INTRODUCTION TO DATA SCIENCE

INSTRUCTOR: KENISHA PRIESTER

WHAT IS DATA SCIENCE??

Harvard
Business
Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

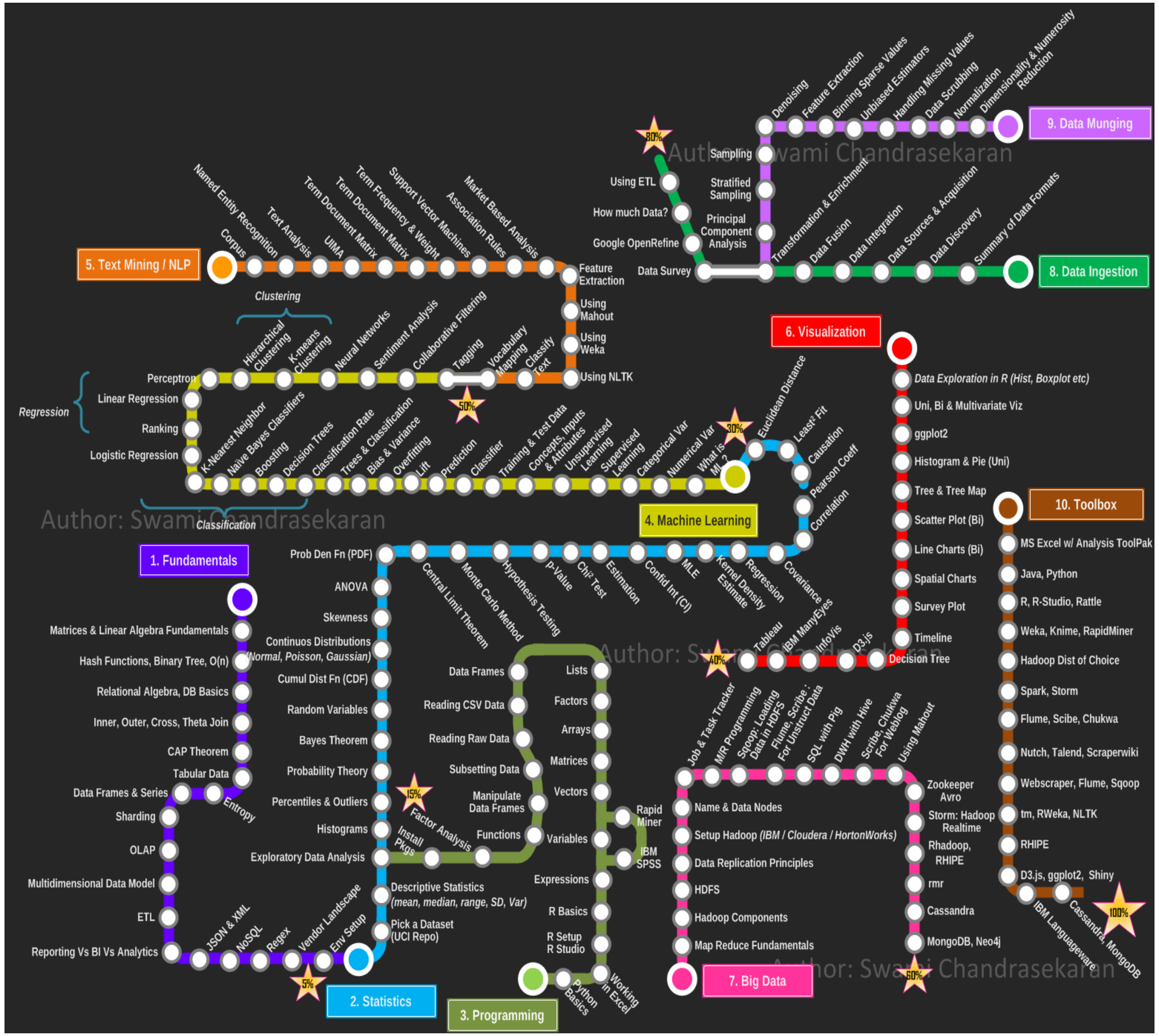
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

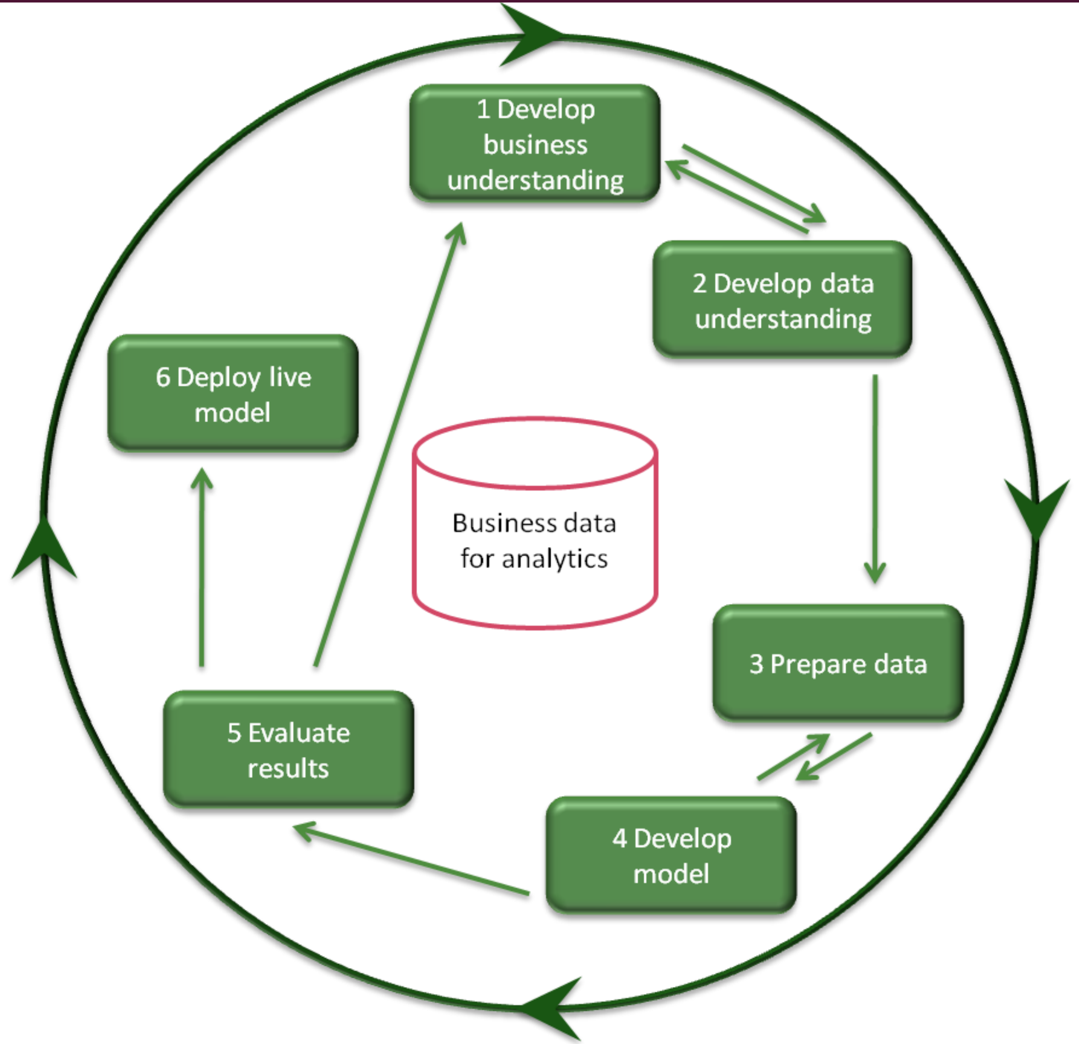
WHAT TO READ NEXT

Big Data: The Management Revolution



- Use data to generate insights for human and/or machine decisions
- Blend of statistics, math (linear algebra), computer science, business, and engineering
- 70-80% data preparation, 20-30% data modeling (the cool stuff)

DATA PROJECT LIFE CYCLE



1 Develop business understanding

2 Develop data understanding

3 Prepare data

4 Develop model

5 Evaluate results

6 Deploy live model

Business data for analytics

DATA SCIENCE IN ACTION

HOW DOES DATA SCIENCE FIT INTO OUR EVERYDAY LIVES?

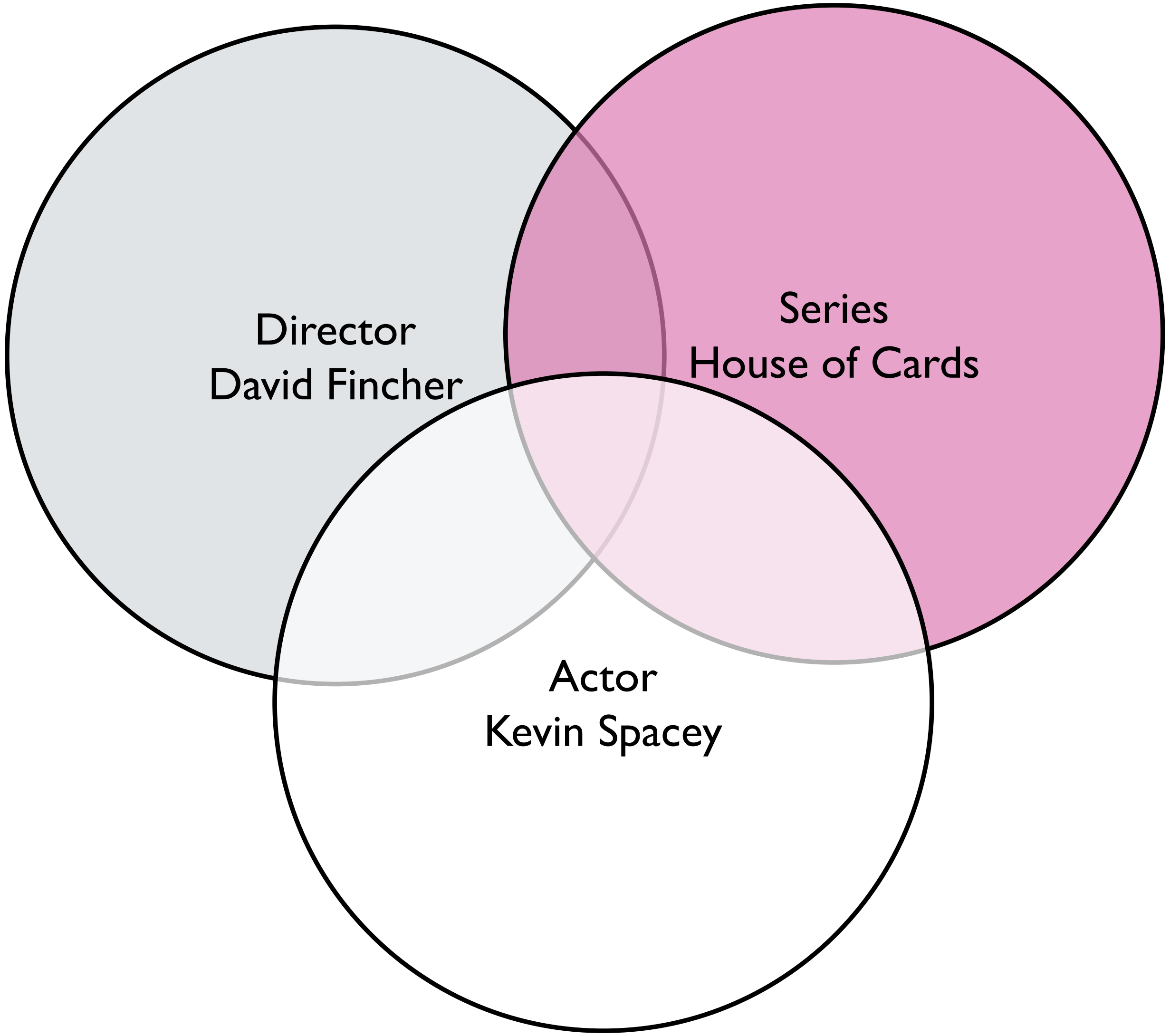
PANDORA®

- In 2000, Music Genome Project created
  - Map “DNA” of songs – 450 features
- Launched Pandora
  - Uses factors like time of day, location, and device used
  - Skip vs Thumbs down

NETFLIX

ATES ARMY

- Bought 2 seasons without watching
- Analyzed viewer data
  - House of Cards (UK) fans like movies with Kevin Spacey directed by David Fincher
  - Less likely to cancel Netflix subscription

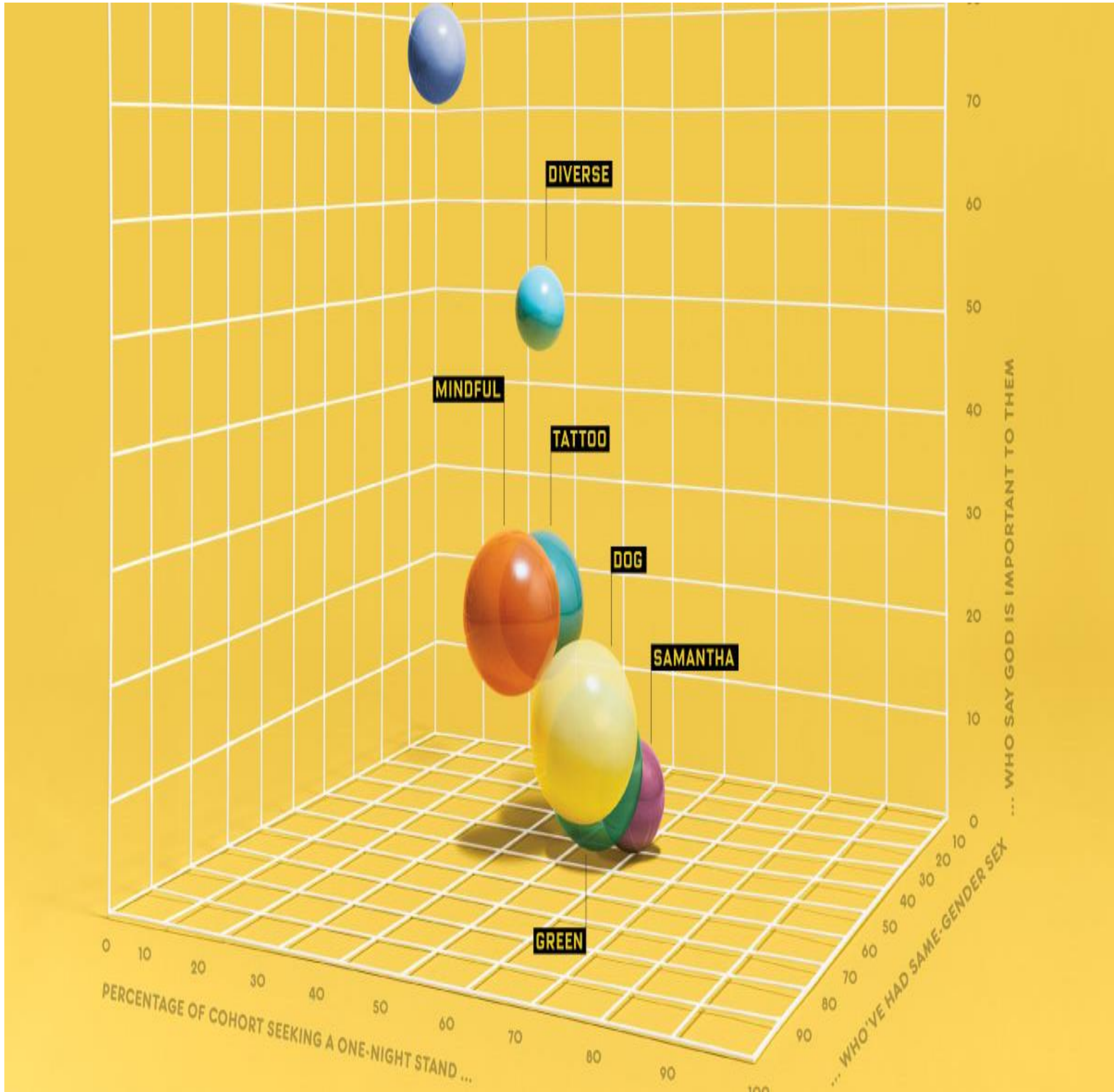


TARGET
6600
Entrance
Not an Entrance
Not an Entrance
Not an Entrance
Exit Only
Exit Only
Exit Only

- Guest ID number for each shopper with shopping habits tracked
- Historical data for baby registries
  - Unscented lotion in 2nd trimester
  - Supplements during first 20 weeks
  - Increase in unscented soap, cotton balls, hand sanitizer, and washcloths for upcoming due date

I Hacked OkCupid

- Breakup 9 months prior
- Only had 6 first dates via OkCupid
- OkCupid user typically only answers 350 profiling questions
  - Must answer the same question in order to be potentially matched
- Made 12 fake accounts
  - Answered questions randomly
  - 20,000 women formed into 7 distinct clusters
- Made 2 real profiles
  - 20 messages per day
  - Collected in-person date information
  - 88th date – found "The One"



COURSE OUTLINE

- Tools: Anaconda (Python & Jupyter Notebook), Tableau

- Explain what data "looks like" (exploratory data analysis)
- Changing data to be clean and consistent (data preparation)
- Transforming non-numerical data for computer to read (feature engineering)
- Feed data into models to understand trends and generate future trend output (predictive analytics/machine learning)
- Show final results using interactive charts and graphs (data visualization)

LET'S GET STARTED!

WELCOME TO THE WORLD OF DATA SCIENCE