10. Bad Data and Big Data

Bad Data

Does the dataset pass the **smell test**?

- invalid entries, anomalous observations, etc.

Data formatted for human consumption, not machine readability

Difficulties with **text processing**

- encoding
- application-specific characters

Bad Data

Collecting data **online**

- legality of obtaining data
- storing offline versions

Detecting **lies** and **mistakes**

- reporting errors (lies or mistakes)
- use of polarizing language

Data and reality

- bad data
- bad reality?

Bad Data

Sources of **bias** and **errors**

- imputation bias
- top/bottom coding (replacing extreme values with average values)
- proxy reporting (head of household for household)

Seeking **perfection**

- academic data
- professional data
- government data
- service data

Bad Data

Data science **pitfalls**

- analysis without understanding
- using only one tool (by choice or by fiat)
- analysis for the sake of analysis
- unrealistic expectations of data science
- it’s on a need-to-know basis and you don’t need to know

Databases vs. files vs. cloud computing

- the cloud will solve all of our problems!

Bad Data

When is **close enough, good enough**?

- completeness
- coherence
- correctness
- accountability

Big Data – A Word of Warning

Big Data is no crystal ball

- "Past performance does not guarantee future results"

Big Data can't dictate personal or organizational values

- The right value answer may be the wrong data science answer
- Data-based conclusions do not live in a vacuum: context matters
- Blind obedience to data-driven results is just as dangerous as rejection based on gut-reaction

Big Data can't solve every problem

- "When all you have is a hammer, everything looks like a nail"

Big Data vs. Small Data

What is the main difference?

- the datasets are **LARGE**
- issues: collection, capture, access, storage, analysis, visualization

Where does the data come from?

- technology advances are lifting the limits on data processing speeds
- information-sensing, mobile devices, cameras and wireless networks

What are the challenges?

- most techniques were built for very small dataset
- direct approach will leave the best analyst waiting years for results

6. **variability**
7. **visualization**

The 5V (7V?) Paradigm

1. **volume:** large amounts of data
2. **velocity:** speed at which data is created, accessed, processed
3. **variety:** different types of available data, can't all be saved in relational databases (tables, pictures,...)
4. **veracity:** quality and accuracy of big data is harder to control
5. **value:** turn the data into something useful

The Big Data Problem

Many computations happen **instantly**, others take a **significant** amount of time.

Crunching very large datasets is a perfect example. Analysis in *R* or *Python* with steadily increasing datasets leads to computer lags. Eventually, the time required becomes **impractically long**.

Optimizing code and using a faster CPU can only provide so much relief.

That is the **Big Data problem**.

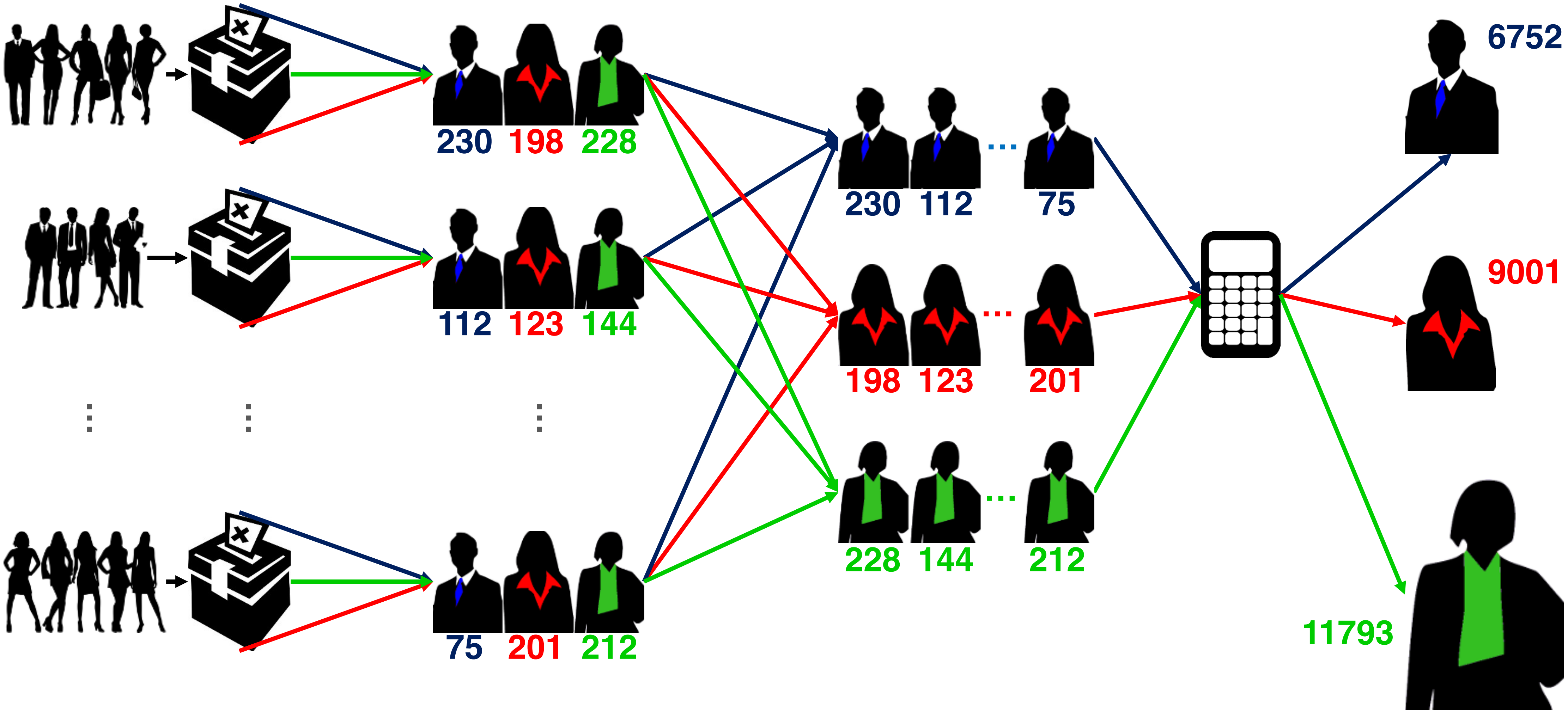
Distributed Computing

Splitting the computations among multiple CPU cores/CPUs can divide the computation time by a factor of 4, or 32, or 1000, or … This allows algorithms to run on big data to keep analytics, smart services, and recommendations updated **daily**, **hourly**, in **real time**.

Election analogy to parallelization:

- counting votes at different polling stations in a riding
- each station simultaneously counts its own votes and reports their total
- the totals of all polling stations are aggregated at Elections HQ
- one person counting all the ballots would eventually get the same result, but it would take *too long* to get the result.

Analogy: Elections



Analogy: Pizzeria

Parallelism gains depend on whether serial algorithms can be **adapted** to make use of **parallel hardware**.

Pizzeria analogy for limitations of parallelization/bottleneck:

- multiple cooks can prepare toppings in parallel
- but baking the crust can't be parallelized
- doubling oven space will increase the number of pizzas that can be made simultaneously but won't substantially speed up any one pizza
- sometimes bottlenecks prevent any gains from parallelism: people line up on both sides of a table to get some soup but there's only one ladle

Good News

Most practical computational tasks can be and are parallelized.

Modern data scientists use frameworks where distributed computing are already implemented (Apache Spark implements *MapReduce*, for instance).

Take some time to think about this potential issue **before** the start of the data collection/data analysis process – it will save headaches in the long run.

Suggested Reading

Bad Data and Big Data

J. Leskovec, A. Rajamaran, and J. D. Ullman, *Mining of Massive Datasets*. Cambridge Press, 2014.

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Issues and Challenges

- Bad Data

Exercises

Bad Data and Big Data

1. As the saying goes, “garbage in, garbage out”. What are the analytical, business, and public policy consequences of making decisions based on bad data?

2. Whether a dataset is considered small or “big” depends not only on the dataset, but also on the available tools.

   Generate increasingly larger random datasets (3 variables + 1 class) to cluster with `kmeans()` and classify with `rpart()`. Keep track of the runtime. How does the runtime vary with the number of observations? At what size do you predict that the algorithms will be too slow and cumbersome for your needs?