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

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

“Big” - when size becomes a **challenge**.

have to learn a new host of tools.
Web Search: How do search engines rank pages?

Online Recommendations: How do Netflix, Amazon, Ebay recommend items and movies that users might like?
My experience: "Data Journalism".

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".

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")

<table>
<thead>
<tr>
<th>Value Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Points</td>
</tr>
<tr>
<td>Penalties</td>
</tr>
<tr>
<td>1stDowns</td>
</tr>
<tr>
<td>Turnovers</td>
</tr>
<tr>
<td>Yards</td>
</tr>
<tr>
<td>TDs</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Value Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass Y/A</td>
</tr>
<tr>
<td>Pass TD/A</td>
</tr>
<tr>
<td>INT/A</td>
</tr>
<tr>
<td>Rush Y/A</td>
</tr>
<tr>
<td>Rush TD/A</td>
</tr>
</tbody>
</table>
```

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

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

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

Classical Averages

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

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).

<table>
<thead>
<tr>
<th>Team</th>
<th>Conference</th>
<th>Totals, Offense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oklahoma</td>
<td>Big 12</td>
<td>587.50</td>
</tr>
<tr>
<td>Oklahoma State</td>
<td>Big 12</td>
<td>578.92</td>
</tr>
<tr>
<td>Memphis</td>
<td>American</td>
<td>571.80</td>
</tr>
<tr>
<td>Louisville</td>
<td>Atlantic Coastal</td>
<td>550.00</td>
</tr>
<tr>
<td>Central Florida</td>
<td>American</td>
<td>539.90</td>
</tr>
<tr>
<td>South Florida</td>
<td>American</td>
<td>525.60</td>
</tr>
<tr>
<td>Ohio State</td>
<td>Big Ten</td>
<td>523.62</td>
</tr>
<tr>
<td>Arkansas State</td>
<td>Sun Belt</td>
<td>498.90</td>
</tr>
<tr>
<td>Toledo</td>
<td>Mid-American</td>
<td>497.00</td>
</tr>
<tr>
<td>Southern Methodist</td>
<td>American</td>
<td>496.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Team</th>
<th>Conference</th>
<th>Totals, Offense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oklahoma</td>
<td>Big 12</td>
<td>613.17</td>
</tr>
<tr>
<td>Oklahoma State</td>
<td>Big 12</td>
<td>594.13</td>
</tr>
<tr>
<td>Louisville</td>
<td>Atlantic Coastal</td>
<td>586.95</td>
</tr>
<tr>
<td>Ohio State</td>
<td>Big Ten</td>
<td>561.18</td>
</tr>
<tr>
<td>West Virginia</td>
<td>Big 12</td>
<td>508.73</td>
</tr>
<tr>
<td>Memphis</td>
<td>American</td>
<td>507.61</td>
</tr>
<tr>
<td>Central Florida</td>
<td>American</td>
<td>503.10</td>
</tr>
<tr>
<td>Missouri</td>
<td>Southeastern</td>
<td>501.24</td>
</tr>
<tr>
<td>Notre Dame</td>
<td>Independent</td>
<td>497.84</td>
</tr>
<tr>
<td>Syracuse</td>
<td>Atlantic Coastal</td>
<td>497.33</td>
</tr>
</tbody>
</table>
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?  ⇒  #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?  \[ \Rightarrow \]  #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? \(\Rightarrow\) #13 with 105 yards allowed per game.

Why so? My COUGAR BIAS you’d say?

Not really, see rushing offense ranks for their opponents:

<table>
<thead>
<tr>
<th>Name</th>
<th>Rush_Yards_Rank</th>
<th>Rush_Attempts_Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Tulsa</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>South Florida</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Tulane</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Navy</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>(Out of .. teams)</td>
<td>130</td>
<td>130</td>
</tr>
</tbody>
</table>
What is Data Science? Venn Diagram.

- Computer Science
- Maths & Statistics
- Business / Domain Expertise

Intersections:
- Machine Learning
- Data Science
- Data Analysis

Traditional Software
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”)

Andrey Skripnikov (University of Houston) All Things ”Data”: Data Science, Data Analysis
June 28, 2018 9 / 16
Might not be a real field, but has REAL JOBS.

Data Scientists Work in Many Industries

- Information Technology: 26%
- Consulting: 14%
- Financial Services: 13%
- Healthcare & Medical: 11%
- Retail / Consumer Products: 9%
- Government: 5%
- Professional Services: 4%
- Communications: 3%
- Engineering: 2%
- Oil & Mining: 2%
- Manufacturing / Ops: 1%
- Aerospace & Defense: 1%
- Management & HR: 1%
- Oil & Gas: 0%
- Construction: 0%
- Accounting: 0%
- Utilities: 0%
- Hospitality & Tourism: 0%
- Trades & Services: 0%
- Natural Resources: 0%
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, also called (Big) Data Engineer:

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

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 Analyst:

- Acquire, process, and summarize data
- Design and create data reports using various reporting tools
- Excel
- SQL
- Business intelligence
- Big data tools (Hive, Pig)
- Knowledge of mathematical statistics
- 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**

<table>
<thead>
<tr>
<th>Skill</th>
<th>Percentage of Job Listings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python</td>
<td>72%</td>
</tr>
<tr>
<td>R</td>
<td>64%</td>
</tr>
<tr>
<td>SQL</td>
<td>51%</td>
</tr>
<tr>
<td>Hadoop</td>
<td>39%</td>
</tr>
<tr>
<td>Java</td>
<td>33%</td>
</tr>
<tr>
<td>SAS</td>
<td>30%</td>
</tr>
<tr>
<td>Spark</td>
<td>27%</td>
</tr>
<tr>
<td>Matlab</td>
<td>20%</td>
</tr>
<tr>
<td>Hive</td>
<td>17%</td>
</tr>
<tr>
<td>Tableau</td>
<td>14%</td>
</tr>
</tbody>
</table>

---

**Skills & Talents**
- Distributed computing
- Predictive modeling
- Story-telling and visualizing
- Math, Stats, Machine Learning

---

**Data Scientist**

- Programming
- Data Visualization & Communication
- Statistics
- Data Preparation
- Machine Learning
- Software & Databases
- Calculus & Linear Algebra
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).