Introduction

Whether you want to be a data scientist or simply learn how to use a little data science to make a better life for yourself

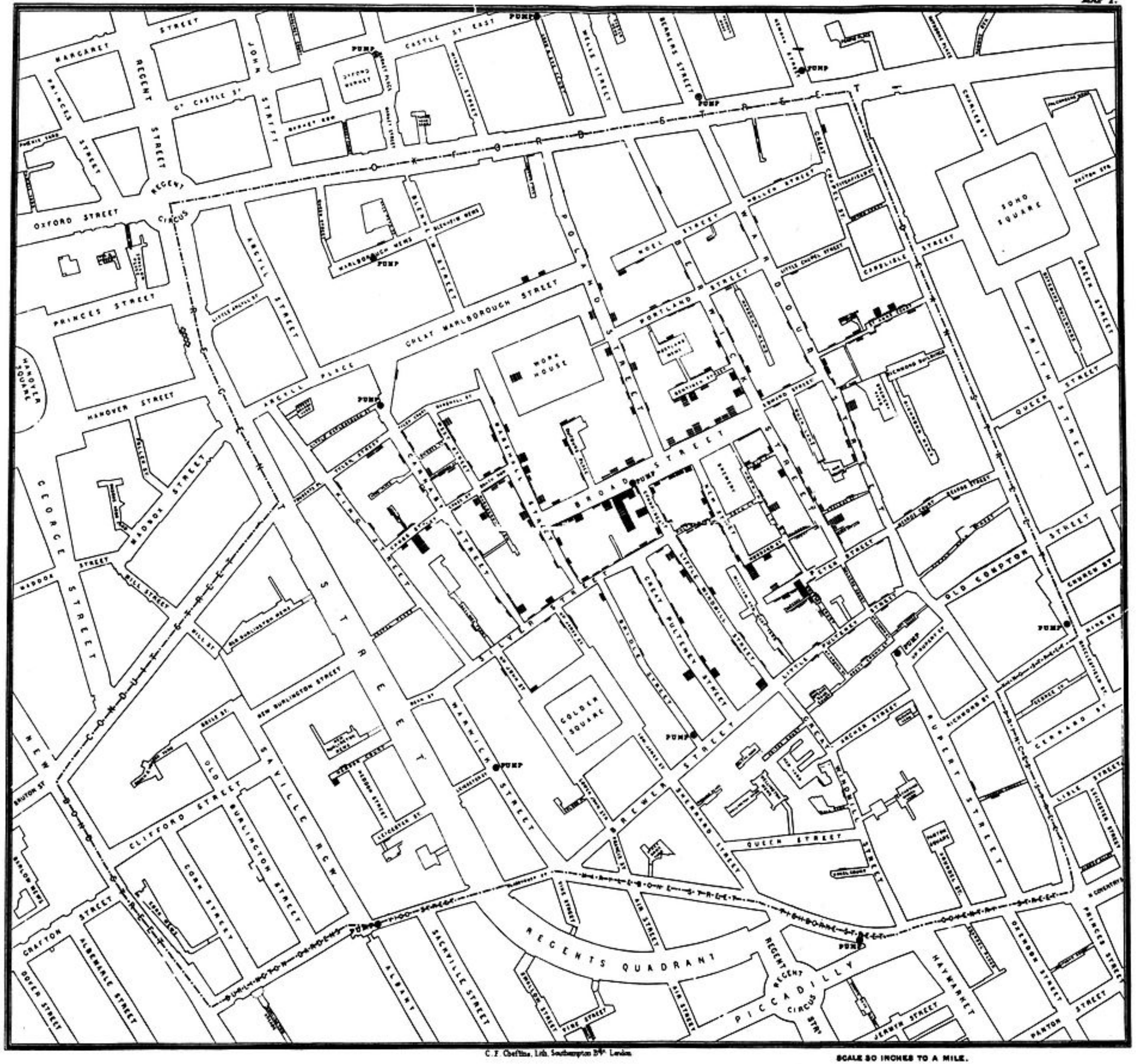
A Famous Example

- Cholera epidemic in Soho, London in 1854
- A nasty disease, sometimes fatal
  - bacterial infection of small intestine
  - Even today – more than a million cases worldwide, >25K deaths
- Nobody knew the origins
  - Common view was miasma theory (bad air)
  - No germ theory of disease yet
- John Snow (1813-1858) was a surgeon
- First encountered cholera in 1831 in Newcastle upon Tyne
- Leading authorities on the use of anesthetics (proper doses of ether and chloroform – Queen Victoria's last two childbirths)
- Skeptical of the miasma theory as the source of cholera

1854 cholera epidemic – Soho, London

- Working together with the Church of England priest Henry Whitehead, Snow interviewed families that had experienced cholera in 1954
- Drew a dot map of the incidences – and looked for patterns
- Had to reshape the patterns based somewhat on what they were told in interviews
    - Deliveries by Southwark and Vauxhaul Waterworks Company of water from polluted sections of the Thames scattered the cholera incidences more widely
    - Five families had taken water not from their local water source but from the Broad Street pump since they thought it tasted better
    - Three other cases school children who lived at a distance but went to school near the Broad Street well

MAP 1.



C. F. Cheffins, Lith, Southampton Bldgs London

SCALE 30 INCHES TO A MILE.

The solution

- Convincing evidence from patterns of incidence that the primary source of the cholera was the Broad Street water pump
- Evidence from microscopic evaluation of the Broad Street water was inconclusive
- Convinced the local council to take the handle off the Broad Street water pump
- Turned out that the Broad Street well was too close (3 feet) from a cesspit that leaked fecal bacteria into the water
- Dot-map process was used in Exeter by another doctor, Thomas Shapter, to locate the source of cholera outbreak
- Regarded as the start of public health and epidemiology
- Snow was a cofounder of the Epidemiological Society of Britain
- Snow drank only boiled water the rest of his life

[Steven Johnson, *The Ghost Map* (Riverhead Books, 2007)]

Lessons

- Patterns can help you understand
- Must render the data in ways that accentuates the relevant information
- The mathematics was simple but powerful
- The social consequences were great
    - Britain took better care of its fresh water and sewage, reducing disease
- The results did not lead to definitive proof but were still extremely helpful
    - Later statistical analysis and microbiological analysis proved Snow right!
    - Confirmatory pattern analysis in another town (Exeter) was partial confirmation of both places – independent cases

Yann LeCun
Director of AI Research at Facebook

- Most of the knowledge in the world in the future is going to be extracted by machines and will reside in machines
- There are just not enough brain cells on the planet to even look or even glance at that data, let alone analyze it and extract knowledge from it
- Knowledge is some compilation of data that allows you to make decisions, and what we find today is that computers are making a lot of decisions automatically
- The amount of human brainpower on the planet is actually increasing exponentially …, but with a very, very, very small exponent. It’s very slow growth rate compared to the data growth rate

[Source: http://www.kdnuggets.com/2015/04/data-scientists-thoughts-that-inspire.html]

Erin Shellman
data scientist in the Nordstrom Data Lab

- Data’s just the world making noises at you
- As a data scientist, even if you don’t have the domain expertise you can learn it, and can work on any problem that can be quantitatively described
- The most interesting types of data are those collected for one purpose and used for another
- Presentation is the ability to craft a story
- Presentation skills are undervalued, but is actually one of the most important factors contributing to personal success and creating successful projects
- What companies want is a person who can rigorously define problems and design paths to a solution

Daniel Tunkelang
Head of Search Quality at LinkedIn

- Intuition is really a well-trained association network
- As data scientists, our job is to extract signal from noise
- Search is the problem at the heart of the information economy
- Our goal is to fail fast. Most crazy ideas are just that: crazy
- Technology is like exercise equipment in that buying the fanciest equipment won’t get you in shape unless you take advantage of it
- Data scientists need to have strong critical-thinking skills and a healthy dose of skepticism

John Foreman
Chief Data Scientist at Rocket Science Group

- Data scientists are kind of like the new Renaissance folks, because data science is inherently multidisciplinary
- It’s essential for a data science team to hire people who can really speak about the technical things they’ve done in a way that nontechnical people can understand
- If you’re solving problems appropriately and you can explain yourself well, you’re not going to lose your job

Goals of this course/techniques for success

- Data is everywhere.
- We can understand our world better by using the data around us.
- Our goal is to learn tools that will help us to describe, organize, and interpret data that we confront in our studies and in our work lives.
- We will use some mathematics in the course, but not too much:
  - Finite mathematics, statistics, probability, even a smidgen of calculus
  - Don’t worry, you can do this! If you passed college-prep math courses in high school, you can succeed in this course!
  - This course will show you some of the value of all that math you learned.
  - Important that you don’t skip sections; do read the material carefully, and do the exercises. Later material builds on earlier material.
  - Being conscientious and orderly is a pathway to success.

Work in Progress/A Course for Pioneers

- What this course is not:
  - A generic intro statistics course or intro data science course
  - A course focused on mathematical techniques for psychology, education, health sciences, biological sciences, physical sciences, engineering
- We want this course to:
  - Provide an intro to quantitative reasoning specifically for CMCI students
  - Provide a quantitative foundation for students who want to major in info sci
- Challenges
  - Finding a good textbook that matches our model curriculum
  - Finding good statistical software that avoids double cognitive load
  - Finding relevant examples

Self-Introduction and Exercise

- Name
- How you would like us to address you
- One nonacademic factoid about yourself that you are willing to share
- Major or anticipated major
- Thoughts about your occupation and career
- Any remarks about how this course might help (or hinder) your career aspirations