A BRIEF INTRODUCTION TO

DATA SCIENCE

WHAT? HOW? WHY?

Alexander Koch — a.koch@maastrichtuniversity.nl

2019

WHAT IS DATA SCIENCE?

DATA noun *da•ta*

1. factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
2. information in digital form that can be transmitted or processed
3. information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

SCIENCE noun *sci•ence*

1. the state of knowing, knowledge as distinguished from ignorance or misunderstanding
2. knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method
   such knowledge or such a system of knowledge concerned with the physical world and its phenomena : NATURAL SCIENCE

DATA SCIENCE

This is why you're here

Extract knowledge and insights from structured and unstructured data

Interdisciplinary: mathematics, statistics, computer science, AI, and information science

No official definition

Buzzword: “Sexiest Job of the 21st Century” — *Harvard Business Review (2012)*

“Sexed-up term for statistics” — *Nate Silver (2013)*

YOUR QUESTIONS

Some answers to my initial e-mail:

- Where can I find public databases and how can I extract data from them?
- Where can I find tutorials and learning materials to learn about data science?
- What’s R and how can I use it?
- How can I visualise my data?
- How do I perform specific types of analyses, such as a prognostic or a differential analysis?
- …

A GENERAL THEME…

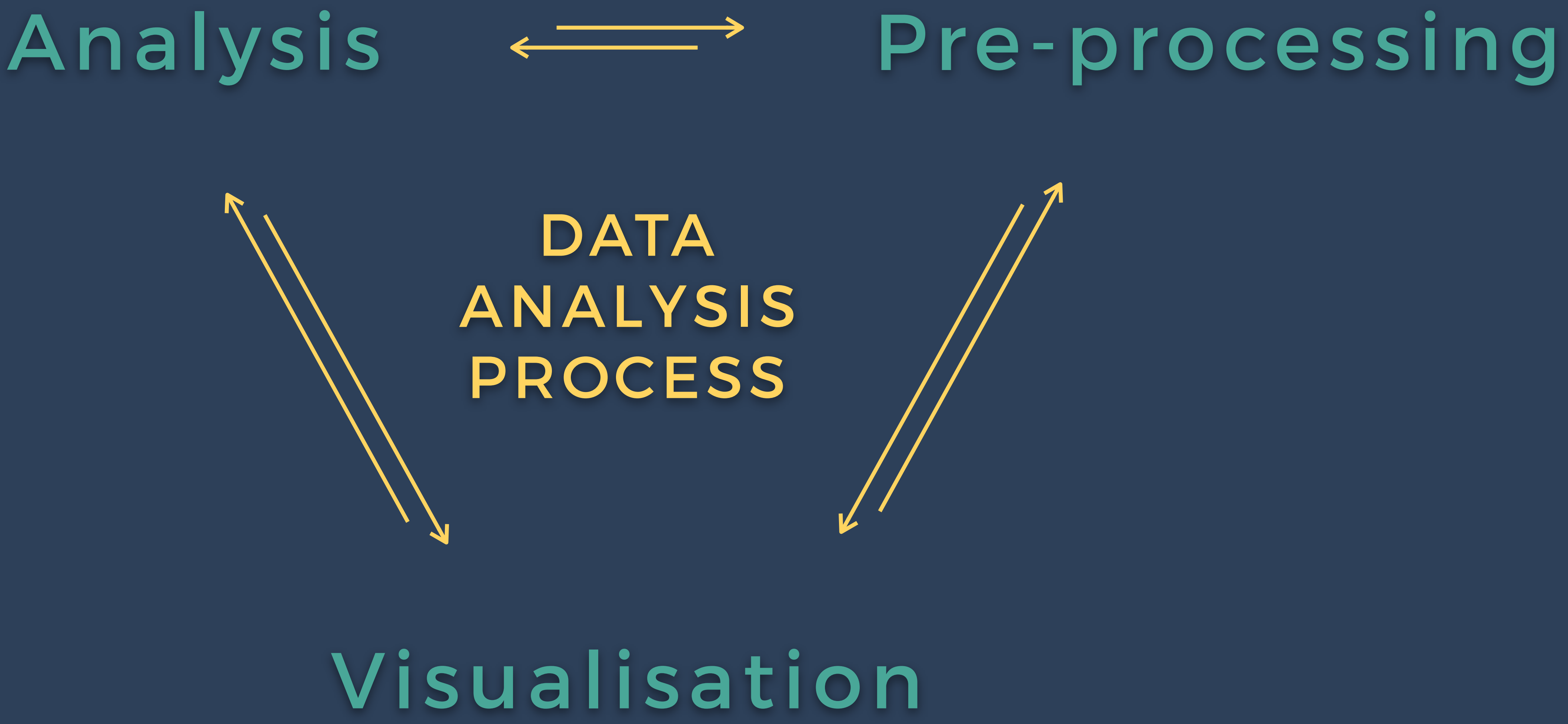
You have a research question that you'd like to answer or a hypothesis you'd like to test, so you…

…set up lab experiments and generate data…
…or find some interesting data online…

…but you don't know how to analyse it, how to extract the knowledge you're after.

WHAT'S THE PLAN?

- Summary of the data analysis process
- Brief overview of some public data portals
- Overview of commonly used tools
- List of useful learning resources
- Hands-on introduction to R

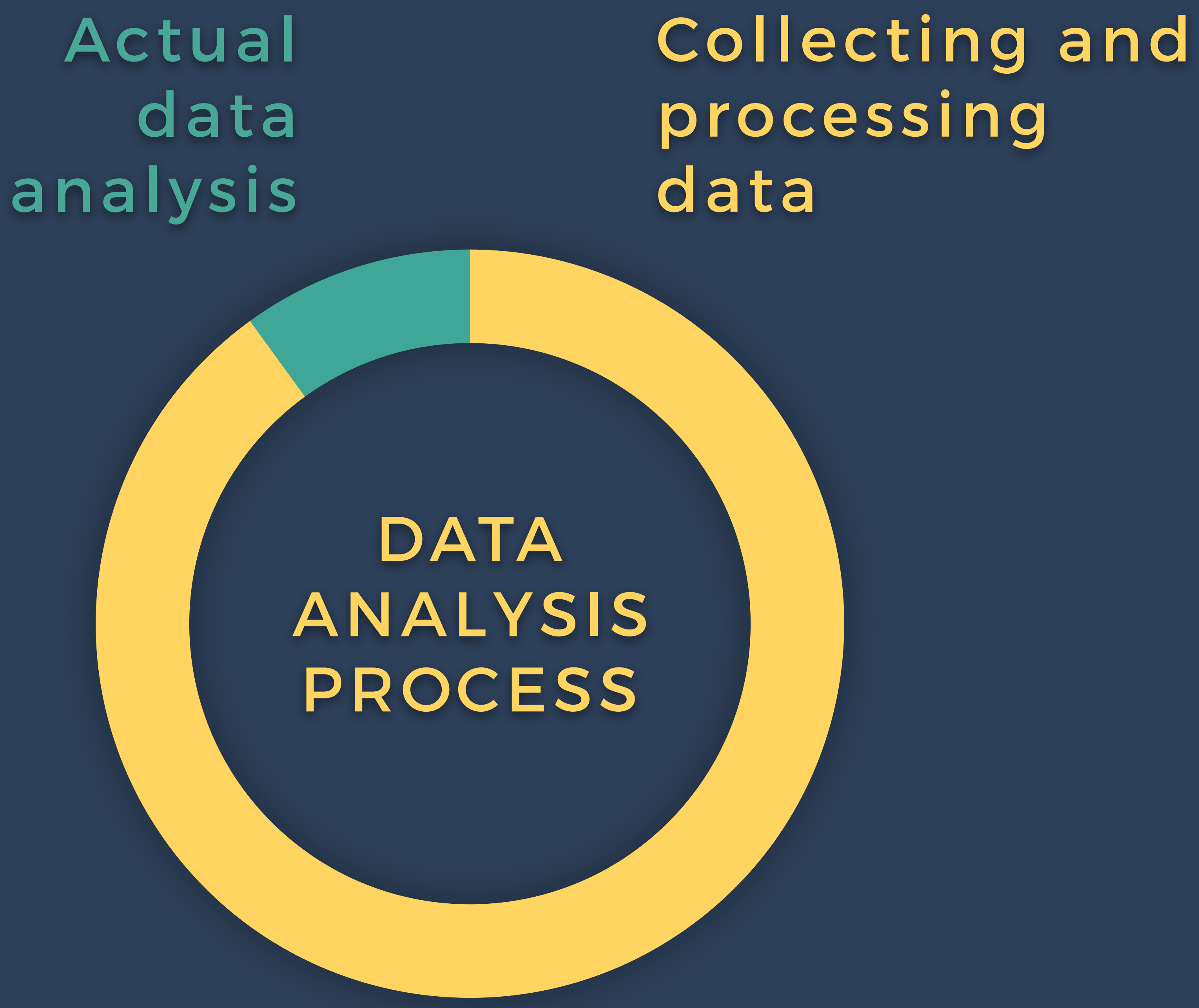


Analysis

Pre-processing

DATA
ANALYSIS
PROCESS

Visualisation


Actual data analysis
Collecting and processing data
DATA ANALYSIS PROCESS

WHERE'S THE DATA?

- Scientific publications
- NCBI's Gene Expression Omnibus (GEO)
  https://www.ncbi.nlm.nih.gov/geo/
  Array and sequencing-based functional genomics data
- Xena data hubs
  https://xenabrowser.net/hub/
  Cancer genomics data (mostly TCGA)
- Genomic Data Commons (GDC) data portal
  https://portal.gdc.cancer.gov/
  Cancer genomics data (mostly TCGA)
- International Cancer Genome Consortium (ICGC)
  https://dcc.icgc.org/
  Cancer genomics data
- Blueprint epigenome project, COSMIC, CIViC, OncoKB, the UCSC genome browser, Ensembl BioMart…

TOOLS

- “Point-and-click” tools
  - Xena browser
  - MEXPRESS
  - cBioPortal
- Programming tools
  - R and RStudio
  - Python
- “In-between” tools
  - Excel

LEARNING RESOURCES

- Online resources
  - https://www.rstudio.com/online-learning/
  - https://www.codecademy.com/learn/learn-r
  - https://www.coursera.org/learn/r-programming
  - https://www.google.com/
  - https://stackoverflow.com/

- Books
  - *“Intuitive biostatistics: a nonmathematical guide to statistical thinking”*, Harvey Motulsky
  - *“The drunkard's walk: how randomness rules our lives”*, Leonard Mlodinow
  - *“Doing Data Science”*, Rachel Schutt & Cathy O’Neil
  - *“An Introduction to Statistical Learning”*, Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani

QUESTIONS?