- Now that we identified what data science is,

However, at the end of the day, the scalable process

should be programmable through utilization of reusable

and reproducible programming interfaces

to systems middleware, analytical tools,

visualization environments and user reporting environments.

This final dimension,

the programming of data science steps in Python,

is what we will be focusing on in this course

with some examples in Python modules for analytics,

data processing and visualization.

let's come back to using data science

to extract value out of data or to solve a question.

How does the data science process function from

the point of asking a question to answering that question?

In other words, can we generalize the process steps that

we started seeing in the soccer data analysis example

and see how data science leads to insights?

After this video, you will be able to list

some of the dimensions of modern data science

and identify why analyzing these dimensions

are important for us as data scientists.

Our experience with building and observing successful

data science projects can be summarized

as a craft with a few distinct components.

Those components can be defined as

the key features or dimensions of data science.

To summarize what we talked about before, we define

data science as a multidisciplinary craft that combines

an interdisciplinary team with an application purpose.

So generally speaking, data science starts with

a team of people with an overarching broad question

and of course some data to explore.

We can say that we start with data and question

and we build a process around how we come up

with a data driven insight.

A process is a conceptual entity in the beginning

and defines the core set of steps to solve the question.

And then from then on,

drilling down through many areas of expertise

often meant blurred lines between the steps.

There are many ways to look at the process.

One way of looking at it is as two distinct activities,

mainly data engineering and data analysis

or computational big data science as I'd like to call it,

as more than simple analytics is being performed

at this stage in general.

A more detailed way of looking at the process reveals five

distinct steps or activities of the data science process,

namely acquire, prepare, analyze,

report and act.

We can simply say data science happens

at the boundary of all these steps.

Ideally, this process should support experimental work

and dynamic scalability on big data and cloud platforms.

This five step process can be used in alternative ways

in real life big data applications if we add

the dependencies of different tools to each other.

The influence of big data pushes for alternative scalability

approaches at each step of the process.

Another way to look at this process is seeing

all these steps with reporting needs in different forms,

or drawing all these activities as an iterating process

including build, explore and scale for big data as steps.

Scalable data analysis needs alternative data management

techniques, systems, analytical tools and methods

as well as nodes of scalability based on dynamic data

and computing load, change in physical infrastructure

and streaming data specific urgencies

arising from special events.

For simplicity, we will refer to the process

as a set of five sequential activities that iterate

in this course intro.