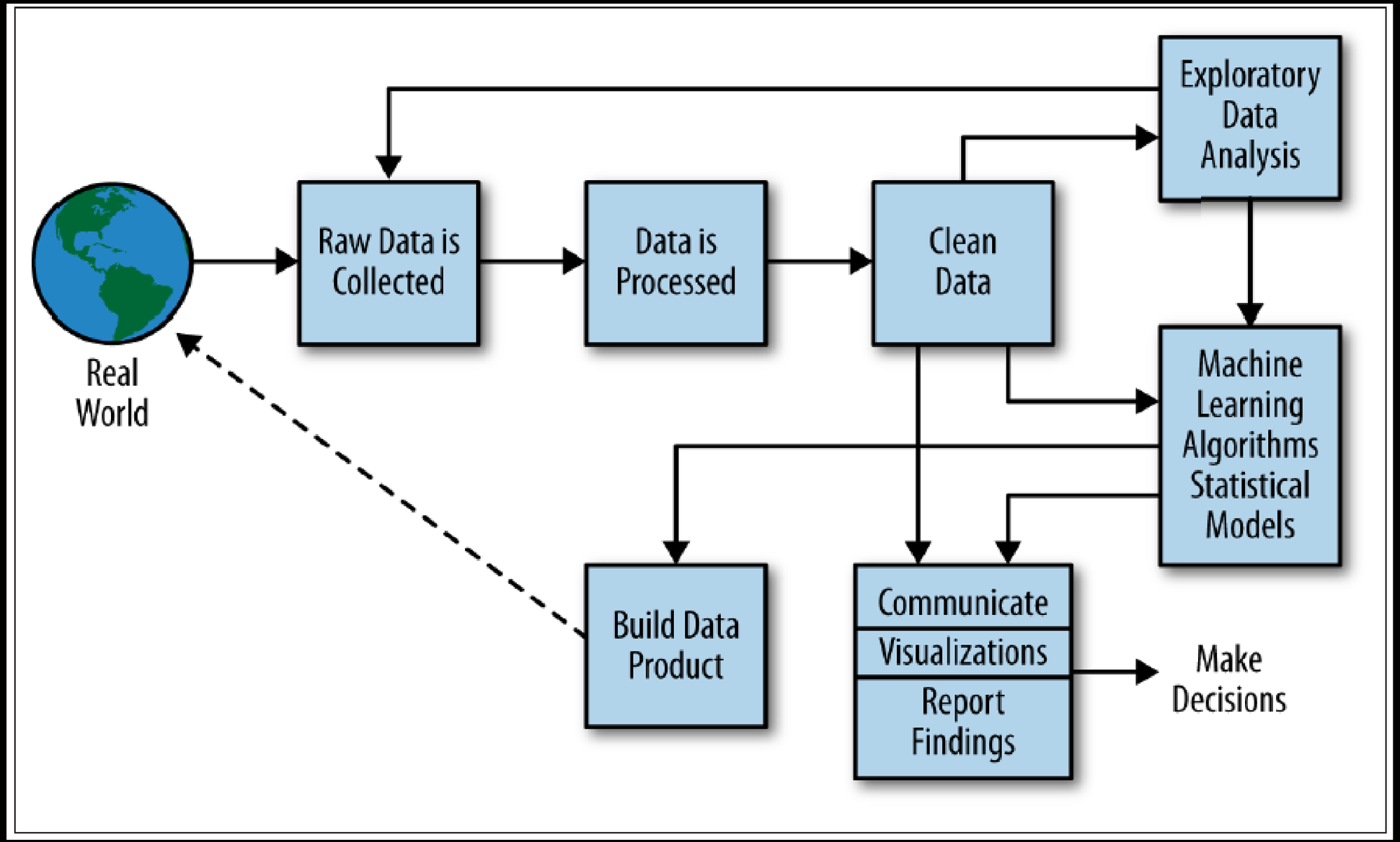
THE DATA SCIENCE PROCESS

The Data Science Process



REAL WORLD

Inside the Real World are lots of people busy at various activities. Some people are using Instagram, others are doing sports, and there are spammers sending spam. Say **we have data on one of these things**.

RAW DATA IS COLLECTED → CLEAN DATA

Specifically, we'll start with raw data—logs, records, or emails. We want to process this to make it clean for analysis.

So we build and use pipelines of data munging: **joining, scraping, wrangling,** or whatever you want to call it. To do this we use tools such as Python, or R.

Eventually we **get the data down to a nice format**, like something with columns:

name | event | year | gender | event time

EXPLORATORY DATA ANALYSIS (EDA)

Once we have this clean dataset, we should be doing some kind of EDA. While doing EDA, we may realize that **it isn't actually clean** because of **duplicates, missing values, absurd outliers, and data that wasn't actually logged or incorrectly logged**. If that's the case, we **may have to go back to collect more data, or spend more time cleaning the dataset**.

MODEL

Next, we design our model to use some algorithm like k-nearest neighbor (k-NN), linear regression, Naive Bayes, or something else. T**he model we choose depends on the type of problem we're trying to solve**, of course, which could be a **classification** problem, a **prediction** problem, or a basic **description** problem.

COMMUNICATE

We then can interpret, visualize, report, or communicate our results. This could take the form of **reporting the results up to our boss or coworkers**, or publishing a paper in a journal and going out and giving academic talks about it.

BUILD DATA PRODUCT (1)

Alternatively, our goal may be to build or prototype a “data product”;
e.g., a spam classifier, or a search ranking algorithm, or a recommendation system.
Now the key here that makes data science special and distinct from statistics is that this data product then **gets incorporated back into the real world**, then, users interact with that product, and that generates more data, which creates a feedback loop.

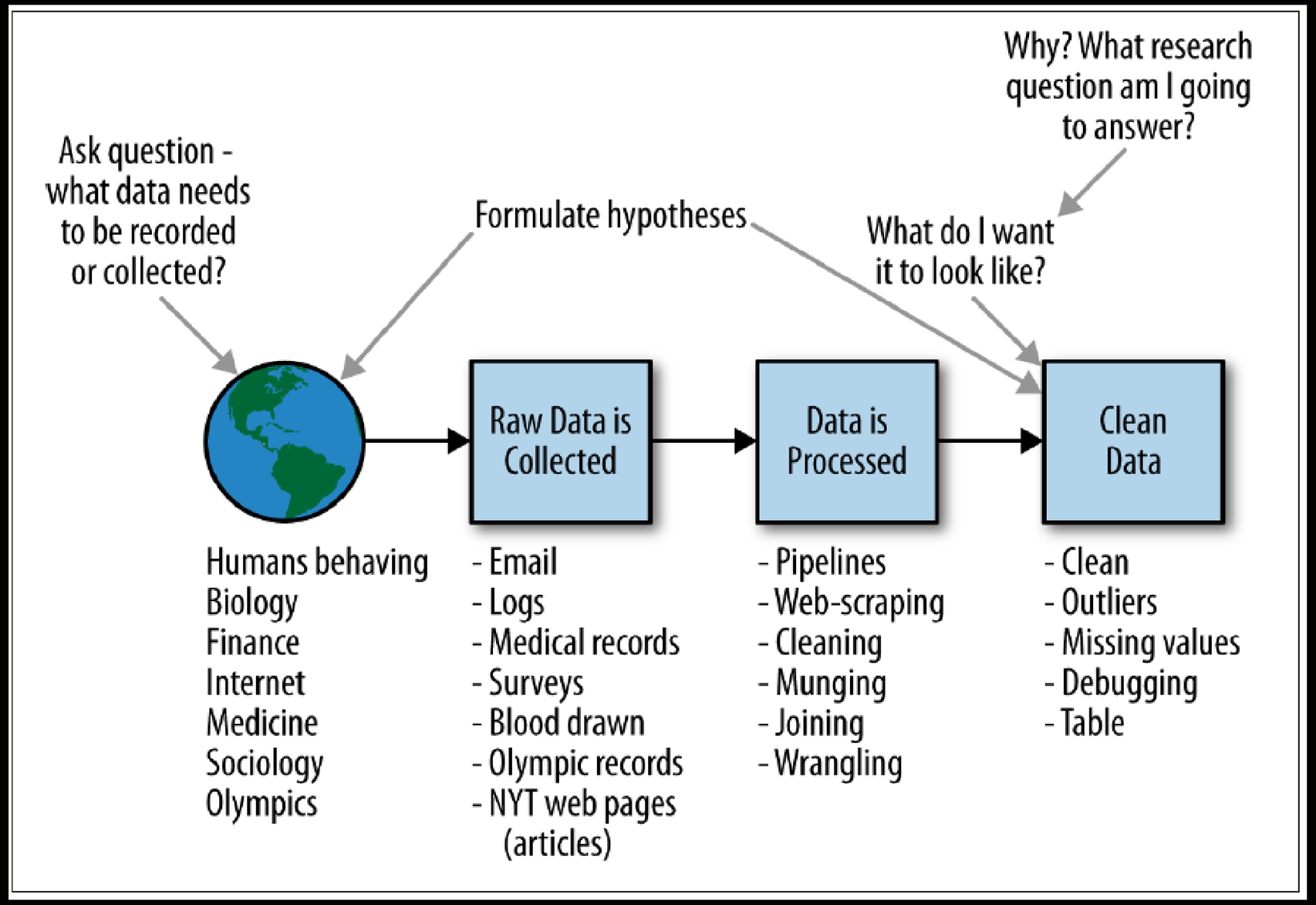
BUILD DATA PRODUCT (2)

Take this loop into account in any analysis you do **by adjusting for any biases your model caused**. Your **models are not just predicting the future, but causing it**!

Previous model so far seems to suggest this will all magically happen without human intervention.

By “human” here, we mean “data scientist.” **Someone has to make the decisions about what data to collect, and why**. That person needs to be **formulating questions and hypotheses** and **making a plan for how the problem will be attacked**. And that someone is the data scientist or our beloved data science team.

A Data Scientist's Role in This Process



WORDS OF INSPIRATION

DATA SCIENCE PROCESS
=
KERANGKA SKRIPSI

Connection to the Scientific Method

We can think of the data science process as an extension of or variation of the scientific method:

- Ask a question.
- Do background research.
- Construct a hypothesis.
- Test your hypothesis by doing an experiment.
- Analyze your data and draw a conclusion.
- Communicate your results.

In both the data science process and the scientific method, not every problem requires one to go through all the steps, but almost all problems can be solved with some combination of the stages. For example, if your end goal is a data visualization (which itself could be thought of as a data product), it's possible you might not do any machine learning or statistical modeling, but you'd want to get all the way to a clean dataset, do some exploratory analysis, and then create the visualization.

TUGAS INDIVIDU

1. Temukan permasalahan yang terjadi di sekitar Anda. Jelaskan dan pahami secara mendetail proses bisnis/alur kejadiannya.
2. Tentukan data apa saja yang Anda butuhkan untuk proses analisis, dan langkah apa saja yang akan Anda lakukan untuk mengumpulkan data yang tersebut.
3. Temukan contoh data yang telah ada dalam berbagai situs yang tersedia sebagai dasar analisis selanjutnya. Pahami data tersebut (variabel, jumlah data, pre-processing/ langkah data cleaning apa saja yang kira-kira akan dibutuhkan)
4. Jelaskan solusi apa yang akan Anda tawarkan dari permasalahan yang anda temukan dan Hipotesis Anda.
5. Temukan minimal 1 paper/ jurnal (Internasional/Nasional) yang pernah menyelesaikan permasalahan yang Anda rumuskan. Karya tulis tersebut bisa membahas mengenai solusi yang sama, atau solusi yang serupa (jika solusi yang anda tawarkan ternyata orisinil atau belum pernah dilakukan sebelumnya). Jelaskan model yang digunakan untuk menyelesaikan permasalahan tersebut.

KELAS A

Berikan jawaban dari poin-poin soal pada slide sebelumnya dalam bentuk essay dengan format DS-A_TUGAS2_NIM.pdf

Upload jawaban di link berikut : bit.ly/DS-A_TUGAS2

Deadline pengumpulan tugas : Minggu, 23 Februari 2020. jam 23:55

KELAS B

Berikan jawaban dari poin-poin soal pada slide sebelumnya dalam bentuk essay dengan format DS-B_TUGAS2_NIM.pdf

Upload jawaban di link berikut : bit.ly/DS-B_TUGAS2

Deadline pengumpulan tugas : Minggu, 23 Februari 2020. jam 23:55

KELAS C

Berikan jawaban dari poin-poin soal pada slide sebelumnya dalam bentuk essay dengan format DS-C_TUGAS2_NIM.pdf

Upload jawaban di link berikut : bit.ly/DS-C_TUGAS2

Deadline pengumpulan tugas : Minggu, 23 Februari 2020. jam 23:55