Statistical Inference and Data Science Process

Statistical Inference

- The world we live in is complex, random, and uncertain.
- It’s one big data-generating machine.
- Data represents the traces of the real-world processes
- Your data collection/sampling is subjective
- You want to understand and describe the underlying processes

Statistical Inference - sources of uncertainty

- Two sources of randomness and uncertainty
  - From the data generating process
  - From the data collection method

Statistical Inference - modeling

- From world you get data
- From data you create a model
- Using mode you describe the world

This is called statistical inference

What is a model?

Humans try to understand the world around them by representing it in different ways.

A model is our attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological, or mathematical.

Artificial Intelligence and Machine Learning

- When you not interested to understand and explain the underlying process but to solve a particular problem
- E.g. prediction
- You might get additional insights that can improve your understanding from your work but the main focus is

Populations and Samples

- Population - all instances
- Sample - a subset of the population

But how do you build a model?

- You have to make underlying assumptions
- You perform EDA to get intuition
- There are no global standards

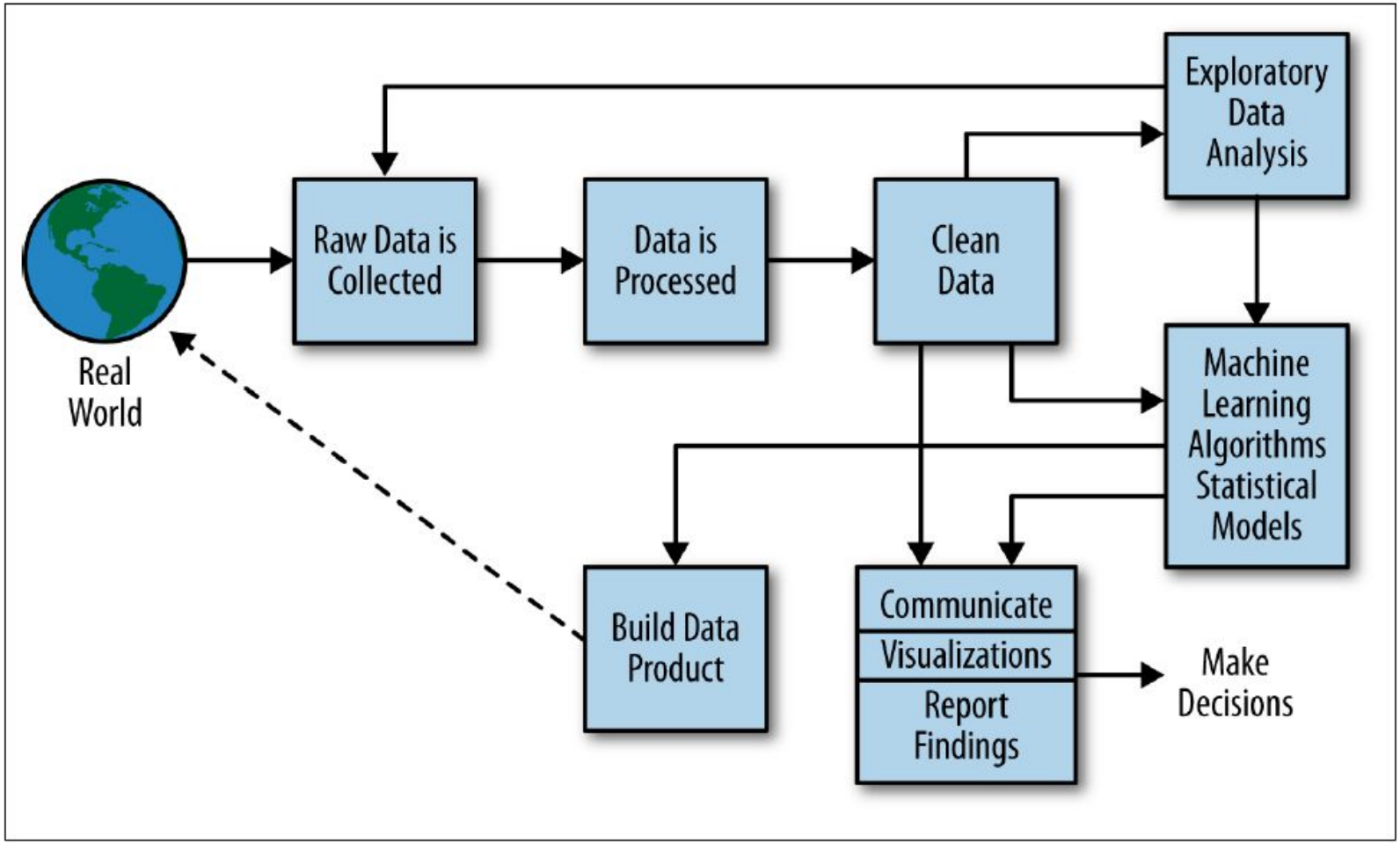

Figure 2-2. The data science process

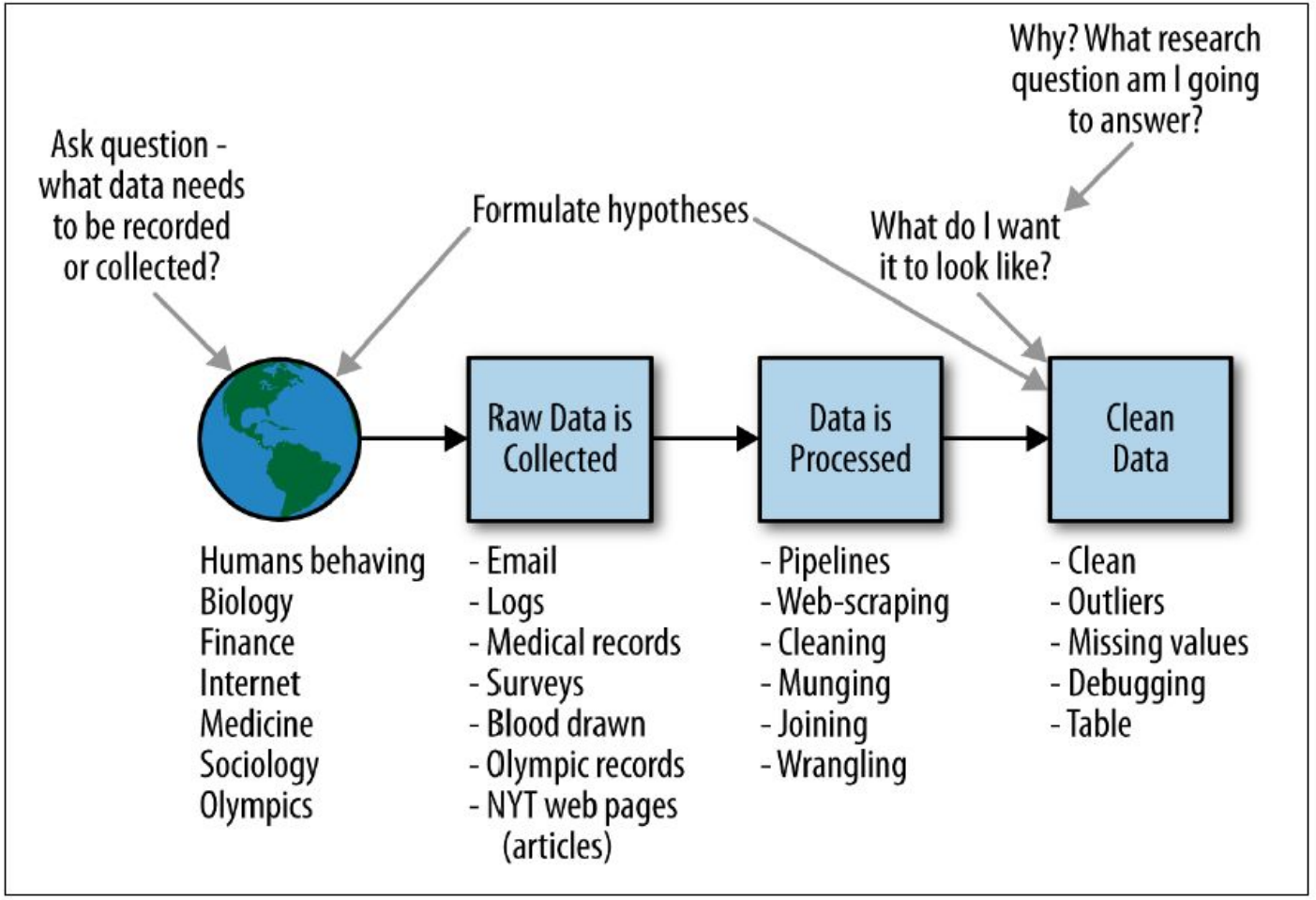

Figure 2-3. The data scientist is involved in every part of this process

References

1. Doing Data Science, Straight Talk From The Frontline by Cathy O'Neil and Rachel Schutt - Chapter 2