

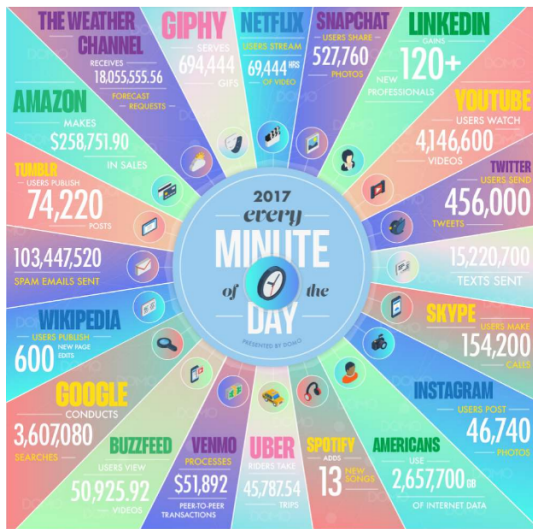
# All Things "Data": Data Science, Data Analytics, Big Data ...

Andrey Skripnikov

Department of Mathematics  
University of Houston

June 28, 2018

# Root of it all: Data.

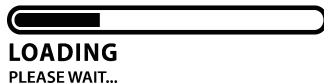


# Big Data: How Big is "Big"?

- **"Big"** is a **moving target**.

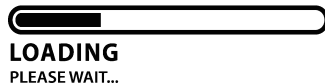
# Big Data: How Big is "Big"?

- "Big" is a **moving target**.
- "Big" - when size becomes **a challenge**.



# Big Data: How Big is "Big"?

- "Big" is a **moving target**.
- "Big" - when size becomes **a challenge**.



have to learn a **new host of tools**.

# Data Science: Tech Companies.



- **Web Search:** How do search engines rank pages?

**Everything is personalized**



- **Online Recommendations:** How do Netflix, Amazon, Ebay recommend items and movies that users might like?

# My experience: "Data Journalism".

Football



... Read More →

## By the numbers: Pirates' defense a sinking ship

**Published on** November 2nd, 2017 | *by Andrey Skripnikov*

UH's motto reads "You are the pride," and its football team epitomized this mantra Saturday by delivering a 28-24 victory over then-No. 17 South Florida — a win to be proud of after suffering back-to-back losses

...

# My experience: "Data Journalism".

## Football



... Read More →

## By the numbers: Pirates' defense a sinking ship

Published on November 2nd, 2017 | by Andrey Skripnikov

UH's motto reads "You are the pride," and its football team epitomized this mantra Saturday by delivering a 28-24 victory over then-No. 17 South Florida — a win to be proud of after suffering back-to-back losses

...

```
> TeamValAndRank("East Carolina",side="Defense",type="Overall",basis="Per Play")
```

	Value	Rank
Points	0.642	130
Penalties	0.111	126
1stDowns	0.354	129
Turnovers	0.013	121
Yards	7.729	130
TDs	0.080	130

```
> PlayerValAndRank("Thomas Sirk",pos = "QB",basis="Per Play")
```

	Value	Rank
Pass Y/A	7.000	68
Pass TD/A	0.038	81
INT/A	0.038	91
Rush Y/A	2.600	47
Rush TD/A	0.047	50



# My experience: Adjust CFB Rankings by Stat. Category.

Issue of College Football: 130 teams, each plays **only ~ 12 opponents** per year, need to **objectively rank ALL of them**.

# My experience: Adjust CFB Rankings by Stat. Category.

Issue of College Football: 130 teams, each plays **only ~ 12 opponents** per year, need to **objectively rank ALL of them**.

**Blue** - **Power 5** conference (traditionally **strong**)  
**Red** - **non-Power 5** (typically **weaker**).

## Classical Averages

	Team	Conference	Totals, Offense
1	Oklahoma	Big 12	587.50
2	Oklahoma State	Big 12	578.92
3	Memphis	American	571.80
4	Louisville	Atlantic Coastal	550.00
5	Central Florida	American	539.90
6	South Florida	American	525.60
7	Ohio State	Big Ten	523.62
8	Arkansas State	Sun Belt	498.90
9	Toledo	Mid-American	497.00
10	Southern Methodist	American	496.73

# My experience: Adjust CFB Rankings by Stat. Category.

Issue of College Football: 130 teams, each plays **only ~ 12 opponents** per year, need to **objectively rank ALL of them**.

**Blue** - **Power 5** conference (traditionally **strong**)

**Red** - **non-Power 5** (typically **weaker**).

## Classical Averages

	Team	Conference	Totals, Offense
1	Oklahoma	Big 12	587.50
2	Oklahoma State	Big 12	578.92
3	Memphis	American	571.80
4	Louisville	Atlantic Coastal	550.00
5	Central Florida	American	539.90
6	South Florida	American	525.60
7	Ohio State	Big Ten	523.62
8	Arkansas State	Sun Belt	498.90
9	Toledo	Mid-American	497.00
10	Southern Methodist	American	496.73

## ADJUSTED Averages

	Team	Conference	Totals, Offense
1	Oklahoma	Big 12	613.17
2	Oklahoma State	Big 12	594.13
3	Louisville	Atlantic Coastal	586.95
4	Ohio State	Big Ten	561.18
5	West Virginia	Big 12	508.73
6	Memphis	American	507.61
7	Central Florida	American	503.10
8	Missouri	Southeastern	501.24
9	Notre Dame	Independent	497.84
10	Syracuse	Atlantic Coastal	497.33

# My experience: What about them Coogs?

Houston Cougars took pride in rush defense, with:

- seniors D'Juan Hines & Matthew Adams (both on NFL rosters now),
- standout sophomore Ed Oliver.

But Houston ranked only #43, with 150 rushing yards allowed per game.

# My experience: What about them Coogs?

Houston Cougars took pride in rush defense, with:

- seniors D'Juan Hines & Matthew Adams (both on NFL rosters now),
- standout sophomore Ed Oliver.

But Houston ranked only #43, with 150 rushing yards allowed per game.

**After the adjustment?**  $\implies$  #13 with 105 yards allowed per game.

# My experience: What about them Coogs?

Houston Cougars took pride in rush defense, with:

- seniors D'Juan Hines & Matthew Adams (both on NFL rosters now),
- standout sophomore Ed Oliver.

But Houston ranked only #43, with 150 rushing yards allowed per game.

**After the adjustment?**  $\implies$  #13 with 105 yards allowed per game.

Why so? My COUGAR BIAS you'd say?

# My experience: What about them Coogs?

Houston Cougars took pride in rush defense, with:

- seniors D'Juan Hines & Matthew Adams (both on NFL rosters now),
- standout sophomore Ed Oliver.

But Houston ranked only #43, with 150 rushing yards allowed per game.

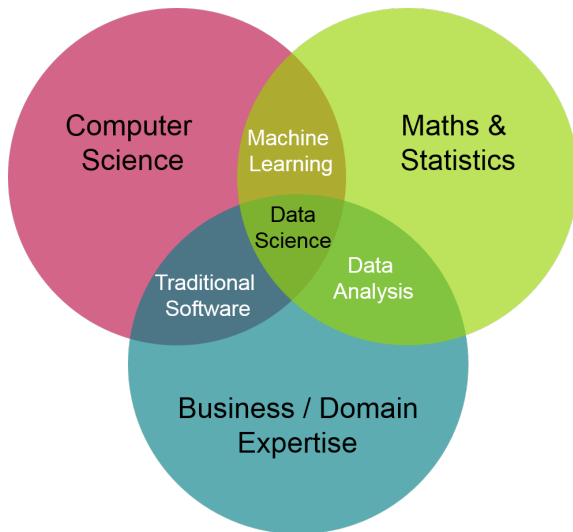
**After the adjustment?**  $\implies$  #13 with 105 yards allowed per game.

Why so? My COUGAR BIAS you'd say?

Not really, see rushing offense ranks for their opponents:

	Name	Rush_Yards_Rank	Rush_Attempts_Rank
1	Arizona	3	13
2	Tulsa	14	7
3	South Florida	8	6
4	Tulane	20	10
5	Navy	2	2
6	(Out of .. teams)	130	130

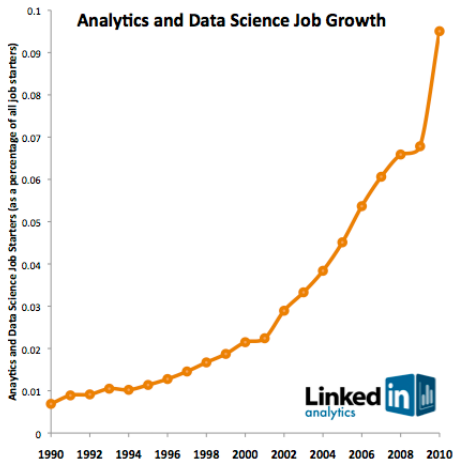
# What is Data Science? Venn Diagram.





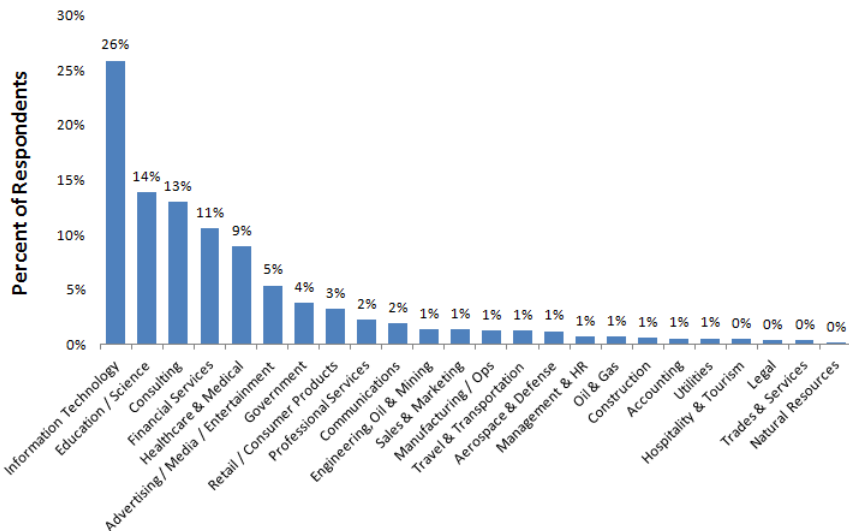
Might not be a real field, but has REAL JOBS.

**"So even if data science isn't a 'real field', it has REAL JOBS."**  
(R. Schutt, C. O'Neil, "Doing Data Science")



Might not be a real field, but has REAL JOBS.

## Data Scientists Work in Many Industries



# Big Three: Data Science, Big Data, Data Analytics.



# Big Three: What your **salary** will look like?

## #5. What your salary will look like



Data Science  
Professional

**\$123,000**



Big Data  
Professional

**\$88,000**




Data Analytics  
Professional

**\$61,000**

# Big Data Specialist, Data Engineer.

**Big Data Specialist**, else called **(Big) Data Engineer**:





**Languages**

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

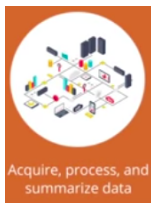
**Skills & Talents**

- ✓ Database systems (SQL & NO SQL based)
- ✓ Data modeling & ETL tools
- ✓ Data APIs
- ✓ Data warehousing solutions

	Data Engineer
Programming	●
Data Visualization & Communication	●
Statistics	●
Data Preparation	●
Machine Learning	●
Software & Databases	●
Calculus & Linear Algebra	●

# Data Analyst.

## Data Analyst:



	Data Analyst
Programming	
Data Visualization & Communication	
Statistics	
Data Preparation	
Machine Learning	
Software & Databases	
Calculus & Linear Algebra	

# Data Scientist: Skills/responsibilities.

**Data Scientist** (on **TOP** of **Data Analyst** skills & duties):

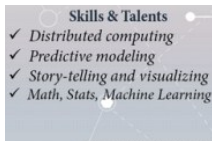


The Ten Most Common  
Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	72%
R	64%
SQL	51%
Hadoop	39%
Java	33%
SAS	30%
Spark	27%
Matlab	20%
Hive	17%
Tableau	14%

Source: Glassdoor Economic Research

glassdoor



	Data Scientist
Programming	●
Data Visualization & Communication	●
Statistics	●
Data Preparation	●
Machine Learning	●
Software & Databases	●
Calculus & Linear Algebra	●

# Final Thoughts.

- ① **Data Science** can be used **everywhere**.
- ② Not a specialization, but a **state of mind**.
- ③ **Ask questions**, use **data**, do a **project**.



# Final Thoughts.

- ① **Data Science** can be used **everywhere**.
- ② Not a specialization, but a **state of mind**.
- ③ **Ask questions**, use **data**, do a **project**.

If you do a project:

- ① **Mess up**. Mess up **A LOT**. **Learn** from it.
- ② **Own up**.
- ③ **Seek the truth** (I mean, **Google** it).