

What data scientists really do, according to 50 data scientists

Hugo Bowne-Anderson
[@hugobowne](#)

12/17/2018

A bit about me

- Hugo Bowne-Anderson, data scientist at DataCamp
 - ◆ Undergrad in sciences/humanities (double math major)
 - ◆ PhD in Pure Mathematics (UNSW, Sydney)
 - ◆ Applied math research in cell biology (Yale University, Max Planck Institute)
 - ◆ Python curriculum engineer at DataCamp
 - ◆ Host of [DataFramed, the DataCamp podcast](#)



Today's topics of discussion

- Data science at a high level
- The nitty gritty of data science
- Future data science

Data
Framed

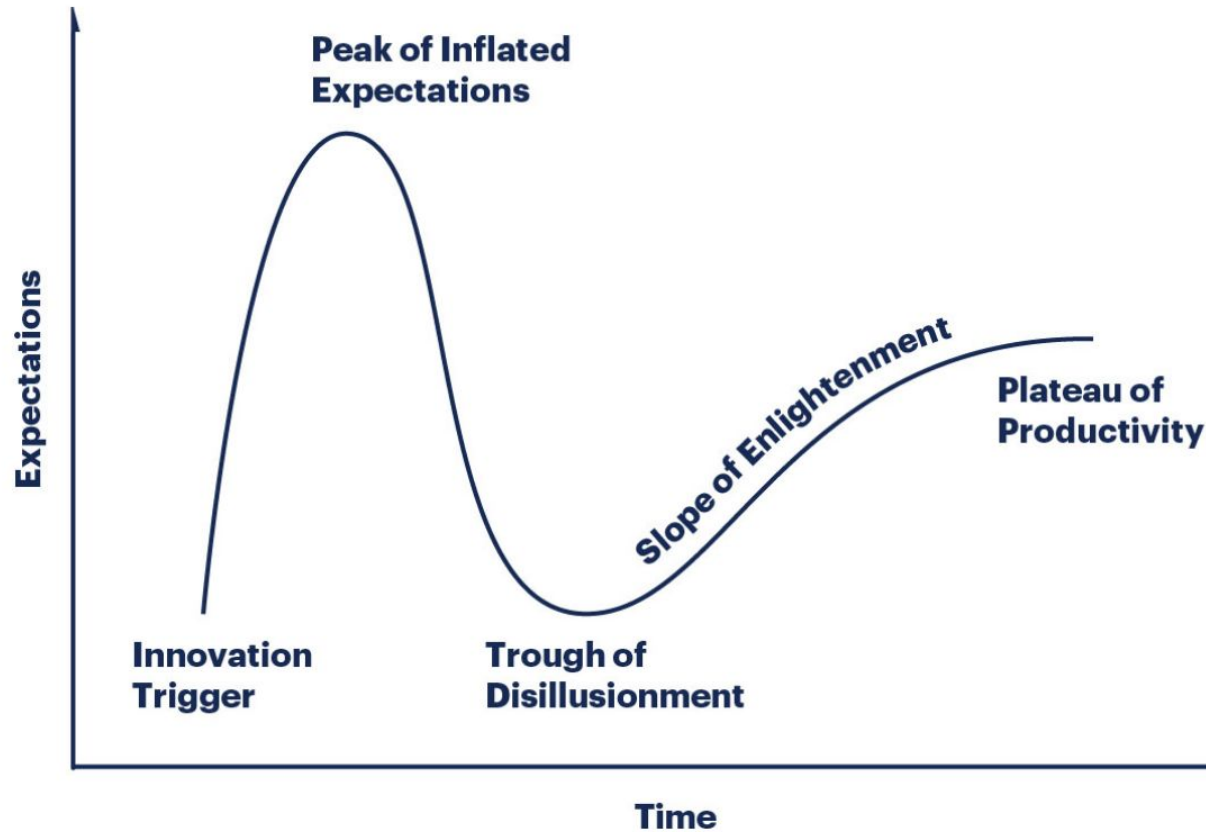
BY  DataCamp

- Data science at a high level
 - ◆ Data science as hype
 - ◆ What data scientists do (take one)
 - ◆ The emergence of modern data science\
 - ◆ Data science and the decision function
- The nitty gritty of data science
- Future data science

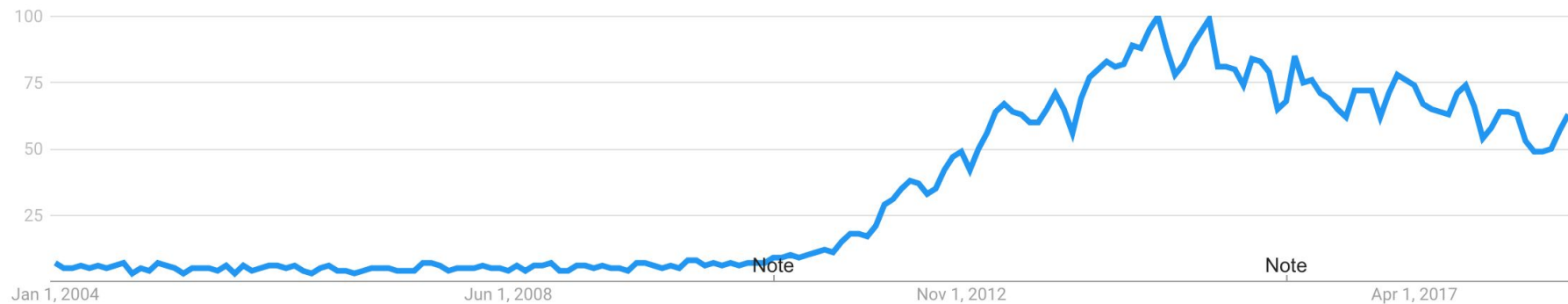
Data
Framed

BY  DataCamp

Where is data science in the Gartner Hype cycle?



Where is big data in the Gartner Hype cycle?

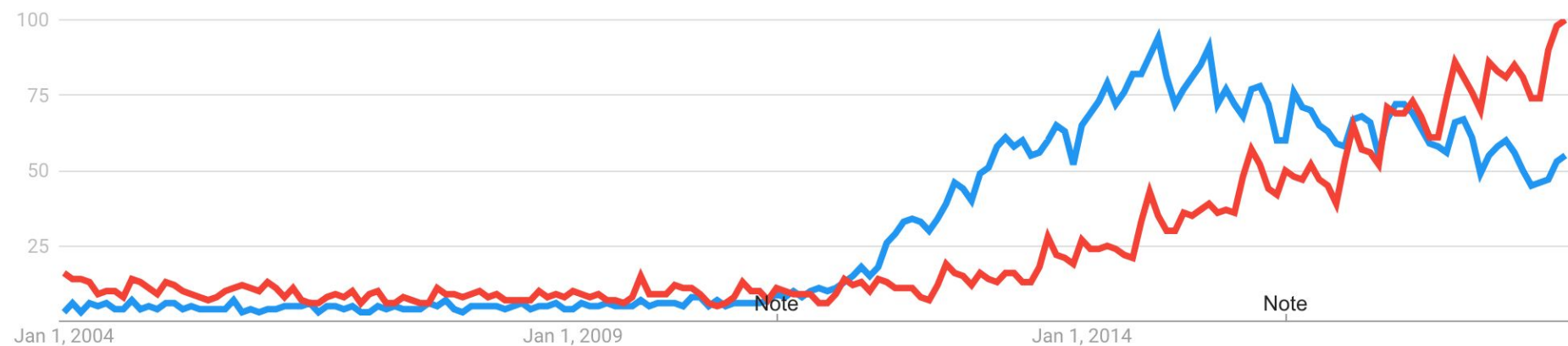


big data

Search term

Google Trends

Where is data science in the Gartner Hype cycle?



big data

Search term



data science

Search term

Google Trends

- Data science at a high level
 - ◆ Data science as hype
 - ◆ What data scientists do (take one)
 - ◆ The emergence of modern data science
 - ◆ Data science and the decision function
- The nitty gritty of data science
- Future data science

Data
Framed

BY  DataCamp

What Data Scientist Actually Do

Today's data scientists do three things...

- Build Venn Diagrams
- Have Imposter Syndrome
- Love Logistic Regression

Venn Diagrams

I google image searched “Venn diagram” and look what I found:



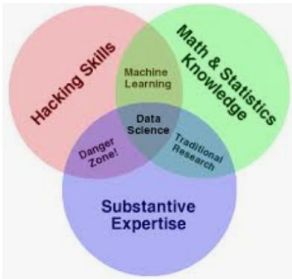
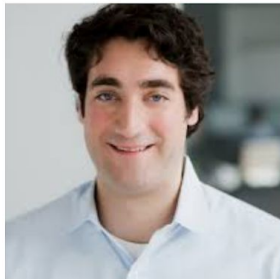

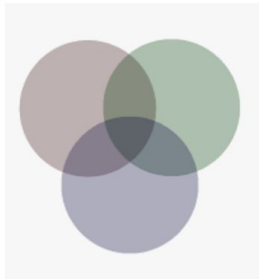
Google search results for "venn diagram".

Search bar: venn diagram

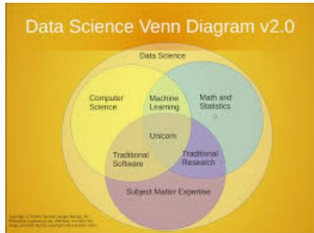
Navigation: All, News, **Images**, Videos, Shopping, More, Settings, Tools

Filters: science venn, data science, conway venn diagram, discipline, data scientist, scientist venn, conway alluvium, danger zone

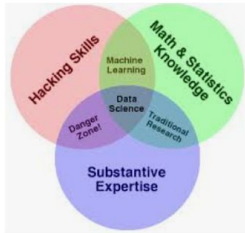
Results:

- 
Drew Conway (@drewconwa...
twitter.com
- 
drewconway (Drew Conway) ...
github.com
- 
The Data Science Venn Diagram ...
drewconway.com
- 
Interview with Drew Conway - A...
simplystatistics.org
- 
Drew Conway
drewconway.com
- 
Drew Conway
drewconway.com

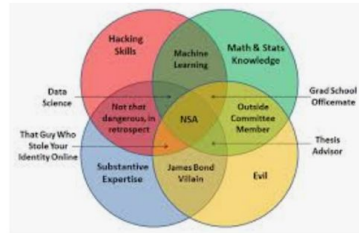
Venn Diagrams



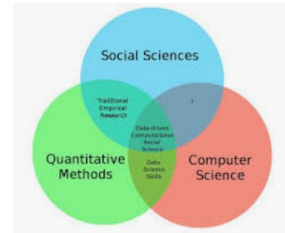
Battle of the Data Science Venn Diagram...
kdnuggets.com



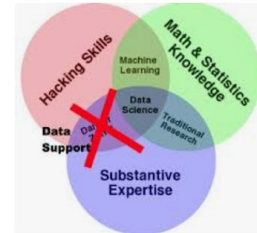
The Data Science Venn Diagram...
drewconway.com



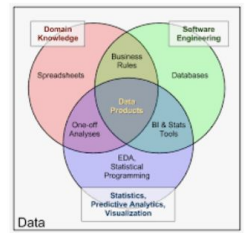
Battle of the Data Science Venn Diagrams
kdnuggets.com



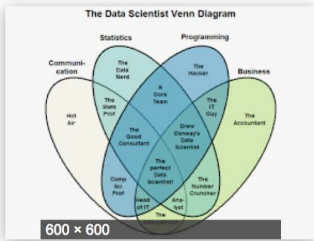
Battle of the Data Science Venn Diagram...
kdnuggets.com



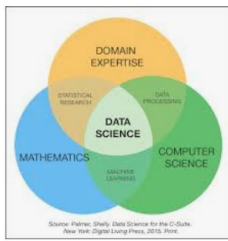
Data Science Venn Diagram
towardsdatascience.com



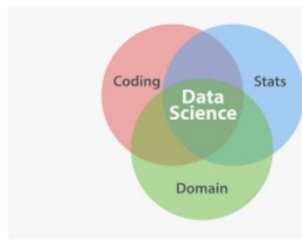
Battle of the Data Science Venn Diagram...
kdnuggets.com



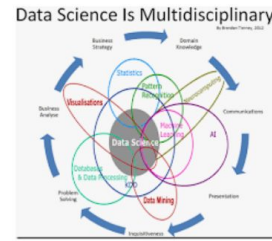
The New Data Scientist Venn Diagram...
whatsthebigdata.com



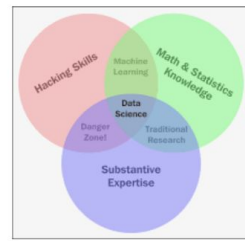
Data Science Venn Diagram...
proofreader.com



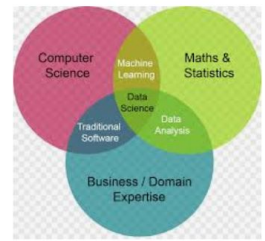
The Data Science Venn Diagram - Data Science...
youtube.com



Battle of the Data Science Venn Diagram...
kdnuggets.com



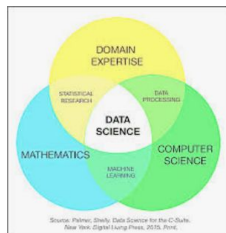
The Data Science Venn Diagram...
hub.packtpub.com



Venn diagram Computer Science...
kisspng.com



Venn diagram of data science unicorn...
science venn diagram - En...



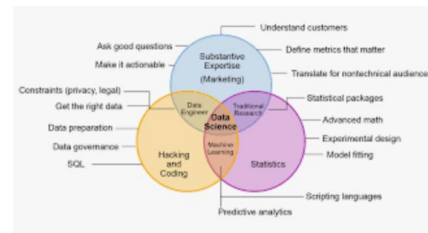
science venn diagram - En...
Data Science Venn Diagram ...



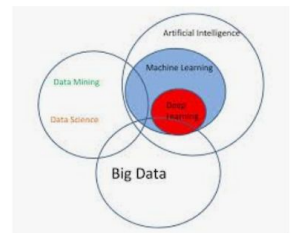
Data Science Venn Diagram ...
The Data Science Venn Diagram...



The Data Science Venn Diagram...
Data Science Venn Diagrams ...

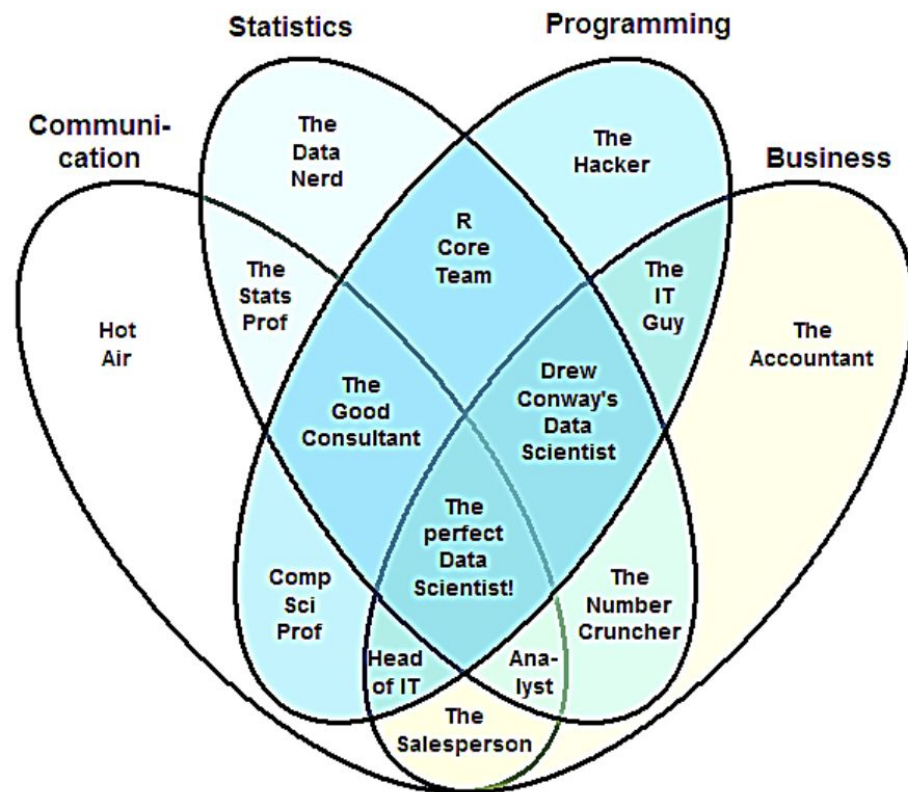


Data Science Venn Diagrams ...
Data Science Venn Diagrams ...

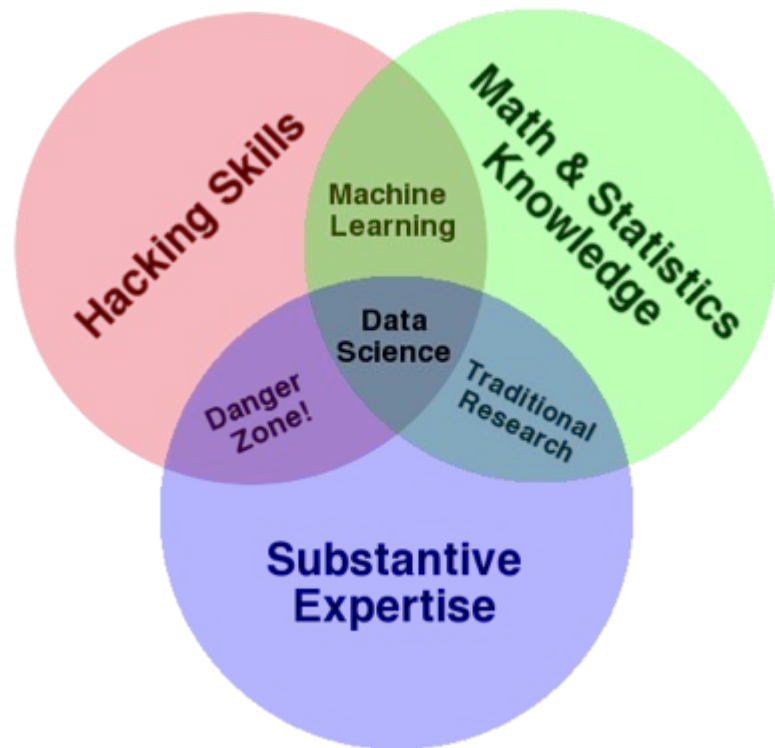


Data Science Venn Diagrams ...

The Data Scientist Venn Diagram



Drew Conway on building data science teams



→ “Data Science” diagram, NOT “Data Scientist” diagram

Data science is ill-defined

- In terms of techniques, career trajectory, job titles
- “Do you think that imprecise ethics, no standards of practice, and a lack of consistent vocabulary are not enough challenges for us today?” -- Hilary Mason
- Brandon Rohrer (Facebook), in [an article](#) called “Imposter Syndrome”:

Our goal isn't to accumulate answers, but to ask better questions. If you are asking questions and using data to find answers, YOU ARE A DATA SCIENTIST. Period.

Brandon

August 22, 2017

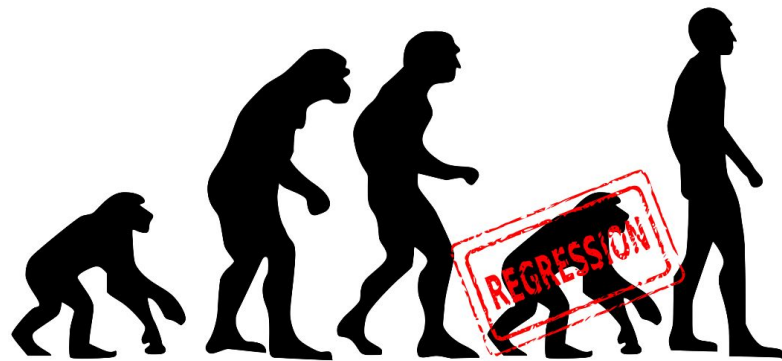
A personal note

- ❖ IPython was born out of deep, hard **failure**
- ❖ The SciPy community saved me!
- ❖ **Impostor syndrome** is **still** very real
- ❖ Scientist, doing software, crossing disciplinary boundaries...

→ Fernando Perez,
co-founder & co-lead of Project Jupyter

Logistic regression

- Simple, communicable, interpretable
- May not perform as well as extreme gradient boosting or a CNN, but what are you optimizing for?
- What is your (business) question?
- Claudia Perlich (DStillery, Two Sigma), Chris Volinsky (AT&T) + many more...

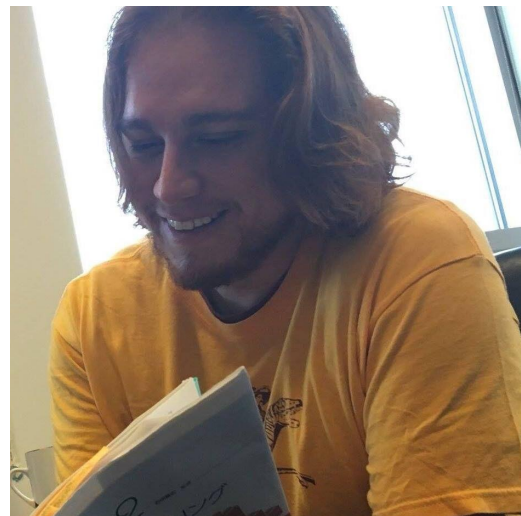


How data scientists think



Humans are horrible at statistics: thought experiment

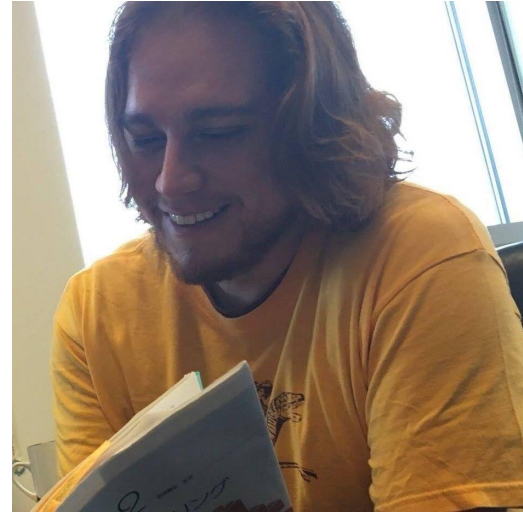
- Police have breathalyzers have a 5% false positive rate, 0% false negative rate.
- 1 in 1,000 drivers are driving drunk.
- Police stop a driver at random & the test is positive.
- What is the probability that the driver is actually drunk?



Michael Betancourt

We are horrible at statistics: thought experiment

- Imagine you do this 1,000 times.
- 999 people will NOT be drunk but, due to false positives, ~50 will test positive.
- 1 person will be drunk and will test positive.
- 1 out of 51 are actually drunk!
- <2% of the positive test results are true positives.
- This is known as [the base-rate fallacy](#).
- It's a data scientist's job to correct these heuristics and biases within organisations AND society.



Michael Betancourt

- FiveThirtyEight, highly reputable pollsters, gave Clinton a 71.4% chance of winning and Trump a 28.6% chance.
- I.e., Trump winning, according to this model, was more likely than flipping two coins and getting two heads, which has a 25% chance.



Allen Downey



Andrew Gelman

Thinking and talking probabilistically



→ The pollsters are getting better!

2018 House Forecast

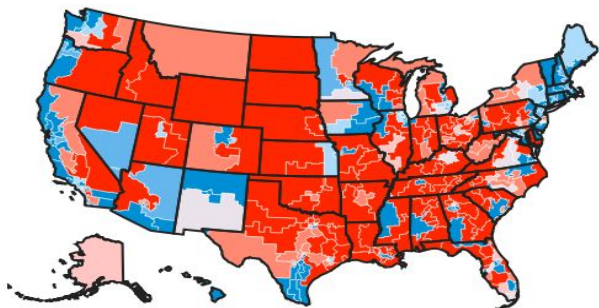
UPDATED NOV. 6, 2018

7 in 8

Chance Democrats
win control (87.9%)

1 in 8

Chance Republicans
keep control (12.1%)



2018 Senate Forecast

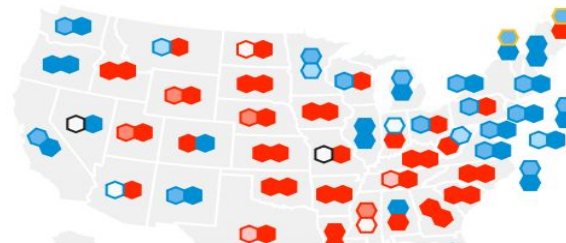
UPDATED NOV. 6, 2018

1 in 5

Chance Democrats
win control (19.1%)

4 in 5

Chance Republicans
keep control (80.9%)



This is normally the point at which you might expect us to give you a throat-clearing “well, actually” about how 1 in 7 chances happen *all the time*. Indeed, they do. One in seven days of the week is a Thursday. None of us woke up this morning screaming “Oh my gosh, I just can’t *believe* it’s a Thursday!” And nobody should really be *that* surprised if Democrats win the Senate next week, or if Republicans keep the House.

[Election Update: Democrats Need A Systematic Polling Error To Win The Senate](#), Nate Silver, FiveThirtyEight

- Data science at a high level
 - ◆ Data science as hype
 - ◆ What data scientists do (take one)
 - ◆ The emergence of modern data science
 - ◆ Data science and the decision function
- The nitty gritty of data science
- Future data science

Data
Framed

BY  DataCamp

- Google
- BuzzFeed
- LinkedIn



1. Lay Solid data foundation to perform robust analytics
2. Use online experiments to achieve sustainable growth
3. Build machine-learning pipelines and personalized data products to make better business decisions



Robert Chang

Airbnb: Customer Lifetime Value (Robert Chang)

Feature Engineering

```
def preprocess_data(df):
    # Convert date to datetime
    df['date'] = pd.to_datetime(df['date'])

    # Extract month and year
    df['month'] = df['date'].dt.month
    df['year'] = df['date'].dt.year

    # Create interaction features
    df['month_year'] = df['month'].astype(str) + df['year'].astype(str)

    # Drop unnecessary columns
    df.drop(['date'], axis=1, inplace=True)

    return df
```

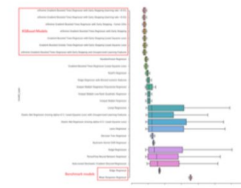
**Feature pipeline
Custom Backfill**

Model Training



**Write transforms
Training**

**Model Selection
&
Model Validation**



**Benchmarking
Compare models**

Productionization



**Virtualenv
Python UDF**

THE DATA SCIENCE **HIERARCHY OF NEEDS**

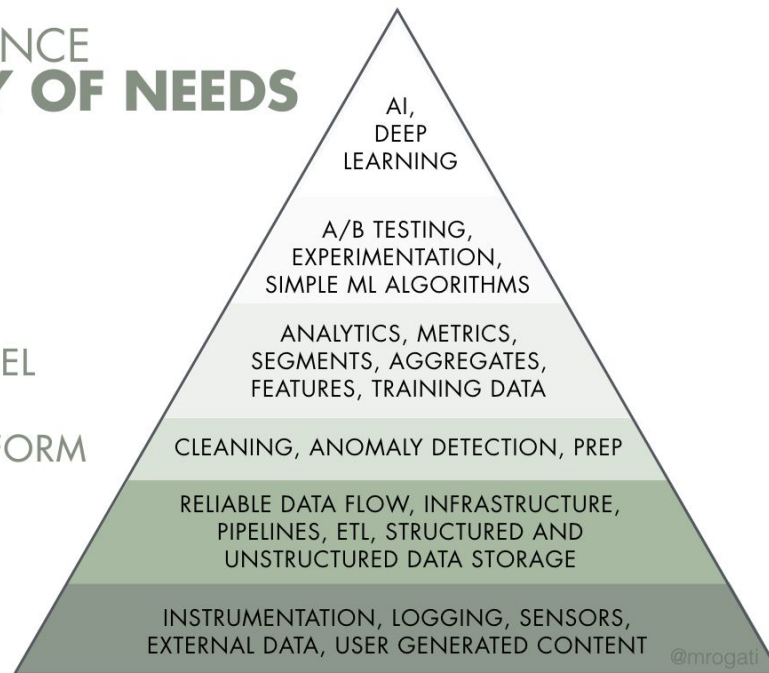
LEARN/OPTIMIZE

AGGREGATE/LABEL

EXPLORE/TRANSFORM

MOVE/STORE

COLLECT



[The AI Hierarchy of Needs, Monica Rogati](#)

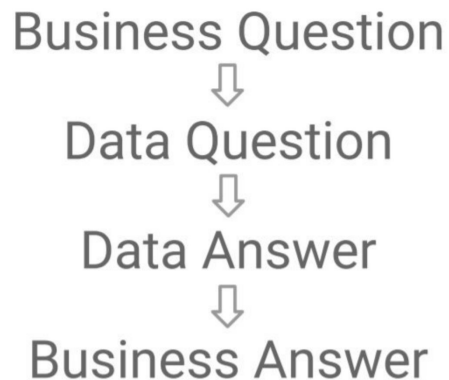
- Data science at a high level
 - ◆ Data science as hype
 - ◆ What data scientists do (take one)
 - ◆ The emergence of modern data science
 - ◆ Data science and the decision function
- The nitty gritty of data science
- Future data science

Data
Framed

BY  DataCamp

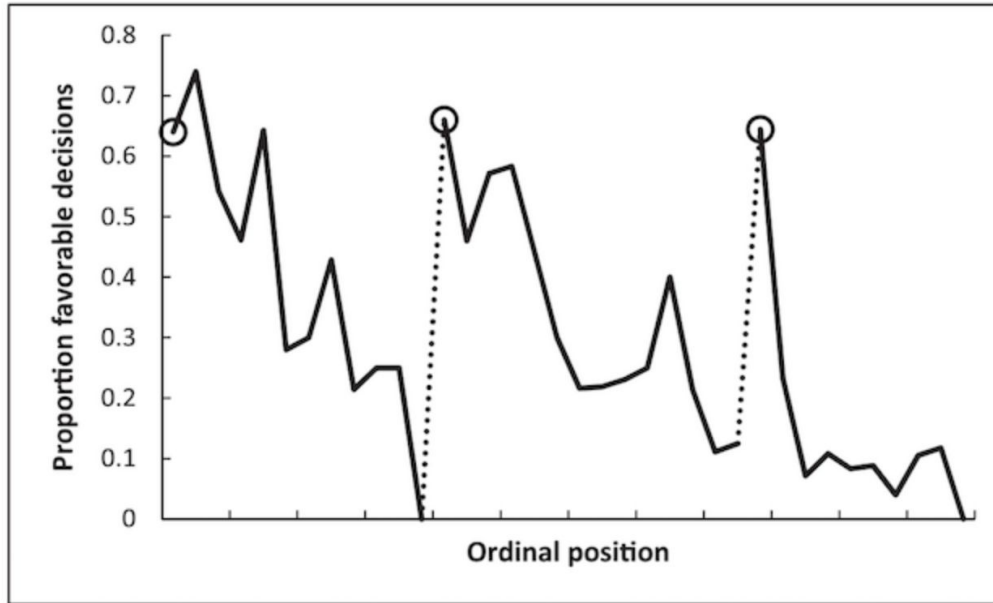
- Data scientists answer business questions & are one of several inputs into the decision-making process.
- Renee Teate (Heliocampus, Becoming a Data Scientist):

Renee's expectation for any data analyst or data scientist:



- Communication skills (and [data translation](#))
- ““Which skill is more important for a data scientist: the ability to use the most sophisticated deep learning models, or the ability to make good PowerPoint slides?” (Jacqueline Nolis, DS consultant & leader, Seattle Area)
- As a data scientist, when do you schedule meetings with decision makers?

When to schedule meetings?



→ [Extraneous factors in judicial decisions](#)
([Danziger et al., PNAS, 2011](#))

Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on x axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

- But wait, there's more:
- “Within each session, unrepresented prisoners usually go last and are less likely to be granted parole than prisoners with attorneys.” -- [Overlooked factors in the analysis of parole decisions](#) (Weinshall-Margel et al., 2011, PNAS)
- Be wary of the narrative fallacy.

Data science embeddings in orgs

- Centre of Excellence: dedicated data science team
- Embedded model: data scientists embedded in every team
- Hybrid model (e.g. Airbnb, DataCamp)
- Data scientist as “decision intelligence” operative



*« Inspiration is cheap, but rigor is expensive.
And if you're not willing to pay up, don't
expect that there is some magical formula
that's going to give it to you. »*

CASSIE KOZYRKOV



BY  DataCamp

Today's topics of discussion

- Data science at a high level
- The nitty gritty of data science
 - ◆ What data scientists actually do
 - ◆ Types of data science
- Future data science

Data
Framed

BY  DataCamp

What Data Scientist Actually Do

Today's data scientist...



Big Data Borat

@BigDataBorat

Following



In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

10:47 PM - 26 Feb 2013

What Data Scientist Actually Do

Today's data scientist...

- Data Collection and Cleaning
- Building dashboards and reports
- Data visualization
- Building models (statistical inference & machine learning)
- Communicate results to stakeholders
- Business decisions are then made!

What Data Scientist Actually Do

This will change...

We should be looking at Automated Machine Learning tools as more like data science assistants, rather than replacements for data scientists.



RANDY OLSON

Data
Framed

BY  DataCamp

Generalist ---> Specialist

- Type A data scientist
 - ◆ Is an analyst, traditional statistician
- Type B data scientist
 - ◆ Is building machine-learning models



Emily Robinson



- “For **consulting** particularly, it makes sense if you are a **generalist**...the skillset of **understanding** what the client **needs** and **figuring out** how to get to the point where you can offer a **solution.**” -- Vicki Boykis



Today's topics of discussion

- Data science at a high level
- The nitty gritty of data science
 - ◆ What data scientists actually do
 - ◆ Types of data science
- Future data science

Data
Framed

BY  DataCamp

Jacqueline Nolis breaks data science into 3 components

1. Business Intelligence
2. Machine Learning
3. Decision Science

1. Business Intelligence (descriptive analytics)

- Taking data company already has
- Getting that data to the right people
- In form of dashboards, reports, emails





« Everyone at any level, C-level, entry level, should be looking and diving into data the same way that you were expected to start using email 20 years ago. »

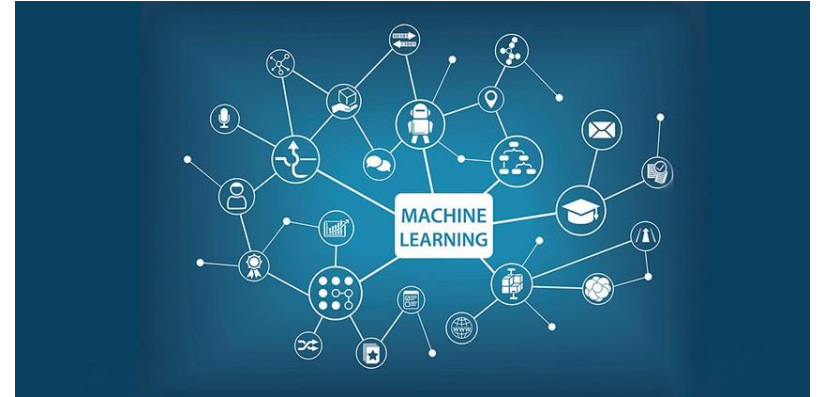
TANYA CASHORALI

Data
Framed

BY  DataCamp

2. Machine Learning (predictive analytics)

- Put models continuously into production
- E.g. LTV at airbnb



3. Decision Making (prescriptive analytics)

- Take the insight discovered from the data science work
- Use it to help company decision making
- E.g. what do you if your data science work tells you a particular type of customer will churn?



Today's topics of discussion

- Data science at a high level
- The nitty gritty of data science
- Future data science
 - ◆ What data scientists need to learn
 - ◆ The future of data science

Data
Framed

BY  DataCamp

Skills To Focus On

- Deep learning
- AI
- Recurrent neural networks



Skills To Focus On

- Asking the right questions
- Turning business questions into data science questions (and answers!)
- Learning on the fly
- Communicating well
- Explaining complex results to non-technical stakeholders
- Critical thinking and quantitative skills will remain in demand



Today's topics of discussion

- Data science at a high level
- The nitty gritty of data science
- Future data science
 - ◆ What data scientists need to learn
 - ◆ The future of data science

Data
Framed

BY  DataCamp

“Do you think that imprecise ethics, no standards of practice, and a lack of consistent vocabulary are not enough challenges for us today?”

Hilary Mason

A person's hand is shown writing on a notepad with a pen. In the background, a laptop is visible on a desk. The entire image is covered with a semi-transparent blue filter. The text "Data ethics or lack thereof" is centered in white.

Data ethics or lack thereof

- We're approaching a consensus that ethical standards need to come from within data science itself, as well as from legislators, grassroots movements and other stakeholders.



“We need to have that ethical understanding, we need to have that training, and we need to have something akin to a Hippocratic Oath.”

Github Senior Machine Learning Data Scientist Omoju Miller

But do we need oaths, checklists and/or codes of conduct? [This conversation](#) is just now heating up!

 DATA SCIENCE

Of oaths and checklists

Oaths have their value, but checklists will help put principles into practice.

By Mike Loukides, Hilary Mason, and DJ Patil. July 17, 2018

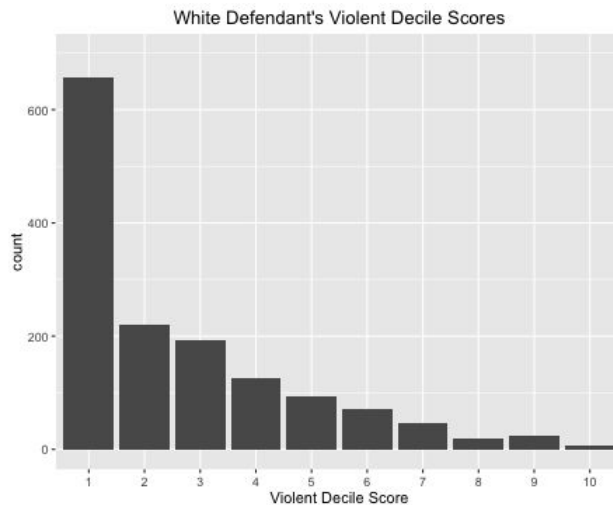
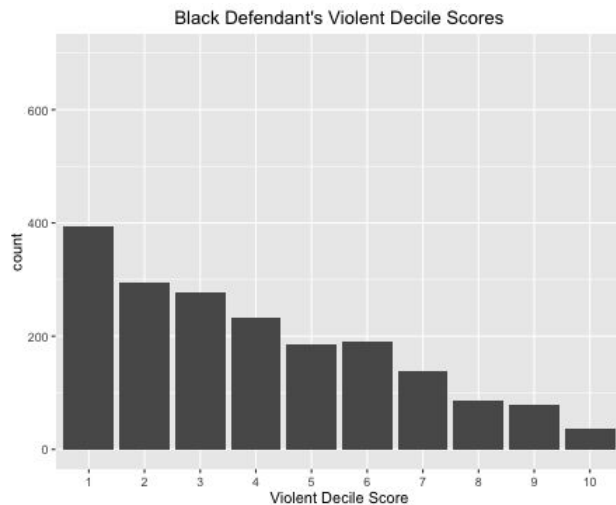
DrivenData has put together [Deon](#), a command line tool, that allows you to easily add an ethics checklist to your data science projects.

Part of this movement involves a re-emphasis on interpretability in models, as opposed to black-box models.



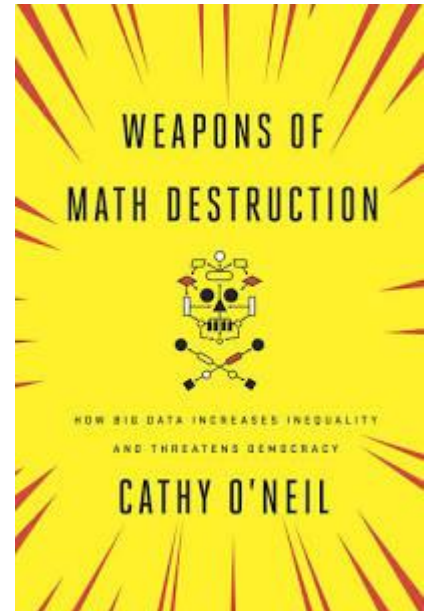
COMPAS Recidivism Risk Score

- Used to predict recidivism rate
- Then used to inform parole decisions



Ethics of Data Science: Cathy O'Neil

- Old definition of DS: “A data savvy, quantitatively minded coding literate problem solver.”
- New definition: “Data science doesn't just predict the future. It causes the future.”
- [Algorithmic audits](#), including a sensitivity analysis
- Ethical matrix: “rows are the stakeholders, the columns are the concerns.”



	Efficiency	Fairness	False +’s	False -’s	Transparency	Predictive Parity	Consistency	Data quality
Court								
Black Defendants								
White Defendants								
Public								
Northpointe								



« We are entering an era where large, complex organizations need to scale interactive computing with data to their entire organization in a manner that is collaborative, secure, and human centered. »

BRIAN GRANGER



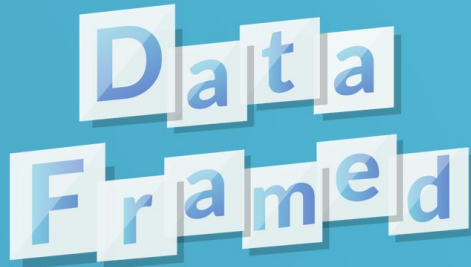
BY  DataCamp

“If you’re using the open source data science stack, then you’re benefiting from others’ work given freely to you. I’d say **it makes perfect sense to give back.**” -- Eric Ma

At the conference & the meetup, I want everyone to feel at home. And that means everyone, including underrepresented groups.



JARED LANDER



BY  DataCamp

Problems with practice

The background image shows a person's hand writing mathematical formulas on a notepad. The formulas include various mathematical symbols and expressions, such as $\frac{1}{x^2}$, $\frac{1}{x^3}$, $\frac{1}{x^4}$, $\frac{1}{x^5}$, $\frac{1}{x^6}$, $\frac{1}{x^7}$, $\frac{1}{x^8}$, $\frac{1}{x^9}$, $\frac{1}{x^{10}}$, $\frac{1}{x^{11}}$, $\frac{1}{x^{12}}$, $\frac{1}{x^{13}}$, $\frac{1}{x^{14}}$, $\frac{1}{x^{15}}$, $\frac{1}{x^{16}}$, $\frac{1}{x^{17}}$, $\frac{1}{x^{18}}$, $\frac{1}{x^{19}}$, $\frac{1}{x^{20}}$, $\frac{1}{x^{21}}$, $\frac{1}{x^{22}}$, $\frac{1}{x^{23}}$, $\frac{1}{x^{24}}$, $\frac{1}{x^{25}}$, $\frac{1}{x^{26}}$, $\frac{1}{x^{27}}$, $\frac{1}{x^{28}}$, $\frac{1}{x^{29}}$, $\frac{1}{x^{30}}$, $\frac{1}{x^{31}}$, $\frac{1}{x^{32}}$, $\frac{1}{x^{33}}$, $\frac{1}{x^{34}}$, $\frac{1}{x^{35}}$, $\frac{1}{x^{36}}$, $\frac{1}{x^{37}}$, $\frac{1}{x^{38}}$, $\frac{1}{x^{39}}$, $\frac{1}{x^{40}}$, $\frac{1}{x^{41}}$, $\frac{1}{x^{42}}$, $\frac{1}{x^{43}}$, $\frac{1}{x^{44}}$, $\frac{1}{x^{45}}$, $\frac{1}{x^{46}}$, $\frac{1}{x^{47}}$, $\frac{1}{x^{48}}$, $\frac{1}{x^{49}}$, $\frac{1}{x^{50}}$, $\frac{1}{x^{51}}$, $\frac{1}{x^{52}}$, $\frac{1}{x^{53}}$, $\frac{1}{x^{54}}$, $\frac{1}{x^{55}}$, $\frac{1}{x^{56}}$, $\frac{1}{x^{57}}$, $\frac{1}{x^{58}}$, $\frac{1}{x^{59}}$, $\frac{1}{x^{60}}$, $\frac{1}{x^{61}}$, $\frac{1}{x^{62}}$, $\frac{1}{x^{63}}$, $\frac{1}{x^{64}}$, $\frac{1}{x^{65}}$, $\frac{1}{x^{66}}$, $\frac{1}{x^{67}}$, $\frac{1}{x^{68}}$, $\frac{1}{x^{69}}$, $\frac{1}{x^{70}}$, $\frac{1}{x^{71}}$, $\frac{1}{x^{72}}$, $\frac{1}{x^{73}}$, $\frac{1}{x^{74}}$, $\frac{1}{x^{75}}$, $\frac{1}{x^{76}}$, $\frac{1}{x^{77}}$, $\frac{1}{x^{78}}$, $\frac{1}{x^{79}}$, $\frac{1}{x^{80}}$, $\frac{1}{x^{81}}$, $\frac{1}{x^{82}}$, $\frac{1}{x^{83}}$, $\frac{1}{x^{84}}$, $\frac{1}{x^{85}}$, $\frac{1}{x^{86}}$, $\frac{1}{x^{87}}$, $\frac{1}{x^{88}}$, $\frac{1}{x^{89}}$, $\frac{1}{x^{90}}$, $\frac{1}{x^{91}}$, $\frac{1}{x^{92}}$, $\frac{1}{x^{93}}$, $\frac{1}{x^{94}}$, $\frac{1}{x^{95}}$, $\frac{1}{x^{96}}$, $\frac{1}{x^{97}}$, $\frac{1}{x^{98}}$, $\frac{1}{x^{99}}$, $\frac{1}{x^{100}}$. A laptop is visible in the background.

- “One assumption people make a lot is that data is objective.” -- Lukas Vermeer, Booking.com.

Many analysts, one dataset: Making transparent how variations in analytical choices affect results

Silberzahn, R; Uhlmann, E L; Martin, D P; Anselmi, P; Ullrich, Johannes

The garden of forking assumptions

- First impressions last and humans look for patterns
- When data scientists perform analyses, they are also prone to confirmation bias
- The same data is often used for inspiration AND validation
 - ◆ “Split your damned data” -- Cassie Kozyrkov
- Humans use confidence and precision as a proxy for accuracy
- The industry lacks consistent workflow (although see projects such as [Cookiecutter data science](#))
- The state of the industry is similar to medicine before randomized control trials.

The current and coming **credibility crisis** in data science

WHAT'S THE SCIENCE IN DATA SCIENCE?

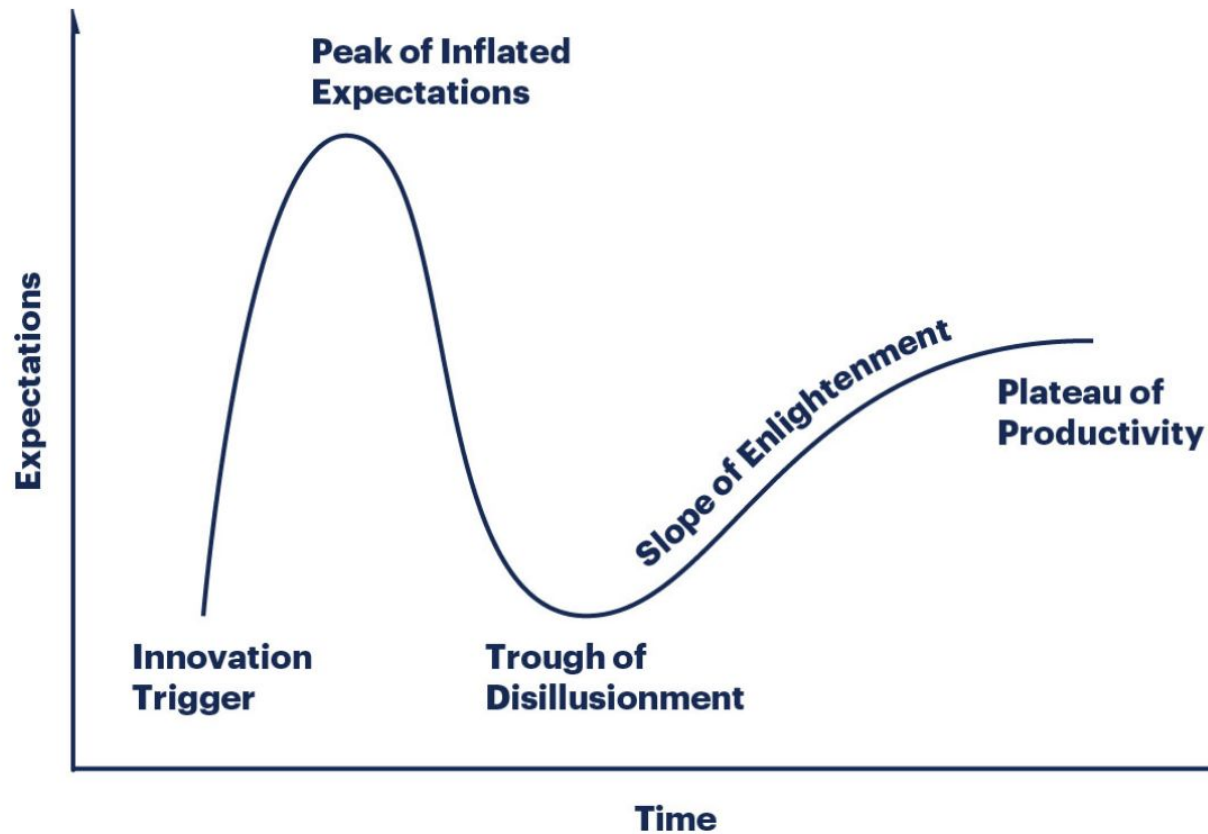
PyData Meetup
Ann Arbor, Michigan
July 19, 2018

Skipper Seabold, Director of Data Science R&D, Product Lead
Civis Analytics
[@jseabold](https://twitter.com/jseabold)

“Will we even have data science in 10 years? I remember a world where we didn’t and it wouldn’t surprise me if the title goes the way of ‘webmaster’.”

Hilary Mason

Where is data science in the Gartner Hype cycle?



What will the slope of enlightenment look like?

“Very few companies expect *only* professional writers to know how to write. So why ask *only* professional data scientists to understand and analyze data, at least at a basic level?”

-- Jonathan Cornelissen, CEO, DataCamp in [HBR](#)

Data literacy is not a either you are or you aren't ready to work with data. It's really a continuum and I think the deeper everyone gets into it, the healthier the entire community can be.



DAVID ROBINSON



BY  DataCamp



DataCamp

Thank you!



Hugo Bowne-Anderson
@hugobowne
@datacamp