

Data Analytics

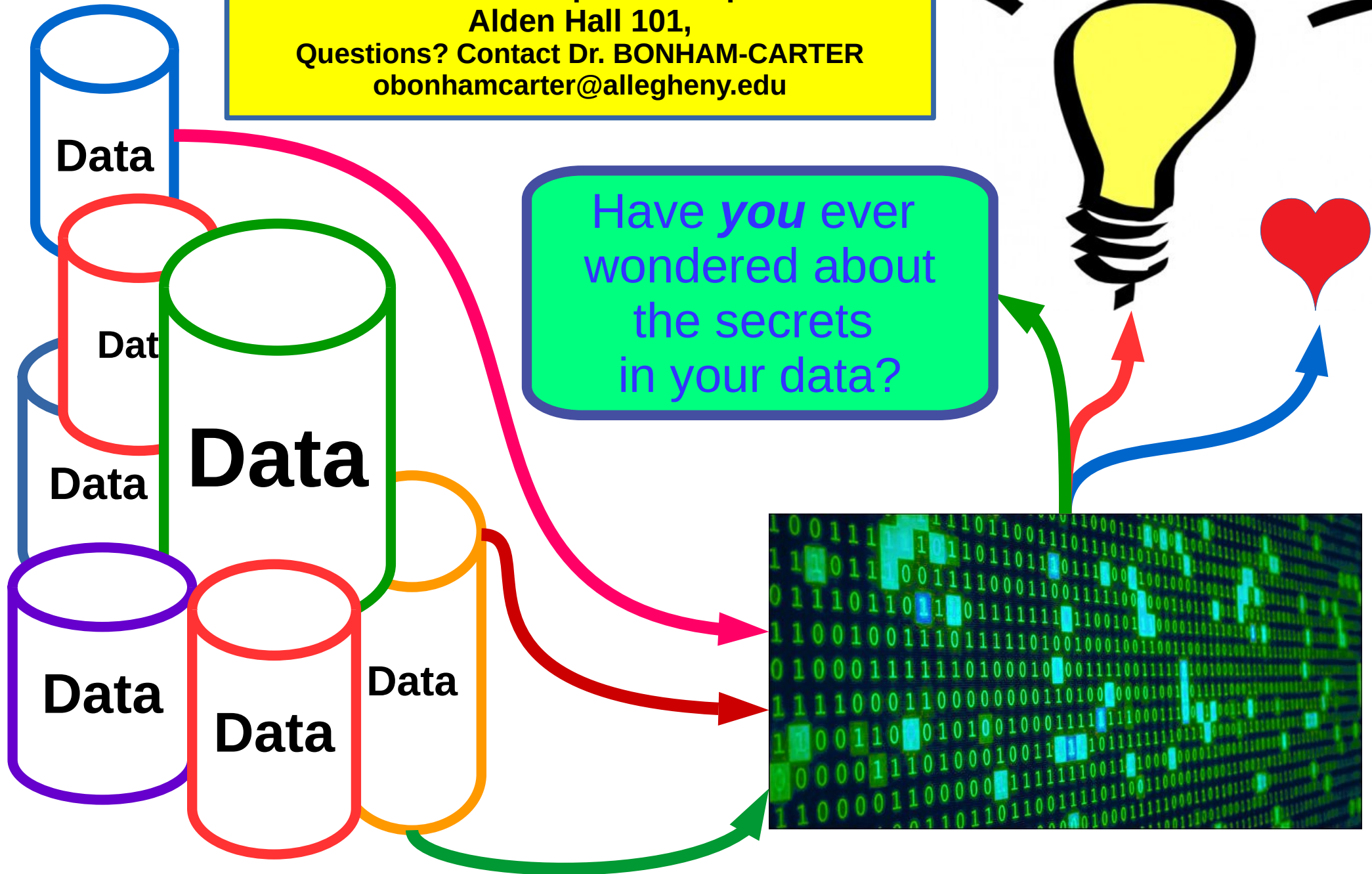
CS301

Introduction to Data Analytics

Week 1: 1st Sept
Fall 2020

Oliver BONHAM-CARTER

Data Analytics CMPSC*301
Lect: T/TH: 11:10 am – 12:25 pm
Lab: F: 3:00 pm – 4:50 pm
Alden Hall 101,
Questions? Contact Dr. BONHAM-CARTER
obonhamcarter@allegheny.edu





Links To Our Class

- Course web site:
<https://www.cs.allegheny.edu/sites/obonhamcarter/cs301.html>
 - Syllabus
 - “*Planning-Your-Time*”, class schedule
- Course calendar
 - <https://calendar.google.com/calendar/b/1?cid=Y184bXN0dDg2cW5oaWNjb3NxYWdibHNINzFva0Bncm91cC5jYWxlbmRhci5nb29nbGUuY29t>
- Zoom meetings for class and lab
 - <https://allegheny.zoom.us/j/95834628670>
 - Also see calendar for Zoom link



Flow in Our Class

Tuesday class

Tuesday group
In-person

Thursday group
Online

Thursday class

Tuesday group
Online

Thursday group
In-person

Friday Lab

Tuesday group
Online

Thursday group
Online



Two Class Groups

- Your group's day determines the weekday of class when are physically present.
- Tuesday group: Physically in class on Tuesdays
- Thursday group: Physically in class on Thursdays
- When you are not in class, it is expected that you will be coming to class via Zoom, or watching the recorded class videos.



Tuesday class

Tuesday group
In-person

Thursday group
Online

Thursday class

Tuesday group
Online

Thursday group
In-person

Friday Lab

Tuesday group
Online

Thursday group
Online

Computers and Information



ALLEGHENY
COLLEGE



- In this class, you will learn how to use machines to understand *trends* in data.
- (Making decisions by data)



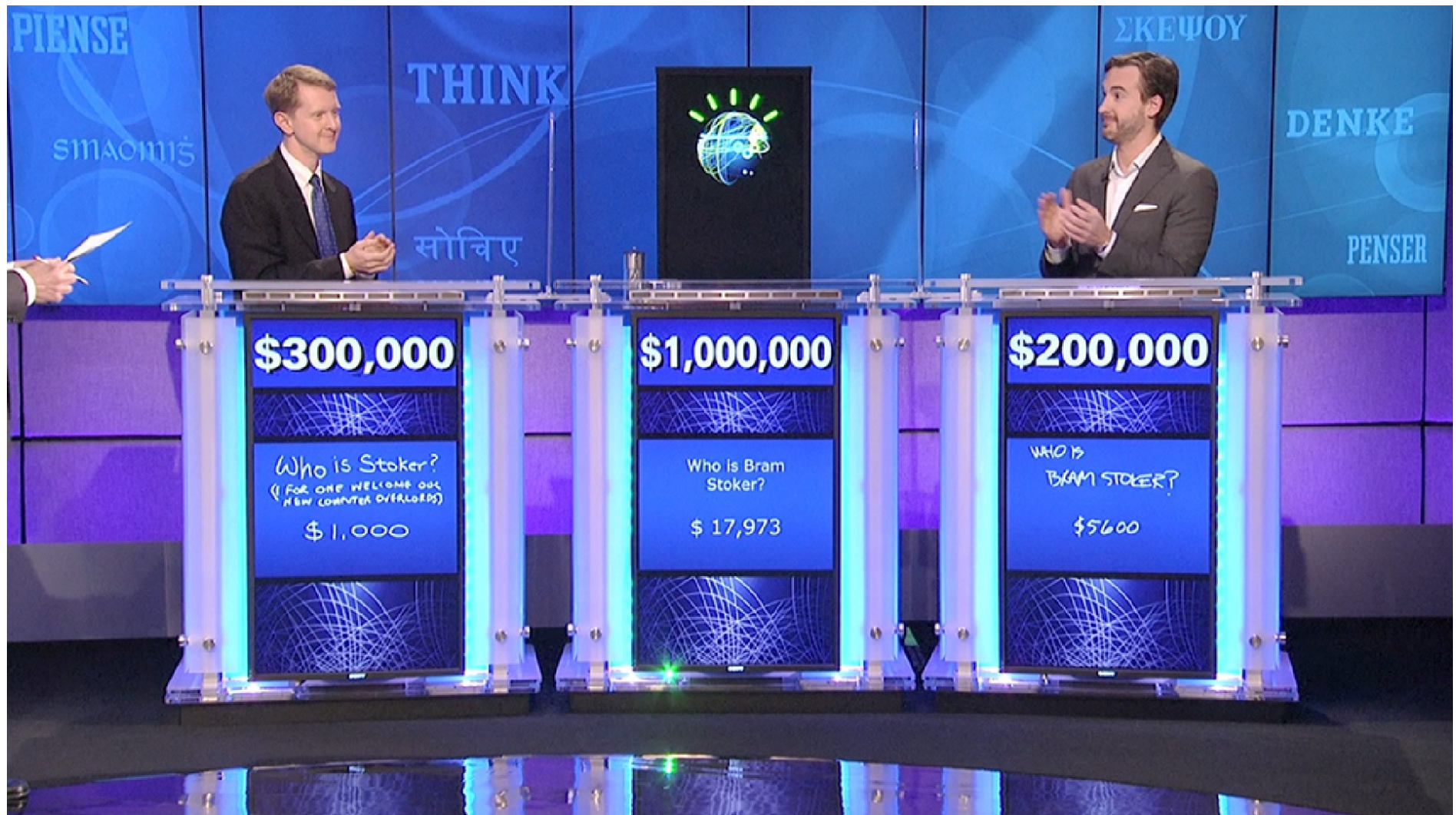


Analytics in Action

- The Jeopardy! Challenge of February 2011
- IBM's Watson beat the show's greatest champions: Ken Jennings and Brad Rutter.



Machines, Data and Information



WATSON for PRESIDENT



Is Watson magic??

<http://watson2016.com/>
(The Electronic Frontier Foundation)

Surrounded by DATA!

- We live in the “Information age”
- Actually, we live in the “Data age” since there is more data available than information
- Data != Information



Surrounded by DATA!

- It is cheap (and free or even lucrative) for businesses to collect data concerning:
 - in e-commerce,
 - customer behaviors,
 - purchase interests,
 - health and medical data.



Meet the Fitbit Family



EVERYDAY FITNESS

ACTIVE FITNESS

PERFORMANCE FITNESS

SMART SCALE

zip

one

flex

alta

chargeHR

blaze

surge

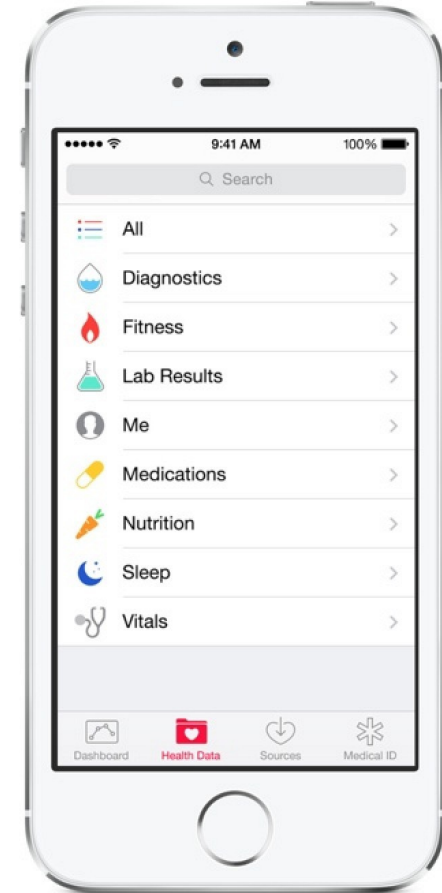
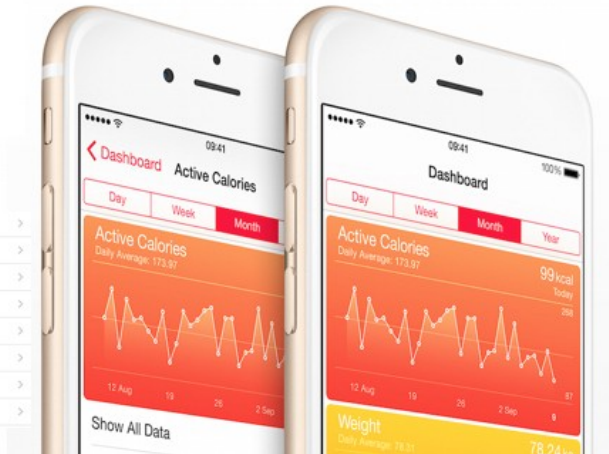
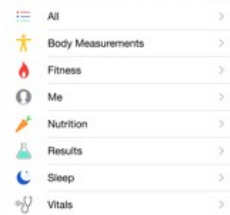
aria

We Voluntarily Give Away Our Data





Our Phones Create Data



- Smart phones constantly monitor us and keep data.
- **Q: How does the iPhone decide whether we are actually getting enough sleep?**
- **Who keeps the data?**



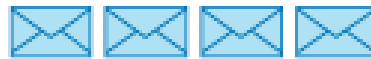
ALLEGHENY
COLLEGE

And So, Data is Increasing

 **65 billion**

Location-tagged payments
made in the U.S. annually

154 billion



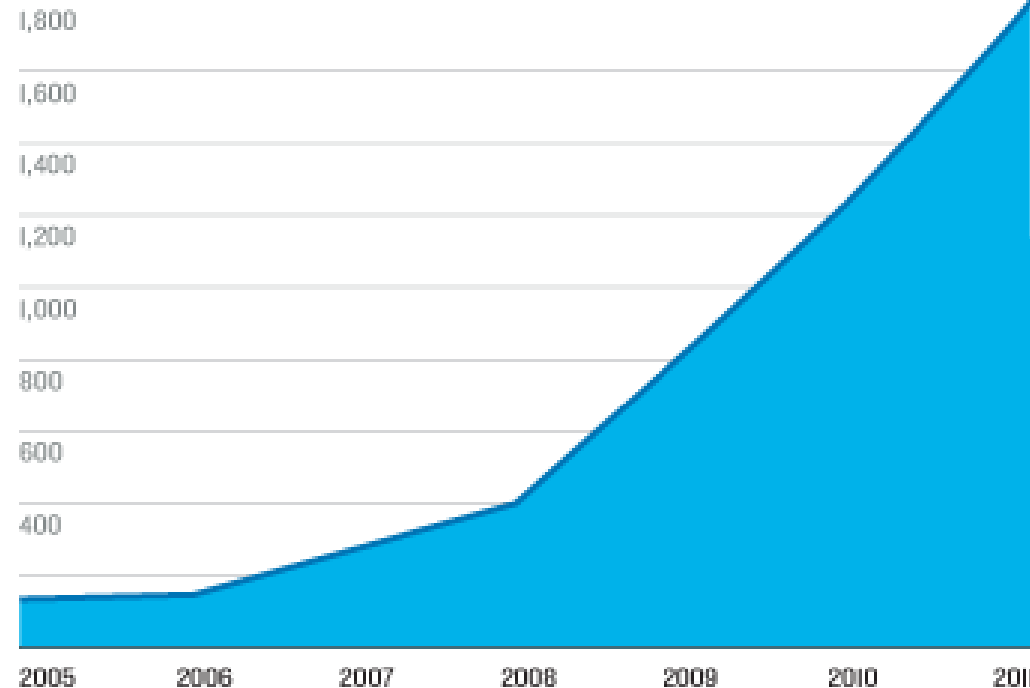
E-mails sent per day

 **87%**

U.S. adults whose location is
known via their mobile phone

Digital Information Created Each Year, Globally

2,000 BILLION GIGABYTES



2,000%

Expected increase in
global data by 2020

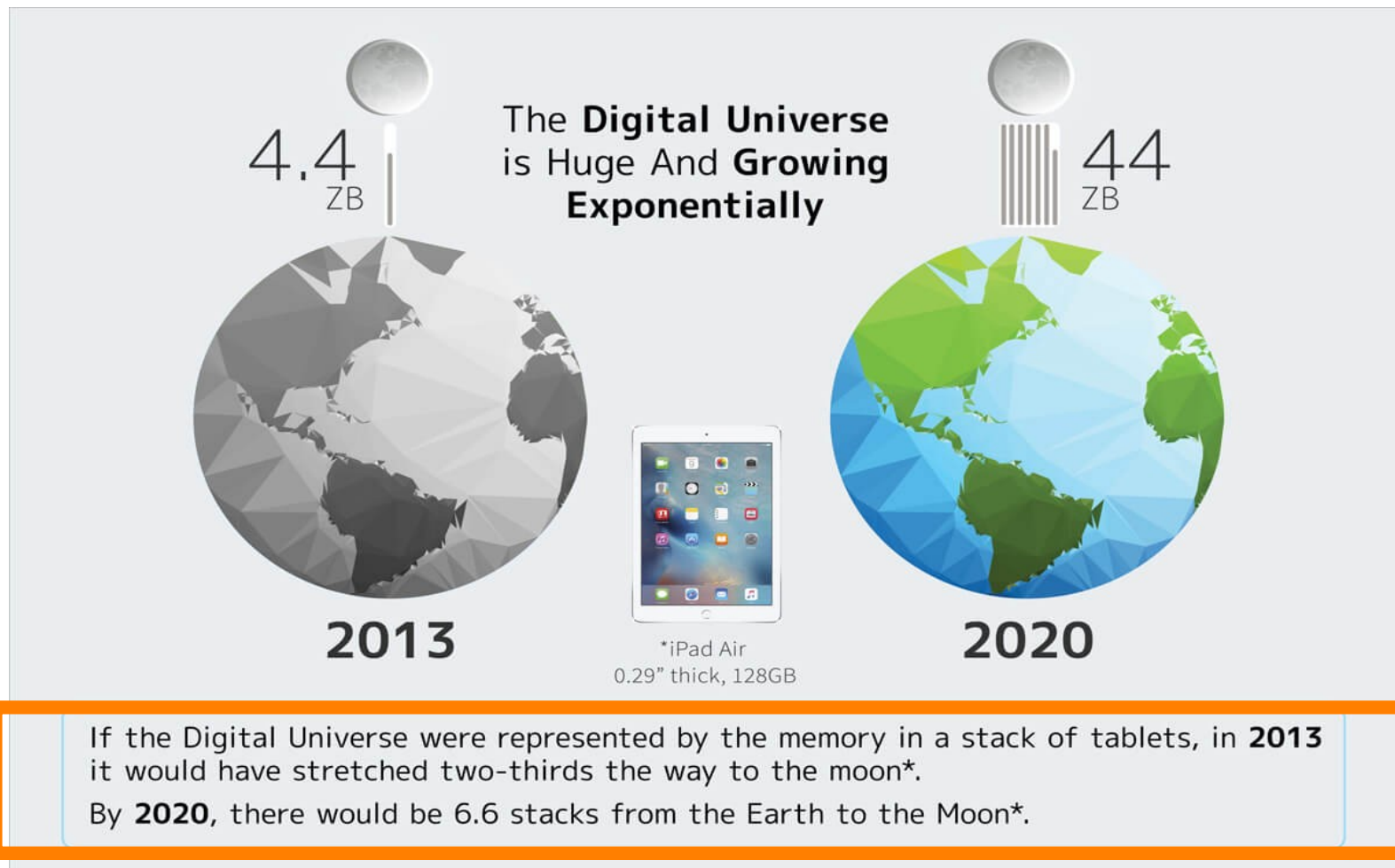
**III
Megabytes**

Video and photos stored
by Facebook, per user

75%

Percentage of all digital
data created by consumers

Data, Data, Data, Data!

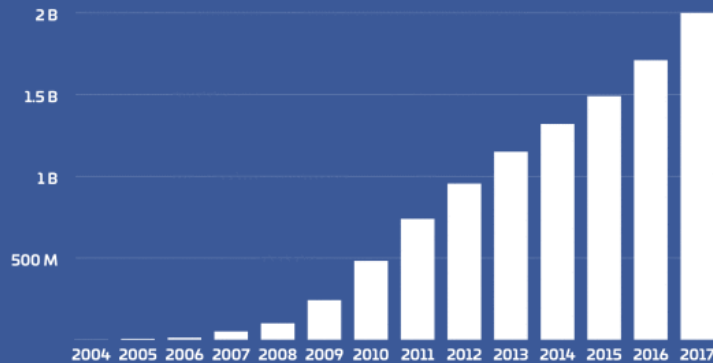


- How much data is there?
 - <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-creat-e-every-day-the-mind-blowing-stats-everyone-should-read/#76dc5de060ba>
 - <https://youtu.be/VLAnBI2B4OY>

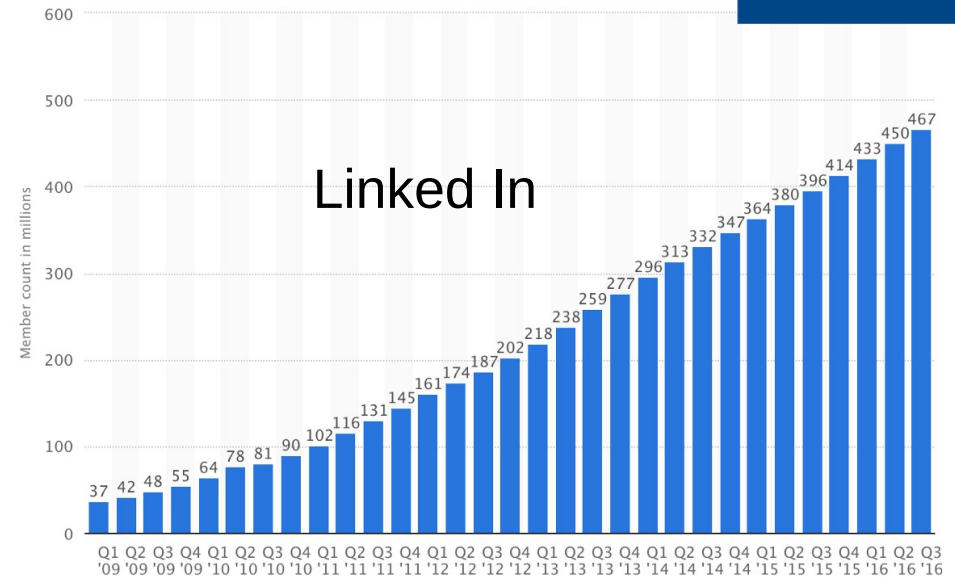


Sources of Data

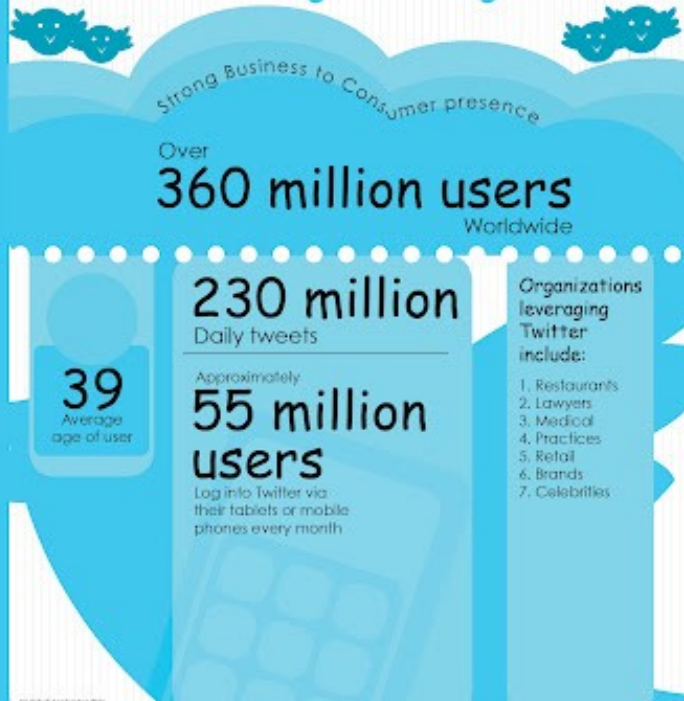
FACEBOOK MONTHLY ACTIVE USERS
JUNE 2017



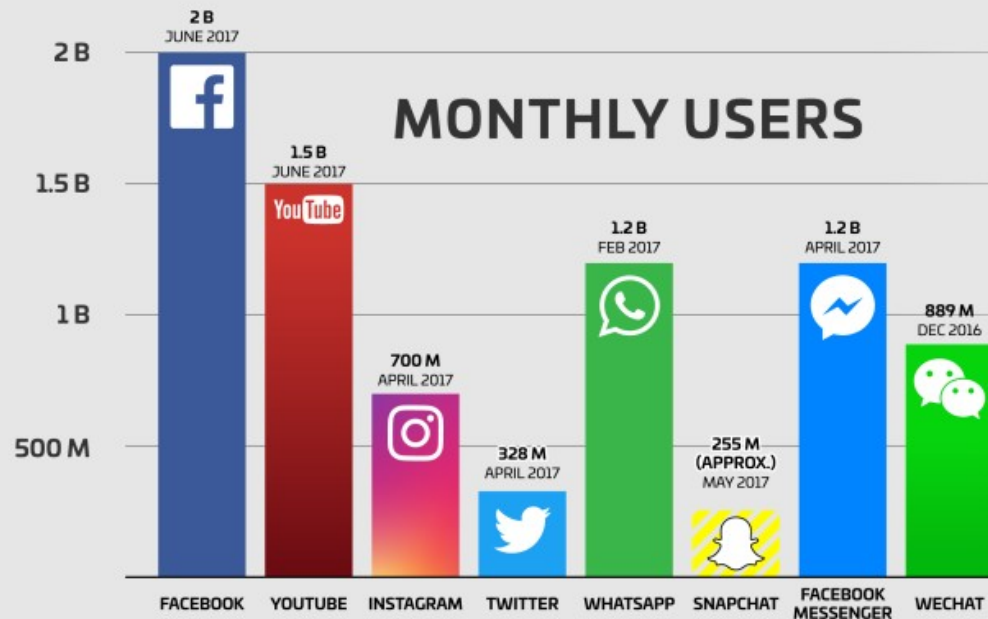
Linked In



twitter fast facts



MONTHLY USERS

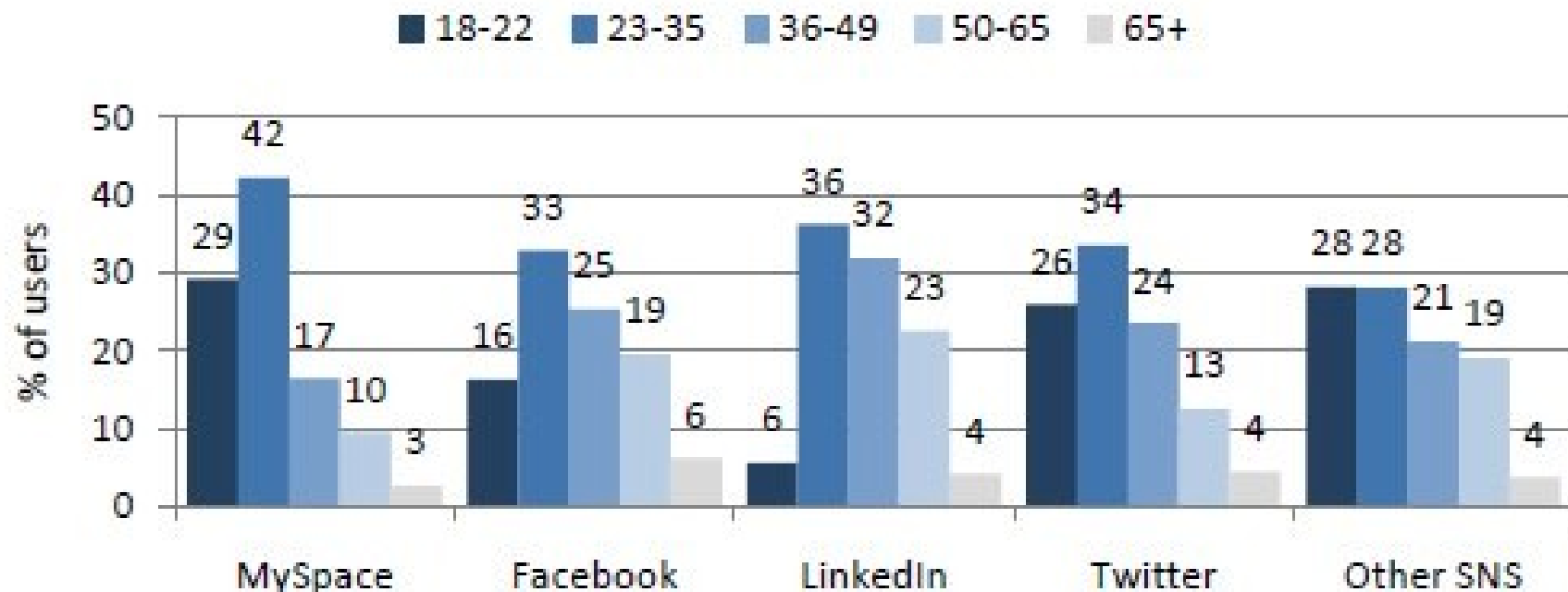




Data of User Ages

Age distribution by social networking site platform

% of social networking site users on each site who are in each age group. For instance, 29% of MySpace users are 18-22 years old.



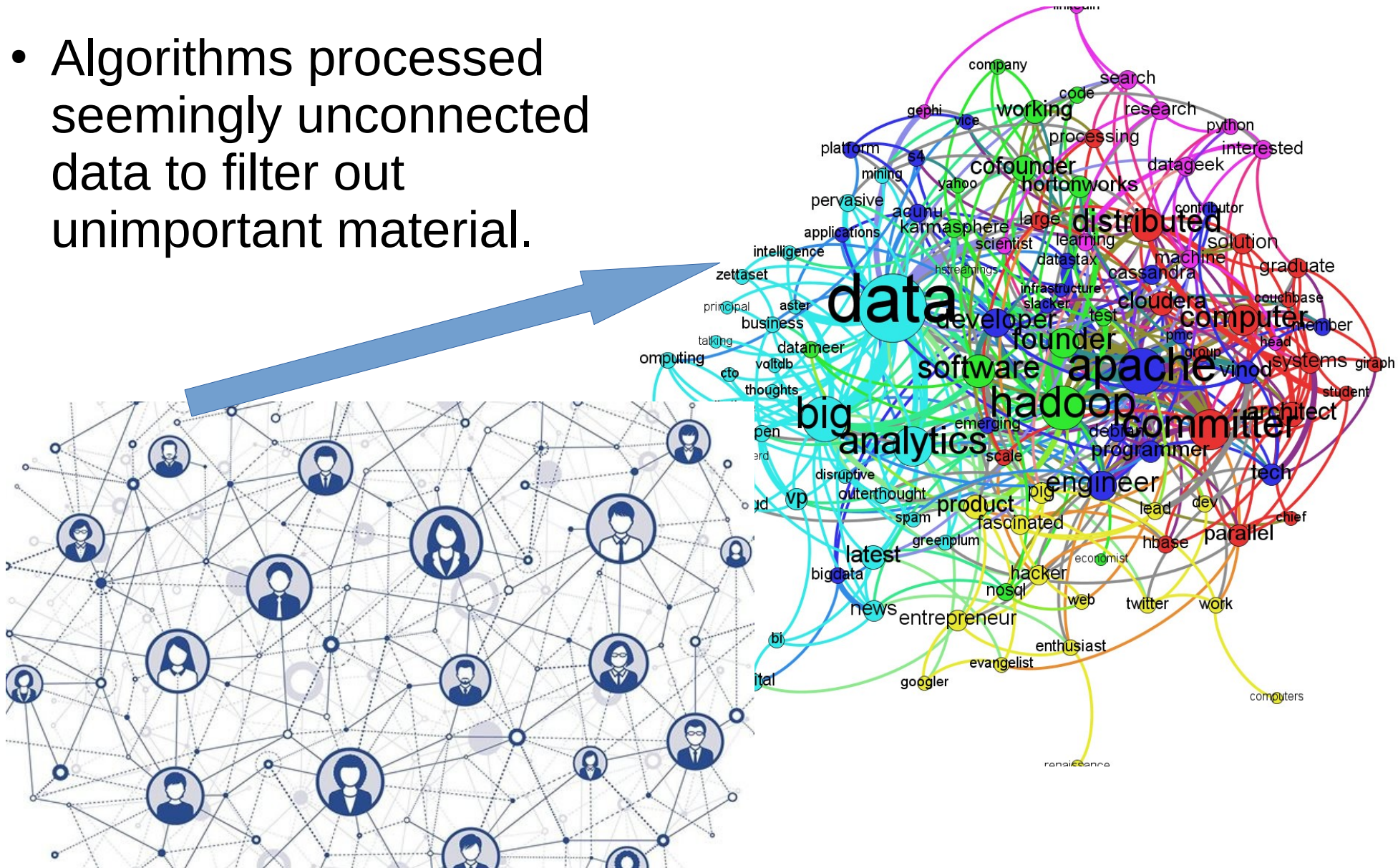
Source: Pew Research Center's Internet & American Life Social Network Site survey conducted on landline and cell phone between October 20-November 28, 2010. N for full sample is 2,255 and margin of error is +/- 2.3 percentage points. N for social network site and Twitter users is 975 and margin of error is +/- 3.5 percentage points.

By the way: These last slides visually describe trends ...

- Graphics have informed us:
 - Which apps are popular
 - Number of people in age groups for social networking sites
 - How much data is created each year, in relation to other years
 - Twitter “fast-facts”
 - Monthly users of services
 - Increases in Linked-In membership
- **How did we learn this information to make these previous visualizations?**

***Seriously, where did this
information come from???***

- Algorithms processed seemingly unconnected data to filter out unimportant material.





How Do We Know?

- The previous graphs came to us via raw Big Data from sites like Google, Facebook, Twitter and others.
- **Raw Data:** Seemingly meaningless clutter-like gibberish in which patterns are masked.

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

-- Gartner





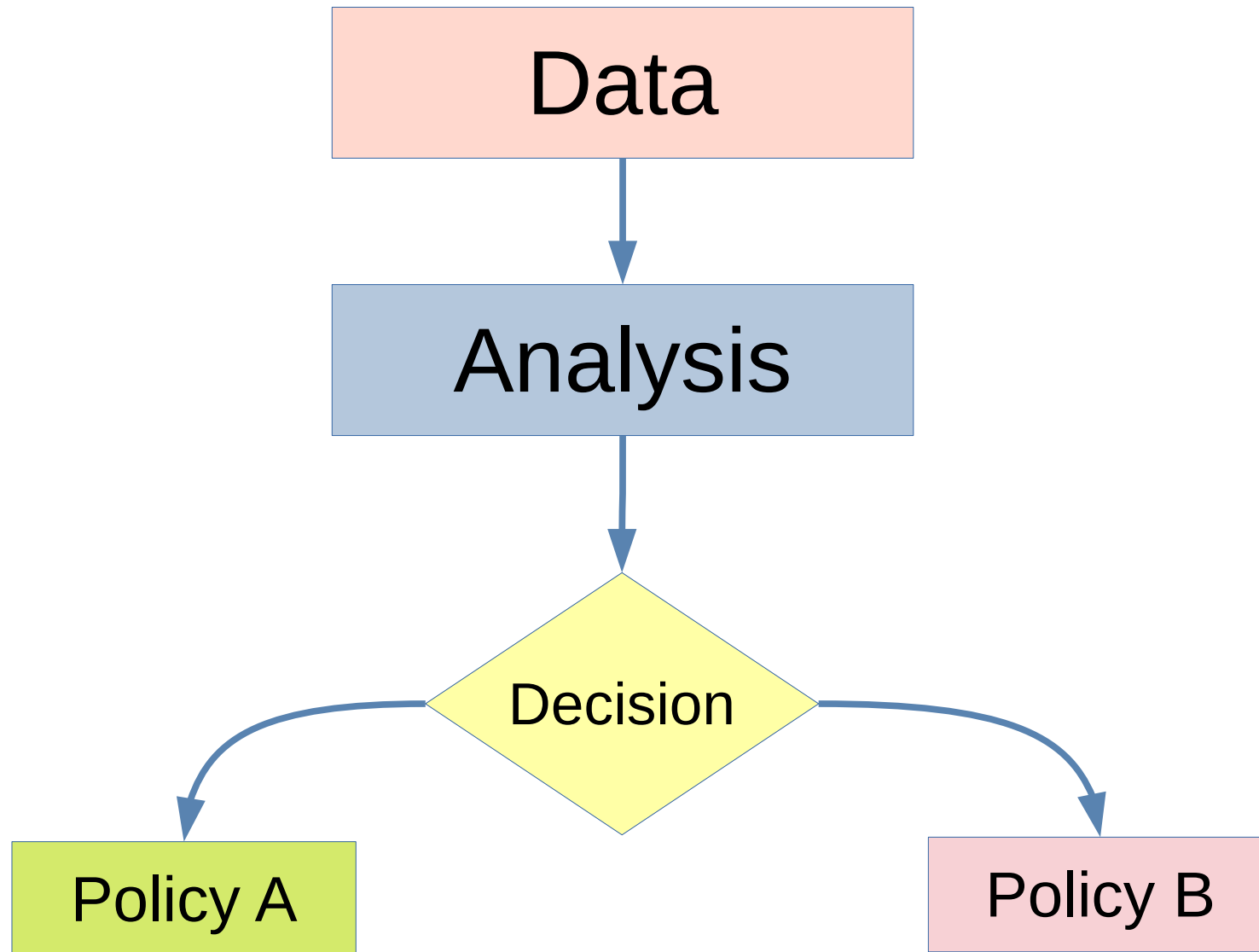
So, It Looks Like We Need **Data** to *Live Intelligently*

- Making *smart* (?) decisions:
 - *Can we make reliable decisions without **data**?*
 - *Is the quality of our society diminished by bad or missing **data**?*
 - *How can we improve commerce, trade without knowledge from **data**?*
 - *How can we make better health decisions without knowledge from **data**?*
- You could give surveys to gather ideas from people but few are likely to respond...

***But, when was the last time
YOU took a survey?***



Policy Creation by Analytics



Thus, Much Interest in Data Analytics

- The present and future are information-driven
- Some of the decisions made after studying trends in a population
 - **Commerce:** what have customers already bought?
 - **Media:** What themes of films, music make money?
 - **Industry:** What products should we make to build, satisfy a market? Which market?
 - **Life Sciences and Medicine:** Reasons for sickness? Bad types of foods? Exposures to toxins?





Your Career Could Be Here!

- *“Big Data & Analytics Is The Most Wanted Expertise By 75% Of IoT (Internet of Things) Providers”*
 - <https://www.forbes.com/sites/louiscolumbus/2017/08/21/big-data-analytics-is-the-most-wanted-expertise-by-75-of-iot-providers/#52082a4e5188>
- *“75% of IoT providers are prioritizing big data and analytics expertise in their hiring decisions.”*
 - <http://www.forbes.com/sites/louiscolumbus/2017/08/21/big-data-analytics-is-the-most-wanted-expertise-by-75-of-iot-providers/>
- *“68% of vendors developing IoT solutions are struggling to find and recruit employees with development expertise.”*
 - <http://www.forbes.com/sites/louiscolumbus/2017/08/21/big-data-analytics-is-the-most-wanted-expertise-by-75-of-iot-providers/>

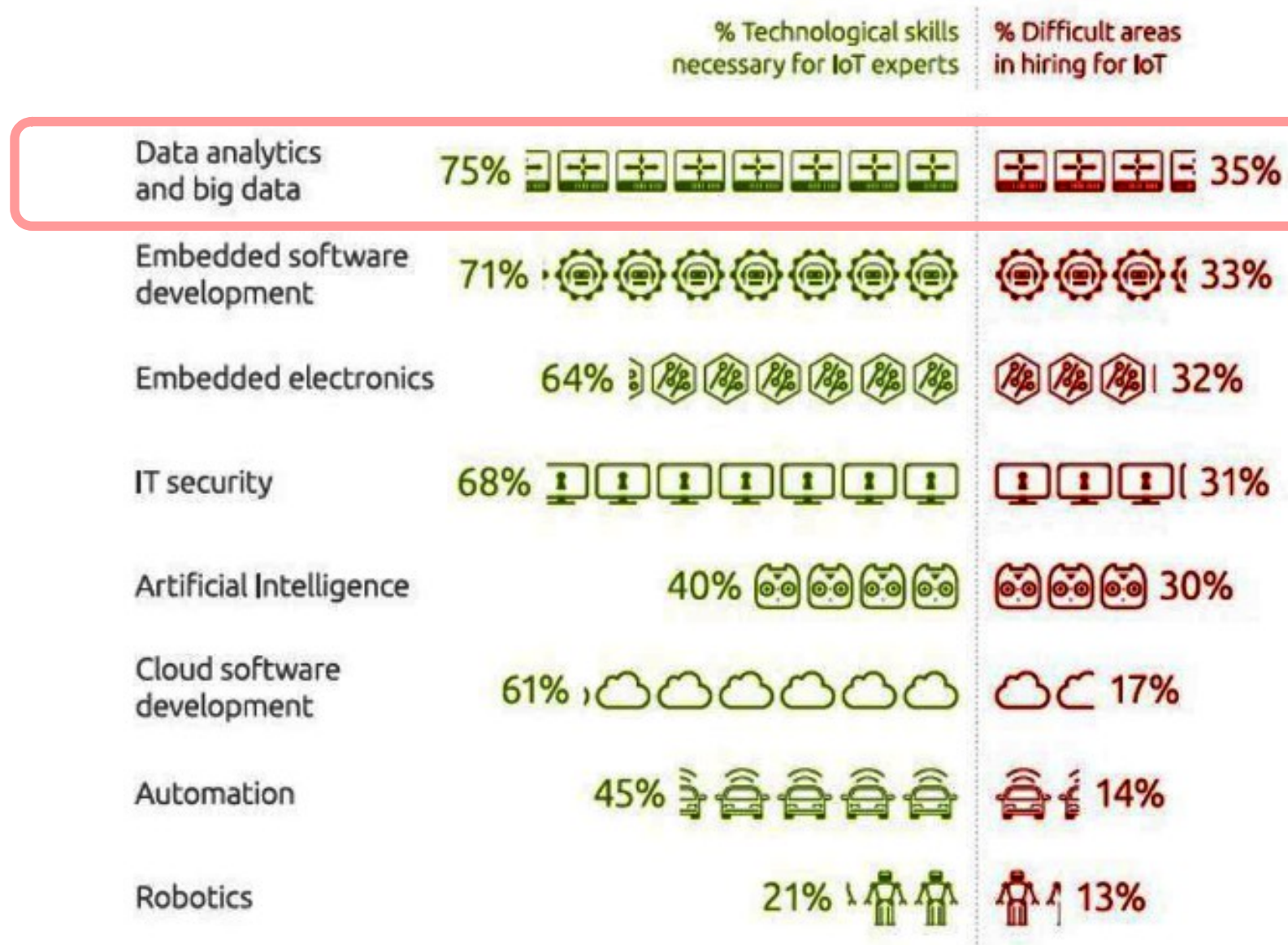


Forbes

- ***“75% of firms are prioritizing big data and analytics expertise in their hiring decisions, stating that having these skills is critical for any candidate to be considered an IoT (Internet of Things) expert.”***



Hard to Hire Skilled People






Glassdoor Informs of Careers

Prodigy Game 4.3★
Game Data Analyst

Job Company Rating

Bonus Points For:

- Degree in Engineering, Computer Science, Stats, Mathematics
- Demonstrated ability to solve hard mathematical, algorithmic, and statistical problems
- Expertise in advanced game analytics - user segmentation, player modelling
- Knowledge of Python/R languages
- Experience working with cloud platforms like AWS (Redshift, Athena, S3)
- Experience working with A/B testing and experimentation
- Significant accomplishments that required both technical and strategic capabilities, such as research projects, open source software contributions, and entrepreneurship




Prodigy Game
Game Data Analyst
Oakville

4.3★

ADURO 3.9★
Data Analyst

Job Company Rating Salary Reviews Why W

- Demonstrated ability to maintain absolute confidentiality
- Proficiency in BI Tools (Sisense, Tableau, PowerPivot, DOMO or other comparable tools)
- Aptitude with SQL, R, and other languages supporting data analysis
- Experience with C based object-oriented programming
- Experience working within a large reporting Data Warehouse
- Experience working with web and application analytics tools (Firebase, Google Analytics)
- Experience using analytics to support product development
- Familiar with source-control repositories and associated practices (Git, GitHub)
- Proven attention to detail and accuracy



ADURO
Data Analyst
Redmond, WA
\$46K-\$78K (Glassdoor Est.) ⓘ

3.9★

- **An Analytics Expert**
 - To apply data analysis skills to help development teams better understand users by applying analytics
 - Find and integrate data from multiple sources to provide analysis
 - Develop tools & methods to ensure data accuracy
 - Collaborate with Data & Analytics team members
 - R skills



Consider This ...

- You are given the lists of words from several main stream-news articles.
- Pick a list to work on with a group of your peers.
- Although the article text cannot be read directly, can you determine the general sense of the article from a list of its words?
- What is the general subject of your article?
 - Are there names of people you recognize in your list? What can you infer about the article from the name(s)?
 - Do the listed nouns support your conclusions?
 - What type of media source would contain such a story?

THINK

Find the data at:

https://www.cs.allegHENY.edu/sites/obonhamcarter/cs301_resources.html



Please Read for Next Class

- Come prepared to discuss
- *Twelve Million Phones, One Dataset, Zero Privacy*, A New York Times opinion piece
- Link:
<https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>

Opinion | THE PRIVACY PROJECT

Twelve Million Phones, One Dataset, Zero Privacy

By Stuart A. Thompson and Charlie Warzel

DEC. 19, 2019

THINK