Introduction to Data Science/ Data Mining for Business Analytics

BRIAN D'ALESSANDRO

VP – DATA SCIENCE, DSTILLERY

ADJUNCT PROFESSOR, NYU

FALL 2014

WHAT IS DATA SCIENCE?

TOO SEXY FOR THIS COURSE?

THE MAGAZINE

October 2012

Harvard Business Review

ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here to register for FREE access »**

Data Scientist: The Sexiest Job of the 21st Century

"Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions.

They find the story buried in the data and communicate it. And they don't just deliver reports:

They get at the questions at the heart of problems and devise creative approaches to them."

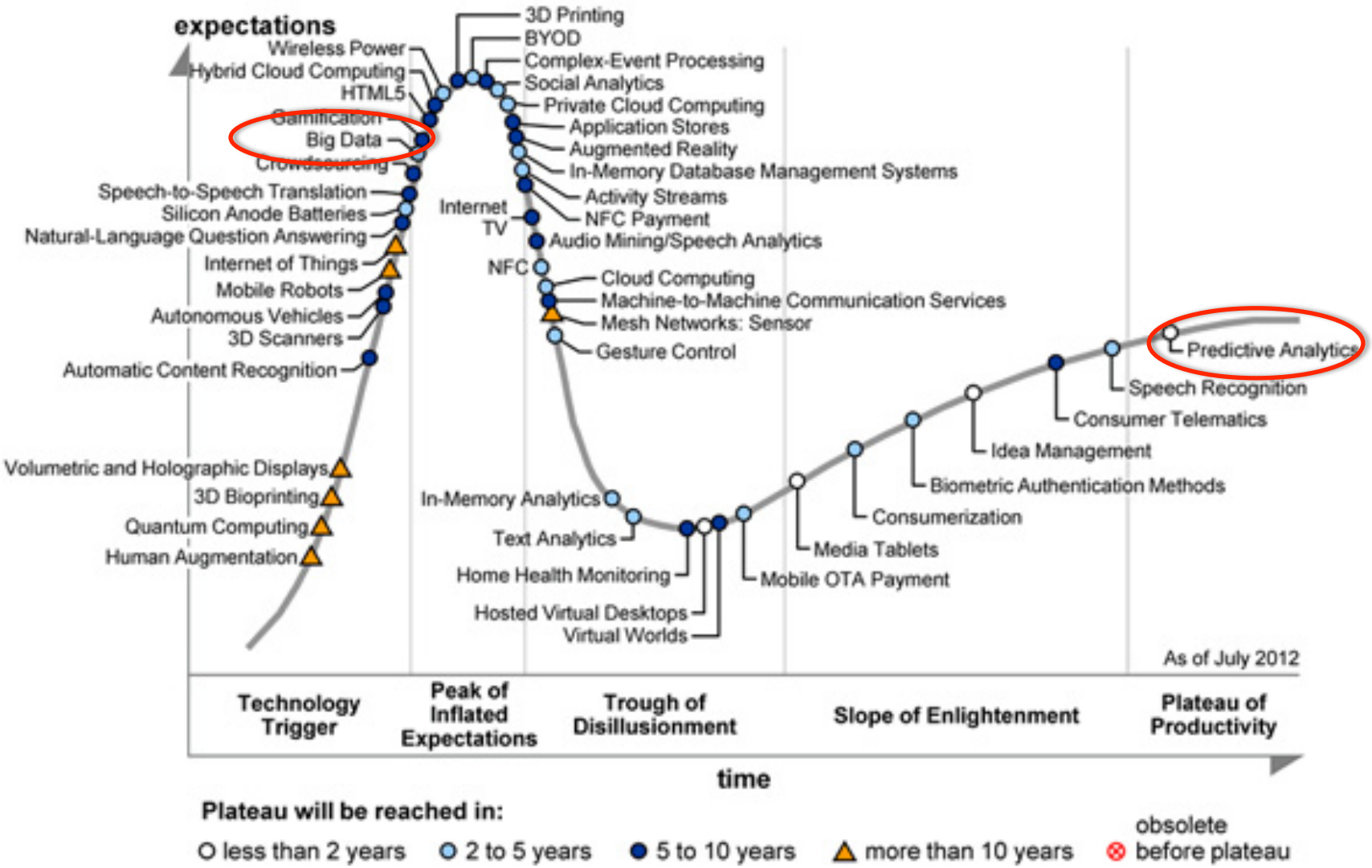
http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

WHY IS IT SO SEXY?

WHO'S BUYING IT?

WHAT VALUE IS IT CREATING?

HYPE OR NOT?



“DATA SCIENCE” IS NEW.

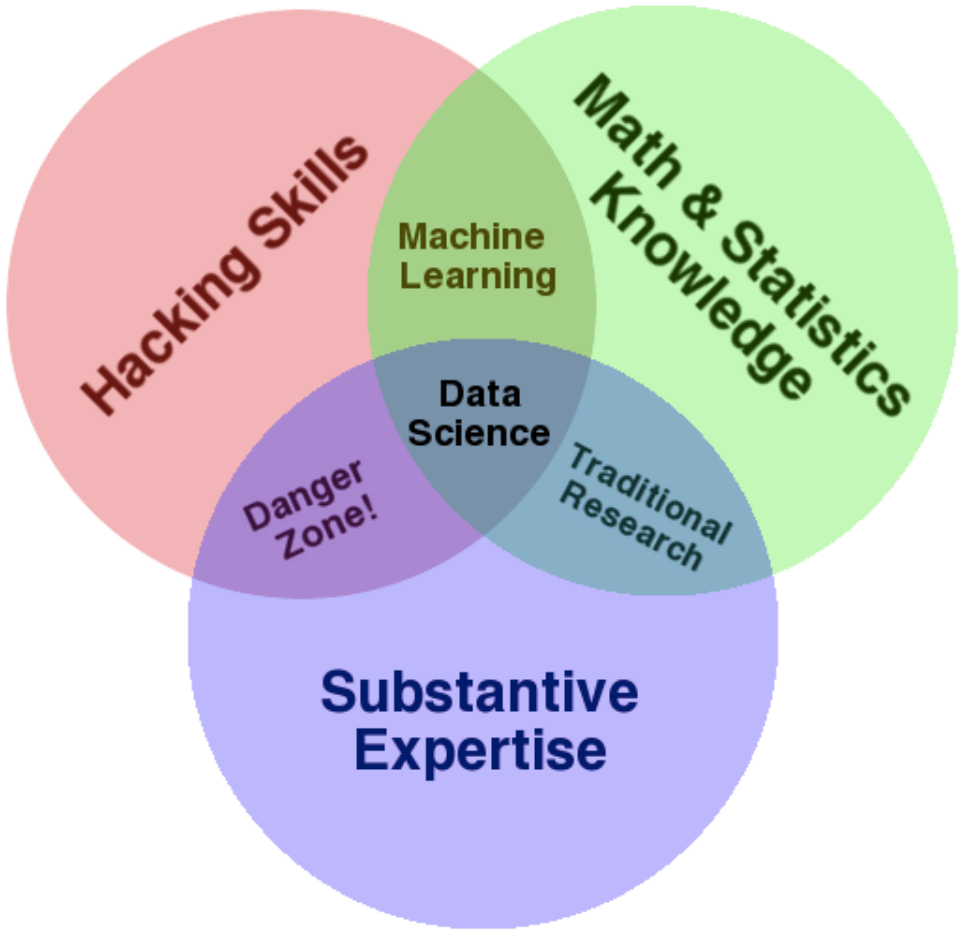
DATA SCIENCE ISN’T.

LETS START TO DEFINE THINGS

What skills do we expect in our data scientists?

LETS START TO DEFINE THINGS

What skills do we expect in our data scientists?



Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

SERIOUSLY, KEEP OUT.

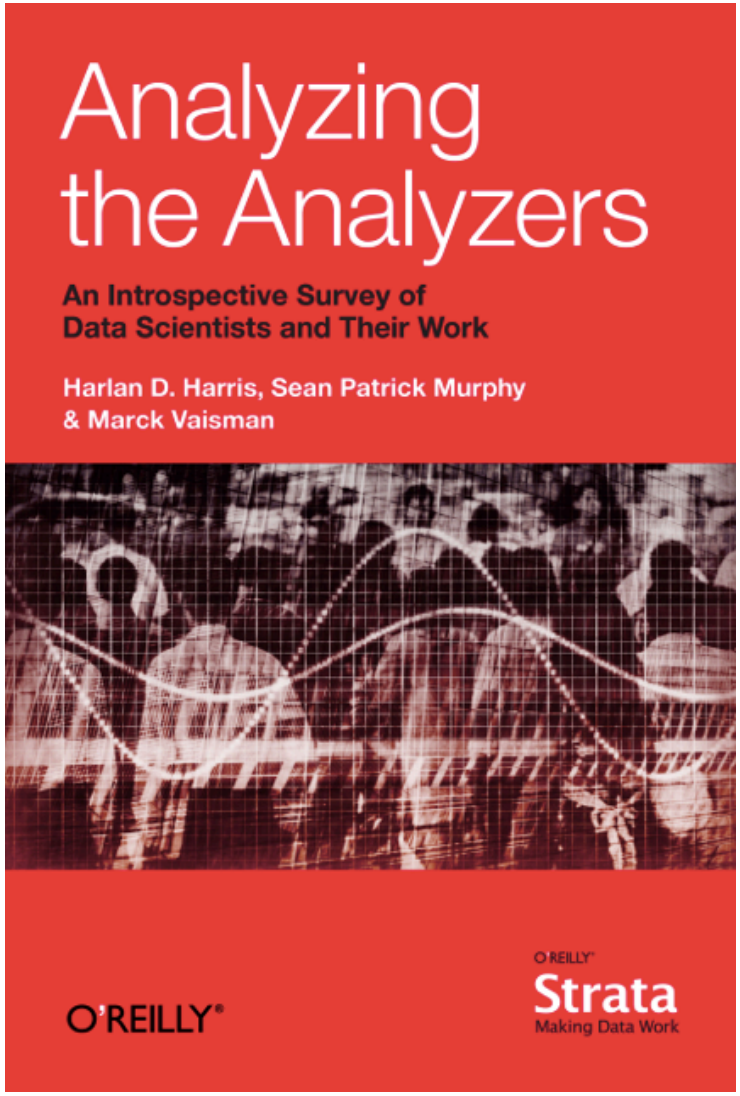
=

Knowing how to build a model, but not knowing what a model really is or how to properly evaluate it.

TOWARDS A DEFINITION

There is no 'one-size-fits-all' type of data scientist.

Luckily, people are using data science to define data science.



RANGE OF DS SKILLS

They're all very similar, but some categorization still helps.

Business	ML / Big Data	Math / OR	Programming	Statistics
Product Developement	Unstructured Data	Optimization	Systems Administration	Visualization
Business	Structured Data	Math	Back End Programming	Temporal Statistics
	Machine Learning	Graphical Models	Front End Programming	Surveys and Marketing
	Big and Distributed Data	Bayesian / Monte Carlo Statistics		Spatial Statistics
		Algorithms		Science
		Simulation		Data Manipulation
				Classical Statistics

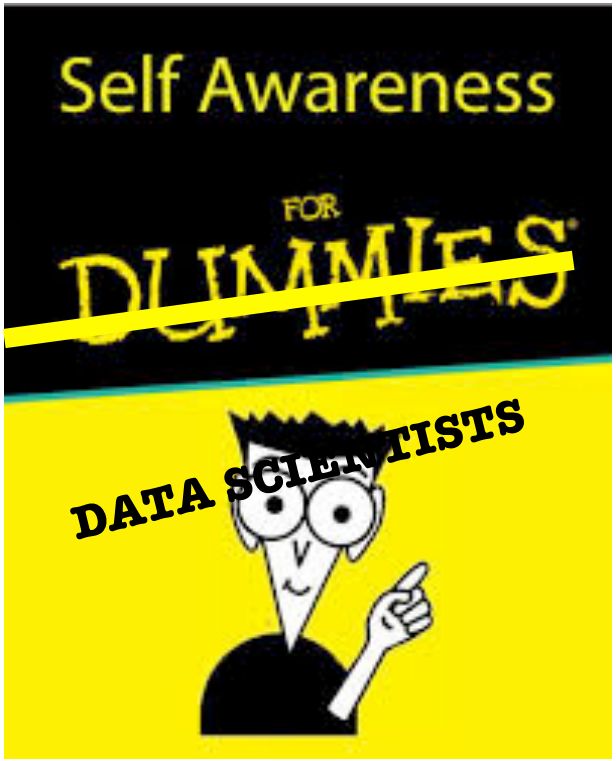
DATA ROLES

In Analyzing the Analyzers, the authors identified 4 types of “data scientists.”

Data Developer	Developer	Engineer	
Data Researcher	Researcher	Scientist	Statistician
Data Creative	Jack of All Trades	Artist	Hacker
Data Businessperson	Leader	Businessperson	Entrepeneur

IT MATTERS

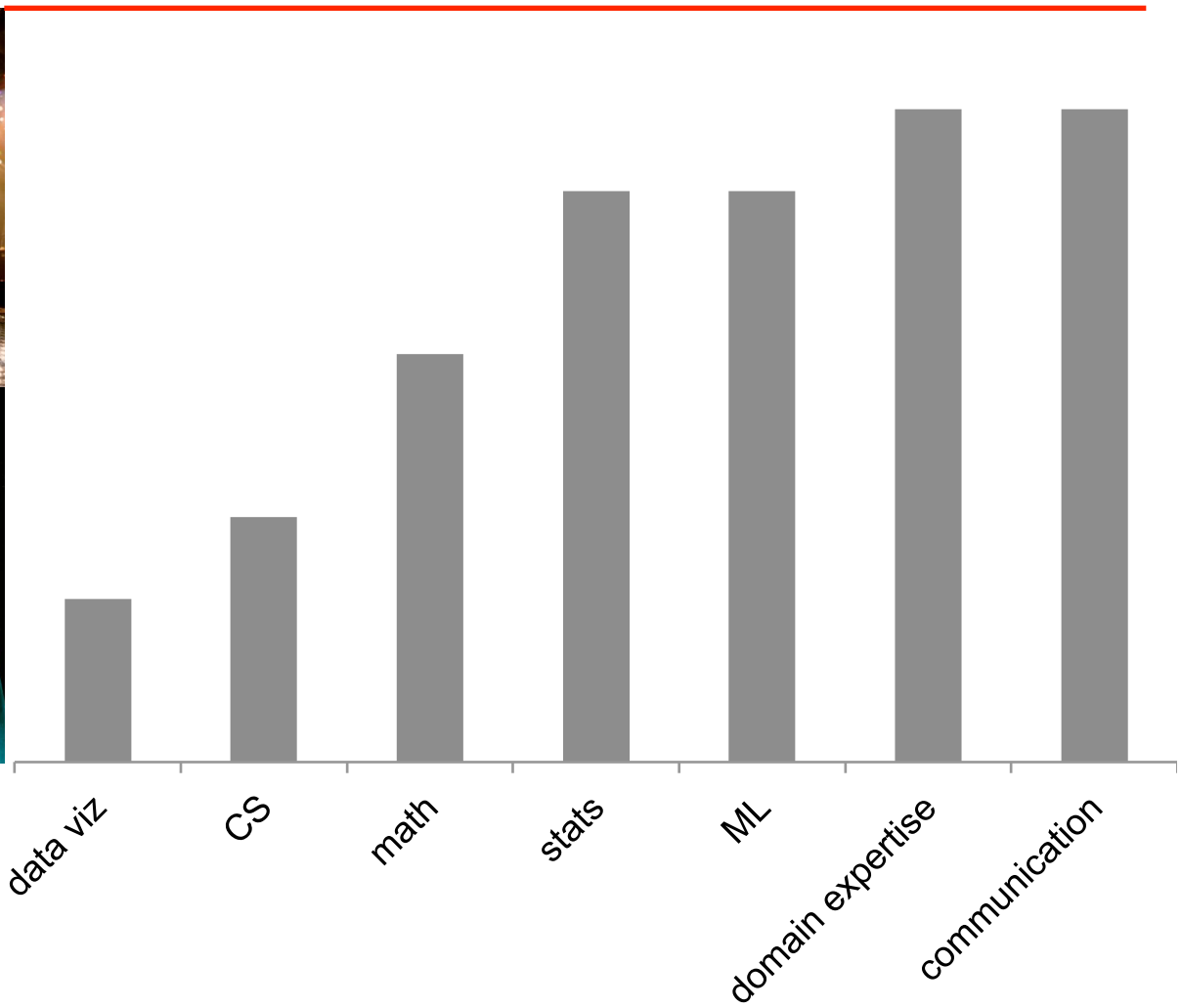
You don't have to fit into one bucket, but you should know where you are…



- Personal skills development
- Choosing the right job (your future boss might not know what a data scientist is, or should be)

DATA SCIENCE PROFILE

What I think I am…



WHY SCIENCE?

We defined 4 data roles, but what is the “science” of data science?

<u>The scientific method:</u> evaluating the merit of a hypothesis with rigorous empirical testing.

I.e.,

Given raw data, constraints and a problem statement, you have an infinite set of models to choose from, with which you will use to maximize performance on some evaluation metric, that you will have to specify.

Every design choice you make can be formulated as a hypothesis, upon which you will use rigorous testing and experimentation to either validate or refute.

BUT ITS STILL AN ART

Outside of modeling competitions, seldom is a well-posed problem and clean dataset presented to you.

Putting the art into your practice means…

- Translating problems into the language of data science
- Formulating reasonable hypotheses
- Developing an intuition for good vs. bad data, good vs. bad models.
- Abstracting problems to identify similarities
- Managing the DS process from end to end

REMINDER

With this course we want to emphasize the *soft* skills of data science

Art => Abstract and intuitive thinking

Science => process

We'll cover necessary DS tools, but with the goal of applying them towards analytic problem solving.

A CASE STUDY

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

A CASE STUDY

Why would they want to predict what is going to happen?

What kind of things might they want to predict?

What data do they have to make predictions?