Statistics Bootcamp using R

DAY 1 INTRODUCTION TO STATISTICS IN BUSINESS

1.1 BASIC VOCABULARY OF STATISTICS & DATA TYPES

GU Zhan (Sam)
Institute of Systems Science
National University of Singapore

issgz@nus.edu.sg

Agenda

Day 1: Introduction to Statistics in Business

- **Basic Vocabulary of Statistics & Data Types**
- Introduction to R
- Data Collection & Summarization

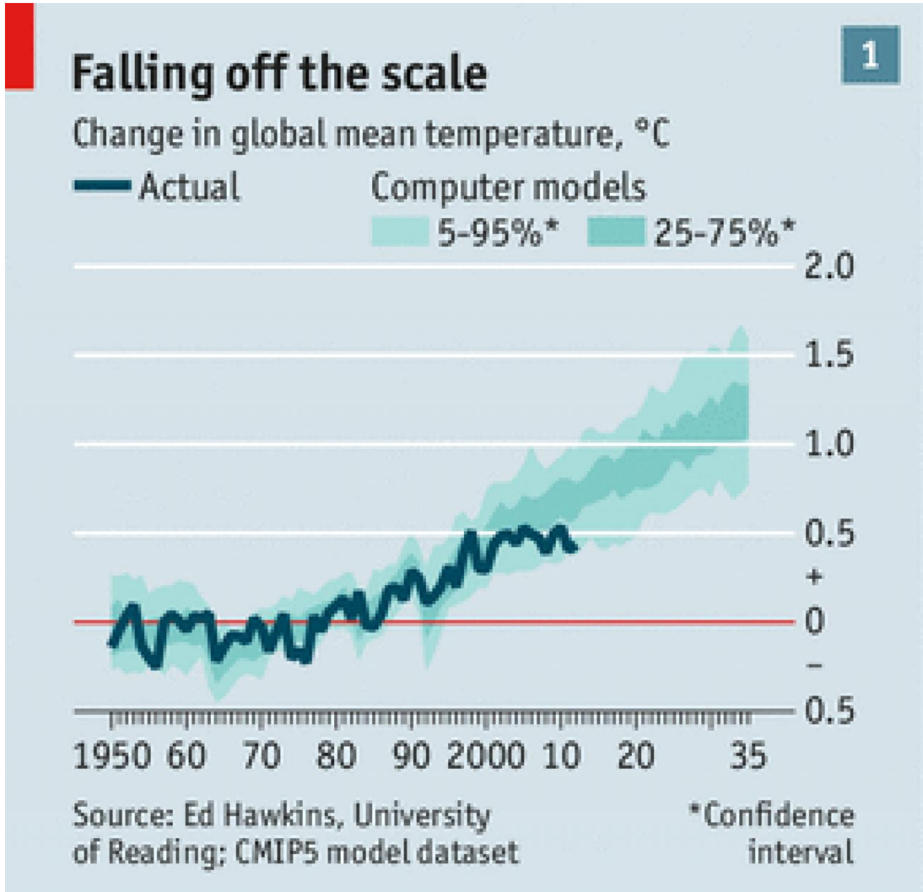
Learning objectives

- Understand statistic's business values
- Understand different data types
- Understand population and sample

How does statistics help in making business decisions?

Provide estimation

with various level of confidence



Source: https://www.economist.com/science-and-technology/2013/03/30/a-sensitive-matter

Make the right choice

by consulting rigorous math

- Coach A:
  25 students
  17 passed lifeguard test
  Cost S$1,200

68%

- Coach B:
  72 students
  57 passed lifeguard test
  Cost S$1,800

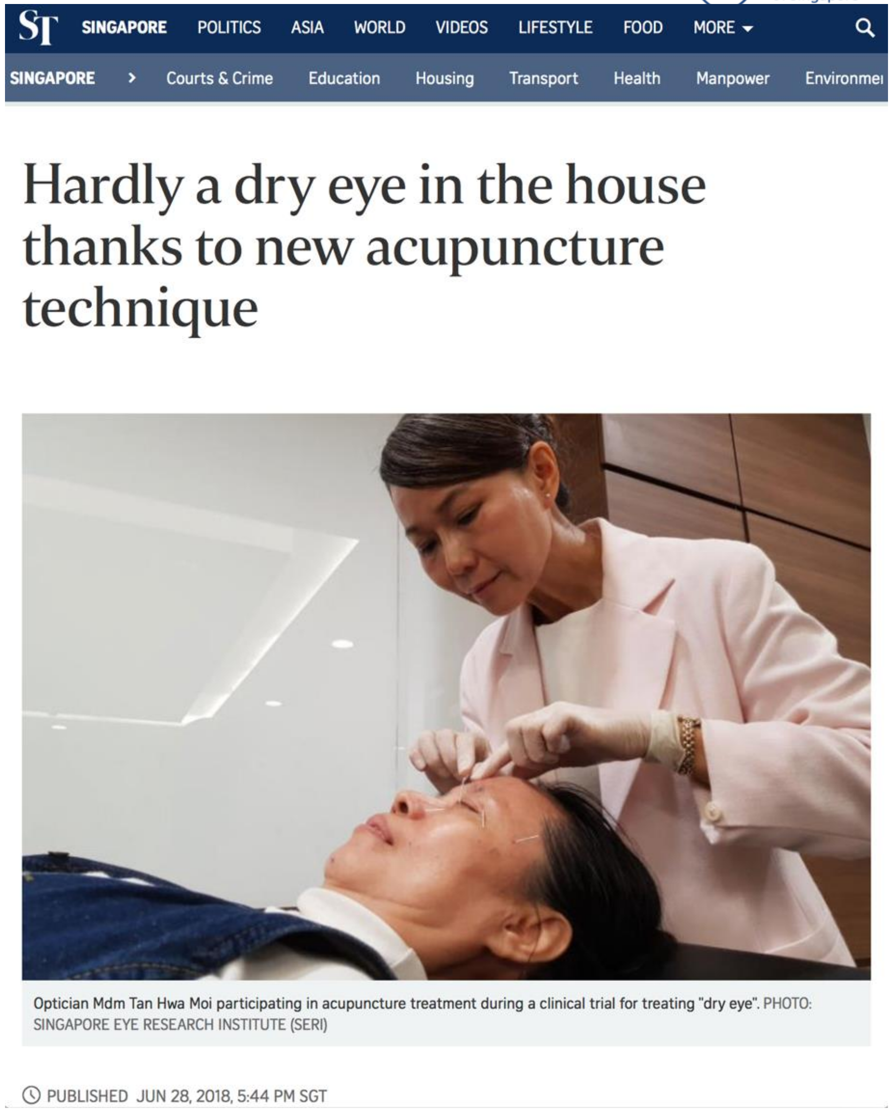
79%

Evaluate medical treatment effectiveness

by doing new study and verified by statistical method

Evaluate improvement before and after process changes

by doing new study and verified by statistical method



Hardly a dry eye in the house thanks to new acupuncture technique

Optician Mdm Tan Hwa Moi participating in acupuncture treatment during a clinical trial for treating "dry eye". PHOTO: SINGAPORE EYE RESEARCH INSTITUTE (SERI)

PUBLISHED JUN 28, 2018, 5:44 PM SGT

These companies use predictive modelling

predictive modelling need statistics

amazon

to recommend books

NETFLIX

to recommend shows

Predictive modelling: the process of developing a mathematical tool or model that generates an accurate prediction

citibank

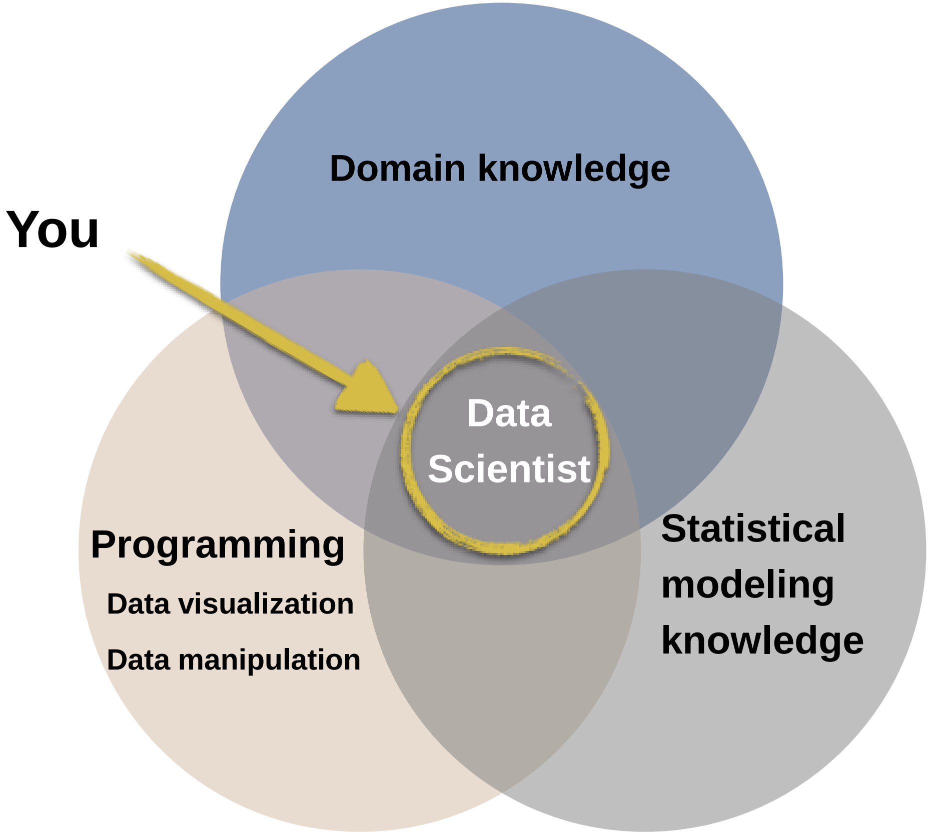
to detect fraud

tinder

to recommend a friend…

Career with statistics?

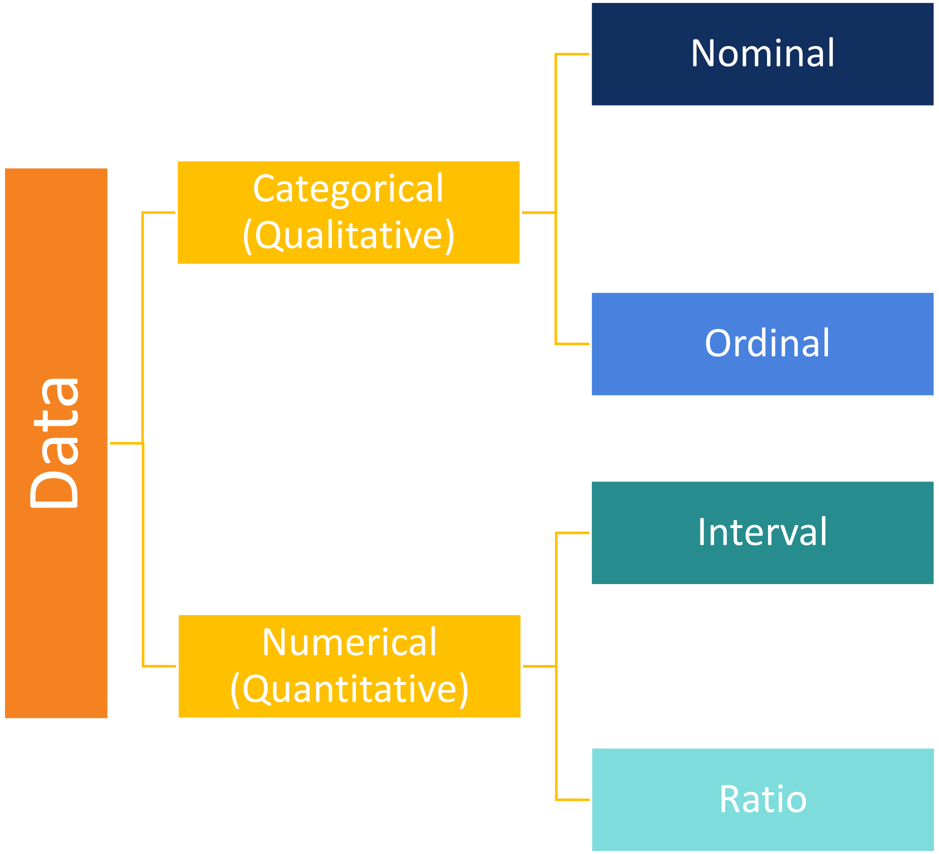
You just need these skills



Data Types

Know your data (structured*)

It comes in various forms



Source: 'Design for Information', by Isabel Meirelles

Qualitative: Nominal

Objects, names and concepts are examples of nominal data. The questions we ask about nominal data are what and where. Nominal data have no implicit quantitative relationship or inherent ordering. Because categorization plays a major role in manipulating nominal data, it is often called categorical data.

Genders;
Blood types;

Qualitative: Ordinal

Ordinal data can be arranged in a given order or rank, such that we can say which comes first or second, which is smaller or larger. Ordinal data provides the order, but not the degree of differences. For example, we might know which country ranks first in relation to apple exports, but not how much more compared to second place.

Floor levels;
Earthquake magnitudes;

Quantitative

Quantitative data can be numerically manipulated, such as with statistical method. Numerical data require that we ask questions of how much, e.g. the number of apple produced, the average size of apple and so on.

Quantitative data can be transformed into ordinal data by classing it

Interval: Temperature, pH, IQ
Ratio: Weight, Salary, GDP

*There is unstructured data like text, audio, video, IoT etc. which falls under big data

The numbers don’t know where they came from

Not all numbers are equal

7, 6, 4, 2, 9, 10
Time duration for 6 tasks (**ratio data**)

7, 6, 4, 2, 9, 10
6 high temperatures in Celsius from a Northeastern US city (**interval data**)

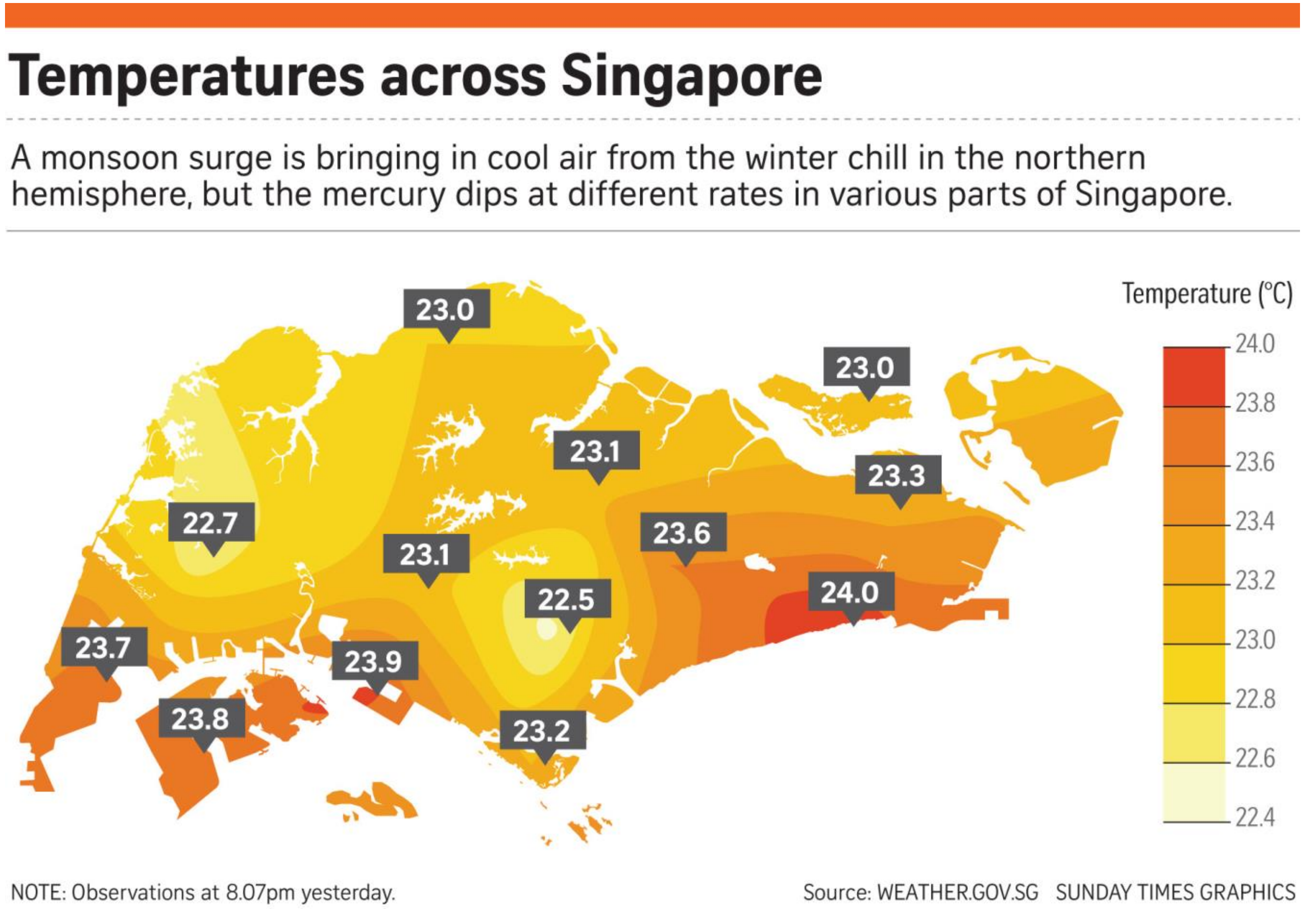
7, 6, 4, 2, 9, 10
6 responses to the likelihood to recommend the hotel (**ordinal data**)

7, 6, 4, 2, 9, 10
6 numbers at the back of football jerseys (**nominal data**)

Source: https://measuringu.com/interval-ordinal/

Cameron highland in Singapore

Guess the data type



Temperatures across Singapore

A monsoon surge is bringing in cool air from the winter chill in the northern hemisphere, but the mercury dips at different rates in various parts of Singapore.

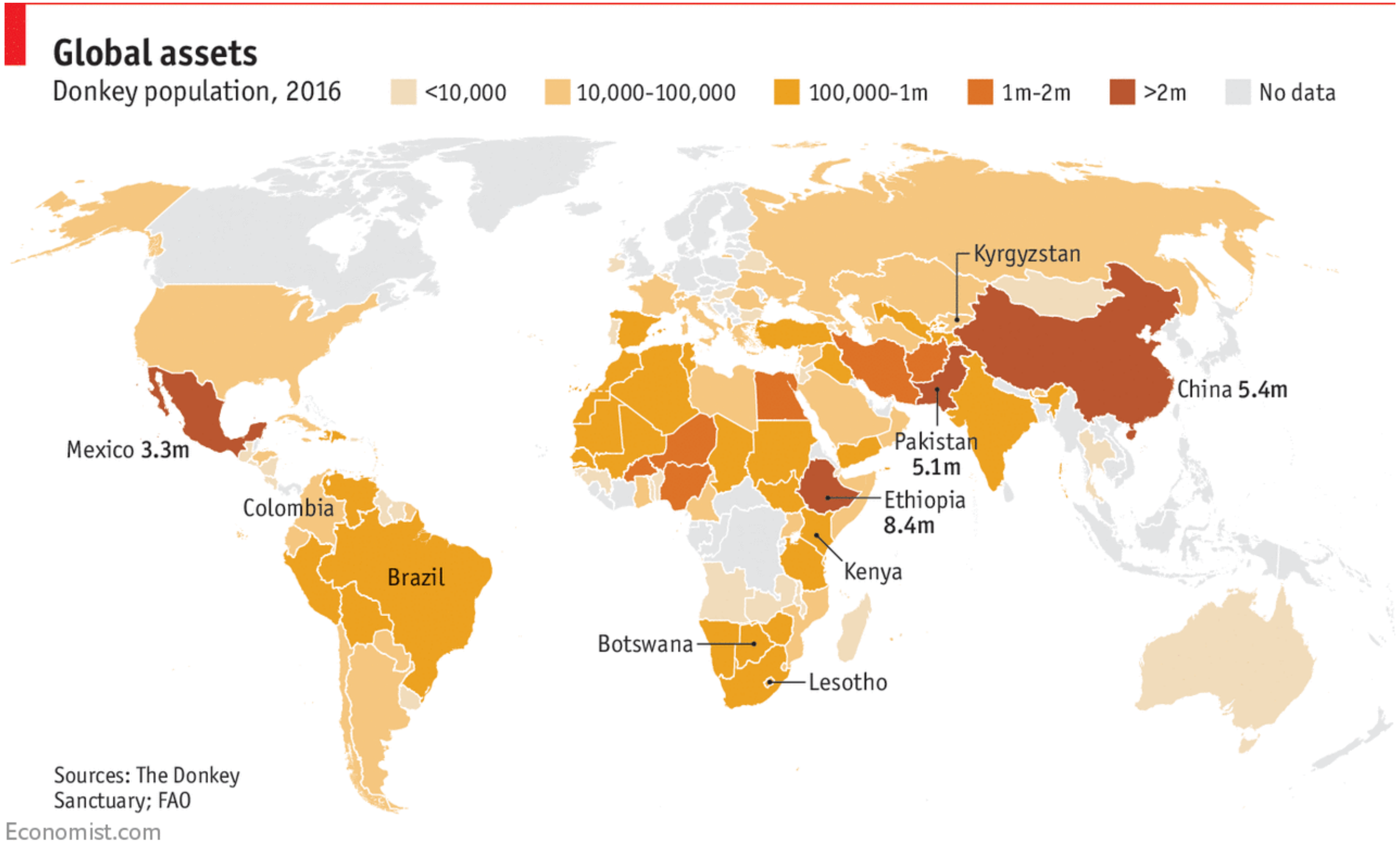
NOTE: Observations at 8.07pm yesterday.

Source: WEATHER.GOV.SG SUNDAY TIMES GRAPHICS

Source: https://www.gov.sg/news/content/the-straits-times---why-temperatures-vary-across-singapore

Where are the donkeys?

Guess the data type



Global assets

Donkey population, 2016

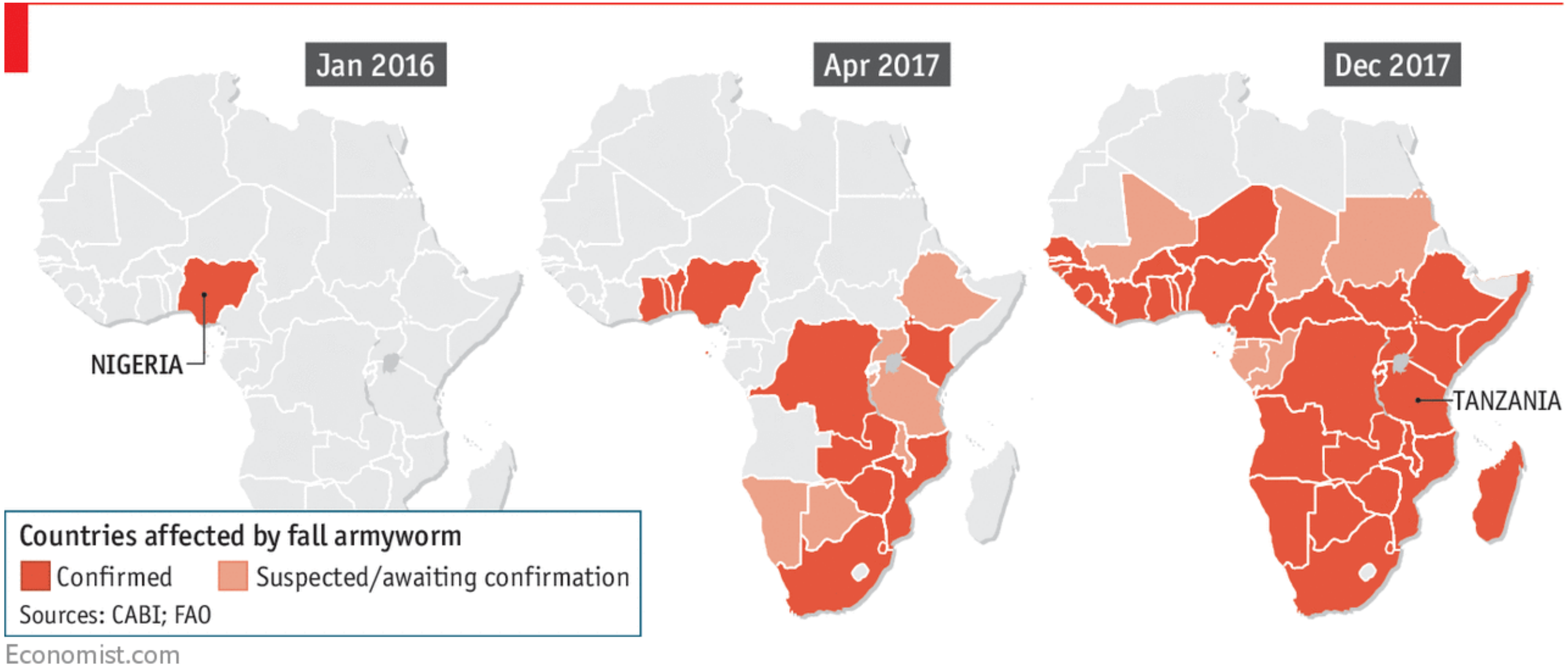
Sources: The Donkey Sanctuary; FAO

Economist.com

Source: https://www.economist.com/graphic-detail/2018/03/30/donkey-skins-are-the-new-ivory

An army of worm

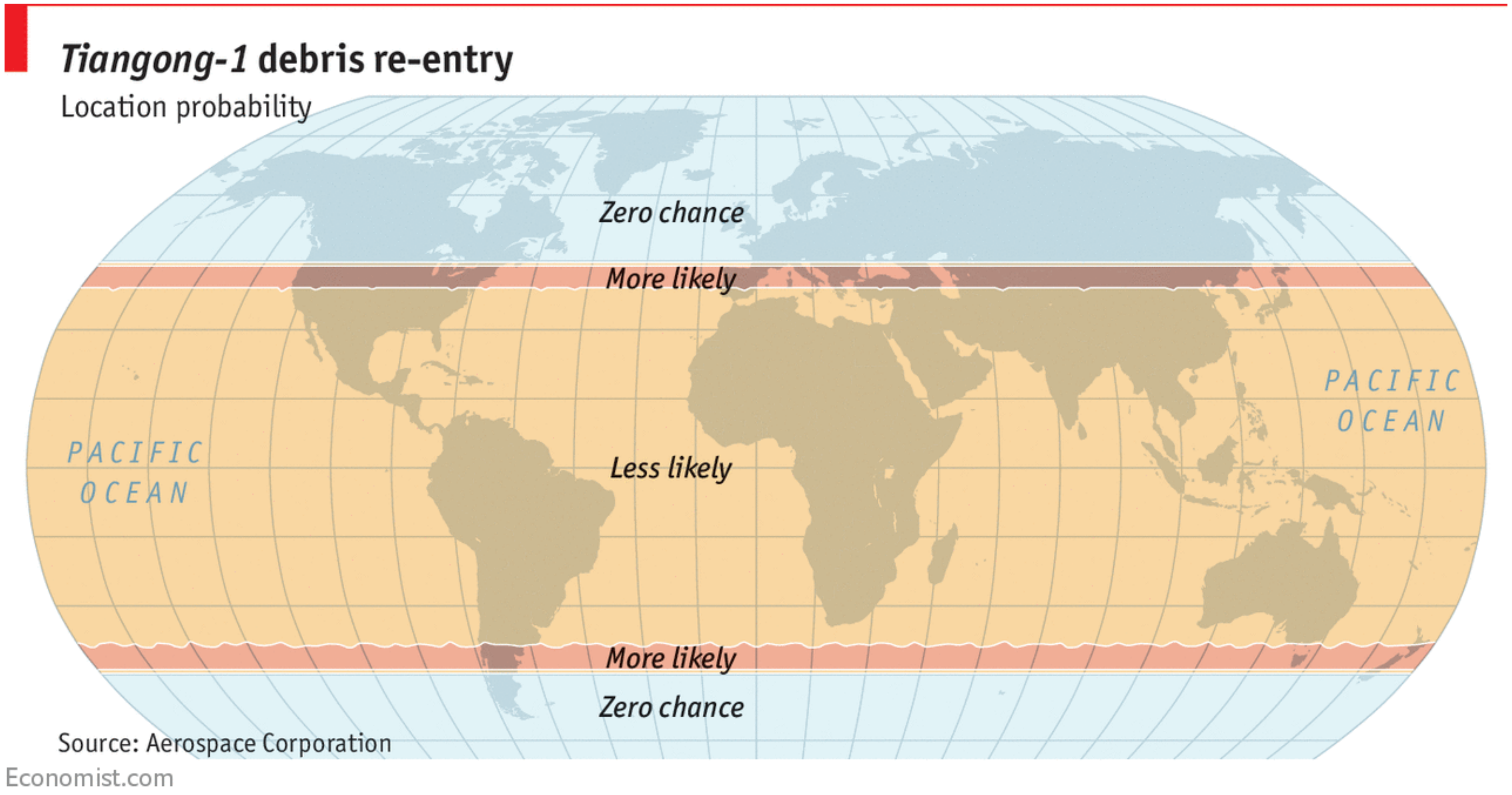
Guess the data type



Source: https://www.economist.com/graphic-detail/2018/01/24/an-army-of-worms-is-invading-africa

The space station came back

Guess the data type



Tiangong-1 debris re-entry

Location probability

Source: Aerospace Corporation

Economist.com

Source: https://www.economist.com/graphic-detail/2018/03/19/an-out-of-control-chinese-space-station-will-soon-fall-to-earth

Basic Vocabulary in Statistics: Population vs. Sample

Population

Sample

Source: https://www.questionpro.com/blog/simple-random-sampling/

Population

A population data set contained all members of a specified group.

Use ‘population’ when you know you have the entire population.
Or, use ‘population’ if you have a sample taken from a population, but you are only interested in this set of data (**descriptive analysis purpose**) and do not want to know anything about the population.

Eg: You are interested in literacy rate among women in Africa

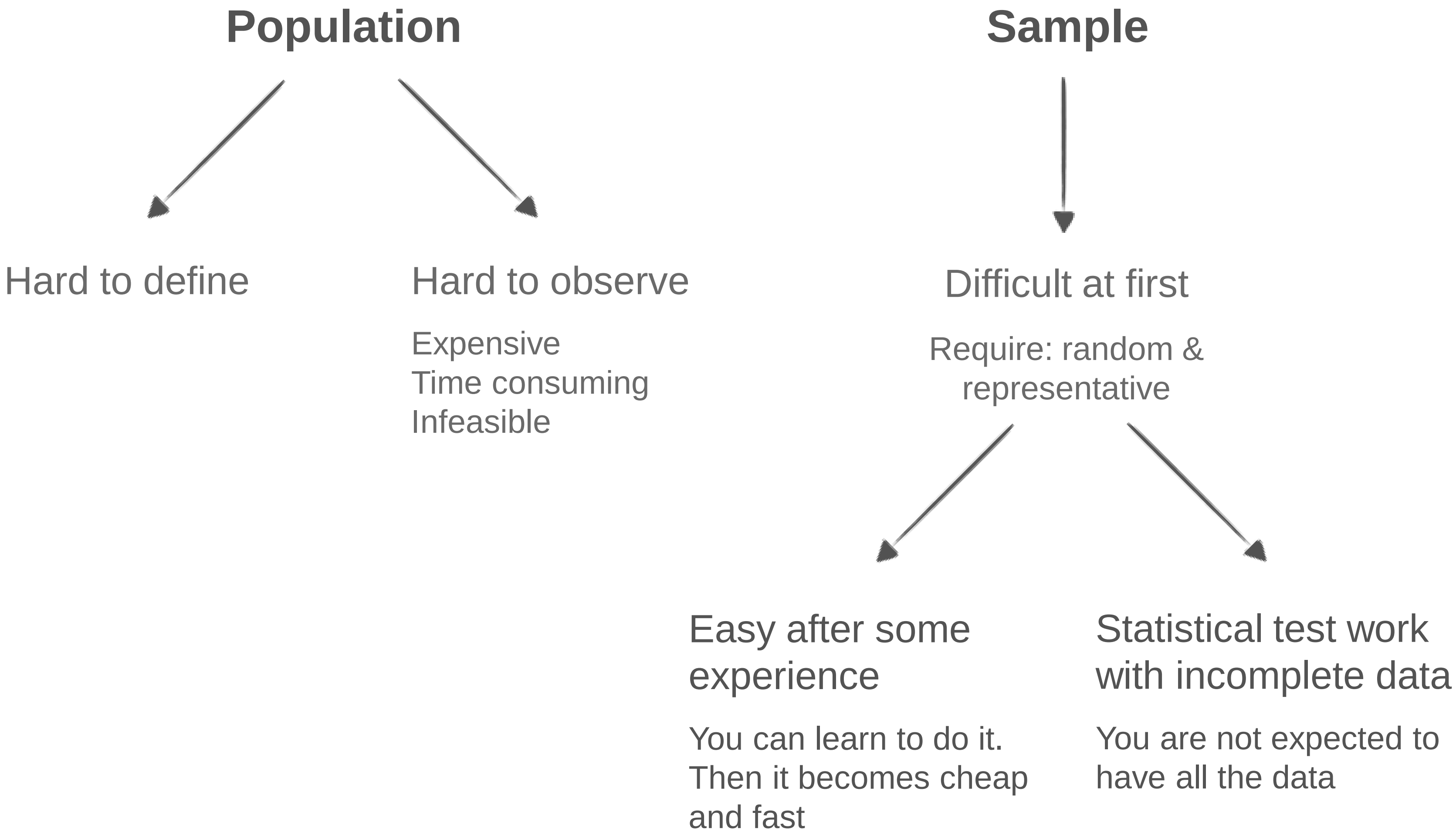
Sample

A sample data set contains a part, or a subset of a population.

Use ‘sample’ when have a data set taken from a population (the size of the sample < the size of the population), and you wish use this data set to understand or make estimation about the population (predictive analysis purpose).

Eg : You take random stratified sample from different African states in proportion with their women population to have the estimate on the literacy rate

Population vs. Sample



Population

Hard to define

Hard to observe

Expensive
Time consuming
Infeasible

Sample

Difficult at first

Require: random &
representative

Easy after some experience

You can learn to do it. Then it becomes cheap and fast

Statistical test work with incomplete data

You are not expected to have all the data

Source: https://www.youtube.com/watch?v=eIZD1BFfw8E

Summary

- Structured Data may come in following types:
  - Categorical (nominal, ordinal)
  - Numerical (interval, ratio)

- Concepts in statistics:
  - Population
  - Sample

End of Lecture Notes