GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1: WELCOME

Neo Ellison
Data Science Partner, Fix It With Code
Lead Data Scientist, Enertiv

LIES!!!

LIES!!

There are three kinds of lies:
lies, damned lies and statistics

– Benjamin Disraeli, Prime Minister of Great Britain
(1868, 1874-1880)

HAS THIS EVER HAPPENED TO YOU?

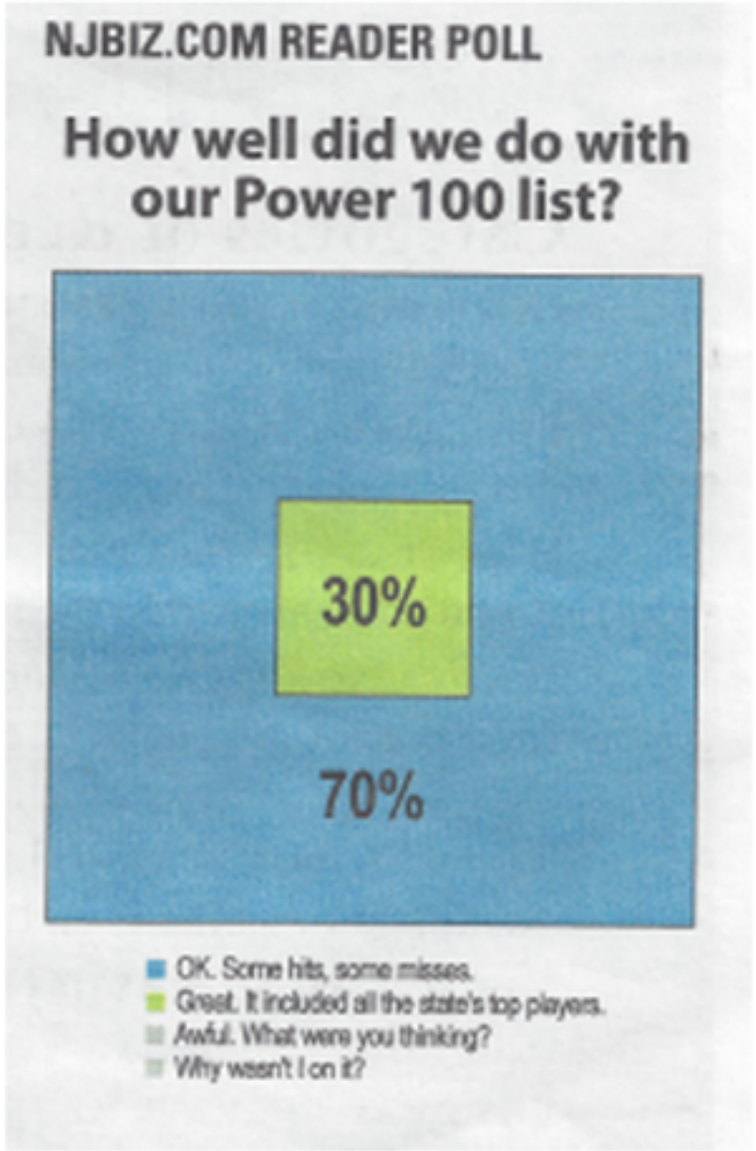
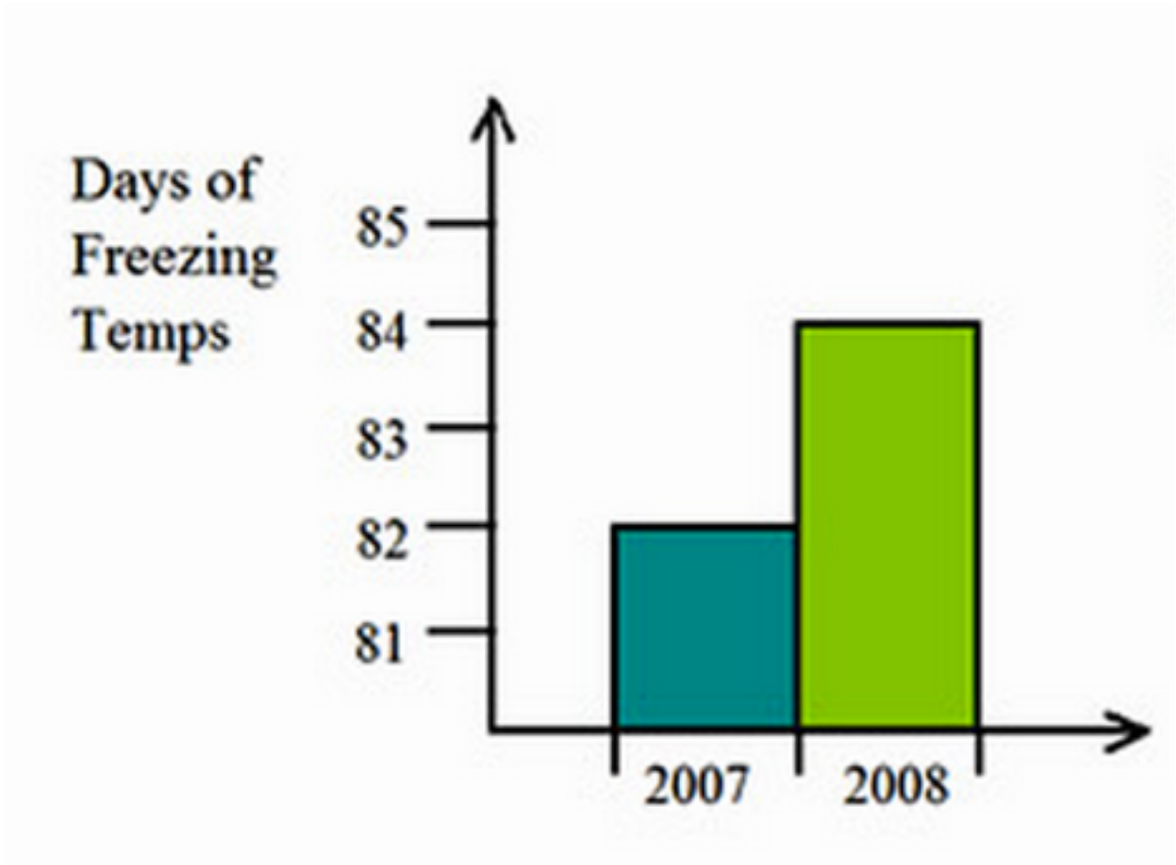
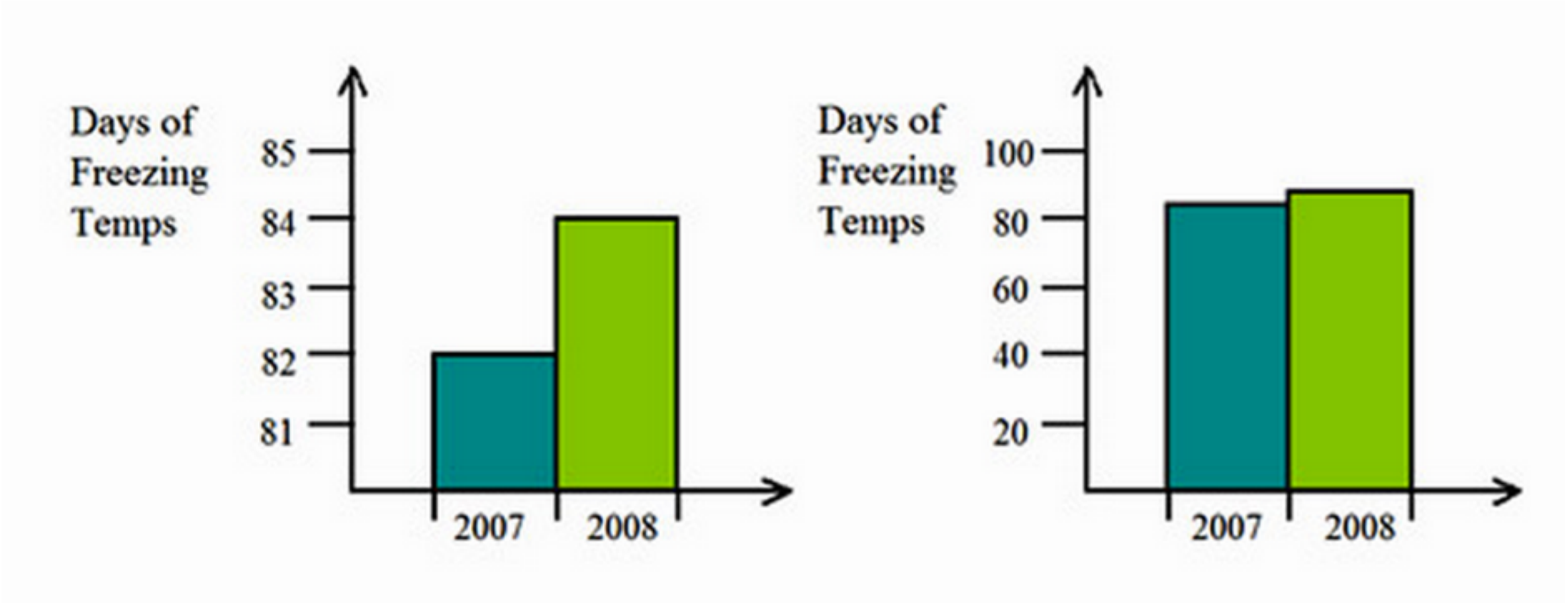


Figure 1. Source: NJBIZ February 4, 2013

HAS THIS EVER HAPPENED TO YOU?



HAS THIS EVER HAPPENED TO YOU?



AGENDA

I. Course Overview
II. What is Data Science
III. Lab
    - Setup Developer Environment
    - iPython

I. COURSE OVERVIEW

Course Overview

Contact Info

Office Hours

Neo Ellison neo@fixitwithcode.com by appointment

Katie Barnwell ktbarnwell@gmail.com 9:30-10:30 M/W

Class M/W 6:30-9:30 PM - 12/10 - 3/9

Location GA East (10 E. 21st St, 4th Floor)

Classroom 4C

COURSE OVERVIEW

Topics

Regression Models and Continuous Variables

Classification, Clustering, and Categorical Variables

Data Visualization, NLP, Bayesian Inference

Data Engineering

Assignments

Dataexplor Challenges

Term Project

COURSE OVERVIEW

Tools

Python Data Science Stack

Scikit-Learn (machine learning)

Numpy, Scipy (linear algebra, numerical computation)

Matplotlib (visualization)

Pandas (modeling, easy-to-use data structures)

iPython (interactive matlab-style interface)

INTRODUCTIONS

3 minutes:
Partner with someone next to you. Introduce yourselves. Then we will reconvene, and you will introduce your partner to the group.

Please share your partner's:

- name
- occupation
- experience with Python and scikit-learn
- what s/he is most excited about learning/doing
- what s/he is most apprehensive about learning/doing
- One Weird Trick to Help Me Remember Your Name

II. WHAT IS DATA SCIENCE

FUN FACT:

- Every Day We Create 2.5 Quintillion Bytes of Data

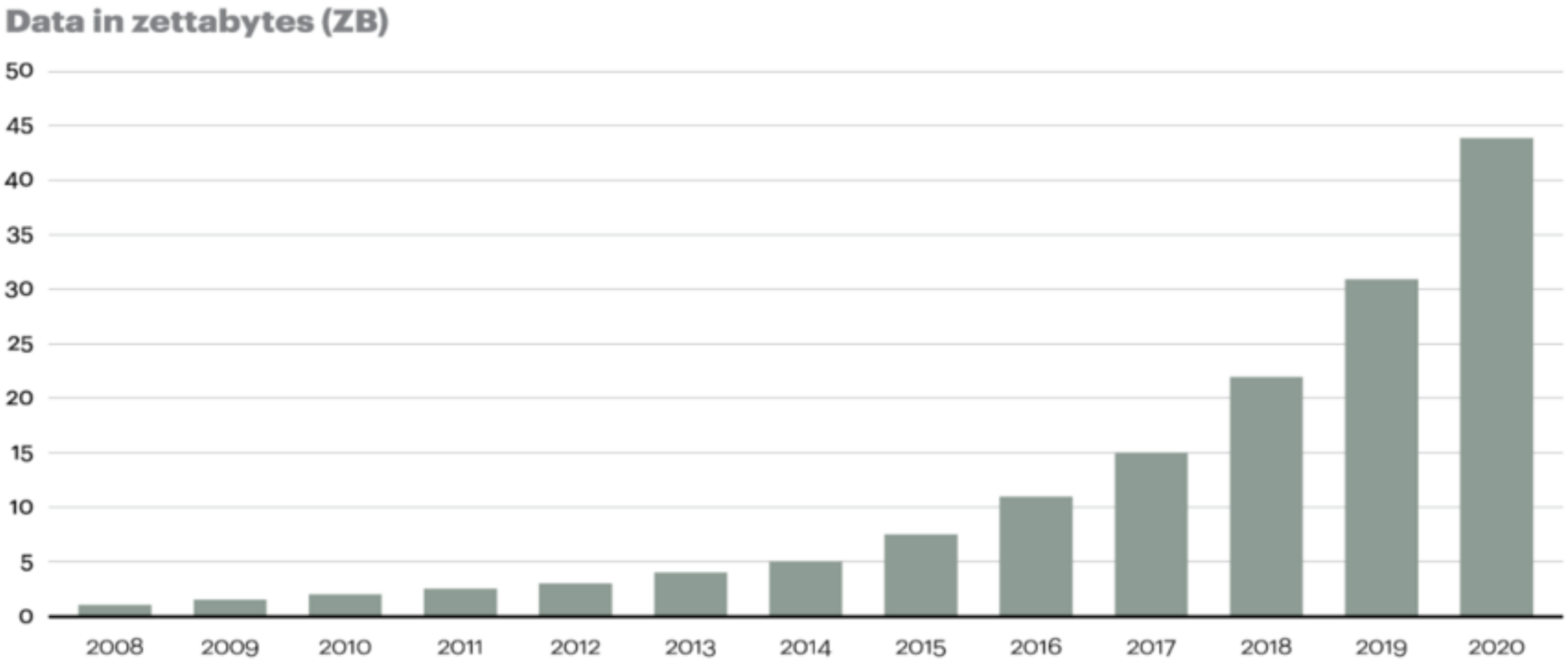
FUN FACT:

- Every Day We Create 2.5 Quintillion Bytes of Data
- 90% of current data was collected in the past two years

FUN FACT:

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



Source: Oracle, 2012

What is Data Science

- A set of tools and techniques used to extract useful information from data.

What is Data Science

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject

What is Data Science

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-oriented subject
- The application of the *Scientific Method* to solving business problems

Who Uses Data Science?

What is a Data Scientist?

- “a data analyst who lives in California”

What is a Data Scientist?

- “a data analyst who lives in California”
- “a business analyst who lives in New York”

What is a Data Scientist?

- “a data analyst who lives in California”
- “a business analyst who lives in New York”
- “a statistician who lives in San Francisco”

What is a Data Scientist?

Michael E. Driscoll
@medriscoll

Follow

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

Reply Retweet Favorite Pocket More

RETWEETS	FAVORITES
34	27

4:57 AM - 18 Jul 2012

What makes a good Data Scientist?

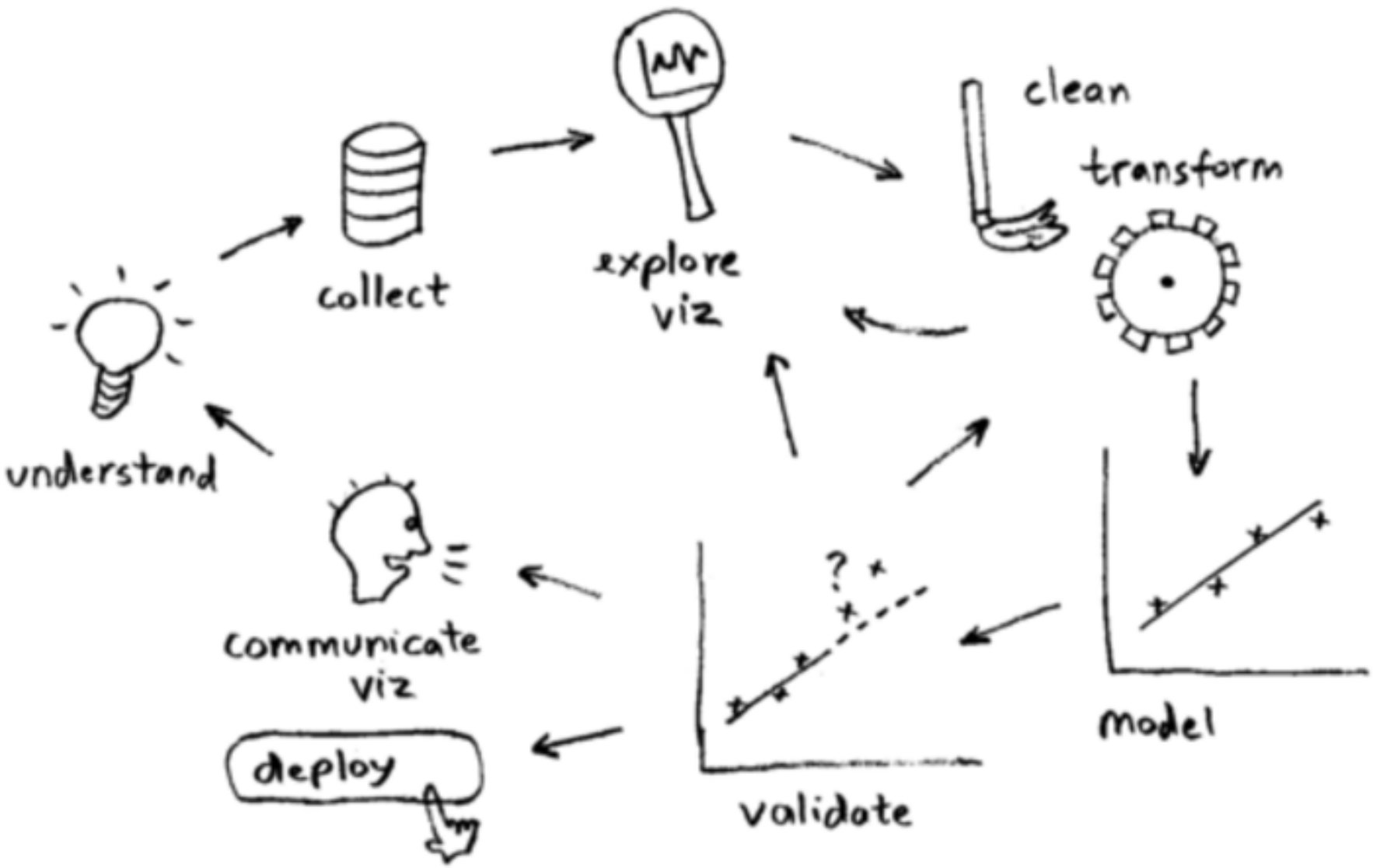
- Statistical inference and Machine Learning knowledge
- Computer Science and Engineering Experience
- Domain expertise
- Communication skills
- Data visualization skills

DATA SCIENCE WORKFLOW

Jeff Hammerbacher, Cloudera:

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

DATA SCIENCE WORKFLOW



source: DATA SCIENCE TOOLBOX SURVEY RESULTS... SURPRISE! R AND PYTHON WIN

DISCUSSION:

Problem: How would you implement “More items to consider” on Amazon.com?

In a small group, define the process an Amazon Data Scientist would work through to curate the “More items to consider” list for a particular user.

III. LAB: SETUP DEVENV