# Wat is data science

Volgens Berkely houdt datascience het volgende in:

“Effective data scientists are able to

* identify relevant questions,
* collect data from a multitude of different data sources,
* organize the information,
* translate results into solutions,
* and communicate their findings in a way that positively affects business decisions.”

Omdat Berkely’s betekenis vooral ingaat op de stappen in data science gebruiken we in dit verslag de volgende betekenis van Wikipedia:

“Data science is an **inter-disciplinary field** that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights from** many structural and unstructured **data**… Data science is a ‘concept to unify statistics, data analysis and their related methods’ in order to ‘understand and analyze actual phenomena’ with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science.”

Al de elementen van deze betekenissen zijn terug te vinden in de opdrachten in dit portfolio. Om te beginnen het inter-disciplinaire onderdeel. De discipline die het eerste opvalt is natuurlijk die van informatica: in data science wordt meestal geprogrammeerd. Daarnaast moet een data scientist ook wat af weten van wiskunde en het wetenschappelijke proces om goede conclusies te kunnen trekken. Met wiskunde maken we van data een feit en met kennis van het wetenschappelijke proces kan daar een goede conclusie getrokken worden. Neem als voorbeeld de analyse van correlatie in opdracht 14: Eerst moet de data opgehaald met behulp van code. Daarna moet er een hypothese opgesteld worden (wetenschappelijk proces) waarna er met een wiskundige functie (chi squared test) gekeken wordt of deze hypothese verwerpen of aangenomen kan worden.

We zien in dit voorbeeld ook meteen het doel van datascience: kennis en inzicht met behulp van data. In opdracht 14 is mijn hypothese dat kleur en giftigheid gecorreleerd zijn. Door middel van datascience weten we uiteindelijk dat de kans hierop zeer hoog is. We kunnen onze hypothese dus aannemen. Kennis van het wetenschappelijk proces is ook belangrijk om ervoor te zorgen dat we begrijpen dat correlatie niet hetzelfde is als causaliteit.