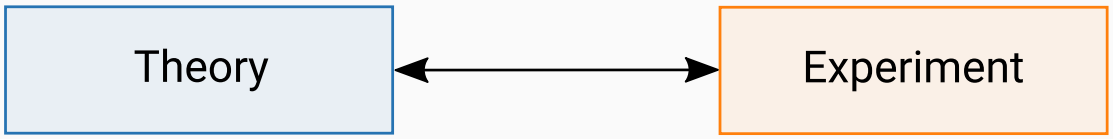
Online CDS 101 *Introduction to Computational and Data Sciences*

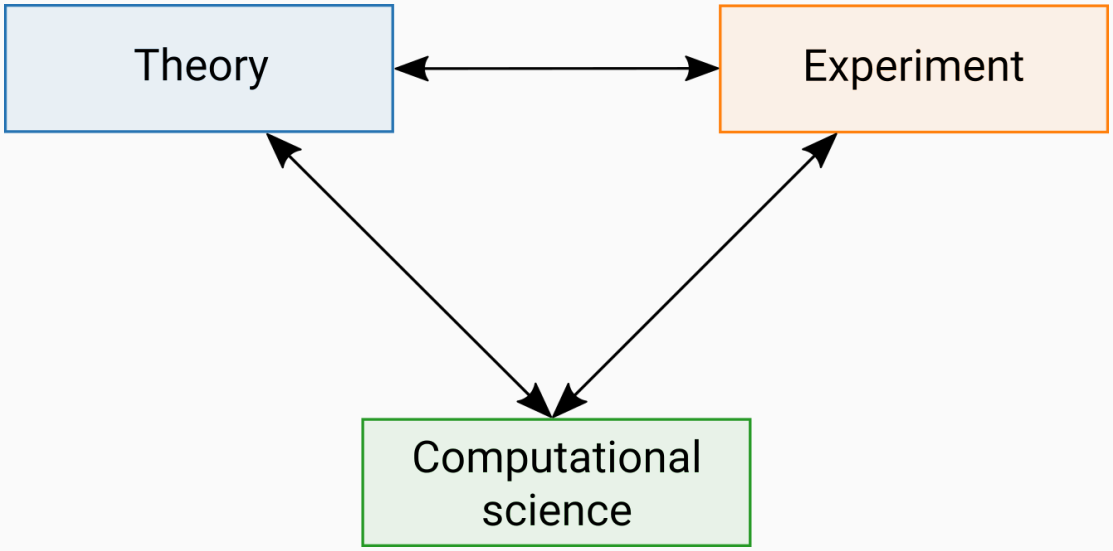
What are the computational and data sciences?

Computation

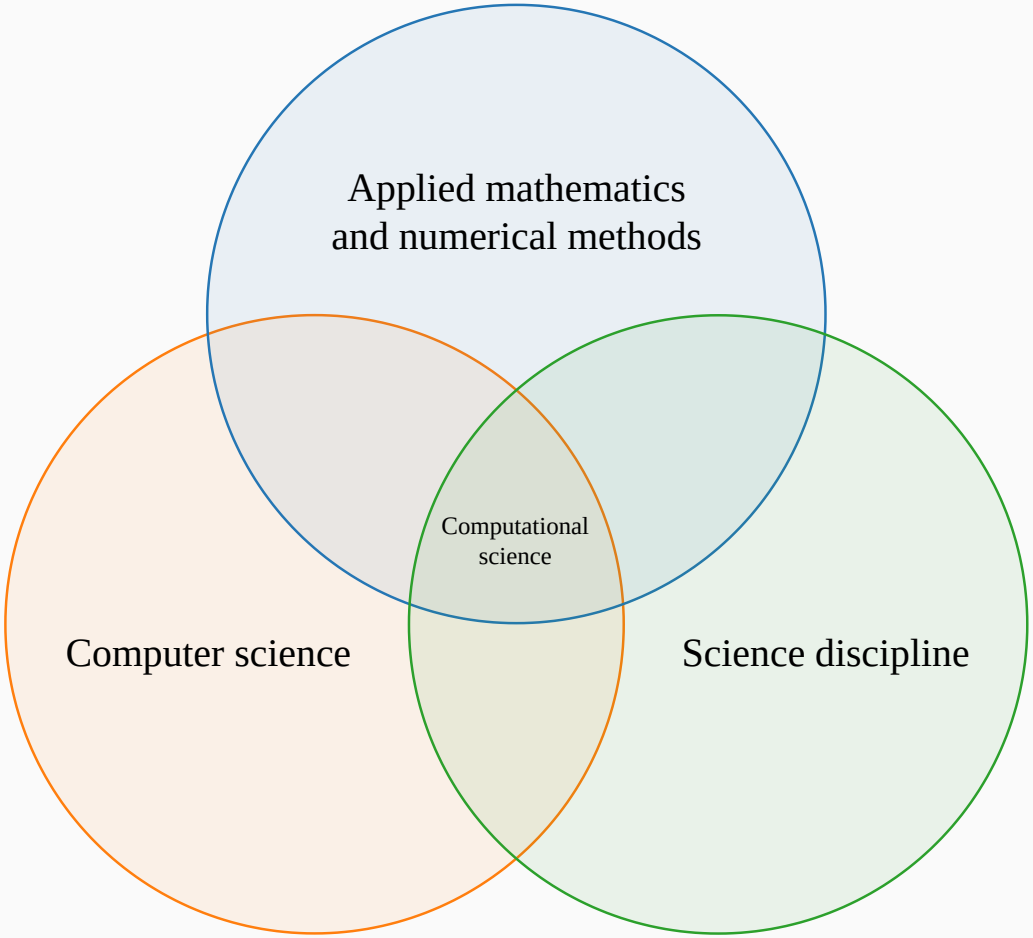
Modes of science



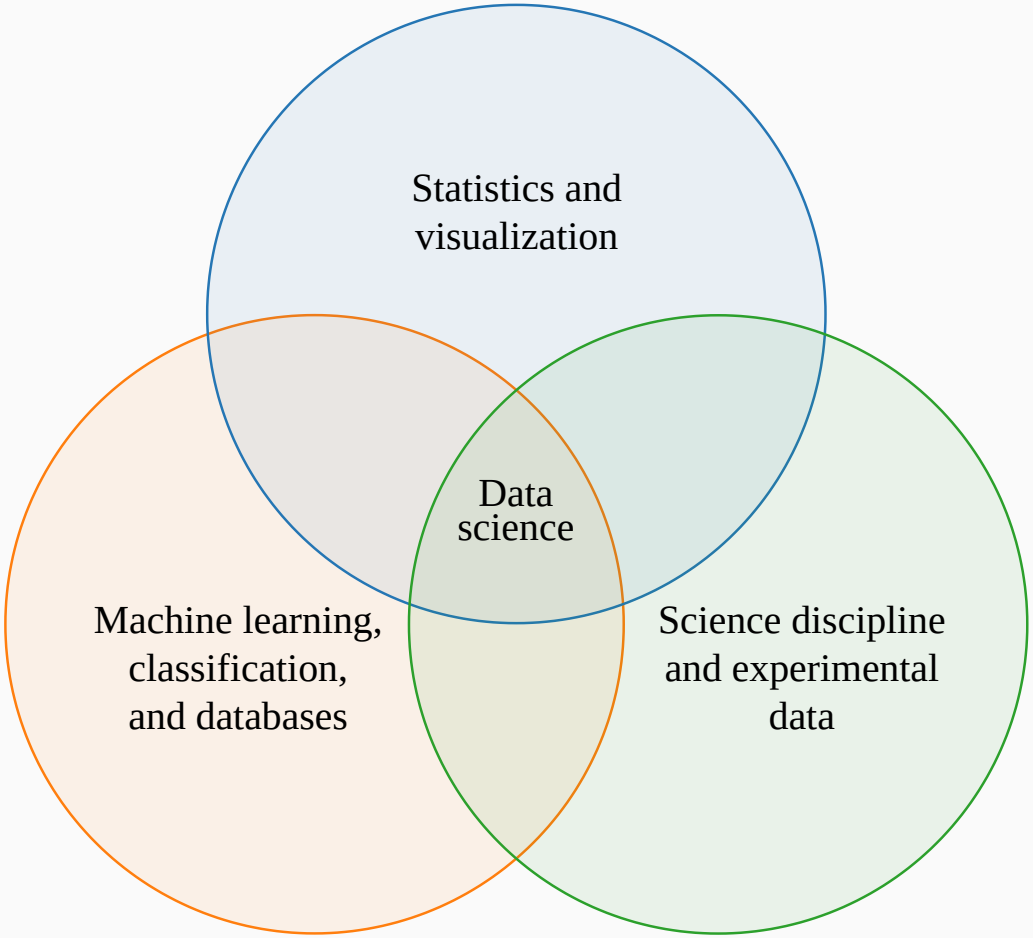
Modes of science



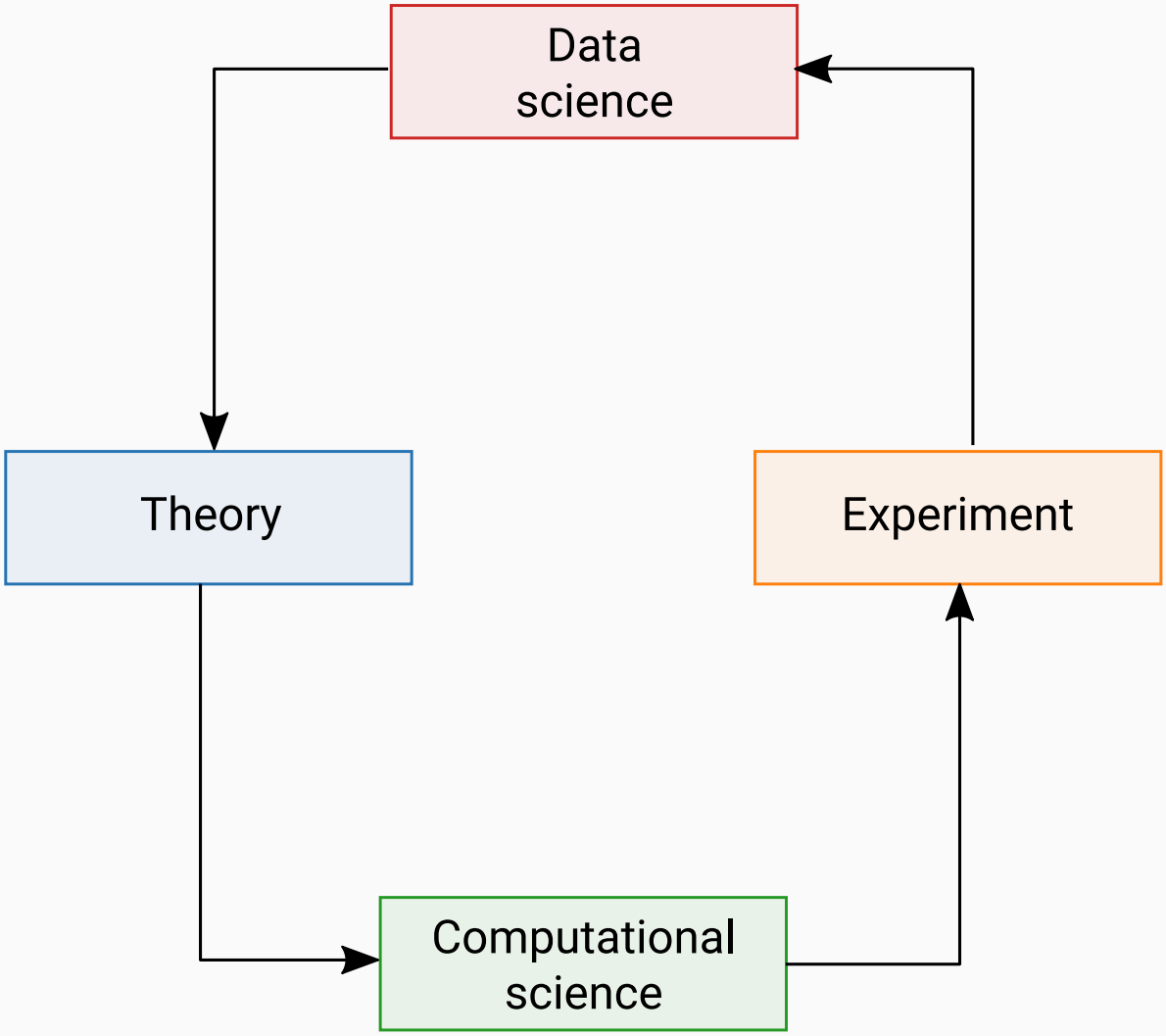
Defining computational science



Defining data science



Computational and data sciences



Big data looms

> *Big data [refers to] data sets that are so big and complex that traditional data-processing application software [is] inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, [and] information privacy [...] There are a number of concepts associated with big data: originally there were 3 concepts volume, variety, velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.*
>
> — Wikipedia

Why are new skills and training needed?

Why are new skills and training needed?

- The average scientific researcher devotes as much as 30% of their time developing and 40% of their time using scientific software (Hannay et. al., in SECSE Conference (2009), pp. 1-8), yet many undergraduate natural science programs do not integrate computational skills into the curriculum

Why are new skills and training needed?

- The average scientific researcher devotes as much as 30% of their time developing and 40% of their time using scientific software (Hannay et. al., in SECSE Conference (2009), pp. 1-8), yet many undergraduate natural science programs do not integrate computational skills into the curriculum
- A lack of computational skills increases the risk of computational errors, which hurt reproducibility and can even invalidate a study's conclusions (Merali, Nature 467, 775 (2010))

Why are new skills and training needed?

- The average scientific researcher devotes as much as 30% of their time developing and 40% of their time using scientific software (Hannay et. al., in SECSE Conference (2009), pp. 1-8), yet many undergraduate natural science programs do not integrate computational skills into the curriculum
- A lack of computational skills increases the risk of computational errors, which hurt reproducibility and can even invalidate a study's conclusions (Merali, Nature 467, 775 (2010))
- Data cleaning is a prerequisite to analyzing data in most contexts and disciplines

Why are new skills and training needed?

- The average scientific researcher devotes as much as 30% of their time developing and 40% of their time using scientific software (Hannay et. al., in SECSE Conference (2009), pp. 1-8), yet many undergraduate natural science programs do not integrate computational skills into the curriculum
- A lack of computational skills increases the risk of computational errors, which hurt reproducibility and can even invalidate a study's conclusions (Merali, Nature 467, 775 (2010))
- Data cleaning is a prerequisite to analyzing data in most contexts and disciplines
- A large chunk of work in the computational and data sciences involves applying a series of data transformations in a certain order

Why are new skills and training needed?

- Automated workflows with error-checking are very important when working with large datasets

Why are new skills and training needed?

- Automated workflows with error-checking are very important when working with large datasets
- Data science methods applied to other fields: medicine, humanities, political science, law, and the list goes on

Why are new skills and training needed?

- Automated workflows with error-checking are very important when working with large datasets
- Data science methods applied to other fields: medicine, humanities, political science, law, and the list goes on
- Researchers in the computational and data sciences often need to communicate results to non-experts, which requires effective visualizations and developing the ability to write and present a clear and compelling story

Focus of this course

- The computational and data sciences are **very** broad

Focus of this course

- The computational and data sciences are **very** broad
- Focus is on the tools, methods, and practices within *data science* category

Focus of this course

- The computational and data sciences are **very** broad
- Focus is on the tools, methods, and practices within *data science* category
    - If you also want to be introduced to the *computational science* side of things, consider taking CDS 130!

Main topics

- Learning a toolset that facilitates reproducible research

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations
- Data cleaning and reshaping (tidying)

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations
- Data cleaning and reshaping (tidying)
- Using statistical tools to interpret data distributions

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations
- Data cleaning and reshaping (tidying)
- Using statistical tools to interpret data distributions
- Inference and simulation

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations
- Data cleaning and reshaping (tidying)
- Using statistical tools to interpret data distributions
- Inference and simulation
- Modeling

Main topics

- Learning a toolset that facilitates reproducible research
- Data visualization
- Data transformations
- Data cleaning and reshaping (tidying)
- Using statistical tools to interpret data distributions
- Inference and simulation
- Modeling
- Special topic: basics of web scraping

Credits

License

Acknowledgments

Content adapted from the Lecture 1: The Computational and Data Sciences slides by John Wallin.