- We just went through why data science matters,

you can ask yourself is,

why am I doing data science in this problem?

What insights do I expect?

The question will lead you to find what you're looking for

and you can design strategies in alignment with your goals.

Our goal is to take the data

and generate insights from it

that we can use to take data-driven actions.

We call these insights actionable insights,

which are very valuable,

and they require knowledge of the subject area

or business domain.

For the soccer scenario,

the insights we are trying to generate

are related to better understanding

of player strengths, performance enhancement,

and attributes of a player's performance.

We can turn this into a question, saying,

I want to find the quickest way

to improve my favorite player's performance.

How do I know what traits impact a player's performance

more than the others?

The coach can then take these insights

and take actions.

To design programs that build upon these insights

to improve team strengths.

So how do we go about taking the steps to go from data

to those most coveted insights?

There are five key steps

in the overall process of data science, namely,

data acquisition, data preparation,

data analysis, presentation, and reporting of insights,

and turning these insights into data-driven actions.

In our soccer example,

the acquire involves downloading the data set,

or importing it to your data science.

Acquisition step leads into

data exploration and visualization,

and other data preparation.

We then analyze the prepared data set

using statistical analysis and machine learning.

Findings from these analyses

are typically turned into reports

and presented to the stakeholders who need to take actions.

Using such insights,

the stakeholders will go and take their actions

and use them.

In general, as we go through the examples,

you will notice there are similar steps

in any data science project,

much like the steps you take when you drive

to the grocery store from your home,

pick the keys, open the door,

take a form of transport, car, bus, cycle,

move, arrive, park if needed, et cetera.

The specifics of each step

can be different,

but the overall process remains more or less the same.

We will look into the data science process

as a generalizable activity the rest of this week,

but let's first continue to look at these steps

in the light of the soccer data analysis example.

As a first step in any data science activity,

and some applications of it.

In the rest of this week,

we will look at the different stages of data science

in a structured manner.

Before we dive into that process,

let us look through a live example

of how data science works in action.

It is much more engaging to take real data sets

and do your own analytics

and come to your own conclusions.

We will take the European soccer data set

and perform analytics on it.

Although how to build some parts of this example

is not very clear to you now,

by the end of this course achieving similar things

should be your goal.

So let's go.

By the end of this video,

you will be able to talk about the Big Picture

of data science through a soccer case study,

generate statistics about a soccer data set,

summarize how data cleaning and correlations

were applied to an existing data set,

recite the data visualization techniques

employed in this study,

explain how clustering similar groups

and plotting these clusters help the case study,

and recall what was used to draw conclusions

based on this data analysis.

Each week in this course,

we will work with a real data set

and work on a case study.

This week we will be using an open data sets

from the popular site Kaggle.

This European database

have more than 25,000 matches,

and more than 10,000 players

for European professional soccer seasons

from 2008 to 2016.

Although we won't be getting into the details of it

for our example,

the data set even has attributes on weekly game updates,

team lineup, and detailed match events.

We will use these data sets to demonstrate

the basic steps of the process we take

for such data science projects,

and use the data set for three main goals.

Form meaningful player groups,

discover other players that are similar

to your favorite athlete,

and form strong teams by using analytics.

Since we are looking for questions to solve in data science

we can formulate the question around these goals.

For example, I would instead say,

how do I form meaningful player groups

to find players similar to my favorite player?

Finally, how do I use this information

to form strong teams?

To go even further, we need to ask ourselves,

why do I want to know about strong teams,

or what is the benefits of using analytics?

In fact, one of the most critical questions

we need to consider that data can come from

that can make this task much easier for us.

We have a dedicated visualization section

in the upcoming weeks.

Analysis is the crux of the matter in data science.

Once the basic preparatory steps are completed,

you get to the algorithms.

Although the third course in our micro masters

will take you for a ride through these algorithms,

in this first course,

we will introduce you to some of them.

There are three key categories,

supervised learning, unsupervised learning,

and semi-supervised learning.

There are vast number of algorithms and techniques

as seen in this diagram

for dimensionality reduction, clustering,

and regression, for example.

The scikit-learn library in Python

provides many tools for machine learning in Python.

We will introduce some of these tools to you

as part of this first course.

Here, we start giving you examples

of using some of these tools

as a part of the soccer data analysis,

and our upcoming case studies.

As a part of this example, we do feature selection.

Feature selection is about selecting attributes

that have the greatest impact

towards the problem you are solving.

It requires some domain knowledge to narrow down

the number of features.

For example, in the soccer use case,

if you are trying to predict a player's performance,

what are the most critical attributes?

Blue attributes on agility, reaction time,

shot power, and sprint speed,

or green attributes on hairstyle, or movie preferences?

Similarly, if you're grouping players into different sets,

what attributes would you choose to assign these groupings

to create complex features?

Narrowing the features has several benefits.

You get models that are easier to interpret,

models get trained much faster,

and you're likely to generalize well

to newer scenarios.

You will find almost every top machine learning algorithm

already implemented in Python.

Different functionalities are organized

into libraries in Python, like scikit-learn,

containing implementations

of fundamental machine learning algorithms.

In the soccer example,

we will utilize a form of clustering algorithm

called K-Means from sklearn.

Clustering means grouping your players

into similar meaningful sets,

based on those attributes you just decided.

The idea of this slide

is to show you how quickly

and concisely you can call sub-routines in Python

that do exactly what you want to do.

We import the right library in Python, for K-means,

and then we use the library to analyze our data.

many different sources.

This diversity of data sources

will only continue to grow as more innovations are made.

The broad categories include relational,

and NoSQL databases,

text files in various data formats,

and live online streams coming from machines' sensors

and online activities.

In our soccer example,

the providers of the data set

gathered the data scattered across many Internet sites

and did a tour of data collection and processing

to make the data ready for analysis.

The data set includes structured data on scores,

lineup, team formation, and events,

as well as data on betting odds

and players' and teams' attributes.

All we had to do in this case

was to take that data set and ingest it into Python.

Data ingestion will be one of our focus areas

in this course.

Python has well-defined methods for ingesting data

from diverse resources.

These sources include various databases,

data access APIs like the Twitter API,

text files, and sensors' data streams.

By the end of this course,

you will know how to efficiently utilize

each of these data sources.

The next step of our data science process

is exploring the data set.

Python has libraries that can assist you

in the data preparation phase

when you want to explore your data sets.

For example, with just one line of command,

as you see here,

you can generate vital statistical summary

of your data sets like mean and standard deviation.

The data preparation also involves data cleaning,

as there are many challenges in the real world data sets.

The cleaning can also build on the statistical analysis

like removing outliers, missing values,

or in general, weeding out unwanted stuff from your data.

Although sometimes removing the unwanted entries

can be a quick solution,

sometimes it can still be a challenge

to decide what to remove.

In those situations you can impute those fields

with known aggregate values

such as mean of the columns, et cetera.

Python offers data cleaning functions

to help with general data cleaning tasks

like finding and removing null values.

In the example notebook on soccer data analysis

we share some examples to get started with,

but we will point out to those functions

as we go through our case studies throughout this course.

In each step of the data science process,

data visualization is an effective way

to capture your team's attention

and convey your message in a minimal time.

Python has several open source data visualization libraries

Notice how clustering is performed here in just one line.

through data science.

We also discussed some of the tools

Python provides for data science.

As a part of this week's materials,

we give you a full Jupyter notebook

that shows our soccer data analysis example.

By the end of this course

you will be able to create similar notebooks of your own

for simple data analysis tasks.

Next, we will generalize this process

for any data science problem

before we start diving into the details of Unix

and Python scripting.

LINK: <https://www.kaggle.com/hugomathien/soccer/version/10>

Don't worry about the details of these methods yet.

Although it requires further expertise

and understanding to pick the right tool

for your analysis,

this course will start a foundation for you

to build up on in the coming courses.

Now let's go back to the data science process

we were following for our soccer case study.

At this point, we are done with clustering.

This means we have grouped players

into meaningful groups

based on the attributes we chose.

Next, we will start interpreting the results.

So how do we analyze the results?

Something to consider are,

do all of them have the same number of players?

When we look at the attributes we had selected,

how do these groups differ?

Plotting these clusters might help with interpreting

and presenting these results.

The graph here helps us with both.

Once we have done all the work of data cleaning,

analyzing, and interpretation,

it's time to present our findings.

A big part of the presentation, or reporting,

is explaining how we interpreted these results.

Look at the graph,

and think of the four lines

as signatures of each of the four clusters

our K-means algorithm has found,

based on the features shown in the x-axis of this plot.

Do any of the groups

have exactly the same signature?

The answer is no.

Each group is unique in the sense that

it differs from the other three

in at least one attribute.

To act upon such findings,

team coaches can use this information

to design customized improvement strategies

for each group.

There are many techniques and best practices

for presentation or visualization of these results.

We need to decide on the graph type,

find the library, or write our own,

have enough details for the picture

to be self-explanatory,

like adding label axes, legends,

and a readable font size.

To summarize, we just explained

a soccer data analysis case study,

and how we used a five step data science process

to generate insights from our original data set.

The process involved an acquiring results,

data preparation, analysis and reporting of results,

can then be used for data-driven actions.

We explained why actionable insights

are the expected product

of any data science project,

and the importance of using data

and domain knowledge

to generate questions that we can answer