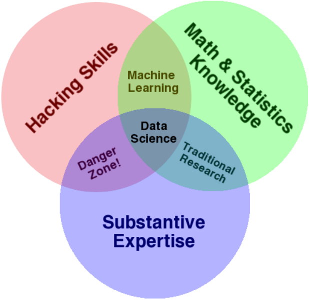
Exploring the World with Data Science

Hal Snyder
drxyzzy@gmail.com
Contributor, SageMathInc.
cloud.sagemath.com

October 29, 2016

Outline



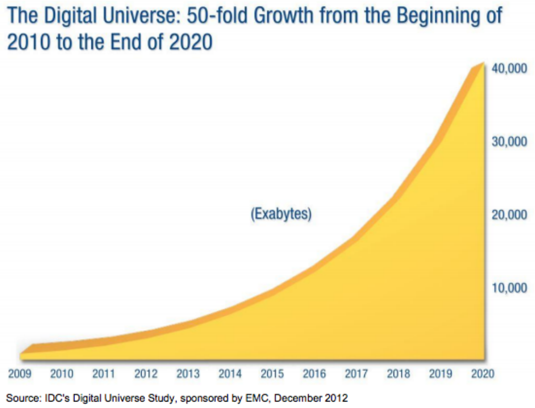
- Examples and Definitions
- No Programming Necessary
- Hints for Aspiring Data Scientists
- Conclusion

Definitions

- **Data Science:** An interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.
- **Machine Learning:** The subfield of computer science that gives computers the ability to learn without being explicitly programmed.
- **Big Data:** A term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.
- **Exploratory Computing:** frequent execution of small fragments of code, in an iterative cycle where code is run to obtain partial results that inform the next bit of code to be written.

The first three definitions are from Wikipedia. Last one is a quote from Fernando Perez, the creator of IPython / Jupyter.

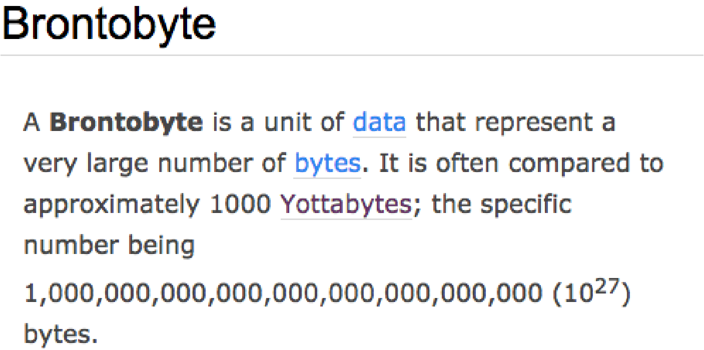
The Ubiquitous Complexity Curve



`www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf`

This shape applies for: all data, cell phone data, data science jobs, research papers, etc. **Why is this a thing now?** That is something to ask whenever a new field of specialization emerges.

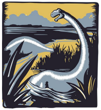
Welcome to the Age of the Brontobyte



Brontobyte

A **Brontobyte** is a unit of data that represent a very large number of bytes. It is often compared to approximately 1000 Yottabytes; the specific number being 1,000,000,000,000,000,000,000,000,000 (10^{27}) bytes.

http://www.computerhope.com/jargon/b/brontobyte.htm
image: https://commons.wikimedia.org/wiki/Category:Brontosaurus



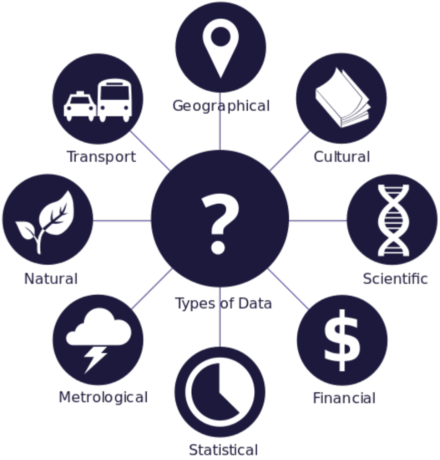
$$y = e^x$$

$$\frac{dy}{dx} = e^x$$

$$\frac{dy}{dx} = y$$

What does this tell you about x and y?

Where Does Data Science Apply?



https://commons.wikimedia.org/wiki/File:Data_types_-_en.svg

medicine business science communications politics sustainability

Example - Elections - Psephology

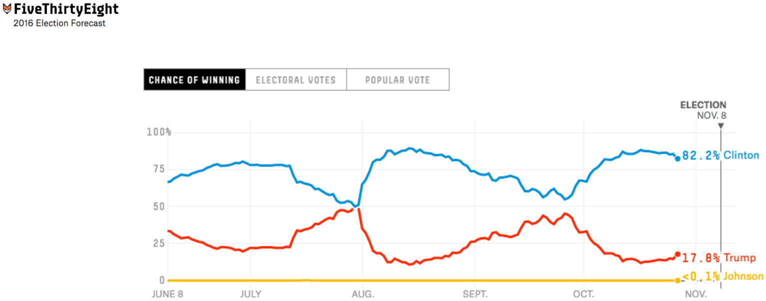
Psephology

Psephology (from Greek psephos, 'pebble', as the Greeks used pebbles as ballots) is a branch of political science which deals with the study and scientific analysis of elections. - *Wikipedia*

Nate Silver successfully called the outcomes in 49 of the 50 states in the 2008 U.S. Presidential election and predicted the winner of all 50 states and the District of Columbia in the 2012 election.

The website *FiveThirtyEight.com* was founded by Nate Silver in 2008. It is a popular source of statistical predictions in politics and sports, as well as offering stats-based commentary on a wider range of subjects.

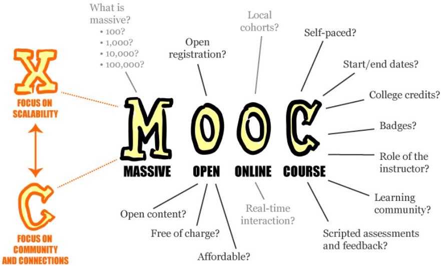
Example - Elections - FiveThirtyEight.com



`http://projects.fivethirtyeight.com` on October 27

Data Science is 25% communication.

A Word About MOOCs



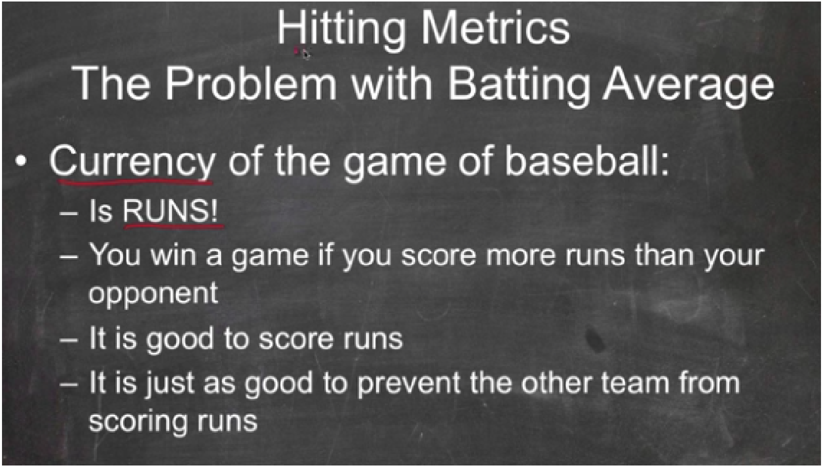
By Mathieu Plourde https://commons.wikimedia.org/w/index.php?curid=26072198

edX.org
coursera.org

complexityexplorer.org
sdgacademy.org

Example - Sabermetrics - Moneyball

Scouts vs. Stats

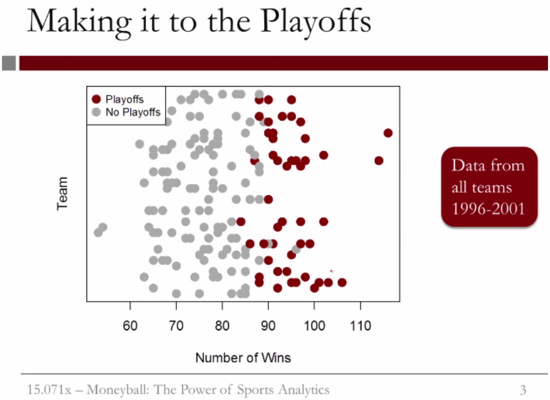


Hitting Metrics
The Problem with Batting Average

- Currency of the game of baseball:
  - Is RUNS!
  - You win a game if you score more runs than your opponent
  - It is good to score runs
  - It is just as good to prevent the other team from scoring runs

SABR101x from Boston University 2015

Example - Sabermetrics - Analytics



Making it to the Playoffs

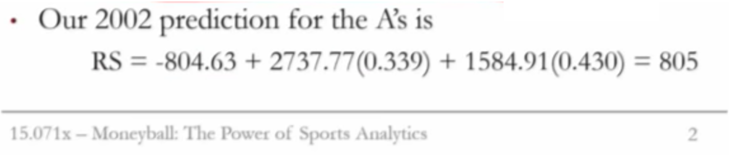
Data from all teams 1996-2001

15.071x – Moneyball: The Power of Sports Analytics 3

MITx: 15.071x The Analytics Edge from MIT 2016

How many games must a team win to make it to the playoffs?
2016 wins: **Cubs 103, Indians 94** for regular season with 162 games.

Example - Sabermetrics - Modeling



- Our 2002 prediction for the A's is

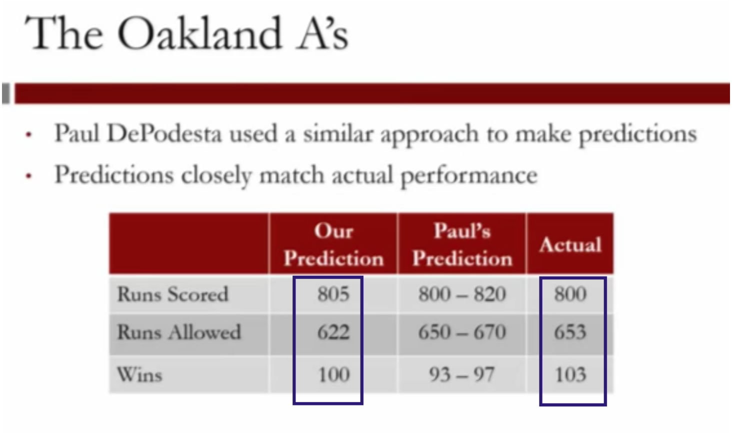
$$RS = -804.63 + 2737.77(0.339) + 1584.91(0.430) = 805$$

15.071x – Moneyball: The Power of Sports Analytics 2

MITx: 15.071x The Analytics Edge from MIT 2016

Two stats with most predictive value for wins:
- **OBP** - percentage player gets on base (inc. walks)
- **SLG** - how far player gets around bases on his turn

Example - Sabermetrics - Predicting



The Oakland A’s

- Paul DePodesta used a similar approach to make predictions
- Predictions closely match actual performance

	Our Prediction	Paul’s Prediction	Actual
Runs Scored	805	800 – 820	800
Runs Allowed	622	650 – 670	653
Wins	100	93 – 97	103

MITx: 15.071x The Analytics Edge from MIT 2016

Predicting Wins for a Given Team Makeup

- Model number of wins as a function of RS (runs scored) and RA (runs allowed)

$$W = 80.881 + 0.106 \times RD$$

- Model RS as function of batter OBP and SLG.

$$RS = -804.63 + 2737.77 \times OBP + 1584.91 \times SLG$$

- Model RA as function of pitcher OOBP and OSLG.

$$RA = -837.38 + 2913.6 \times OOBP + 1514.2 \times OSLG$$

- To predict for year 2002, apply model to player stats for 2001 season.

Data Science Scenario

- Ask a question.
- Look for available data.
- Import data. (Tidying.)
- Explore. What are the variables? Is there noise? Are there missing values?
- Revise question to something the data can answer.
- Split data into training and test sets.
- Create models based on training set. Do this manually or with machine learning.
- Evaluate models based on test set.
- Make decisions based on best models.

Example - Epidemiology - Ebola Outbreak 2014

Looking at a couple more examples...

- The outbreak started with two-year old child who died of the disease in December 2013.
- Approximately 40% of the people who suffered from the disease died.
- Because the outbreak occurred in an urban environment, it spread faster than previous outbreaks, and caused more fatalities, 10,000 by early 2015.
- Hospital workers especially vulnerable. WHO reported that ten percent of the dead were healthcare workers.

Example - Ebola Data Challenges

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data. (Dasu and Johnson, 2003)

Three types of challenges were identified for data scientists attempting to resolve the tragedy:

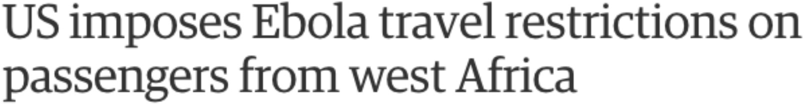
- **Diagnosis.** Months passed between the first Ebola case and its reporting. Tests needed to be cheap, usable by untrained staff, not require electricity, use reagents stable above $104^{\circ}F$, and give quick results.
- **Epidemiological.** On-the-ground surveys, police and hospital reports were collected too slowly to curb the spread of the disease.
- **Treatment.** There is no approved vaccine for Ebola.

Excerpted from Data Science and Ebola, Aske Plaat, Universiteit Leiden, April 13, 2015

https://arxiv.org/abs/1504.02878

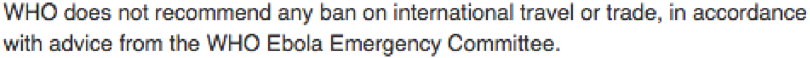
Example - Epidemiology - Travel Restrictions

Headline from *The Guardian* on Tuesday 21 October 2014:



US imposes Ebola travel restrictions on passengers from west Africa

Statement issued by World Health Organization on August 11, 2014, and reiterated on November 7, 2014:

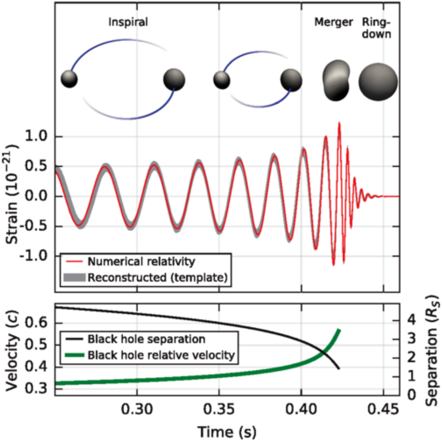


WHO does not recommend any ban on international travel or trade, in accordance with advice from the WHO Ebola Emergency Committee.

Isolation separates sick people with a contagious disease from people who are not sick.

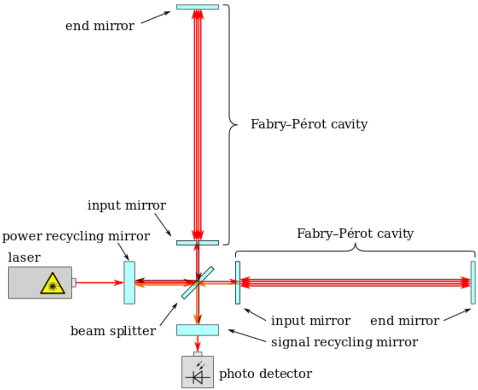
Quarantine separates and restricts the movement of people who were exposed to a contagious disease to see if they become sick.

Example - Physics - Gravitational Waves



https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.116.061102

Example - Physics - Interferometer



https://commons.wikimedia.org/wiki/File:LIGO_simplified.svg

Example - Physics - LIGO

LIGO

On September 14, 2015, the two detectors of the Laser Interferometer Gravitational Wave Observatory simultaneously observed a transient gravitational wave signal - sent after two orbiting black holes spiraled into one another more than a billion light years from Earth. The signal sweeps upwards in frequency from 35 to 250 Hz with a **peak gravitational-wave strain of $\mathbf{1.0 \times 10^{-21}}$**.

The waves from the black hole merger were brief, lasting mere milliseconds. But the output from that collision was **50 times greater than all the power put out by all the stars in the observable universe.**

The paper announcing the result was published in February of 2016 and listed **more than 1,000 co-authors**. LIGO is the largest and most ambitious project ever funded by the NSF.

`http://www.symmetrymagazine.org/article/ligo-sees-gravitational-waves`

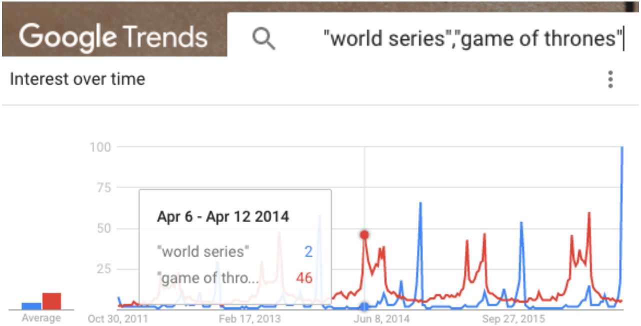
Example - Physics - Data Science and LIGO

- Machine learning is used in LIGO to recognize and predict noise in the data. The AI can recognize noise spikes introduced by airplanes flying overhead or by misalignment of the mirrors along the laser path, and not mistake those as true signals.
- LIGO is so sensitive, it can detect if the 4km distance between its mirrors changes by 1/10,000 the width of a proton, or about 10 zeptometers.
- In addition to more than 80 research institutions, there are more than 44,000 active volunteer LIGO participants running Einstein@home software on personal computers.

No Programming Necessary - Three Links

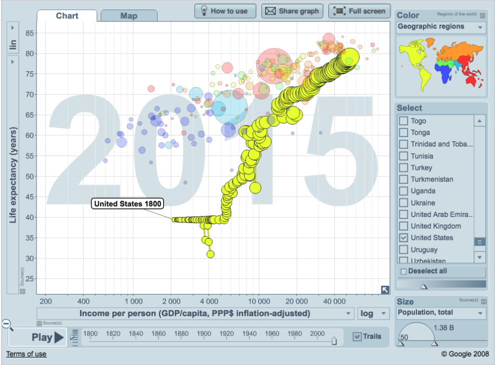
Here are some sites to explore where the visualization tools are already quite advanced.

- Gapminder. `http://www.gapminder.org`
- Zooniverse. `https://www.zooniverse.org/`
- Google Trends. `https://www.google.com/trends/`



Hans Rosling's Gapminder

Here is the end result of running animated display for Wealth & Health of the US from 1800 to 2015:



Color: Geographic regions

Select:
- □ Togo
- □ Tonga
- □ Trinidad and Toba...
- □ Tunisia
- □ Turkey
- □ Turkmenistan
- □ Uganda
- □ Ukraine
- □ United Arab Emira...
- □ United Kingdom
- ☑ United States
- □ Uruguay
- □ Uzbekistan

□ Deselect all

Size: Population, total

☑ Trails

https://www.gapminder.org/world

Look for any of Hans Rosling's inspirational talks at TED.com and his 3-minute youtube video, *The Joy of Stats*.

Zooniverse - Citizen Science

Here is one project offered through the Zooniverse portal:

Welcome to

Cyclone Center

Tropical cyclones are still a mystery.

We need your help to decipher them.

`https://www.cyclonecenter.org`

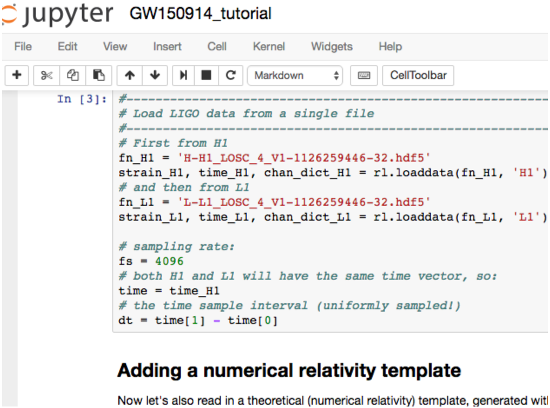
Getting Started in Data Science

Hints for lifelong learners:

- Online courses. Data Science is HOT
  edX.org - 184 of 1293 courses match "data science"
  coursera.org - 488 of 1295 courses match "data science"
- Interactive Notebooks. Game-changers for exploratory computing.
  Jupyter notebooks (used to be IPython)
  R Markdown
  SageMathCloud - aka SMC Sage worksheets)
- **R language** and ggplot plotting system
- **python language**, pandas dataframe toolset, matplotlib for plotting
- Join the data science community. **GitHub, Kaggle, Hacker News, Twitter**

Sample Jupyter Notebook

Here is one of several notebooks available at the LIGO Open Science Center:



In [3]:

```
#-----------------------------------------------------------------
# Load LIGO data from a single file
#-----------------------------------------------------------------
# First from H1
fn_H1 = 'H-H1_LOSC_4_V1-1126259446-32.hdf5'
strain_H1, time_H1, chan_dict_H1 = rl.loaddata(fn_H1, 'H1')
# and then from L1
fn_L1 = 'L-L1_LOSC_4_V1-1126259446-32.hdf5'
strain_L1, time_L1, chan_dict_L1 = rl.loaddata(fn_L1, 'L1')

# sampling rate:
fs = 4096
# both H1 and L1 will have the same time vector, so:
time = time_H1
# the time sample interval (uniformly sampled!)
dt = time[1] - time[0]
```

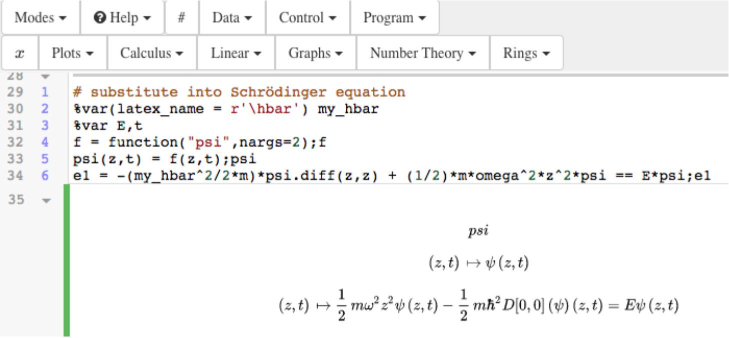
Adding a numerical relativity template

Now let's also read in a theoretical (numerical relativity) template, generated wit

https://losc.ligo.org/s/events/GW150914/LOSC_Event_tutorial_GW150914.html

Sample Sage Worksheet

Here a sample Sage Worksheet showing the ability to do symbolic calculations in theoretical physics:



```
# substitute into Schrödinger equation
%var(latex_name = r'\hbar') my_hbar
%var E,t
f = function("psi",nargs=2);f
psi(z,t) = f(z,t);psi
el = -(my_hbar^2/2*m)*psi.diff(z,z) + (1/2)*m*omega^2*z^2*psi == E*psi;el
```

$$psi$$

$$(z,t) \mapsto \psi\left(z,t\right)$$

$$(z,t) \mapsto \frac{1}{2}\, m\omega^{2} z^{2} \psi\left(z,t\right) - \frac{1}{2}\, m\hbar^{2} D[0,0]\left(\psi\right)\left(z,t\right) = E\psi\left(z,t\right)$$

Online Databases

Just as the trend for Open Source Software emerged in the 1990's, we seem to be at the beginning of a trend for Open Data, and for many of the same reasons. Data is plentiful; the skill to comprehend and interpret it is limited.

- `http://www.seanlahman.com/open-source-sports/` - sports
- `https://www.bioconductor.org/` - genomics
- `http://opendata.cern.ch/?ln=en` - particle physics
- `http://www.unglobalpulse.org/projects` - sustainability

Online Database Portals

Portal sites are emerging, linking to large numbers of data sets.

- https://data.gov/

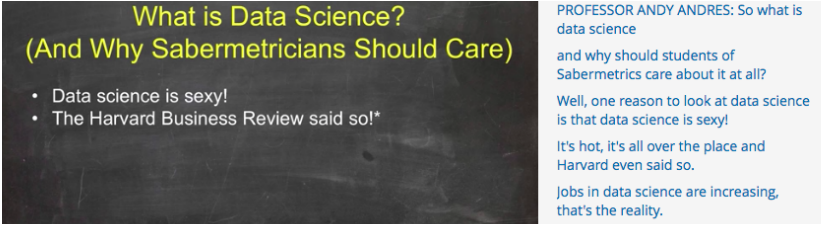
GET STARTED

SEARCH OVER 192,252 DATASETS

- https://data.oecd.org/
- https://www.ncdc.noaa.gov/data-access
- http://www.who.int/gho/en/
- http://www.ipcc-data.org/
- https://data.gov.uk/

Conclusion - Can Data Science Make Us Smarter?

- The complexity of our world is growing exponentially.
- Many of the biggest challenges facing us are interconnected.
- Planetary boundaries represent fixed limits.
- Data Science can help us see more variables and more interdependencies, and make more evidence-based choices.



What is Data Science?
(And Why Sabermetricians Should Care)

- Data science is sexy!
- The Harvard Business Review said so!*

PROFESSOR ANDY ANDRES: So what is data science

and why should students of Sabermetrics care about it at all?

Well, one reason to look at data science is that data science is sexy!

It's hot, it's all over the place and Harvard even said so.

Jobs in data science are increasing, that's the reality.

THE END

THANK YOU
Never stop learning

Hal Snyder
drxyzzy@gmail.com
twitter: @HalDroid
https://www.linkedin.com/in/hsnyder
https://github.com/DrXyzzy
SMC - https://cloud.sagemath.com