Data@Nite

ELI5: Data Science

Chris Hua

1/29/2017

Examples

Almost everything can be a data science problem...

- Residency matching
-
- FedEx problem

Residency matching

- How can we *fairly* match medical school graduates with hospitals for residencies?
- Goal: to match applicants to hospitals so that the final result is "stable", i.e. no applicant A and hospital H such that both:
  - No A is unmatched or would prefer to go to H over the hospital he is currently matched with
  - H has a free slot or would prefer A over one of the candidates currently filling one of its slots.
- Won 2012 Nobel Prize in Economics

FedEx problem

- FedEx needed to pick a hub location that is close to everywhere
- You could use data science for this:

The FedEx problem

Kent E. Morrison

(Submitted on 23 Oct 2014)

The original shipping strategy of FedEx is to fly all packages to a hub location during the afternoon and evening, sort them there, and then fly them to their destinations during the night for delivery the next day. This leads to interesting mathematical questions: Given a population represented by points in Euclidean space or on a sphere, what is the location of the point of the hub that minimizes the total distance to all the points? Is such a point unique? Then using census data from 2000 we examine how close the FedEx hub in Memphis is to the hub for the U.S. population.

Figure 1:

FedEx problem

20 pages of math and data processing later:

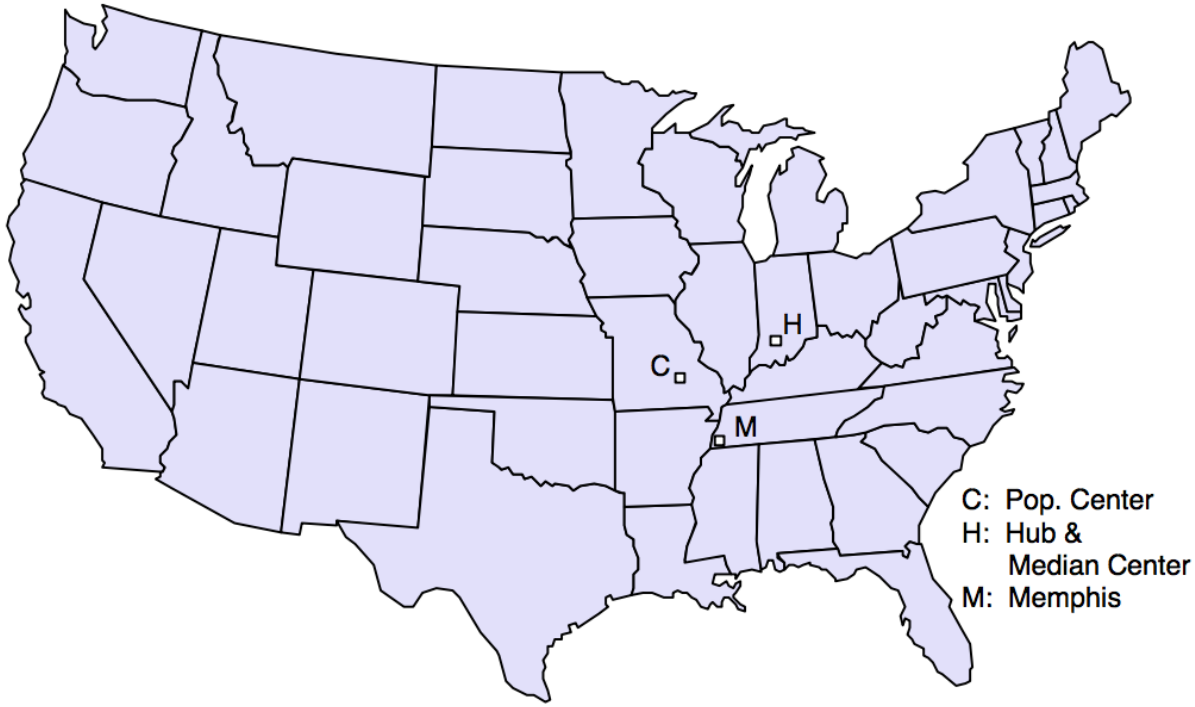

Figure 5: Location of the 2000 Census Bureau population center (C), the 2000 U.S. population hub (H), and the FedEx Memphis hub (M). The median center is very close to H.

Figure 2:

FedEx problem

graycat 673 days ago [-]

Yes, at FedEx, we considered that *problem* for about three seconds before we noticed that we also needed:

(1) A suitable, existing airport at the hub location.

(2) Good weather at the hub location, e.g., relatively little snow, fog, or rain.

(3) Access to good *ramp* space, that is, where to park and service the airplanes and sort the packages.

(4) Good labor supply, e.g., for the sort center.

(5) Relatively low cost of living to keep down prices.

(6) Friendly regulatory environment.

(7) Candidate airport not too busy, e.g., don't want arriving planes to have to circle a long time before being able to land.

(8) Airport with relatively little in cross winds and with more than one runway to pick from in case of winds.

(9) Runway altitude not too high, e.g., not high enough to restrict maximum total gross take off weight, e.g., rule out Denver.

(10) No tall obstacles, e.g., mountains, near the ends of the runways.

(11) Good supplies of jet fuel.

(12) Good access to roads for 18 wheel trucks for exchange of packages between trucks and planes, e.g., so that some parts could be trucked to the hub and stored there and shipped directly via the planes to customers that place orders, say, as late as 11 PM for delivery before 10 AM.

So, there were about three candidate locations, Memphis and, as I recall, Cincinnati and Kansas City.

The Memphis airport had some old WWII hangers next to the runway that FedEx could use for the sort center, aircraft maintenance, and HQ office space. Deal done -- it was Memphis.

Figure 3: Or not

Data science

- Using data, statistics, and logic to answer business questions
- Involved in all steps of the business process
  - Unsupervised analysis
  - Experiment design
  - Experiment analysis
  - Feature creation

Finance

- Quant funds (Two Sigma, RenTech, etc) make trading strategies from lots of weak signals
- Banks/insurance use data to manage risk
- Consulting: it's data science without data, or science

Marketing

- Fundamental question: Who should we target, and how can we reach them?
  - Customer lifetime value and segmentation
  - Targeting, A/B testing, optimization

Healthcare

- Predictive analytics - readmission rates
- Epidemiology - disease spread modelling
- Diagnosis - sort of

Industrial/Manufacturing/Farming

- Lot of operations research, manufacturing optimization
- Oil field discovery, accident/anomaly detection
- Crop productivity, environmental analysis

Tech

- Product analytics: what features should we implement, and are they working out
- Content: what should we show to who (e.g. FB Newsfeed)
- Lots of data, easy measurement, inherently real time update

Role breakdowns

- Business analyst: little programming, some SQL, generally financial analysis, basically a consultant
- Data analyst: some programming, some stats, lots of SQL, mostly analysis
- Data scientist: mix of everything
  - "Analysis" DS: focused on analysis and stats, some programming, little to no ML
  - "Building" DS: lot of programming, solid stats, at minimum can prototype ML models
- Machine learning engineer: deep understanding of ML algorithms, implementation, tuning/optimization
- Data engineer: lot of programming and databases, some statistics
- Also: quant researcher, statistician, quantitative analyst, software engineer ...

Technical stuff

- Programming: mostly R / Python, sometimes Scala/C++ for performance
- Databases: mostly use SQL, also Hadoop, etc

What to do

- Learn to program (try DataCamp)
- Work with open datasets (shoutout Data For Democracy)
- Consider getting a masters/PhD, turns out this is actually really hard