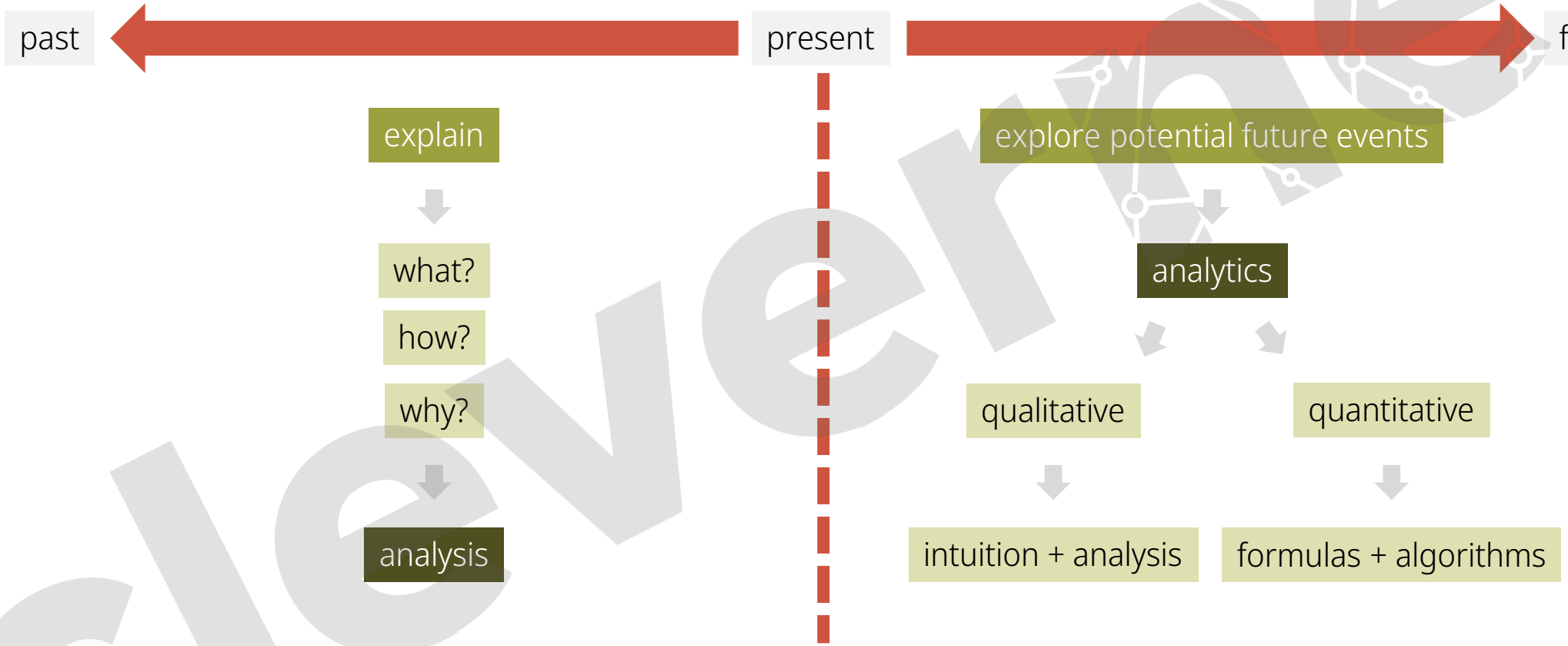
the world of data

general concepts

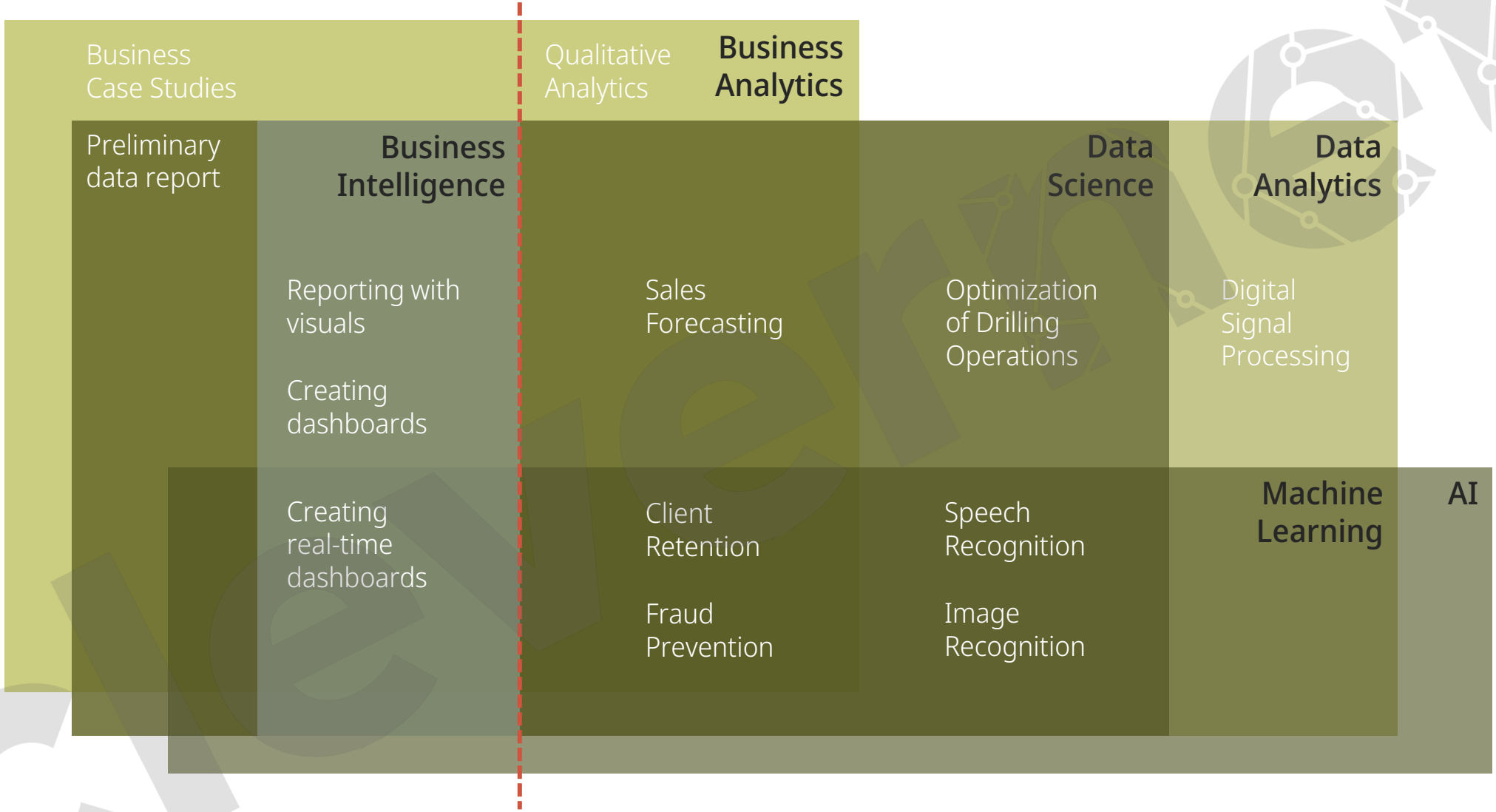
the world of data

analysis vs analytics



the world of data

datascience diagram

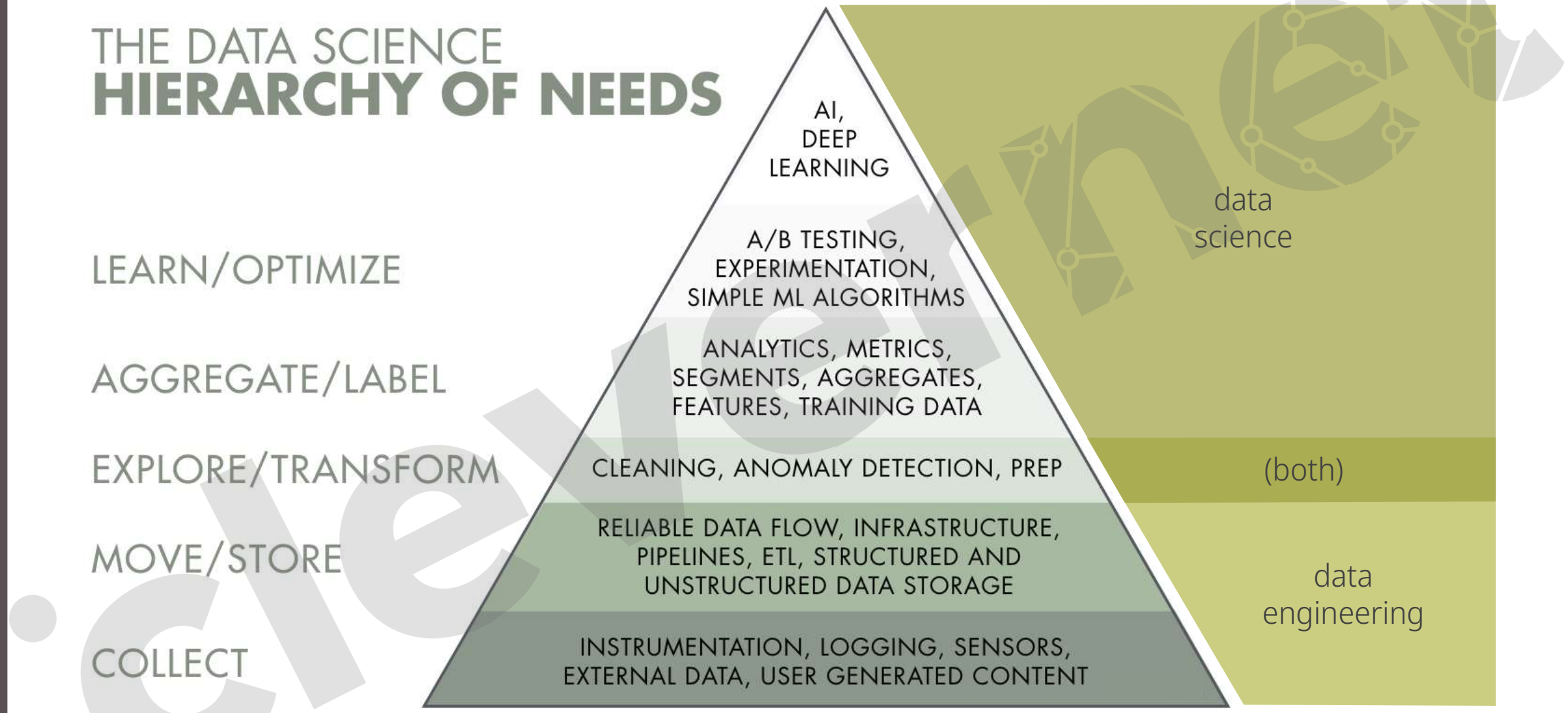


*"More often than not, companies are **not** ready for AI.*

Maybe they hired their first data scientist to less-than-stellar outcomes, or maybe data literacy is not central to their culture.

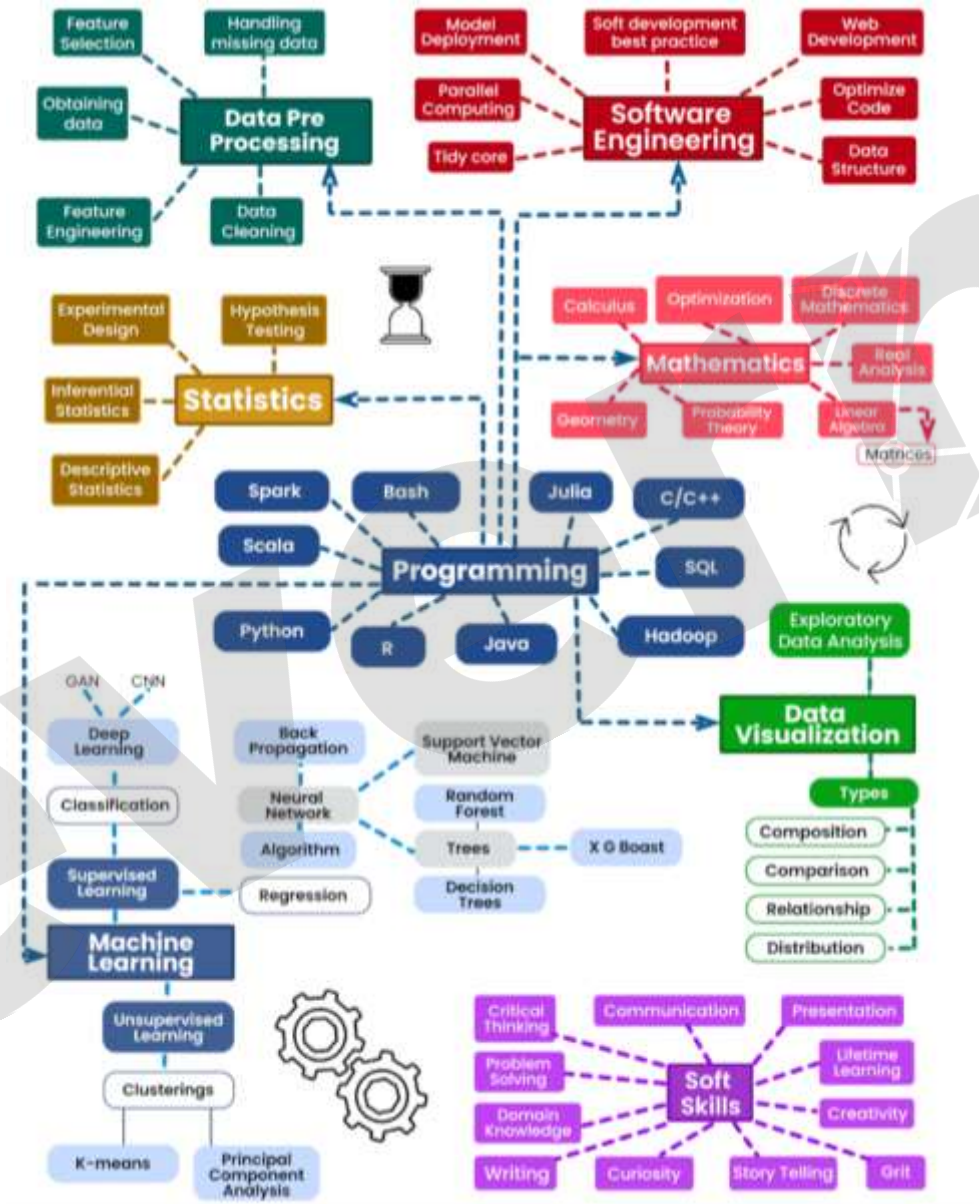
*But the most common scenario is that **they have not yet built the infrastructure** to implement (and reap the benefits of) the most basic data science algorithms and operations, much less machine learning."*

Link to the article

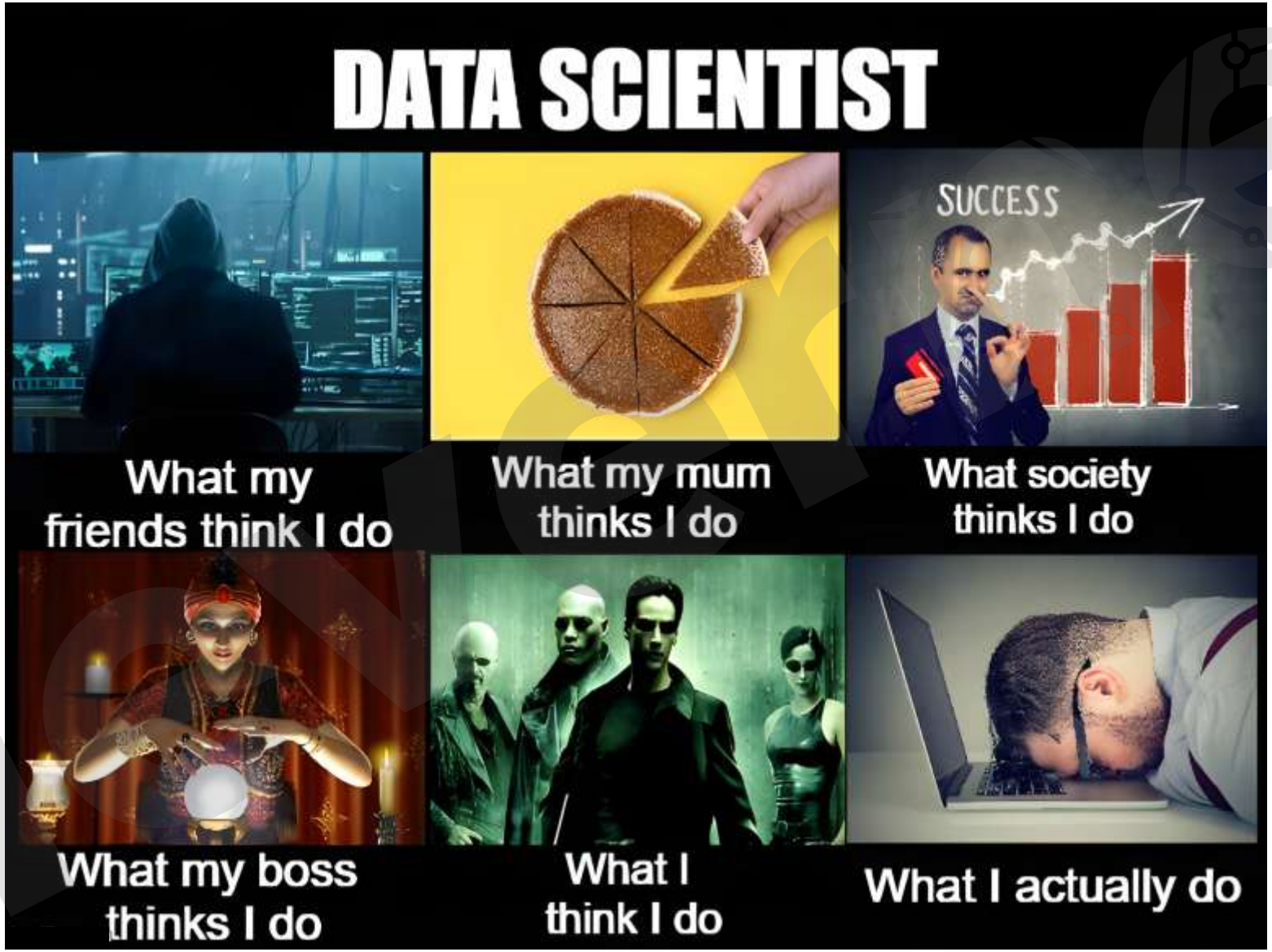

THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
data science
(both)
data engineering

the world of data

data science landscape



the world of data



the world of data

the world of data

data professionals types

The Legends

A Legend has been around since before data science was really a thing, so probably studied math extensively, is a great programmer, and has mastered several languages (including almost-forgotten ones, like C++).

The Star DS Managers

A Star DS Manager concerns him or herself with overall team productivity, removing roadblocks, tooling, etc., hence is not really focused on the data science itself.

The Vertical Experts

A Vertical Expert has lots of experience in a particular domain and is valuable for obvious back-knowledge from the get-go, but can find it difficult to think outside the box on standard problems or questions.

the world of data

data professionals types

The Statisicians

A Statistician is extremely literate in statistics and might have specific experience in the area of finance, but not be as well-versed in working with really huge datasets.

The ML Engineers

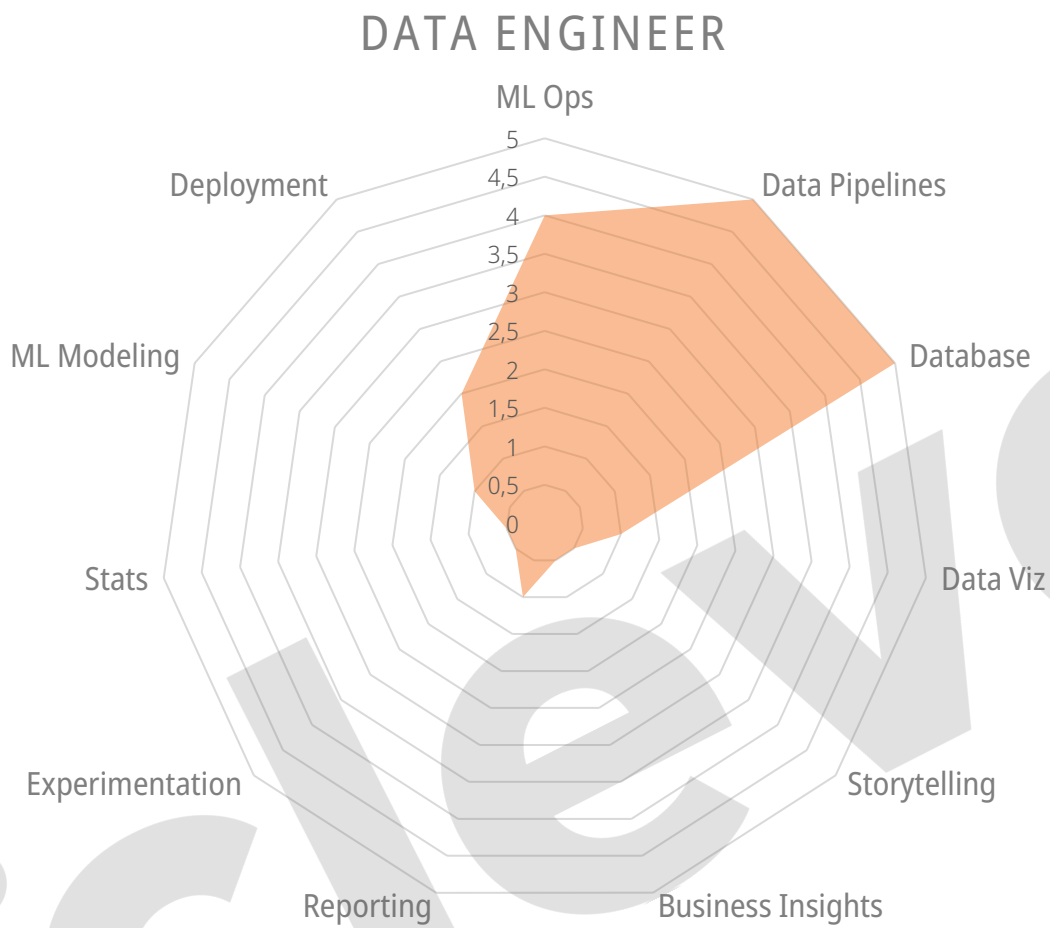
A Machine Learning (or ML) Engineer brings an understanding of model development, software architecture, and model deployment that can be essential for certain types of businesses and datasets.

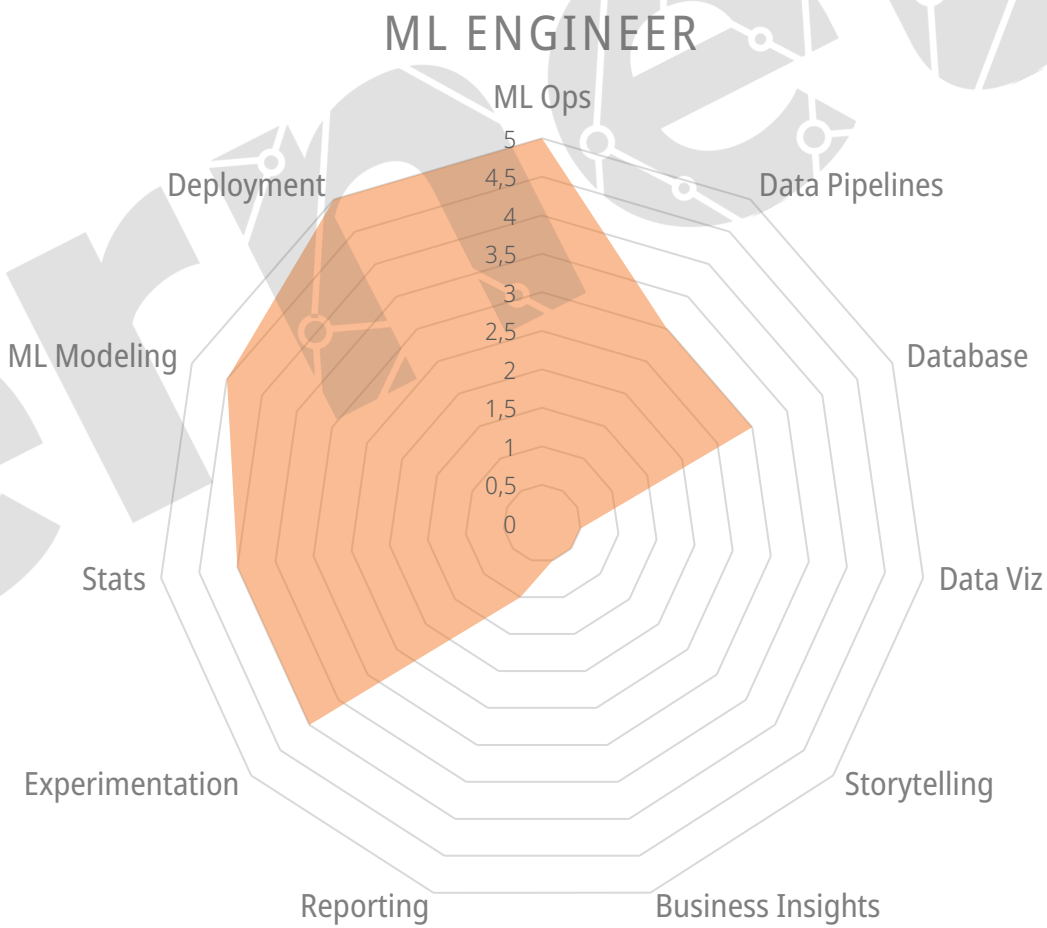
The Generalists

A Generalist knows a little something about everything and is a good fit for larger teams where responsibilities are always shifting, but also for a one-man show at a smaller company where versatility is essential.

the world of data

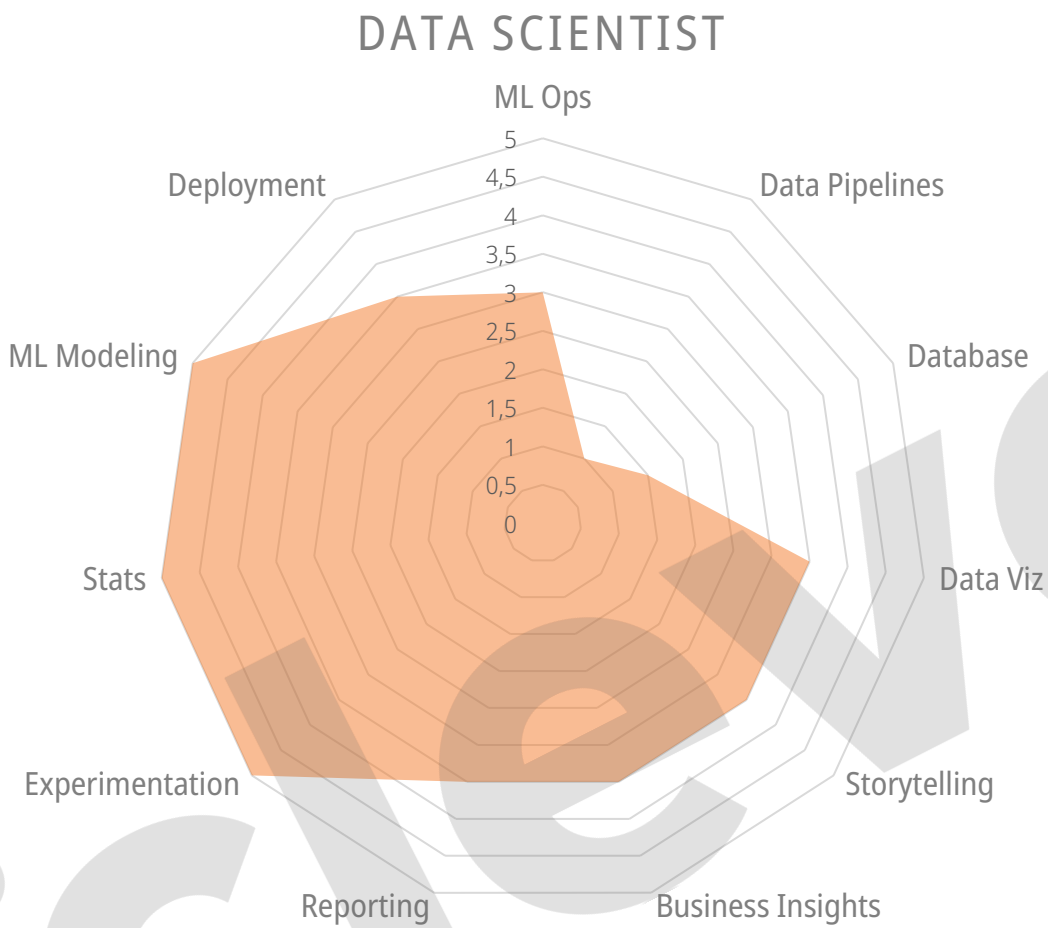
data professionals skills

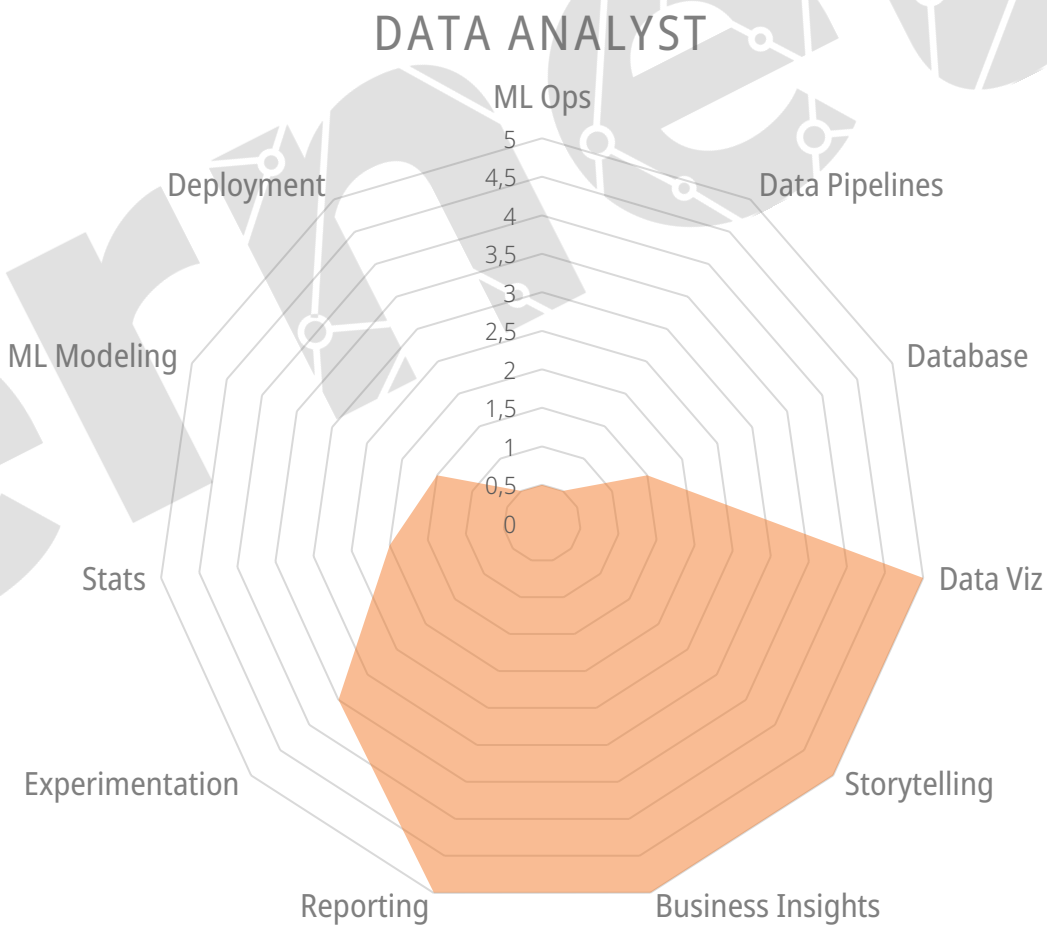




the world of data

data professionals skills

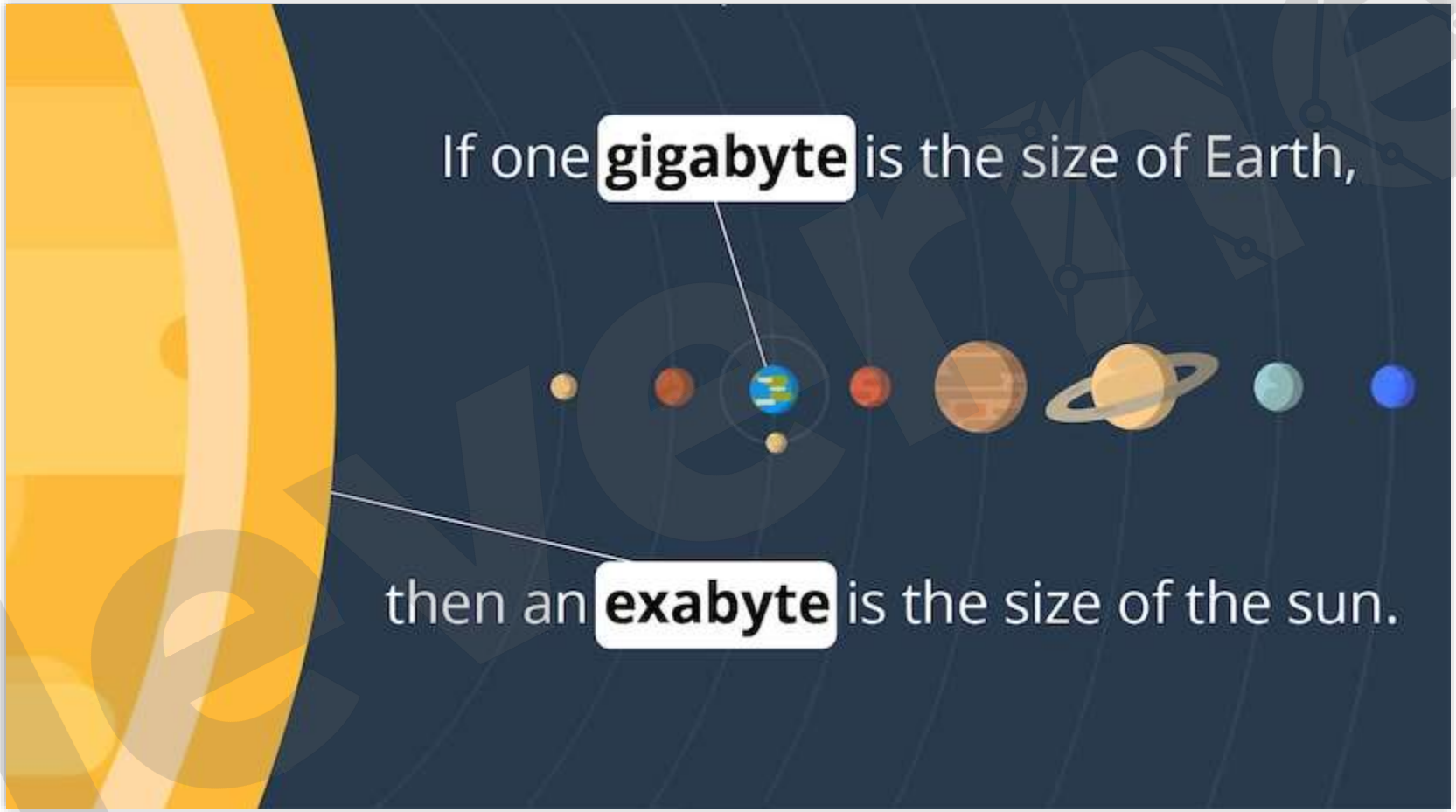




the world of data

DS & big data - an overview of the 10 V's

DS & big data


If one gigabyte is the size of Earth,
then an exabyte is the size of the sun.

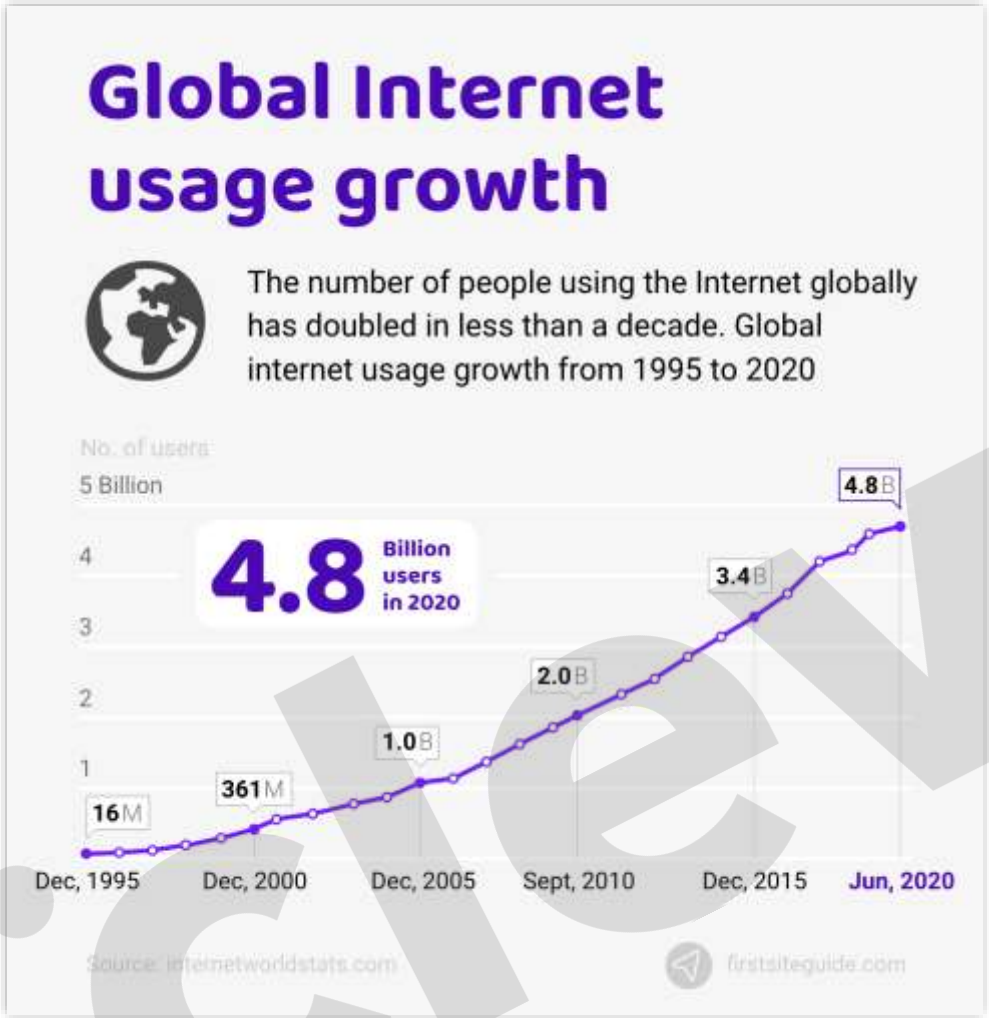
DS & big data

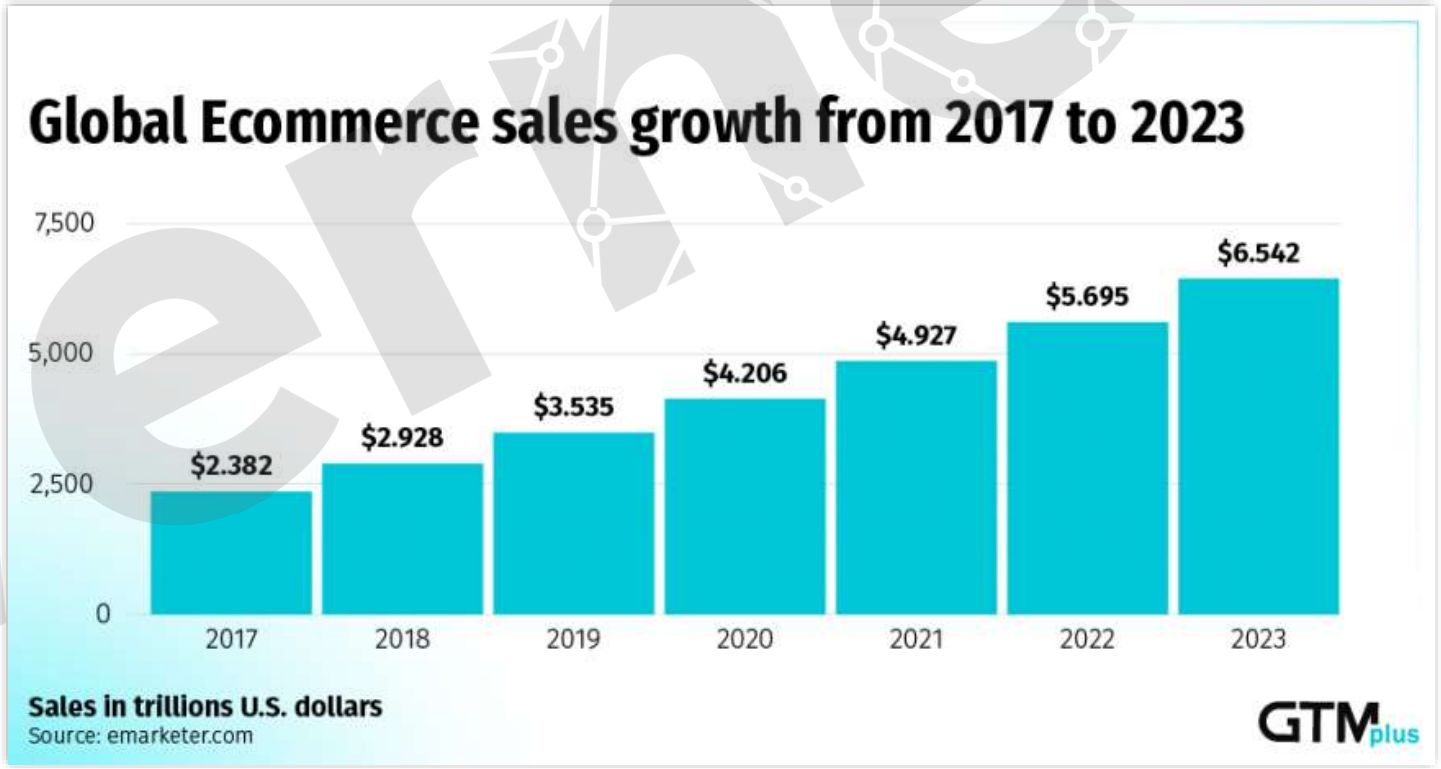
one phone's:	generates:	multiplied by:		result:	solution:
SMS	50 gigabytes of data per month	5 billion smartphones	=	250 exabytes 	BIG DATA
Calls					
Email					
Photos					
Videos					
Browsing					
Music					

A lot for the human mind to process

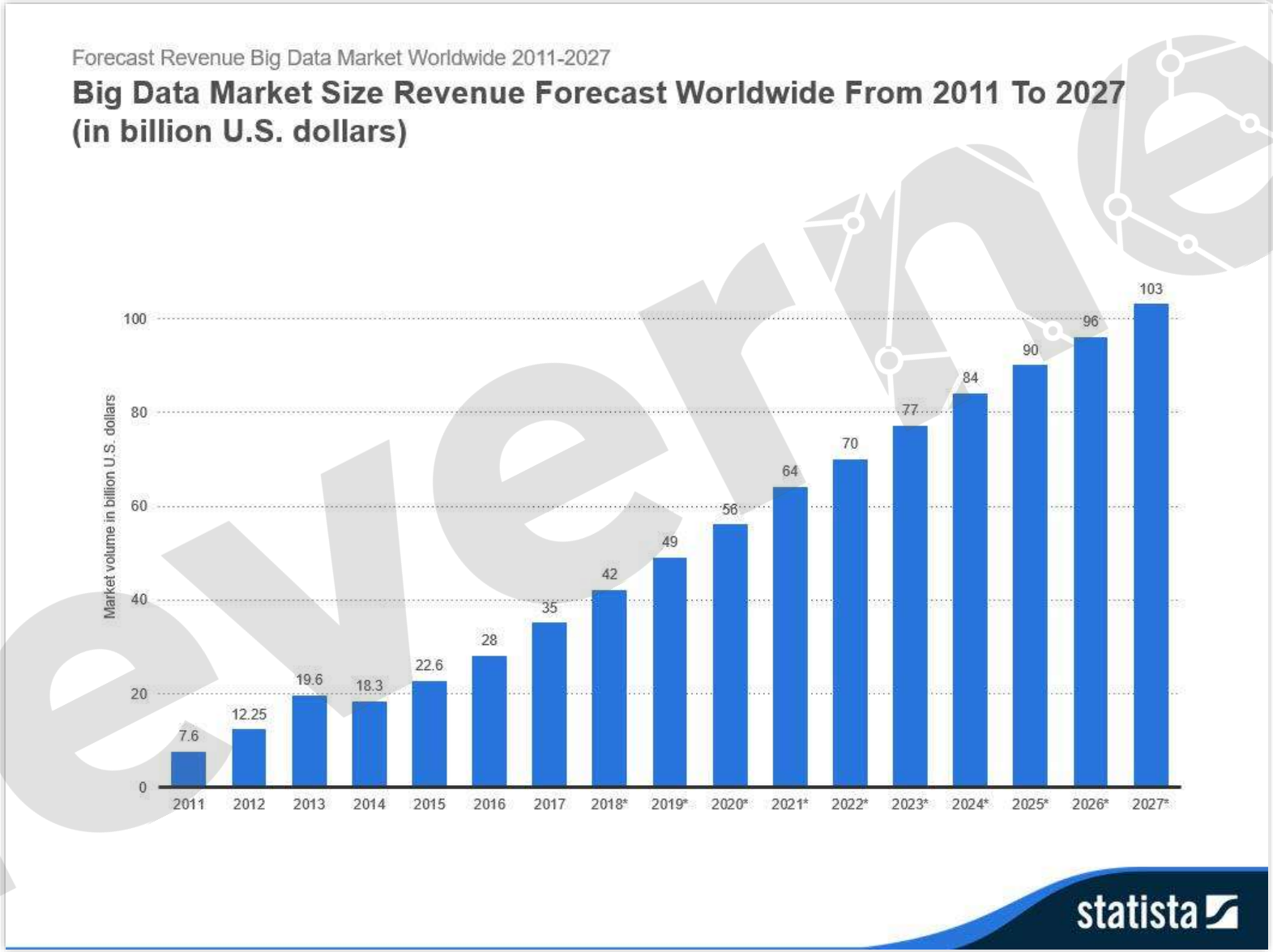
Unmanageable by traditional computing systems

DS & big data





DS & big data


Forecast Revenue Big Data Market Worldwide 2011-2027
Big Data Market Size Revenue Forecast Worldwide From 2011 To 2027
(in billion U.S. dollars)
Market volume in billion U.S. dollars
0
20
40
60
80
100
7.6
12.25
19.6
18.3
22.6
28
35
42
49
56
64
70
77
84
90
96
103
2011
2012
2013
2014
2015
2016
2017
2018*
2019*
2020*
2021*
2022*
2023*
2024*
2025*
2026*
2027*
statista

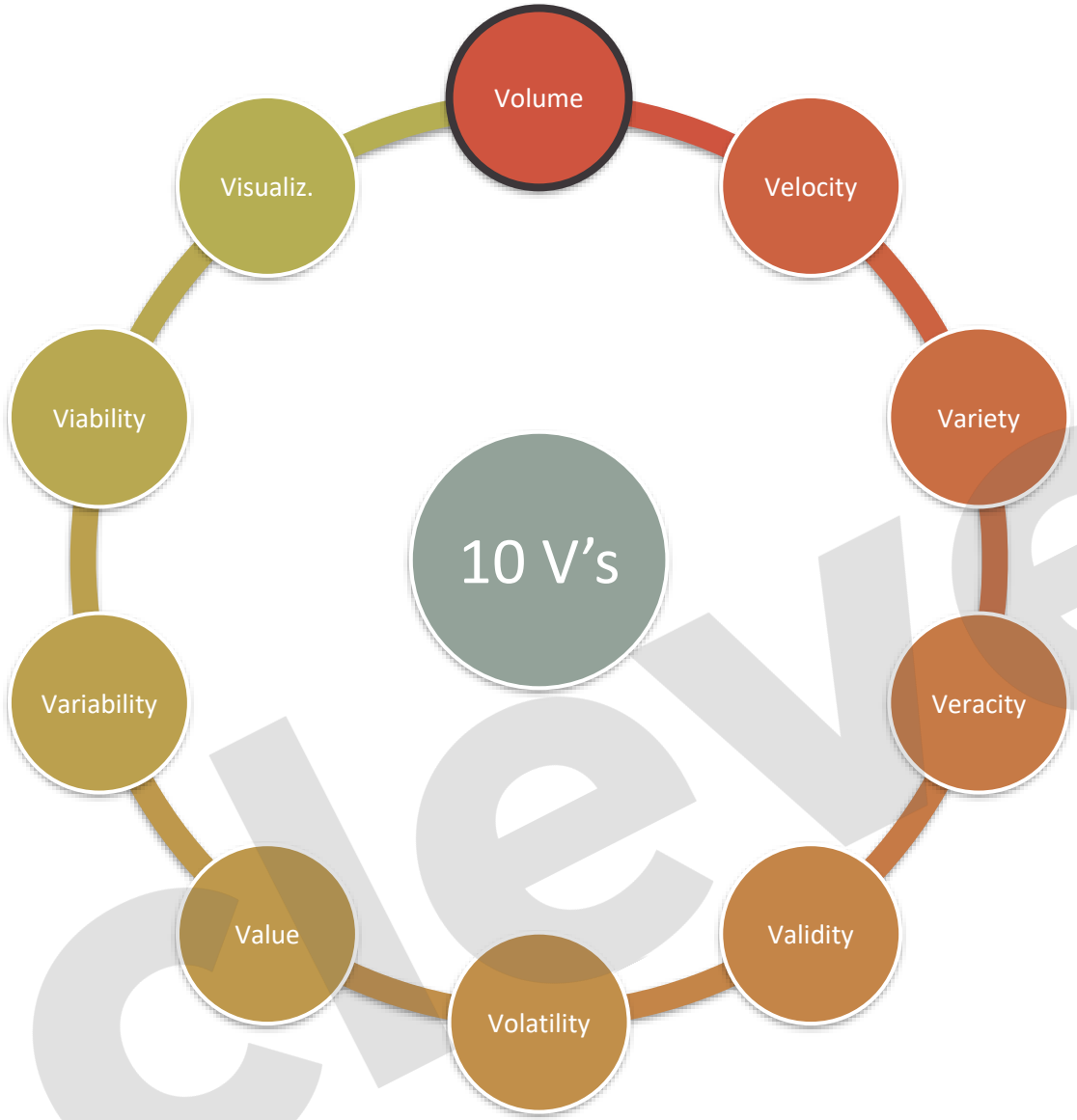
DS & big data

“90% of all data ever created was created in the past two years.

It is expected to double every year.”

DS & big data

how to classify as big data



volume

Examples of volume:

- self driving cars
- social networks
- customer loyalty cards
- GPS device data
- online shopping transactions
- healthcare patient records

"Volume" is when the data itself becomes a part of the problem.

How much data we have?
How rapidly is it growing?

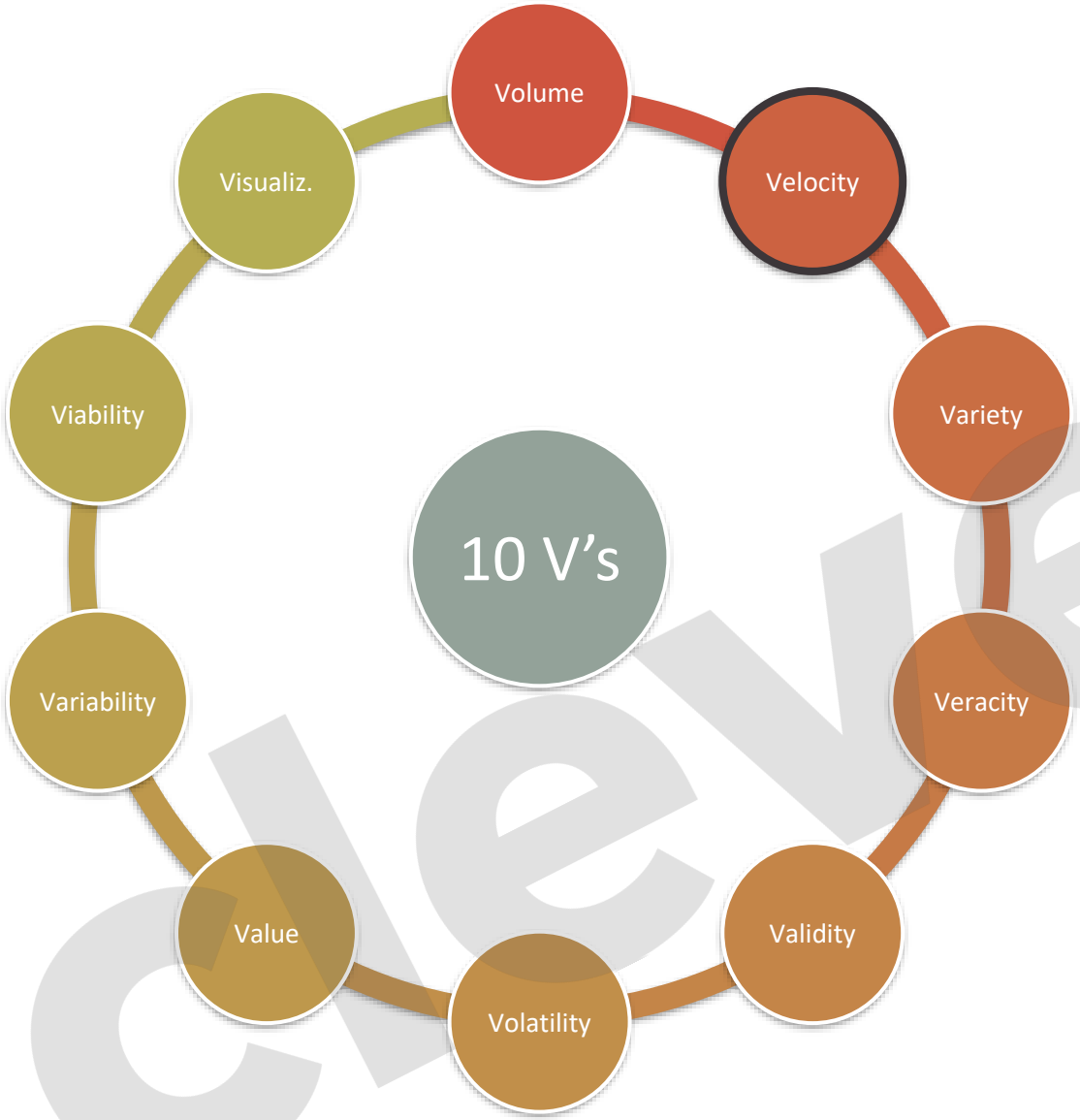
This creates distinct storage, management and processing demands.

Traditional approaches will not work: RDBMS's will fail to store/retrieve, or will take too much time, or will increase the computational costs to unbearable amounts.

What's "big" to an organization might not be for another.

DS & big data

how to classify as big data



velocity

Velocity may be defined as the amount of time it takes to process the data and generate insights or at which speed the data is accessible.

This speed must be in real-time or near real-time.

Velocity refers to data in motion:

- How quickly is the data coming to an organization
  - Social networks, sensors, click-stream data, feedbacks
- How long does it take for the organization to process it

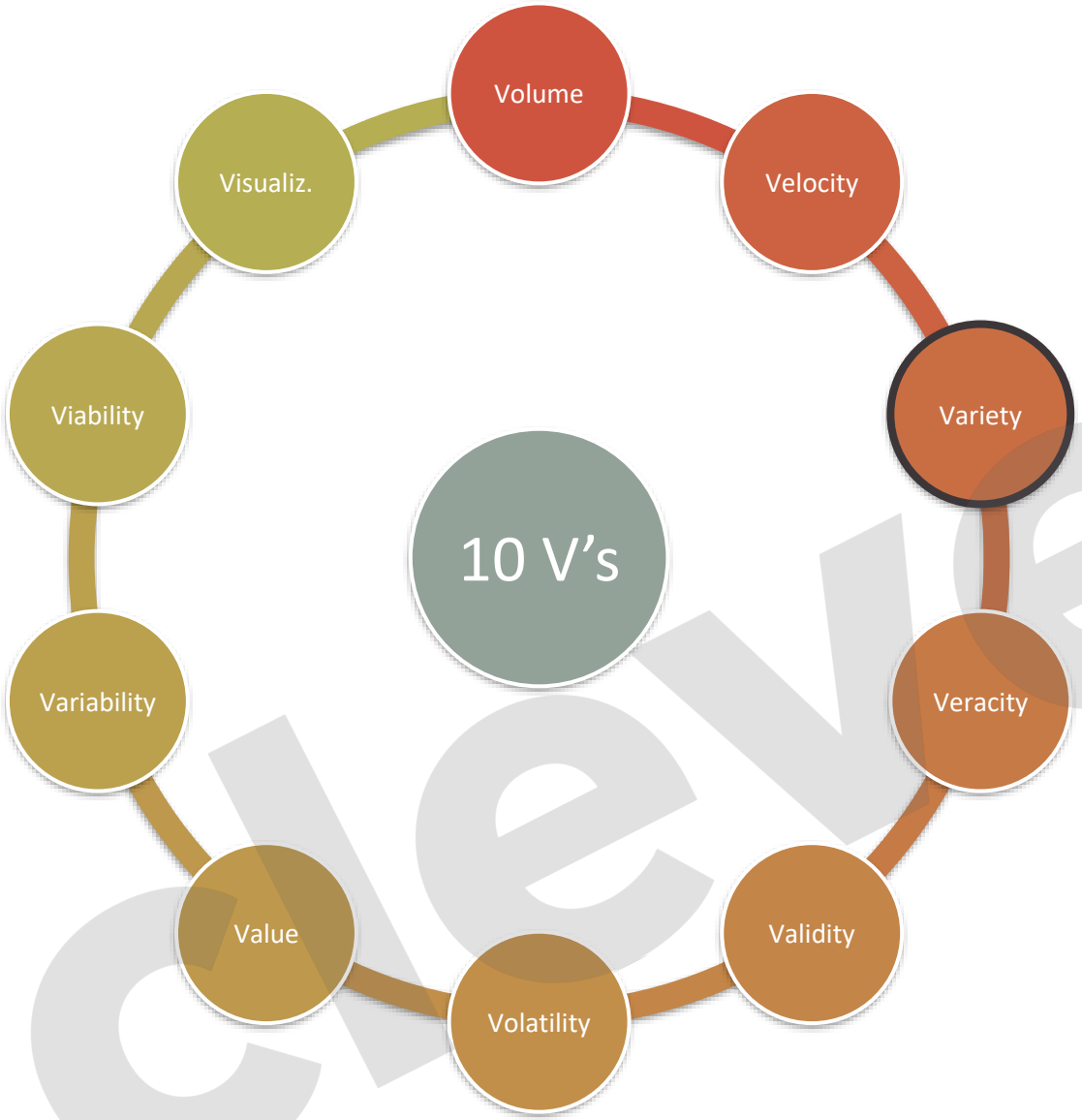
For some applications, the value life span of the data might be short, so if the analysis is not done on time the insights will only have historic value.

Examples of velocity:

- Targeted ads
- Recommendation systems
- Search

DS & big data

how to classify as big data



variety

Data can have different types:

- Structured - RDBMS's
- Semi-structured - XML, JSON, Logs, Sensor data, RSS
- Unstructured - Books, E-mails, Image, Video, Sound
  - On average, it's 80% of the organization's data
  - It has the fastest growing rate
  - Special processing is required
- Metadata - machine generated, usually appended to a data source, crucial to deal with semi and unstructured data types.

And can come from multiple sources:

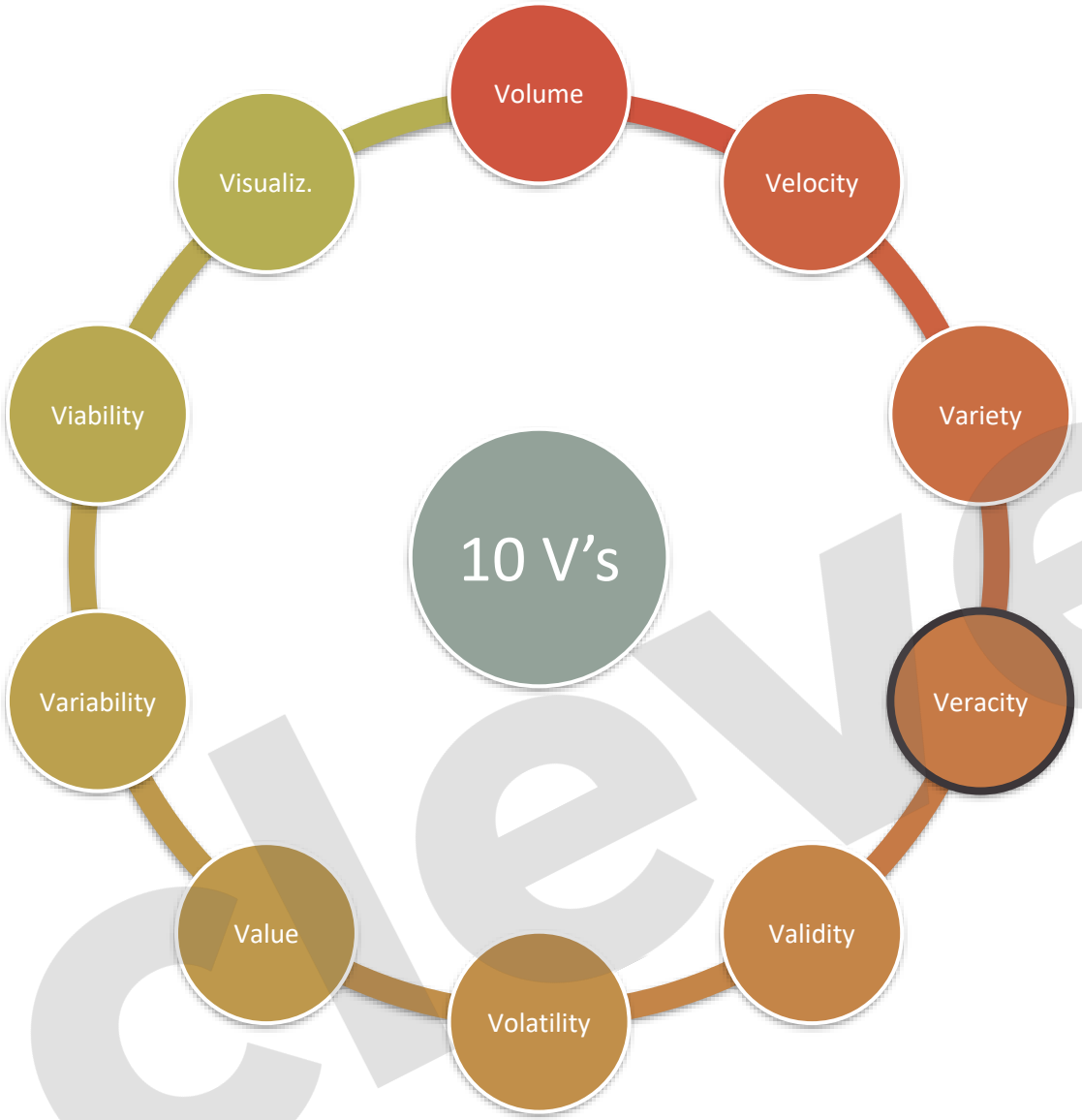
- Inside the organization
- Data vendors
- Social networks, Websites
- IoT

It's one of the biggest challenges nowadays:

- **Storing, processing and retrieving quickly and cost-effectively**
- **Analysing it all together**

DS & big data

how to classify as big data



veracity

Veracity refers to the quality of the data, it's the trustworthiness of the data in terms of accuracy

It may also be referred to the following:

- Authenticity, Accountability, Origin, Security, Reputation

And is the uncertainty about:

- Consistency, Completeness, Latency, Biases, Noise

Asserting value is a big challenge, sometimes only possible with big data, since it needs several preprocessing techniques to cleanse, enrich or remove invalid data.

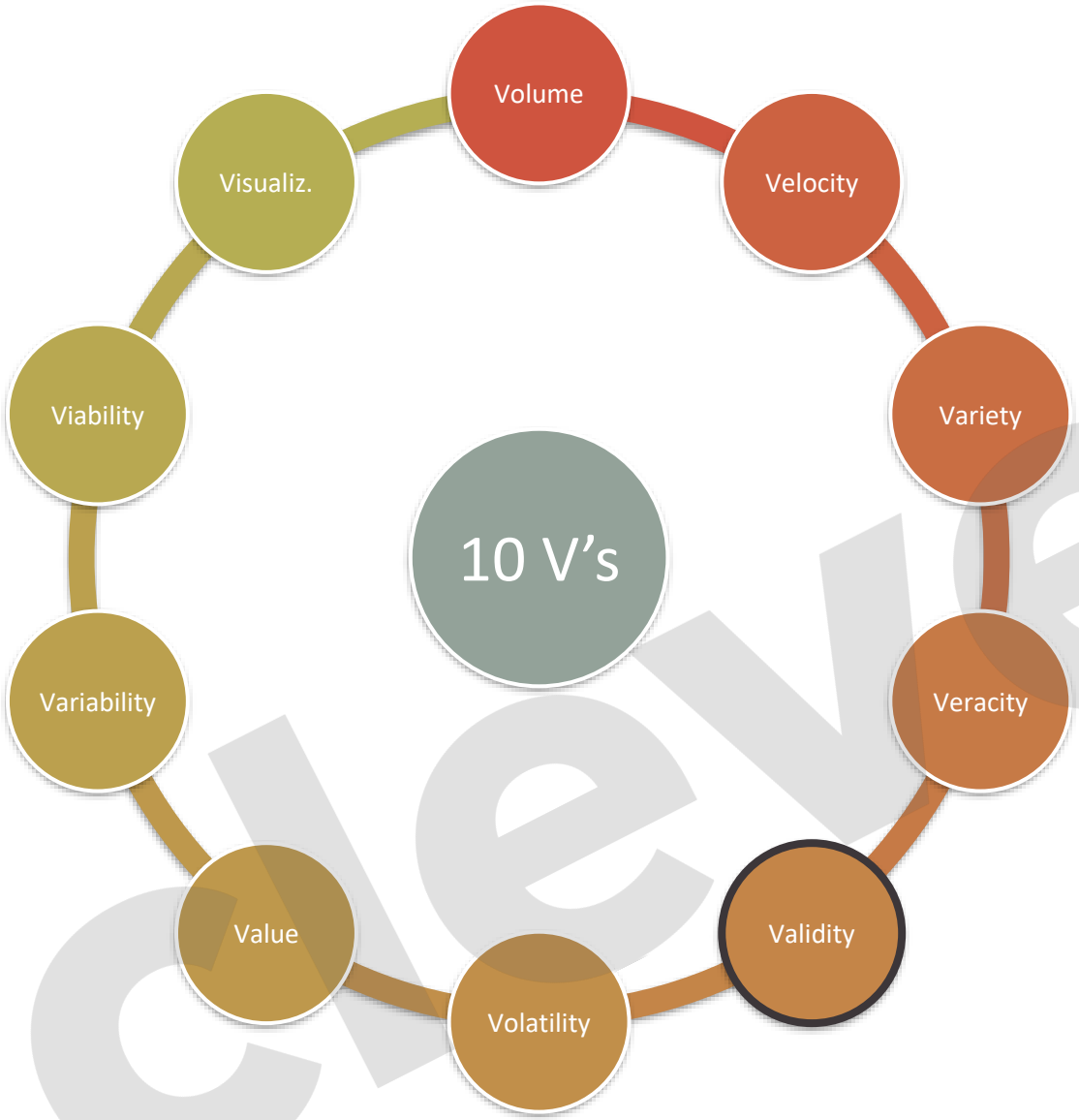
Veracity is critical in automated decision-making or unsupervised ML algorithms.

On average, on most organizations, 70% of the resources are spent in cleaning and preprocessing.

Related: data scientists use the mantra "garbage in, garbage out".

DS & big data

how to classify as big data



validity

Validity is closely related with veracity.

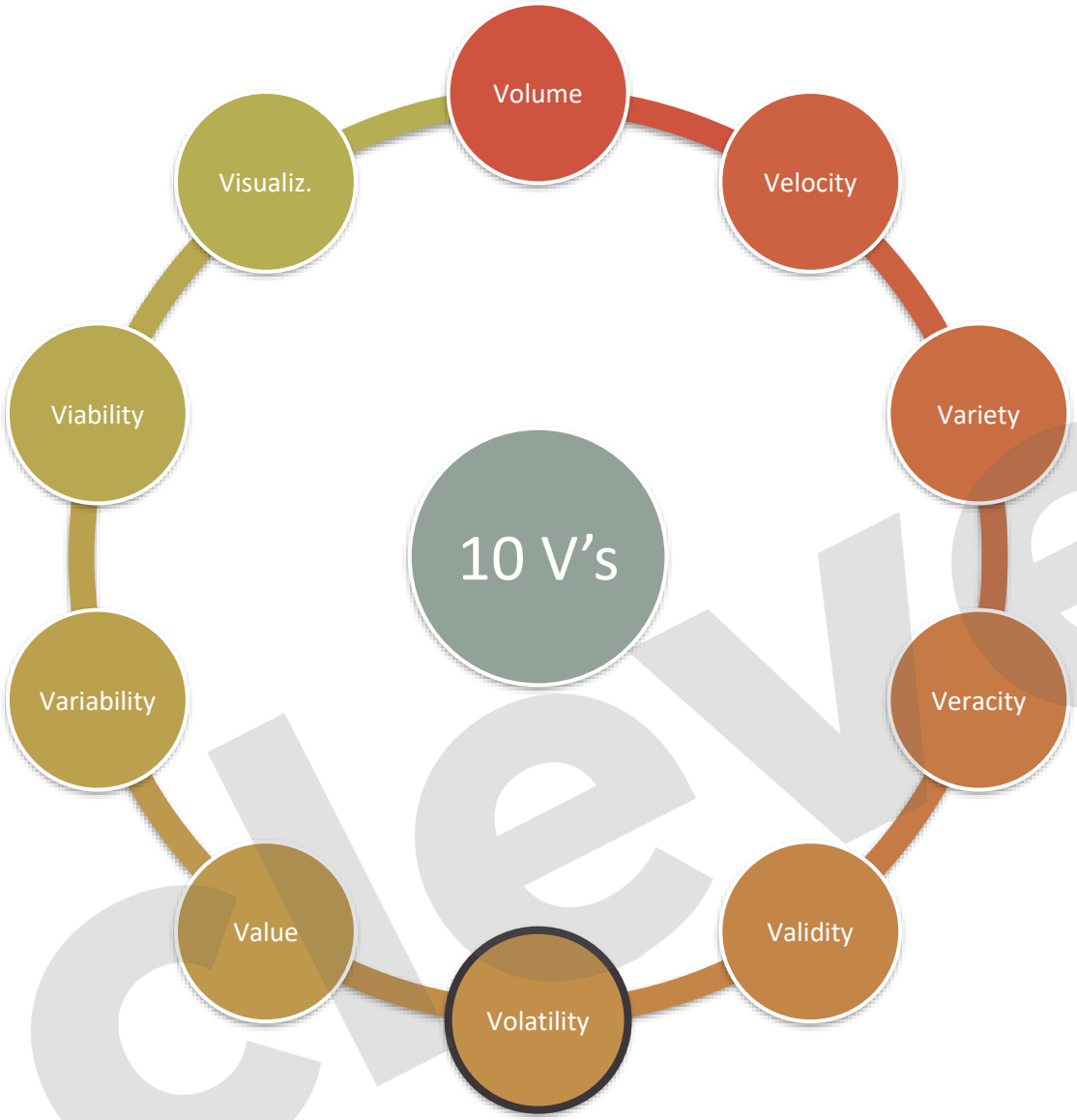
It refers to the following:

- Trustworthiness
- Authenticity
- Accountability
- Correctness
- Appropriateness
- Precision
- Accuracy

An example is the analysis of stock market data in real-time.

DS & big data

how to classify as big data



volatility

Volality is a subset of validity, since it only deals with the time aspect.

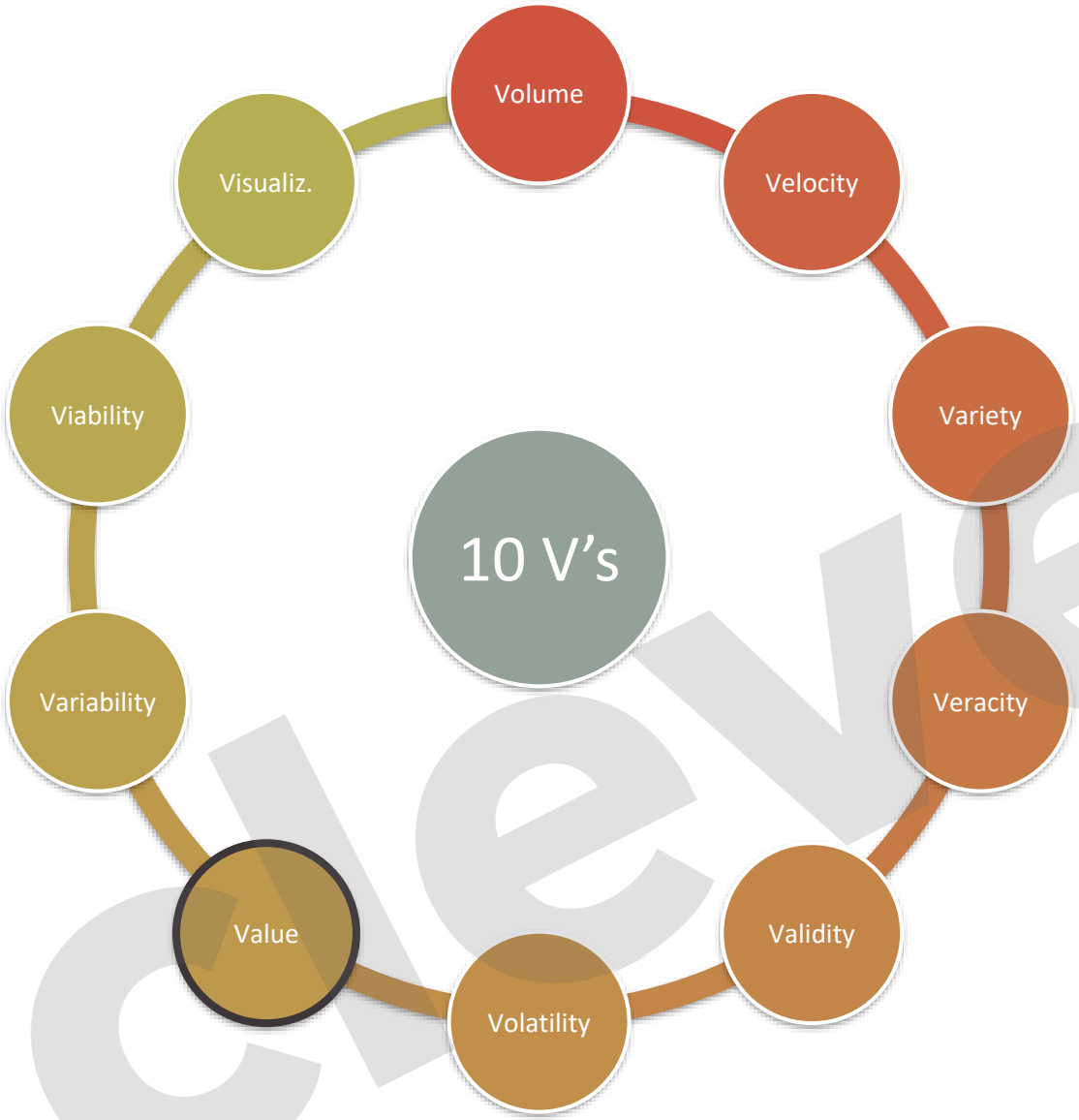
Volatility refers to how long the data is valid and for how long should it be stored before it loses it's validity

If that's the case, big data might be the only solution, due to near real-time processing and output, to leverage this short term data with preexisting data.

An example of this is retargeting.

DS & big data

how to classify as big data



value

Value refers to the usefulness of the data for the organization. **Data itself is not valuable, the analysis and how the organizations use it is what's valuable.**

It implies that just having data is of no use unless we can derive meaningful insights.

Higher the data accuracy, higher the value.

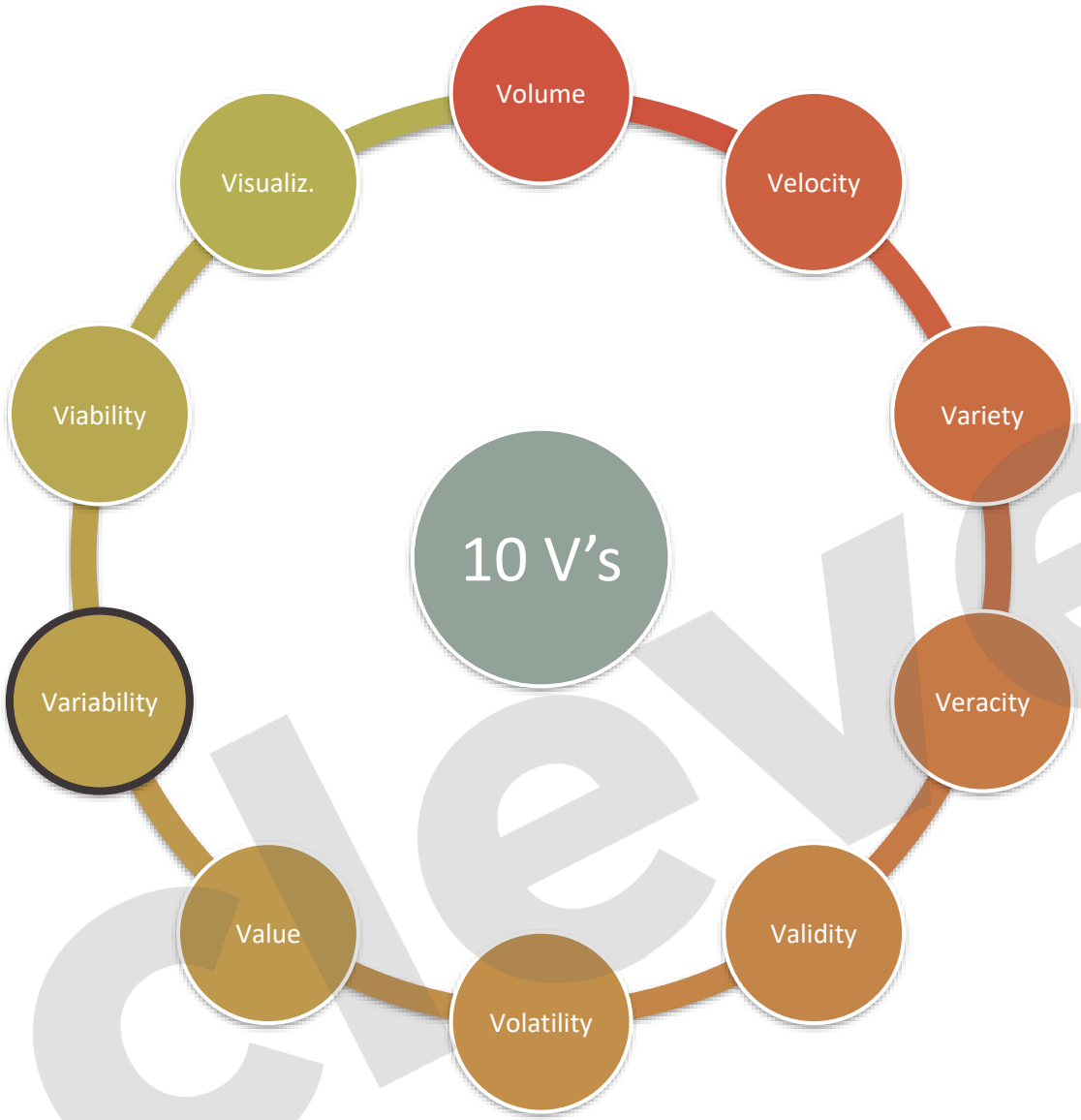
Lesser the data processing time, higher the value, so time and value are inversely related.

Delayed results hinder the quality and speed of informed decision-making.

Related: it's often said that "big data means big business".

DS & big data

how to classify as big data



variability

Variability is different than variety.

Variability refers to the data whose meaning is constantly changing.

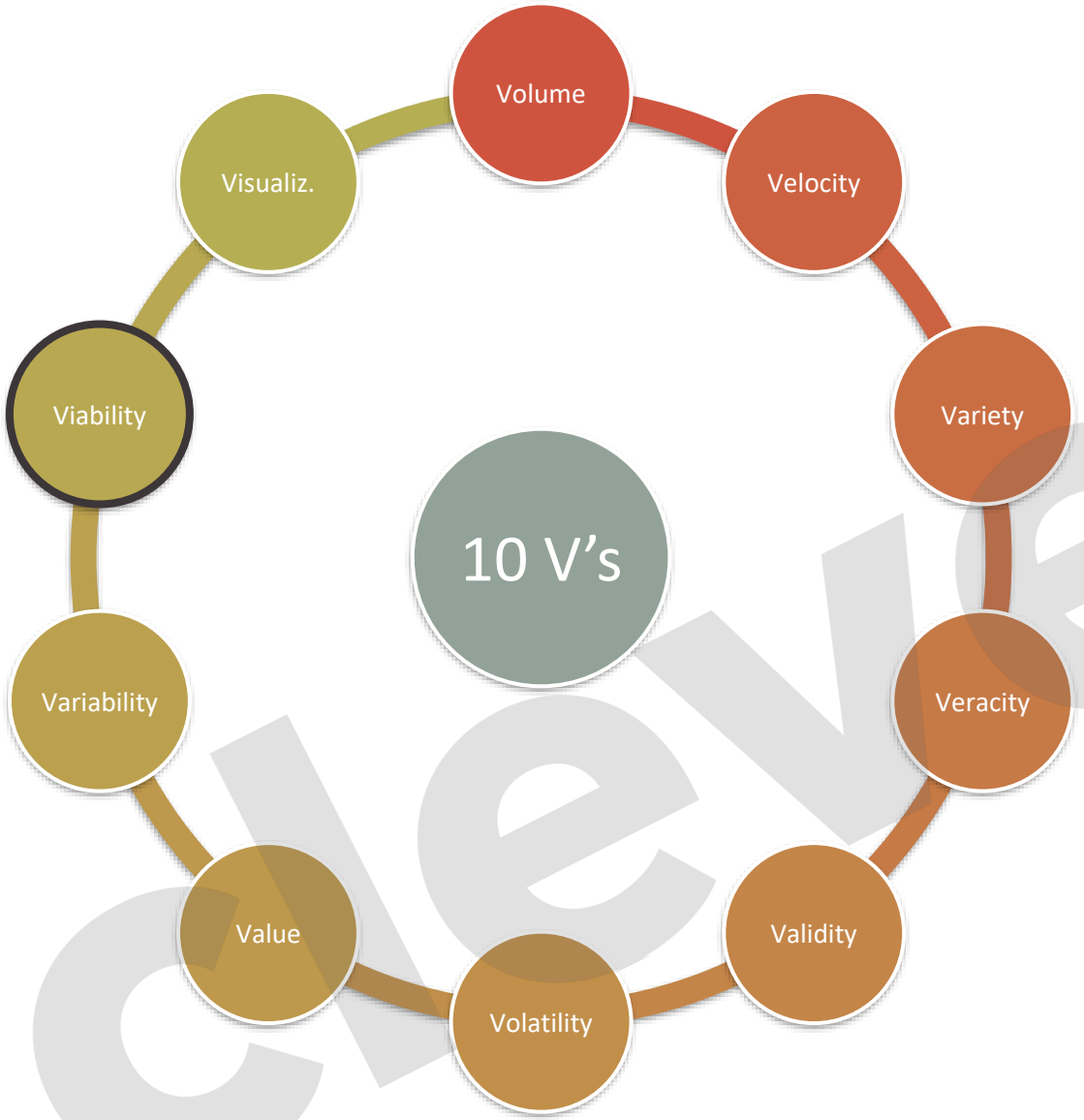
Example: a coffee shop serves cappuccinos and expressos (has variety), but the cappuccino isn't always equal, since it might use a different coffee blend (that's variability).

In data, a good example is sentiment analysis. Given some text, the same algorithm will perform differently if the context of the text is different.

Another example is social listening. If a law enforcement agency is listening to the social chatter on a social network, only with big data they'll be able to distinguish signals from noise.

DS & big data

how to classify as big data



viability

If the data we're collecting isn't just massive, but also has multi-dimensional variables or features, big data can be the solution to uncover which of these are actually important to the analysis, thus reducing processing/analysis costs.

The goal is to uncover the hidden relationships between these variables in order to determine which are actually relevant to the business problem.

In other words, big data can help deciding, in the early stages, which data features matter or can contribute to better insights.

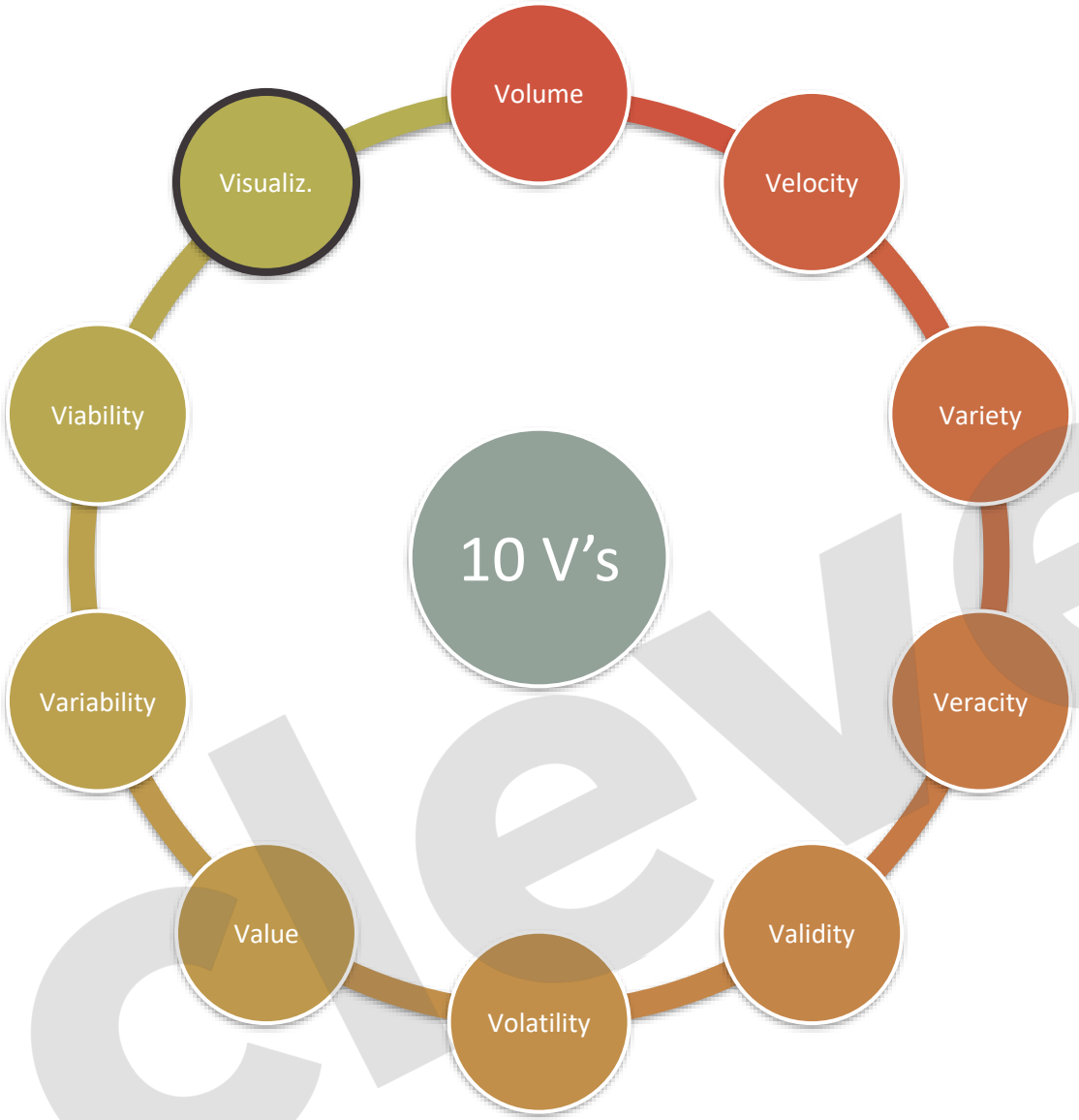
Examples:

- In restaurant reviews, maybe the 1-5 stars rating isn't as important as the review itself (or vice-versa)
- In NLP, the top 60-80% of the most frequent words often provide more value than the top 10-20% most frequent ones.

Related: a lot of data scientists agree that "around 5% of the relevant variables will produce 95% of the benefit".

DS & big data

how to classify as big data



visualization

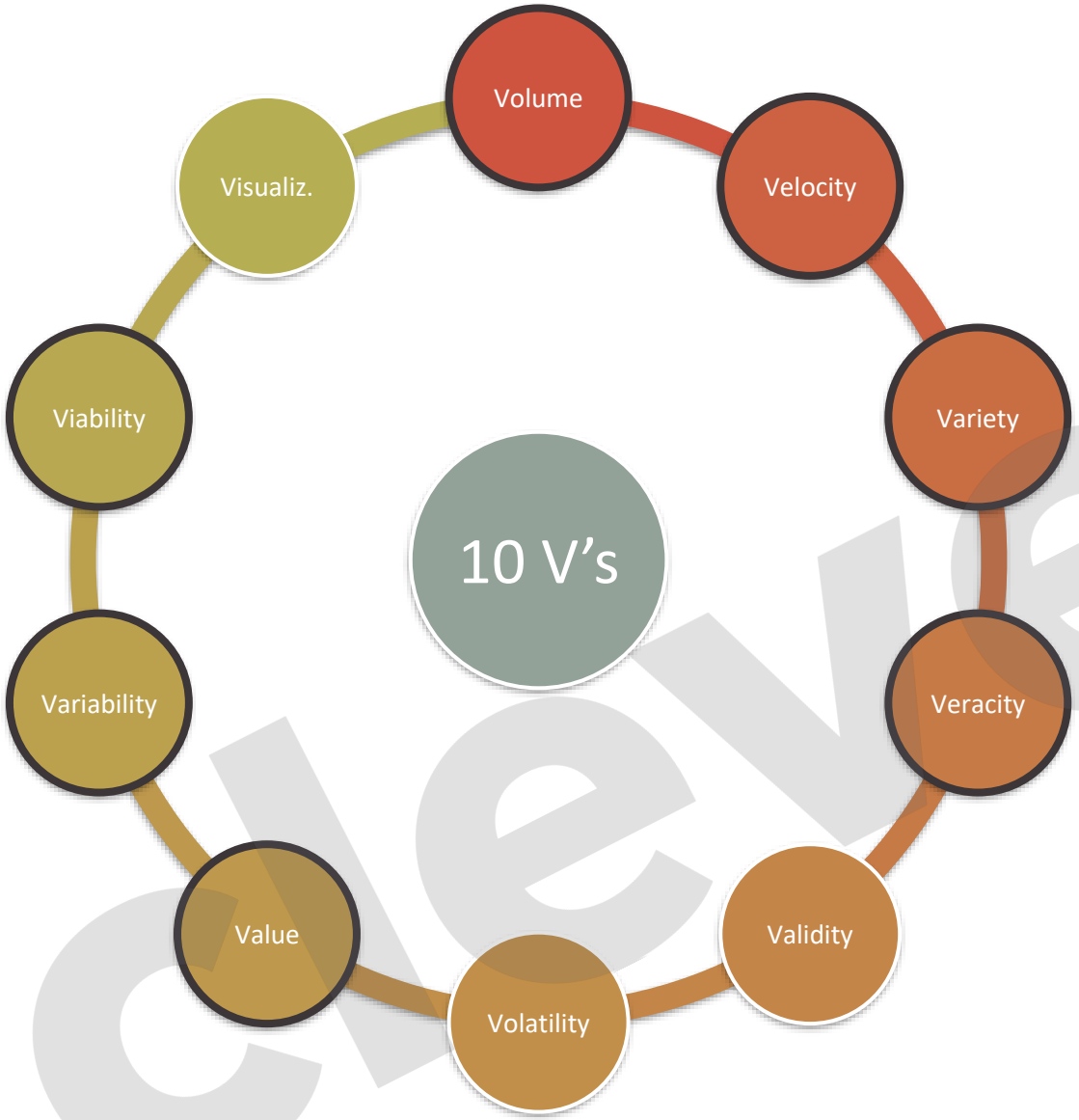
Visualization refers to the communication aspect of big data.

Visualization can help answer questions and uncover unseen trends in the data

If data is massive, ever-growing, raw, noisy, messy and in many forms maybe the only way of visualizing it is through big data.

DS & big data

how to classify as big data



volume

Is the data size massive?

velocity

How quickly do I need outputs?

variety

How many data sources do I have to work with?

veracity

Do I need to absolutely trust the data?

value

Is the data more valuable if processed quicker?

variability

Will I be working with variable data, like text, video or images?

viability

Is the data so complex that I'm struggling to uncover insights?

the world of data

DS real life examples

DS real life examples

wheather forecast

Landslides are very difficult to predict using the basic warning signs.

The Melbourne University developed an advanced tool which predicts whether a landslide is likely to occur two weeks in advance, and also the magnitude of the upcoming destruction.

This big data based tool has helped relocating people before the calamity happens, thus preventing human, animal and material losses.

DS real life examples

Starbucks uses big data for it's mobile app.

17 million users generate all sorts of data: purchases (type of coffee they like the most), geolocation (where to they buy it) and time (when do they buy it).

Based on this data, the app suggests new products, send personalized offers such as birthday discounts and even tries to guess what the person will order based on their history.

Others: Netflix, Spotify, Social Networks, Retail stores, etc.

DS real life examples

United Healthcare uses big data to detect medical fraud, identity threat and waste monitoring.

It predicts the likelihood of success in disease management programs, depending on the patient's reactions.

"We are one of the most diversified healthcare companies in the world, serving 85 million individuals worldwide. So we have data that touches every aspect of the healthcare industry: member, claims, hospital, provider, clinical, operational, financial etc., and at a scale that is unparalleled within the healthcare industry.
I would say that analytics is at the core of everything we do. It helps us understand our data better, get to root-causes of problems quickly, build innovative solutions and actively adapt to the changing healthcare landscape."

DS real life examples

logistics

UPS uses big data to optimize routes dynamically. The system, called ORION (On-road Integrated Optimization and Navigation) can automatically change routes in real time.

By doing so, UPS can anticipate traffic jams and bad weather conditions, and lead to a reduction of about 100 million delivery miles and 100,000 metric tons of carbon emissions.

Big data is also used to try to find answers:

"Should we have more access points? Should we introduce lockers? Should we allow drivers to release shipments without signatures? Data, technology, and analytics will improve our ability to answer those questions in individual locations - and those benefits can come from using the information we collect from our customers in a different way".

DS real life examples

travel / tourism

Airbnb says that big data has helped them design the best model to enable a better match between a host and a guest.

It juggles over 11 petabytes of data and considers the host's preferences on the duration of stay and if they want their place to be continually occupied or if they prefer to have breaks between guests.

In terms of guests, 4 categories are vital factors that influence a guest to select a venue:

- Behavioral Aspect: determined by how the user interacts with the Airbnb website.
- Dimensional Factor: device used, language and location preferred
- Sentiment: lodging reviews, survey results, and ratings are vital deciding factors
- Imputed: sorts the location preference of the traveler, for example, city vs. local towns

They are the creators of the now open source Apache Airflow, a monitor/scheduler or workflows.

DS real life examples

government / law enforcement

NYPD uses big data to protect its citizens.

It identifies crime trends, threats and prevent crimes by analysing data such as emails, social chatter, fingerprints, former police investigations and other public databases.

The system is based on tracking and stopping smaller crimes to stop bigger crimes, and to pinpoint hotspots where crimes are clustered.

Using sophisticated computer models and algorithms, it predicts places of expected criminal activity. The predictions are narrowly tailored, limited to 500-by-500-square foot areas, and usually updated on a daily basis. Police officers are then deployed to those identified areas to deter crime from happening.

It helped reduce crime in NYC by 75% and made it America's safest big city.

DS real life examples

insurance / banking

Is using driver behavioral analytics to create tailored insurance packages.

It offers the 'DriveSafe' app for drivers under 24. The app records journeys and shares journey data through an Internet-enabled app with AXA.

The data generated by the app sends AXA journey information showing your ability to drive safely: for example, keeping within the speed limit.

Using these scores, AXA can, then, offer insurance discounts for drivers.

DS real life examples

customer profiling

The company keeps the records of all the customer data such as what they purchased, whether they used any card, responses to the survey conducted, support issues, etc. in the guest ID. They also collected some additional details like customer's age, religion, educational background, marital status, no. of children, income and job details. Some other personal details like when you last moved, are you divorced, etc.

After analyzing this large amount of data, the insights gained by Target indicated to them the purchase pattern of pregnant women in the different phases of their pregnancy. They concluded that in the initial period of pregnancy, the women purchased various supplements for calcium, magnesium, and zinc. After some months they start purchasing oversized jeans, sanitizers, etc.

This information helped Target to provide personalized product suggestions to that specific group of customers.

DS real life examples

price optimization

Walmart has developed its own data analytics hub which is known as Data Cafe. The Data Cafe is fed up with more than 40 petabytes of customer data which helps them in understanding the market trend. This advanced analysis indicated the grocery team of Walmart that the sale of a particular item is suddenly reduced because of its irrelevant prices.

Thus, this algorithm alerts them whenever there is a sudden decrease in the sales of a product so that appropriate actions can be taken.

DS real life examples

The Ikea app allows customers to preview their products diretly at their homes. This helps them choose the best product for the available space, the best matching colors, etc.

This has contributed to a big boost in their online shopping platforms, since customers no longer need to physically see the product.

DS real life examples

NETFLIX

Aside from the well-known recommendation system, Netflix intelligently utilized the power of their data to run predictive analysis to learn what exactly their customers would be receptive and interested to watch.

By analyzing over 30 million plays a day as well as over 4 million subscriber ratings and 3 million searches, they were able to make winning bets on developing widely-acclaimed hits such as ‘House of Cards’ and ‘Arrested Development’.

DS real life examples

human resources

Google's people analytics teams dug deep into their data and analyzed employee performance reviews and feedback surveys amongst many data sources to better understand how to 'build a better boss'.

This helped to create a list of data-driven insights into what employees valued and helped to improve the manager quality 75% of their lowest-performing managers.

But that's not all, Google's use of analytics extended to making key decisions to enhance employee welfare – such as extending maternity leave to cut their new mother attrition rates in half.

DS real life examples

personalized marketing

Coca-Cola cleverly leverages the power of image recognition technology and data analytics to target users based on the photos they share socially – giving them insights into the individuals drinking their products, where they are from and how (and why) their brand is being mentioned.

The personalized ads served this way enjoyed a 4x greater click-through rate versus other methods of targeted advertising.

DS real life examples

personalized experiences

DBS has invested heavily into AI and data analytics to provide their customers with hyper-personalized insights and recommendations to allow customers to make better financial decisions.

This means providing intelligent banking capabilities that include:

- Offering investment proposals on financial products & instruments
- Stock recommendations based on an investor's portfolio
- Notifications on favorable FX rates
- Unusual transactions notifications

To ensure this evolution is effective and lasting, the bank trained over 16,000 employees in big data and data analytics to truly transform the company into a data-driven organization.

DS real life examples

Uber

With predictive analytics, Uber is able to analyze historical data and key metrics that include the number of ride requests and trips getting fulfilled in different parts of a city as well as the time and day where this is happening.

This analysis helps the company to gain insight into areas that have a supply crunch, allowing them to pre-emptively inform drivers to move to areas ahead of time in order to capitalize on the inevitable rise in demand.

DS real life examples

fraud detection

IRS has modernized its fraud-detection protocols in the digital age. To the dismay of privacy advocates, the agency has improved efficiency by constructing multidimensional taxpayer profiles from public social media data, assorted metadata, emailing analysis, electronic payment patterns and more.

Based on those profiles, the agency forecasts individual tax returns; anyone with wildly different real and forecasted returns gets flagged for auditing.

DS real life examples

customer segmentation

When singles match on Tinder, they can thank the company's data scientists. A carefully-crafted algorithm works behind the scenes, boosting the probability of matches. Once upon a time, this algorithm relied on users' Elo scores, essentially an attractiveness ranking.

Now, it prioritizes matches between active users, users near each other and users who seem like each other's "types" based on their swiping history.

applied data science

essential DS packages

essential DS packages

pandas

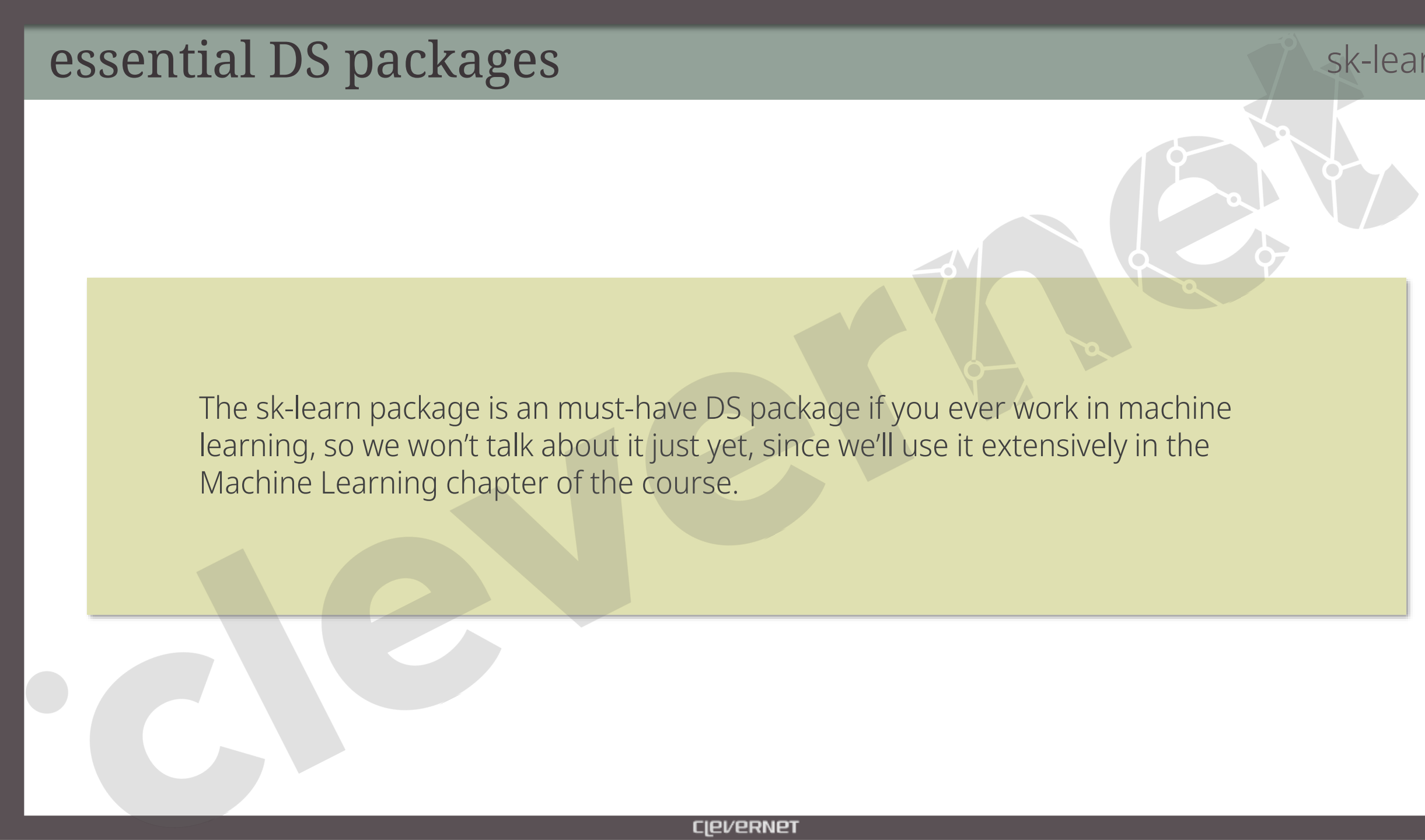
essential DS packages

numpy

essential DS packages

sk-learn

essential DS packages



The sk-learn package is an must-have DS package if you ever work in machine learning, so we won't talk about it just yet, since we'll use it extensively in the Machine Learning chapter of the course.

essential DS packages

visualization

essential DS packages

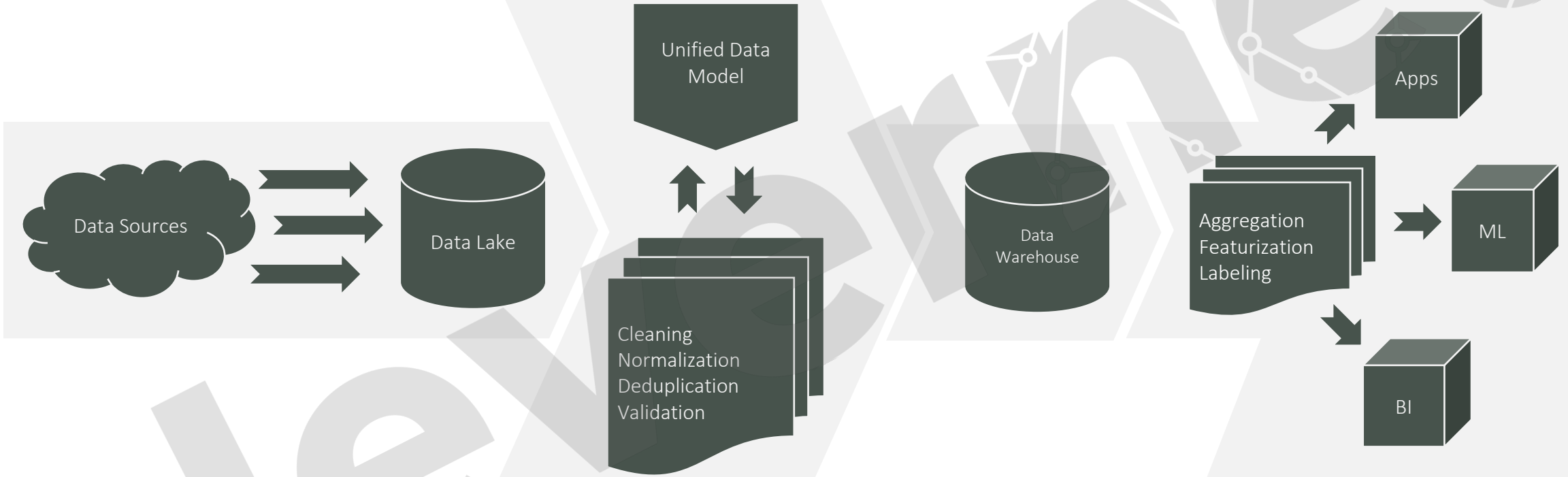
what's EDA & to automate it

applied data science

data engineering

data engineering

the data pipeline



ETL process

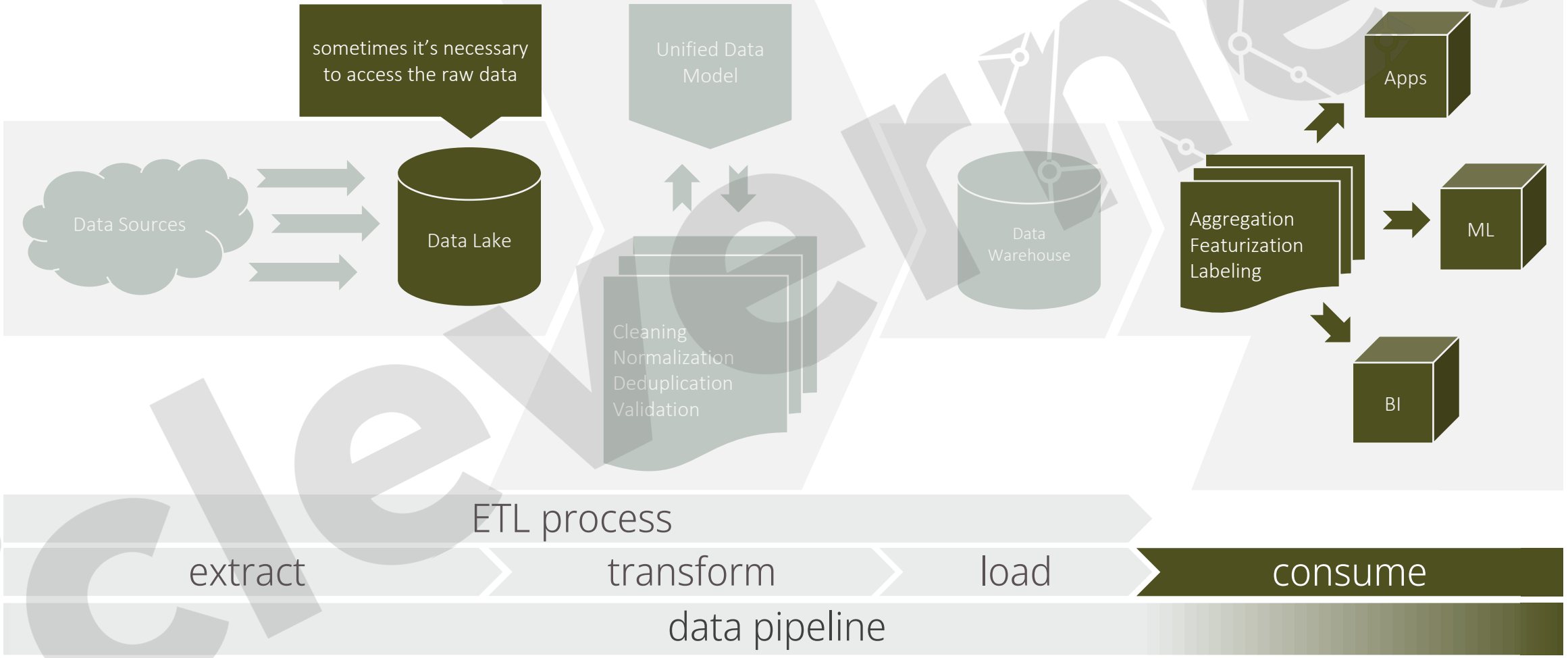
extract | transform | load | consume

data pipeline

data engineering

the data pipeline

in a perfect world, as a data scientist these would be your touch points:



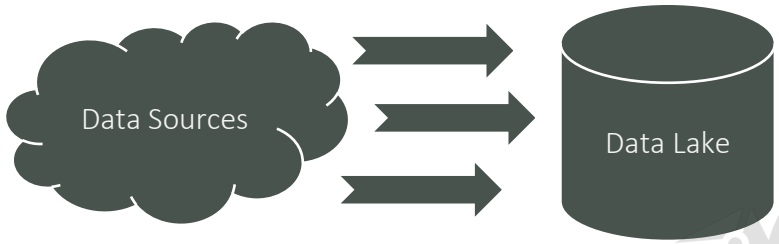
Sadly, the reality is that either:

- you're self employed and have no budget for a data engineer
- the data engineers at work have no time for you (there's always some super complicated support ticket that they need to address)
- "Oh, I'm sure that you can do it by yourself!", says your paternalistic manager thinking you're already some form of an artificial intelligence
- you're attending this python training and the instructor thinks you'll benefit greatly from learning the basics

data engineering

the ETL process

the ETL process



list datasources	databases, flat files, excel files, pdf's, web scraping, API requests, media, etc
build connectors	get or build the necessary connectors to tap into the data
extract to data lake	extract the data as is into a temporary db. If small volume, memory can be used

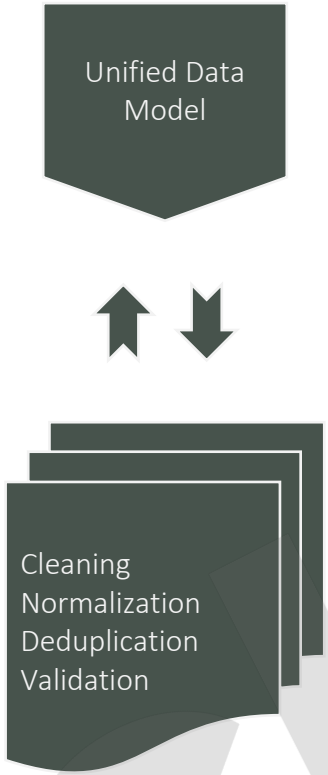
---

Moving on to the sweet stuff!

Open the project folder on jupyter lab and start by opening the `0-briefing.ipynb` file to get acquainted with what you're going to build.

the ETL process

transform



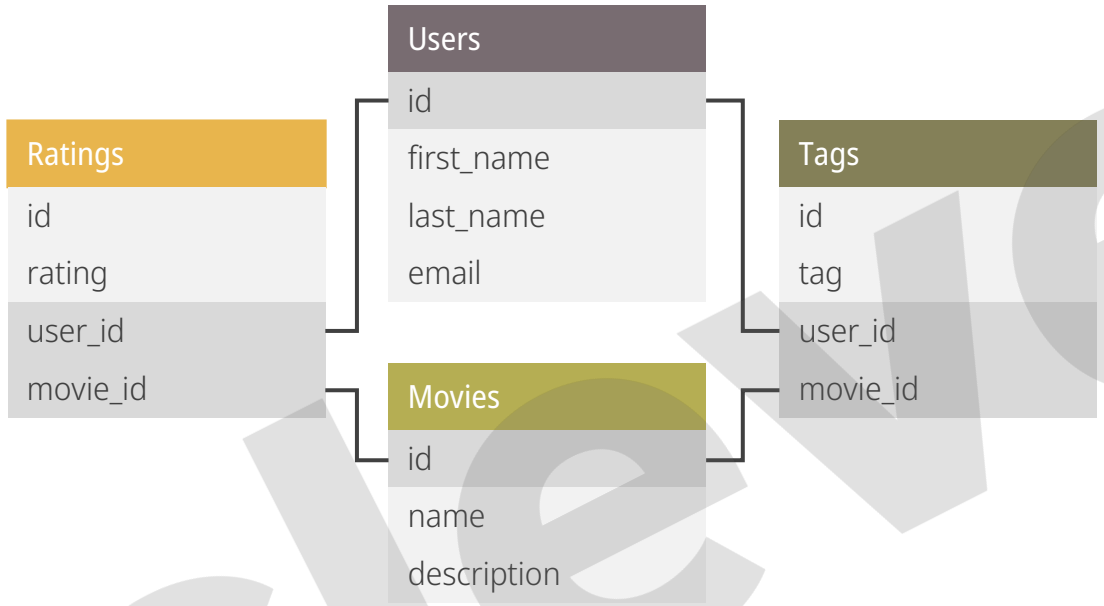
inspect & understand	know your data
find relationships	find ways to join the different sources together
join data sources	construct the joining task
clean	remove duplicates, fill missing values, rename/remove columns, ...
normalize	format dates, convert datatypes, concatenate values, ...
deduplicate	replace the duplicates with references
validate	apply data quality principles: correct addresses, postal codes, ...

the ETL process

transform - unified data model

types of a unified data model:

relational database



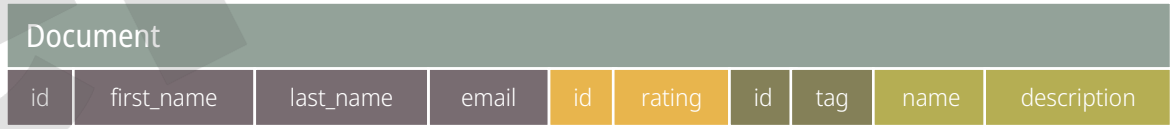
easier to build

scales vertically

less duplicated data

better data integrity

document database



scales horizontally

no need for later joins

easier data faceting

better search capabilities

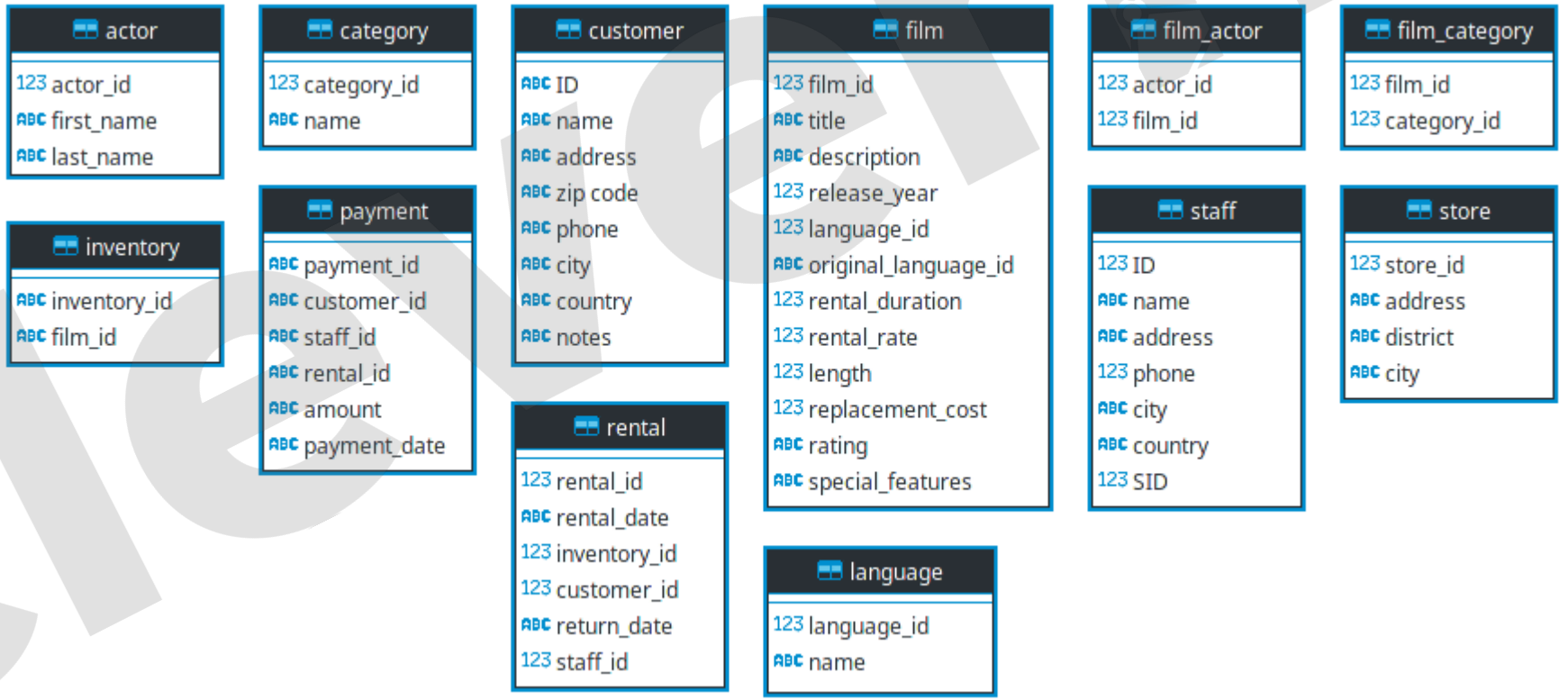
drill-down ready

the ETL process

transform - unified data model

exercise:

inspect & understand	browse the database and start to know your data
find relationships	rearrange the tables below to find a relationship flow
~~join data sources~~	ignore this for now

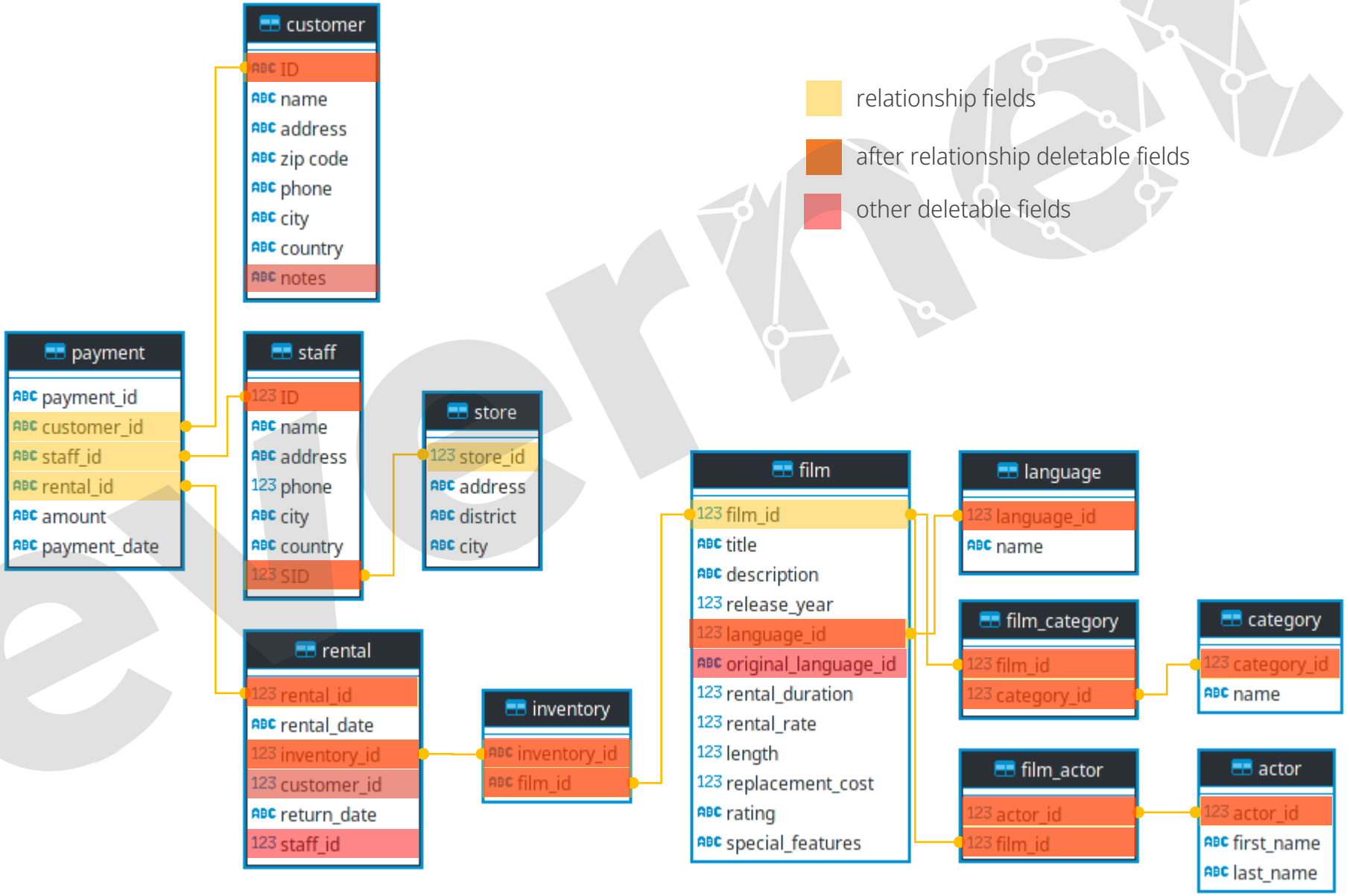


the ETL process

transform - unified data model

exercise:

- ~~inspect & understand~~
- find relationships
- ~~join data sources~~



the ETL process

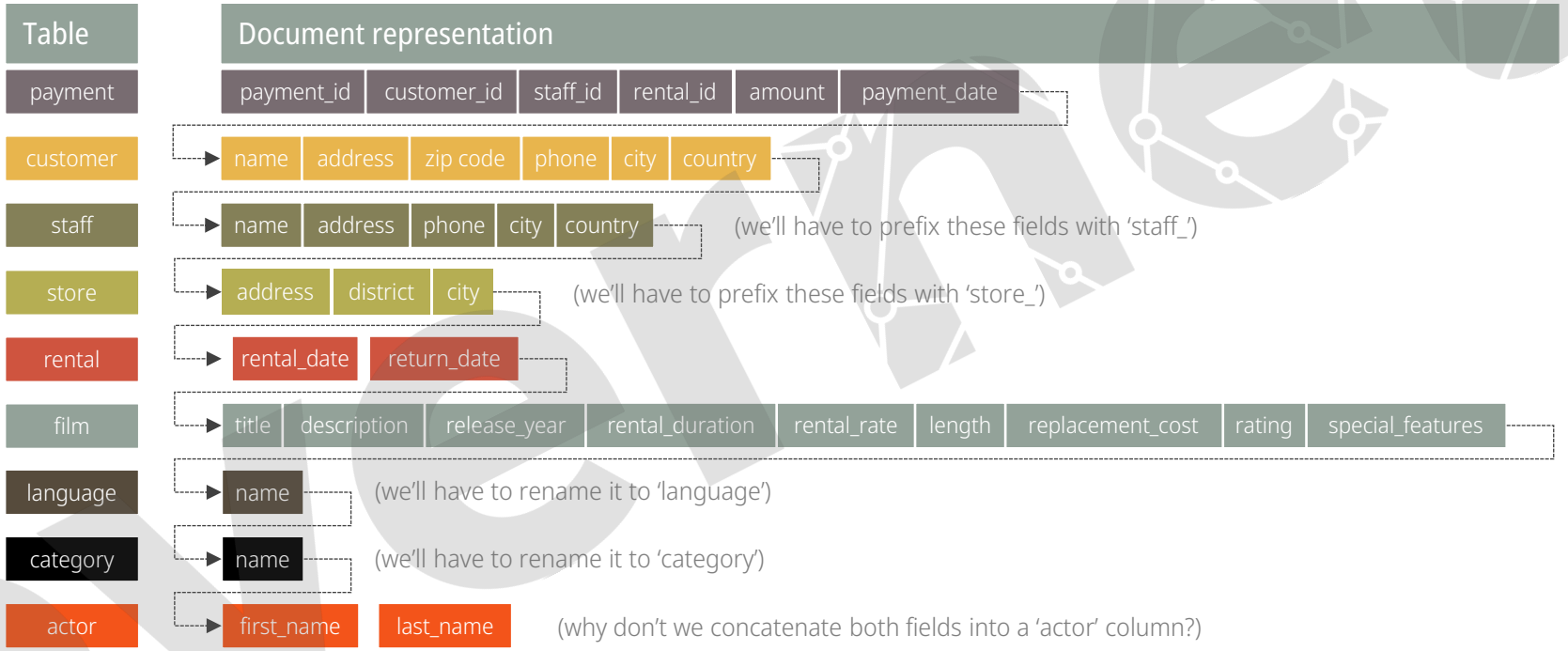
transform - unified data model

exercise:

~~inspect & understand~~

~~find relationships~~

join data sources



Moving on to the sweet stuff!

Open the project folder on jupyter lab and open the `2-transform-model.ipynb` file.

the ETL process

load

Now that we've finished transforming our data, it's time to export it to the data warehouse and make it available to everyone else: analysts, business intelligence users, data scientists, ...

We won't be using a real data warehouse, but we'll use something that represents what a data warehouse might look like for our project.

Moving on to the sweet stuff!

Open the project folder on jupyter lab and start by opening the `3-load.ipynb` file to get acquainted with what you're going to build.

applied data science

machine learning

machine learning

what is

- Supervised machine learning algorithms that are most often used in classification tasks
- Known to be fast learners, solve with ease problems in real-time, as well as handle sparse data
- Preferred when faced with tasks involving text analysis, such as spam filtering, categorizing articles, sentiment analysis, etc.
- Also suitable for recommendation systems
- Through the Bayesian approach we estimate the probability a certain hypothesis is true given past data (evidence)

types of classifiers

- Gaussian - most suitable when features are continuous (have a normal distribution)
- Bernoulli - assumes binary-valued features (true/false, yes/no)
- Multinomial - most suitable when features are discrete (relies on the frequency); best suited for text classification
- Complement - as above, but for imbalanced datasets

spam or ham example

- Consider a training dataset consisting of 20 ham and 20 spam messages.
- This is the word count in those messages:

Word	Count in ham messages	Count in spam messages
dear	5	3
deadline	3	1
lecture	7	10
notes	9	9
assignment	6	7
student	15	0
Total	**45**	**30**

spam or ham example

- According to the table, the word "student" has not appeared in any of the 20 spam messages.
- Therefore, even if an incoming message that contains the word *student* is spam, it will not be considered as such by the model, as the conditional probability of the message being spam would always be 0.
- The remedy is to introduce the so-called smoothing parameter; let's set its value equal to 1 (Laplace smoothing).
- The purpose of this parameter is to increase the count of each word so that the new counts become the ones in the table to the right.

Word	Count in ham messages	Count in spam messages
dear	6	4
deadline	4	2
lecture	8	11
notes	10	10
assignment	7	8
student	16	1
Total	**51**	**36**

spam or ham example

- The marginal probabilities of an email being ham or spam are 50/50, or:

$$P(\text{ham}) = P(\text{spam}) = \frac{1}{2}$$

- Let's calculate the probability of a message with the words *dear*, *deadline* and *student* being spam or ham:

$$P(\text{ham}|\text{dear, deadline, student}) \propto P(\text{dear, deadline, student}|\text{ham})P(\text{ham})$$
$$P(\text{spam}|\text{dear, deadline, student}) \propto P(\text{dear, deadline, student}|\text{spam})P(\text{spam})$$

- Or:

$$P(\text{ham}|\text{dear, deadline, student}) \propto \frac{6}{51} \text{ x } \frac{4}{51} \text{ x } \frac{16}{51} \text{ x } \frac{1}{2} \approx 1.4 \times 10^{-3}$$

$$P(\text{spam}|\text{dear, deadline, student}) \propto \frac{4}{36} \text{ x } \frac{2}{36} \text{ x } \frac{1}{36} \text{ x } \frac{1}{2} \approx 8.6 \times 10^{-5}$$

- To substitute the proportionality sign with an equal sign, we need to further divide by P(dear, deadline, student). The way we calculate this quantity is as follows:

$$P(\text{dear, deadline, student}) = P(\text{dear, deadline, student}|\text{ham})P(\text{ham}) + P(\text{dear, deadline, student}|\text{spam})P(\text{spam})$$

spam or ham example

- Therefore, we need to add the two results we have just obtained:

$P(\text{dear, deadline, student}) \approx 1.4 \times 10^{-3} + 8.6 \times 10^{-5} = 1.486 \times 10^{-3}$

- Finally, the conditional probabilities for the message to belong to the ham or spam classes is the following:

$$P(\text{ham}|\text{dear, deadline, student}) \approx \frac{1.4\ x\ 10{-}3}{1.486\ x\ 10{-}3} \approx 94\%$$

$$P(\text{spam}|\text{dear, deadline, student}) \approx \frac{8.6\ x\ 10{-}5}{1.486\ x\ 10{-}3} \approx 5.7\%$$

- If the two results are now rounded, their sum would indeed equal 100%. In this way, not only did we learn which class the message belongs to, but we also managed to calculate the probabilities of the message belonging to either class.

Let's Play?

1. Let's open *06-0-Naive-Bayes-YouTube.ipynb* to get ready for step 2

2. Try to solve the exercise: *06-1-Naive-Bayes-Exercise.ipynb*

3. More indepth examples: *06-3-Naive-Bayes-Indepth.ipynb*

machine learning

regression analysis

what is

- One of the most common methods of prediction
- Used when we have a causal relationship between variables (cause -> effect)
- Used in supervised machine learning
- The process:
  - Get sample data
  - Design a model
  - Make predictions on the population

dependent variable (predicted) — independent variable (predictors)

$$Y = F(x_1, x_2, \ldots, x_k)$$

Y is a function of the independent variables

linear regression

- Linear approximation of a causal relationship between two or more variables

simple linear regression equation

$$y = b_0 + b_1 x_1$$

- b_0: constant
- b_1: quantifies the effect of the independent (x) in the dependent (y)
- y: estimated / predicted value
- x_1: independent variable

example

$$y = b_0 + b_1 x_1$$

- b_0: minimum wage
- b_1: 100€ for each year of education
- y: predicted salary
- x_1: years of education

linear regression - geometrical representation

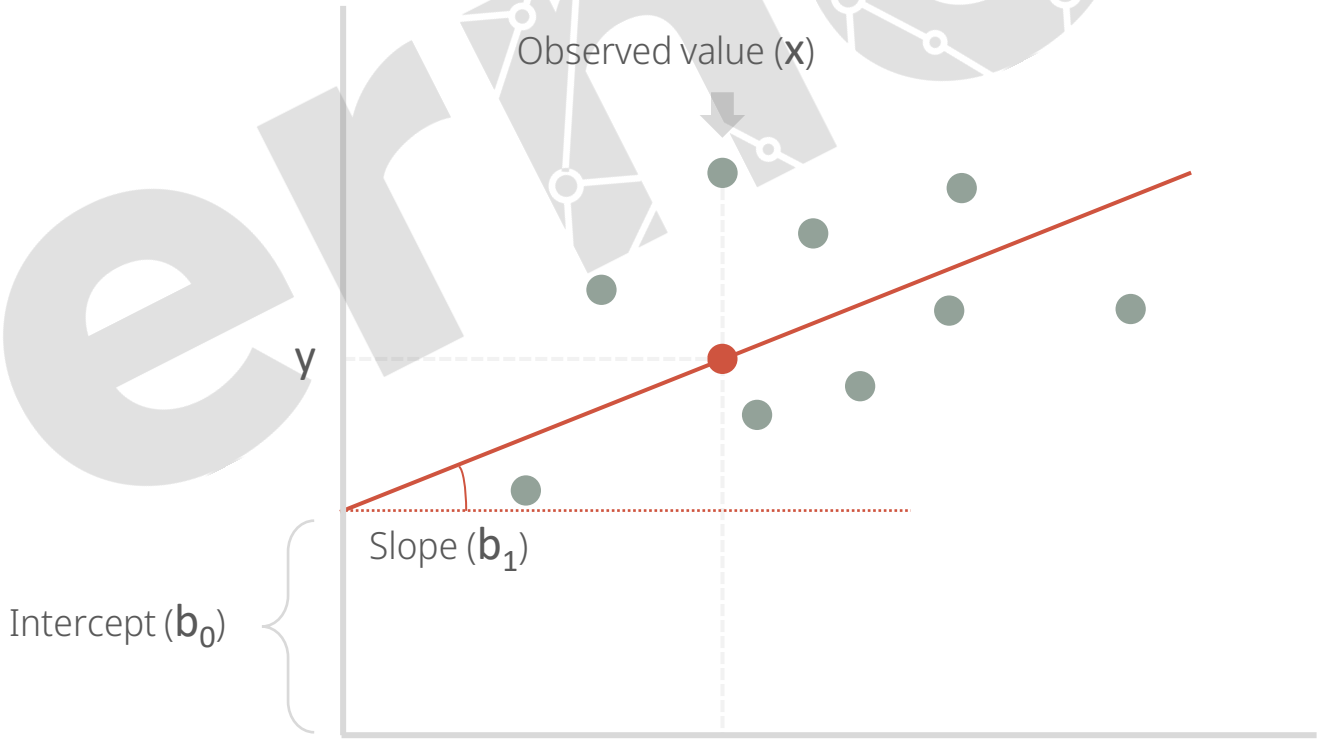
quantifies the effect of the independent (x) in the dependent (y)

constant

$$y = b_0 + b_1 x_1$$

estimated / predicted value

independent variable



correlation vs regression

"correlation does not imply causation"

correlation	regression
relationship	one variable affects the other
move together	cause and effect
symmetrical: $p(x,y) = p(y,x)$	one way only
single point in a chart	a line in a chart

Let's Play?

1. Let's open *07-0-Simple-Linear-Regression.ipynb* to get ready for step 2

2. Try to solve the exercise: *07-01-Simple-Linear-Regression-Exercise.ipynb*

3. More indepth examples: *07-03-Linear-Regression-Indepth.ipynb*

4. A full practical example: *07-04-Linear-Regression-Practical-Example.ipynb*

ridge regression

- Similar to Linear Regression but includes a tool called penalty term, that prevents issues like overfitting and multicollinearity.

overfitting and multicollinearity

- Overfitting happens when the models capture too much noise, due to insignificant data or data that doesn't contribute to the learning process.
- Multicollinearity is often the cause of overfitting. It occurs when the independent variables are too correlated.
- The solution is to apply regularization techiques.

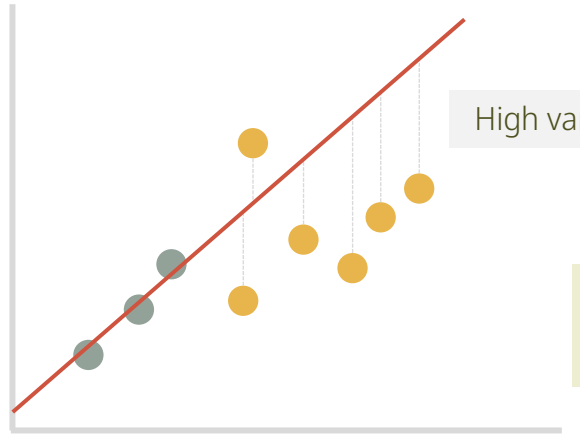
regularization techniques

- Regularization prevents overfitting by including additional information.
- The goal is to make the model simpler to avoid fixating on insignificant data.
- Ridge regression uses the L-2 regularization technique.
- L-2 is a hyperparameter between 0 and infinity, and the way to get the best L-2 value is through cross-validation.

example

- Imagine the following trained linear regression:

- Now what happens when we use test data:

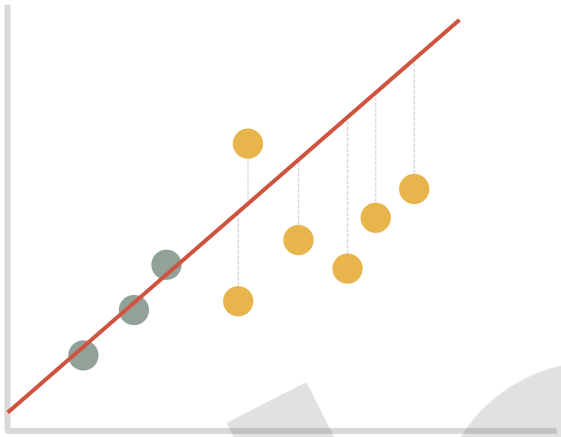


- It failed to make accurate predictions

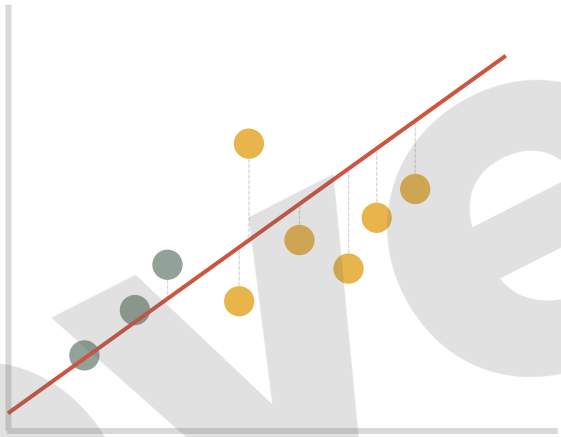
example (cont.)

- How can we fix this? By using a ridge regression, which will increase the bias in order to decrease the variability:

Before



After



- Note that the method will improve the accuracy in the prediction phase, but will increase the error in the training phase.

lasso regression

- Very similar to ridge regression, but instead of using the L-2 regularization technique, it uses the L-1.
- This technique might lead to automatic feature selection, since it can eventually equal some features to zero, thus removing them from the model.
- The L-2 used in ridge also approximates some less important features from zero, but they'll never be zero, so no features are ever removed.
- Therefore, lasso regressions are suitable for big datasets with many features, since it can remove insignificant features, thus reducing variance.

Let's Play?

1. Let's open *07-05-Ridge-Lasso-Regressions-Practical-Example.ipynb* for a full practical example comparing linear, ridge and lasso regressions trying to predict baseball players salaries.

k-nearest neighbors (KNN's)

- It's an intuitive, interpretable, easy to implement and widely used classification and regression algorithm
- The principle behind it is to categorize a data point based on the samples that are closest to it (the nearest neighbors)
- It's very good in identifying patterns, so it's also good to identify samples that fall outside of those patterns, such as outliers

Let's Play?

1. Let's open *07-06-KNNClassifier-Practical-Example.ipynb* for a practical example
2. Let's open *07-07-KNNRegressor-Practical-Example.ipynb* for a practical example

3. A comparison between KNN and Linear Regression on linear data: *07-08-KNNRegressor-Linear-Regression-Linear.ipynb*
4. A comparison between KNN and Linear Regression on non-linear data: *07-09-KNNRegressor-Linear-Regression-Non-Linear.ipynb*

5. Try to solve the exercise: *07-10-KNNRegressor-Exercise.ipynb*

machine learning

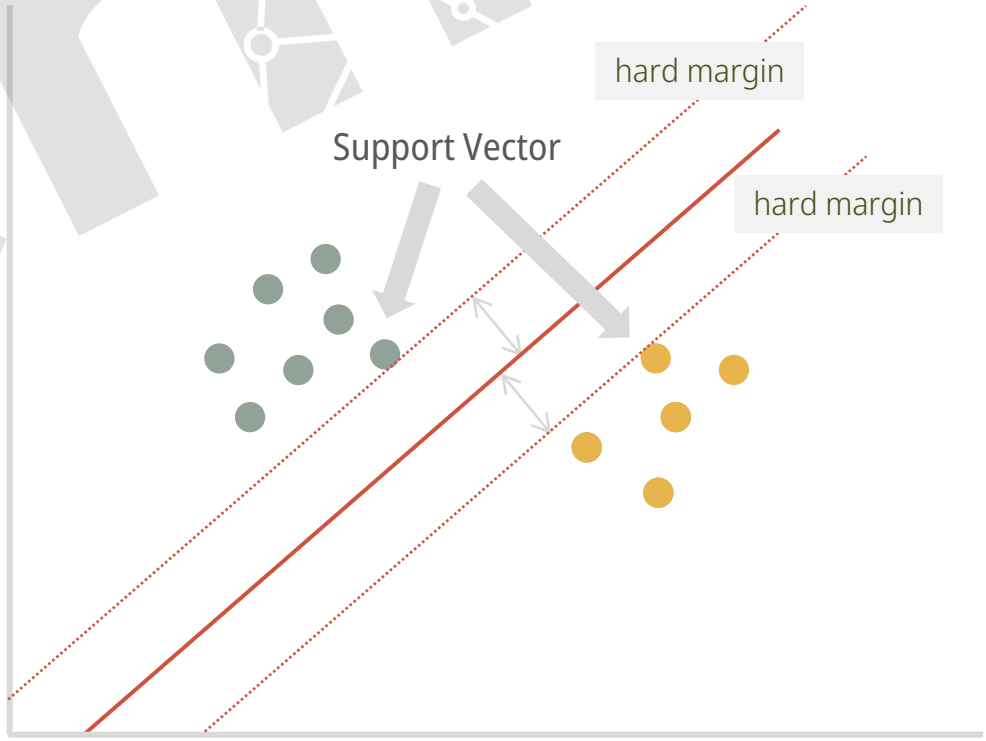
support vector machines

what is

- Supervised machine learning algorithm used for either classification or regression
- For classification, SVM's create an hyperplane that maximizes the differences between the classes
- For regression, an hyperplane is also used, but this time it isn't looking for an exact match, allowing some points to be misclassified in order to reach a better overall solution, without overfitting the data
- Can solve non linear problems with the *kernel trick*
- It's difficult to interpret
- Training data can take a long time

hard margin problem

- How to know where the hyperplane is located?
- We can achieve that by calculating the hard margin: two lines that touch the boundaries of each class, for example:



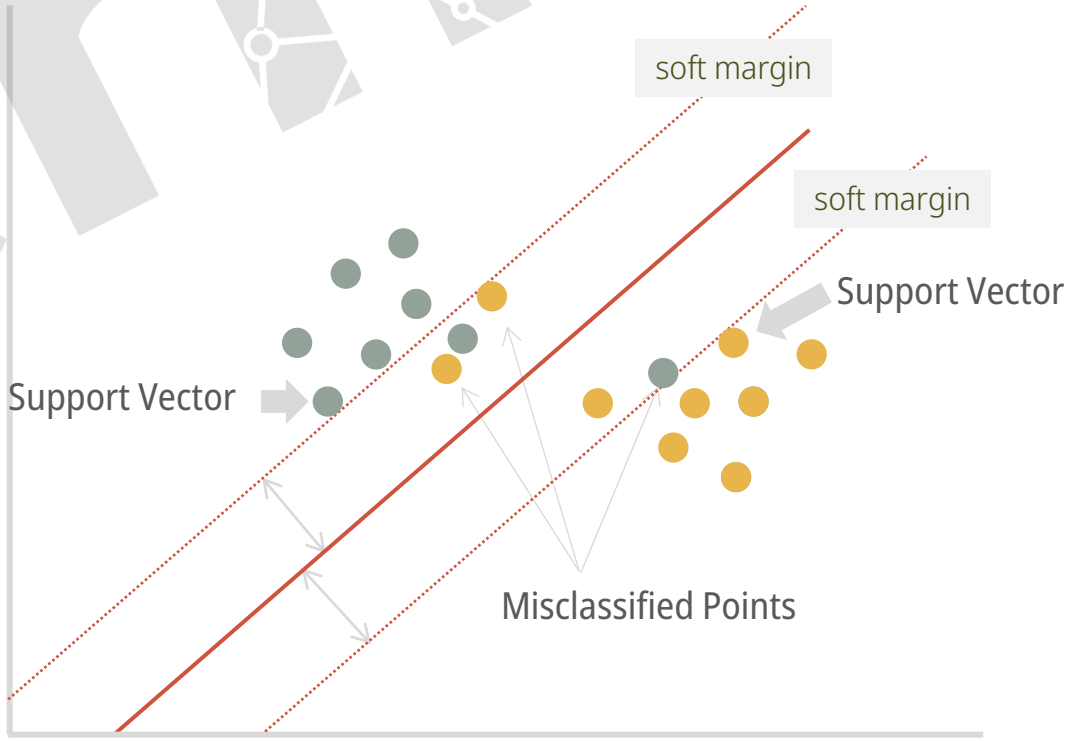
soft margin problem

- What if the classes aren't totally linearly separable, like the plot below?

- In that case, we need to apply the soft margin problem, which will misclassify some points in order to separate the classes by increasing the margin.
- That increase value is called C, and it's not simple to get. The best way to decide the best value is by using cross-validation.

soft margin problem

- Here's an example of the soft margin applied to the previous plot:



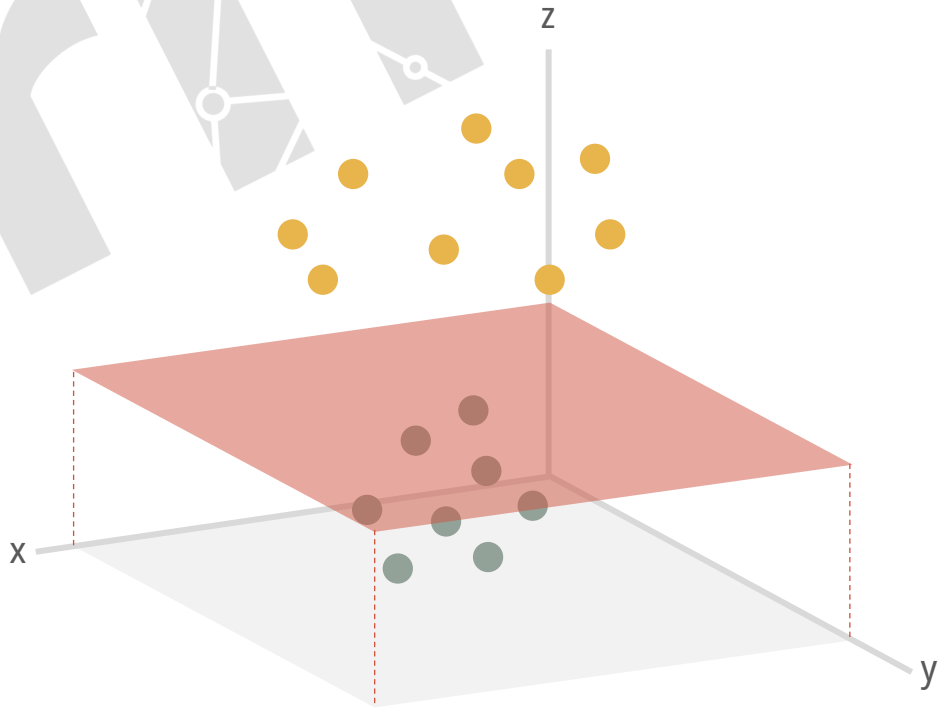
the kernel trick

- What if the classes aren't linearly separable at all, like the plot below?

- Well, they aren't linearly separable on a two dimension plot, but what if we turn the plot into 3D?

the kernel trick

- That way we could have one class above the other, and draw the hyperplane between them, like the image on the side:



Let's Play?

1. Let's open *08-0-Support-Vector-Machines-Practical-Example.ipynb* for an example.

2. Let's open *08-1-Support-Vector-Machines-Ploting-Kernels.ipynb* for an additional example on kernel ploting.

3. More indepth examples: *08-2-Support-Vector-Machines-Indepth.ipynb*

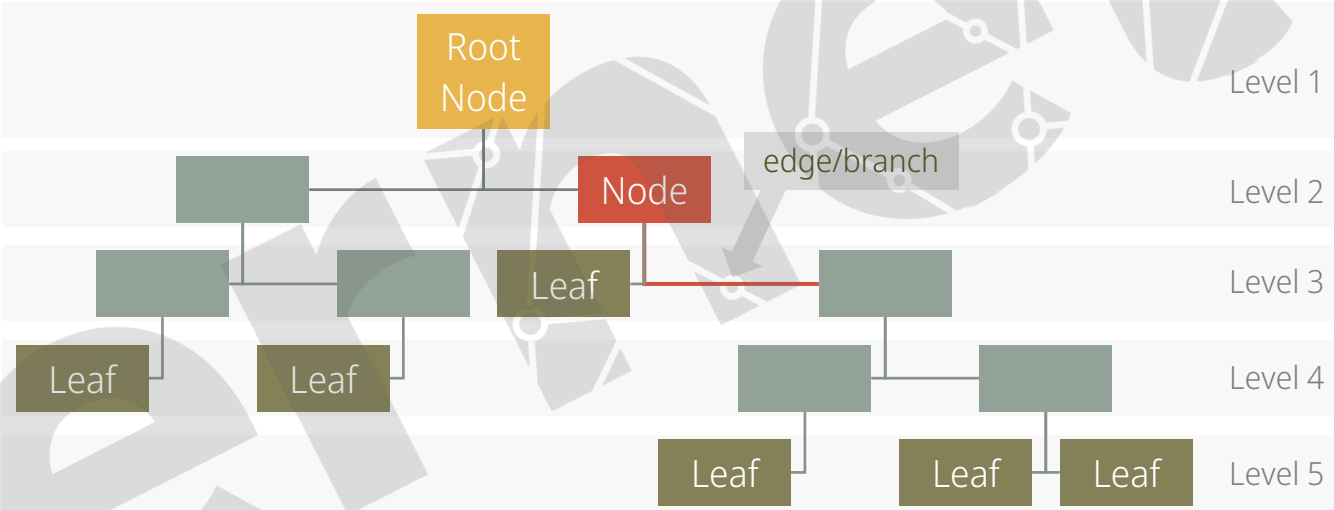
machine learning

decision trees

what is

- Supervised learning algorithm, suitable for both classification and regression
- Are the building blocks of random forests
- There are different types: ID3, C4.5, CART, CHAID or MARS
- Works with both numerical or categorical features
- Pros:
  - simple to understand and interpret
  - built-in feature selection
  - require little to no data preprocessing (no scaling or standardizing needed)
  - fast on large datasets (especially in the predicting phase)
- Cons:
  - Overfitting (can be minimized with *pruning*)
  - As with other ML algorithms, it struggles with unbalanced datasets

what is a tree?



- Edges can only connect nodes from two adjacent levels
- A node can only be connected to another node from a upper level
- If a tree never has more than two nodes per level, is called a binary tree (like the one above)
- The number of levels determine the tree's height
- Nodes represent questions, edges answers and leaves outcomes

what is pruning

- Method to simplify the outcome of a tree
- Removes the sub-trees that are not necessary for the classification, thus reducing the complexity of the classifier
- Improves predictive accuracy, thus diminishing overfitting
- Depends on a hyperparameter called *ccp_alpha*, that will define the pruning strength
    - Reasonable values are usually 0.1, 0.01 and 0.001

Let's Play?

1. Let's open *09-0-Decision-Trees-Practical-Example.ipynb* for an example.

2. In the same notebook, play with the pruning value and try to spot what changes in the decision tree.

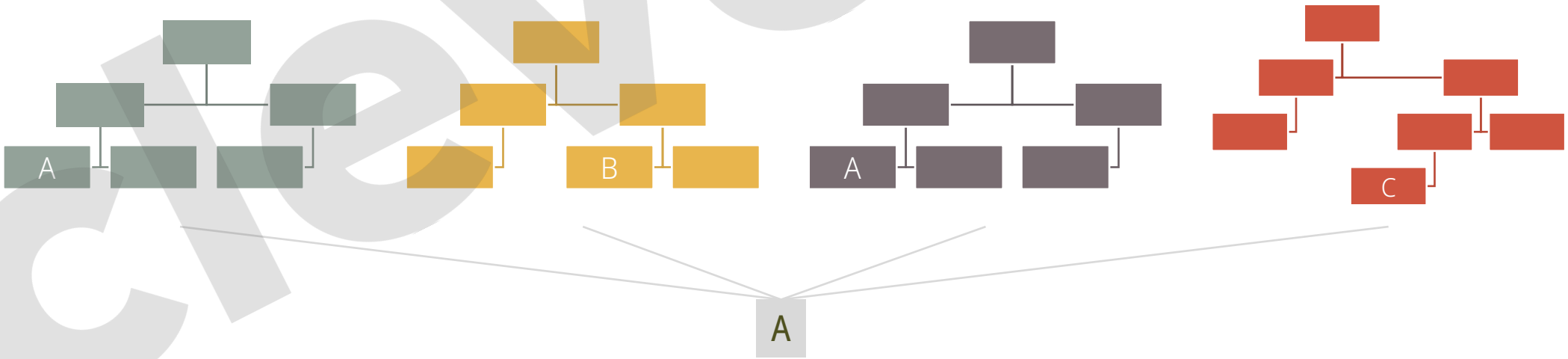
machine learning

what is

- Supervised learning algorithm, suitable for both classification and regression
- A collection of many individual decision trees
- Gives very good accuracy for a non neural network model
- More performant than decision trees
- More difficult to interpret than decision trees

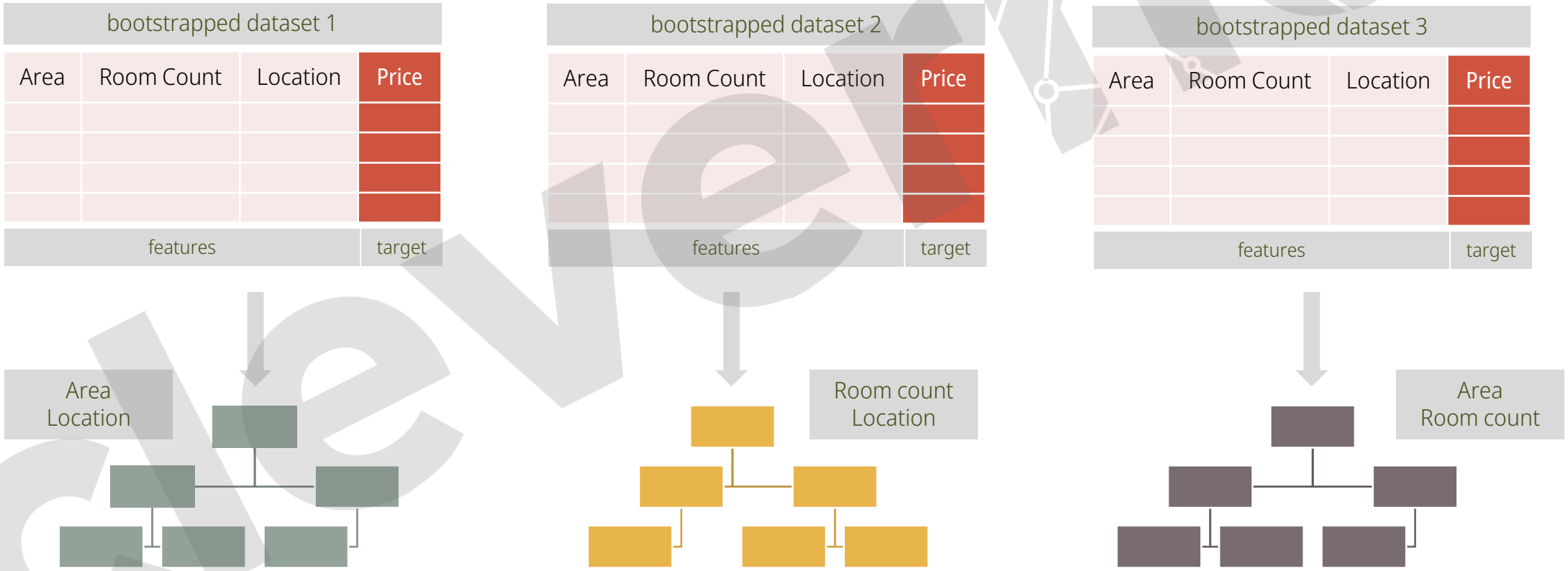
how do they work

- Given a number of trained trees, the most common outcome between them will be the output of the random forest



bootstrapping

- Given a training set, bootstrapping aims to create many slightly different datasets of the same size that follow the same distribution of the original
- Each of these will be used by a decision tree, but each decision tree will consider different features, as the example below:



Let's Play?

1. Let's open *09-1-Decision-Trees-Random-Forests-Practical-Example.ipynb* for an example.

2. More indepth examples: *09-2-Random-Forests-Indepth.ipynb*

machine learning

xgboost

what is

- Supervised learning algorithm, suitable for both classification and regression
- One of the most powerful machine learning algorithms
- Operates on decision trees
- A common technique in ensemble learning (use many models to make predictions together

Decision Trees	Bagging	Random Forest	Boosting	Gradient Boosting	XGBoost
A graphical representation of possible solutions to a decision based on certain conditions	Emsemble learning meta-algorithm combining predictions from multiple decision trees through a majority vote mechanism	Bagging based algorithm where only a subset of features are selected at random to build a forest or collection of decision trees	Models are built sequentially by minimizing the errors from previous models while increasing (boosting) the influence of high-performing models	Employs gradient descent algorithm to minimize errors in sequential models	Optimized gradient boosting thorugh parallel processing, tree pruning, handling missing values and regularization to avoid overfitting / bias

Let's Play?

1. Let's open *11-XGBoost.ipynb* for an indepth explanation and two examples.

what is

- Statistical technique that groups observations based on their features or variables
- Unsupervised learning technique:
  - we don't have train data to match the predictions
  - we don't know if the number of clusters is correct
  - the output we get is something that we have to name ourselves
- The goal is to maximize similarity of observations within each cluster and maximize dissimilarity between clusters
- Good starting point to explore data and identify patterns
- We'll be using the *K-means* clustering technique

Let's Play?

1. Let's see some basic pratical examples:

   - 12-0-K-Means.ipynb
   - 12-1-K-Means.ipynb
   - 12-2-K-Means-Optimal-Number-Clusters.ipynb
   - 12-3-K-Means-Market-Segmentation.ipynb

2. Do an exercise:

   - 12-4-K-Means-Exercise-Part1.ipynb
   - 12-4-K-Means-Exercise-Part2.ipynb

3. More indepth examples:

   - 12-5-K-Means-Indepth.ipynb

rferreira@clevernet.pt
967 270 033