CSC 120: Applied Data Analytics

Instructor: Dr. Jason Willwerscheid

Fall 2022

What is Applied Data Analytics?

- Related buzzwords: machine learning; data mining; big data.

What is Applied Data Analytics?

- Related buzzwords: machine learning; data mining; big data.
- “Data science” is a useful umbrella term.

What is Applied Data Analytics?

- Related buzzwords: machine learning; data mining; big data.
- “Data science” is a useful umbrella term.
- What is data science?

What is Applied Data Analytics?

- Related buzzwords: machine learning; data mining; big data.
- “Data science” is a useful umbrella term.
- What is data science?
- Not well-defined.

What is Applied Data Analytics?

- Related buzzwords: machine learning; data mining; big data.
- “Data science” is a useful umbrella term.
- What is data science?
- Not well-defined.
- Instead of proposing a definition, I will draw some contrasts with classical statistics.

Classical Statistics

Image Removed.

Classical Statistics

Question: Which sleep-inducing drug works better?

Classical Statistics

Data:

```
##  Patient Drug Increase in Sleep
##        1    1               0.7
##        2    1              -1.6
##        3    1              -0.2
##        4    1              -1.2
##        1    2               1.9
##        2    2               0.8
##        3    2               1.1
##        4    2               0.1
```

Classical Statistics

Method: Student's paired t-test

Classical Statistics

Results:

```
t.test(dat$‘Increase in Sleep‘[dat$Drug == 1],
       dat$‘Increase in Sleep‘[dat$Drug == 2],
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dat$‘Increase in Sleep‘[dat$Drug == 1] and dat$‘
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

Classical Statistics

Answer: The difference in sleep times is unlikely to be due to chance. We can be reasonably confident that Drug 2 leads to more sleep.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.

- Machine learning methods tend to be computationally intense; processing power gets part of the way there, but we also need efficient methods.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.

- Machine learning methods tend to be computationally intense; processing power gets part of the way there, but we also need efficient methods.

- Just knowing “what’s in there” is difficult; we need various ways to visualize large datasets.

Data Science

Image Removed.

Data Science

Question: How are individuals *genetically* predisposed towards asthma?

Genetics 101

- The **genome** is the sequence of nucleotides (A, C, G, and T) that comprise the genetic instructions in your DNA; e.g., `ATTCCTATCGAA...`

Genetics 101

- The **genome** is the sequence of nucleotides (A, C, G, and T) that comprise the genetic instructions in your DNA; e.g., `ATTCCTATCGAA...`
- This sequence is nearly identical for each person, but there are sites in the DNA that vary among individuals. So individual A might have `ATTCC{T}ATCGAA` at a certain location while individual B has `ATTCC{C}ATCGAA` at the same location.

Genetics 101

- The **genome** is the sequence of nucleotides (A, C, G, and T) that comprise the genetic instructions in your DNA; e.g., `ATTCCTATCGAA...`
- This sequence is nearly identical for each person, but there are sites in the DNA that vary among individuals. So individual A might have `ATTCC{T}ATCGAA` at a certain location while individual B has `ATTCC{C}ATCGAA` at the same location.
- This kind of variation is known as a **single nucleotide polymorphism** or SNP.

Genetics 101

- The **genome** is the sequence of nucleotides (A, C, G, and T) that comprise the genetic instructions in your DNA; e.g., `ATTCCTATCGAA...`
- This sequence is nearly identical for each person, but there are sites in the DNA that vary among individuals. So individual A might have `ATTCC{T}ATCGAA` at a certain location while individual B has `ATTCC{C}ATCGAA` at the same location.
- This kind of variation is known as a **single nucleotide polymorphism** or SNP.
- There are millions of SNPs in the human genome.

Genetics 101

Image Removed.

Genome-wide association studies

- Data consists of the full genome for a number of individuals.
- Datasets can be *enormous*: typical studies include hundreds of thousands of SNPs for hundreds or thousands of individuals, and there have been studies including over a million individuals.

Genome-wide association studies

- The idea is to find the SNPs that are linked to the trait we're studying.
- Some traits are relatively simple, but a trait like asthma can have *hundreds* of causal SNPs.

Image Removed.

Genome-wide association studies

The data is not unproblematic:

- Sequencing technology is never 100% accurate, so there needs to be quality control.

Genome-wide association studies

The data is not unproblematic:

- Sequencing technology is never 100% accurate, so there needs to be quality control.
- Biased datasets have led to inaccurate conclusions: most GWASs have been over-represented by individuals of European ancestry, and results are not automatically generalizable to, say, individuals of African ancestry.

Genome-wide association studies

Methods:

- At one level, the statistical methods used are very classical: t-tests and p-values are used to determine the effect of each SNP.

Genome-wide association studies

Methods:

- At one level, the statistical methods used are very classical: t-tests and p-values are used to determine the effect of each SNP.
- But there are problems that require new methods.

Genome-wide association studies

Methods:

- At one level, the statistical methods used are very classical: t-tests and p-values are used to determine the effect of each SNP.
- But there are problems that require new methods.
- For example, selecting SNPs that have $p < .05$ doesn't work when you are testing millions of them. This is the problem of **multiple testing**.

Genome-wide association studies (GWAS)

- **Fine mapping** is another new and interesting problem: SNPs that are near one another on the genome are *correlated*, so it's tricky to find the ones that are truly causal:

Image Removed.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).
- As the scale of datasets becomes much larger, storage and retrieval present problems.
- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).

- As the scale of datasets becomes much larger, storage and retrieval present problems.

- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.

- Machine learning methods tend to be computationally intense; processing power gets part of the way there, but we also need efficient methods.

Data Science

- The data often presents interesting problems *in itself* (how to collect it; whether we can trust it).
- As the scale of datasets becomes much larger, storage and retrieval present problems.
- Classical statistical methods were developed with small datasets in mind; large datasets make new machine learning methods both possible and necessary.
- Machine learning methods tend to be computationally intense; processing power gets part of the way there, but we also need efficient methods.
- Just knowing "what's in there" is difficult; we need various ways to visualize large datasets.

Course Requirements

- There are no prerequisites for this course.

Course Requirements

- There are no prerequisites for this course.
- There is no required textbook. All readings will be from materials that are freely available online.

Course Requirements

- There are no prerequisites for this course.
- There is no required textbook. All readings will be from materials that are freely available online.
- All required software is freely available online.

Course Requirements

- There are no prerequisites for this course.
- There is no required textbook. All readings will be from materials that are freely available online.
- All required software is freely available online.
- A (reliable) laptop is required.

Course Format

There will be a regular weekly routine:

- M/Tu: Lecture-based, with slides
- W: Discussion-based, with live code demonstrations
- F: Lab (collaborative problem solving)

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.
- Appreciate that real-world datasets are rarely unproblematic; understand some of the issues that arise in data collection and data cleaning.

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.
- Appreciate that real-world datasets are rarely unproblematic; understand some of the issues that arise in data collection and data cleaning.
- Be comfortable using the statistical programming language R to explore datasets via “wrangling” and visualization.

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.
- Appreciate that real-world datasets are rarely unproblematic; understand some of the issues that arise in data collection and data cleaning.
- Be comfortable using the statistical programming language R to explore datasets via “wrangling” and visualization.
- Be familiar, at a conceptual level, with basic statistical principles of data modelling, including ethical issues and common misunderstandings.

Course Objectives

By the end of the course, you will:

- Know how to find relevant datasets in your areas of interest.
- Appreciate that real-world datasets are rarely unproblematic; understand some of the issues that arise in data collection and data cleaning.
- Be comfortable using the statistical programming language R to explore datasets via “wrangling” and visualization.
- Be familiar, at a conceptual level, with basic statistical principles of data modelling, including ethical issues and common misunderstandings.
- Have an in-depth understanding of the uses and abuses of *one* modelling technique (regression).

Course Objectives

It's also useful to spell out what the course does *not* aim to do. In particular, you will not necessarily:

- Have technical mastery in R (this is not a programming course).

Course Objectives

It's also useful to spell out what the course does *not* aim to do. In particular, you will not necessarily:

- Have technical mastery in R (this is not a programming course).
- Have a detailed mathematical understanding of hypothesis testing, linear regression, etc. (this is not a statistics course).

Course Objectives

It's also useful to spell out what the course does *not* aim to do. In particular, you will not necessarily:

- Have technical mastery in R (this is not a programming course).
- Have a detailed mathematical understanding of hypothesis testing, linear regression, etc. (this is not a statistics course).
- Be exposed to a panoply of machine learning techniques such as clustering, support vector machines, and neural networks. These techniques are very powerful, but they require a good statistical foundation if they are to be used responsibly.

Course Schedule

Image Removed.

Assignments for This Week

1. Read through the syllabus on Canvas/GitHub. We can discuss on Friday if there are questions or concerns.
2. Sign up for a GitHub account and complete the software setup detailed in Assignment 1. Troubleshooting on Friday.
3. Request your personal data (Assignment 0). Due Tuesday.