STATISTICAL INFERENCE, AND EXPLORATORY DATA ANALYSIS

DATA GENERATING PROCESS

can you give an example of one of them ?

DATA GENERATING PROCESS

Data represents the traces of the real-world processes.
You, the data scientist, the observer, are turning the world
into data, and this is an utterly **subjective**, not objective, process.

STATISTICAL INTERFERENCE

from the world to the data, and then from the data back to the world

STATISTICAL INTERFERENCE

to simplify those captured traces **into something more comprehensible**, to something that somehow captures it all in a much **more concise way**, and that something could be **mathematical models or functions of the data, known as statistical estimators**.

POPULATION VS SAMPLE VS OBSERVATION

POPULATION ?

POPULATION VS SAMPLE VS OBSERVATION

STATISTICAL INFERENCE POPULATION

in statistical inference population **isn’t used to simply describe only people**. It could be **any set of objects or units**, such as tweets or photographs

POPULATION VS SAMPLE VS OBSERVATION

OBSERVATION

if we could **measure the characteristics or extract characteristics of all those objects**, we’d have a complete set of observations.

POPULATION VS SAMPLE VS OBSERVATION

SAMPLE

subset of the units of size N in order to examine the observations to draw conclusions and make inferences about the population.

EXAMPLE POPULATION

all emails sent last year by employees at a huge corporation, BigCorp

EXAMPLE OBSERVATION

a list of things: the sender’s name, the list of recipients, date sent, text of email, number of characters in the email, number of sentences in the email, and the length of time until first reply

Decide carefully.

SAMPLE

be aware of this **sampling mechanism** because it can introduce **biases into the data**, and distort it, so that the subset is not a “mini-me” shrunk-down version of the population. **Once that happens, any conclusions you draw will simply be wrong and distorted**.

SAMPLE :

(1) make a **list of all the employees** and select **1/10th of those people at random**, then take all the email they ever sent

(2) sample 1/10th of all email sent each day at random

OBSERVATION VS SAMPLE

ESTIMATE : DISTRIBUTION OF EMAILS SENT BY ALL INDIVIUALS AT BIGCORP

USING

How many email messages each person sent.

BIG DATA

POPULATIONS AND SAMPLE OF BIG DATA

we can record all users' actions all the time,
don't we observe everything?
Is there really still this notion of population and sample?

01

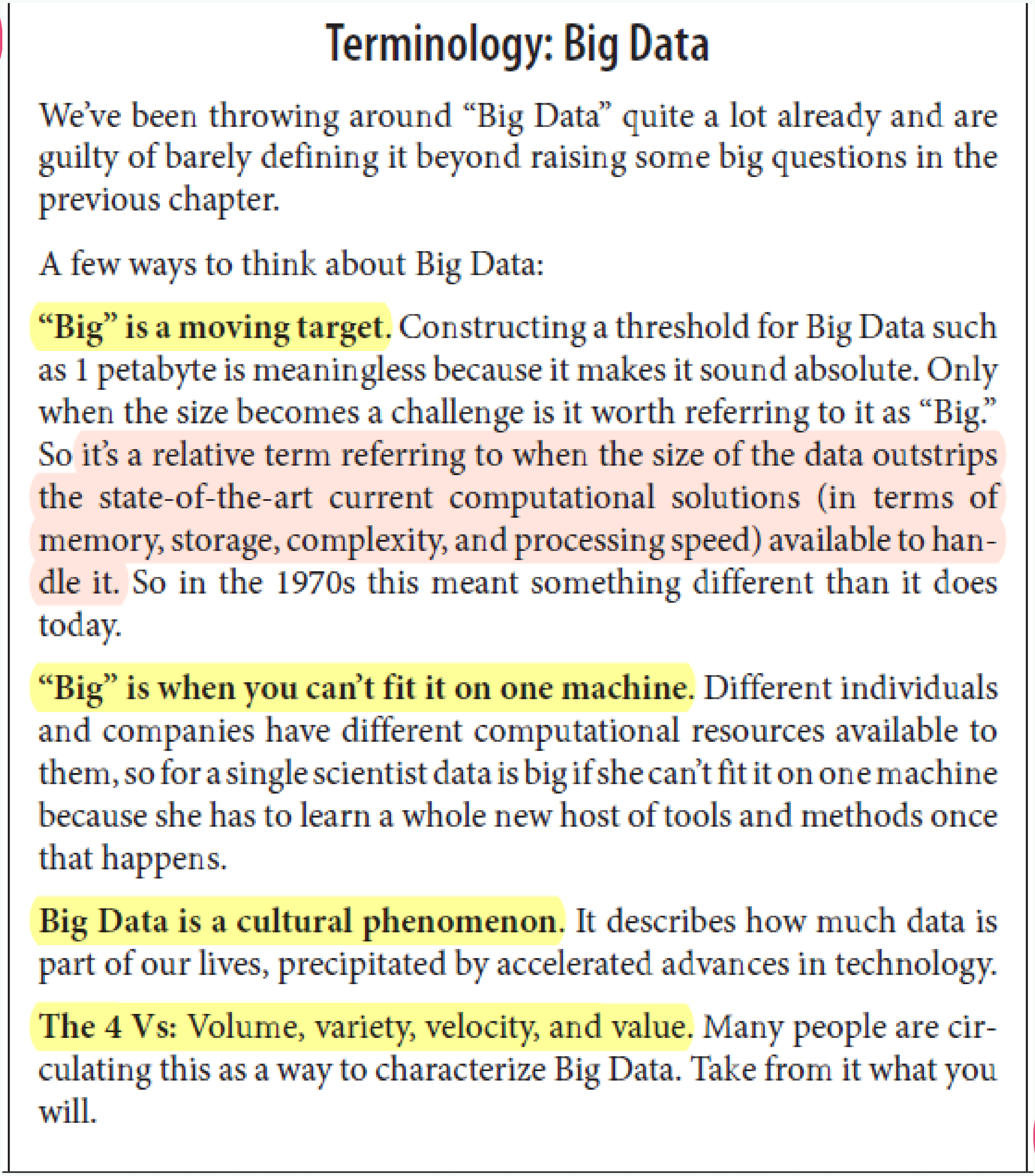
SAMPLING SOLVES SOME ENGINEERING CHALLENGES

How much data you need at hand really depends on what your goal is.

02

BIAS

if you didn’t have context and know about **something**, you wouldn’t know enough to interpret this data properly



Terminology: Big Data

We've been throwing around "Big Data" quite a lot already and are guilty of barely defining it beyond raising some big questions in the previous chapter.

A few ways to think about Big Data:

"Big" is a moving target. Constructing a threshold for Big Data such as 1 petabyte is meaningless because it makes it sound absolute. Only when the size becomes a challenge is it worth referring to it as "Big." So it's a relative term referring to when the size of the data outstrips the state-of-the-art current computational solutions (in terms of memory, storage, complexity, and processing speed) available to handle it. So in the 1970s this meant something different than it does today.

"Big" is when you can't fit it on one machine. Different individuals and companies have different computational resources available to them, so for a single scientist data is big if she can't fit it on one machine because she has to learn a whole new host of tools and methods once that happens.

Big Data is a cultural phenomenon. It describes how much data is part of our lives, precipitated by accelerated advances in technology.

The 4 Vs: Volume, variety, velocity, and value. Many people are circulating this as a way to characterize Big Data. Take from it what you will.

BIG DATA POPULATION

BIG DATA CAN MEAN BIG ASSUMPTIONS

the new approach of Big Data is letting “**N=ALL**.”

BIG DATA POPULATION

“N=ALL” OFTEN GETS TRANSLATED INTO THE IDEA THAT DATA IS OBJECTIVE

data is objective or that “**data speaks**”

EXAMPLE

Say you decided to **compare women and men with the exact same qualifications** that have been hired in the past, but then, looking into what happened next you learn that those **women have tended to leave more often, get promoted less often, and give more negative feedback on their environments** when compared to the men.

Your model might be likely to hire the man over the woman next time the two similar candidates showed up, rather than looking into the possibility that the company doesn’t treat female employees well.

BIG DATA POPULATION

DATA DOESN'T SPEAK FOR ITSELF. DATA IS JUST A QUANTITATIVE, PALE ECHO OF THE EVENTS OF OUR SOCIETY.

ignoring causation can be a flaw, rather than a feature. **Models** that ignore causation **can add to historical problems instead of addressing them**

What is a Model?

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

MODELLING

Before you get too involved with the data and start coding, it's useful to draw a picture of what you think the underlying process might be with your model.

What comes first?
What influences what?
What causes what?
What's a test of that?

MODELLING

Some prefer to express these kinds of relationships in terms of math

if you have two columns of data, x and y, and you think there's a linear relationship, you'd write down y = β0 +β1 x.

MODELLING

Other people prefer pictures and will first draw a diagram of data flow

possibly with arrows, showing how things affect other things or what happens over time. This gives them an abstract picture of the relationships before choosing equations to express them.

How do you build a Model?

The best thing to do is start simply and then build in complexity. Do the dumbest thing you can think of first. It’s probably not that dumb.

MODELLING

REMEMBER, IT'S ALWAYS GOOD TO START SIMPLY.

There is a trade-off in modeling between simple and accurate. Simple models may be easier to interpret and understand. Oftentimes the crude, simple model gets you 90% of the way there and only takes a few hours to build and fit, whereas getting a more complex model might take months and only get you to 92%.

EXPLORATORY DATA ANALYSIS (EDA)

as the first step toward building a model. “easiest” and lowest level data analysis process

EXPLORATORY DATA ANALYSIS (EDA)

as the first step toward building a model. “easiest” and lowest level data analysis process

EDA IS A CRITICAL PART OF THE DATA SCIENCE PROCESS

Characteristics

NO HYPOTHESIS & NO MODEL

“exploratory” aspect means that your understanding of the problem you are solving, or might solve, is changing as you go.

BASIC TOOLS: PLOTS, GRAPHS AND SUMMARY STATISTICS

Generally, it’s a method of systematically going through the data

MINDSET ABOUT RELATIONSHIP WITH THE DATA

to understand the data—gain intuition, understand the shape of it, and try to connect your understanding of the process of the data itself

EDA
VS
DATA VISUALIZATION

EDA

is done toward the beginning of analysis, the graphics are solely done for you to understand what’s going on

DATA VISUALIZATION

is done toward the end to communicate one’s findings.

EXAMPLE

Suppose you are trying to develop a ranking algorithm that ranks content that you are showing to users. To do this you might want to develop a notion of "popular."

EXAMPLE

Before, you need to decide how to quantify popularity (which could be, for example, highest frequency of clicks, or the post with the most number of comments, or comments above some threshold, or some weighted average of many metrics). **You need to understand how the data is behaving**.

iterative cycle

EDA STEPS

1

Generate questions about your data.

2

Search for answers by visualising, transforming, and modelling your data.

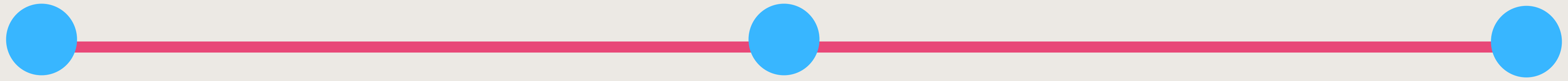
3

Use what you learn to refine your questions and/or generate new questions.

DS - A

TUGAS 1

BIT.LY/DS-A_TUGAS1



1

PPT dan Contoh bisa didownload di: github.com/DS-ifupnyk/kelas2020

2

Buat EDA seperti contoh studi kasus. Data bisa diambil dari berbagai sumber. Minimal 10.000 baris

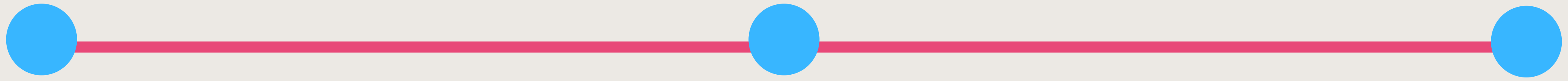
3

Hasil EDA dalam bentuk narasi/essay (.pdf) diupload maksimal Kamis, 13 Februari 2020 jam 23.55

DS - B

TUGAS 1

BIT.LY/DS-B_TUGAS1



1

PPT dan Contoh bisa didownload di: github.com/DS-ifupnyk/kelas2020

2

Buat EDA seperti contoh studi kasus. Data bisa diambil dari berbagai sumber. Minimal 10.000 baris

3

Hasil EDA dalam bentuk narasi/essay (.pdf) diupload maksimal Rabu, 12 Februari 2020 jam 23.55

DS - C

TUGAS 1

BIT.LY/DS-C_TUGAS1

1

PPT dan Contoh bisa didownload di: github.com/DS-ifupnyk/kelas2020

2

Buat EDA seperti contoh studi kasus. Data bisa diambil dari berbagai sumber. Minimal 10.000 baris

3

Hasil EDA dalam bentuk narasi/essay (.pdf) diupload maksimal Kamis, 13 Februari 2020 jam 23.55