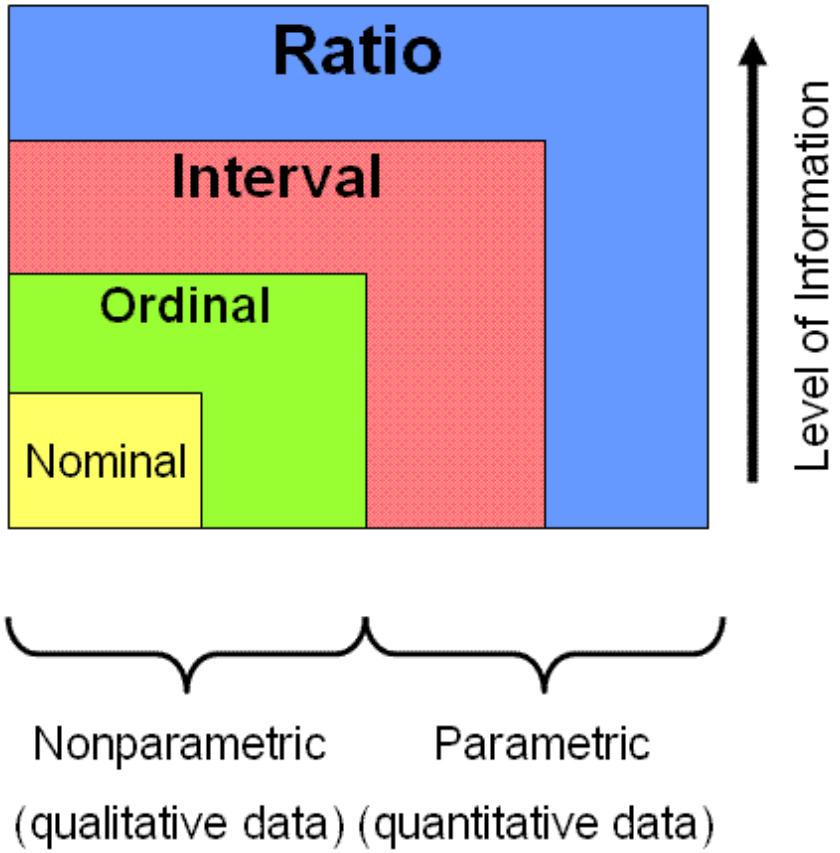
Statische-analyse + Grafische weergave

Meetniveaus & hun kenmerken

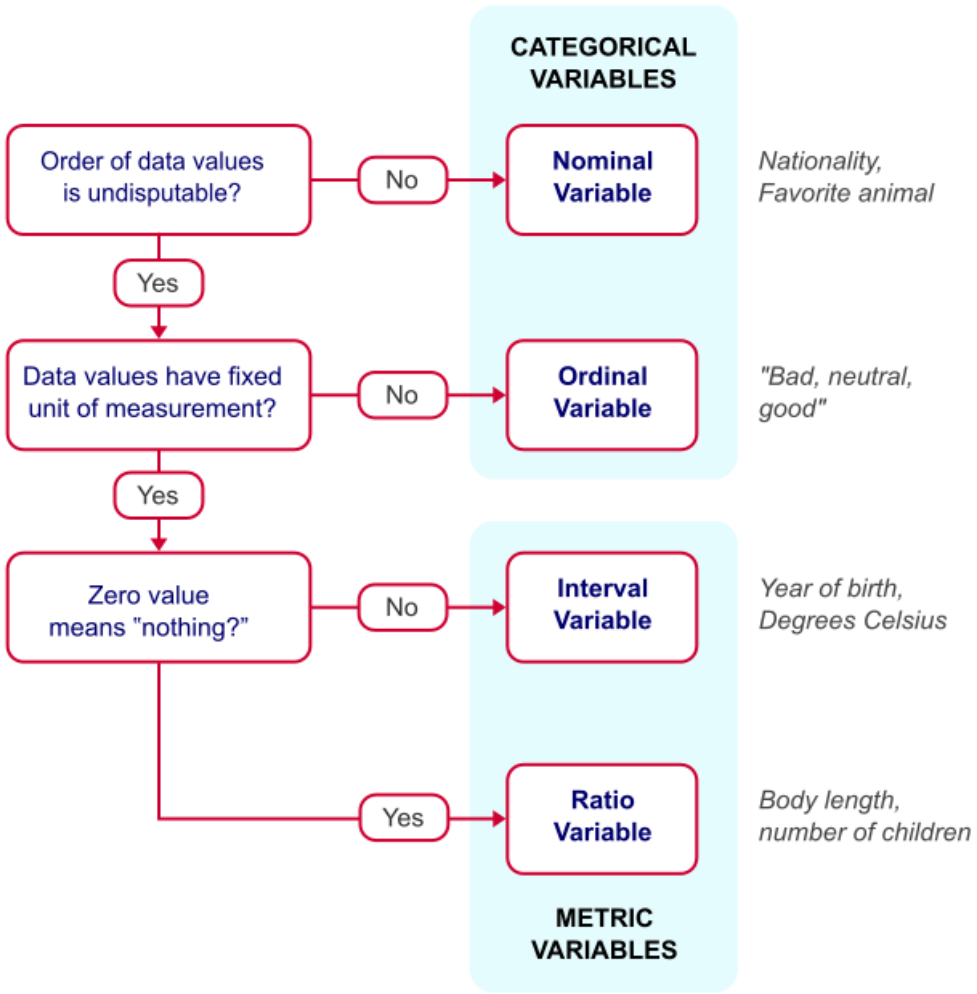
		Scale	**Rationiveau**
		Intervalniveau	Verhouding blijven gelijk
	Ordinaal niveau	Gelijke verschillen	Gelijke verschillen
Nominale niveau	Ordening	Ordening	Ordening
Onderscheid	Onderscheid	Onderscheid	Onderscheid
Geslacht	*Opleidingsniveau*	*Intelligentie*	*Leeftijd*

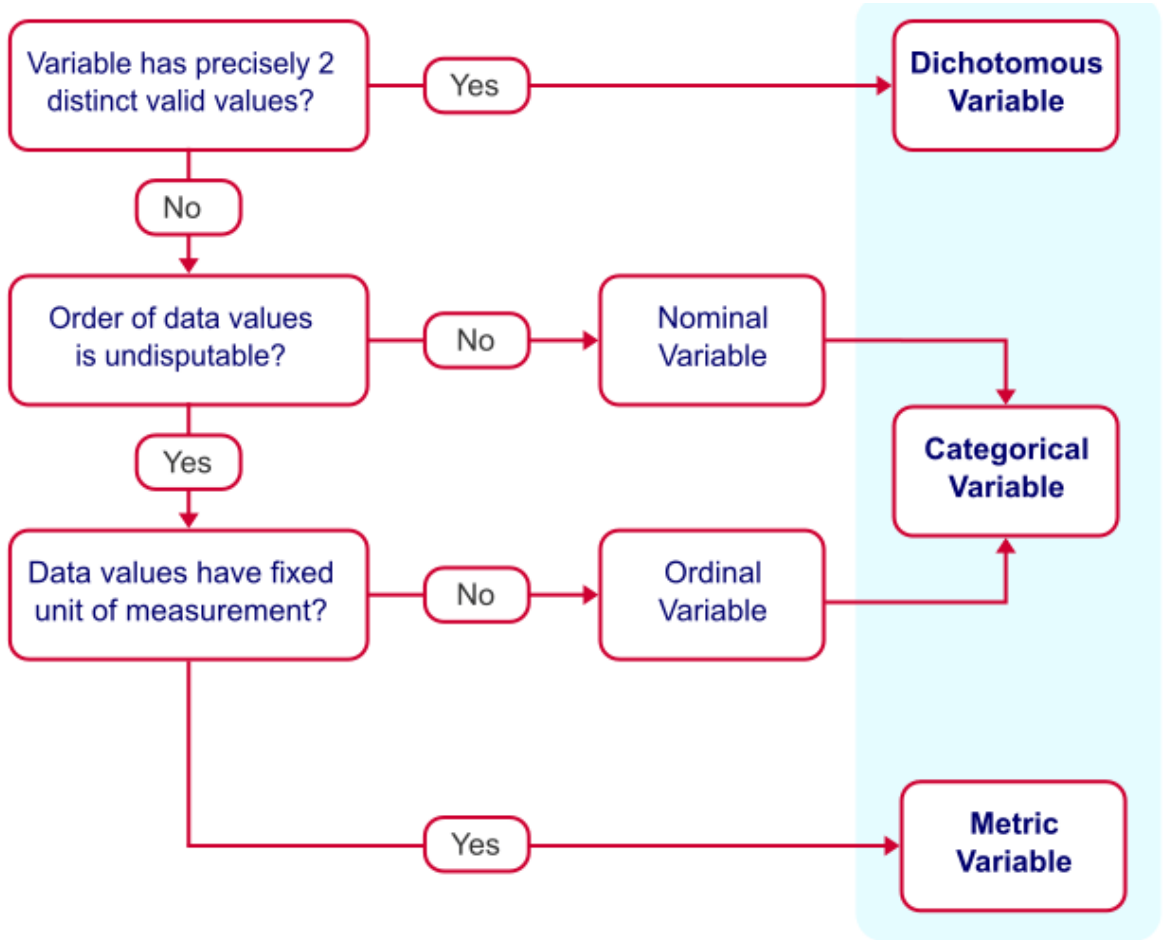
LEVELS OF DATA: LEVELS OF MEASUREMENTS/OBSERVATIONS

SCALE		EXAMPLE
Nominal		Gender
Ordinal		Position in race
Interval		Temperature (in Fahrenheit)
Ratio		Money



LEVELS OF DATA: What data to measure or observe?





LEVELS OF DATA: What data to measure or observe?

Differences between measurements, true zero exists

Ratio Data

Differences between measurements but no true zero

Interval Data

Quantitative Data

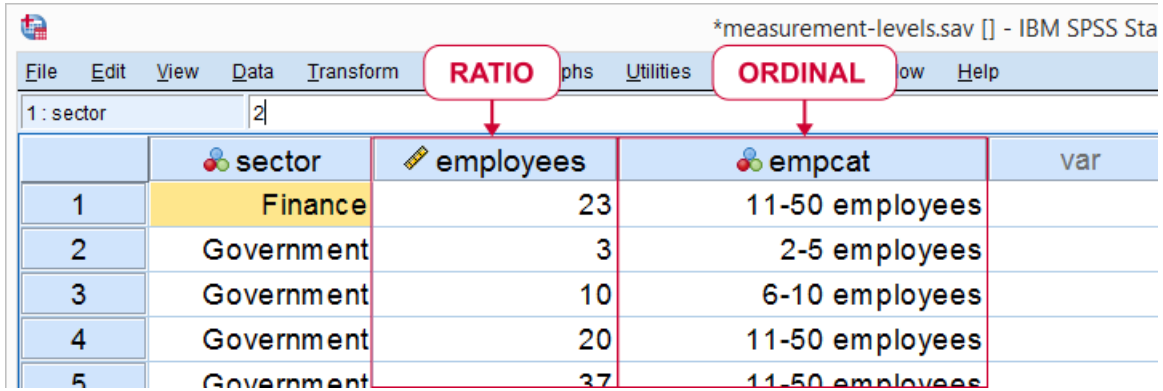
Ordered Categories (rankings, order, or scaling)

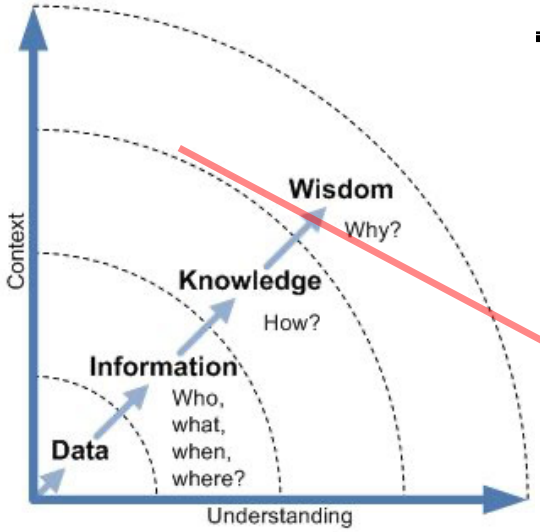
Ordinal Data

Categories (no ordering or direction)

Nominal Data

Qualitative Data





Context Independent

Wisdom is the recognition that knowledge patterns arise from fundamental principles and the understanding of what those principles are.

Wisdom

UNDERSTANDING PRINCIPLES

Knowledge is represented by patterns among data, information and possibly other knowledge. These patterns don't actually constitute knowledge until they are understood.

Knowledge

UNDERSTANDING PATTERNS

Information represented by relationships between data and possibly other information

Information

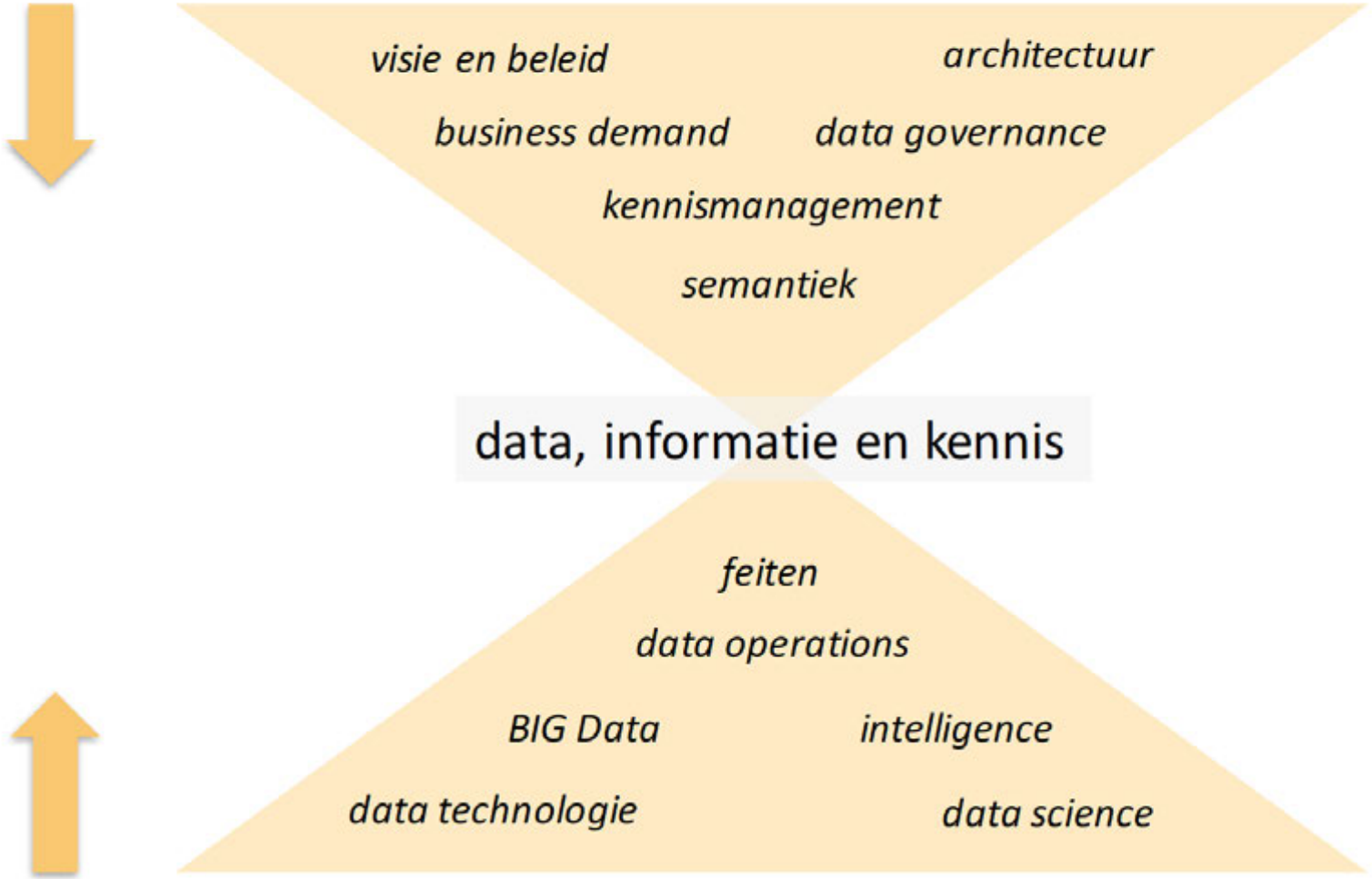
UNDERSTANDING RELATIONS

Data is an item or event out of context with no relation to Other things.

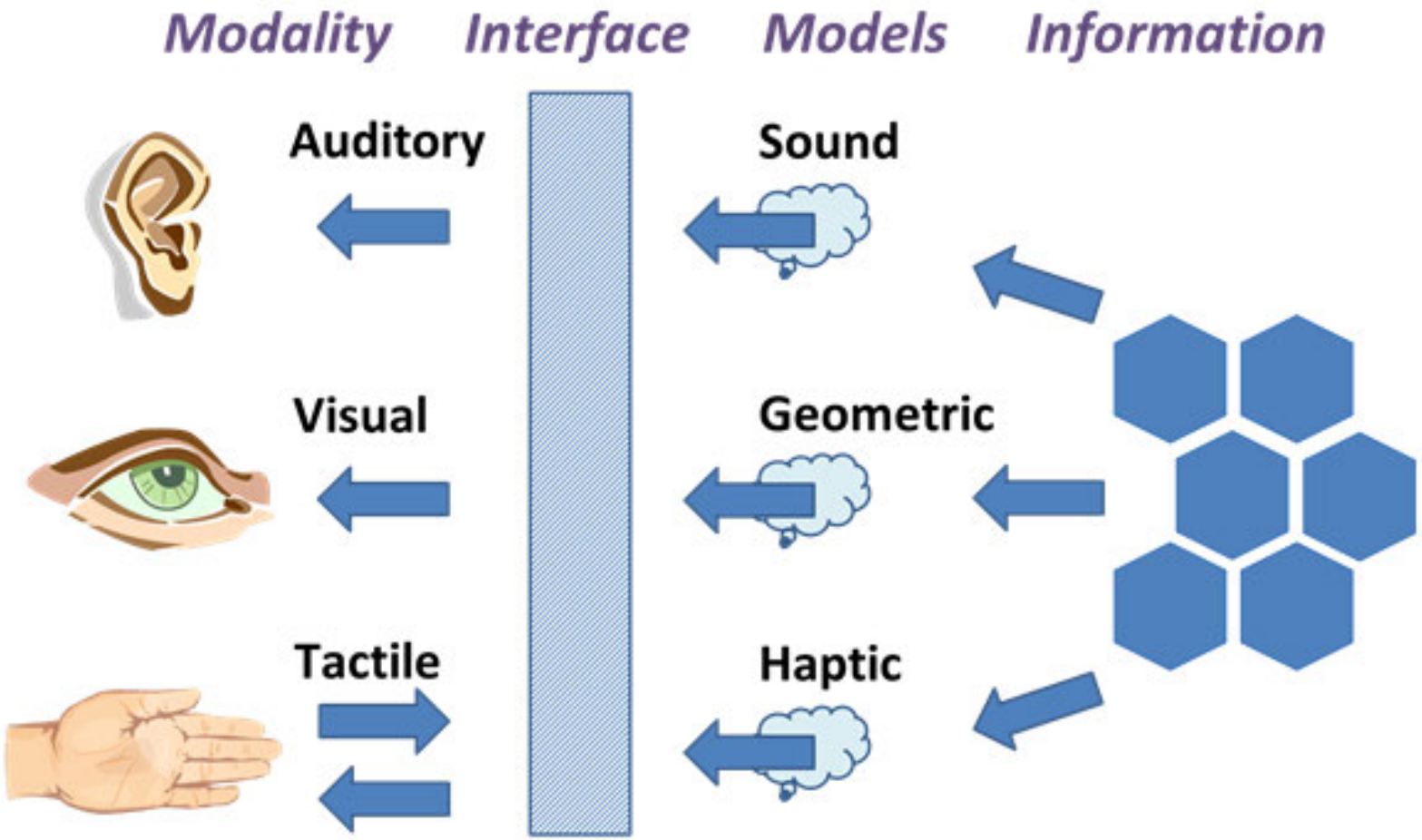
Data

Understanding

Data, informatie en kennis is wat ons verbindt



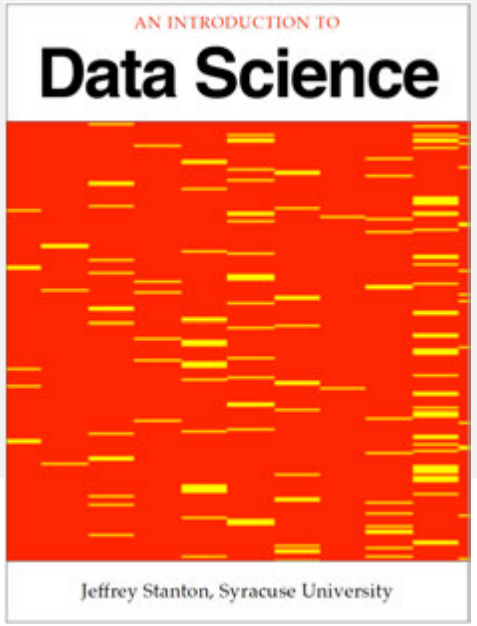
http://intelligence.agconnect.nl/content/van-data-naar-informatie


Modality
Interface
Models
Information
Auditory
Visual
Tactile
Sound
Geometric
Haptic

DATA-DRIVEN: Data versus Information

Data [gegevens]

Raw Facts
No Context
Numbers
Symbols

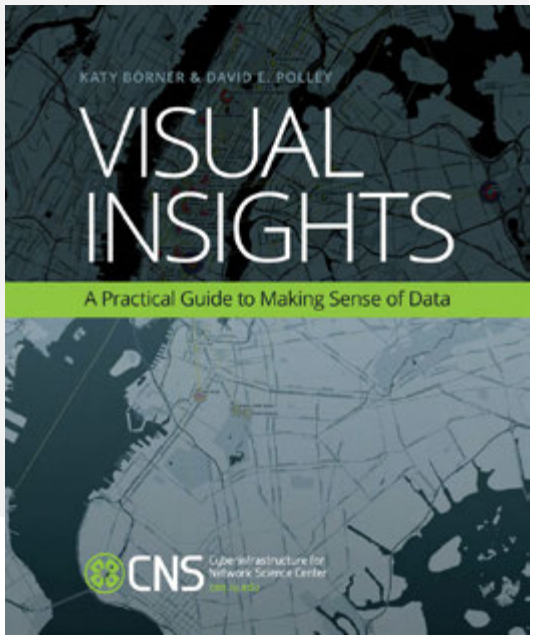


Information

Data with structure = processed data

Value-added to Data through

- Summarisation
- Organisation
- Analysing



TYPES OF DATA: Structured vs Unstructured [organized vs unorganized]

Data Structuring

Structured (organized) data: This is data that can be thought of as observations and characteristics. It is usually organized using a table method (rows and columns).

Unstructured (unorganized) data: This data exists as a free entity and does not follow any standard organization hierarchy.

Structured Data
High Degree of organization, such as a relational database

Column	Value
Patient	Joe Brown
Date of Birth	02/13/1972
Date Admitted	02/05/2014

Unstructured Data
Information that is difficult to organize using traditional mechanisms

"The patient came in complaining of chest pain, shortness of breath, and lingering headaches…smokes 2 packs a day… family history of heart disease…has been experiencing similar symptoms for the past 12 hours…."

DATA STRUCTURING: Generalized Form of a Data Table

Data Table [DATA MATRIX]

A generalized version of the data table is shown.

This table can represent any number of observations described over multiple variables.

This table describes a series of observations (from o1 to on) where each observation is described using a series of variables (from x1 to xp). A value is provided for each variable of each observation.

		Variables				
Observations		x_1	x_2	x_3	...	x_p
	o_1	x_{11}	x_{12}	x_{13}	...	x_{1p}
	o_2	x_{21}	x_{22}	x_{23}	...	x_{2p}
	o_3	x_{31}	x_{32}	x_{33}	...	x_{3p}
	...	...	...	...	...	...
	o_n	x_{n1}	x_{n2}	x_{n3}	...	x_{np}

Most data that exists in text form, including server logs and Facebook posts, is unstructured

Scientific observations, as recorded by careful scientists, are kept in a very neat and organized (structured) format: THE DATA TABLE

A genetic sequence of chemical nucleotides [ACGTATTGCA] is unstructured even if the order of the nucleotides matters

DATA STRUCTURING: Observations versus Variables

Data Table [DATA MATRIX]

A generalized version of the data table is shown. This table can represent any number of observations described over multiple variables.

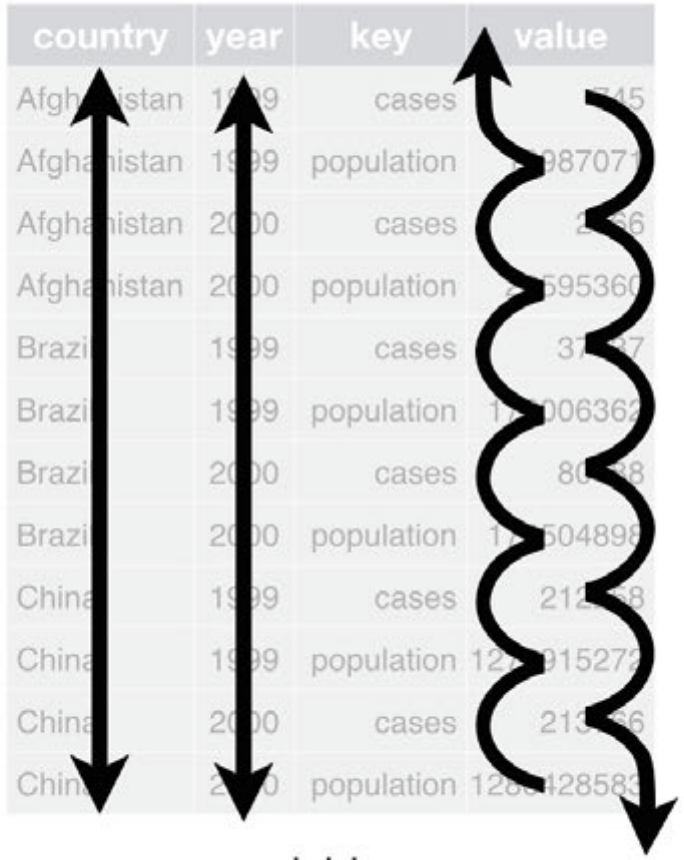
This table describes a series of observations (from o1 to on) where each observation is described using a series of variables (from x1 to xp). A value is provided for each variable of each observation.

		Variables				
Observations		x_1	x_2	x_3	...	x_p
	o_1	x_{11}	x_{12}	x_{13}	...	x_{1p}
	o_2	x_{21}	x_{22}	x_{23}	...	x_{2p}
	o_3	x_{31}	x_{32}	x_{33}	...	x_{3p}
	...	...	...	...	...	...
	o_n	x_{n1}	x_{n2}	x_{n3}	...	x_{np}

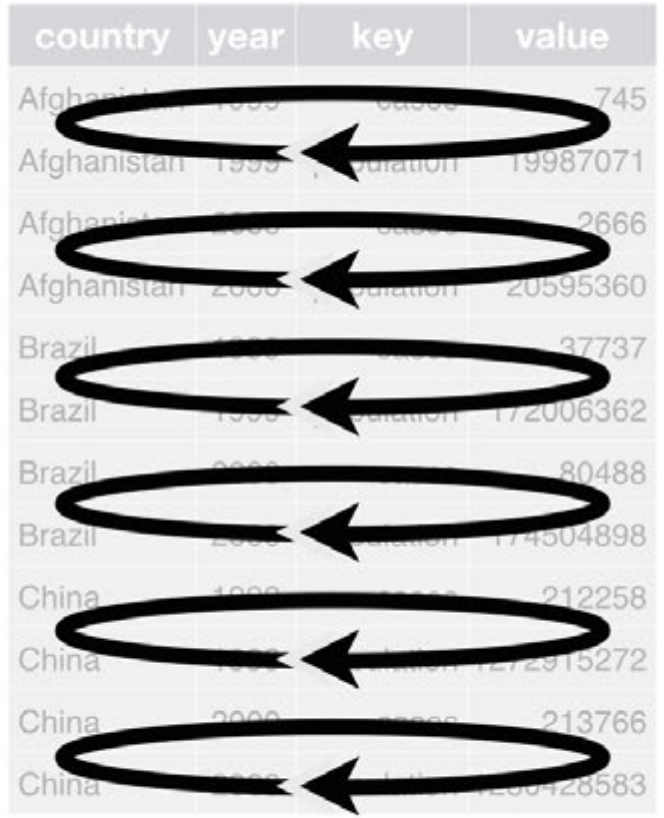
Patient ID	Treated	Age	Outcome	Random
1	Yes	Young	Positive	0.24
2	No	Young	Positive	0.85
3	Yes	Old	Negative	0.64
4	No	Old	Negative	0.70
5	No	Old	Negative	0.87
6	No	Old	Negative	0.72
7	No	Old	Negative	0.86
8	No	Young	Negative	0.16
9	No	Young	Positive	0.17

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2



variables



observations

STRUCTURED DATA FORMATS

- CSV, comma separated values

```
sepal_length,sepal_width,petal_length,petal_width,species
5.1,3.5,1.4,0.2,setosa
4.9,3,1.4,0.2,setosa
4.7,3.2,1.3,0.2,setosa
4.6,3.1,1.5,0.2,setosa
```

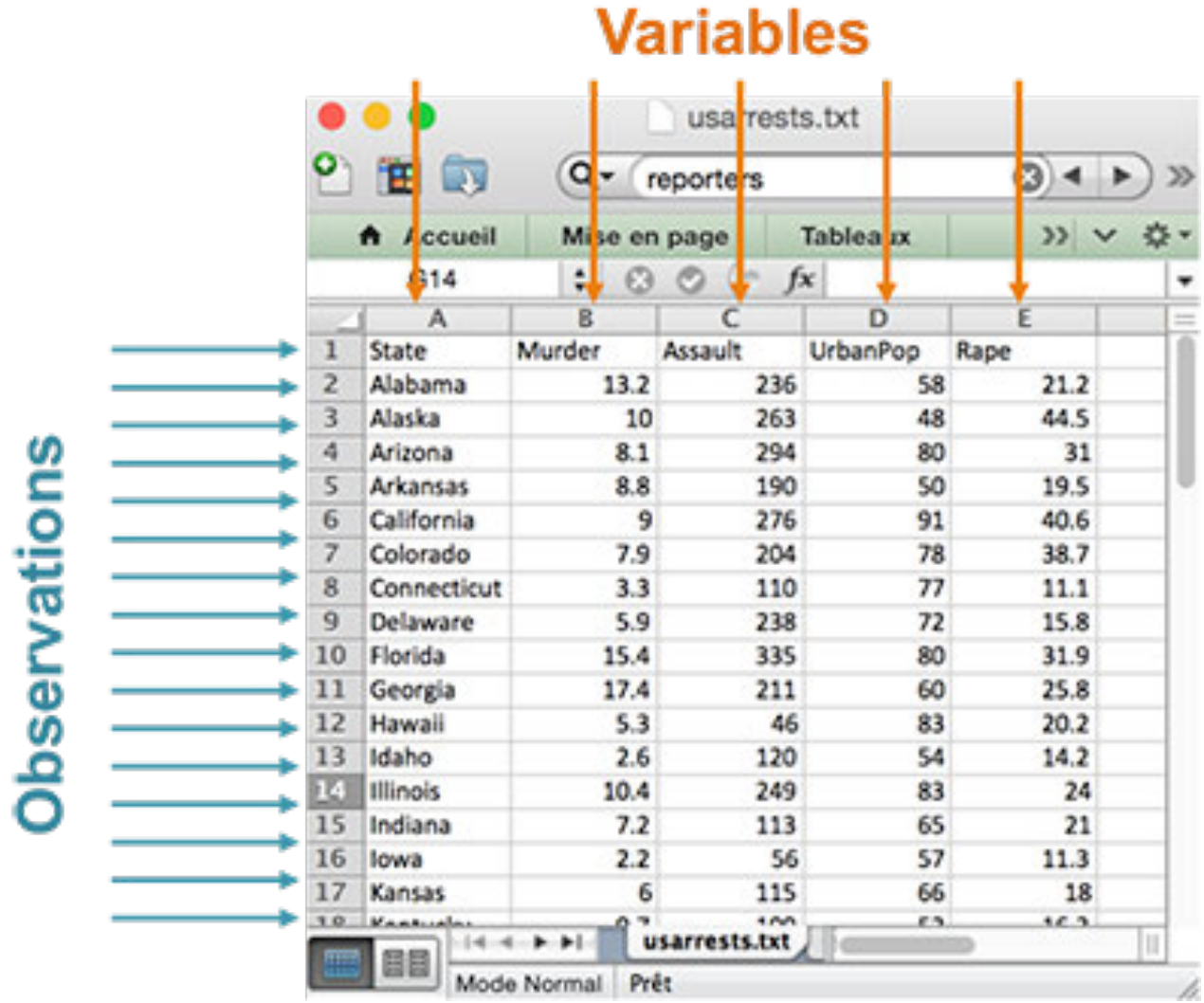
- hierarchies, e.g., Newick tree format

```
(A,B,(C,D)E)F;
```

- networks, e.g., GraphViz (DOT)

```
digraph graphname {
    a -> b -> c;
    b -> d;
}
```

Tidy data



	A	B	C	D	E
1	State	Murder	Assault	UrbanPop	Rape
2	Alabama	13.2	236	58	21.2
3	Alaska	10	263	48	44.5
4	Arizona	8.1	294	80	31
5	Arkansas	8.8	190	50	19.5
6	California	9	276	91	40.6
7	Colorado	7.9	204	78	38.7
8	Connecticut	3.3	110	77	11.1
9	Delaware	5.9	238	72	15.8
10	Florida	15.4	335	80	31.9
11	Georgia	17.4	211	60	25.8
12	Hawaii	5.3	46	83	20.2
13	Idaho	2.6	120	54	14.2
14	Illinois	10.4	249	83	24
15	Indiana	7.2	113	65	21
16	Iowa	2.2	56	57	11.3
17	Kansas	6	115	66	18

KINDS OF DATA

- STRUCTURED DATA
  - lists
  - $n \times p$ tables, arrays
  - hierarchies (e.g., organization chart)
  - networks (e.g., travel routes, hypertext = links)
- Generic data-interchange formats: XML, JSON

- UNSTRUCTURED DATA
  - text
  - images
  - video
  - sound
- Often can be made structured by, e.g., parsing language, segmenting images, etc.

BIG DATA

- A crucial part of the rise of Data Science is the steep increase in the amount and availability of data

- Big Data refers not only to the quantity but also to the quality of the data:
  - VOLUME: lots of it
  - VELOCITY: fast (streaming)
  - VARIETY: all kinds, not nice and "clean"
  - VERACITY: can it be trusted?

JSON

- Similar to XML but simpler

```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    },
    {
      "type": "mobile",
      "number": "123 456-7890"
    }
  ],
  "children": [],
  "spouse": null
}
```

XML

- same example:

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumber>
    <type>home</type>
    <number>212 555-1234</number>
  </phoneNumber>
  <phoneNumber>
    <type>fax</type>
    <number>646 555-4567</number>
  </phoneNumber>
  <gender>
    <type>male</type>
  </gender>
</person>
```

PARSING

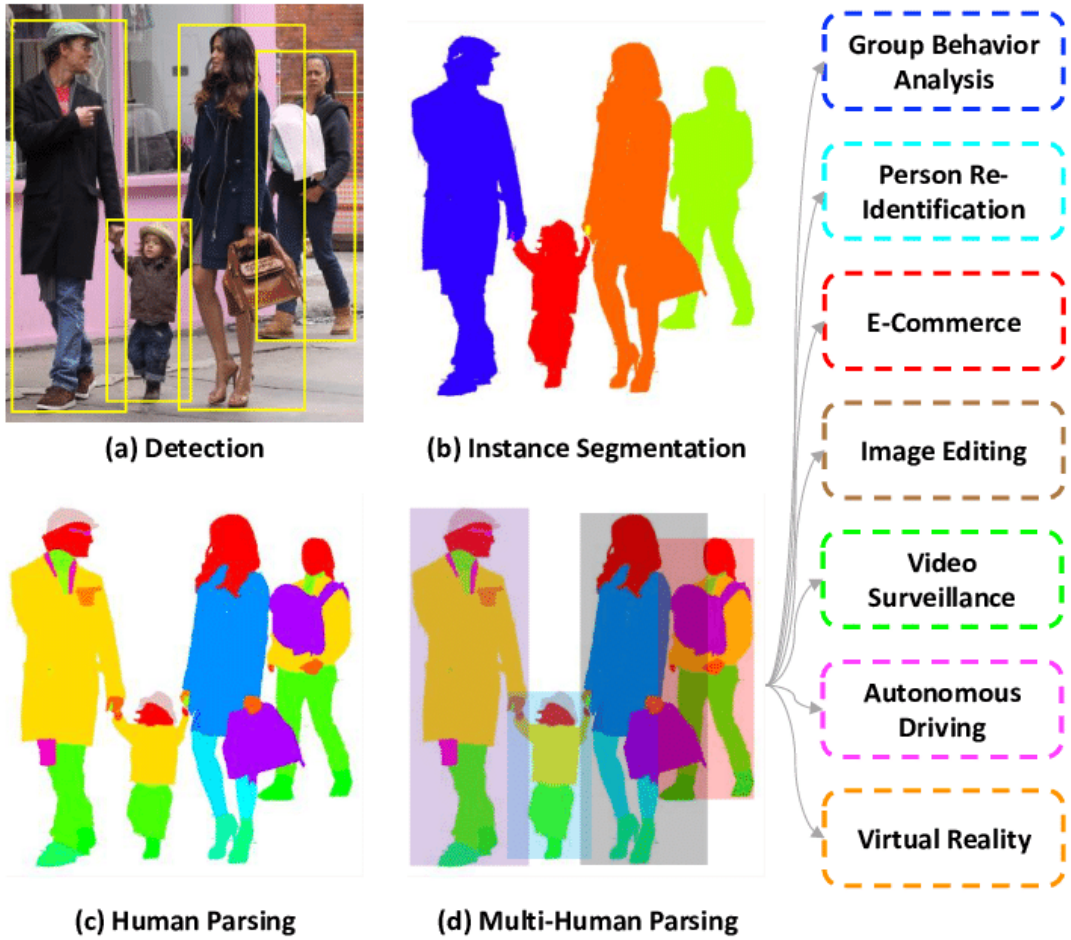
- Given a known grammar, unstructured text data can be parsed
- “It ain’t over till the fat lady sings”

  ((it, (ain't, over)), (till, ((the, (fat, lady)), sings)))

- Similarly, images can be segmented into parts

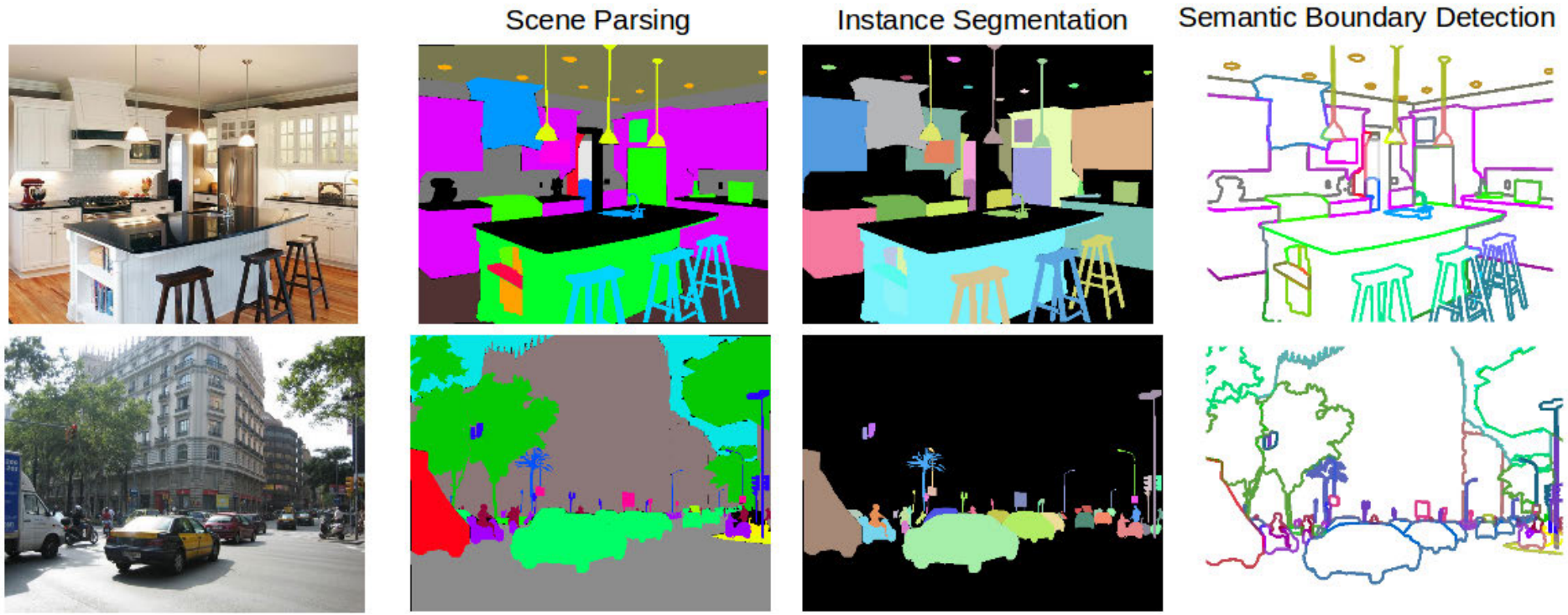
PARSING

- Similarly, images can be segmented into parts



PARSING

- Similarly, images can be segmented into parts



Human Perception

vs

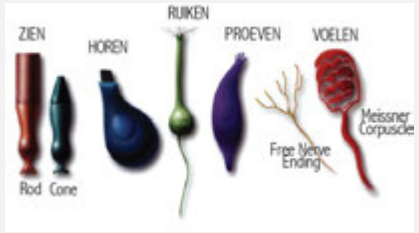
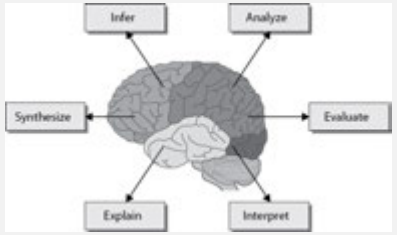
Data (Types)

HOGESCHOOL ROTTERDAM

Human Perception vs Data

Data VIZ
requires insight into human **perception & cognition**

NEUROETHOLOGIE	Neuro-ethologisch perspectief: wat maakt ons humaan?

Gewaarwording & Perceptie	Cognitie & Semiotiek	Gedrag & Communicatie	Theory of Mind (ToM)
			
Biologie / neuro-wetenschappen	Biologie / Neuro-wetenschappen	Biologie / Psychologie	Psychologie / Sociologie

Human Perception vs Data

Data VIZ
requires bottom-up + top-down approach
to human perception and/or Cognition

Analytic or Holistic approach:

From problems to behavior **("top-down")**

What problems must organisms solve?

What strategies do they use to solve these problems?

What mechanisms are used to implement the strategies?

Synthetic approach:

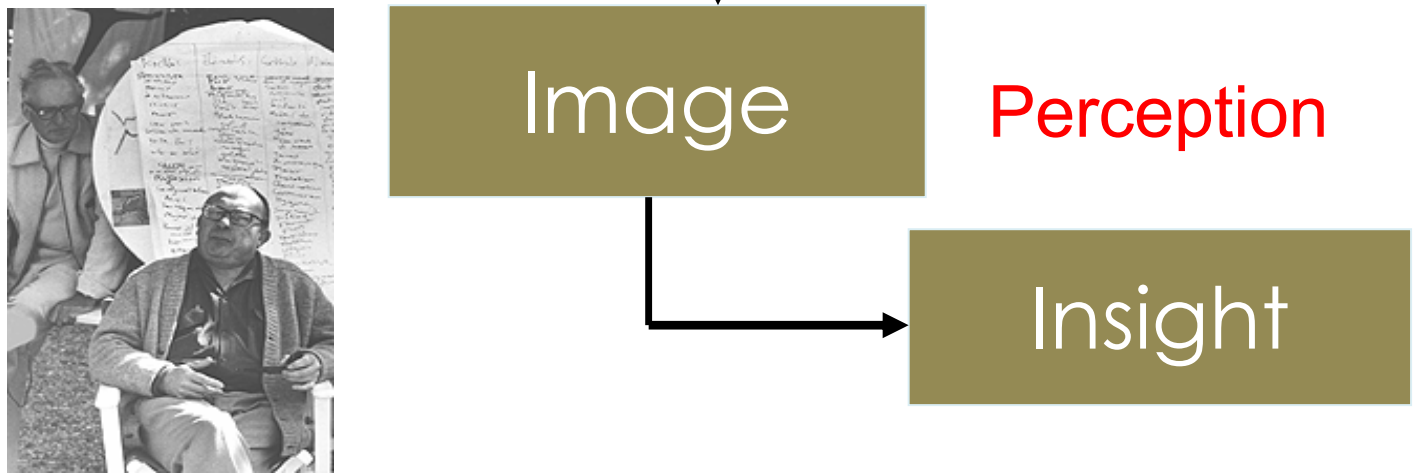
from behavior to causes **("bottom-up")**

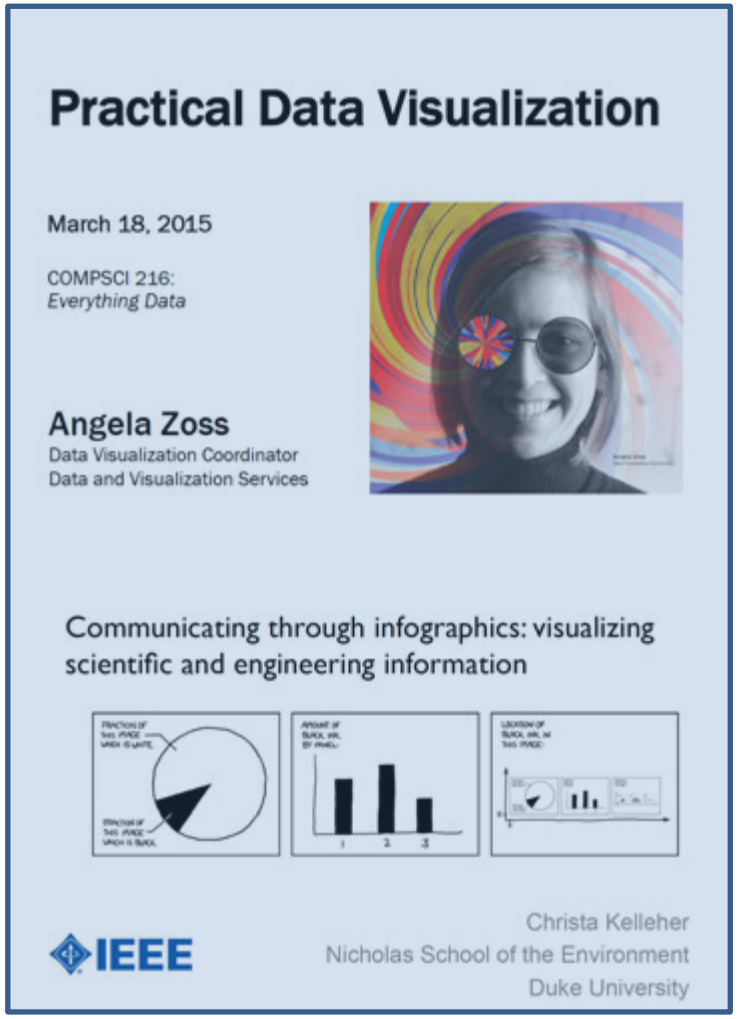
What do organisms do? (Description)
How do they do it? (Proximate causes)
- Developmental mechanisms
- Physiological mechanisms
- Behavioral mechanisms

Why do they do it? (Ultimate causes)
- Functions (adaptive values)
- Evolution (history of the species)

DATA STRUCTURING: data worden pas inzichtelijk als (beeld)figuur (graphical visualization) of GRAAF (graph)

Jacques Bertin who wrote the classic works of **graphical visualization** "Semiology of Graphics" states that the "transformation from numbers to insight requires two stages"





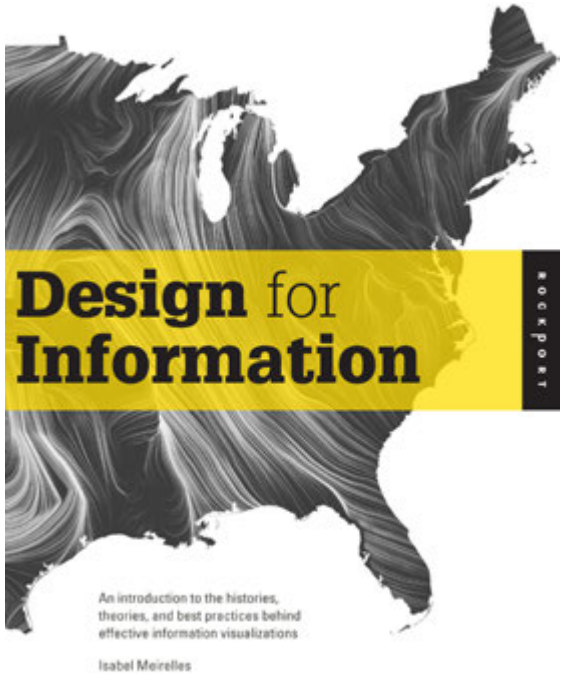
SEE ALSO: http://www.cs.wright.edu/~jgalli/hfe306/Data_Visualization_Quenin.ppt

DATA STRUCTURING: Preattentive Processing [perceptual level]

HOGESCHOOL ROTTERDAM

THREE-STAGE MODEL OF PERCEPTUAL PROCESSING

A schematic overview of the simplified information-processing model of human visual perception proposed by Collin Ware.[14]



Bottom-up information drives pattern building

Top-down attentional processes reinforce relevant information

STAGE 1

Billions of neurons work in parallel to extract millions of **features** that are processed rapidly and simultaneously, such as color, texture, orientation, and so on.

STAGE 2

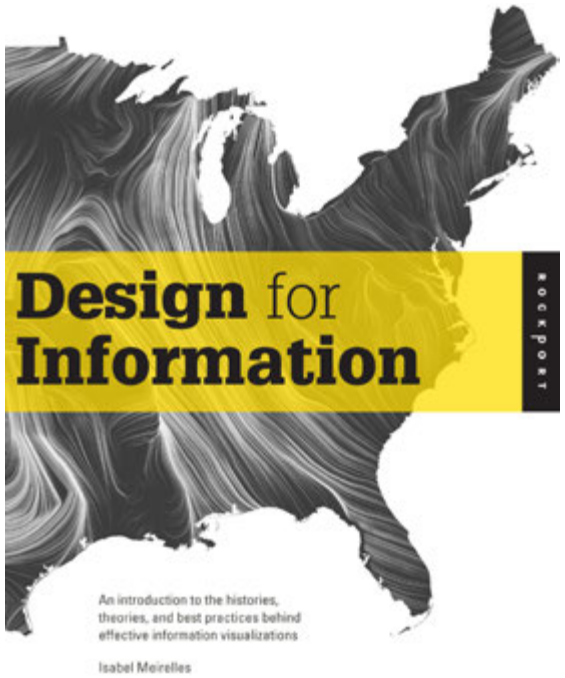
Patterns are extracted serially and slowly, such as regions of the same color, and regions of the same texture. The pattern-finding process leads to two pathways: object perception, and locomotion and action.

STAGE 3

At the highest level of perception, we are able to hold between one and three **objects** at any instance in our working visual memory. Patterns that provide answers to the visual query construct the objects in conjunction with information stored in our long-term memory and that are related to the task at hand.

DATA STRUCTURING: Preattentive Processing [perceptual level]

Stage 1: Rapid parallel processing to extract basic features;
Stage 2: Slow serial processing for extraction of patterns and structures;
Stage 3: Sequential goal-oriented processing with information reduced to a few objects and held in working visual memory to form the basis for visual thinking.



18596746321475030608030504090
70502769843010215346748950213
06057204020503090845064201040
70204070835061305080239245798

18596746321475030608030504090
70502769843010215346748950213
06057204020503090845064201040
70204070835061305080239245798

18596746321475030608030504090
70502769843010215346748950213
06057204020503090845064201040
70204070835061305080239245798

18596746321475030608030504090
70502769843010215346748950213
06057204020503090845064201040
70204070835061305080239245798

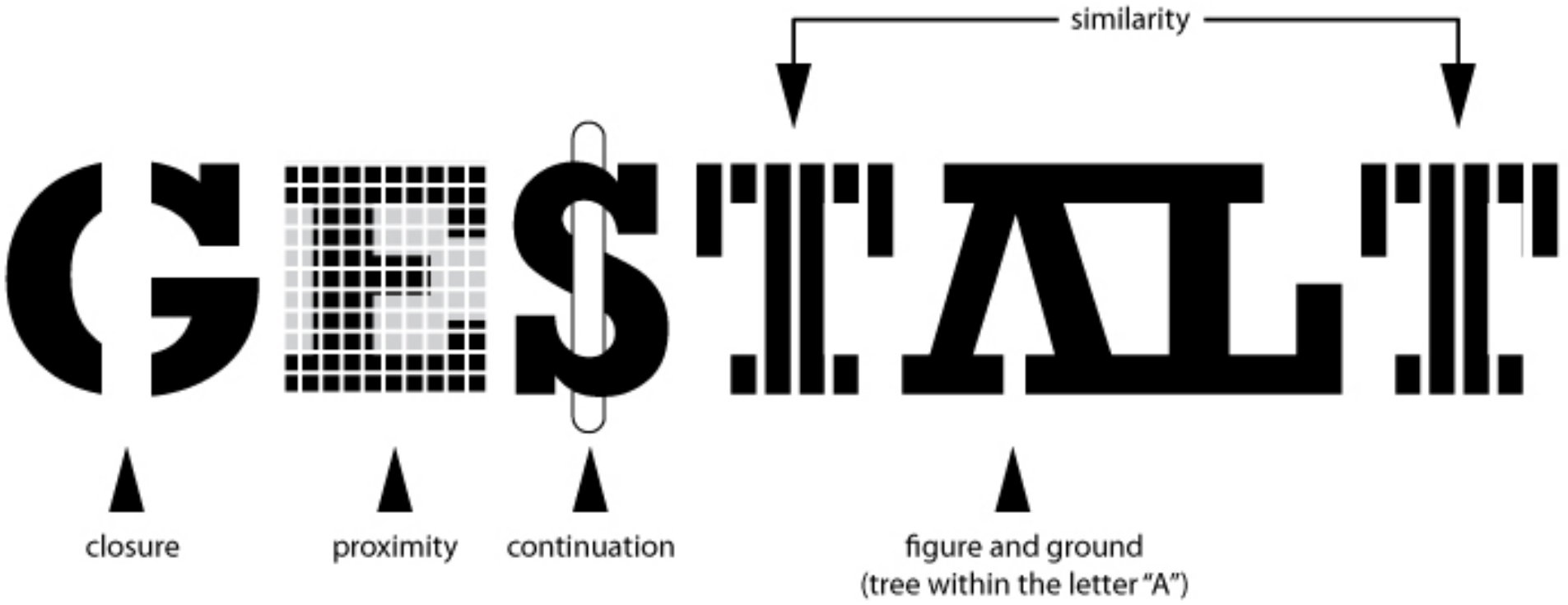
MAPPING is most effective way to stucture Data {visually}

Graphical Mapping Methods

There are 7 dominant graphical methods used primarily in thematic maps for representing all sorts of qualitative and quantitative data:

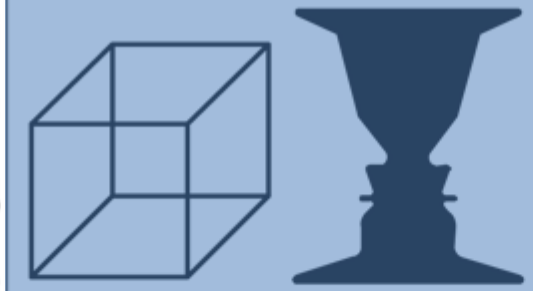
1. Dot Distribution Maps
2. Graduated Symbol Maps
3. Isometric & Isopleth Maps
4. Flow And Network Maps
5. Choropleth Maps
6. Area & Distance Cartograms
7. Tree Maps

Gestalt psychologie classificeert perceptie volgens (top-down) groepering principes

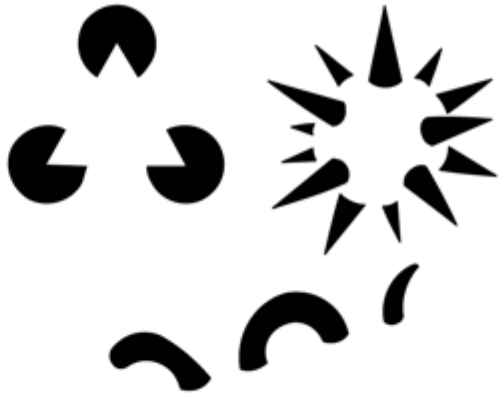


Gestalt classificeert Perceptie

Emergence



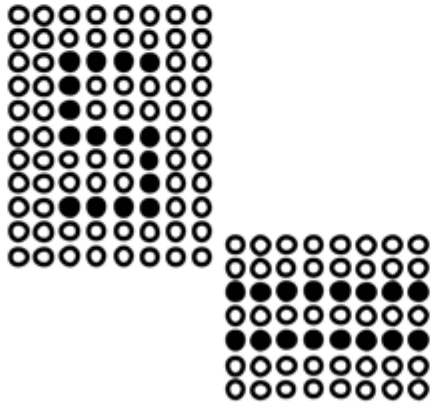
Multistability
Figure/Background selection



Reification
Illusory contours

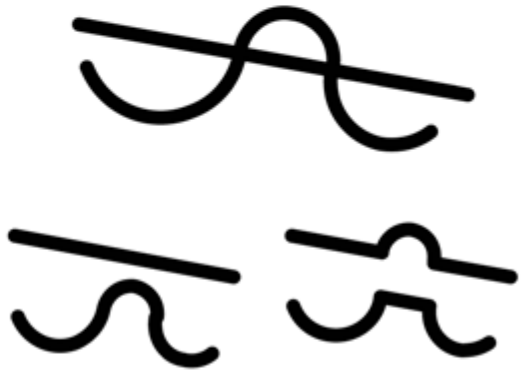
Invarriance

Closure



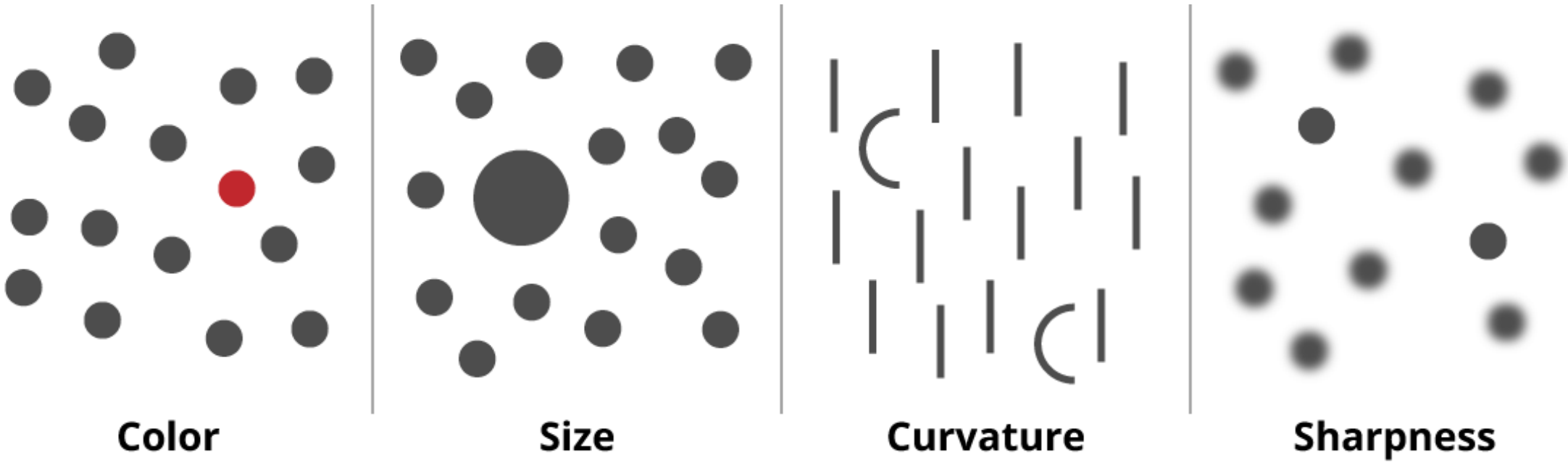
Similarity

Proximity

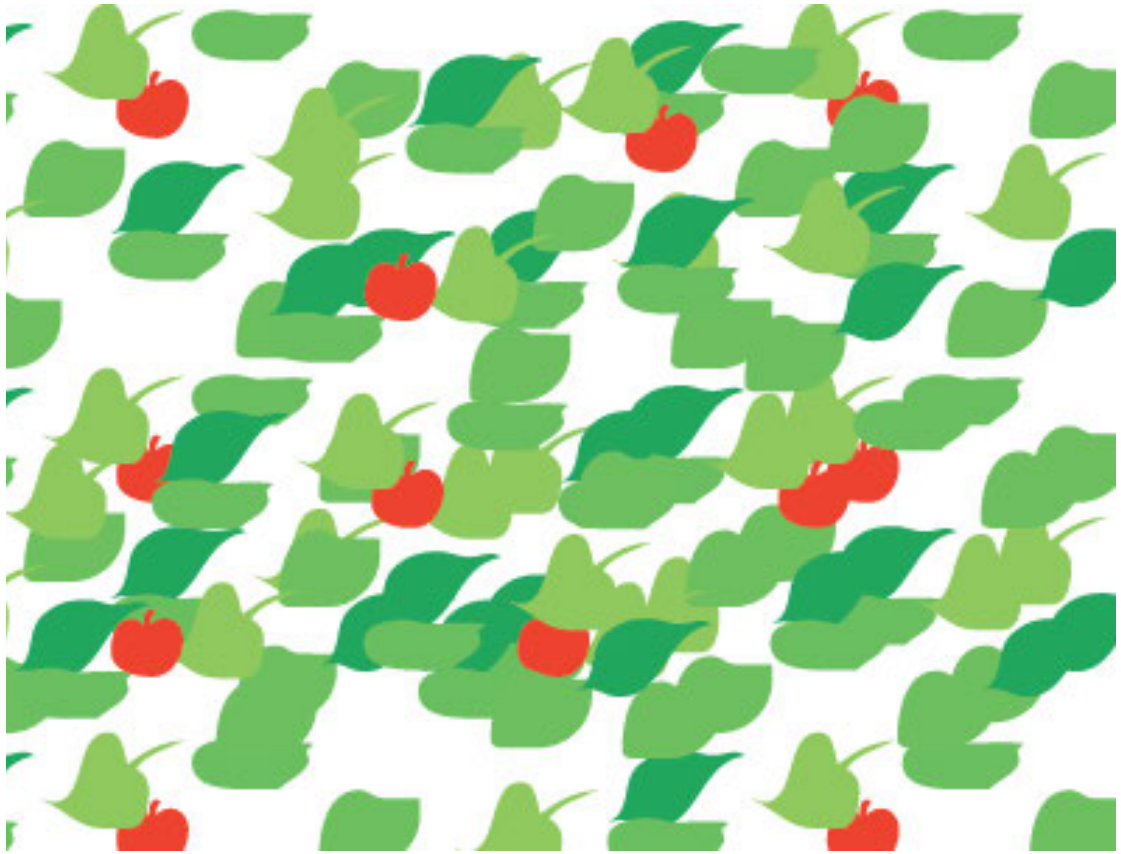


We see this...but not this
Continuity

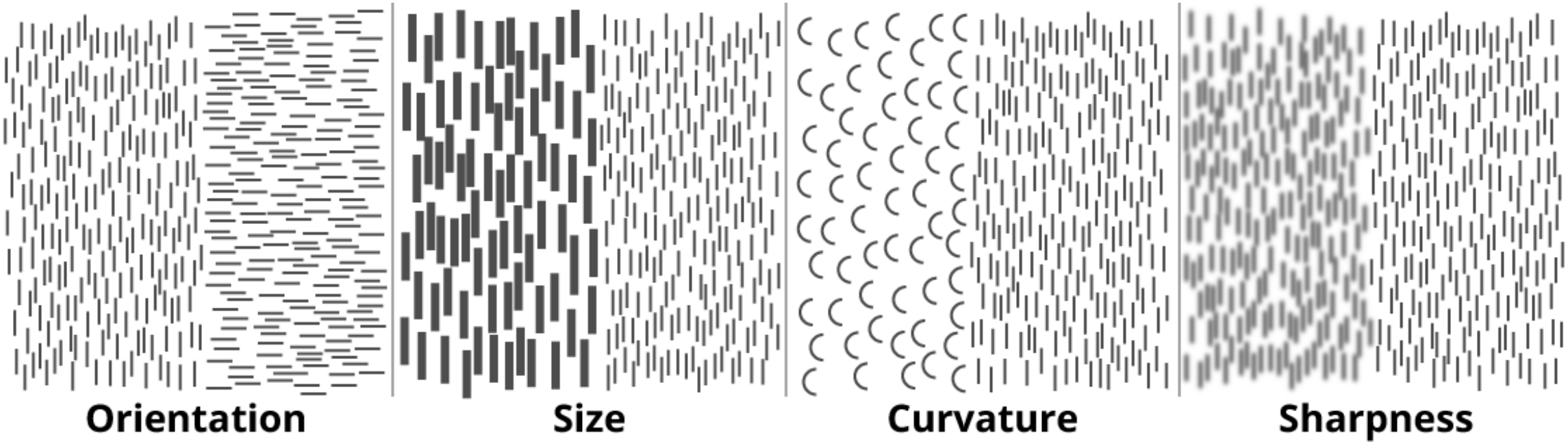
Gestalt is geen PoP-out fenomeen



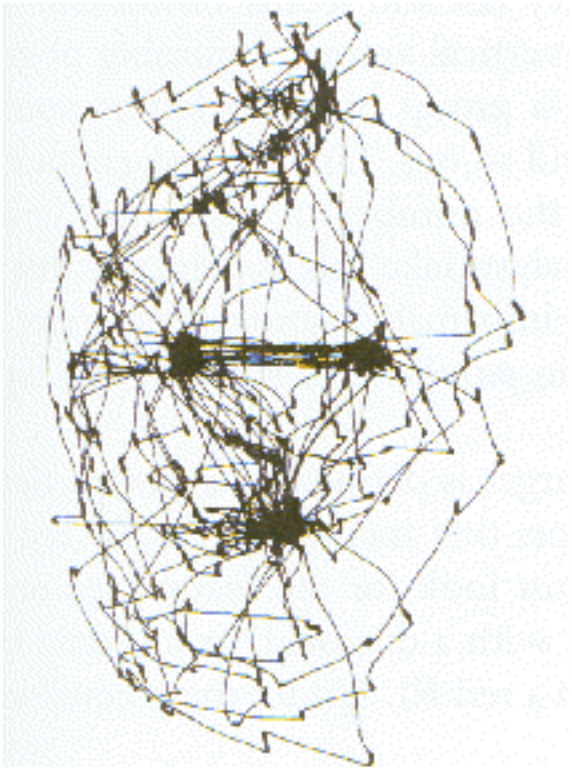
Gestalt is geen chromatisch fenomeen

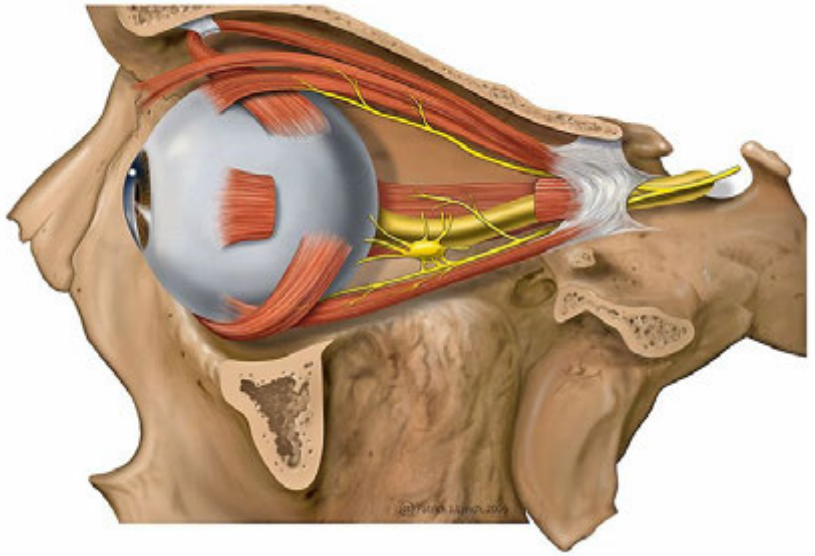


Gestalt is geen Textuur fenomeen

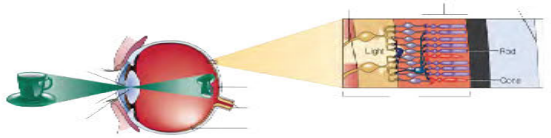


Gestalt is een passief fenomeen





Gestaltprincipes zijn een bijzondere vorm van *passieve* visuele waarneming
ze vereisen dan ook *geen* oogbewegingen

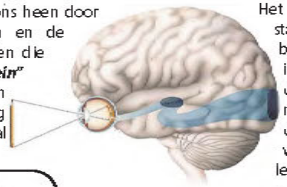


1 Paradox van het Zien

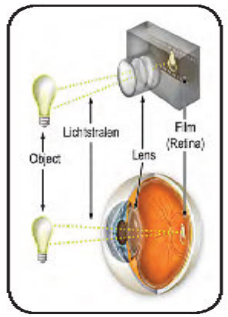
Zien is meer dan fotos maken

Het oog als camera

We zien de wereld om ons heen door middel van onze ogen en de daarmee verbonden delen die tezamen het ***"visuele-brein"*** vormen. Het proces van de visuele gewaarwording delen we op in een aantal stadia.



Het zuiver optische stadium (het afbeeldings-proces) in het oog wordt vaak vergeleken met de werking van een camera, waarin ook een lens een afbeelding vormt op een lichtgevoelige filmlaag. Hoewel deze vergelijking onjuist is, leidt zij wel tot zinvolle vragen.



Het eerste stadium is zuiver ***"optisch"***: het vormen van een scherpe afbeelding door de lens van het oog op het netvlies ***(retina)***, dat de lichtgevoelige cellen bevat. Volgende stadia zijn het omzetten van een afbeelding in zenuwsignalen en het verwerken van deze signalen door de hersenen.

Hoe is het mogelijk dat we met ons oog scherp en "rechtop staand" zien?

Bij het maken van een foto moeten we de camera zoveel mogelijk stil houden. Wanneer niet de camera, maar het gefotografeerde object beweegt, wordt de foto onscherp. Ook wordt het object "op zijn kop" afgebeeld. Ten slotte moet de belichting stijd zeer kort zijn.

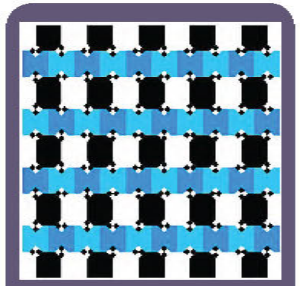
Het oog beweegt in het hoofd, het hoofd beweegt op ons lichaam, dat zich weer verplaatst in de ruimte. Het netvlies is continu belicht. Toch hoef je niet muisstil te zitten om goed te kunnen zien!

Zien

We nemen onze omgeving scherp waar terwijl we lopen, fietsen of autorijden, en ook bewegende objecten kunnen we scherp zien. Terwijl retinale-afbeeldingen juist omgekeerd, instabiel en plat (tweedimensionaal, 2D) zijn.

In de wereld om ons heen vinden we aanwijzingen dat dieren (inclusief primaten zoals wijzelf) niet reageren op de afbeeldingen in hun ogen, maar op een "brein-veranderende-versie" ervan.

Wanneer de ogen van een kikker operatief worden omgedraaid t.o.v. zijn kop, dan kan hij geen vliegen meer vangen. Maar, na een paar dagen is de kikker hersteld. Dus het kikker-brein maakt dat wat de ogen aan beelden binnen halen geschikt om er iets mee te kunnen doen in de "buitenwereld".



Bij het kijken naar het "blokpatroon", zoals hierboven afgebeeld, zullen velen de indruk hebben dat de *horizontale* blauwe balken scheef lopen.

Er is een prijs die we betalen voor deze "Brein-veranderde-versie". Zien is "niet *natuurgetrouw*". Dit fenomeen kennen we als *"Gezichtsberog"*.

In dit cahier zullen we stap voor stap nagaan hoe wetenschappers / kunstenaars *"gezichtsbedrog"* bestudeerd hebben. Beide concluderen dat ons brein zich gedraagt als een ***"verhalen verteller"***.

Gezichtsbedrog reflecteert ons vermogen om te komen tot een (be)grijpbare realiteit.

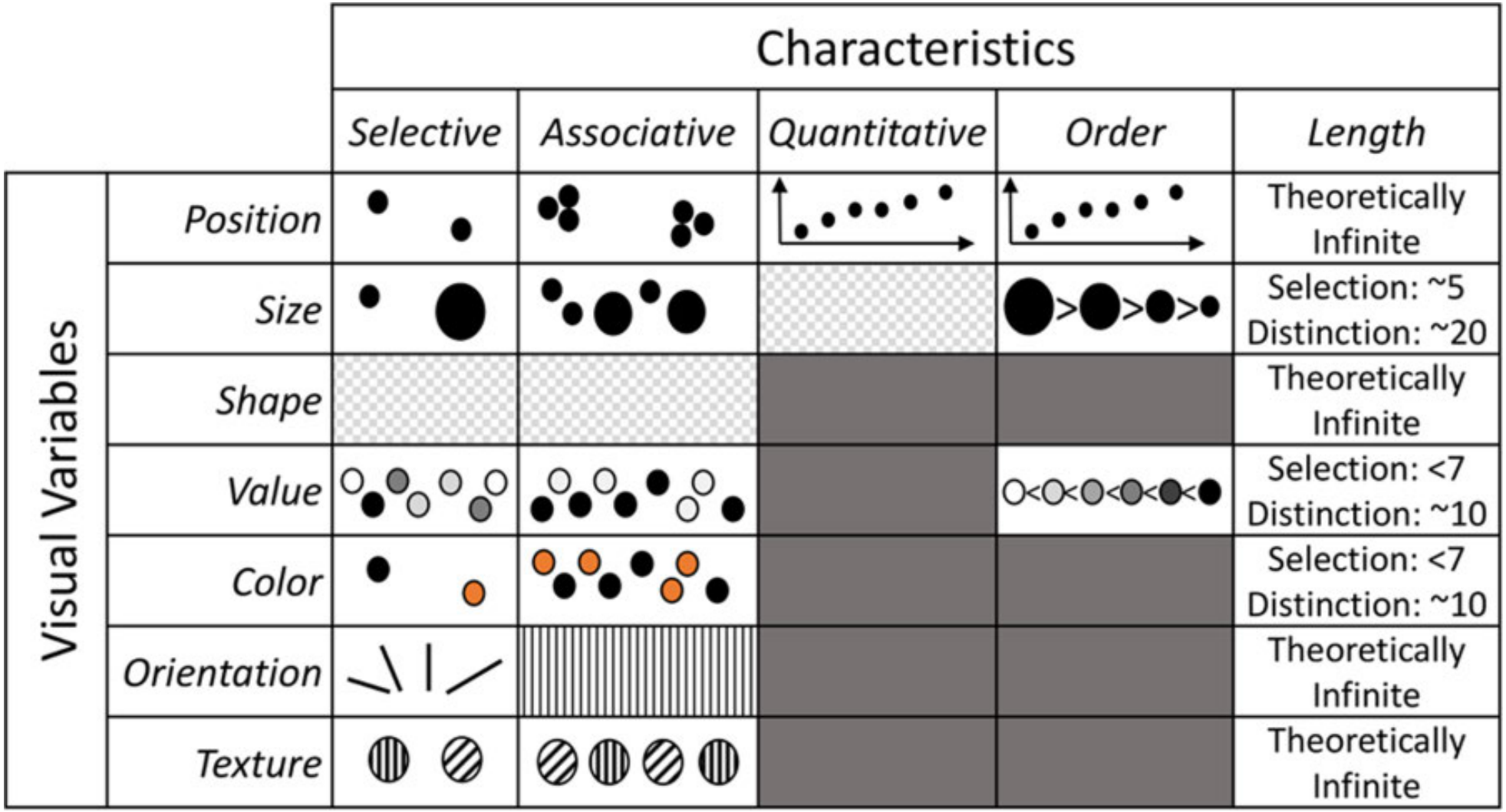
Gezichtsbedrog wordt vaak omschreven als "*onverwachte valkuilen van het zien*."

Het zijn echter "*ogenschijnlijke*" weeffouten die alleen kunnen bestaan in de visuele ruimte van onze hersenen als gevolg meerduidigheid of incompleetheid van de zintuigelijke informatie.

Ons visuele brein kiest voor de meest voor hand liggende interpretatie ---of vult zelf aan--- door gebruik te maken van ingebouwde "kennis" in ons brein over de tastbare wereld om ons heen.

http://www.mbfys.ru.nl/~robvdw/PRACTICUM_ILLUSIONS/DOCs/GezichtsBedrogNovember2009.pdf

DATA STRUCTURING: Bertin's 7 Visual Variables (1967)



SEE ALSO: http://www.cs.wright.edu/~jgalli/hfe306/Data_Visualization_Quenin.ppt

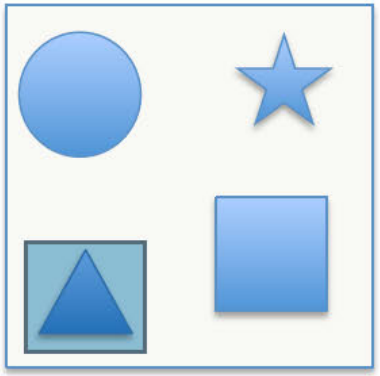
DATA STRUCTURING: Visual Variables

Visual Variable	Author	Example
Size	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Shape	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Lightness/ value	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Color *(hue+saturation)*	Bertin (1967/83).	
Orientation	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Texture	Bertin (1967/83), Morrison (1974), MacEachre (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Tyner (2010).	
Location	Bertin (1967/83), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Hue	Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	

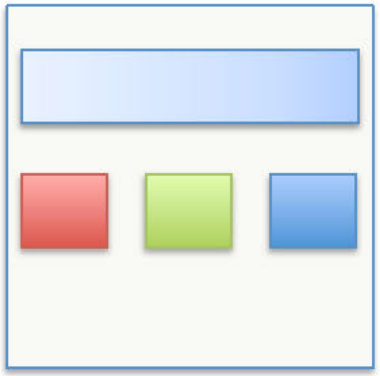
Visual Variable	Author	Example
Saturation/ intensity	Morrison (1974), MacEachren (1995), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Arrangement	Morrison (1974), MacEachren (1995), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Focus/ crispness	MacEachren (1995).	
Resolution	MacEachren (1995).	
Transparency	MacEachren (1995).	
Spacing	Slocum et al (2010).	
Perspective Height	Slocum et al (2010).	

SEE ALSO: http://www.iag-aig.org/attach/30dee1f85f7bd479367f1f933d48b701/V61N1_2FT.pdf

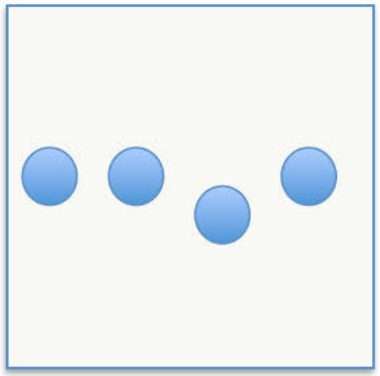
DATA STRUCTURING: Preattentive Processing [perceptual level]



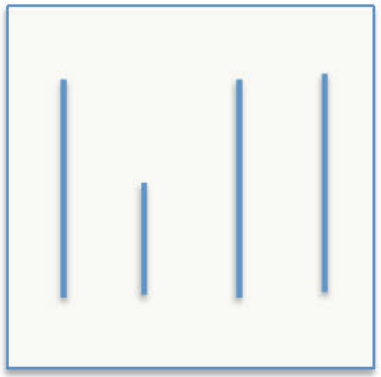
Shape



Color

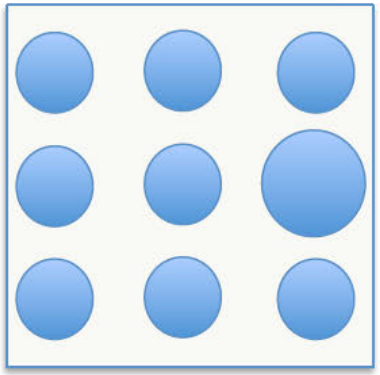


Position

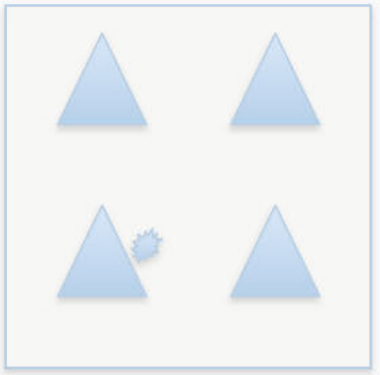


Line Length

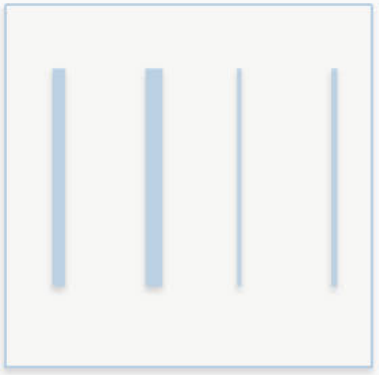
Orientation



Size

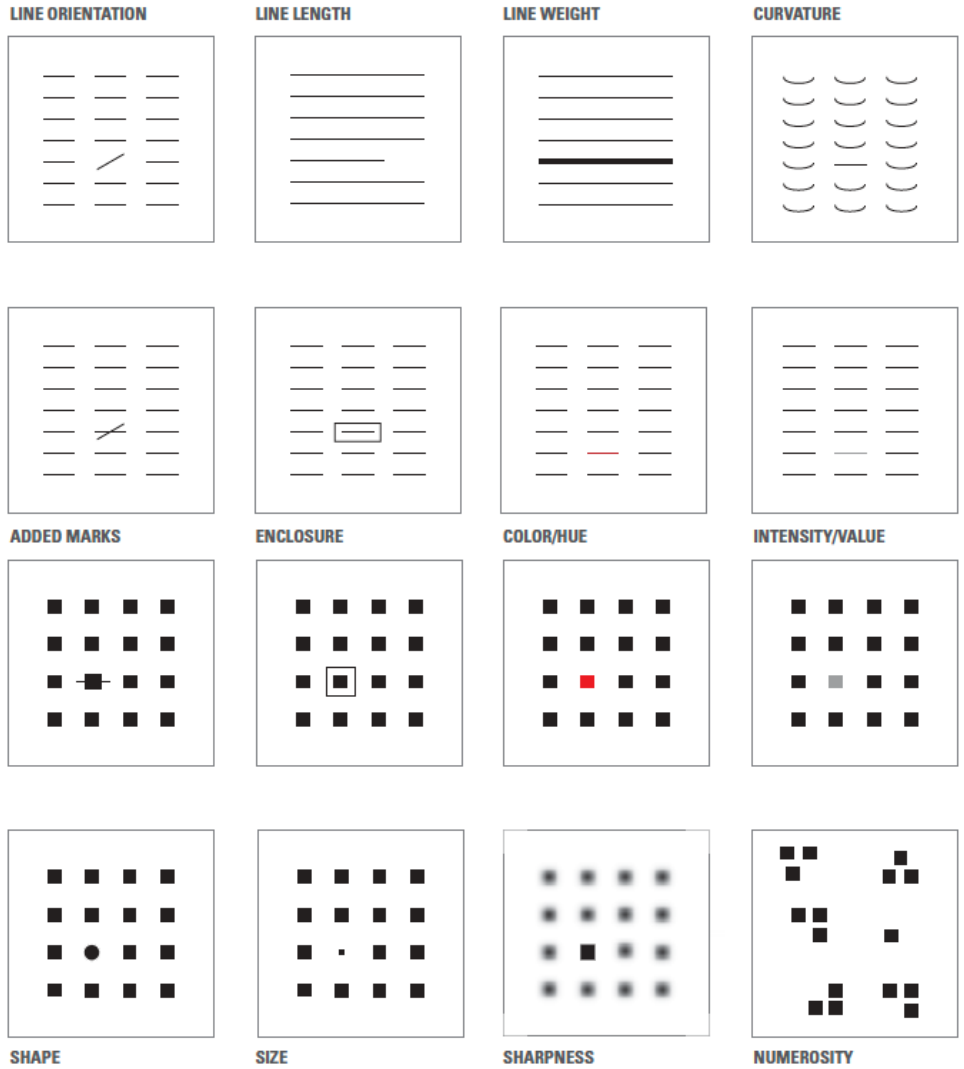


Added Marks



Line Width

DATA STRUCTURING: Preattentive Processing [data type]



Shape

Color

Position

Line Length

Orientation

Size

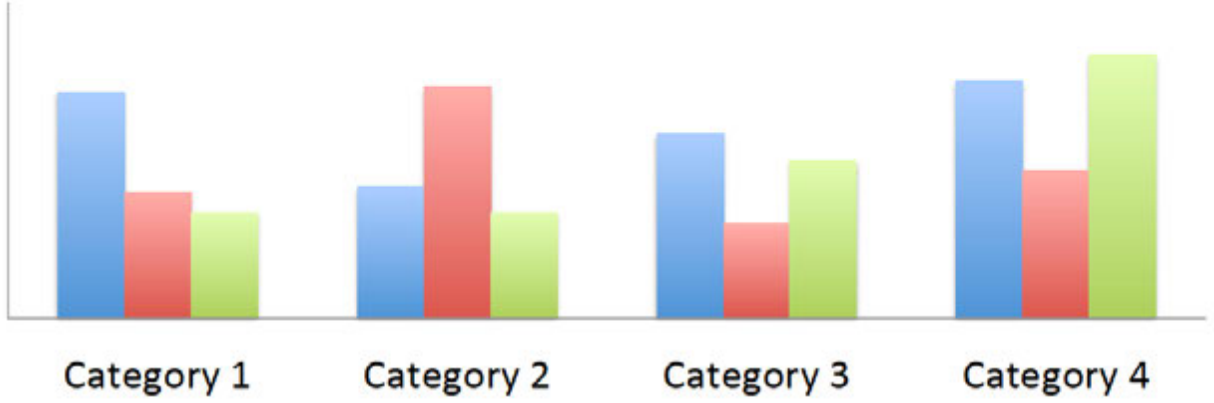
Added Marks

Line Width

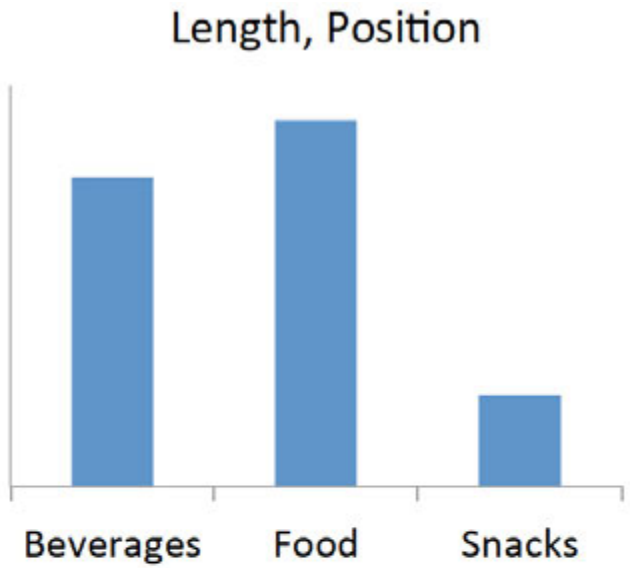
DATA STRUCTURING: Relationships & Patterns [applied]

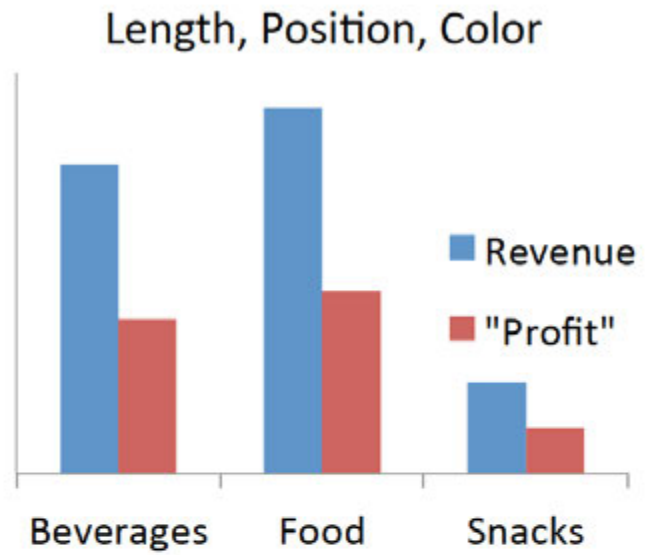
- Tables:
  - Accuracy
  - Lookup
- Charts
  - Story
  - Summarize lots of data

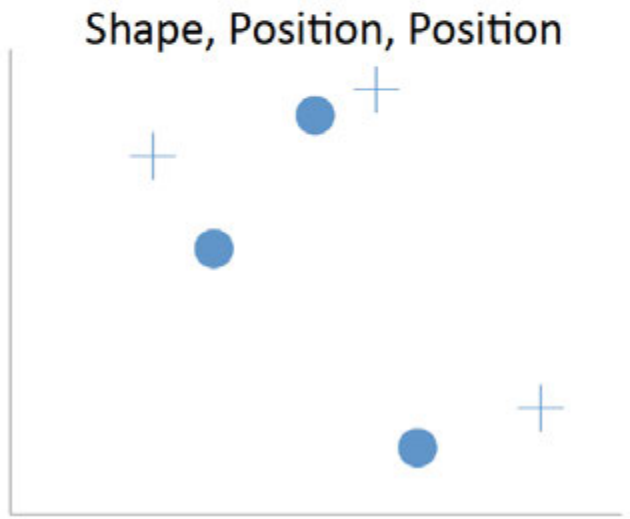
Lender	Adjustable	Fixed
Bank 1	7%	5.25%
Bank 2	7.25%	5.5%

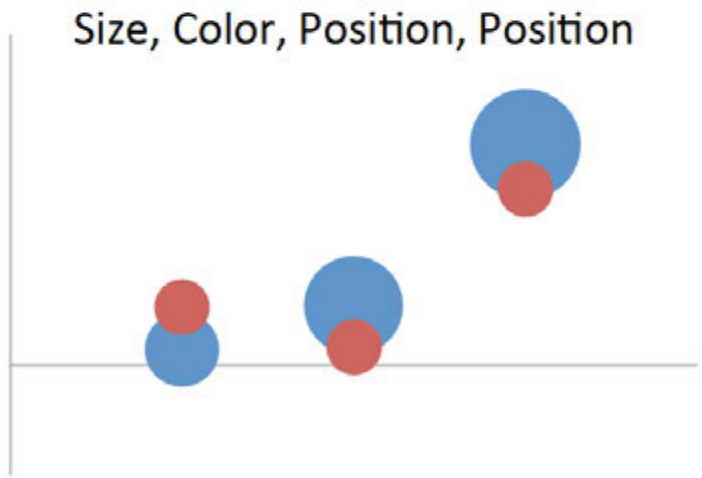


DATA STRUCTURING: Relationships & Patterns [applied: CHARTS]

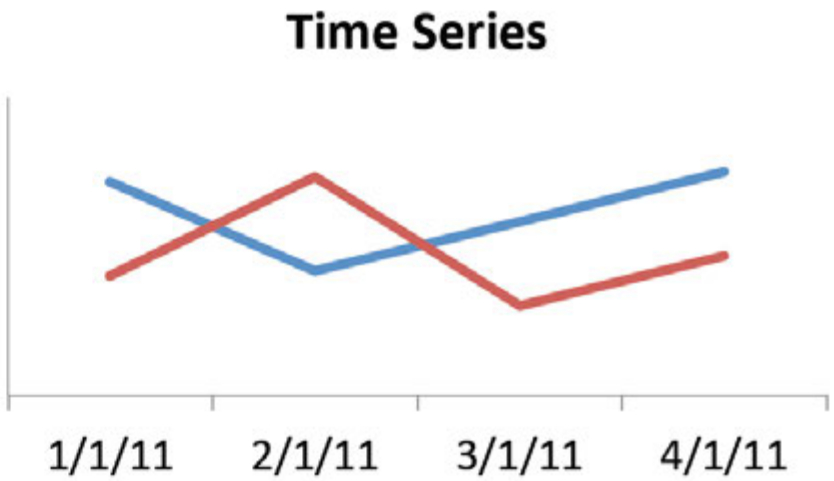


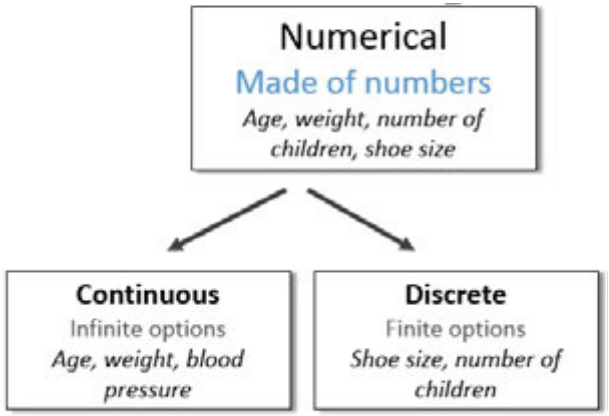


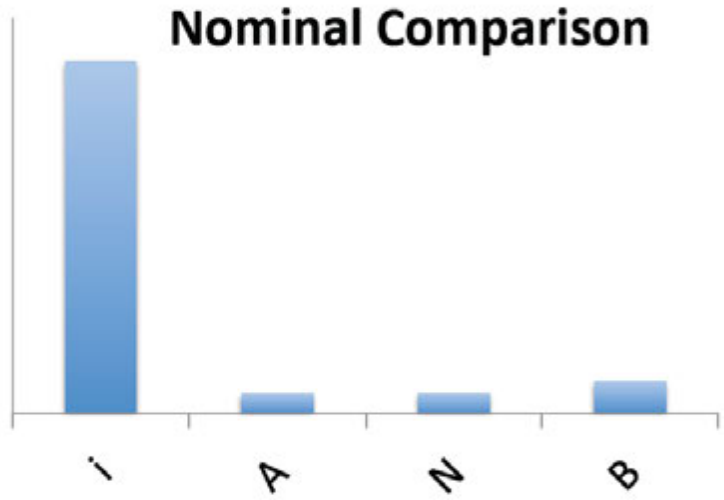


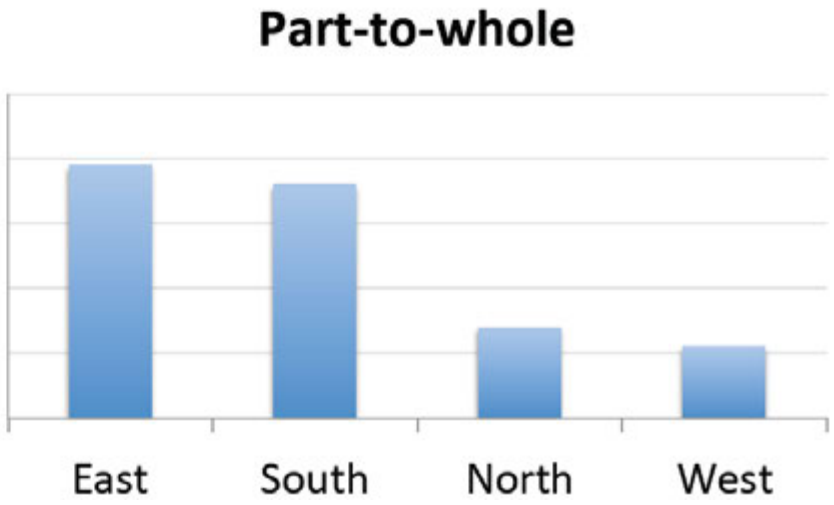


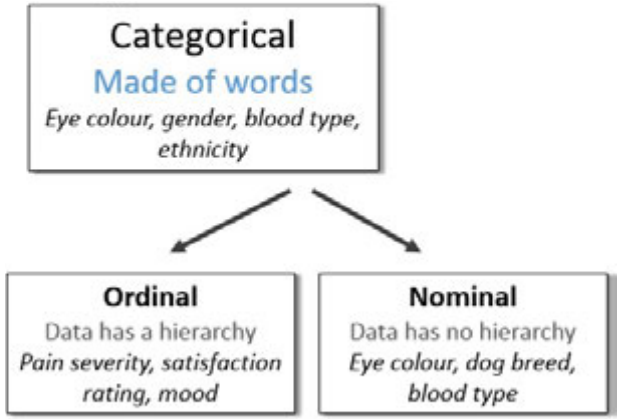
DATA STRUCTURING: Relationships & Patterns [applied: CHARTS]

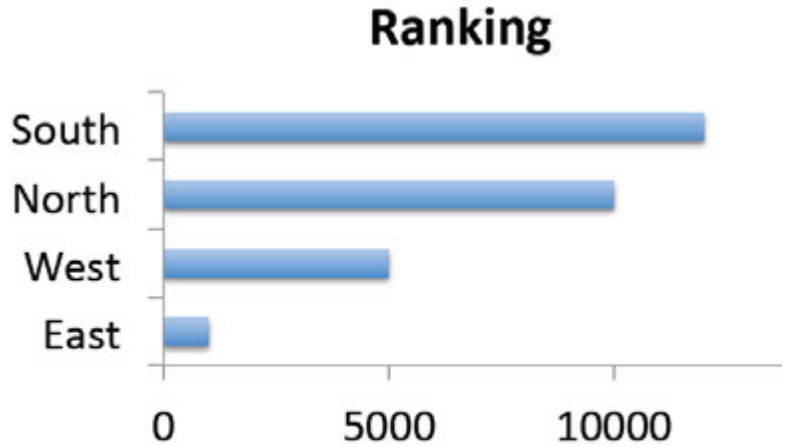








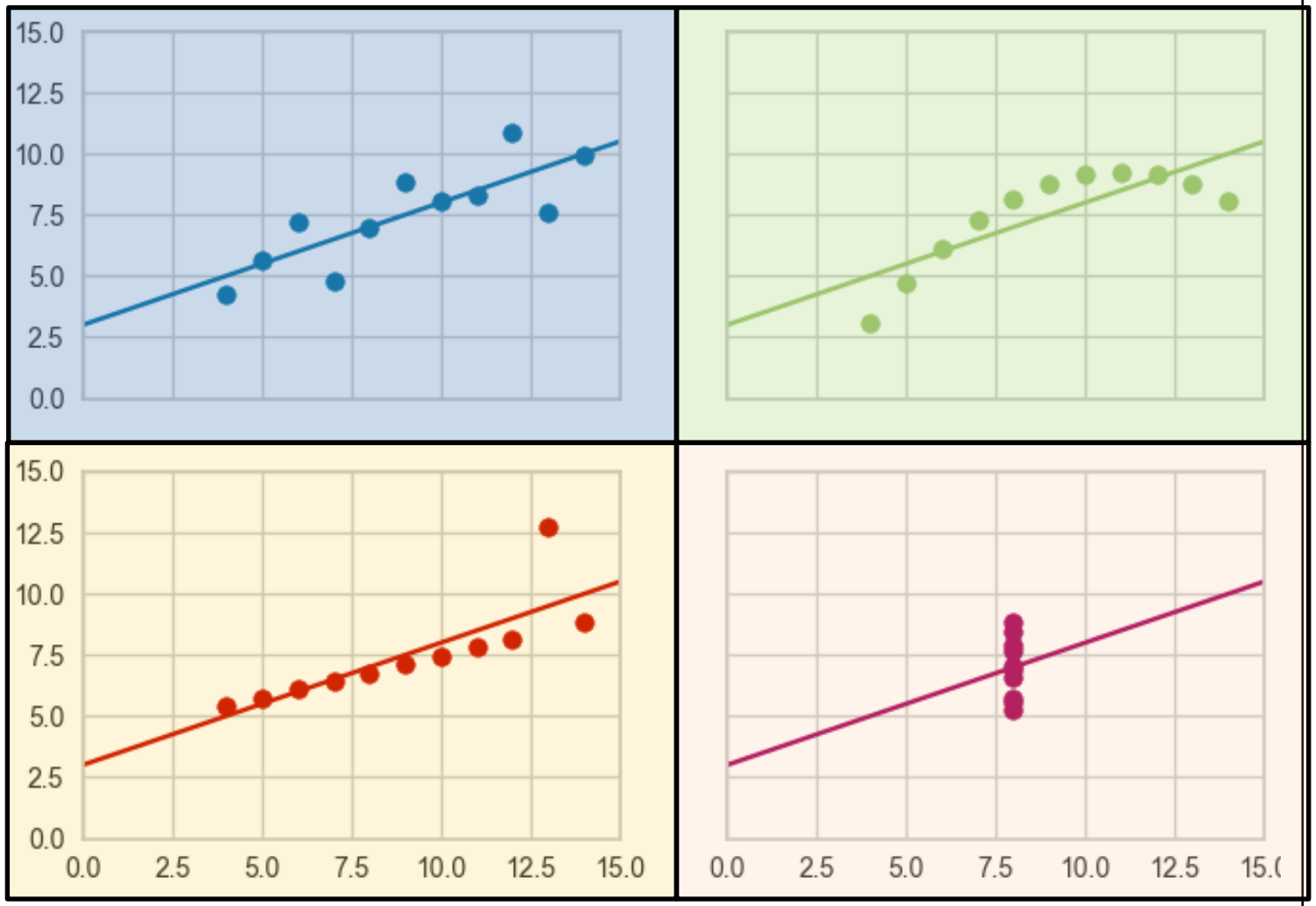




DATA STRUCTURING: Relationships & Patterns [Anscombe Quartet]

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



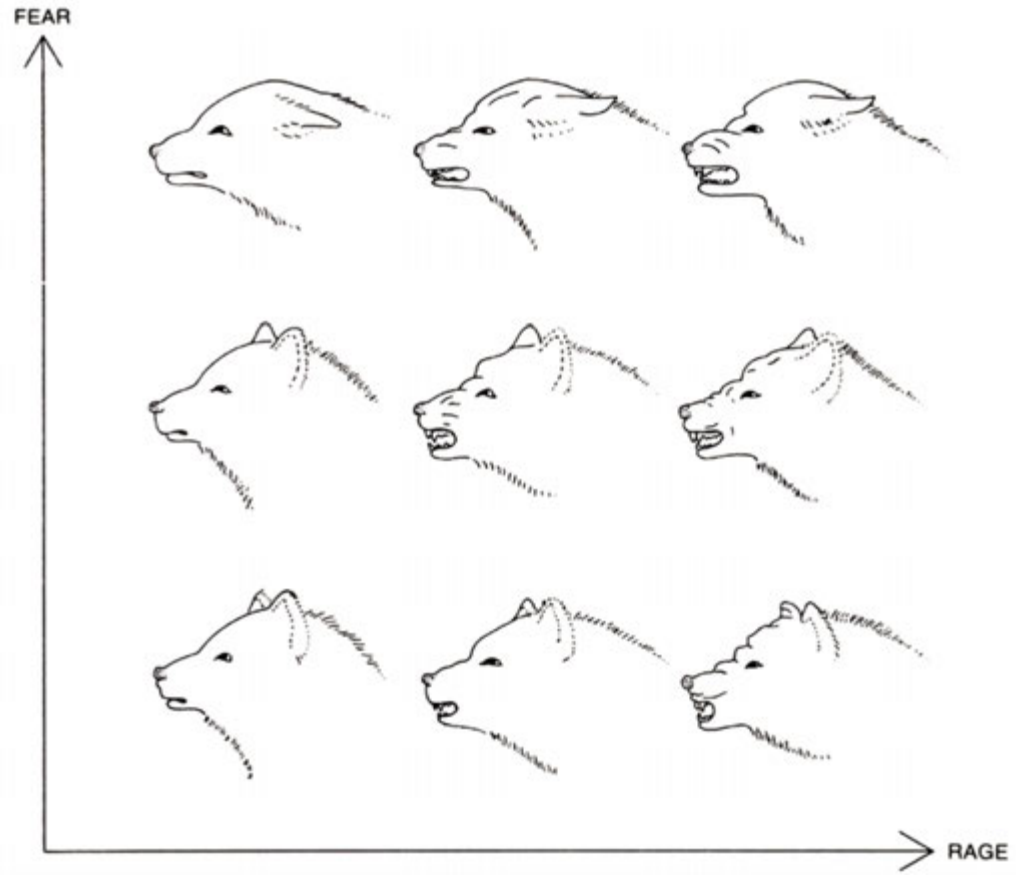
DATA STRUCTURING: Relationships & Patterns [applied]

when the image is the data

the visual medium is ideal for depicting multivariate data

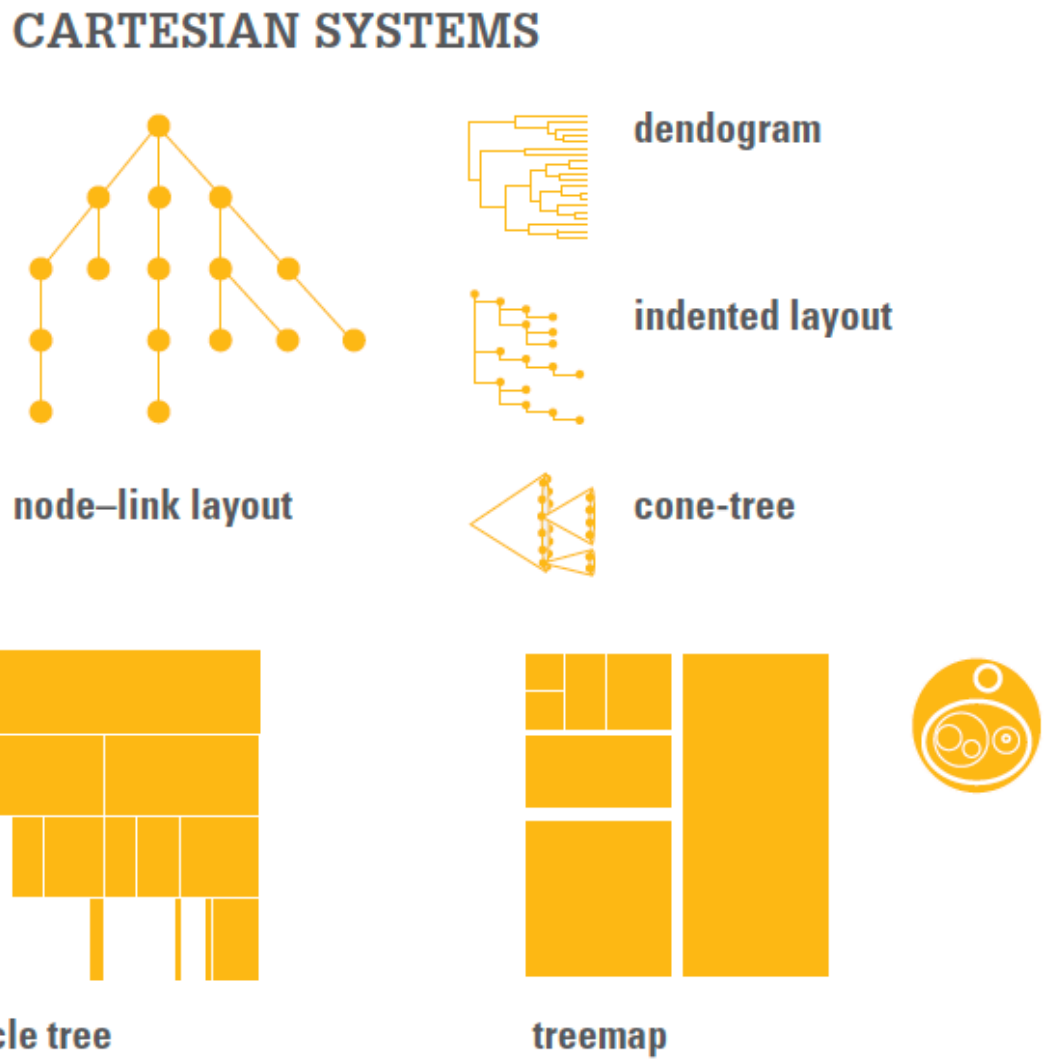
arguably univariate and bivariate data should be tabularized, within reason

this example shows a plot for a case where data cannot be easily parametrized



http://www.edwardtufte.com/tufte/posters

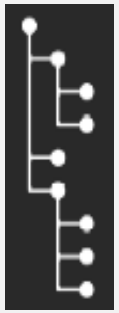
DATA STRUCTURING: Relationships & Patterns [HIERARCHIES]



Graph Drawing

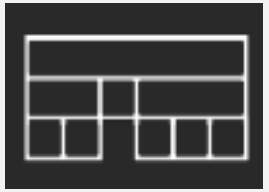
4 Major tree visualizations

Indented lists

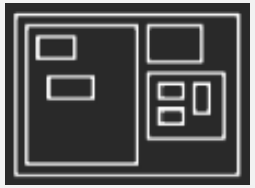


Node-link trees

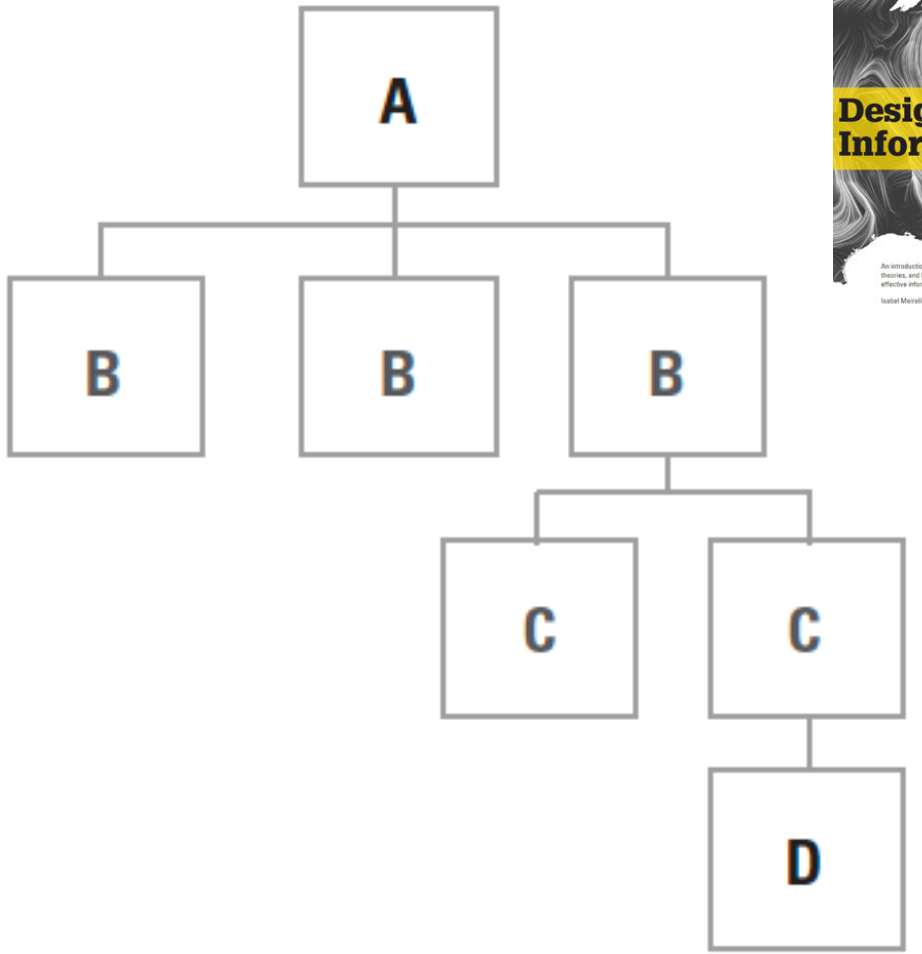
Layered diagrams



Treemaps



DATA STRUCTURING: Graph (graaf) [treemap versus intented list]



DATA STRUCTURING: TREE MAP [example]

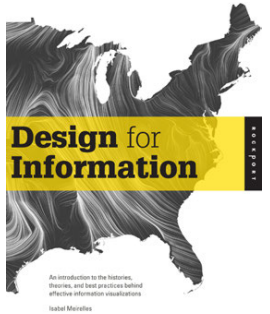
Canada	United Kingdom	Germany	India
United States	Netherlands	Austria	Australia
Mexico	France	Italy	New Zealand
Brazil	Spain		
Argentina			

https://newsmap-js.herokuapp.com

https://github.com/IJMacD/newsmap-js

AUTHOR Marcos Weskamp (concept, design, frontend and backend coding) and Dan Albritton (backend coding)
COUNTRY United States
DATE 2004
MEDIUM Online, real-time interactive application
URL http://newsmap.jp
DOMAIN News coverage aggregated by Google News API
TASK To provide an overview of online news stories and reveal underlying patterns in news reporting around the world
STRUCTURE The visualization uses the treemap technique. The algorithm renders the inner-division shapes closer to rectangles, facilitating readability of text.

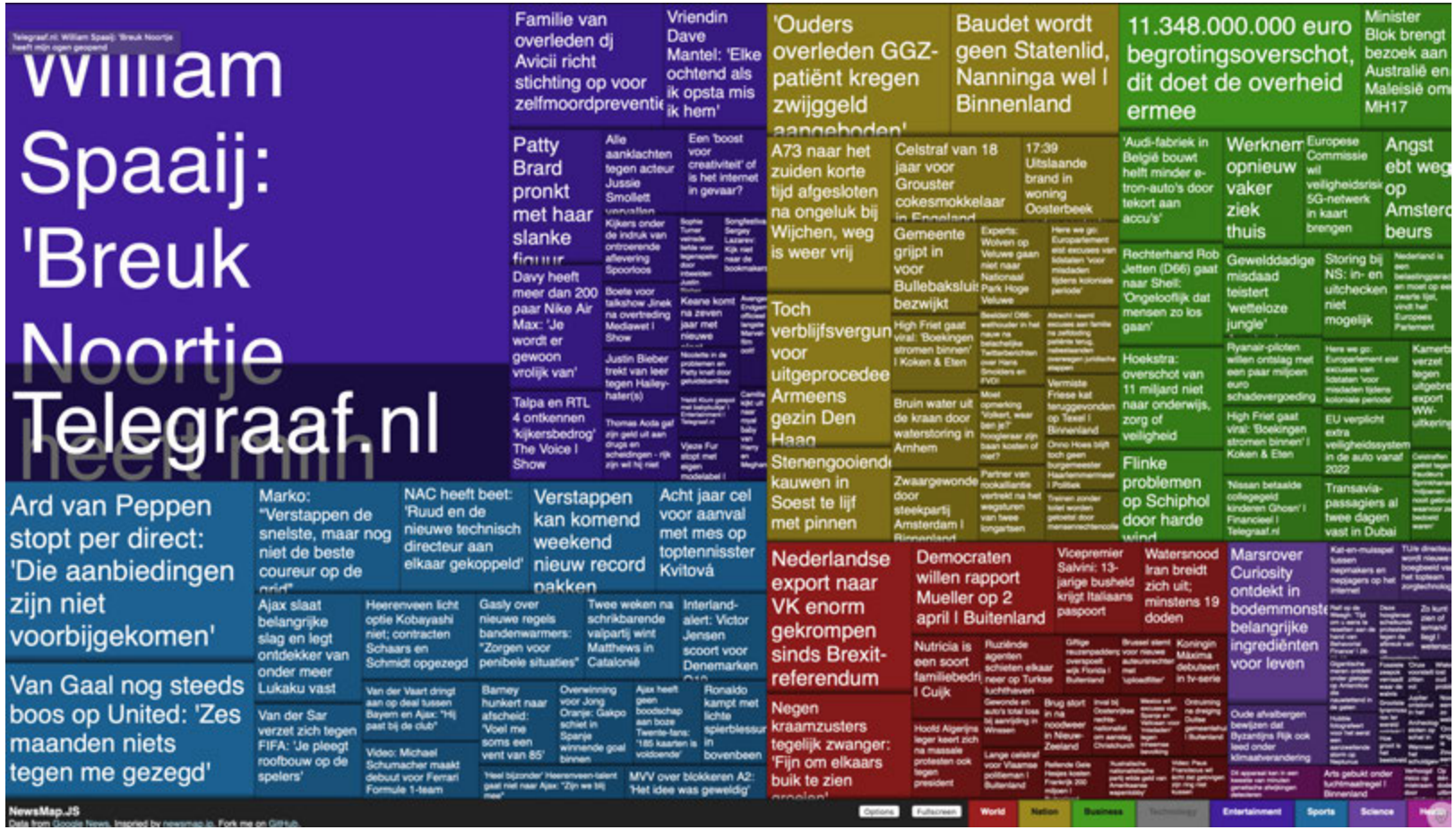
DATA TYPE AND VISUAL ENCODING
Categorical: News segments
Encoding: Color hues and spatial grouping
Categorical: Countries
Encoding: Label and enabled by selection
Temporal: News age: how old the news is
Encoding: Color value
Quantitative: Number of related stories
Encoding: Area size
Nominal: Title of news story
Encoding: Type size relative to the quantitative data

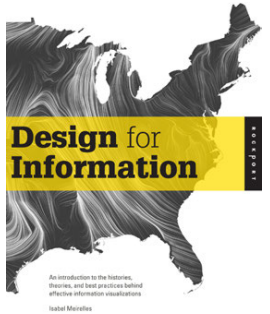


DATA STRUCTURING: TREE MAP [example]

https://github.com/IJMacD/newsmap-js

https://newsmap.ijmacd.com

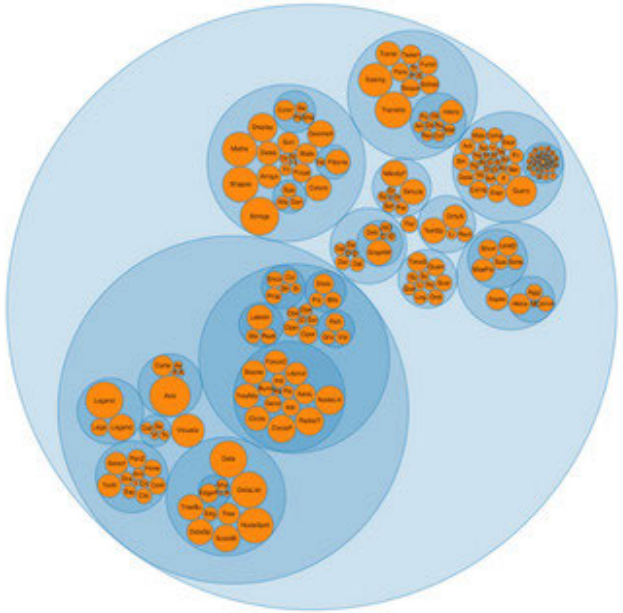




DATA STRUCTURING: TREE MAP [example]

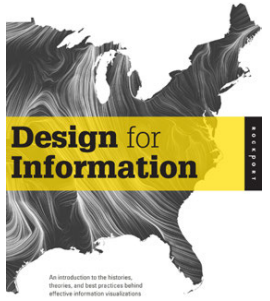
A

B

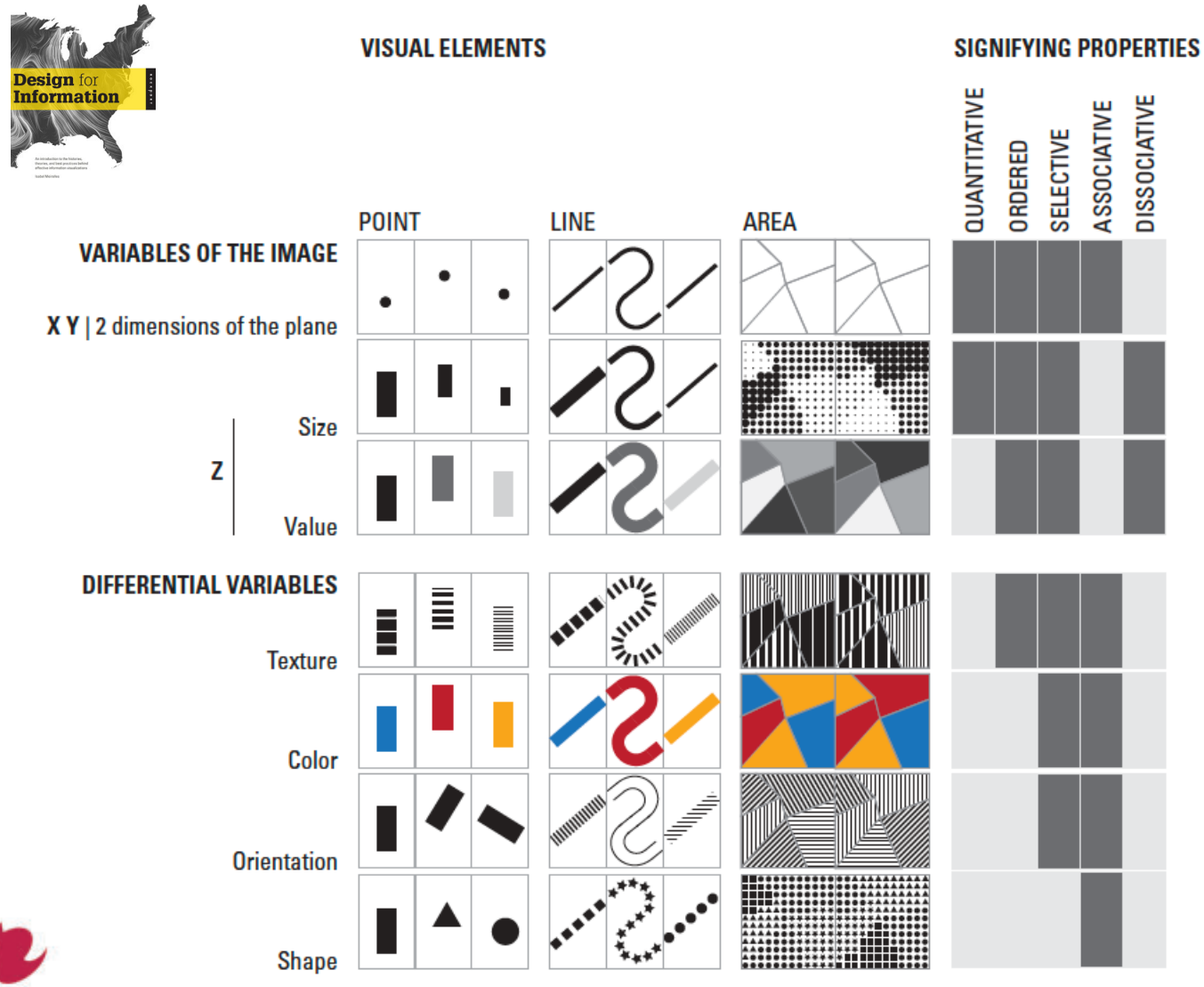


C

D



DATA STRUCTURING: Spatial relationships [MAPPING]



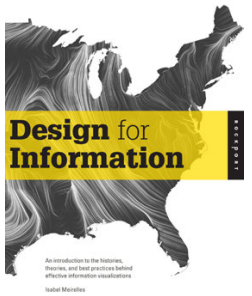
Visual Encoding

The data attribute of dimension is one of the most important characteristics when considering how to conceptualize visual marks in cartography.

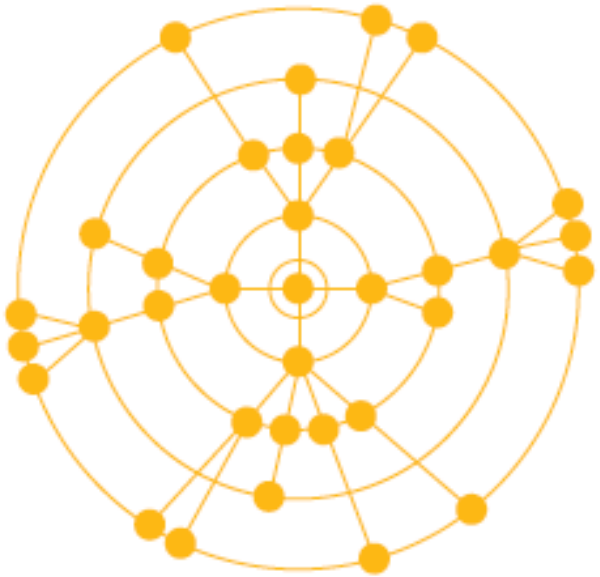
The basic graphic elements of visual representation are:

Point has no dimension
provides a sense of place.
Line has one dimension [X]
provides a sense of length and direction.
Plane has two dimensions [X,Y]
provides a sense of shape and scale.
Volume has 3 dimensons [X,Y,Z]
provides a sense of space, shape and scale.

DATA STRUCTURING: Relationships & Patterns [HIERARCHIES]



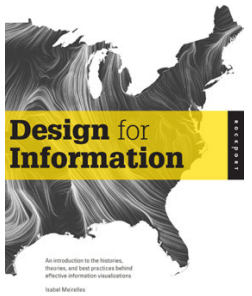
POLAR SYSTEMS



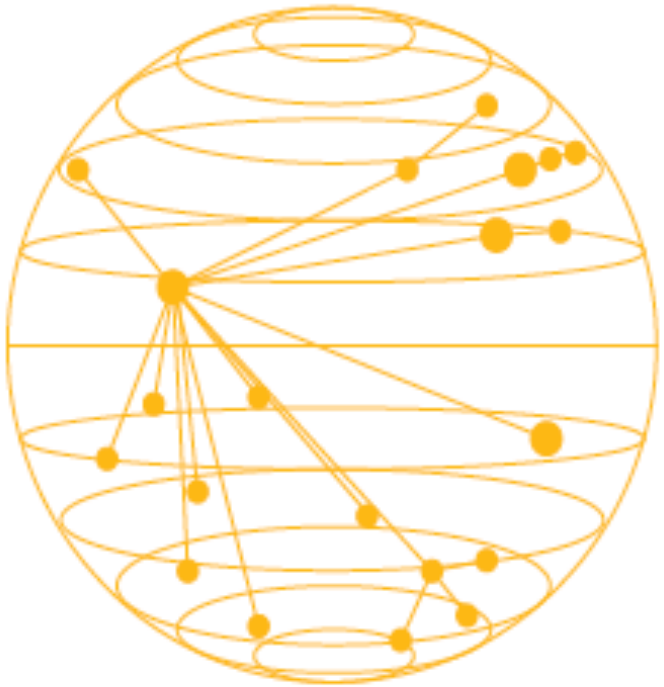
node–link radial layout

radial icicle or sunburst

DATA STRUCTURING: Relationships & Patterns [HIERARCHIES]



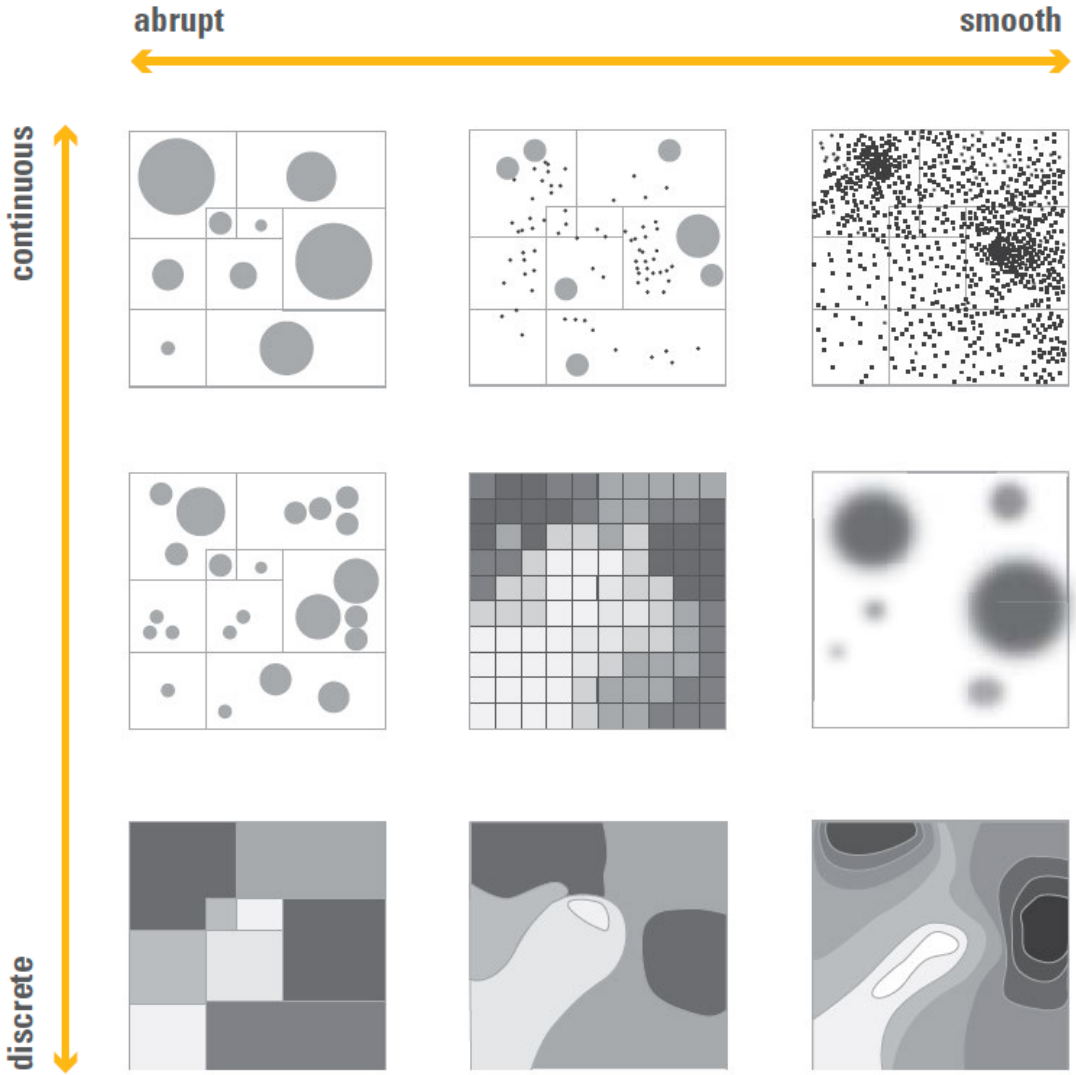
OTHER GEOMETRIES



3D hyperbolic tree

vonoroi treemap

DATA STRUCTURING: Spatial relationships [MAPPING]



Visual Encoding

Visual encoding is the process of matching the phenomena to be visualized, which is provided by the dataset (data scale and attributes), to the most suitable type of representation (graphical elements and visual properties). Visual encoding in cartography is often called symbolization.

The data attributes highly significant for spatial relationships are discrete versus continuous data & abrupt (low resolution) versus smooth data(high resolution).

DATA STRUCTURING: Mapping

Graphical Methods

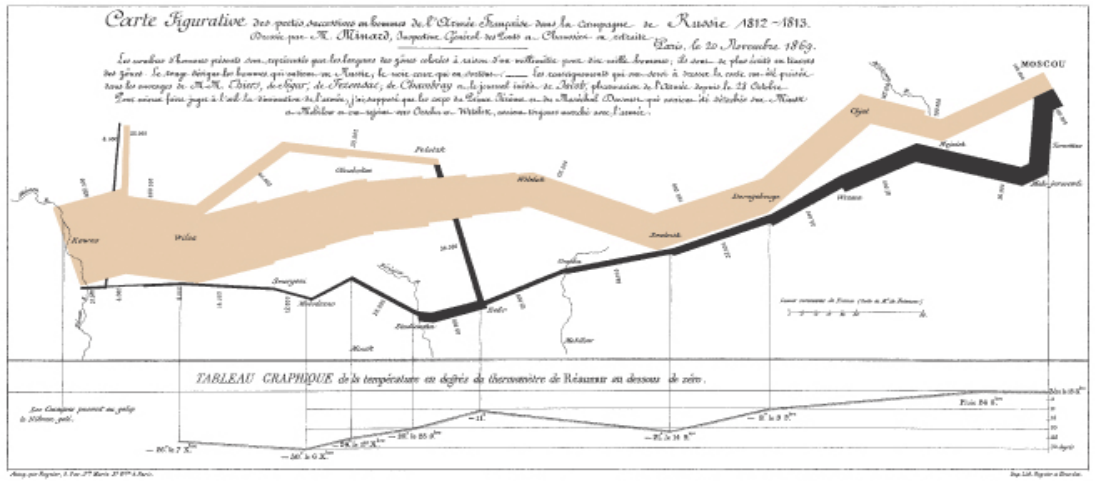
There are 6 graphical methods used primarily in thematic maps for representing all sorts of qualitative and quantitative data:

1. Dot Distribution Maps
2. Graduated Symbol Maps
4. Isometric & Isopleth Maps
5. Flow And Network Maps
3. Choropleth Maps
6. Area & Distance Cartograms

Maximization of useful information on a limited display

Probably the best statistical graphic ever drawn, this map by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the Polish-Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. Exquisitely printed in two colors on fine archival paper, 22" by 15".

Minard's sources. Minard's biography.



Napoleon's March to Moscow The War of 1812 Charles Joseph Minard

DATA STRUCTURING: **Mapping** Dot distribution maps

HOGESCHOOL ROTTERDAM

The New York Times

Mapping the 2010 U.S. Census

Browse population growth and decline, changes in racial and ethnic concentrations and patterns of housing development.

http://www.nytimes.com/projects/census/2010/map.html

View More Maps

Distribution of racial and ethnic groups in 2010

One dot = 5,000 people

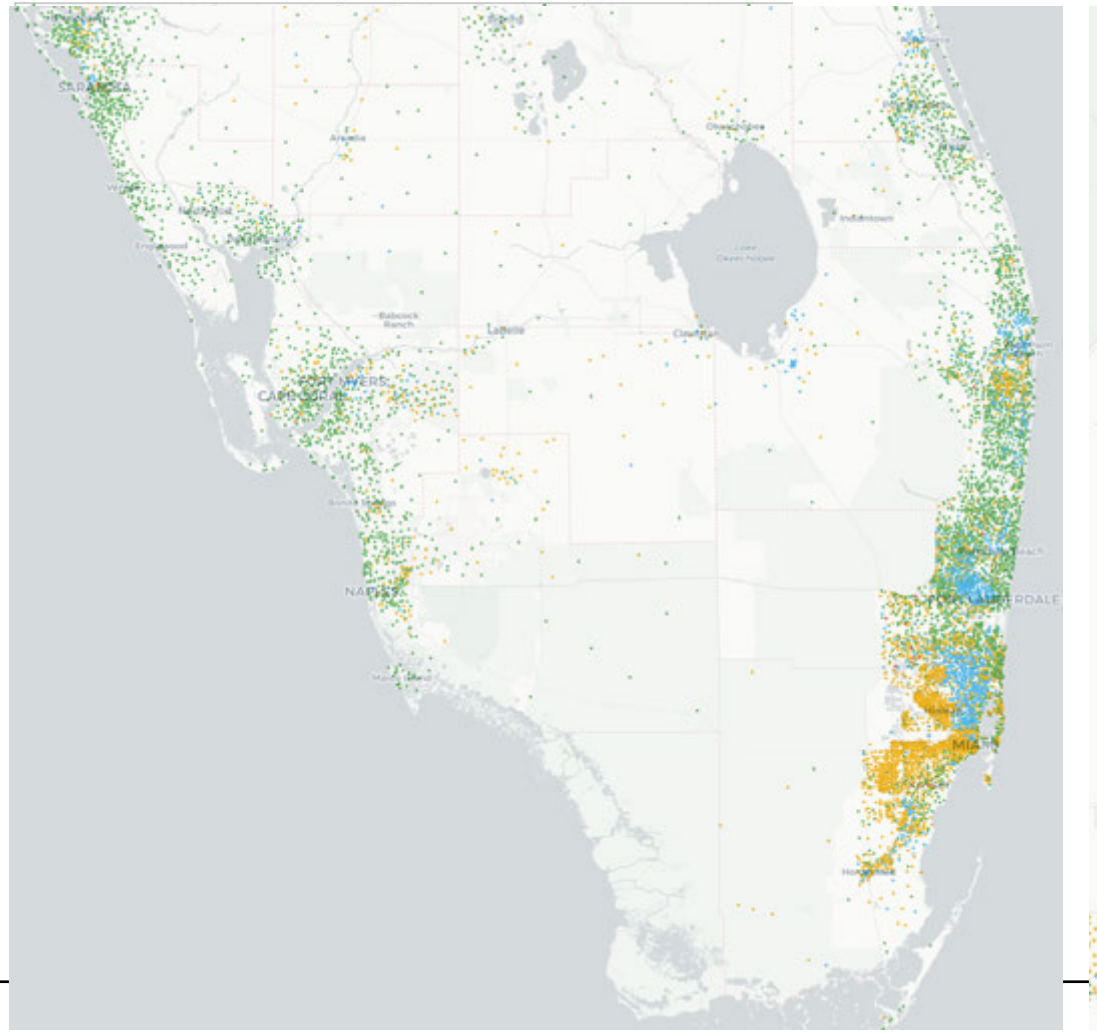
- White
- Black
- Hispanic
- Asian
- Native American
- Other

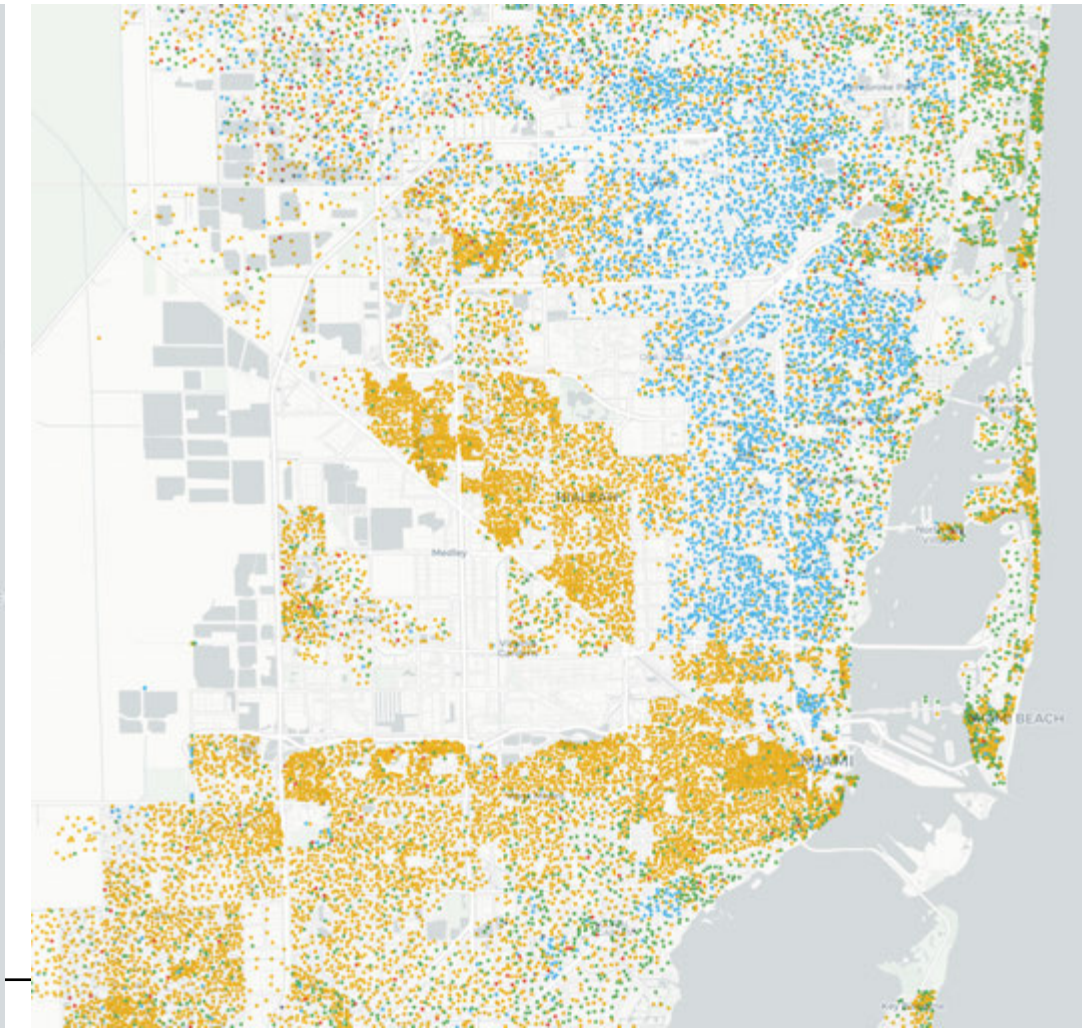
Zoom to a State

Missouri

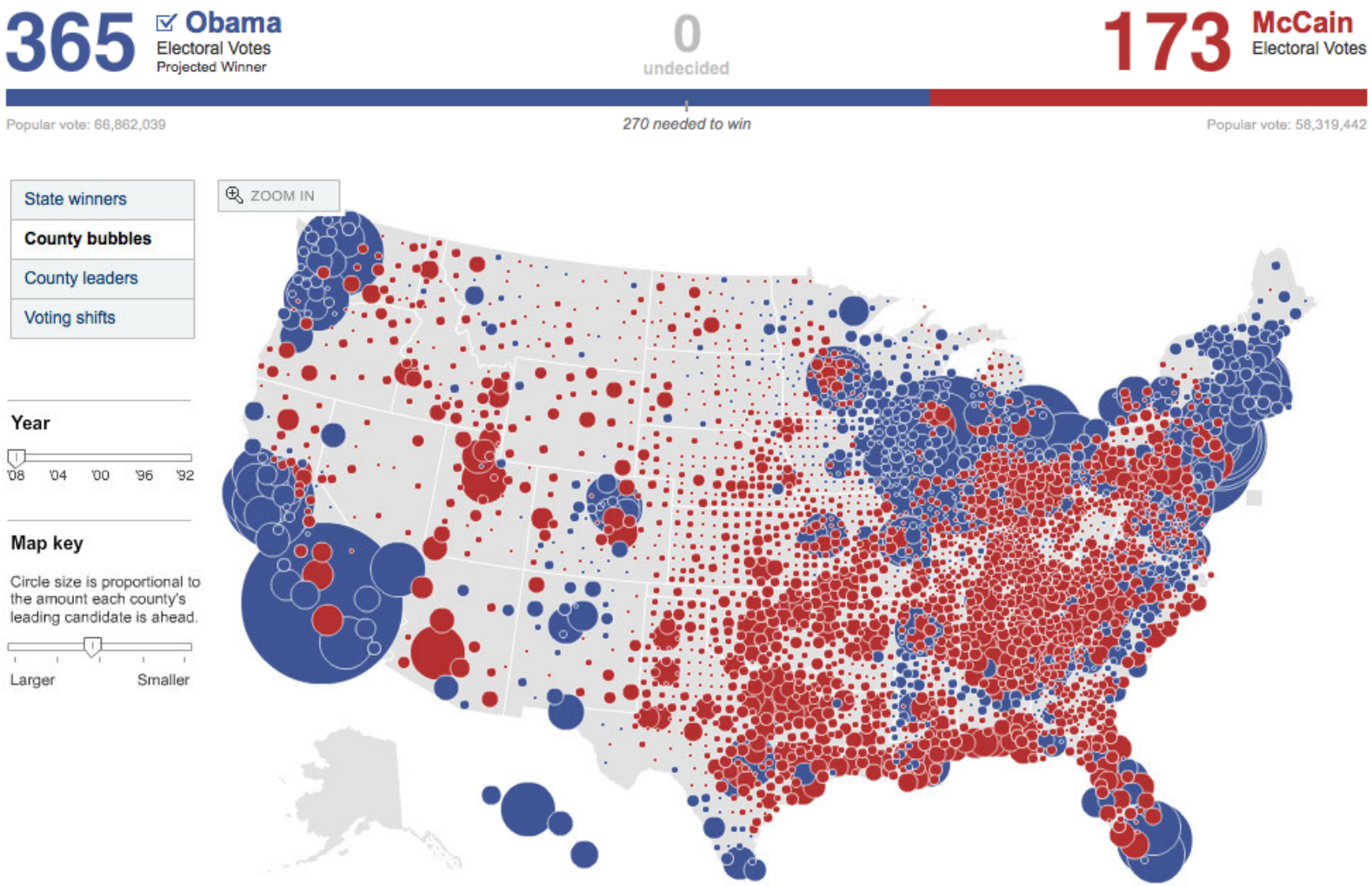
2010 POPULATION	CHANGE FROM 2000
5,988,927	**+7.0%**

RACE/ ETHNICITY	SHARE OF POP.	CHANGE FROM 2000
Whites:	81%	+4%
Blacks:	11%	+10%
Hispanics:	4%	+79%
Asians:	2%	+59%
Native Amer.:	0%	+3%
Multiracial:	2%	+48%
Other groups:	0%	+35%



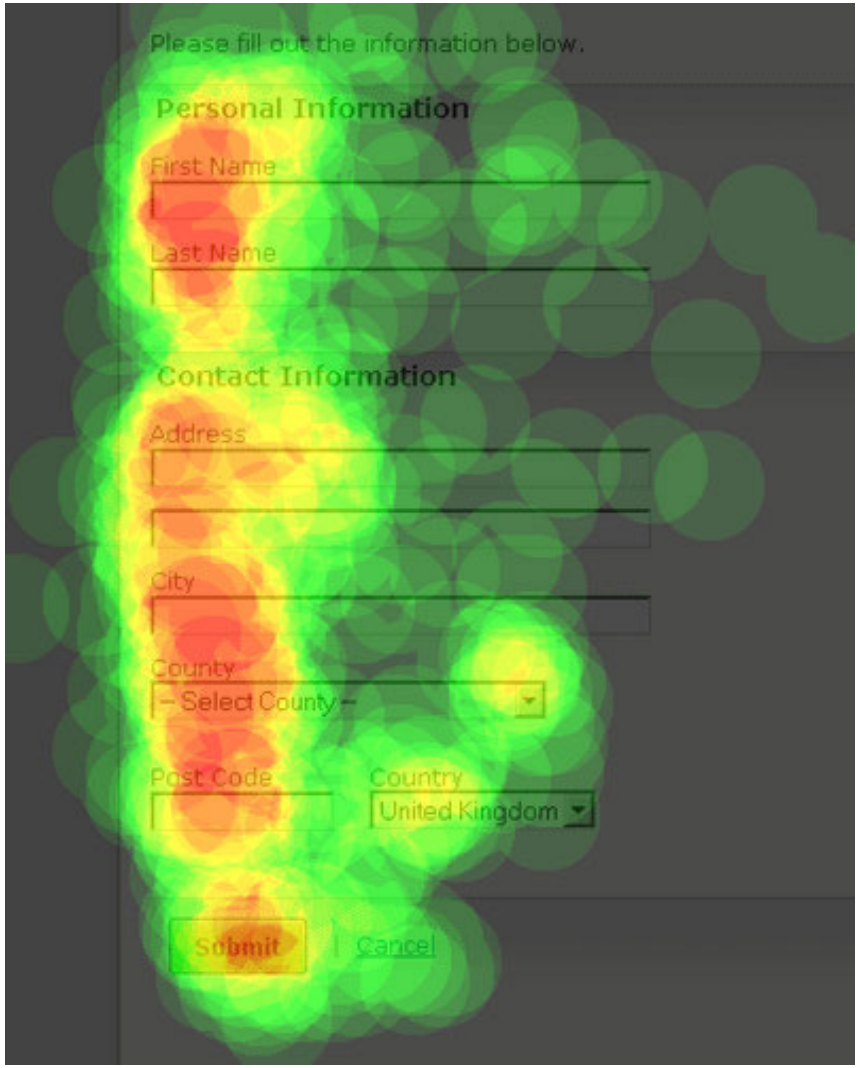


DATA STRUCTURING: **Mapping** Graduated symbol maps

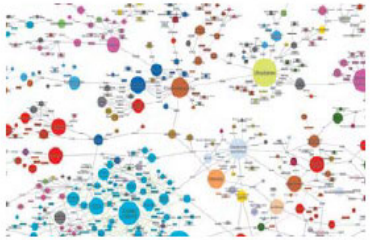


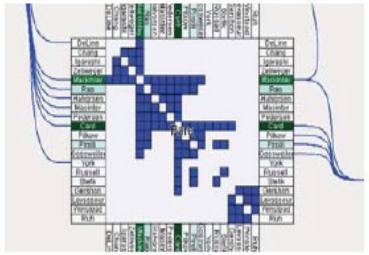
https://www.nytimes.com/elections/2008/results/president/map.html

DATA STRUCTURING: **Mapping** Isometric and isopleth maps



DATA STRUCTURING: **Mapping** Flow and network maps





LINEAR:

Nodes are organized linearly and the links are usually arcs connecting nodes.
Con: It's hard to identify clusters and is only feasible for small datasets.

FORCE DIRECTED:

There are many algorithms that use an iterative process to locate nodes according to physical forces.
Con: There are too many node occlusions and link crossings in dense areas.

CIRCULAR:

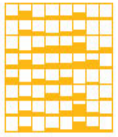
Nodes are organized around the circumference and usually grouped by categories. Links cross the circle and are usually bundled so as to simplify the crossings.
Con: It's hard to identify clusters.

COMMUNITY STRUCTURE:

The focus is on community structures.

GEOGRAPHY BASED:

Spatial location of a node is provided by its geo position.



MATRIX:

Grid of nodes with link information positioned within the cell.

SANKEY TYPE DIAGRAMS:

Nodes are organized vertically and the links horizontally.

FORCE DIRECTED:

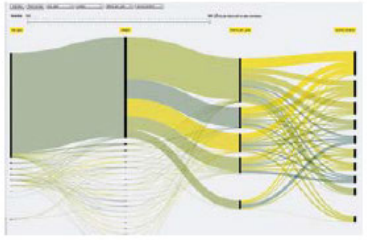
Force directed graphs centered on a node.

POLAR OR RADIAL:

Nodes are organized around a central node, with their position related to the number of hops it takes to reach it.

RADIAL COMMUNITY STRUCTURE:

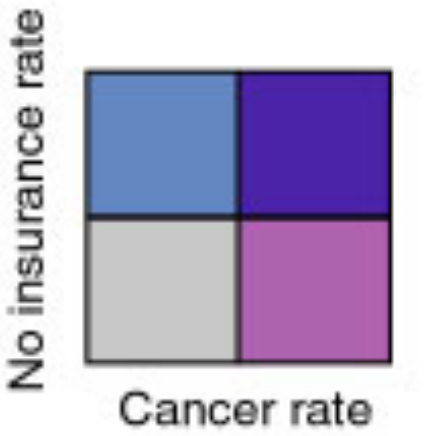
Nodes are organized around a central community

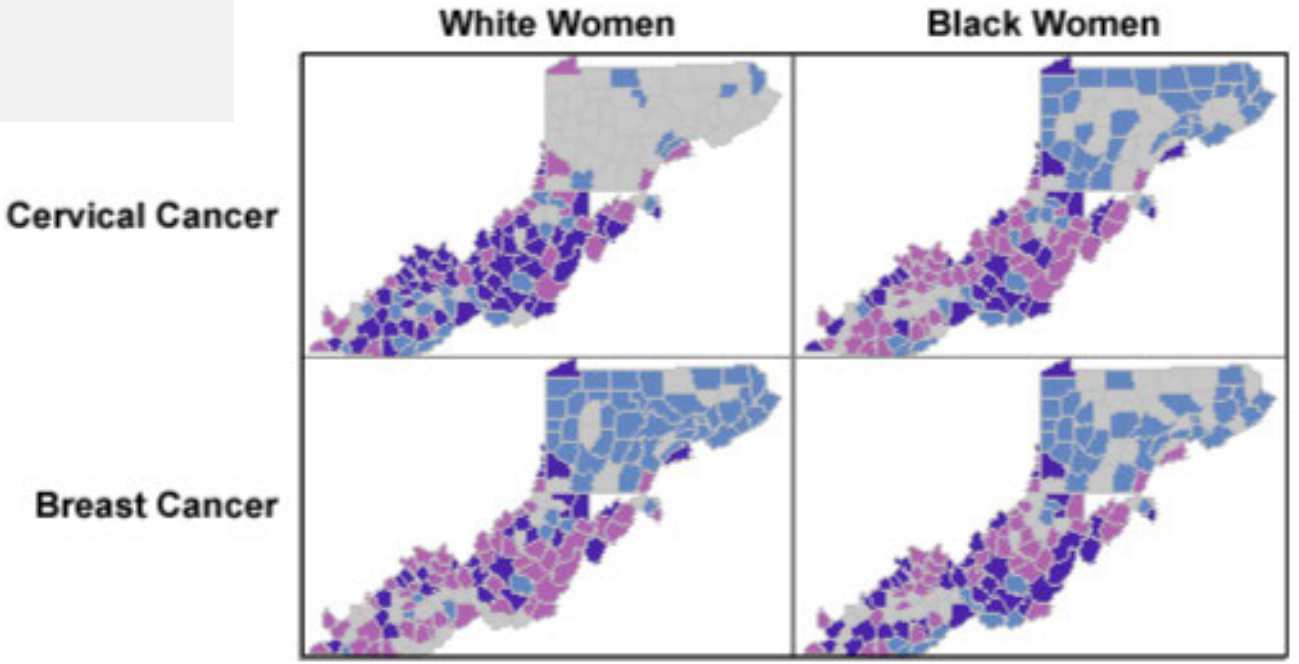


DATA STRUCTURING: **Mapping** Choropleth (Multivariate) maps

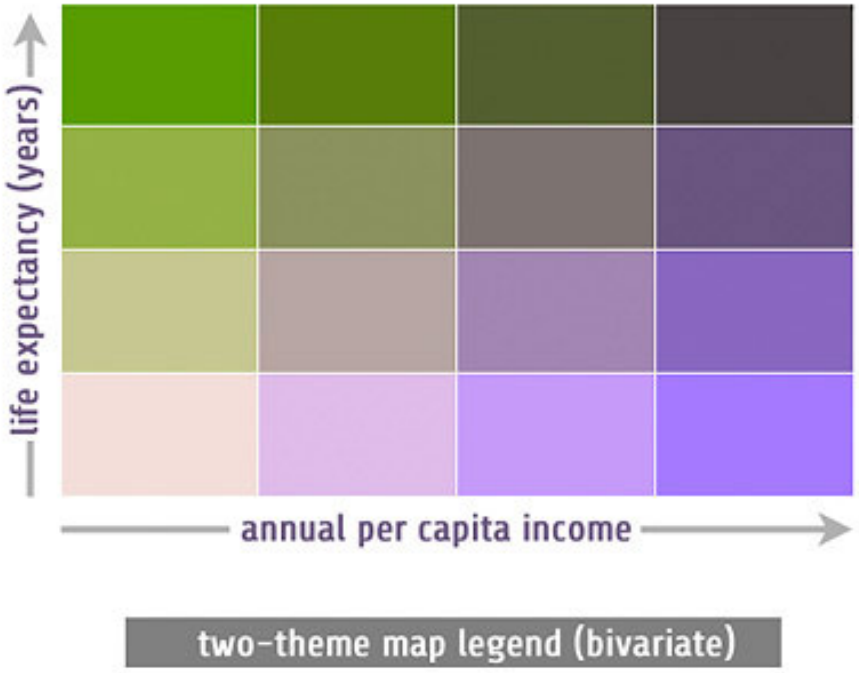
Multivariate Data

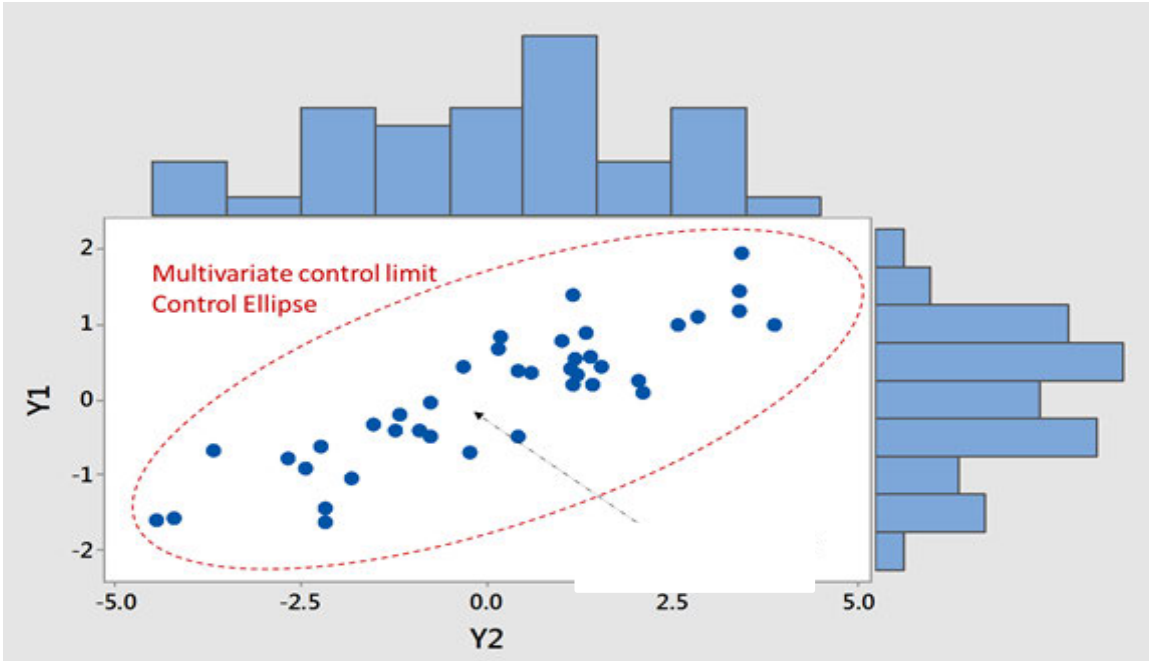
Thematic maps can depict several sets of nonspatial data simultaneously. When a thematic map portrays exclusively one set of data, it is called univariate. If it shows two distinct sets of data, it is called bivariate, and for more than two sets, maps are called multivariate.





DATA STRUCTURING: **Mapping** Choropleth (Multivariate) maps

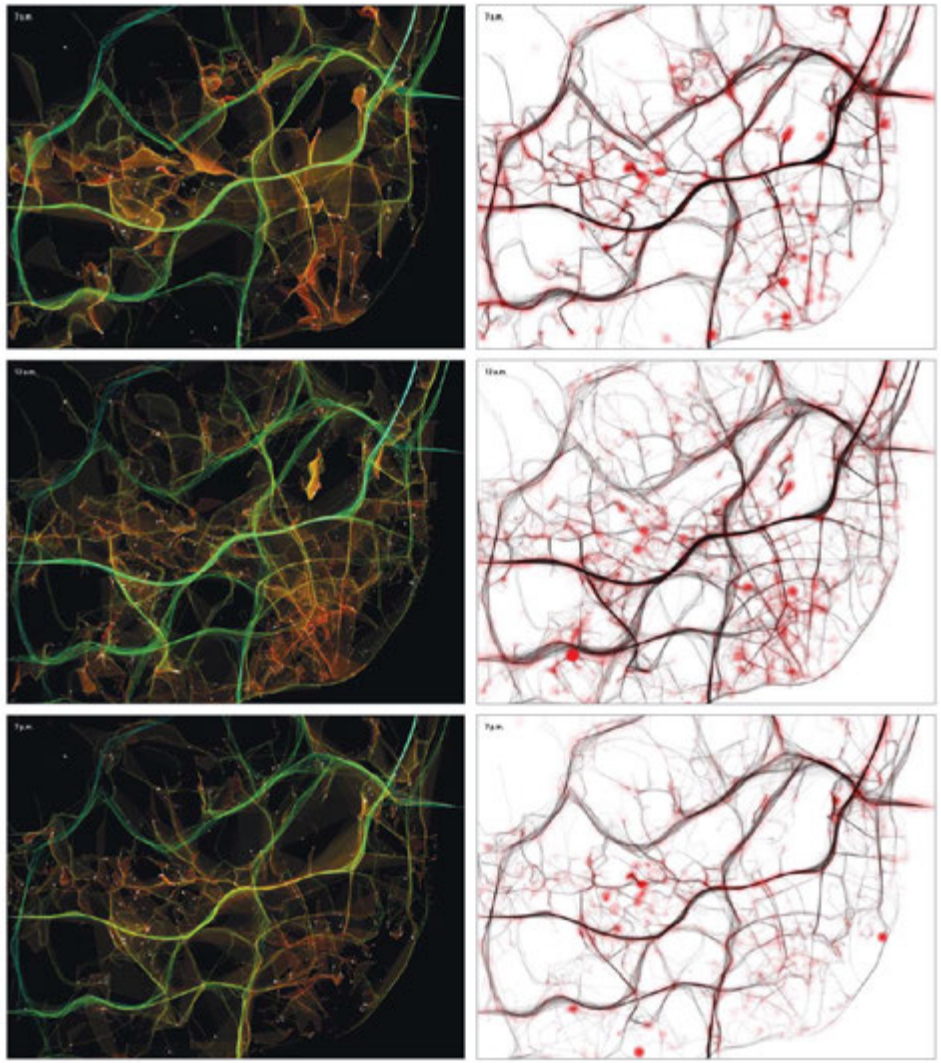




Multivariate Data

Thematic maps can depict several sets of nonspatial data simultaneously.
When a thematic map portrays exclusively one set of data, it is called univariate.
If it shows two distinct sets of data, it is called bivariate,
and for more than two sets, maps are called multivariate.

DATA STRUCTURING: **Mapping** Area and distance cartograms



Start your own IoT Project

HOGESCHOOL ROTTERDAM

To Do's

What you need to do (next lesson)

Make a new Mini Project Repository on your GitHub Account
by providing insights on DATA you learned in this lesson

What you need to do (next lesson)

- Idea is to come up with an **idea** of a IoT data science solution:
  - starting point **A** (need, data, possibilities)
  - aim **B** (added value! not just a technical solution)
  - you need to come up with:
    - **both A and B**, as well as
    - the **solution** to get from A to B

- *Not good ideas:*
  - *take clean data and a given task (e.g., prediction) and apply a machine learning method*
  - *take data and visualize it without any particular aim or goal*
  - *copy an existing data science project*

  => **try to be creative!**

Example Mini Projects

- These are primarily to give you an idea of suitable topics
- IT IS NOT ADVISABLE TO ACTUALLY CHOOSE ONE OF THESE
- Instead, they are just examples to give you the flavour
- Elements of the project (will be evaluated based on these):
  - Data: sources, wrangling, management
  - Data analysis: statistics, machine learning
  - Communication of results: summarization & visualization
  - Operationalization: added value, end-user point of view

Example Mini Projects

EXAMPLE #1: HOW TO BE POPULAR ON SOCIAL MEDIA

1. Create a system that recommends ways to boost social media popularity.
2. Instagram/twitter/facebook data from open APIs. Challenges: extracting data following accepted policies. Finding other relevant data sources. Can involve big data problems.
3. Predicting popularity using machine learning tools. In case of text content, e.g., word clouds, clustering, regression.
4. For example, visualization of the data so that a user user can identify good strategies: e.g., an interactive hashtag exploration tool that highlights common choices based on co-occurrence.
5. The more popularity, the better the system. Usability also important (considering target group, e.g., teenagers wouldn't appreciate boring statistical graphics).

Example Mini Projects

EXAMPLE #2: OK BUT FIRST COFFEE

1. Sometimes the coffee maker in Gurula (student space) is out of coffee just when you need it the most. It's easy enough to make more, but you'd like to know beforehand.
2. Create your own data by installing a webcam (or using one that may already be there) pointing at the coffee maker. Be sure not to violate anyone's privacy and be careful about controlling access to the camere feed. Alternative: use user input.
3. Try using computer vision to predict the amount of coffee left.
4. Share information through an API or an app (possibly a simple visualization, or just the number of cups left).
5. No coffee = bad, coffee = good. Also, possibly recommend how many cups to brew based on expected consumption, or predict peaks in demand.

Example Mini Projects

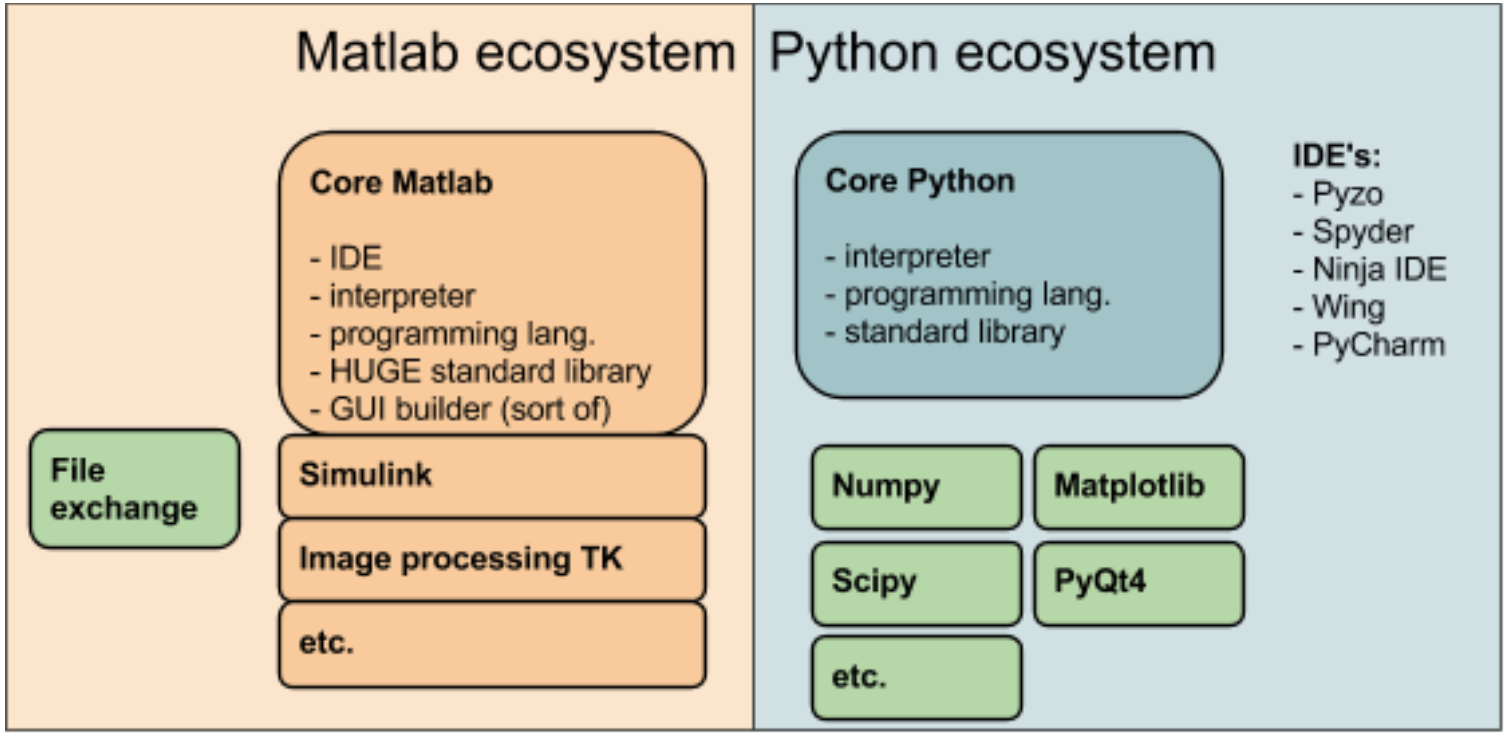
EXAMPLE #3: RENT-A-FLAT

1. The demand for rental flats in Helsinki is leading to high, and increasing rental prices. However, the rents per area vary greatly. Which areas are underrated (and "underrented")?
2. Helsinki open https://dev.hel.fi/apis/open311/ . Other possible data sources can include info about important services (bars, computer stores, ...) including info about their popularity.
3. Summarization of data to generalize to future rents.
4. Visualization of GIS data, possibly over time. In addition to rental prices, the visualization can display, e.g., crime statistics, availability of services, "hipster index"
5. Information that supports individual users (with different preferences) in finding a cheap flat. Could also be used by city officials to support planning.

Next Time:

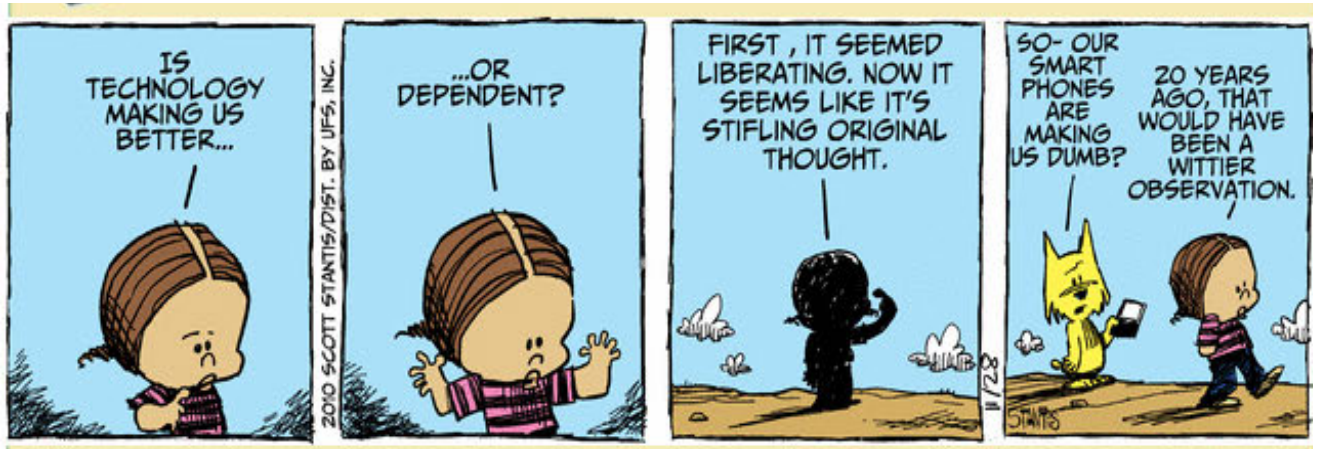
Which data tools should you use?

Matlab vs Python



https://pyzo.org/python_vs_matlab.html

HOGESCHOOL ROTTERDAM



This lesson was developed by:
Rob van der Willigen
CMD, Hogeschool Rotterdam
NOV 2019

Creative Commons License Types

	Can someone use it commercially?	Can someone create new versions of it?
Attribution	👍	👍
Share Alike	👍	Yup, AND they must license the new work under a Share Alike license.
No Derivatives	👍	👎
Non-Commercial	👎	Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike	👎	Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives	👎	👎

SOURCE
http://www.masternewmedia.org/how-to-publish-a-book-under-a-creative-commons-license/

http://creativecommons.org/licenses/

http://empoweringthenatives.edublogs.org/2012/03/15/creative-commons-licenses/