Introduction to Data Science

Jay Urbain, PhD

Topics

- Course introduction
- Data Deluge
- What is data science?
- Data science skills
- Role of data scientist
- Need for data science
- Why become a data scientist
- eScience
- Application examples

Introduction to Data Science

- This course provides an introduction to applied data science including data preparation, exploratory data analysis, statistical inference, predictive modeling, factor analysis, and data visualization.

- Review:
  - Course description
  - Course outcomes
  - Grading policy (syllabus)
  - Topics (next slide)

- Structure:
  - Lectures
  - ~weekly hands-on lab assignments using Jupyter notebooks + final 2-week project
  - Quizzes as needed
  - Midterm, final

Data Deluge

- The availability and cost of collecting and storing data has plunged.
- Businesses, governments, and society are trying to tap its vast potential, gain competitive advantage.

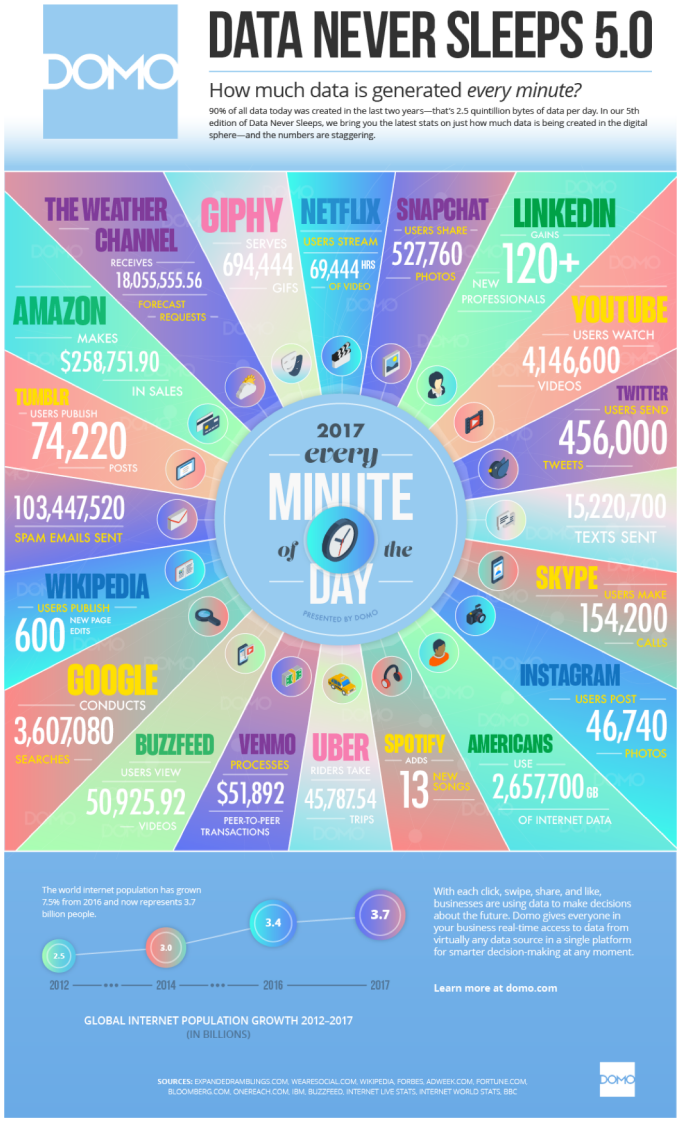
How much data do we create?

- There are 2.5 quintillion bytes (10^{18}) of data created each day at our current pace.
- Pace is only accelerating with the growth of the Internet of Things (IoT).
- Over the last two years alone 90 percent of the data in the world was generated.
- Every time we interact with the Internet, we generate data.
  - Google search
  - Text messages
  - 32 billion people are active on Facebook **daily**
  - Continuous location update
  - ???

Forbes, May 21, 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#2959a86f60ba

How much data do we create?

https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1



What is Data Science?

- "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."
- "The next sexy job"
  - Hal Varian, Google's Chief Economist, NYT, 2009:

- "They need to find nuggets of truth in data and then explain it to the business leaders"
  - Richard Snee, EMC

- "Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information."
  - Jeffrey Stanton, Syracuse University

- Extract insights from messy data
  - Joel Grus

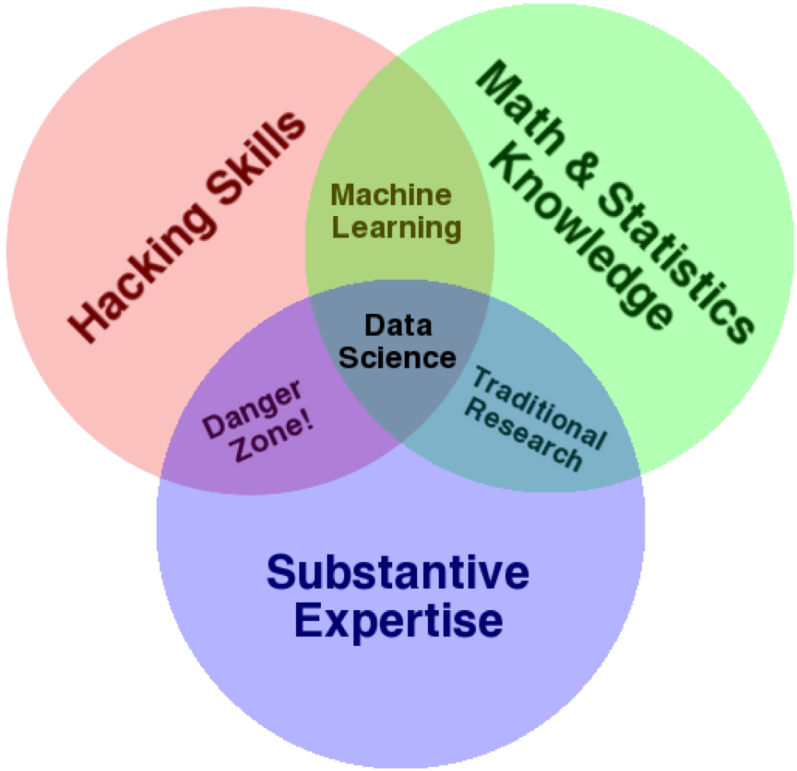
- Extract actionable knowledge from data.
  - Jay Urbain, MSOE ;-)

Data Science Skills

- Data Munging - parsing, scraping, formatting, querying data
- Statistics - traditional analysis
- Machine learning – modeling, prediction
- Communication + Visualization – reports, tables, graphs, etc.
- Scientific –
  - Scientific discovery and building knowledge
  - Requires motivating questions about the world and hypotheses that can be brought to data and tested statistically.
  - search for answers, prove/disprove hypothesis from data.

Data Science Venn Diagram

- Drew Conway, 2013.

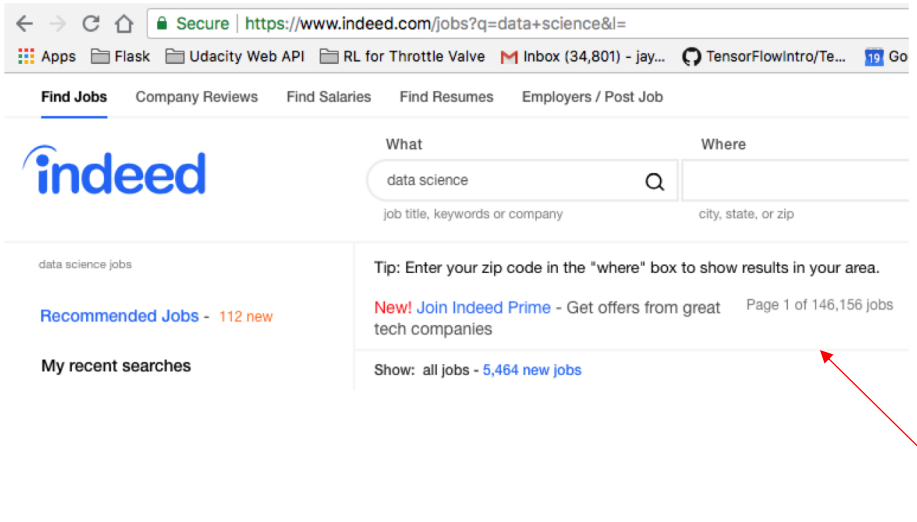


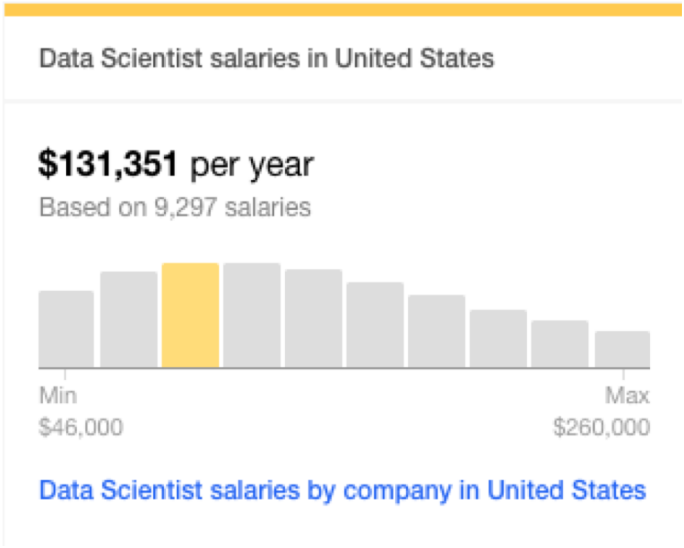
Role of Data Scientist

- Data scientists assigned to point projects
- Here's a question. Is there a "signal" in the data to help us answer that question.
- Need to integrate data sources
  - ETL pipeline is too heavy weight.
- Do predictive modeling from signal(s)
  - Monitor data streams to predict equipment failure.
  - Buy Google keywords – what keywords to buy?

Why Data Science?

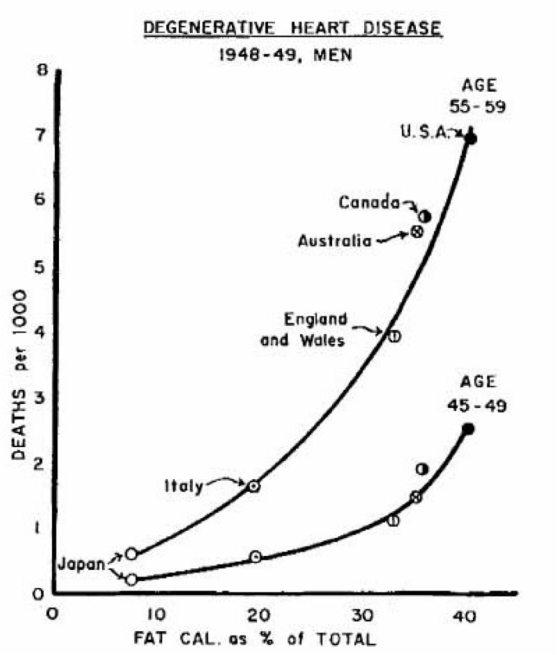
- Knowledge discovery is fun
- Opportunities





Data makes everything clearer

- Seven Countries Study (Ancel Keys, UCB **1925,28**)
- 13,000 subjects total, 5-40 years follow-up.



Data Science: Why all the Excitement?

e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

Note: Did not hold. As more people heard about Google Flu Trends, more people searched.

Why the all the Excitement?

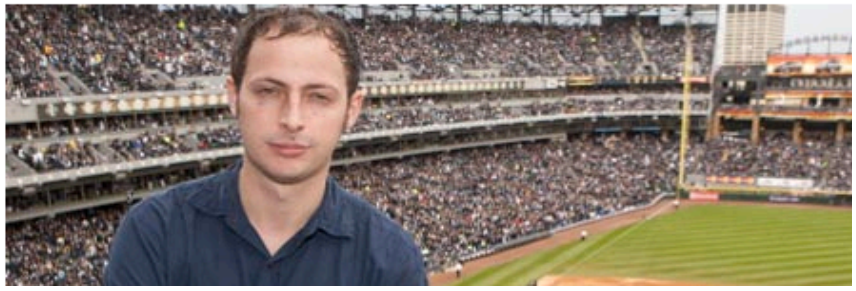
elections2012

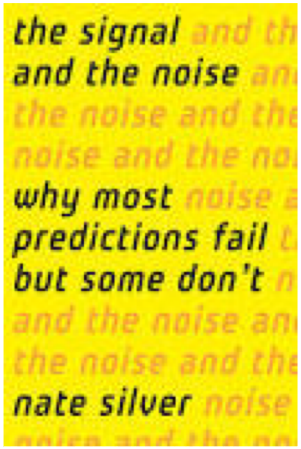
Live results | President | Senate | House | Governor | Choose your

Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST

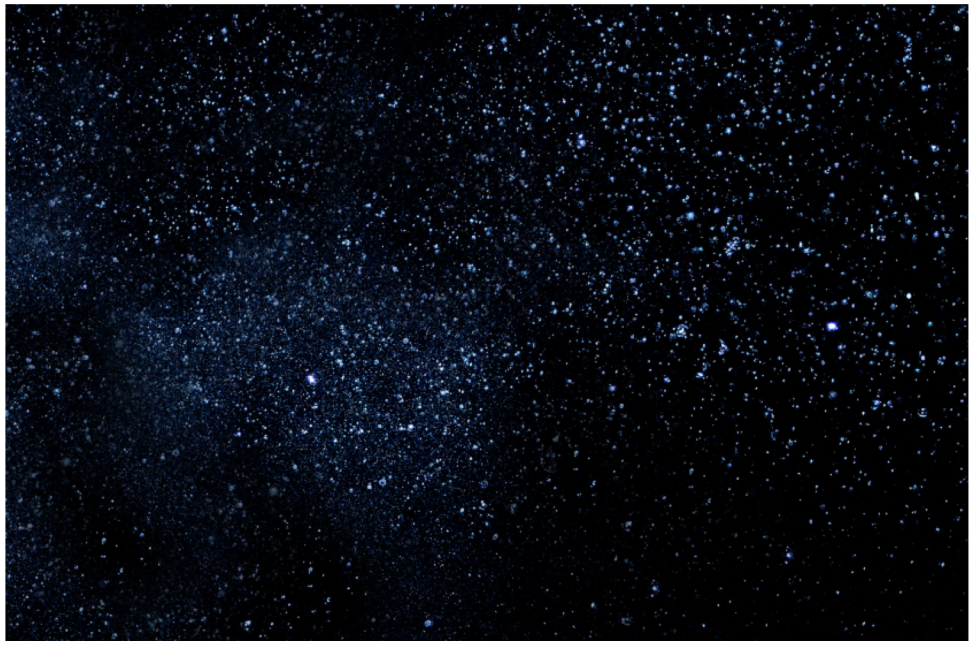




Revolutionized politics. Silver did not do well for 2016 election, but others did.

A long time ago, science was only empirical

- People counted things – stars, crop yield, ROI



A few hundred years ago, scientists developed theoretical approaches.

- Derive equations to describe phenomena

1. $\nabla \cdot \mathbf{D} = \rho_V$
2. $\nabla \cdot \mathbf{B} = 0$
3. $\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$
4. $\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J}$

$$T^2 = \frac{4\pi^2}{GM} a^3$$

can be expressed as simply

$$T^2 = a^3$$

If expressed in the following units:

- T Earth years
- a Astronomical units AU (a = 1 AU for Earth)
- M Solar masses $M_\odot$

Then $\frac{4\pi^2}{G} = 1$

$$H(t)|\psi(t)\rangle = i\hbar \frac{\partial}{\partial t}|\psi(t)\rangle$$

eScience = Data Science

- Empirical
- Theoretical
- Computational
  - Computational methods of discovery

Illumina HiSeq 2000 Sequencer

~1TB/day

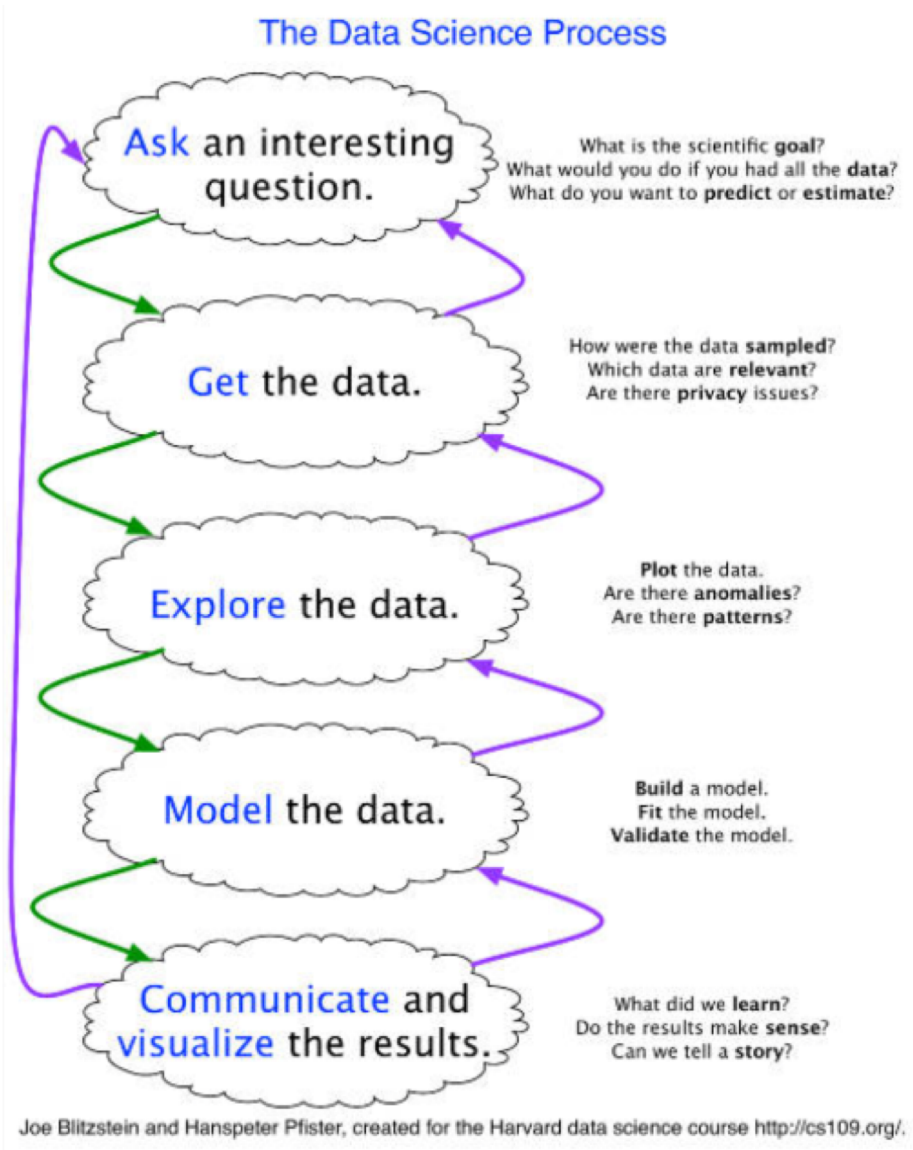
Major labs have 25-100 of these machines

eScience is about the analysis of data

- The automated or semi-automated extraction of knowledge from massive volumes of data.
- Simply too much of it to look at.
- Too complex to understand hidden relationships.
- It's not just a matter of volume. The Three V's of Big Data:
  - Volume: number of rows / objects / bytes
  - Variety: number of columns / dimensions / sources
  - Velocity: number of rows / bytes per unit time
- More V's: - Veracity: Can we trust this data?
- Data analysis has replaced data acquisition as the new bottleneck to discovery.

Computational approach

- Knowledge discovery



Example 1: Medical Costs

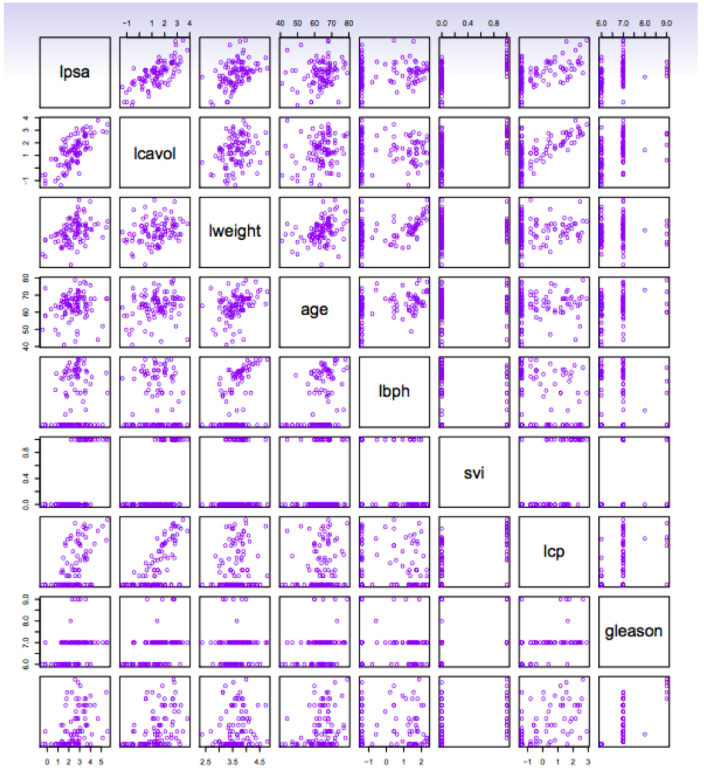
MGH Cancer Center

- Where do patient costs come from?
- Is there some property for treating the most expensive patients?
- Are they sicker?
  - **Really no difference in degree of sickness.**
- So what are the factors that represents this cost factors?
  - **Additional treatment that is not making the patients better.**

Example: Disease Management

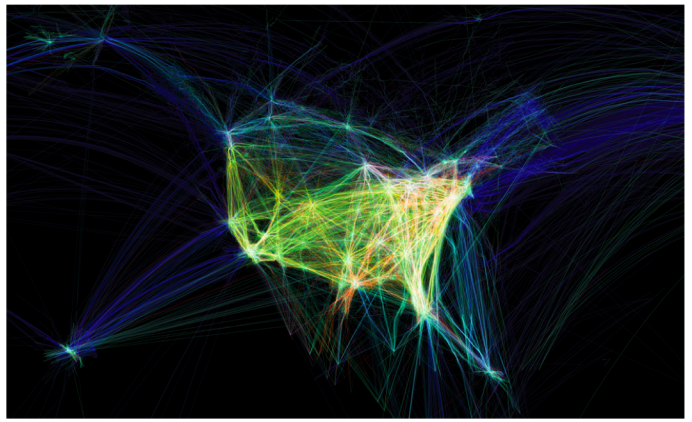
- Given a patient with a history of a progressive disease, can we identify similar patients in earlier stages?
- Can we treat patients prospectively and achieve better outcomes?

Paired indicators of Prostate Cancer



Example: Flight prediction/flight delays

- Weather, congestion, flights delayed, mechanical failure.
- Understanding why a flight is delayed: challenging.
- Big Data problem – many types of information to integrate, and diverse data sets disagree.
- People need help solving this integration and data *veracity* problem.

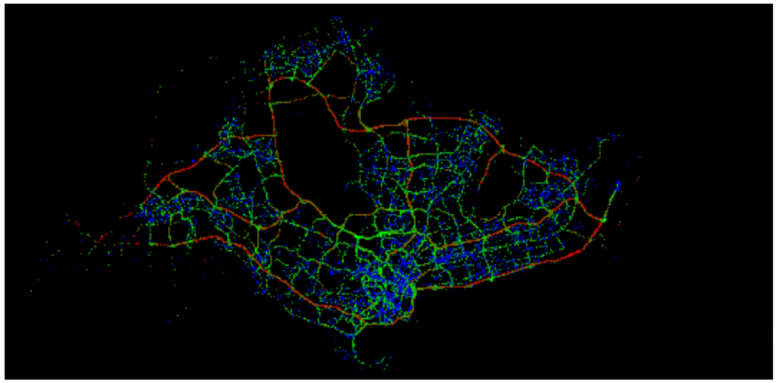


Example: Data analytics for transportation

Daniel Rus, MIT

- Time in transportation has doubled or tripled in some cases
- Data-driven transportation via data streams
  - Modeling, prediction, and controlling
- Improve level of service
- Enhance sustainability
- Personalization
- Optimization

Traffic Visualization: Speed/ Congestion

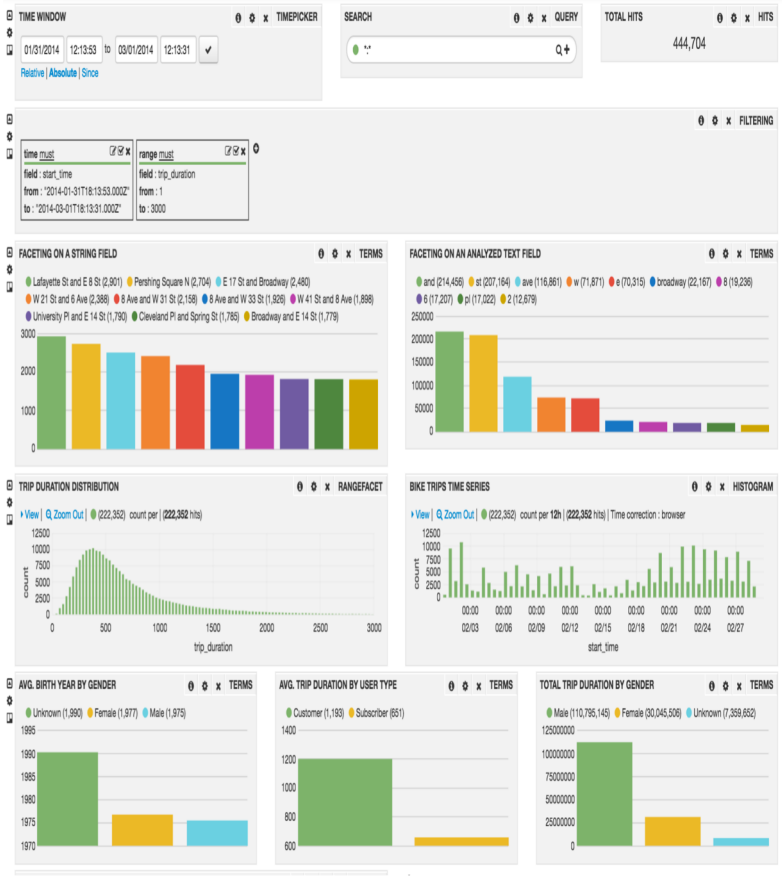


Slow Medium Fast

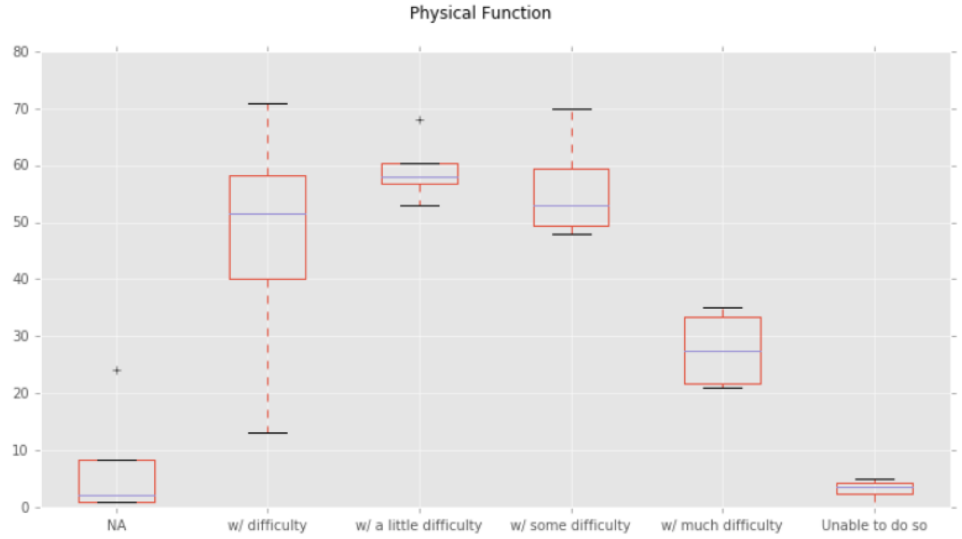
Example: Data analytics for bike sharing

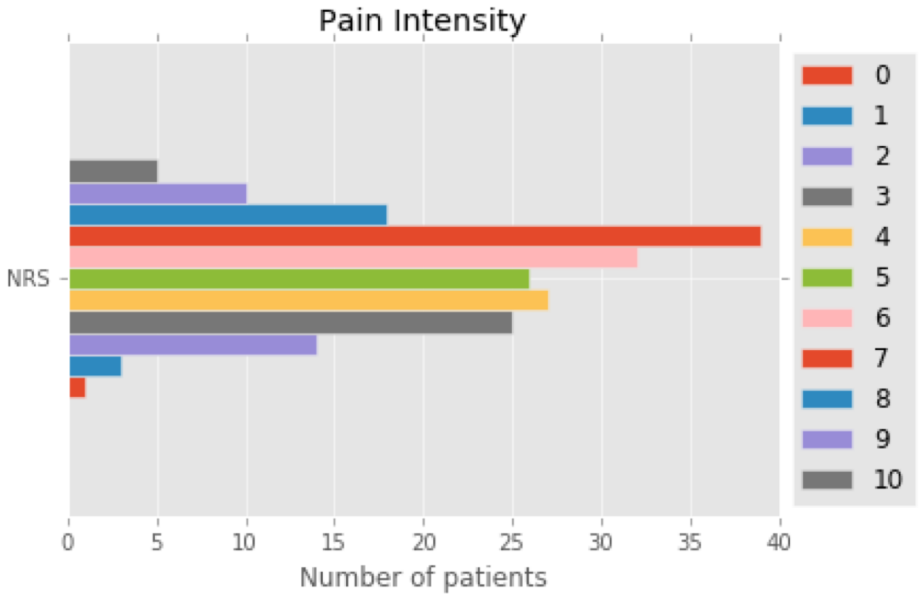
Faceting by:

- Time
- Location
- Date
- Gender
- Station
- Duration

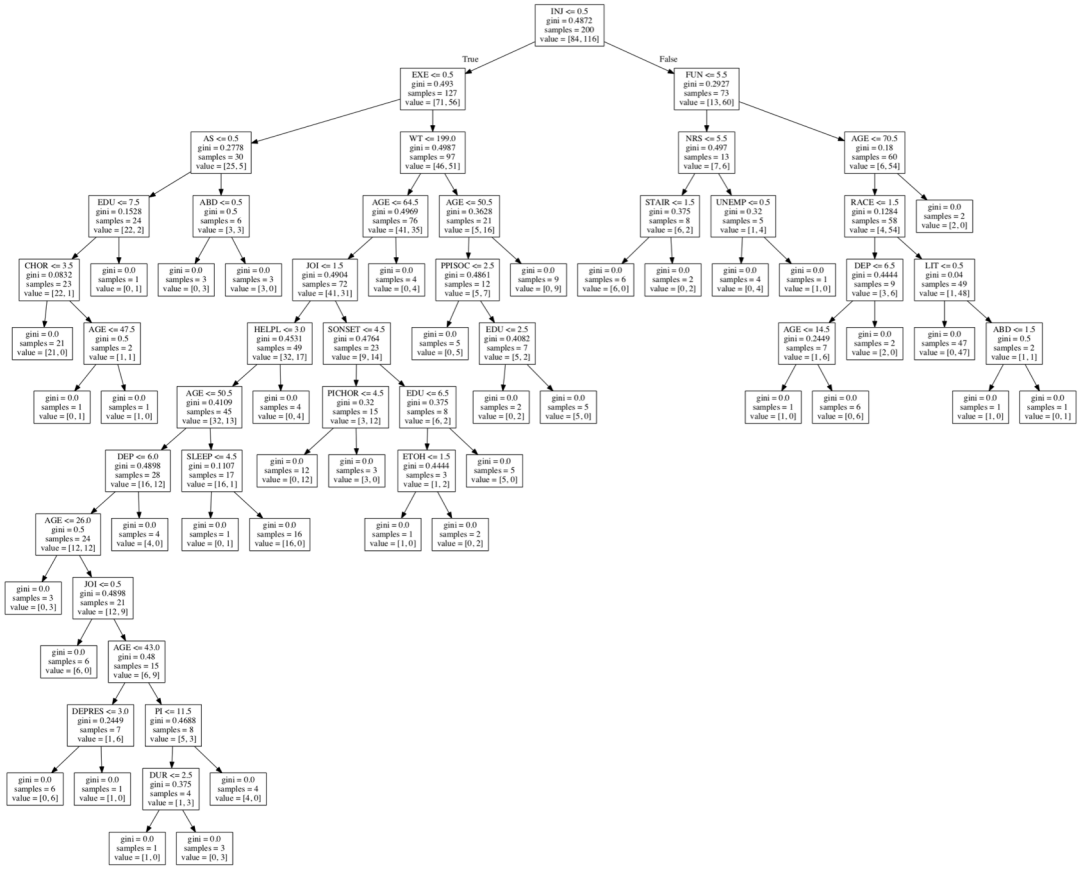


Example: Analyzing chronic back pain data



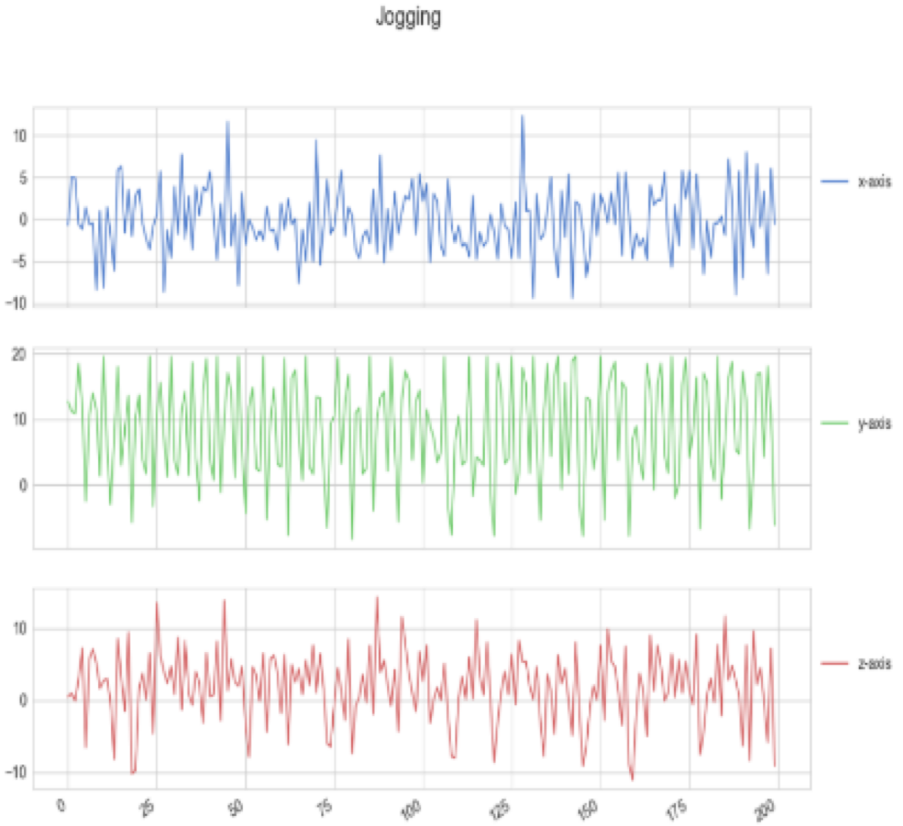


Example: Analyzing comorbid factors



Temporal sensor analysis and activity tracking

	user	activity	timestamp	x-axis	y-axis	z-axis
0	33	Jogging	49105962326000	-0.694638	12.680544	0.503953
1	33	Jogging	49106062271000	5.012288	11.264028	0.953424
2	33	Jogging	49106112167000	4.903325	10.882658	-0.081722
3	33	Jogging	49106222305000	-0.612916	18.496431	3.023717
4	33	Jogging	49106332290000	-1.184970	12.108489	7.205164



Example: And ultimately, the web graph

- Web graph abstraction
- *Everything is interconnected*