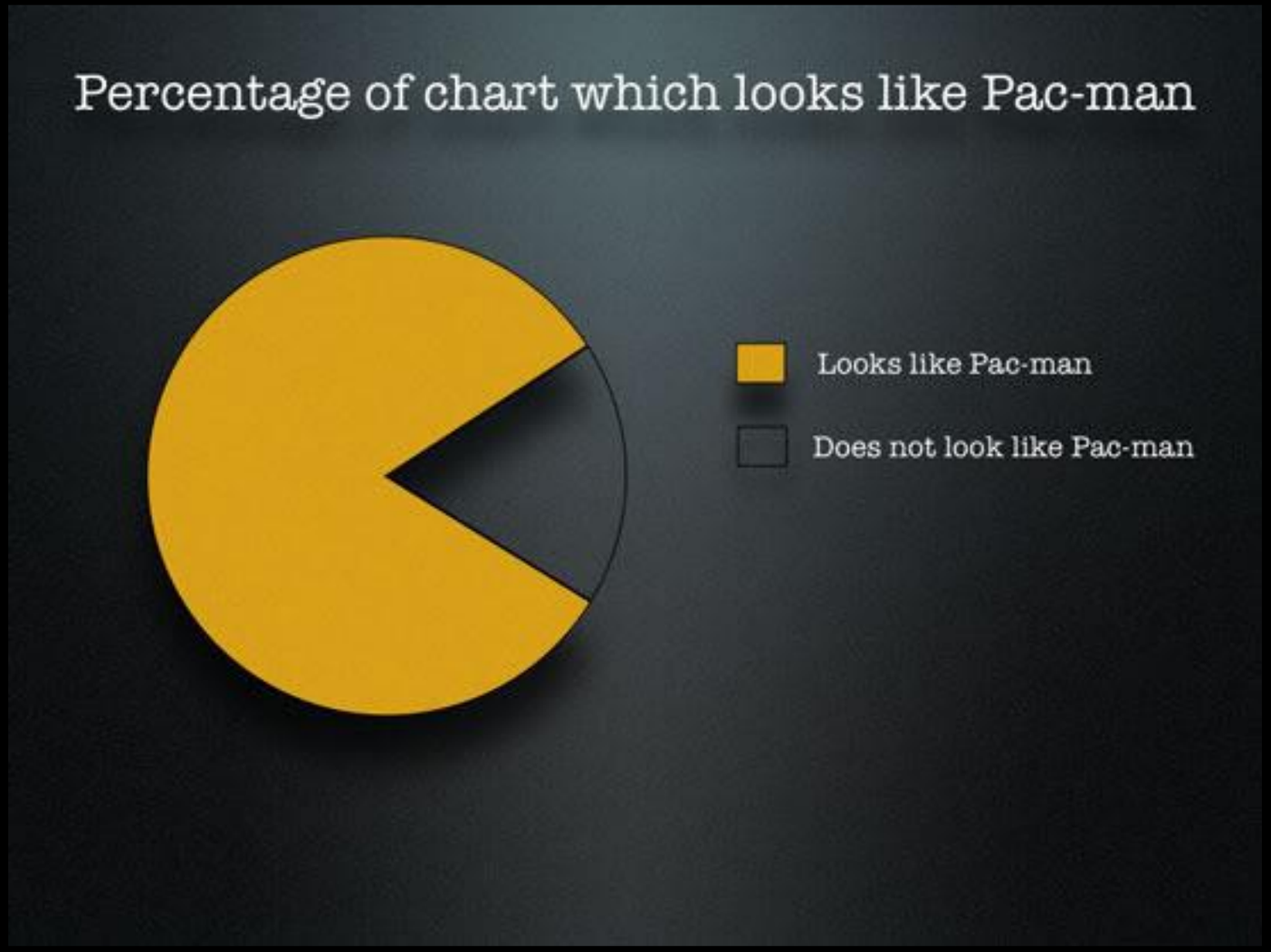
DATA SCIENCE



GENERAL ASSEMBLY

WHAT IS DATA SCIENCE?

Jim Byers, Technical Program Manager DS/BI

HELLO!

JIM BYERS

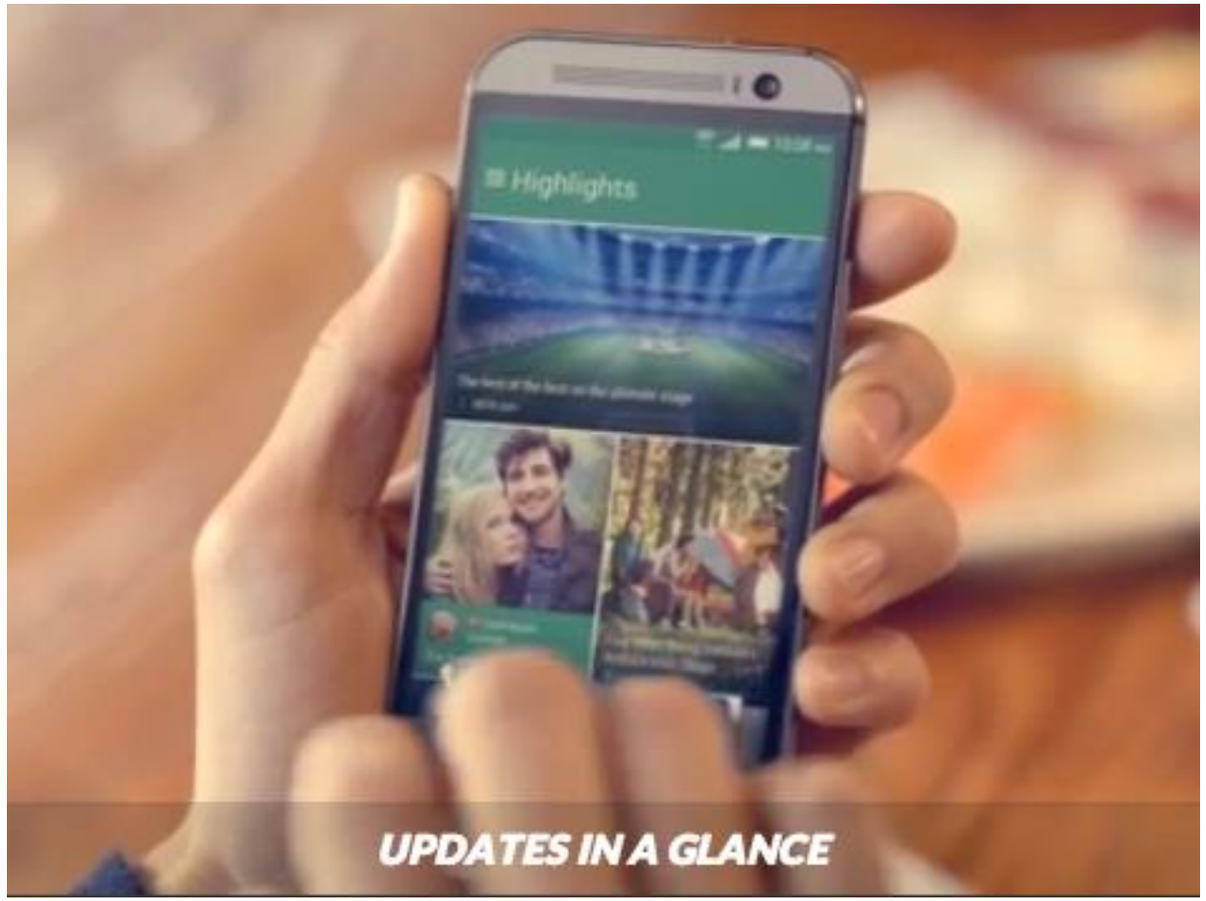
- Data Science / Analytics Technical Program Manager
  - Disney/ABC
- Loves pulling information and insights from data
- Career path: engineering, sales, product management, business development, program management, **data analysis, data science**

BY THE END OF THIS YOU WILL BE ABLE TO:

- Communicate what data science is and why we care
- Describe three problems that we can solve with data science
- List at least 3 skill areas a good data scientist needs to have
- Be able to approach problems with a data science workflow

WHAT IS DATA SCIENCE AND WHY DO WE CARE?

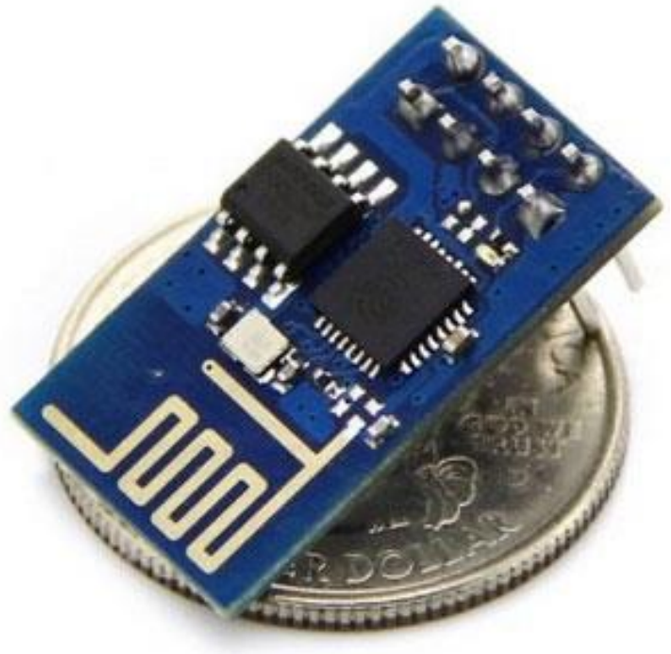
Data, data and more data sources every day



- **Data collection to the web with a $3 wifi & processor board**

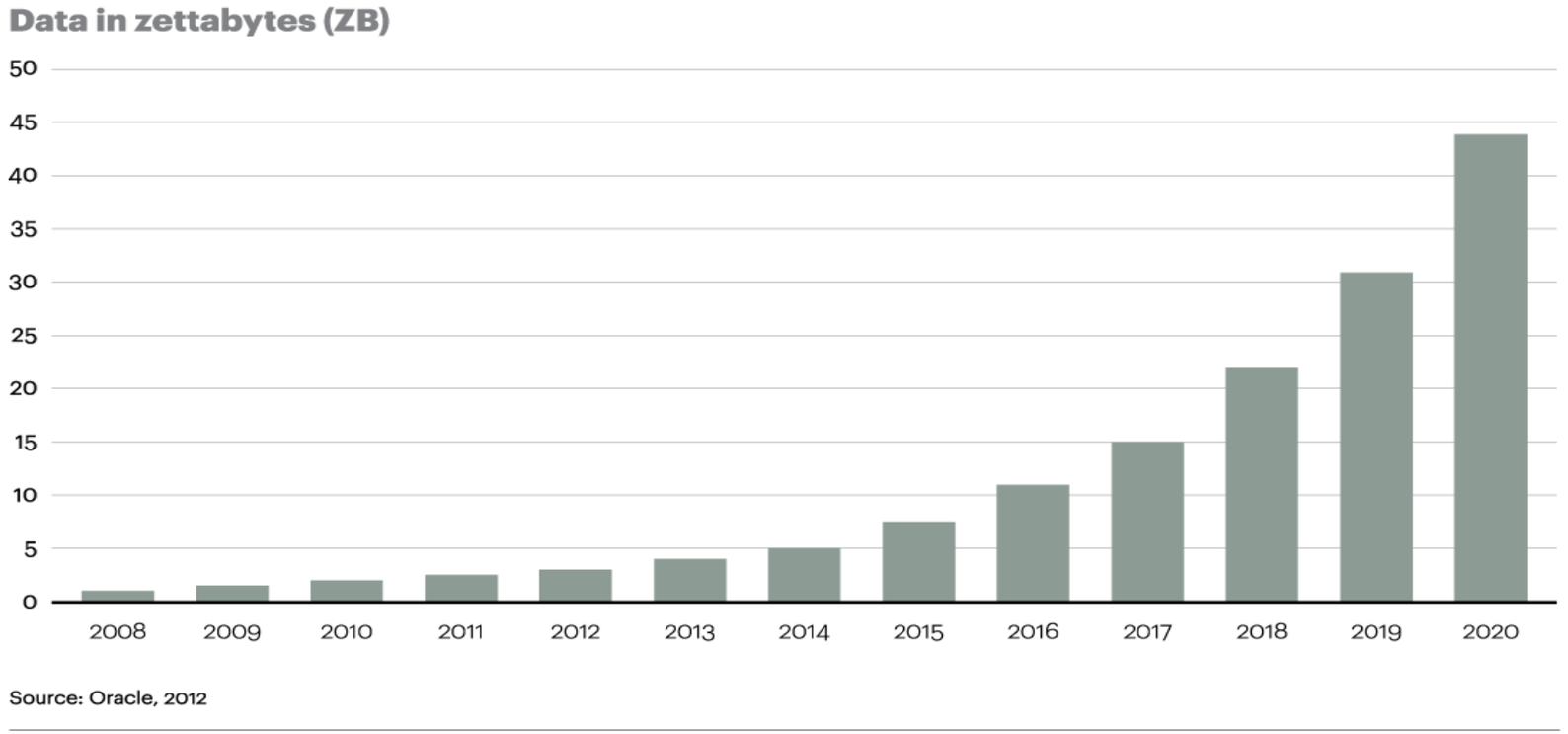
- **Jim's temperature and humidity data logged to the web**

https://public.tableau.com/profile/publish/HomeOffice-TemperatureandHumidity/Jimshomeoffice-TemperatureandHumidity



THE DATA EXPLOSION – GIGANTIC AMOUNTS OF DATA

- Gigantic amounts of data are being produced, doubling every two years
- 2014 will produce > 5,000,000,000,000,000,000,000 bytes (5 Billion Terabytes) of data



- But what we want is information from the data
- Less than 1% of the data is analyzed

WHAT THE HECK IS DATA SCIENCE ANYWAY?

- Data science attempts to understand the world through empirical, testable, and incorrect models
- Data science is the intersection of computer science, statistics, applied math, and machine learning

WHY DO WE CARE ABOUT DATA SCIENCE?

- Because it works!
- The world is full of situations where we want to:
  - Forecast likely outcomes given that we know now
  - Segment things into groups
  - Recommend something based on the likelihood it will be viewed or clicked on
- Data science is ***uniquely*** able to solve these problems

EXAMPLES OF DATA SCIENCE IN ACTION

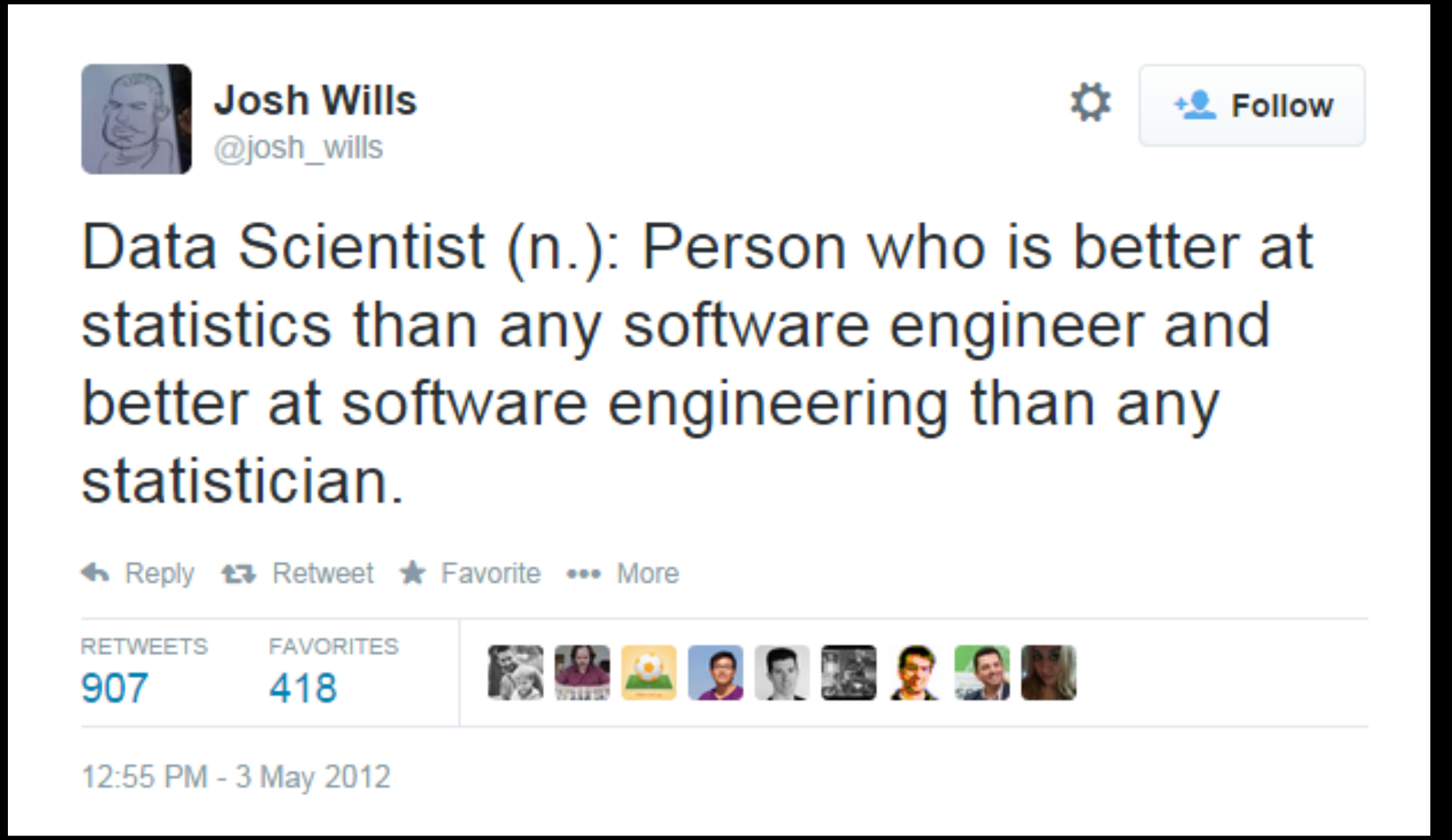
- Facebook facial recognition in photos
- Netflix/Amazon/Spotify recommendations
- Siri/Echo/Cortana voice recognition assistants
- Building art with Neural Networks - https://github.com/jcjohnson/neural-style
- Faceswap - https://www.youtube.com/watch?v=UngUWA43q5o
- Stock Market - https://www.quantopian.com/ - building crowd source hedge funds
- Helping people
  - https://www.drivendata.org/ - determine who is a good bet to give money to for a micro loan
  - http://www.datakind.org/projects
- Help find missing children - http://www.datakind.org/projects/finding-30000-missing-children
- Find correlations from sickness, grades, and attendance and try to find ways to improve them
  - http://coolculture.org/webfm_send/62
  - https://www.theengineroom.org/datakind-tests-project-accelerator/ (Kevin's previous company)
- Additional examples
- https://www.kaggle.com/wiki/DataScienceUseCases

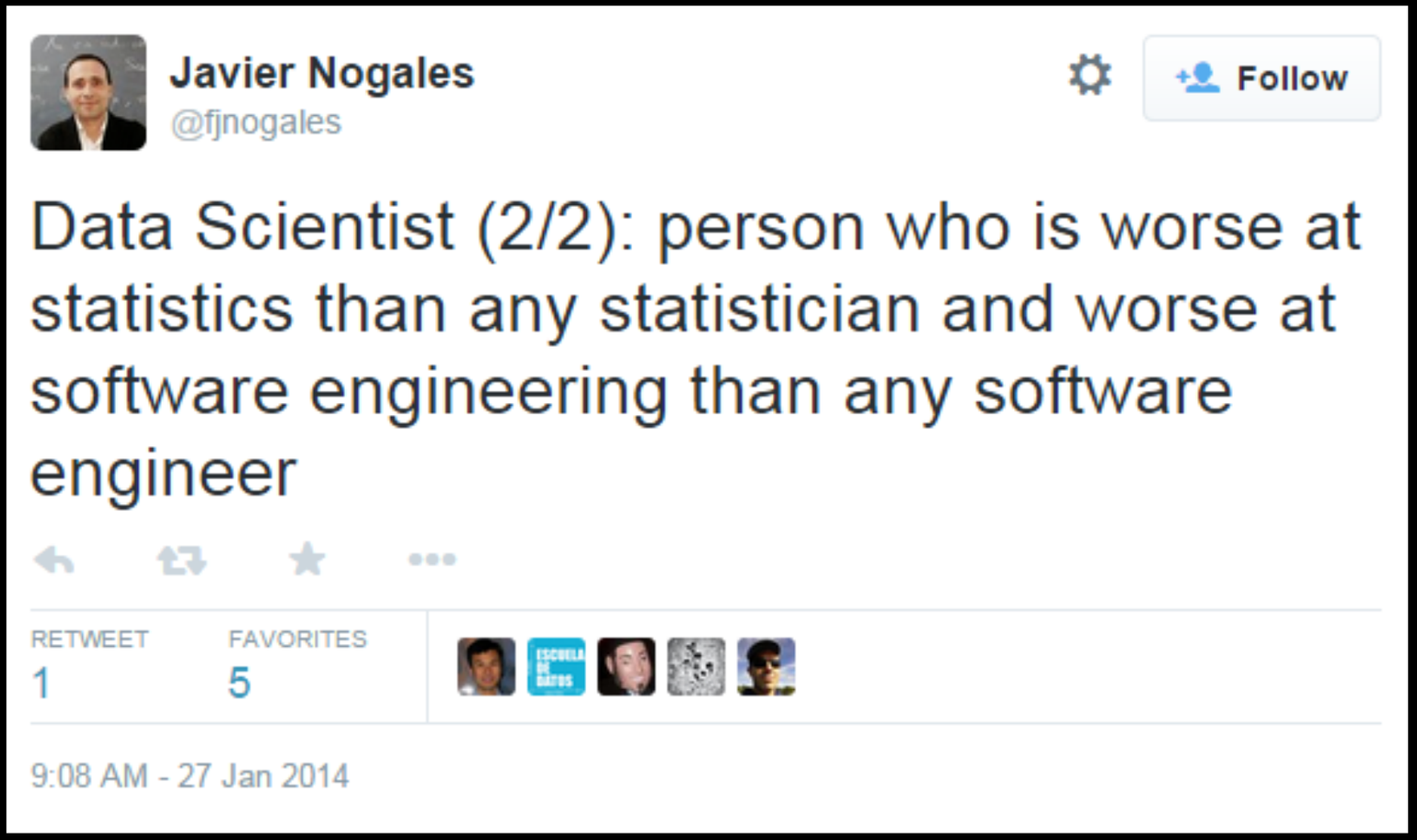
SOME OF THE TECHNIQUES APPLIED IN DATA SCIENCE

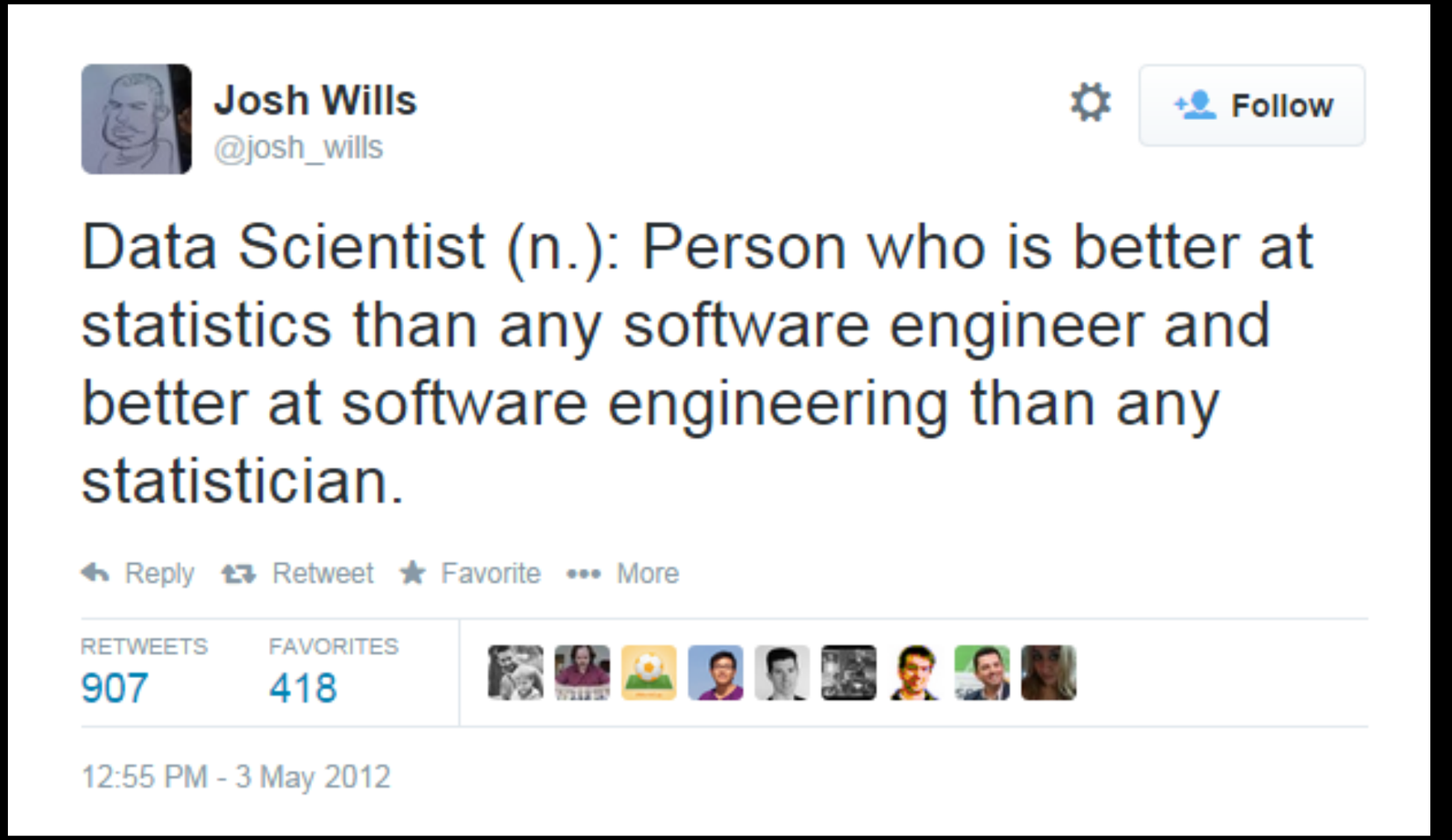
Forecast and prediction from *numeric* values	”What are our sales going to be next year given the trend in the sales of our product lines?”	Regression
Segmentation and cluster analysis	“What is a good grouping of our customers that I can use to think about how best to appeal to them?”	K-Means, DBSCAN
Spam filter	”Should this email message be classified as spam?”	Naïve Bayes
Matching web site users of similar interest	”What group is this new web page likely to appeal to”	Nearest neighbor, SVC, Ensemble Classifiers

SKILLS REQUIRED TO BE AN EFFECTIVE DATA SCIENTIST


Zvi
@nivertech
Follow
"Data Scientist" is a Data Analyst who lives in California.
Reply
Retweet
Favorite
More
RETWEETS
140
FAVORITES
40
9:55 PM - 14 Mar 2012


Josh Wills
@josh_wills
Follow
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
Reply
Retweet
Favorite
More
RETWEETS
907
FAVORITES
418
12:55 PM - 3 May 2012


Javier Nogales
@fjnogales
Follow
Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer
RETWEET
1
FAVORITES
5
9:08 AM - 27 Jan 2014


Josh Wills
@josh_wills
Follow
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
Reply
Retweet
Favorite
More
RETWEETS
907
FAVORITES
418
12:55 PM - 3 May 2012

WHAT IS A DATA SCIENTIST?

“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

WHAT IS A DATA SCIENTIST?

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data "in production"**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

HIERARCHY OF SKILLS

- **See Jim's hierarchy of analyst, Business Intelligence analyst and Data scientist skills**

DATA SCIENCE WORKFLOW

A DATA SCIENCE WORKFLOW MODEL

1. Identify the problem
2. Acquire the data
3. Understand data

   Fix, Parse, and analyze the data
4. Refine the data
5. Build and test model
6. Present the results, disseminate information

   - Share findings, visualizations and models

WHAT WE LEARNED

- What data science is and why we care
- The types of problems that can we solve with data science
- What makes a good data scientist: The skills required
- How to think about the data science workflow

QUESTIONS?

JIM'S EXAMPLE – SEGMENT USERS BY USE BEHAVIORS

- **Understanding the problem:** Worked with product manager on what question they were needing answered and what decisions they would make with answer
- **Obtain the data**: wrote SQL database query to get a random sample of users across 4 metrics of interest
- **Understand the data**: Used SPSS to do factor analysis and found that I could use just two metrics instead of four since some were highly correlated with others
- **Preparing the data** – Other than making sure I had no blank values, I used the data as is (turned out to be a mistake)
- **Analyze the data**
  - I wrote a program based on publically available libraries and some code examples on the web to implement a K-Means analysis on the data

JIM'S EXAMPLE – SEGMENT USERS BY USE BEHAVIORS

- **Analyze the data continued**
  - Analyzed the data and found that 4 clusters was the most useful
  - Plotted the results of the clustering and found that although clusters made sense there were some obvious outliers
  - Went back to "preparing the data" and removed the outliers
  - Re-Analysed the data to get updated clustering
- **Visualization:** Created scatter chart with color coding for each cluster and named the clusters so they would be easy to internalize and remember
- **Disseminate information:** Discussed the results with the product manager and wider product team and defined next steps (breakdown by country)

INTRO TO DATA SCIENCE

ADDITIONAL RESOURCES AND WHAT WE HAVE LEARNED TOGETHER

ADDITIONAL RESOURCES

- How to Lie With Statistics - Darrell Huff
- What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics - Andrew Vickers
- Teaching Statistics: A Bag of Tricks - Andrew Gelman and Deborah Nolan
- An introduction to Statistical Learning: with applications in R - James Gareth
- Python Machine Learning - Sebastian Raschka