Questions & Answers

Introduction to Data Science (JBP010)

18th of December 2017

R.J. Klaasse Bos

Hi there 👋!

To help you get most out of this slide deck, please take note of the following structure:

- **Part 1:** [Question + Answer] on a single slide
- **Part 2:** Question slide (without answer) followed by [Question + Answer]

The latter may be helpful to test your memorisation super powers..

Happy memorising! ;-)

And of course… 

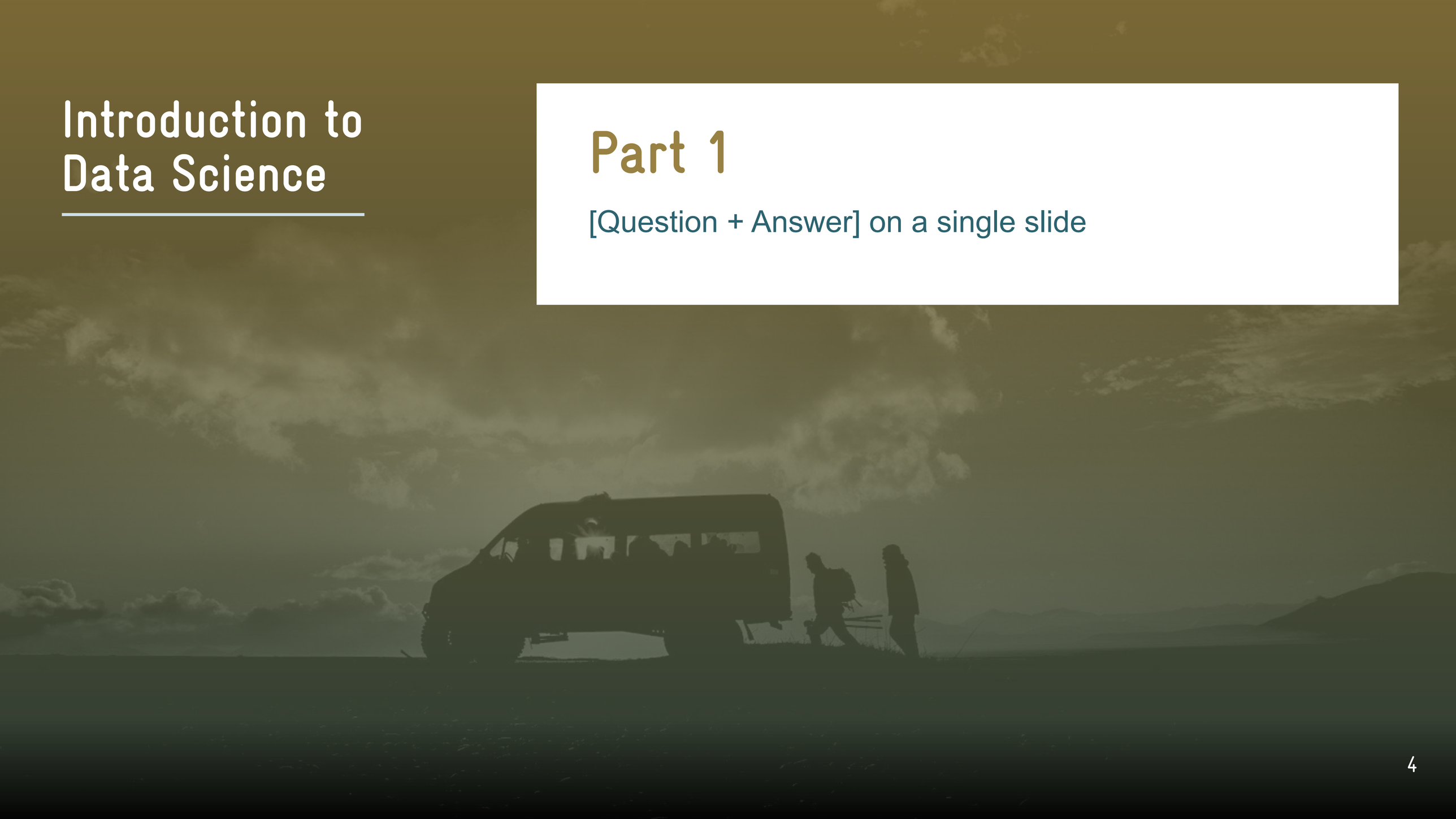
*"A **merry Christmas** and the **best wishes** for the new year!"*

And now back to why you are really here..

Introduction to Data Science

Part 1

[Question + Answer] on a single slide



Question 1:
What is Data Science?

A set of **fundamental principles** that guide the **extraction** of **knowledge** from data

Question 2:
What is Data Engineering?

A set of **technologies** that facilitates the **access** to **data** and its **processing**

in order to serve Data Science.

Question 3:
What is Big Data?

Data processing tools that handle **Volume**, **Velocity** and **Variety**

that traditional data processing systems cannot support.

Question 4:
What is Data Mining?

Tools and **processes** that guide the **extraction** of **knowledge** via Data Science principles.

Question 5:
What is **date-driven decision making?**

The practice of **basing decisions** on the **analysis** of **data** rather than purely on **intuition.**

Question 6:

What organisations do for creating a **data-driven culture**?

They..

acquire *(focus on keeping data clean),*

process *(democratise data access)*,

and **leverage** *(use data in decision-making and for product creation)*

..data in order to navigate the competitive landscape.

Question 7:

What “acquire” means in data-driven organizations?

- Focus on **keeping** the **data clean** (typically 80% of the work) both on a small and **large** scale
- Storing data with a **predefined purpose**

Question 8:

What “process” means in data-driven organizations?

Democratize access to **data** to encourage **experiments** and **hypotheses testing**

with the aim of **improving** an **organization** and its **processes.**

Question 9:

How do Data Scientists **spend most** of their **time**?

80% of the work consists of **cleaning** the **data**

Question 10:
What is a Data Scientist?

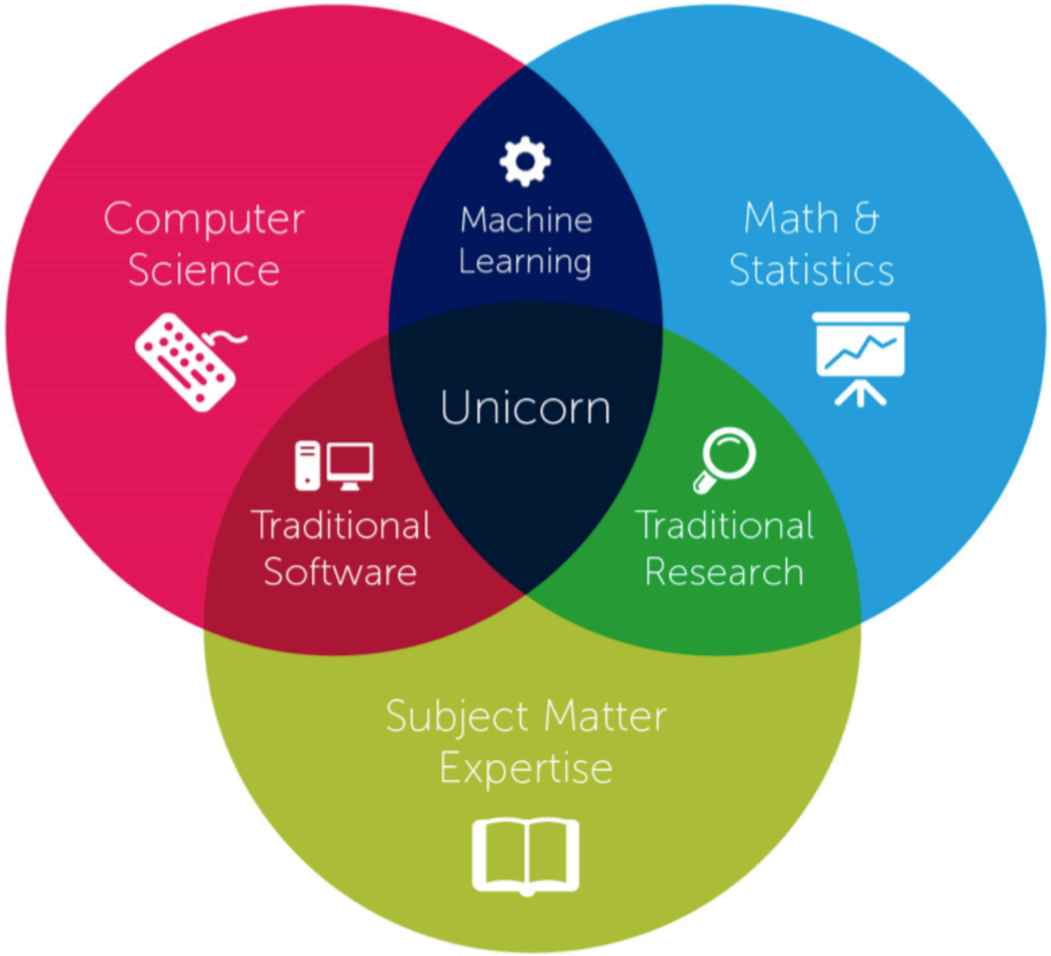
A **computer scientist**, **mathematician** and **storyteller** combined with a **curious attitude, domain knowledge** and **expertise.**

Question 11:
What is a data science unicorn?

Someone who masters the field of **Computer Science,**

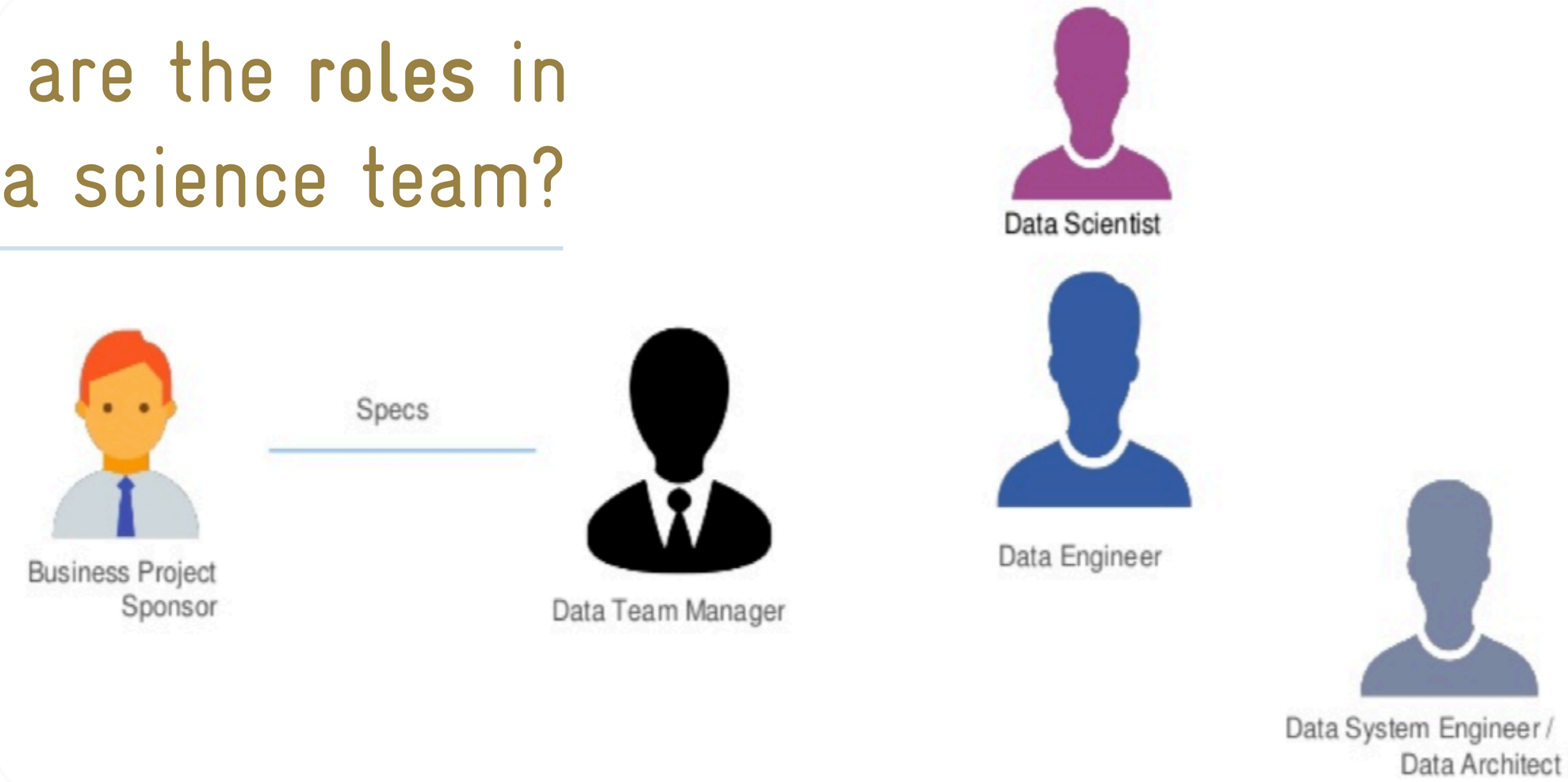
Math & Statistics and combines that with

Subject Matter Expertise.



Question 12:

What are the **roles** in a data science team?



Question 13:

What does a data scientist do?

A data scientist has a background in **statistics** and **machine learning,** produces **actionable insights** and is a (data) **storyteller.**

Question 14:

What does a data engineer do?

A data engineer has a background in **signal processing**, **computer science** and **algorithms** and focuses on **cleaning**, **processing** and **storing** data.

Question 15:

What does a **data architect** do?

A data architect has a background in **system administration** and **distributed computing architectures** and focuses on systems that **acquire**, **store** and **retrieve** data.

Question 16:

What does a **data team manager** do?

A data team manager **checks** the **steps** of a data science project and the **correctness** of the results, knows how to **build strong teams**, provides **help** and overall **project direction** and has an understanding of how **data** can help **shape** a **team's decision.**

Question 17:

Which are the timeless attributes for tools selection?

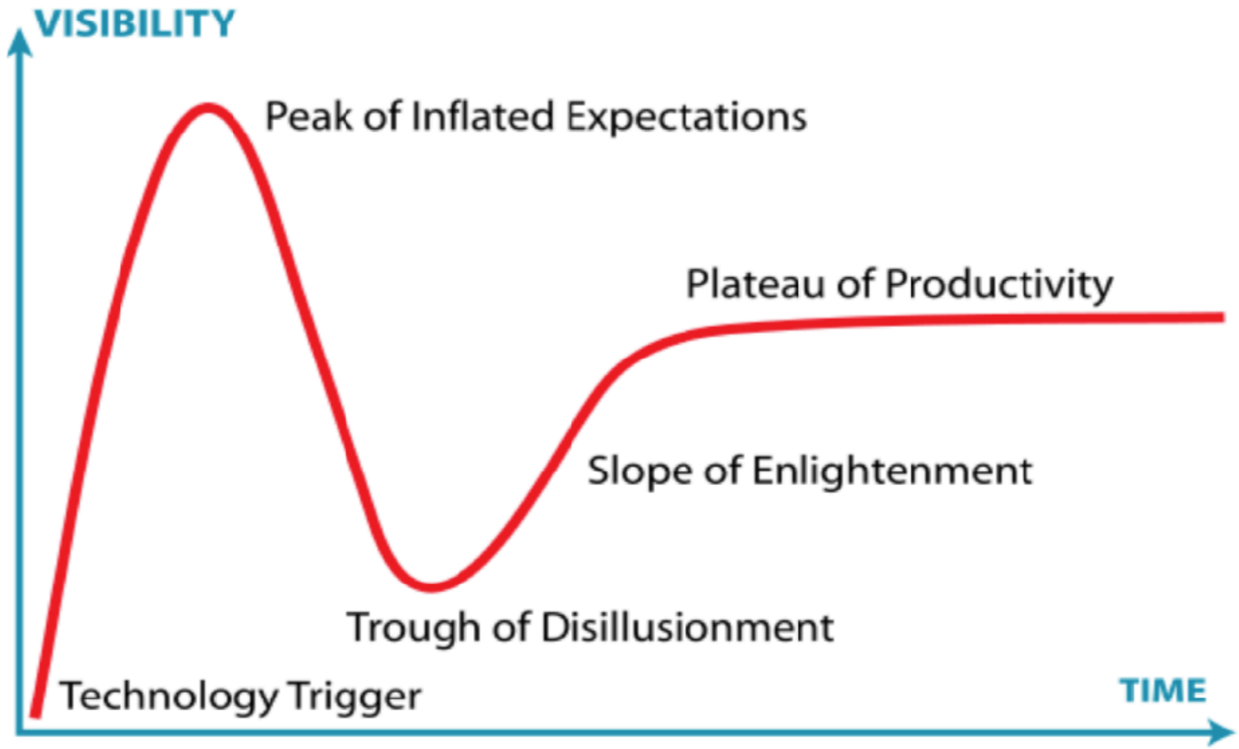
The tools are:

- **Powerful**
- **Easy to use**
- **Easy to learn**
- **Support teamwork** and
- **Beloved by the community**

Question 18:

Which are the phases of the hype-hope curve?

- **Technology Trigger**
- **Peak** of **Inflated Expectations**
- **Trough** of **Disillusionment**
- **Slope** of **Enlightenment**
- **Plateau** of **Productivity**



Question 19:

Which are the questions to understand if a Data Science project is **hype** or **hope**?

1. **What question** are you trying to answer?
2. Do you have (**relevant**!) **data** to answer the question
3. Can we **use** the answer?

Question 20:

Which types of Analyst Archetypes are there?

1. **Hackers**
2. **Scripters**
3. **Application users**

Question 21:

Can you highlight **3 key characteristics** of a **Hacker**?

1. **Most proficient programmers**
(comfortable manipulating data, masters multiple (scripting) languages)
2. **Most diverse and complex workflows**
(typically jobs at scale using distributed computing resources)
3. **Less sophisticated statistical models**
(due to larger data sets)

Question 22:

Can you highlight **3 key characteristics** of a **Scripter**?

1. **Capable of simple manipulations**
(filtering, aggregating, simple SQL-queries)
2. **Applies the most sophisticated models**
(spends relatively a lot of time on this)
3. **Uses the same tool for visualisation and analysis**
(allows him/her to iterate between the two tasks)

Question 23:

Can you highlight **3 key characteristics** of an **Application user**?

1. Uses **spreadsheet** or **data analysis software**
(e.g. Excel, SPSS and SAS)
2. **Dependent on others for pulling data**
(receives data as spreadsheet from IT team)
3. Works on **smaller datasets**

Question 24:

Which are the **three steps** of the **sandwich** method?

1. Focus on the **strengths**
2. **Provide criticism**

3a. **Reiterate positive** comments

3b. Mention **expected results** if the criticism is acted upon

Question 25:

Which **questions** can data science answer?

- **Descriptive** (summaries without interpretation)
- **Exploratory** (*not* been tested whether results hold on another sample)
- **Inferential** (tested whether results hold on another sample)
- **Predictive** (predict measurements for individuals)
- **Causal** (average effect)
- **Mechanistic** (deterministic effect)

Question 26:

Can you name **3 characteristics** of a **good (data) question**?

1. **Of interest to an audience** (identity depends on the context and environment)
2. **Not already been answered**
3. **Answerable** (data availability, feasibility, ethics)

Question 27:

Which are the types of analytics?

- **Descriptive** *(what happened)*,
- **Diagnostic** *(why did it happen)*,
- **Predictive** *(what will happen),*
- **Prescriptive** *(what should I do)*

Question 28:

What is descriptive analytics?

- Outcome of a **descriptive / exploratory** question
- Shows what **happened** or is **happening now**
- Uses **email reports** / **dashboards**
- Requires **considerable human intervention**

Question 29:

What is diagnostic analytics?

- Outcome of a **inferential** question
- Looks at the **past performance** to answer: **why** did it happen?
- Uses **analytical dashboards**
- Requires **considerable human intervention**

Question 30:

What is **predictive analytics**?

- Outcome of a **predictive** question
- Predicts what **will happen**
- Yields **predictive forecasts**
- Requires **minimal human intervention**

Question 31:

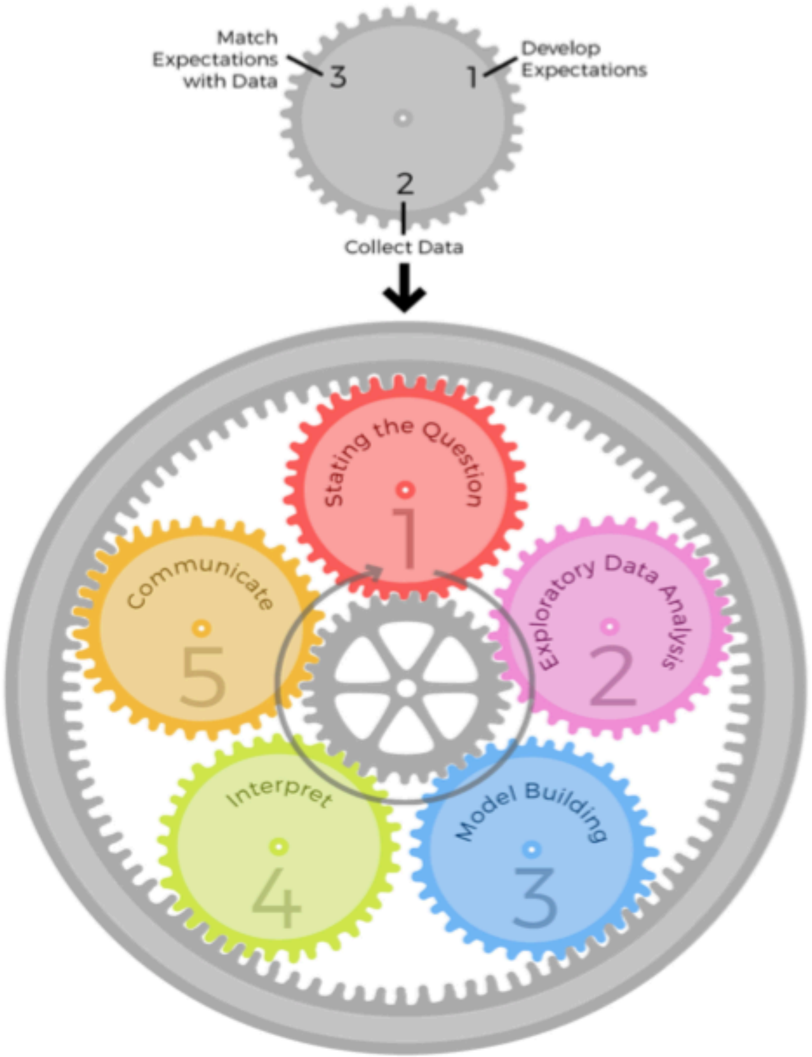
What is prescriptive analytics?

- Outcome of a **causal / mechanistic** question
- Looks at the **past performance** to answer: **what actions should be taken?**
- Uses **analytical dashboards**
- Requires **no human intervention (automated)**

Question 32:

Which are the **steps** of the Data Analysis process?

1. **Stating** and **refining** the **question**
2. **Exploring** the **data**
3. **Building** formal statistical / predictive **models**
4. **Interpreting** the results
5. **Communicating** the results



Question 33:
What is Map-Reduce?

A **parallel computing programming** paradigm developed at Google used to **preprocess** and **analyse** web pages.

Question 34:
What is Hadoop?

An open-source **distributed big data file system** by Google File System

and implemented at Yahoo!

Question 35:
What is Spark?

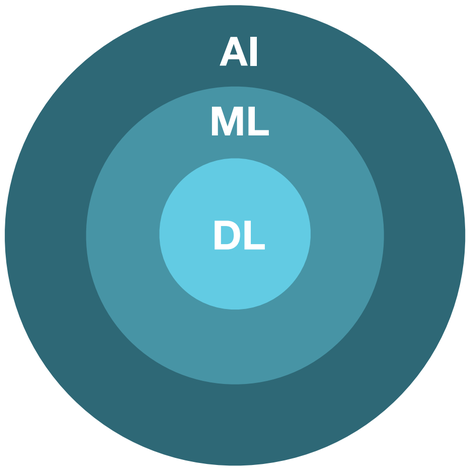
A new generation of **distributed big data processing system** that is **~10x faster** than MapReduce or Hadoop

Question 36:

What is the difference between Artificial Intelligence, Machine Learning and Deep Learning?

At the highest level, Machine Learning (ML) is a **subset** of Artificial Intelligence (AI).

In the same way, Deep Learning (DL) is a **subset** of Machine Learning.



Exists since: AI (1950), ML (1980), DL (2010)

Essence: Solve problems humans are typically better at (AI),

learn from data to make predictions (ML), mimic the human brain (DL).

Question 37:

Which are the types of Machine Learning?

There are two types of Machine Learning algorithms:

- **Supervised Learning** (*computer learns from examples*) which can be divided into:
  - Classification (*categorical target variable*)
  - Regression (*numerical target variable*)
- **Unsupervised Learning** *(computer learns the structure of the data / identifies clusters)*

Question 38:

Which are the **inputs** of a **Machine Learning algorithm**?

Inputs to ML-algorithms are **features** (i.e. the columns in a dataset).

- **Numeric format** (e.g. dates, income)
- **Signal processing** (sound/image input)
- **(Text) probability and statistics:**
  (word frequency, co-occurrences = which words are grouped together)

Question 39:

Why ensemble learning works?

By **combining** the predictions of multiple classifiers (i.e. ensembles) the **error term** of the prediction is **reduced** significantly.

Question 40:

What are the **trade-offs** in machine learning?

There are five trade-off in machine learning:

1. **Interpretability** (e.g. Decision Tree vs Random Forest)
2. **Fast** (training and testing times)
3. **Accurate** predictions
4. **Simple** to understand models (e.g. Linear Regression vs Deep Learning)
5. **Scalable** (for large data-sets)

Question 41:
Why are features important?

In a way features are the **building blocks** for any machine learning algorithm.

Using **linear algebra** features are used to fit a model to the data.

Question 42:

How can we make data linearly separable using SVM?

By applying the **Kernsel Density trick** (i.e. mapping data to a higher dimension)

Question 43:

Can you mention 3 good points of traditional statistical analyses?

1. The **parameters** of a statistical model can be easily **interpreted**
2. Statistical models are **easy** to **understand**.
3. Findings can be generalised over a **population** rather than a **sample**.

Question 44:

Can you mention 3 good points of Machine Learning?

1. It oftentimes leads to a **higher performance** than statistical models.
2. More **accurate** and **informative** alternative on **small data** sets
3. It can also be used for **large** and **complex** datasets

Question 45:

Which are the questions you should ask for creating a **business case**?

1. What is you **goal**?
2. **What** is **stopping you** from reaching that goal?
3. **How much change** is needed to overcome the problem?
4. Are you **certain** this will **solve** the **problem**?

Question 46:

Which are the 3 components of persuasion and their meaning?

1. **Logos =** the ideas make sense from the audience's point of view
2. **Ethos =** your reputation/credibility; what are you known for?
3. **Pathos =** emotional connection (e.g. by telling a personal story)

Source: https://www.youtube.com/watch?v=O2dEuMFR8kw (02:07)

Question 47:

Can you mention 5 out of 10 points to make better presentations?

1. **Turn off the computer/technology** (great ideas don't come from the computer)
2. **Put the audience first** (make them care emotionally, intellectually, aesthetically)
3. **Have a solid structure** (ideal > reality > problem > solution > next step)
4. **Remove the nonessential** (everything has a reason)
5. **Hook them early** (just get started, no formalities in advance)

Source: https://www.youtube.com/watch?v=YbV3b-I1sZs

Question 48:

What is the **mission** of start-up **Connecterra**?

“**Connecting the analogue to understand everything on earth”** (by applying AI/ML/Data Science) Ida, one of their products, is a dairy farmer’s assistant that provides meaningful insights to farmers based on cows behaviour (eating, walking, standing).

Question 49:

What is the **business-model** of the start-up **Connecterra**?

Their service to farmers is on a **subscription basis**: €7,50 per cow / month.

Question 50:

Which are the **opportunities** for Data Science in **insurance**?

The biggest opportunity is in insurance **claims** (risk based pricing / fraud detection).

Other opportunities are in **marketing** (e.g. churn), reducing **overhead** and optimising their profit **margin**.

Question 51:

How do Data Scientists at ABN Amro reduce the **variance** of **churn prediction** models?

They make use of **ensemble models** such as Random Forest. The basic idea there is that you train a classifier on **multiple random samples** of the data based on a random subset of features for each iteration.

Question 52:

What **train flow** do Data Scientists at Rabobank use?

They use a rather standard data science approach:
Select features from historic data, **prepare** the **data** (dummies for categorical variables, imputation for NAs, scaling for better performance, add interactions), **build** a **ML-model** using the H2O package, **validate** hypotheses with a statistical **test** and eventually evaluate the **model performance**.

Question 53:

Can you mention 2 **examples** of **personalisation**?

Netflix’s movie recommendation system, chatbots, Google searches, your Facebook feed, Amazon’s product recommendations, Spotify’s Discovery Weekly playlist, and many more…

Question 54:

Which are **common features** to detect **stress**?

Common features to detect stress are: **heart rate** (variability), **skin conductance level** and **respiration rate**.

Question 55:

Which are typical data sources used to rank hotels?

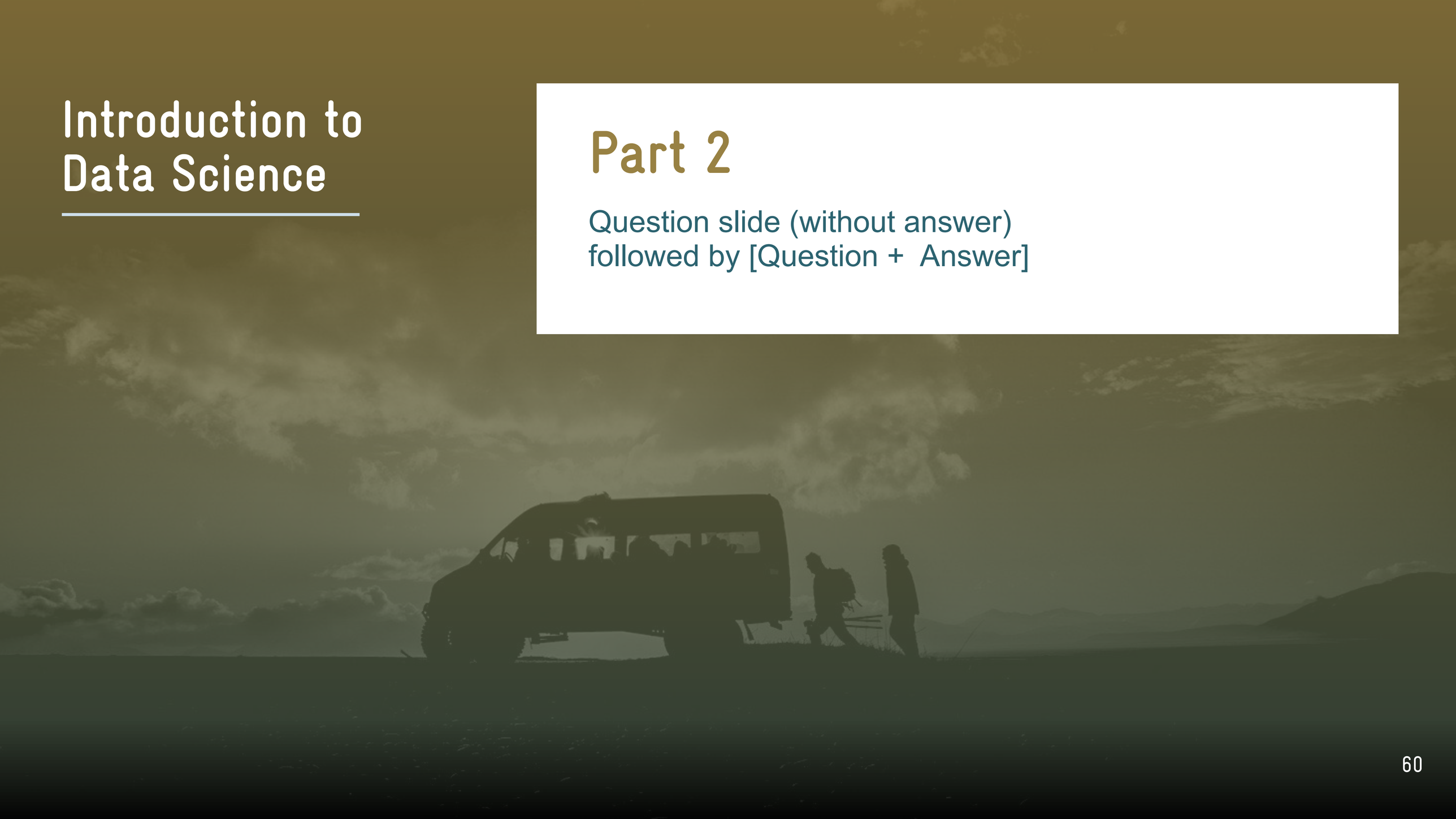
Trivago uses the following data sources to rank hotels:

- **Hotel** meta-data (location)
- **Ratings** (star rating, cleanliness, breakfast, etc.)
- **Facilities** (laundry, room service, pool, etc.)
- **Images** (hotel / surroundings)
- **Personal click behaviour** (how the user interacted on the site)

Introduction to Data Science

Part 2

Question slide (without answer)
followed by [Question + Answer]



Question 1:
What is Data Science?

Question 1:
What is Data Science?

A set of **fundamental principles** that guide the **extraction** of **knowledge** from data

Question 2:
What is Data Engineering?

Question 2:
What is Data Engineering?

A set of **technologies** that facilitates the **access** to **data** and its **processing**

in order to serve Data Science.

Question 3:
What is Big Data?

Question 3:
What is Big Data?

Data processing tools that handle **Volume**, **Velocity** and **Variety**

that traditional data processing systems cannot support.

Question 4:
What is Data Mining?

Question 4:
What is Data Mining?

Tools and **processes** that guide the **extraction** of **knowledge** via Data Science principles.

Question 5:
What is date-driven decision making?

Question 5:
What is **date-driven decision making?**

The practice of **basing decisions** on the **analysis** of **data** rather than purely on **intuition.**

Question 6:

What organisations do for creating a **data-driven culture**?

Question 6:

What organisations do for creating a **data-driven culture**?

They..

acquire *(focus on keeping data clean),*

process *(democratise data access)*,

and **leverage** *(use data in decision-making and for product creation)*

..data in order to navigate the competitive landscape.

Question 7:

What “**acquire**” means in data-driven organizations?

Question 7:

What “acquire” means in data-driven organizations?

- Focus on **keeping** the **data clean** (typically 80% of the work) both on a small and **large** scale
- Storing data with a **predefined purpose**

Question 8:

What “process” means in data-driven organizations?

Question 8:

What “process” means in data-driven organizations?

Democratize access to **data** to encourage **experiments** and **hypotheses testing**

with the aim of **improving** an **organization** and its **processes.**

Question 9:

How do Data Scientists **spend most** of their **time**?

Question 9:

How do Data Scientists **spend most** of their **time**?

80% of the work consists of **cleaning** the **data**

Question 10:
What is a Data Scientist?

Question 10:
What is a Data Scientist?

A **computer scientist**, **mathematician** and **storyteller** combined with a **curious attitude, domain knowledge** and **expertise.**

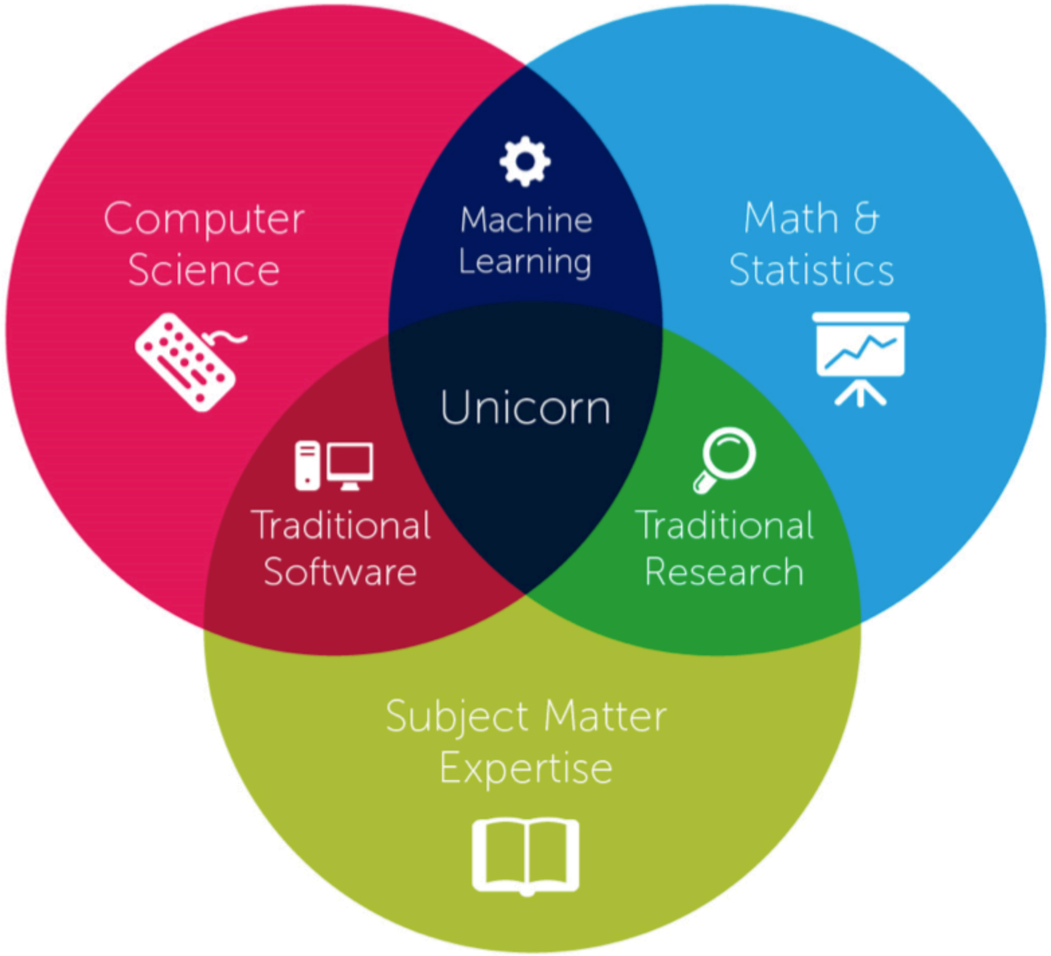
Question 11:
What is a **data science unicorn**?

Question 11:
What is a data science unicorn?

Someone who masters the field of **Computer Science,**

Math & Statistics and combines that with

Subject Matter Expertise.

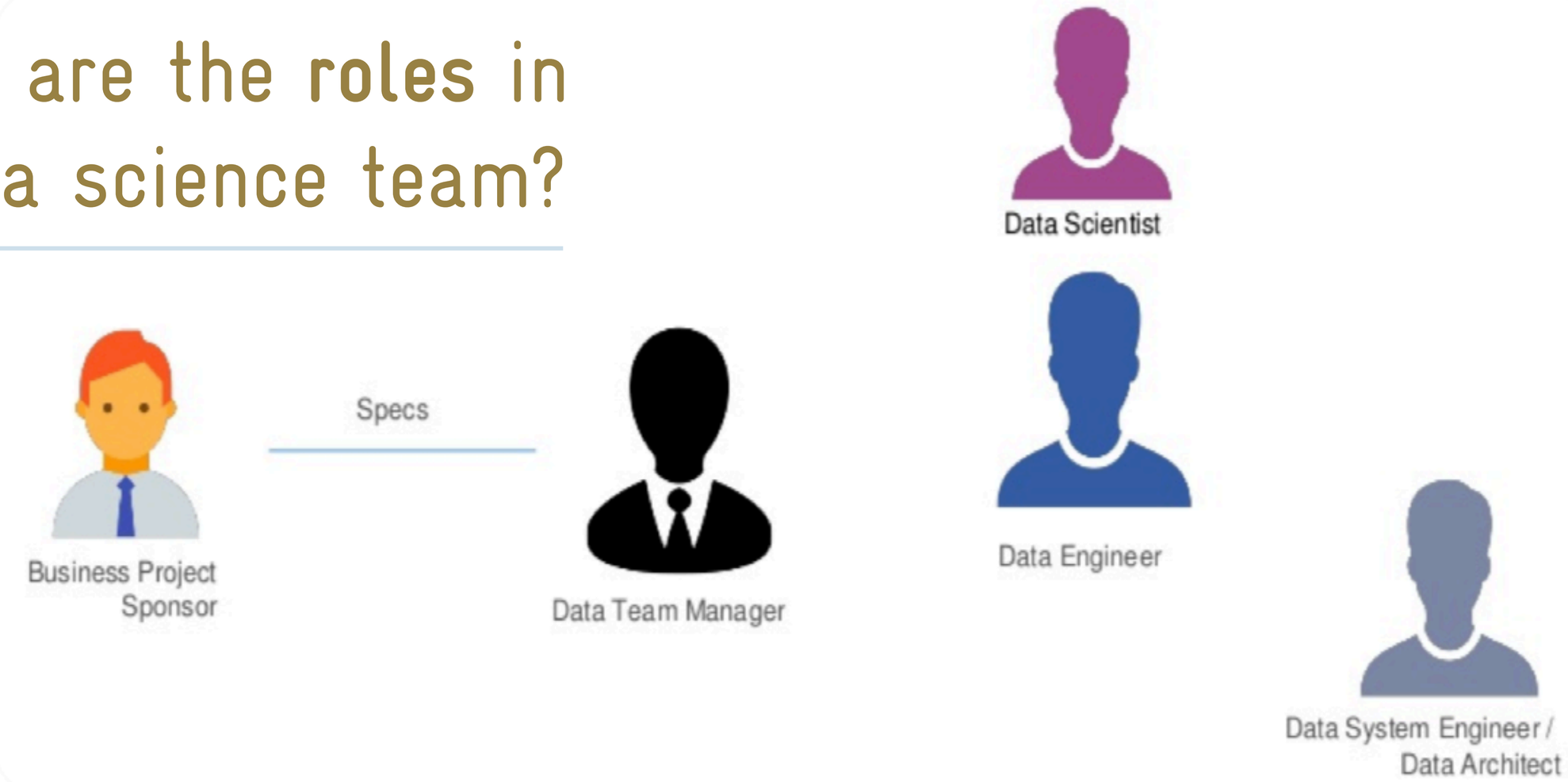


Question 12:

What are the **roles** in a data science team?

Question 12:

What are the **roles** in a data science team?



Question 13:

What does a data scientist do?

Question 13:

What does a **data scientist** do?

A data scientist has a background in **statistics** and **machine learning,** produces **actionable insights** and is a (data) **storyteller.**

Question 14:

What does a data engineer do?

Question 14:

What does a data engineer do?

A data engineer has a background in **signal processing**, **computer science** and **algorithms** and focuses on **cleaning**, **processing** and **storing** data.

Question 15:

What does a **data architect** do?

Question 15:

What does a **data architect** do?

A data architect has a background in **system administration** and **distributed computing architectures** and focuses on systems that **acquire**, **store** and **retrieve** data.

Question 16:

What does a **data team manager** do?

Question 16:

What does a data team manager do?

A data team manager **checks** the **steps** of a data science project and the **correctness** of the results, knows how to **build strong teams**, provides **help** and overall **project direction** and has an understanding of how **data** can help **shape** a **team's decision.**

Question 17:

Which are the **timeless attributes** for **tools selection**?

Question 17:

Which are the timeless attributes for tools selection?

The tools are:

- **Powerful**
- **Easy to use**
- **Easy to learn**
- **Support teamwork** and
- **Beloved by the community**

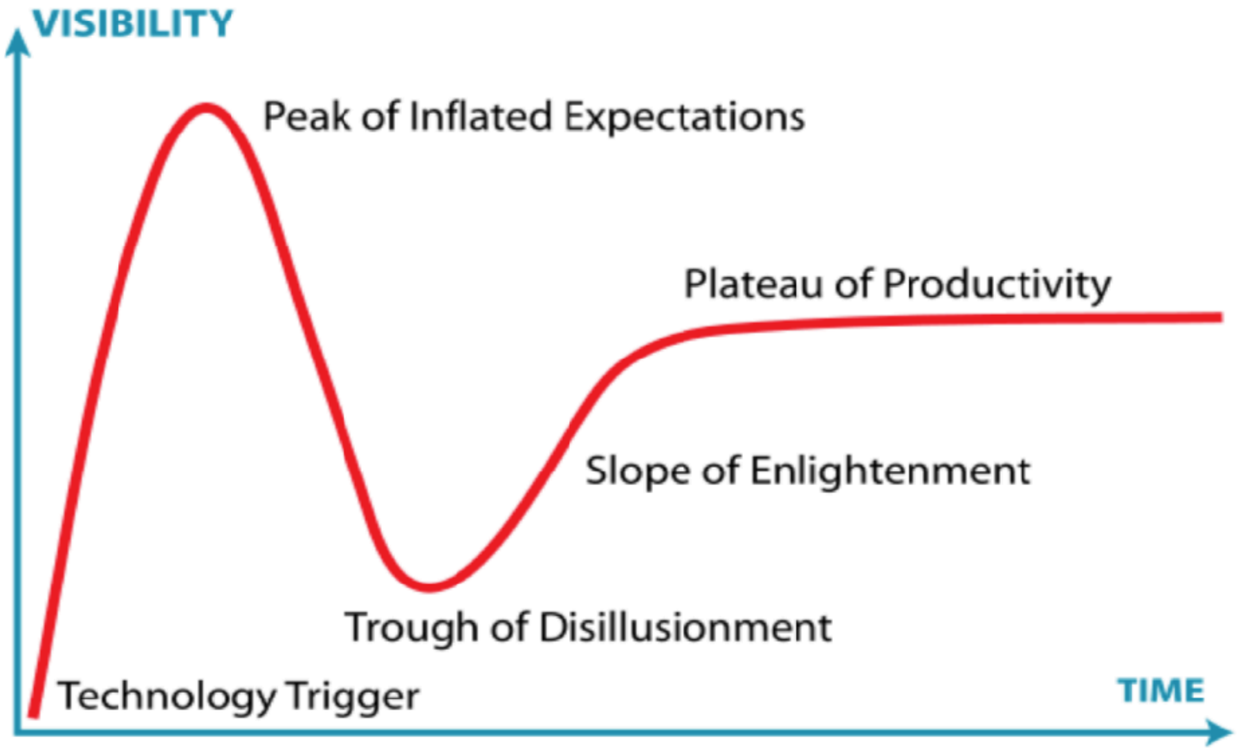
Question 18:

Which are the **phases** of the **hype-hope curve**?

Question 18:

Which are the phases of the hype-hope curve?

- **Technology Trigger**
- **Peak** of **Inflated Expectations**
- **Trough** of **Disillusionment**
- **Slope** of **Enlightenment**
- **Plateau** of **Productivity**



Question 19:

Which are the questions to understand if a Data Science project is **hype** or **hope**?

Question 19:

Which are the questions to understand if a Data Science project is **hype** or **hope**?

1. **What question** are you trying to answer?
2. Do you have (**relevant**!) **data** to answer the question
3. Can we **use** the answer?

Question 20:

Which types of Analyst Archetypes are there?

Question 20:

Which types of Analyst Archetypes are there?

1. **Hackers**
2. **Scripters**
3. **Application users**

Question 21:

Can you highlight **3 key characteristics** of a **Hacker**?

Question 21:

Can you highlight 3 key characteristics of a Hacker?

1. **Most proficient programmers**
(comfortable manipulating data, masters multiple (scripting) languages)
2. **Most diverse and complex workflows**
(typically jobs at scale using distributed computing resources)
3. **Less sophisticated statistical models**
(due to larger data sets)

Question 22:

Can you highlight **3 key characteristics** of a **Scripter**?

Question 22:

Can you highlight **3 key characteristics** of a **Scripter**?

1. **Capable of simple manipulations**
(filtering, aggregating, simple SQL-queries)
2. **Applies the most sophisticated models**
(spends relatively a lot of time on this)
3. **Uses the same tool for visualisation and analysis**
(allows him/her to iterate between the two tasks)

Question 23:

Can you highlight **3 key characteristics** of an **Application user**?

Question 23:

Can you highlight 3 key characteristics of an Application user?

1. Uses **spreadsheet** or **data analysis software**
(e.g. Excel, SPSS and SAS)
2. **Dependent on others for pulling data**
(receives data as spreadsheet from IT team)
3. Works on **smaller datasets**

Question 24:

Which are the **three steps** of the **sandwich** method?

Question 24:

Which are the **three steps** of the **sandwich** method?

1. Focus on the **strengths**
2. **Provide criticism**

3a. **Reiterate positive** comments

3b. Mention **expected results** if the criticism is acted upon

Question 25:

Which **questions** can data science answer?

Question 25:

Which **questions** can data science answer?

- **Descriptive** (summaries without interpretation)
- **Exploratory** (*not* been tested whether results hold on another sample)
- **Inferential** (tested whether results hold on another sample)
- **Predictive** (predict measurements for individuals)
- **Causal** (average effect)
- **Mechanistic** (deterministic effect)

Question 26:

Can you name **3 characteristics** of a **good (data) question**?

Question 26:

Can you name 3 characteristics of a good (data) question?

1. **Of interest to an audience** (identity depends on the context and environment)
2. **Not already been answered**
3. **Answerable** (data availability, feasibility, ethics)

Question 27:

Which are the types of analytics?

Question 27:

Which are the types of analytics?

- **Descriptive** *(what happened)*,
- **Diagnostic** *(why did it happen)*,
- **Predictive** *(what will happen),*
- **Prescriptive** *(what should I do)*

Question 28:

What is descriptive analytics?

Question 28:

What is descriptive analytics?

- Outcome of a **descriptive / exploratory** question
- Shows what **happened** or is **happening now**
- Uses **email reports** / **dashboards**
- Requires **considerable human intervention**

Question 29:

What is **diagnostic analytics**?

Question 29:

What is diagnostic analytics?

- Outcome of a **inferential** question
- Looks at the **past performance** to answer: **why** did it happen?
- Uses **analytical dashboards**
- Requires **considerable human intervention**

Question 30:

What is **predictive analytics**?

Question 30:

What is **predictive analytics**?

- Outcome of a **predictive** question
- Predicts what **will happen**
- Yields **predictive forecasts**
- Requires **minimal human intervention**

Question 31:

What is **prescriptive analytics**?

Question 31:

What is prescriptive analytics?

- Outcome of a **causal / mechanistic** question
- Looks at the **past performance** to answer: **what actions should be taken?**
- Uses **analytical dashboards**
- Requires **no human intervention (automated)**

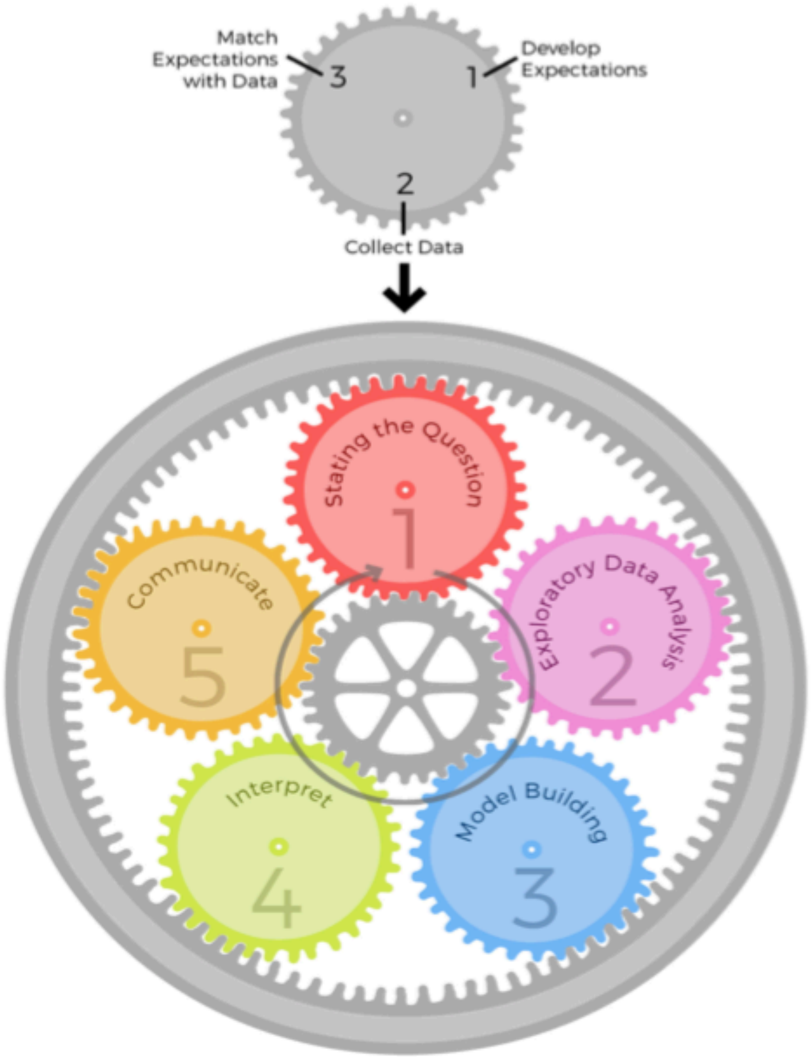
Question 32:

Which are the **steps** of the Data Analysis process?

Question 32:

Which are the **steps** of the Data Analysis process?

1. **Stating** and **refining** the **question**
2. **Exploring** the **data**
3. **Building** formal statistical / predictive **models**
4. **Interpreting** the results
5. **Communicating** the results



Question 33:
What is Map-Reduce?

Question 33:
What is Map-Reduce?

A **parallel computing programming** paradigm developed at Google used to **preprocess** and **analyse** web pages.

Question 34:
What is Hadoop?

Question 34:
What is Hadoop?

An open-source **distributed big data file system** by Google File System

and implemented at Yahoo!

Question 35:
What is Spark?

Question 35:
What is Spark?

A new generation of **distributed big data processing system** that is **~10x faster** than MapReduce or Hadoop

Question 36:

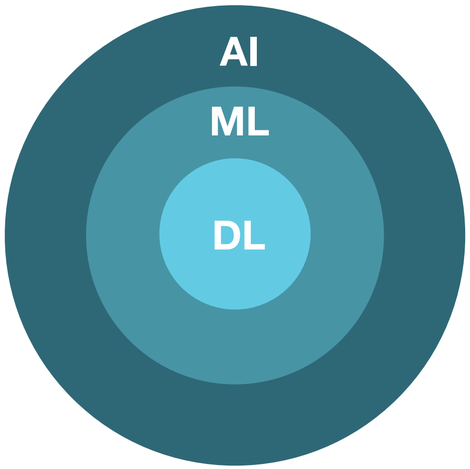
What is the difference between **Artificial Intelligence**, **Machine Learning** and **Deep Learning**?

Question 36:

What is the difference between Artificial Intelligence, Machine Learning and Deep Learning?

At the highest level, Machine Learning (ML) is a **subset** of Artificial Intelligence (AI).

In the same way, Deep Learning (DL) is a **subset** of Machine Learning.



Exists since: AI (1950), ML (1980), DL (2010)

Essence: Solve problems humans are typically better at (AI),

learn from data to make predictions (ML), mimic the human brain (DL).

Question 37:

Which are the types of Machine Learning?

Question 37:

Which are the types of Machine Learning?

There are two types of Machine Learning algorithms:

- **Supervised Learning** (*computer learns from examples*) which can be divided into:
  - Classification (*categorical target variable*)
  - Regression (*numerical target variable*)
- **Unsupervised Learning** *(computer learns the structure of the data / identifies clusters)*

Question 38:

Which are the **inputs** of a **Machine Learning algorithm**?

Question 38:

Which are the **inputs** of a **Machine Learning algorithm**?

Inputs to ML-algorithms are **features** (i.e. the columns in a dataset).

- **Numeric format** (e.g. dates, income)
- **Signal processing** (sound/image input)
- **(Text) probability and statistics:**
  (word frequency, co-occurrences = which words are grouped together)

Question 39:

Why ensemble learning works?

Question 39:

Why ensemble learning works?

By **combining** the predictions of multiple classifiers (i.e. ensembles) the **error term** of the prediction is **reduced** significantly.

Question 40:

What are the **trade-offs** in machine learning?

Question 40:

What are the **trade-offs** in machine learning?

There are five trade-off in machine learning:

1. **Interpretability** (e.g. Decision Tree vs Random Forest)
2. **Fast** (training and testing times)
3. **Accurate** predictions
4. **Simple** to understand models (e.g. Linear Regression vs Deep Learning)
5. **Scalable** (for large data-sets)

Question 41:
Why are features important?

Question 41:
Why are features important?

In a way features are the **building blocks** for any machine learning algorithm.

Using **linear algebra** features are used to fit a model to the data.

Question 42:

How can we make data **linearly separable** using **SVM**?

Question 42:

How can we make data linearly separable using SVM?

By applying the **Kernsel Density trick** (i.e. mapping data to a higher dimension)

Question 43:

Can you mention 3 good points of **traditional statistical analyses**?

Question 43:

Can you mention 3 good points of traditional statistical analyses?

1. The **parameters** of a statistical model can be easily **interpreted**
2. Statistical models are **easy** to **understand**.
3. Findings can be generalised over a **population** rather than a **sample**.

Question 44:

Can you mention 3 good points of Machine Learning?

Question 44:

Can you mention 3 good points of Machine Learning?

1. It oftentimes leads to a **higher performance** than statistical models.
2. More **accurate** and **informative** alternative on **small data** sets
3. It can also be used for **large** and **complex** datasets

Question 45:

Which are the questions you should ask for creating a **business case**?

Question 45:

Which are the questions you should ask for creating a **business case**?

1. What is you **goal**?
2. **What** is **stopping you** from reaching that goal?
3. **How much change** is needed to overcome the problem?
4. Are you **certain** this will **solve** the **problem**?

Question 46:

Which are the **3 components** of **persuasion** and their meaning?

Question 46:

Which are the 3 components of persuasion and their meaning?

1. **Logos =** the ideas make sense from the audience's point of view
2. **Ethos =** your reputation/credibility; what are you known for?
3. **Pathos =** emotional connection (e.g. by telling a personal story)

Source: https://www.youtube.com/watch?v=O2dEuMFR8kw (02:07)

Question 47:

Can you mention **5** out of **10** points to make **better presentations**?

Question 47:

Can you mention 5 out of 10 points to make better presentations?

1. **Turn off the computer/technology** (great ideas don't come from the computer)
2. **Put the audience first** (make them care emotionally, intellectually, aesthetically)
3. **Have a solid structure** (ideal > reality > problem > solution > next step)
4. **Remove the nonessential** (everything has a reason)
5. **Hook them early** (just get started, no formalities in advance)

Source: https://www.youtube.com/watch?v=YbV3b-I1sZs

Question 48:

What is the **mission** of start-up **Connecterra**?

Question 48:

What is the **mission** of start-up **Connecterra**?

“**Connecting the analogue to understand everything on earth”** (by applying AI/ML/Data Science) Ida, one of their products, is a dairy farmer’s assistant that provides meaningful insights to farmers based on cows behaviour (eating, walking, standing).

Question 49:

What is the **business-model** of the start-up **Connecterra**?

Question 49:

What is the **business-model** of the start-up **Connecterra**?

Their service to farmers is on a **subscription basis**: €7,50 per cow / month.

Question 50:

Which are the **opportunities** for Data Science in **insurance**?

Question 50:

Which are the **opportunities** for Data Science in **insurance**?

The biggest opportunity is in insurance **claims** (risk based pricing / fraud detection).

Other opportunities are in **marketing** (e.g. churn), reducing **overhead** and optimising their profit **margin**.

Question 51:

How do Data Scientists at ABN Amro reduce the **variance** of **churn prediction** models?

Question 51:

How do Data Scientists at ABN Amro reduce the **variance** of **churn prediction** models?

They make use of **ensemble models** such as Random Forest. The basic idea there is that you train a classifier on **multiple random samples** of the data based on a random subset of features for each iteration.

Question 52:

What **train flow** do Data Scientists at Rabobank use?

Question 52:

What **train flow** do Data Scientists at Rabobank use?

They use a rather standard data science approach:
Select features from historic data, **prepare** the **data** (dummies for categorical variables, imputation for NAs, scaling for better performance, add interactions), **build** a **ML-model** using the H2O package, **validate** hypotheses with a statistical **test** and eventually evaluate the **model performance**.

Question 53:

Can you mention 2 **examples** of **personalisation**?

Question 53:

Can you mention 2 **examples** of **personalisation**?

Netflix’s movie recommendation system, chatbots, Google searches, your Facebook feed, Amazon’s product recommendations, Spotify’s Discovery Weekly playlist, and many more…

Question 54:

Which are **common features** to detect **stress**?

Question 54:

Which are **common features** to detect **stress**?

Common features to detect stress are: **heart rate** (variability), **skin conductance level** and **respiration rate**.

Question 55:

Which are typical data sources used to rank hotels?

Question 55:

Which are typical data sources used to rank hotels?

Trivago uses the following data sources to rank hotels:

- **Hotel** meta-data (location)
- **Ratings** (star rating, cleanliness, breakfast, etc.)
- **Facilities** (laundry, room service, pool, etc.)
- **Images** (hotel / surroundings)
- **Personal click behaviour** (how the user interacted on the site)

One more thing...

That was it!

In other words, time to showcase your memorisation **super powers** (or just insert these (171!) slides as back-up slides..)

Anyhow, good luck! 🌟