Ve572 Lecture 1

Manuel and Jing

UM-SJTU Joint Institute

May 15, 2018

- Course Description:

  Over the last decade, we have seen the emergence of "big data". This course is to introduce students to the basics of big data methods and tools. The course begins with a broad and practical introduction to topics in data visualisation, data analysis, machine learning and data mining. The course then examines and discusses the current big data ecosystem/infrastructural technologies. Students are expected to obtain a good knowledge of the statistical programming language R and the most cutting-edge big data technologies and frameworks, such as Hadoop, Spark and Drill.

- Who should take this class?

  The prerequisite for this class is computer/programming knowledge at the level of Ve482 (or above), and statistics knowledge at the level of Ve401 (or above). Both undergraduates and graduate ECE students are welcome to take the course.

- **Instructor:**

  Manuel Charlemagne
  Jing Liu

- **Lectures:**

  Tuesday (12:10pm – 1.50pm) in CRQ-312
  Thursday (12:10pm – 1.50pm) in CRQ-312

- **Office Hours:**

  Monday (08:30am – 09:40am) in JI-Building 441A
  Monday (02:30pm – 03:40pm) in E2-104
  Thursday (02:30pm – 03:40pm) in E2-104 Even Weeks Only

- **Email:**

  charlem@sjtu.edu.cn and stephen.liu@sjtu.edu.cn

- **Teaching Assistant/s:**

  See Canvas for his/her contact information

- To improve communication between the students and the teaching team please observe the following guidelines:
  - Any student facing a special situation likely to impact his studies, such as serious illness or full time work, is expected to contact the instructor as early as possible in order to discuss it and see if any solution can be found.
  - When sending an email related to this course please include the tag [Ve572] in the subject e.g. Subject: [Ve572] special request
  - When contacting an instructor for a grade issue or any other major problem send a carbon copy (cc) to the other instructor. Not doing it might result in omissions, not up-to-date grades etc. If such problem occurs and there is no record of the issue the request will be **automatically rejected**.
  - Never attach a large file ($>$ 2 MB) to an email, use Canvas Dropbox instead and only include a link in the email.
  - Keep in touch with the teaching team, feedbacks and suggestions will be much appreciated.

- Lab:

  10% There will be 4 compute lab sessions.

- Assignment:

  20% Assignments will be given in the form of problem sets, and may require extra reading and using of a computer.

- Project:

  30% The project has two major components.

Methods	10%
Tools	20%

- Exam:

  40% There will be two exams:

Midterm	25%	
Final	15%	Not accumulative.

- For this course, the grade will be curved to achieve a median grade of "B^+".

- Honesty and trust are important. Students are responsible for familiarising themselves with what is considered as a violation of honour code.
- Assignments/labs/projects are to be solved by each student individually. You are encouraged to discuss problems with other students, but you are advised not to show your written work to others. Copying someone else's work is a very serious violation of the honour code.
- Students may read resources on the Internet, such as articles on Wikipedia, Wolfram MathWorld or any other forums, but you are not allowed to post the original assignment question online and ask for answers. It is regarded as a violation of the honour code.
- Only a single sheet, A4 size with your original handwritten notes on both sides, is allowed during the written exams.
- Since it is impossible to list all conceivable instance of honour code violations, the students has the responsibility to always act in a professional manner and to seek clarification from appropriate sources if their or another student's conduct is suspected to be in conflict with the intended spirit of the honour code.

- White. (2015)

  Hadoop: the definitive guide.

- Some Additional Material:
  - James et al. (2017)

    An introduction to Statistical learning: with Applications in R
  - Grolemund and Wickham (2016)

    R for Data Science
  - Walkowiak. (2016)

    Big Data Analytics with R.
  - Sitto and Presser (2015)

    Field Guide to Hadoop.
  - Holmes (2014)

    Hadoop in Practice.

- Teaching Schedule:

Week	Topics
1	Introduction Data Science Wars: R Vs Python
2	Data types and structures in R Lab - Linux + Scripting + Java (Checklist)
3	Data import and visualisation in R Data Wrangling in R
4	Data Analysis in R Large Datasets in R
5	Concurrent programming in R Machine Learning
6	Data Mining Google's Big Data Platforms and Services (Optional)

7	Lab - Relational Database Management System **First Midterm Exam**
8	Distributed Systems Lab - Setting up Hadoop
9	Hadoop Ecosystem Lab - Using Hadoop
10	MapReduce, Yarn, and Pig Hadoop distributed file system
11	Hadoop Spark Hadoop Drill
12	Bring Hadoop to R More sophisticated Hadoop uses (Optional)
13	**Final Exam**

Motivation

- Hal Varian (2009)

Quotation

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would have guessed that computer engineers would have been the sexy job of the 1990s."

- Harvard Business Review (2012)

Data Scientist: The Sexiest Job of the 21st Century

- Glassdoor (2016-2018)

The best jobs of the year

- Linkedin (2017 Dec)

20 Fastest-Growing Jobs

- What will this course provide?

1. A broad and practical introduction to statistical methods for Big Data
2. A detailed discussion on ecosystem/infrastructural technologies for Big Data

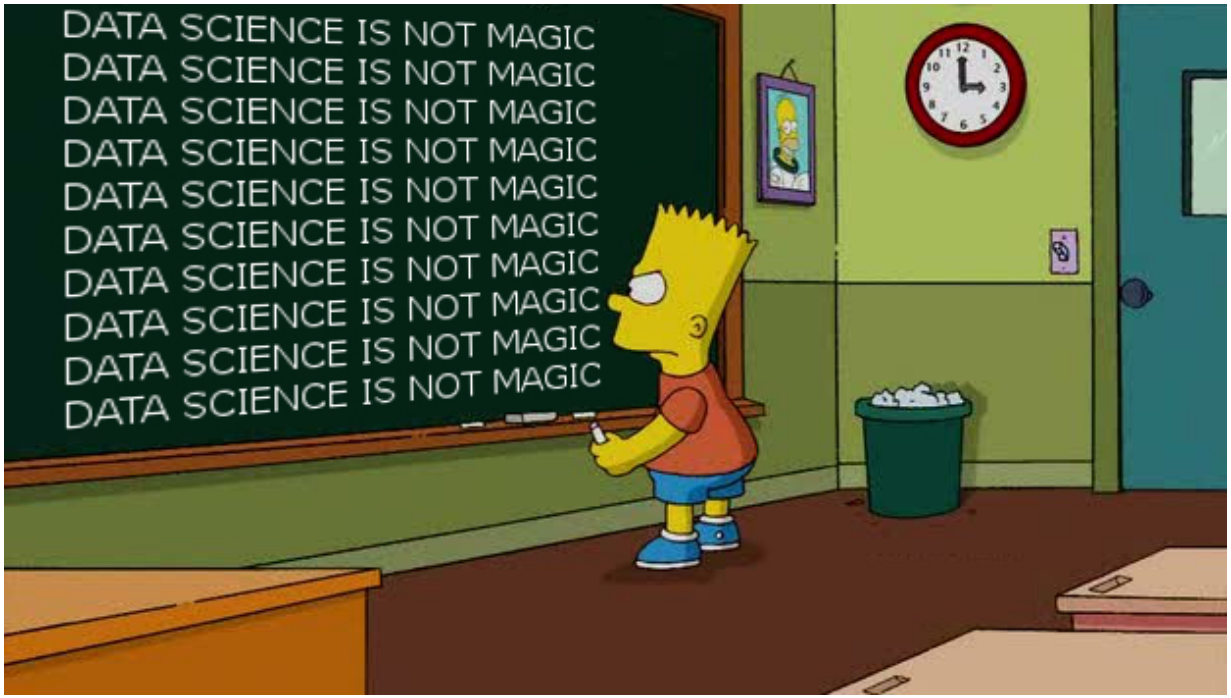
- What will this course not focus on?

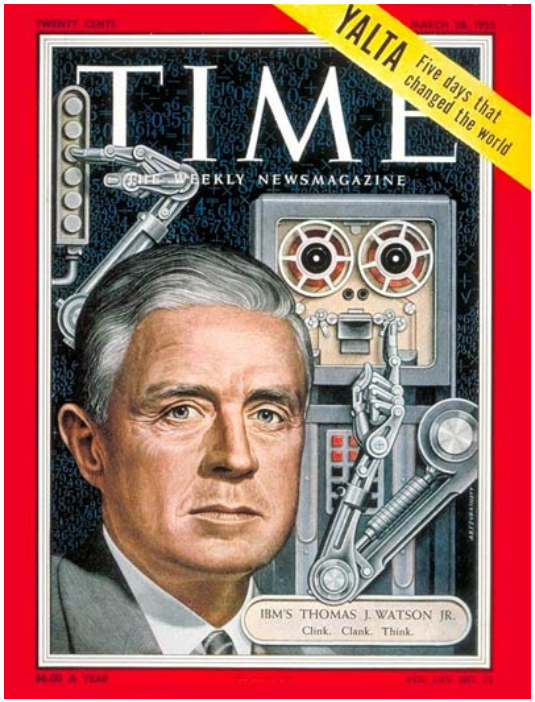
1. Mathematical derivation
2. Business aspects, privacy and other ethical issues

- Technical skills this course expect you to know or learn quickly on your own:
  - Linux
  - Python
  - Java
- Technical skills that you can expect to learn in this course:
  - R
  - Hadoop

Q: What does "Big Data" mean?

Q: Is "Big Data" overrated?





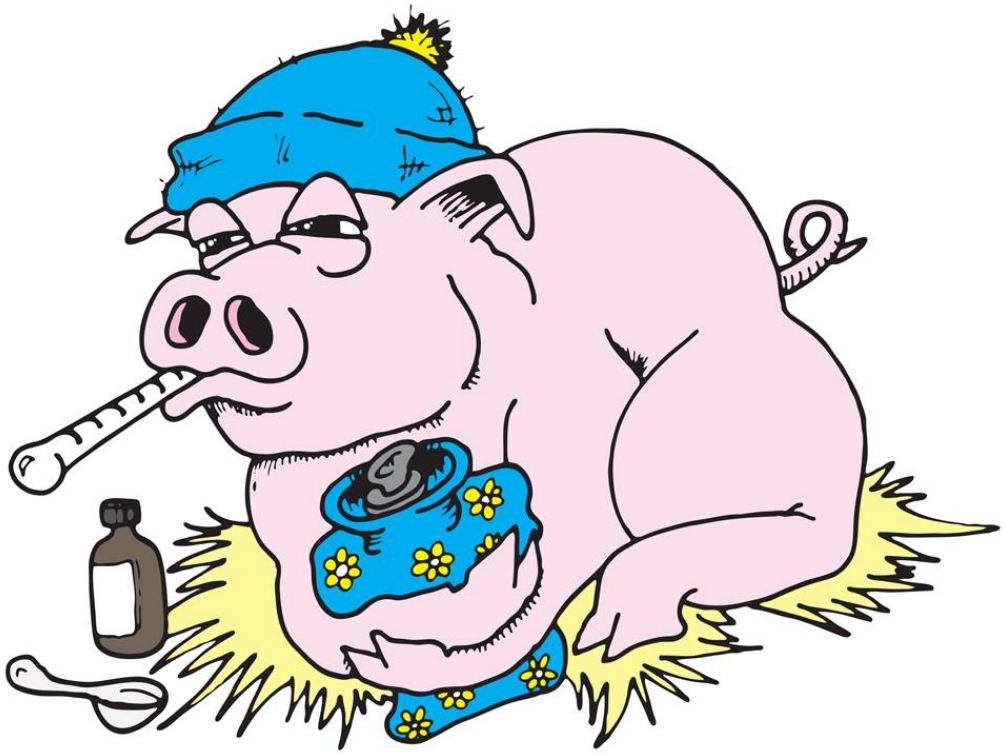
Quote

"I think there is a world market for maybe five computers."

Q: What is the first success story about Big Data that have impressed you?

Q: How many of you have heard of "Google Flu Trends" (GFT)?

- In 2009, a new virus, H1N1, was discovered and spread very quickly



- At the time no vaccine against it was readily available. The only hope public health authorities had was to slow its spread. But to do that, they needed to know where it already was.

- In USA, the government requested that doctors to report new flu cases.

- With a rapidly spreading disease, a two-week lag is an eternity. This delay completely blinded the public health agencies at the most crucial moments.
- A few weeks before H1N1 made headlines, Google published a paper

Detecting Influenza Epidemics Using Search Engine Query Data

in which it gives a model. Like the government, Google could tell where the flu had spread, but unlike the government they could tell it in near real time, one day or two, not a week or two after the fact.

- It was remarkable for three reasons:

1. Simplicity

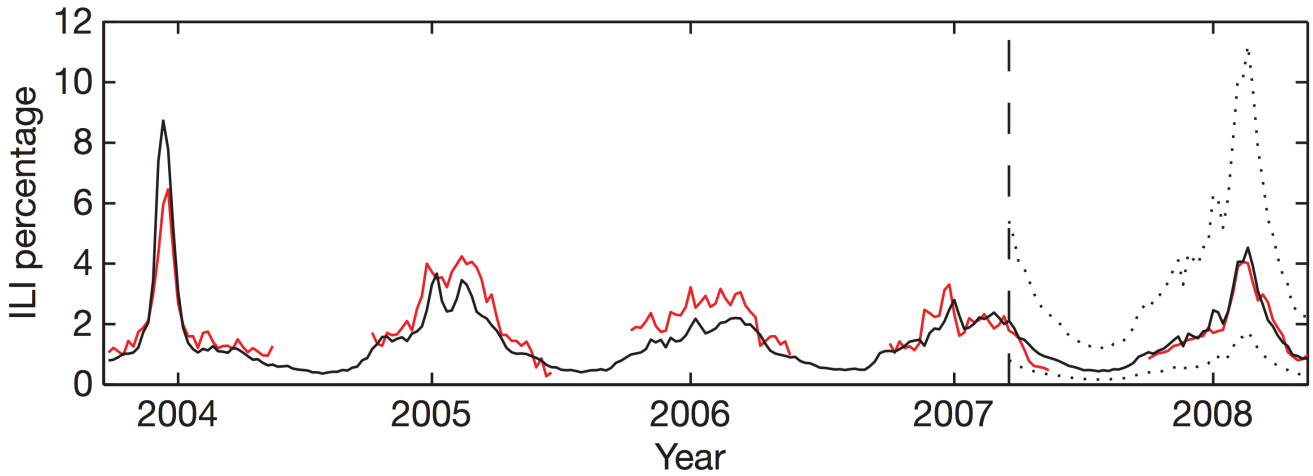
$$\text{logit}\,(P) = \beta_0 + \beta_1\,\text{logit}\,(Q_1) + \beta_2\,\text{logit}\,(Q_2) + \cdots \beta_{45}\,\text{logit}\,(Q_{45}) + \varepsilon$$

Not only the model is relatively simple but Google did so little to obtain it.

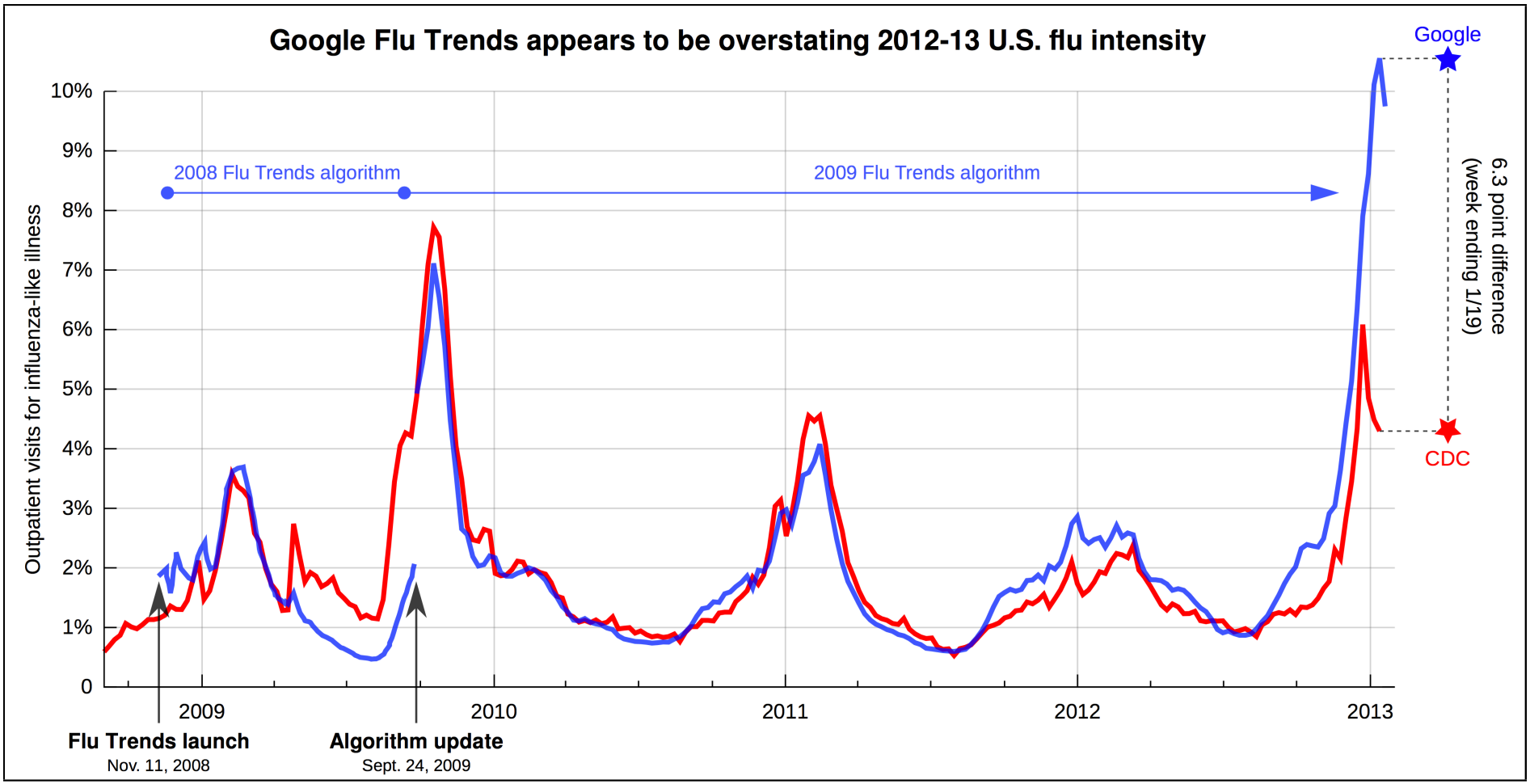
2. Data size

$$\underbrace{50{,}000{,}000}_{queries} \times \underbrace{154}_{locations} \times \underbrace{182}_{weeks} = 1{,}393.7 \text{ billion data points} \approx 11{,}149.60\text{GB}$$

3. Initial performance



- However, something that is too good to be true, is often not true!



- Once the media got hold of it, after a study published in the journal Science,

  The Parable of Google Flu: Traps in Big Data Analysis

  Google was shaken by negativity,

  - Google Flu Trends is no longer good at predicting flu
  - Google Flu Trends Gets It Wrong Three Years Running
  - Why Google Flu Is A Failure
  - Data Fail! How Google Flu Trends Fell Way Short
  - Google's Flu Project Shows the Failings of Big Data

  and Google Flu Trends was discontinued in 2015.

  The remains of the Google Flu Trends

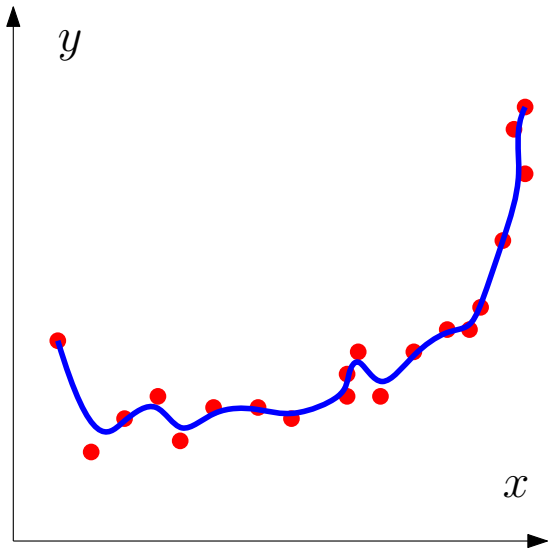
Q: What was the problem?

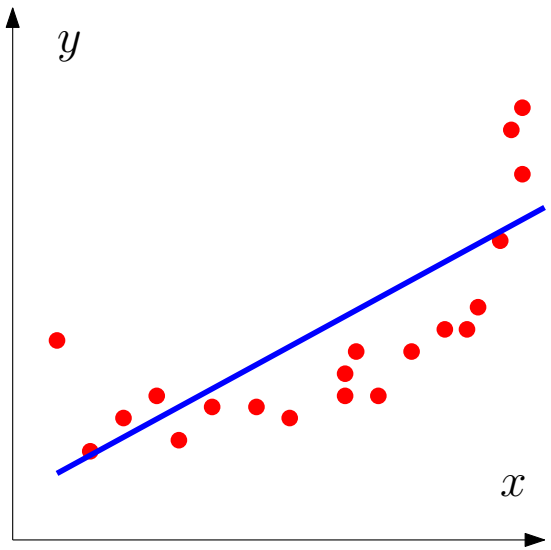
- Problems that Google Flu Trends had
  - Overfitting to seasonal terms unrelated to the flu
  - Correlation is not the same as causation

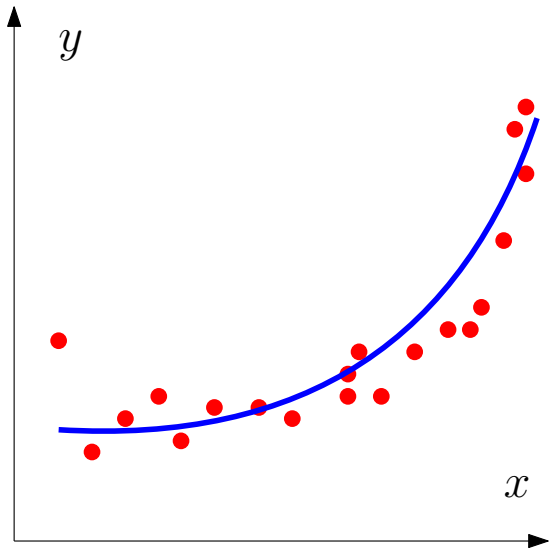
Definition

Overfitting is a conclusion, often a model, drawn from a particular data set that is too specific. Underfitting is a conclusion drawn from a particular data set that is too simplistic.

Q: Is each of the followings a overfit, underfit or "about right"?

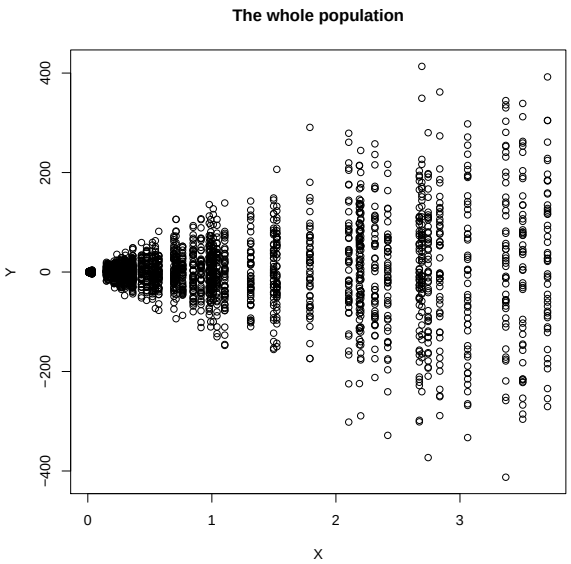


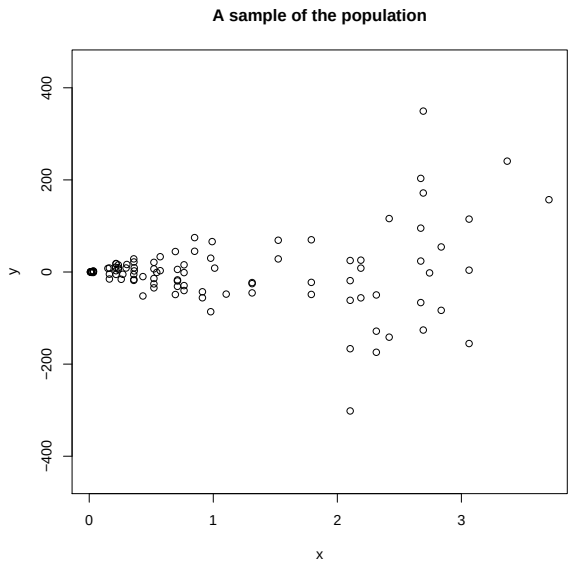


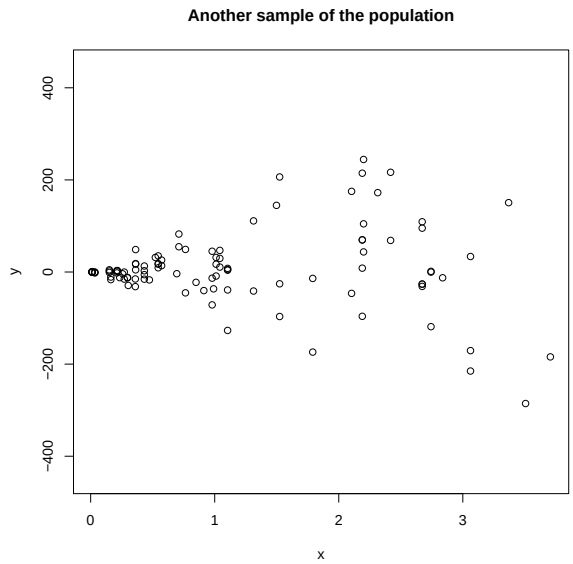


Q: Why is overfitting/underfitting problematic?

- Data is only a portion of the whole picture, which we hope is representative!







- The population is generated by taking $Y_i \sim \text{Normal}\,(\mu = X_i, \sigma = 50X_i)$.
- Notice some variability could be due entirely by chance instead of coming from the actual relationship between X and Y.
- An overfitted model often leads to an inflated prediction error.

- Notice how overfitting inflates the error when x is large.