1

WHAT IS DATA SCIENCE?

What is data science?

- Definitions of data science
- Our definition of data science
- Importance of data science
- Who are data scientists?
- An nice ambiguity: science with data or science of data ?
- Research vs practice
- Data science = data curation + data analytics

Some definitions of data science

"…work concerned with the collection, preparation, analysis, visualization, management, and preservation of large collections of information …

Data scientists play active roles in … four related areas: data architecture, data acquisition, data analysis, and data archiving … Data Science is an applied activity and data scientists serve the needs and solve the problems of data users."

— Jeffrey Stanton, *Introduction to Data Science*, 2013

•

"… the science of planning for, acquisition, management, analysis of, inference, and discovery from data"

—"Final Report from StatSNSF [Support for the *Statistical Sciences* at NSF] subcommittee"
Iain Johnstone, Fred Roberts, Co-Chairs July 18, 2014.

Other terms related to data science

Other terms related to data science are:

- *informatics*, *data mining*, *knowledge discovery*, *eScience*, *cyberinfrastructure*, *data analytics*, *data curation*, etc.
- As well as, of course, the colloquial phrase “big data”.

Our definition of data science

Data science is concerned with all aspects of the **creation, management, analysis,** and **communication** of data focusing particularly on the application of computational methods to digital data

The data science objective: *extracting useful knowledge from data*

The importance of data science today

Several things have combined to create this revolution:

1. vast quantities of digital data from a wide variety of sources and often arriving in real time
2. stunning advances new analytical strategies and computational methods
3. easy to use powerful analysis tools
4. high-speed global communications
5. access to distributed high performance computing from almost anywhere in the world
6. a large community of experts in the new strategies for data management and analysis

The importance of data science today

Those things create exciting new opportunities for major advances in many industries, professions, and disciplines. . .

. . .including medicine, health, engineering, defense, safety, agriculture, business, the arts, community services, government services, and more

Data science — the with/of ambiguity

Existing definitions of data science reflect an ambiguity in the phrase “data science”.

Specifically: Is data science to be understood as the science *of* data, or as science *with* data?

Wikipedia has defined data science as:

1) “. . . the study of the generalizable extraction of knowledge from data.” — Wikipedia c. 2013-14
2) “. . . the extraction of knowledge from data” — Wikipedia c. 2014-15

The literal difference is clear:

In 1) doing data science is doing research on computational methods, that is: developing new tools and techniques, studying those tools and techniques, and studying their applications.

In 2) doing data science is *using* data science methods to do research in *other* fields (genetics, medicine, engineering, marketing, etc.).

Our definition deliberately avoids the issue. [Again:]

Data science is concerned with all aspects of the **creation, management, analysis,** and **communication** of data focusing particularly on the application of computational methods to digital data

The data science objective: *extracting useful knowledge from data*

Data science = Data Curation + Data Analytics

Data science has two components:

Data curation: Ensuring that data can be efficiently and reliably found and used

Data analytics: Employing specific techniques to extract knowledge from data

Data curation is concerned primarily with the *management of data* in order to better support *the analysis of data*

It includes among many other things: acquisition and collection, modeling, workflow, provenance, validity and integrity, metadata, preservation, integration, retrieval, re-use, policy, standards, identifiers, format conversions, processing levels, supporting reproducibility, etc.