
DOGBERT CONSULTS
MY DATA-MINING SOFTWARE HAS FOUND ANOTHER MESSAGE FROM GOD.
www.dilbert.com scottadams@aol.com
IT SAYS YOU'VE BEEN STEALING LUNCHES FROM THE REFRIGERATOR IN THE BREAK ROOM.
© 1999 United Feature Syndicate, Inc.
THEN IT SAYS, "HA HA. THAT WASN'T PUDDING!"


November 2011

# Data Mining: Example (myth?)

- What products are sold together with diapers in a grocery store/supermarket?
  - Answer: Beer
- Highest volume on Friday afternoons
  - By men between the ages of 25 and 35.
- What did the supermarket do as a consequence?
  - They put the beer display next to the diapers.
- Beer sales skyrocketed.

# Data Mining: Example

- What item saw the greatest increase in sales before hurricanes?



# Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes to .....
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

# What types of data?

- World Wide Web
  - Billions of documents, Access logs
  - Linked structure (Web graph)
- Financial interactions
  - ATM/Credit card
  - Deposits/Withdraws
- User interactions
  - Phone call records
- Sensor technologies
  - Wearable sensors, smartphones,….
- Internet of Things
  - Smart devices communicating with one another

# What types of data?

- Business transactions
- Social media sites
- Digital pictures and videos
- Cell phone GPS signals
- Scientific Data
- ….

# How much data?

- Every day, we create 2.5 quintillion ($10^{18}$) bytes of data
- 90% of the data in the world today has been created in the last two years alone.

# How much data?

| SI decimal prefixes | | Binary usage |
|---|---|---|
| Name (Symbol) | Value | |
| Kilobyte (KB) | $10^3$ | $2^{10}$ |
| Megabyte (MB) | $10^6$ | $2^{20}$ |
| Gigabyte (GB) | $10^9$ | $2^{30}$ |
| Terabyte (TB) | $10^{12}$ | $2^{40}$ |
| Petabyte (PB) | $10^{15}$ | $2^{50}$ |
| Exabyte (EB) | $10^{18}$ | $2^{60}$ |
| Zettabyte (ZB) | $10^{21}$ | $2^{70}$ |
| Yottabyte (YB) | $10^{24}$ | $2^{80}$ |

# How much data?

- YouTube
  - July 2011 - 48 hours of video uploads/minute
  - 1 hr of video = 80GBytes (640 x 480 x 30fps x 8bpp)
  - With 10:1 compression ratio = 8Gbytes
  - 2014: 300 hours/min
  - 2017: 500 hours/min
- More video is uploaded to YouTube in 60 days than the 3 major US networks created in 60 years.
- 1.5 billion active users
- 1 billion hours of videos watched per day

YouTube Stats

# How much data?

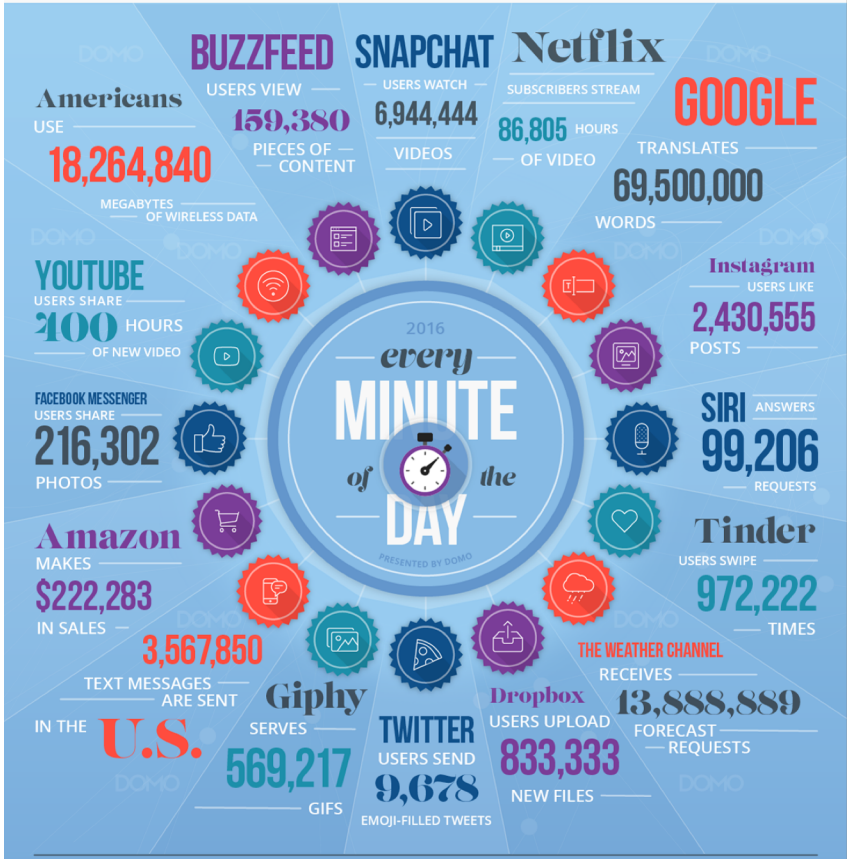
- Facebook
  - Over 2 billion(monthly) active users (1 billion daily users)
  - 6 new profiles are created every second
  - 300 million photos are uploaded per day (2015)
- Twitter
  - 336 million monthly active users
  - 500 Million tweets per day (2018)
  - 6000 tweets per second (2018)
- Flickr
  - Over 10 Billion images (2015)
  - Up to 25 Million added per day (high traffic day)
  - 75 million photographers
- Digital Images
  - 1 trillion photos taken in 2015
  - Over 6 billion smart phones by 2020 (2.6 Billion in 2015)

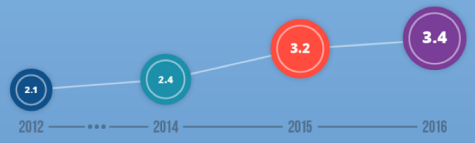
# Machine-to-Machine Data

- Self-Driving Cars
  - 3 PBytes per car per year
- Flying Cars
- Sensors
  - 1Trillion sensors on the Internet by 2020
  - Songdo (South Korea) Smart City
- Smart "things"
  - Windows, homes, hotels
  - Bridges
  - Tractors
  - TV

# DATA NEVER SLEEPS 4.0

**How much data is generated every minute?** In the fourth annual edition of Data Never Sleeps, newcomers like Giphy and Facebook Messenger illustrate the rise of our multimedia messaging obsession, while veterans like Youtube and Snapchat highlight our insatiable appetite for video. Just how many GIFs, videos, and emoji-filled Tweets flood the internet every minute? See for yourself below.




GLOBAL INTERNET POPULATION GROWTH 2012–2016
(IN BILLIONS)

Data has become the new enterprise currency. The ability to collect, analyze, and leverage it effectively will distinguish the best from the rest. Domo helps you stay ahead by bringing your data and people together in the cloud, where everyone in your organization can easily access the information they need to make faster, better-informed decisions and optimize business performance.

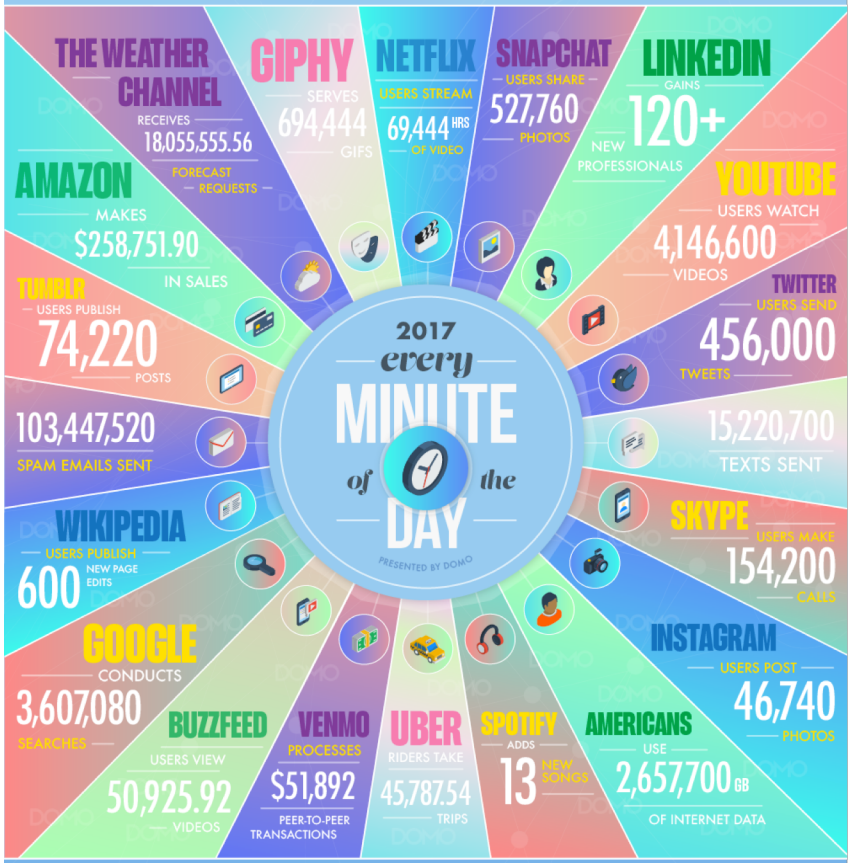
**Learn more at www.domo.com**

**SOURCES:** SNAPCHAT, NETFLIX, GOOGLE, INSTAGRAM, TINDER, THE WEATHER COMPANY, DROPBOX, GITHUB, GIPHY, YOUTUBE, BUZZFEED, AMAZON, CTIA, MARY MEEKER'S 2016 INTERNET TRENDS REPORT, USA TODAY, GLOBAL WEB INDEX

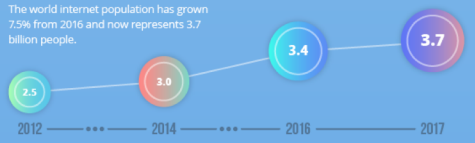
# DATA NEVER SLEEPS 5.0

## How much data is generated *every minute?*

90% of all data today was created in the last two years—that's 2.5 quintillion bytes of data per day. In our 5th edition of Data Never Sleeps, we bring you the latest stats on just how much data is being created in the digital sphere—and the numbers are staggering.




The world internet population has grown 7.5% from 2016 and now represents 3.7 billion people.




GLOBAL INTERNET POPULATION GROWTH 2012–2017
(IN BILLIONS)

With each click, swipe, share, and like, businesses are using data to make decisions about the future. Domo gives everyone in your business real-time access to data from virtually any data source in a single platform for smarter decision-making at any moment.

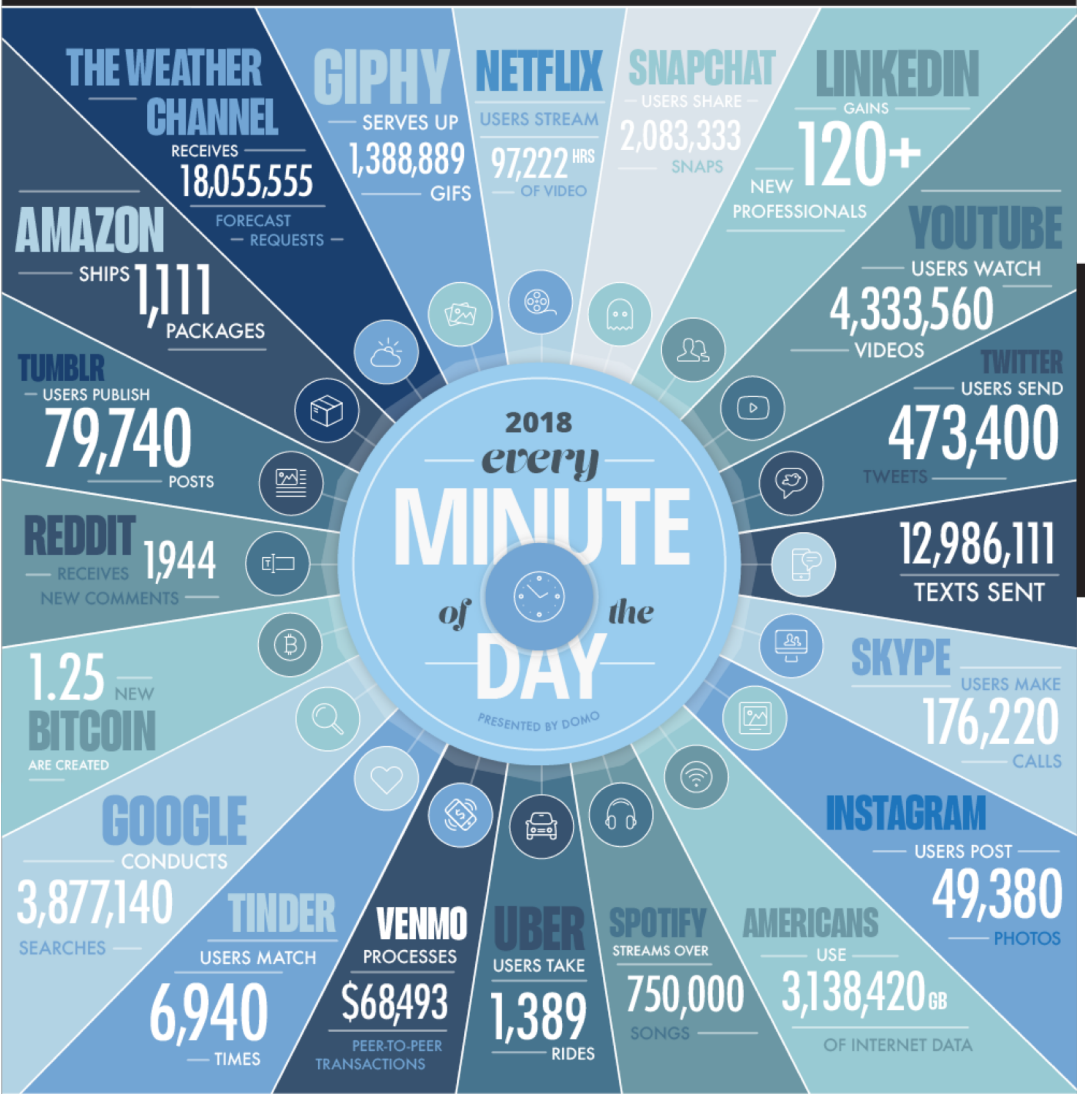
**Learn more at domo.com**

**SOURCES:** EXPANDEDRAMBLINGS.COM, WEARESOCIAL.COM, WIKIPEDIA, FORBES, ADWEEK.COM, FORTUNE.COM, BLOOMBERG.COM, ONEREACH.COM, IBM, BUZZFEED, INTERNET LIVE STATS, INTERNET WORLD STATS, BBC

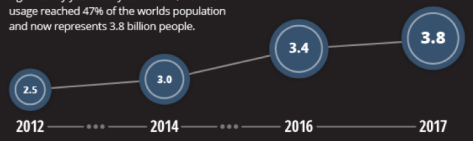
# DATA NEVER SLEEPS 6.0

## How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, but they're not slowing down. By 2020, it's estimated that for every person on earth, 1.7 MB of data will be created every second. In our 6th edition of Data Never Sleeps, we once again take a look at how much data is being created all around us every single minute of the day—and we have a feeling things are just getting started.




The world's internet population is growing significantly year-over-year. In 2017, internet usage reached 47% of the worlds population and now represents 3.8 billion people.



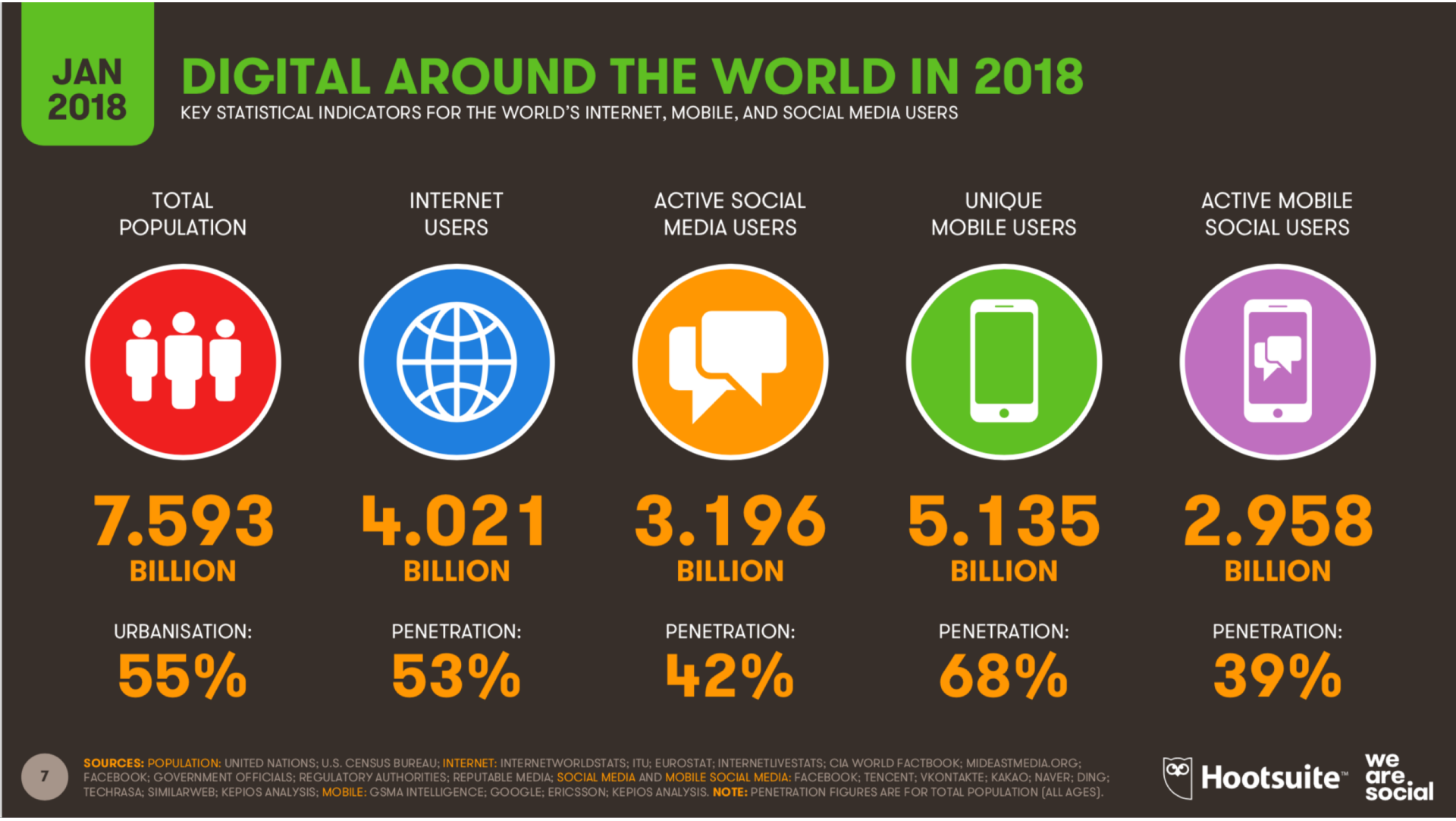

GLOBAL INTERNET POPULATION GROWTH 2012–2017
(IN BILLIONS)

The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

**Learn more at domo.com**

**SOURCES:** STATISTA, LINKEDIN, INTERNET LIVE STATS, EXPANDED RAMBLINGS, SLASH FILM, RIAA, BUSINESS OF APPS, INTERNATIONAL TELECOMMUNICATIONS UNION, INTERNATIONAL DATA CORPORATION

# Internet Trends



# Internet Trends

JAN 2018

## GLOBAL ANNUAL DIGITAL GROWTH

YEAR-ON-YEAR CHANGE IN KEY STATISTICAL INDICATORS

| INTERNET USERS | ACTIVE SOCIAL MEDIA USERS | UNIQUE MOBILE USERS | ACTIVE MOBILE SOCIAL USERS |
|---|---|---|---|
|  |  |  |  |
| **+7%** | **+13%** | **+4%** | **+14%** |
| SINCE JAN 2017 | SINCE JAN 2017 | SINCE JAN 2017 | SINCE JAN 2017 |
| **+248 MILLION** | **+362 MILLION** | **+218 MILLION** | **+360 MILLION** |



SOURCES: POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU; INTERNET: INTERNETWORLDSTATS; ITU; EUROSTAT; INTERNETLIVESTATS; CIA WORLD FACTBOOK; MIDEASTMEDIA.ORG; FACEBOOK; GOVERNMENT OFFICIALS; REGULATORY AUTHORITIES; REPUTABLE MEDIA; SOCIAL MEDIA AND MOBILE SOCIAL MEDIA: FACEBOOK; TENCENT; VKONTAKTE; KAKAO; NAVER; DING; TECHRASA; SIMILARWEB; KEPIOS ANALYSIS; MOBILE: GSMA INTELLIGENCE; GOOGLE; ERICSSON; KEPIOS ANALYSIS. GROWTH DATA: WE ARE SOCIAL & HOOTSUITE'S *DIGITAL IN 2017* REPORT.

Hootsuite™ we are social

# Internet Trends

JAN 2018

## INTERNET USE

BASED ON ACTIVE INTERNET USER DATA, AND ACTIVE USE OF INTERNET-POWERED MOBILE SERVICES

TOTAL NUMBER OF ACTIVE INTERNET USERS

**4.021**
**BILLION**

INTERNET USERS AS A PERCENTAGE OF THE TOTAL POPULATION

**53%**

TOTAL NUMBER OF ACTIVE MOBILE INTERNET USERS

**3.722**
**BILLION**

MOBILE INTERNET USERS AS A PERCENTAGE OF THE TOTAL POPULATION

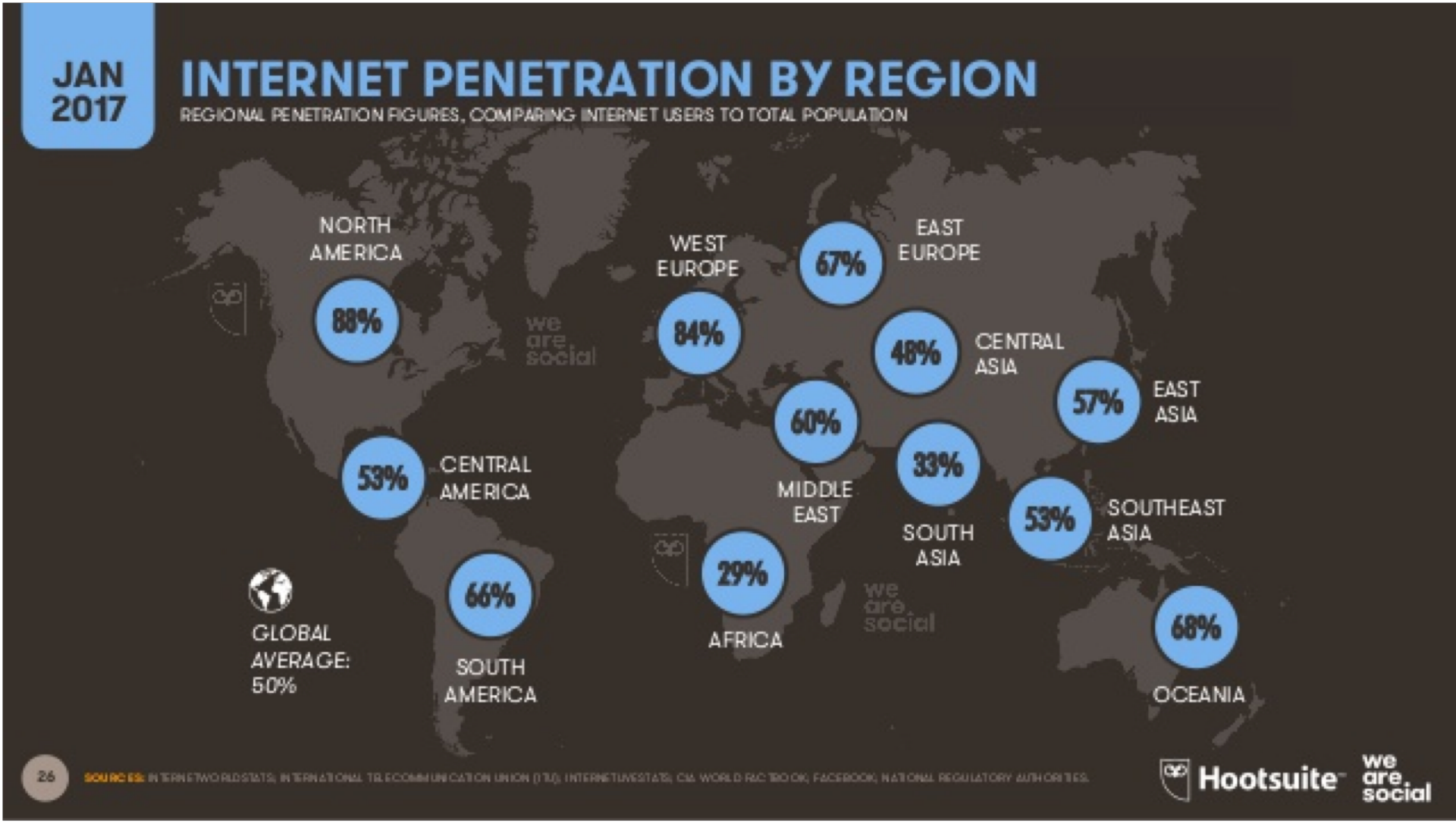
**49%**



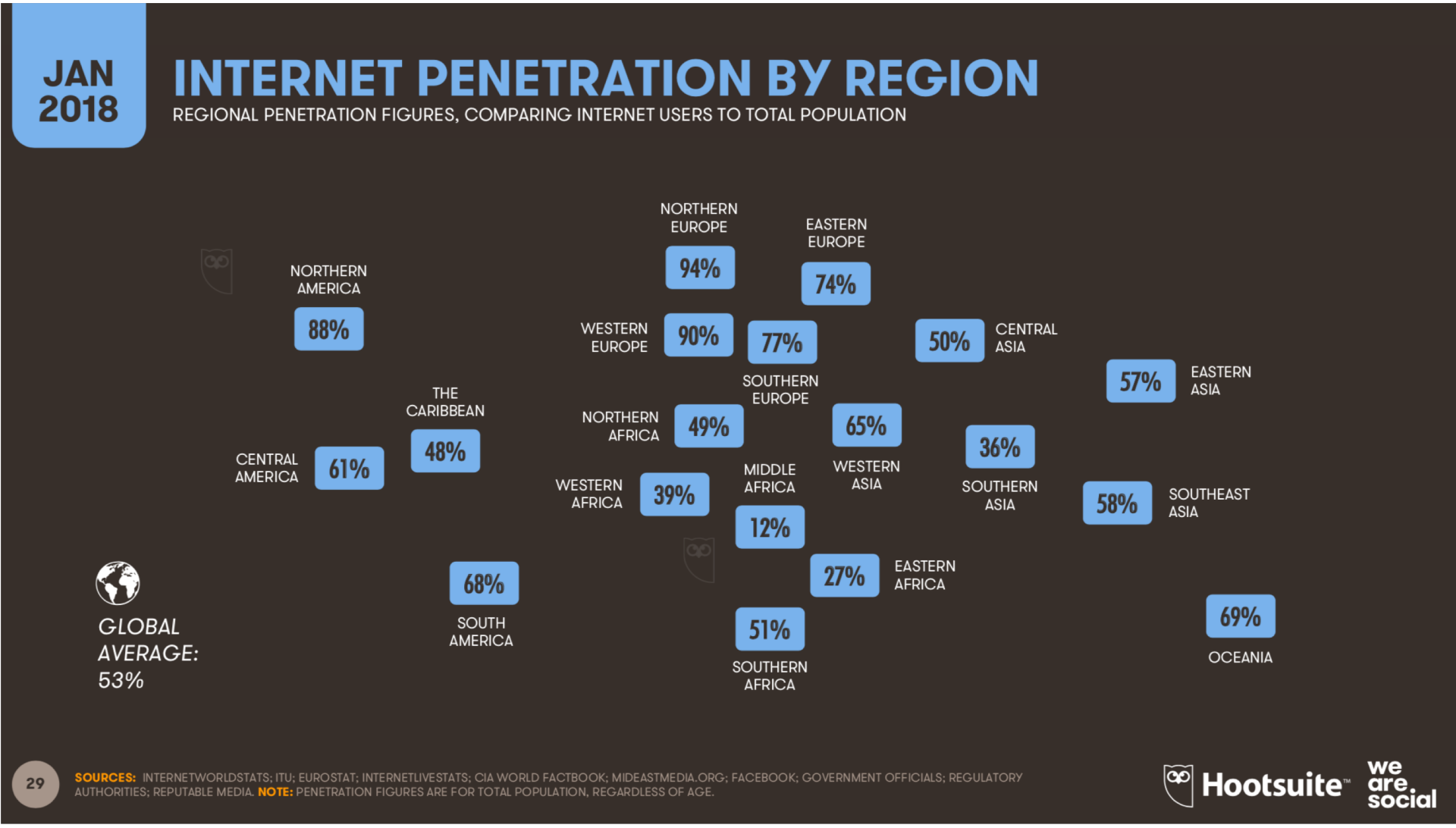
**SOURCES:** INTERNETWORLDSTATS; ITU; EUROSTAT; INTERNETLIVESTATS; CIA WORLD FACTBOOK; MIDEASTMEDIA.ORG; FACEBOOK; GOVERNMENT OFFICIALS; REGULATORY AUTHORITIES; REPUTABLE MEDIA.. **NOTES:** PENETRATION FIGURES ARE FOR FULL POPULATION, REGARDLESS OF AGE.

Hootsuite™ we are social

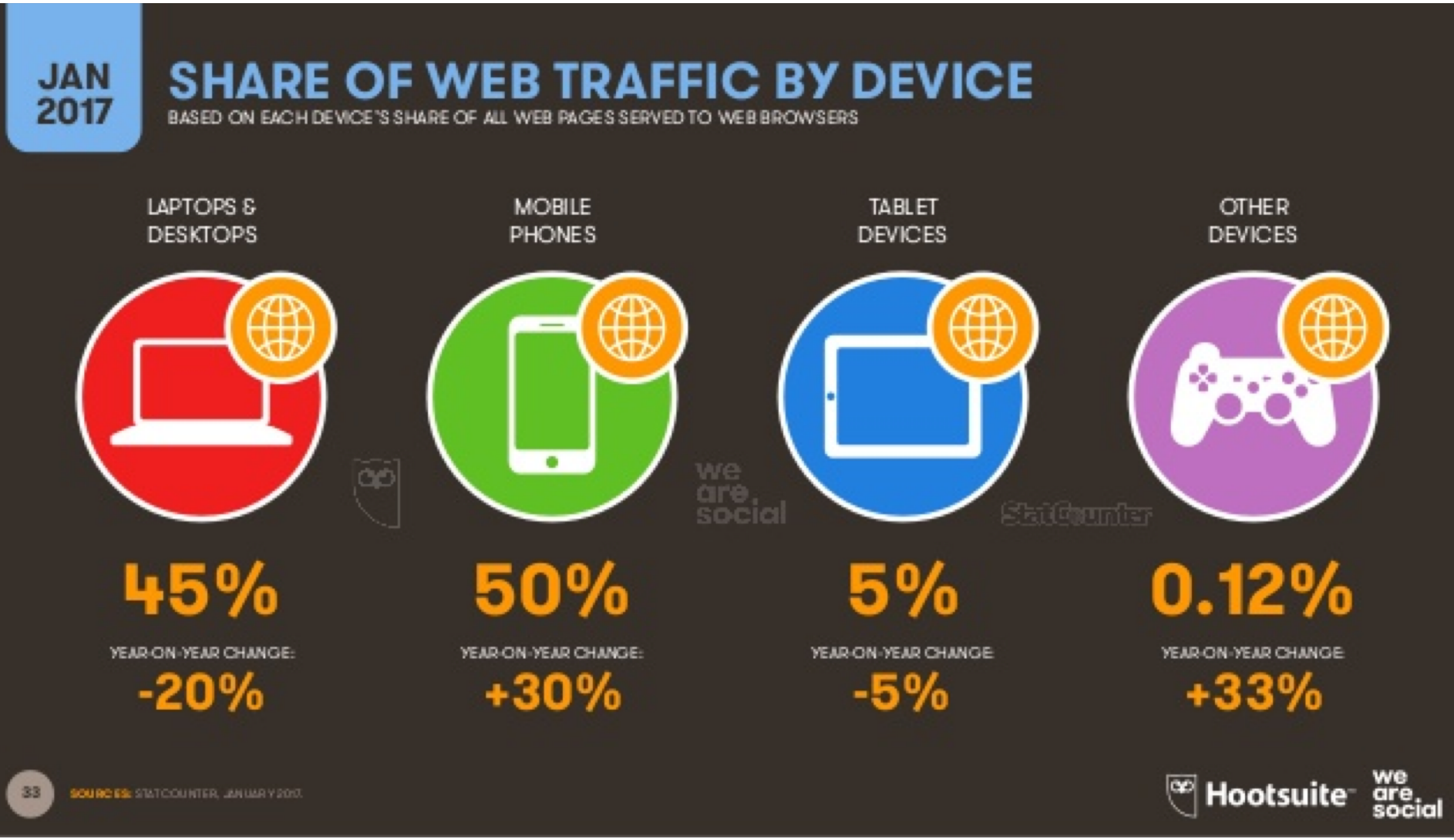
# Internet Trends



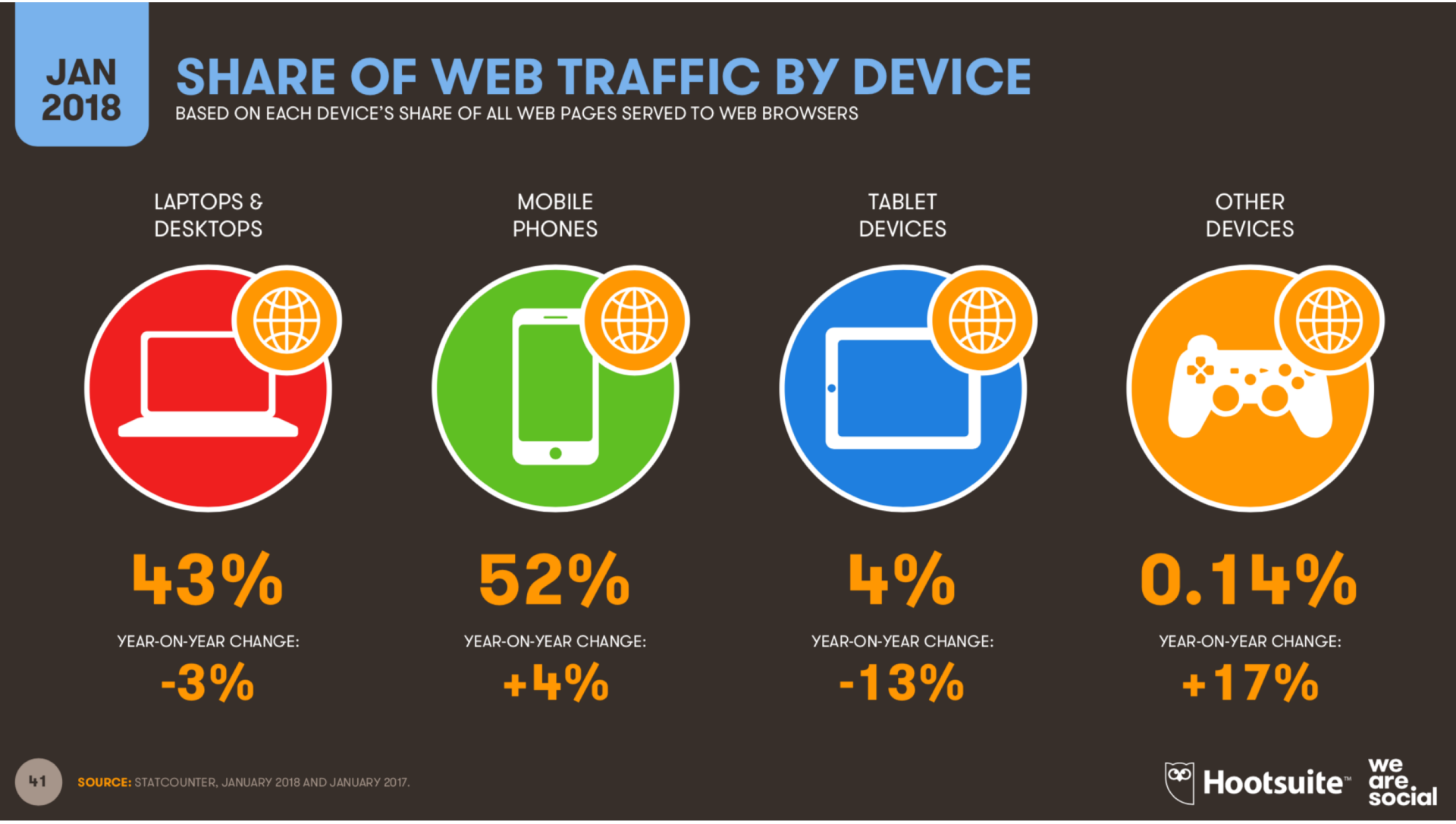
# Internet Trends



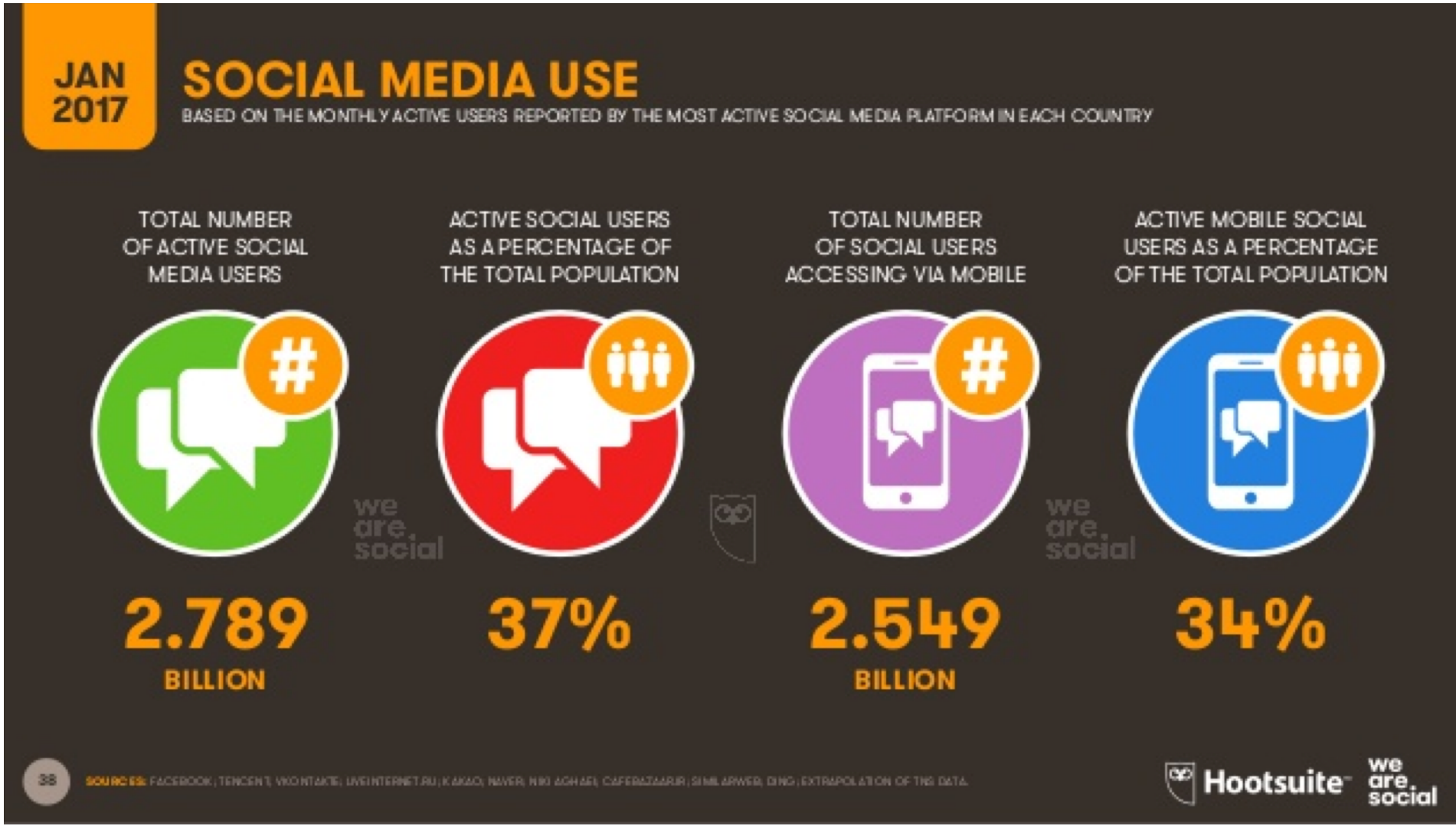
# Internet Trends



# Internet Trends



# Internet Trends



# Internet Trends

JAN 2018

## SOCIAL MEDIA USE

BASED ON THE MONTHLY ACTIVE USERS REPORTED BY THE MOST ACTIVE SOCIAL MEDIA PLATFORM IN EACH COUNTRY

TOTAL NUMBER OF ACTIVE SOCIAL MEDIA USERS

**3.196**
**BILLION**

ACTIVE SOCIAL USERS AS A PERCENTAGE OF THE TOTAL POPULATION

**42%**

TOTAL NUMBER OF SOCIAL USERS ACCESSING VIA MOBILE

**2.958**
**BILLION**

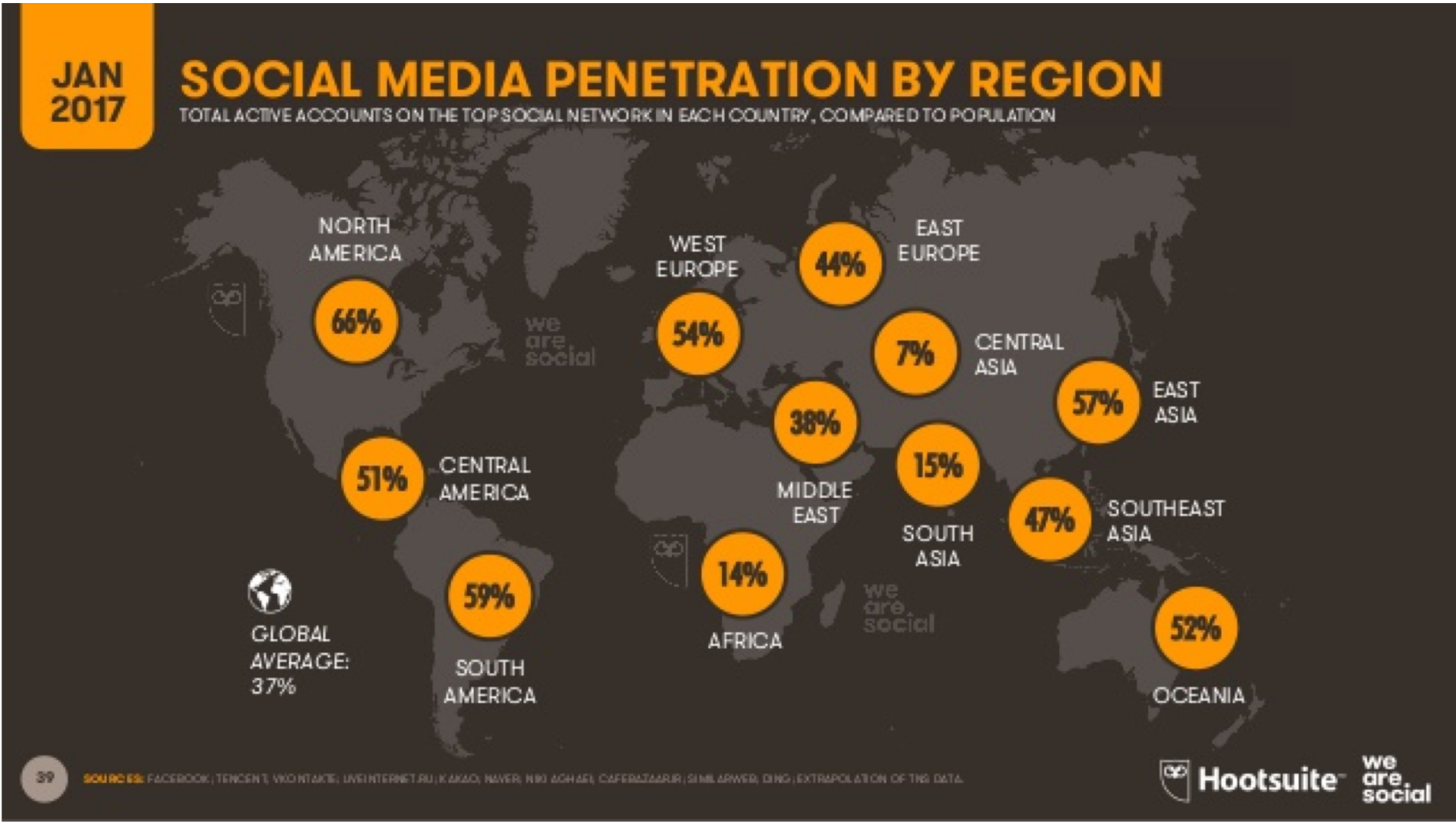
ACTIVE MOBILE SOCIAL USERS AS A PERCENTAGE OF THE TOTAL POPULATION

**39%**



**SOURCES:** FACEBOOK; TENCENT; VKONTAKTE; KAKAO; NAVER; DING; TECHRASA; SIMILARWEB; KEPIOS ANALYSIS.
**NOTE:** PENETRATION FIGURES ARE FOR TOTAL POPULATION, REGARDLESS OF AGE.

# Internet Trends

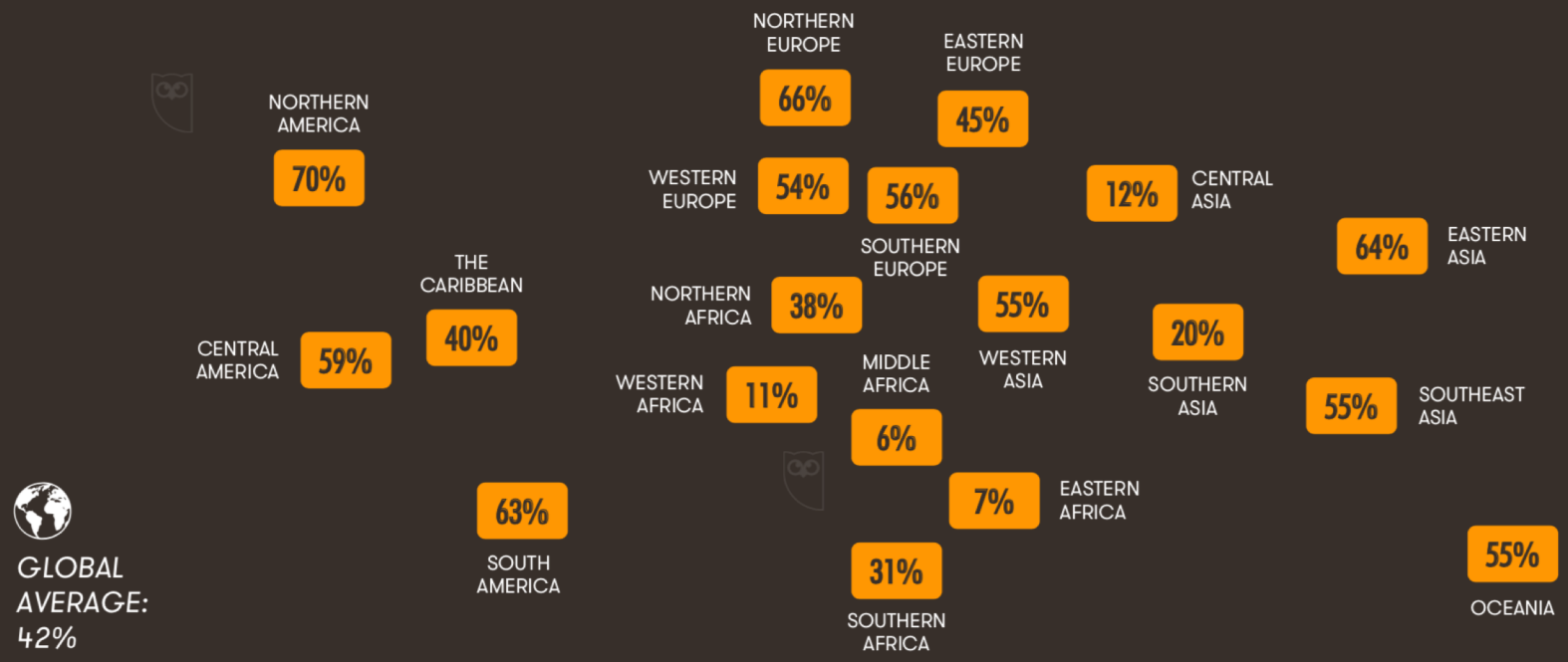


# Internet Trends

JAN 2018

## SOCIAL MEDIA PENETRATION BY REGION

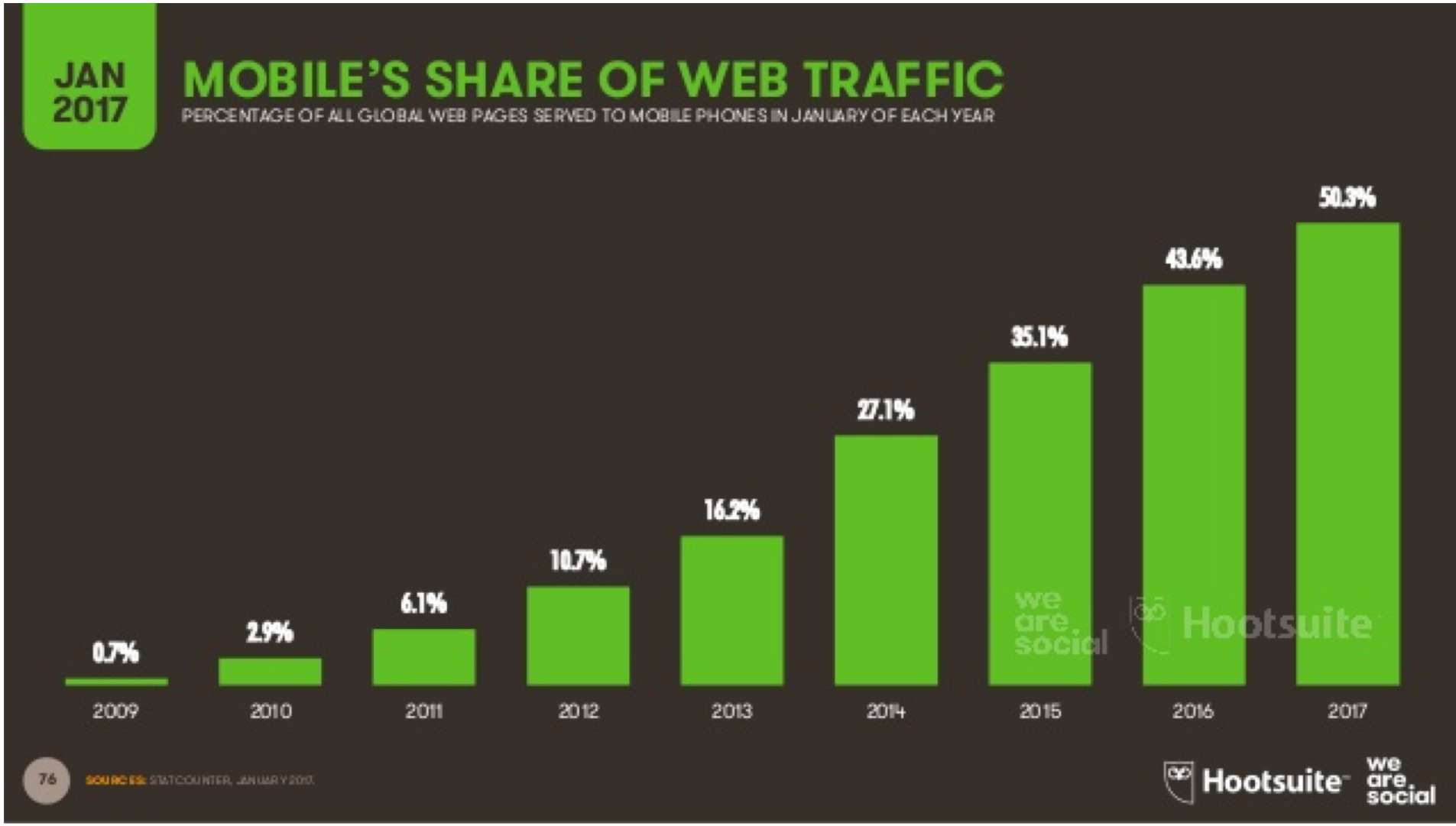
TOTAL ACTIVE ACCOUNTS ON THE MOST ACTIVE SOCIAL NETWORK IN EACH COUNTRY, COMPARED TO POPULATION






**SOURCES:** FACEBOOK; TENCENT; VKONTAKTE; KAKAO; NAVER; DING; TECHRASA; SIMILARWEB; KEPIOS ANALYSIS.
**NOTE:** PENETRATION FIGURES ARE FOR TOTAL POPULATION, REGARDLESS OF AGE.

# Internet Trends



# Internet Trends

JAN 2018

## SHARE OF WEB TRAFFIC BY DEVICE

BASED ON EACH DEVICE'S SHARE OF ALL WEB PAGES SERVED TO WEB BROWSERS

| LAPTOPS & DESKTOPS | MOBILE PHONES | TABLET DEVICES | OTHER DEVICES |
|---|---|---|---|
|  |  |  |  |
| **43%** | **52%** | **4%** | **0.14%** |
| YEAR-ON-YEAR CHANGE: | YEAR-ON-YEAR CHANGE: | YEAR-ON-YEAR CHANGE: | YEAR-ON-YEAR CHANGE: |
| **-3%** | **+4%** | **-13%** | **+17%** |



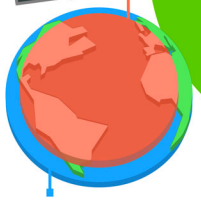
**SOURCE:** STATCOUNTER, JANUARY 2018 AND JANUARY 2017.

# The FOUR V's of Big Data

## Volume
SCALE OF DATA

**40 ZETTABYTES**
[ 43 TRILLION GIGABYTES ]
of data will be created by 2020, an increase of 300 times from 2005


**6 BILLION PEOPLE**
have cell phones



WORLD POPULATION: 7 BILLION

It's estimated that
**2.5 QUINTILLION BYTES**
[ 2.3 TRILLION GIGABYTES ]
of data are created each day

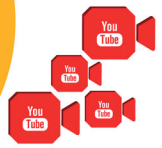
Most companies in the U.S. have at least
**100 TERABYTES**
[ 100,000 GIGABYTES ]
of data stored

## Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
**150 EXABYTES**
[ 161 BILLION GIGABYTES ]

By 2014, it's anticipated there will be
**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**
are watched on YouTube each month

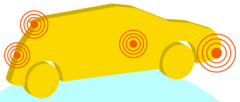



**400 MILLION TWEETS**
are sent per day by about 200 million monthly active users

## Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
**1 TB OF TRADE INFORMATION**
during each trading session

Modern cars have close
**100 SENSORS**
that monitor items such
fuel level and tire pressu



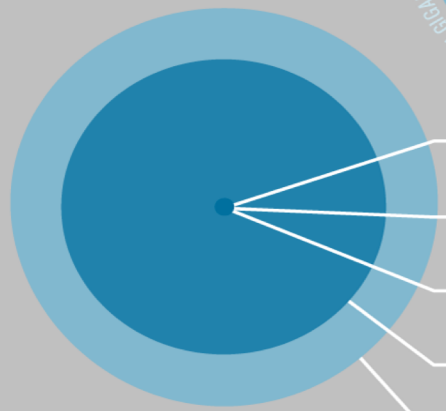
By 2016, it is projected there will be
**18.9 BILLION NETWORK CONNECTIONS**
– almost 2.5 connections per person on earth

## Veracity
UNCERTAINTY OF DATA

Poor data quality costs the US economy around
**$3.1 TRILLION A YEAR**

## CHARACTERISTICS (V'S) OF BIG DATA

GLOBAL INTERNET TRAFFIC IN 2013 WAS APPROXIMATELY 5,000,000,000,000,000 BYTES (5 BILLION GIGABYTES).




Global internet population GREW 14.3% BETWEEN 2011 & 2013

3 BILLION
The number of people who have access to the internet today equals that of the world's population in 1960


**Sources:** McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

# Spatial Data

- Geographic information is any item that is georeferenced
  - Atomic form

    *<location, time, property>*
  - Also called **geospatial** information
  - May be augmented with "quality" or goodness of the information

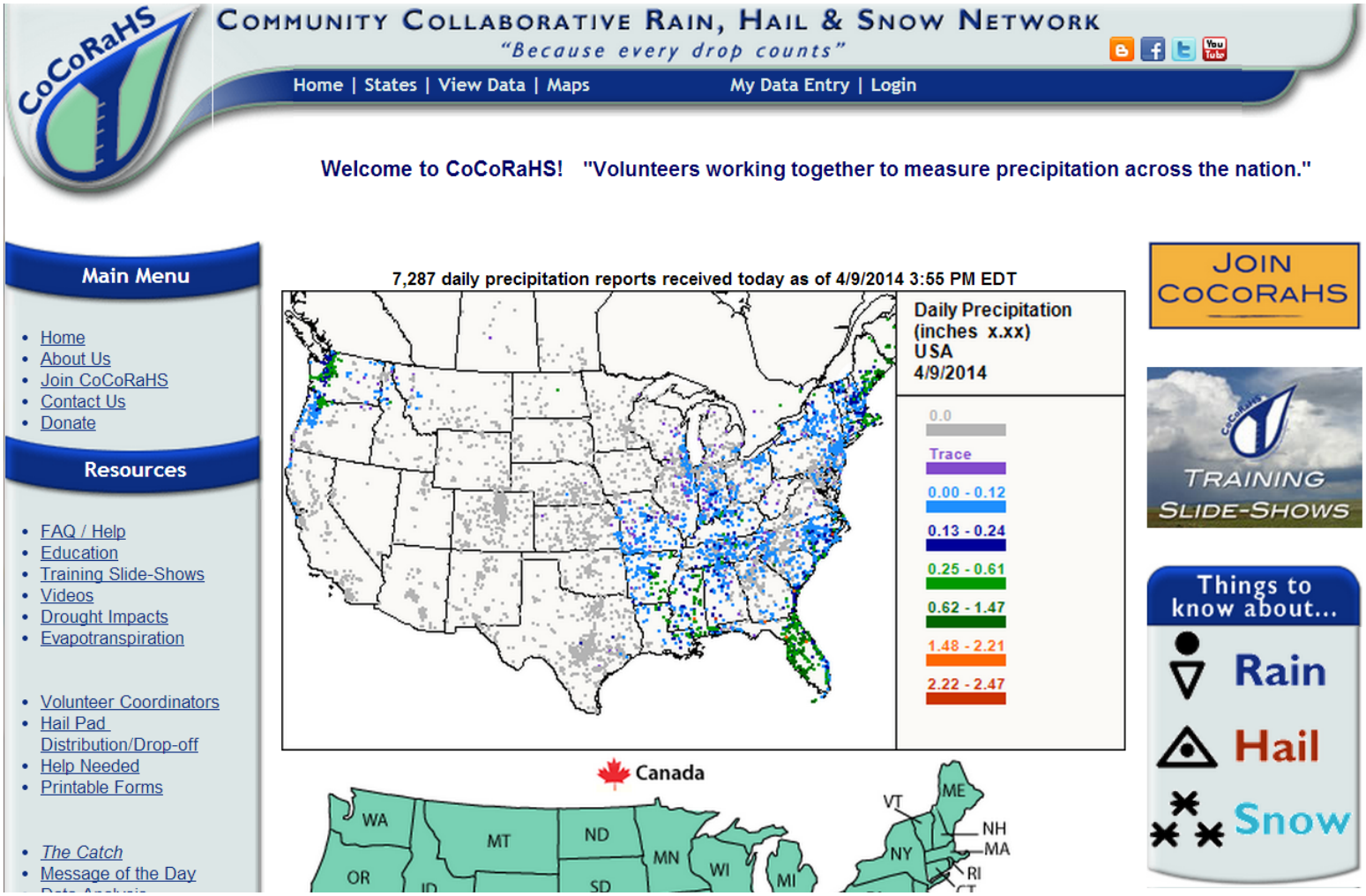
    *<location, time, property, goodness>*
  - May be further augmented with images, audio or video

- Geographic information typically
  - Created by government authorities
    - USGS, NGA, military in many countries, state and local governments
  - Disseminated to users
    - Generally with restrictions
    - At cost of production or reproduction?
    - Restrictions since 9/11
  - Top-down process: information bottlenecks for both collection and processing

# Volunteer Geoinformatics
# Citizen Science

- Networks of amateur observers
- Possibly trained, skilled
  - Christmas Bird Count
    - Thousands of volunteer participants
    - Protocols
  - Project GLOBE
    - An international network of school children
    - Reporting environmental conditions
    - Central integration and redistribution
  - Project BudBurst
    - Monitor Plant phenology
    - More than 2900 people already registered
    - More than 3900 species being monitored

# Volunteer Geoinformatics Example: www.cocorahs.org



# Volunteer Geoinformatics

- Why do people do this?
  - Self-promotion
    - Exhibitionism as information remains identified with source
  - Altruism
    - A belief that everything on the Web can be found and *will* be used to good effect
  - A desire to fill gaps in available data
    - Especially in areas where data are not available or where access is denied for security
  - Sharing with friends, relatives
    - But accessible by all
- Human Sensors
  - 7 billion "intelligent" sensors
  - Informed and capable observers
    - With rich local knowledge
    - With individual processing and interpretations
    - Uplink technology
      - Broadband Internet
      - Mobile phone
    - Information capture technology
      - Webcam
      - Mobile phone with camera/video capability

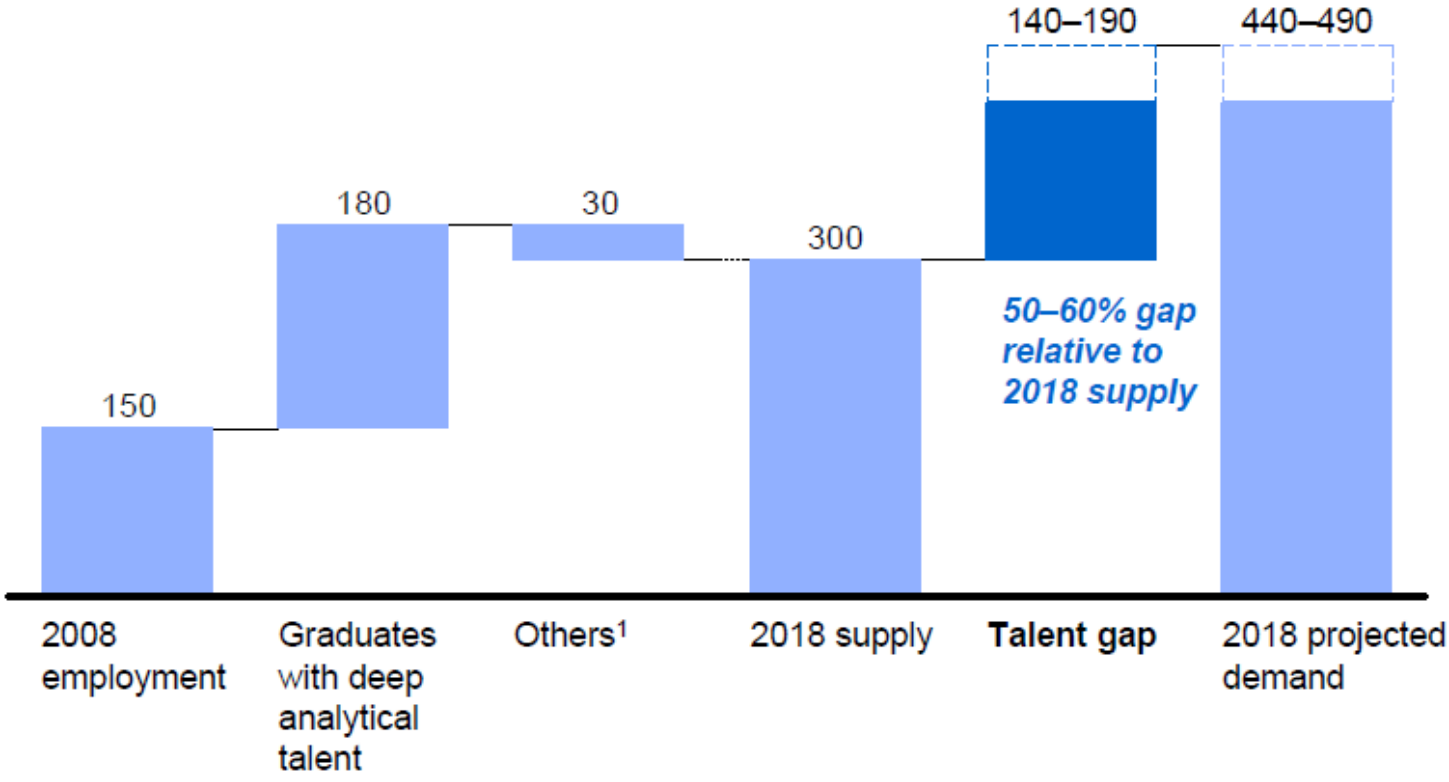
# Looking Ahead

- 163 Zettabytes of data generated per year by 2025 (IDC)
- Revenues for big data and business analytics (BDA) will grow from $130B billion in 2016 to $203B in 2020 (IDC)

# Demand for Data Mining

**Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018**

Supply and demand of deep analytical talent by 2018
Thousand people




1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

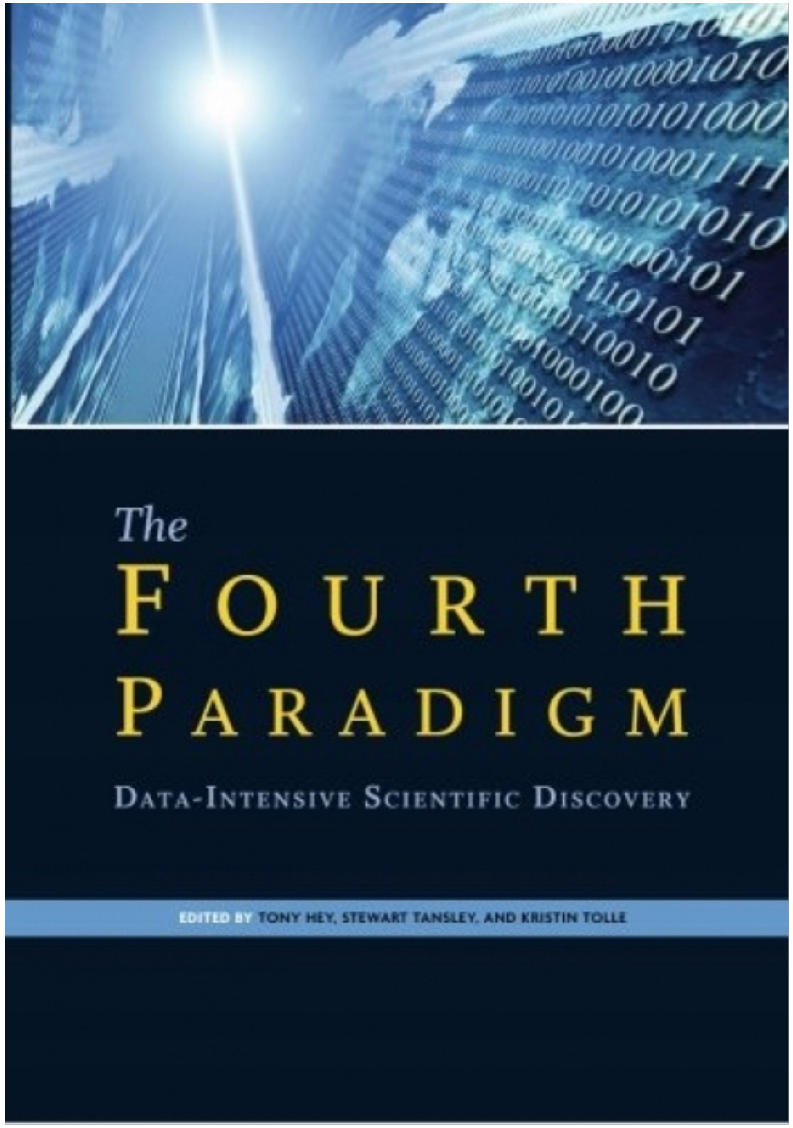
# The world's most valuable resource is no longer oil, but data

*The data economy demands a new approach to antitrust rules*


Print edition | Leaders ›
May 6th 2017

# Evolution of Sciences

- Before 1600: Empirical science
  - Gaining knowledge by observation
  - They are sometimes experimental
- 1600-1950s: Theoretical science
  - Each discipline grew a *theoretical* component.
  - Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s: Computational science
  - In this period, most disciplines grew a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - It traditionally meant simulation.
  - It grew out of our inability to find closed-form solutions for complex mathematical models.



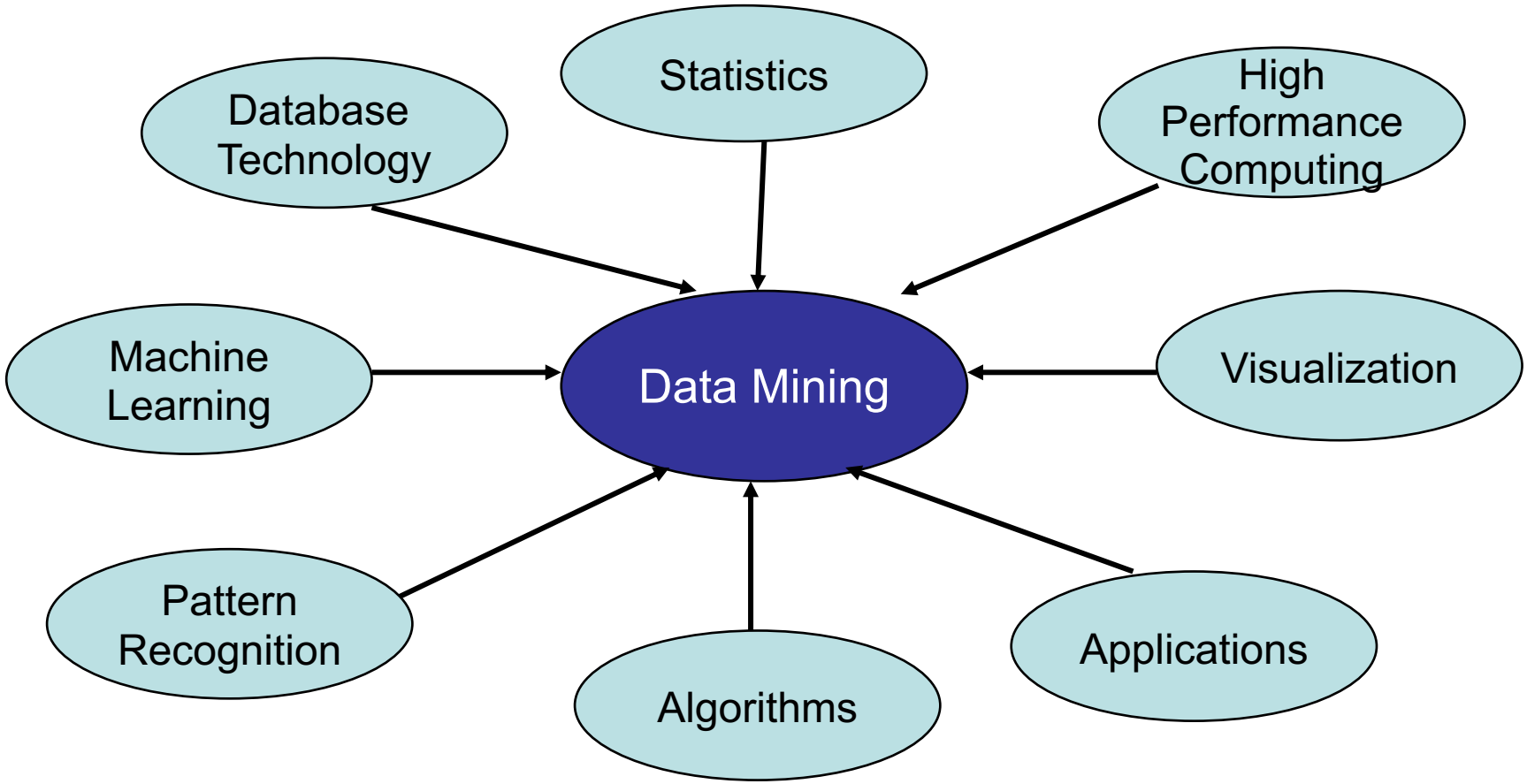

Unify experimental, theoretical and simulation approaches!

# Evolution of Sciences

- 1990-now: Data science
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes.
  - X-info and Comp-X (e.g. bioinformatics, computational ecology)
  - Data exploration is the major new challenge.

# What is Data Mining?

# Confluence of Multiple Disciplines




Data mining overlaps with:

- Databases: Large-scale data, simple queries
- Machine learning: Small data, Complex models
- CS Theory: (Randomized) Algorithms

# Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

# What is Data Mining?

- Multiple definitions

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large datasets

- Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- Alternative names
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, information harvesting, business intelligence, etc.

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

# What is data mining?

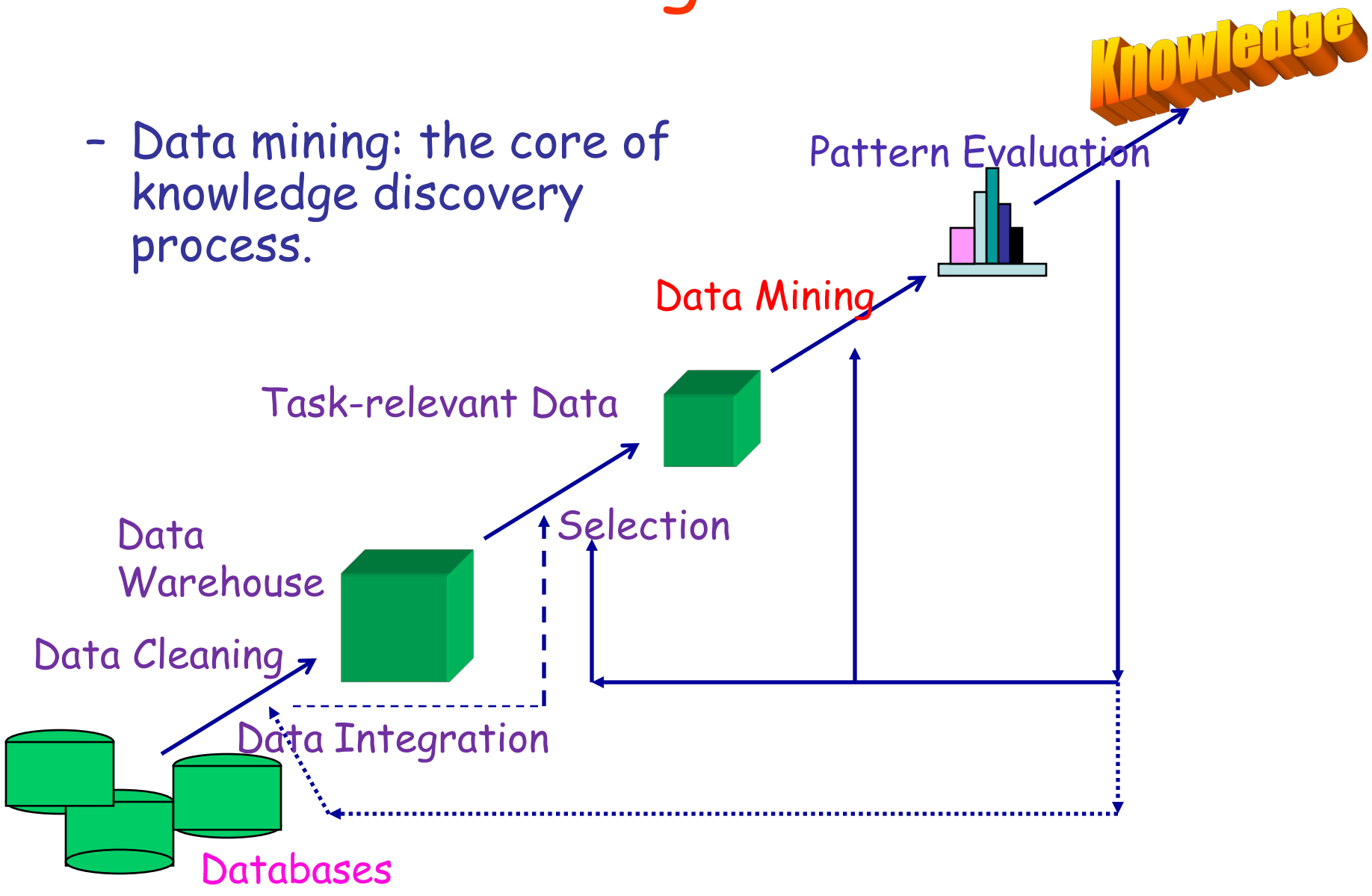
- Novel: previously unknown, not obvious
- Valid: broadly applicable (on new data) with some certainty
- Meaningful: humans should be able to understand
- Useful: should be possible to act on the result (actionable)

# What is (not) mining?

- What is NOT data mining?
  - Look up phone number in a phone directory
  - Query a web search engine for information about "Amazon"
- What is data mining?
  - Find certain names that are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Predict if a customer will consume over $100 in a store

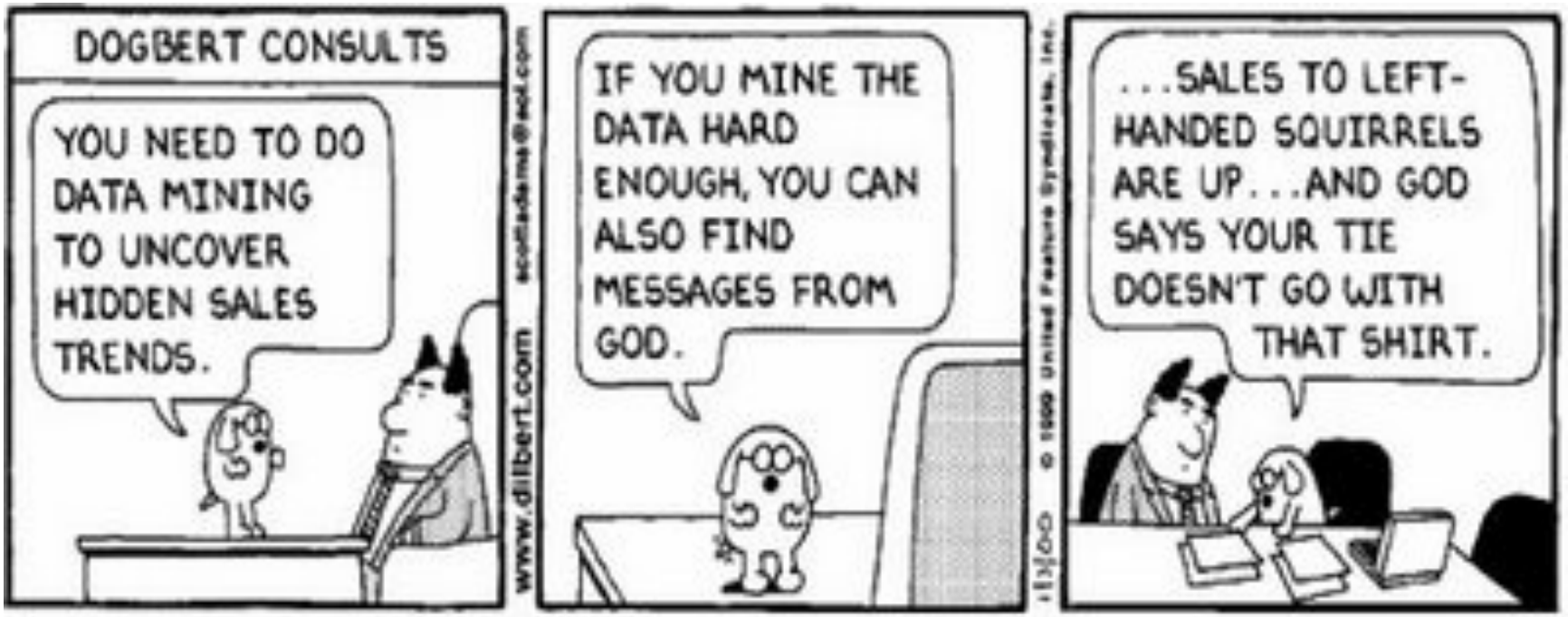
# Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



# Are All "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- Interestingness measures
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- Objective vs. subjective interestingness measures
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.


DOGBERT CONSULTS
YOU NEED TO DO DATA MINING TO UNCOVER HIDDEN SALES TRENDS.
www.dilbert.com
IF YOU MINE THE DATA HARD ENOUGH, YOU CAN ALSO FIND MESSAGES FROM GOD.
...SALES TO LEFT-HANDED SQUIRRELS ARE UP...AND GOD SAYS YOUR TIE DOESN'T GO WITH THAT SHIRT.

# Meaningful Patterns

- A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless

- Statisticians call it Bonferroni's principle:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find meaningless patterns

# Meaningful Patterns

- Find (unrelated) people who have stayed at the same hotel on the same day at least twice
- $10^9$ people being tracked
- 1,000 days
- Each person stays in a hotel 1% of time (1 day out of 100)
- Hotels hold 100 people (so $10^5$ hotels)
- If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?

# Meaningful Patterns

- Expected number of "suspicious" pairs of people:

$$250{,}000$$

- Too many combinations to check
- We need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

# Data Mining Tasks

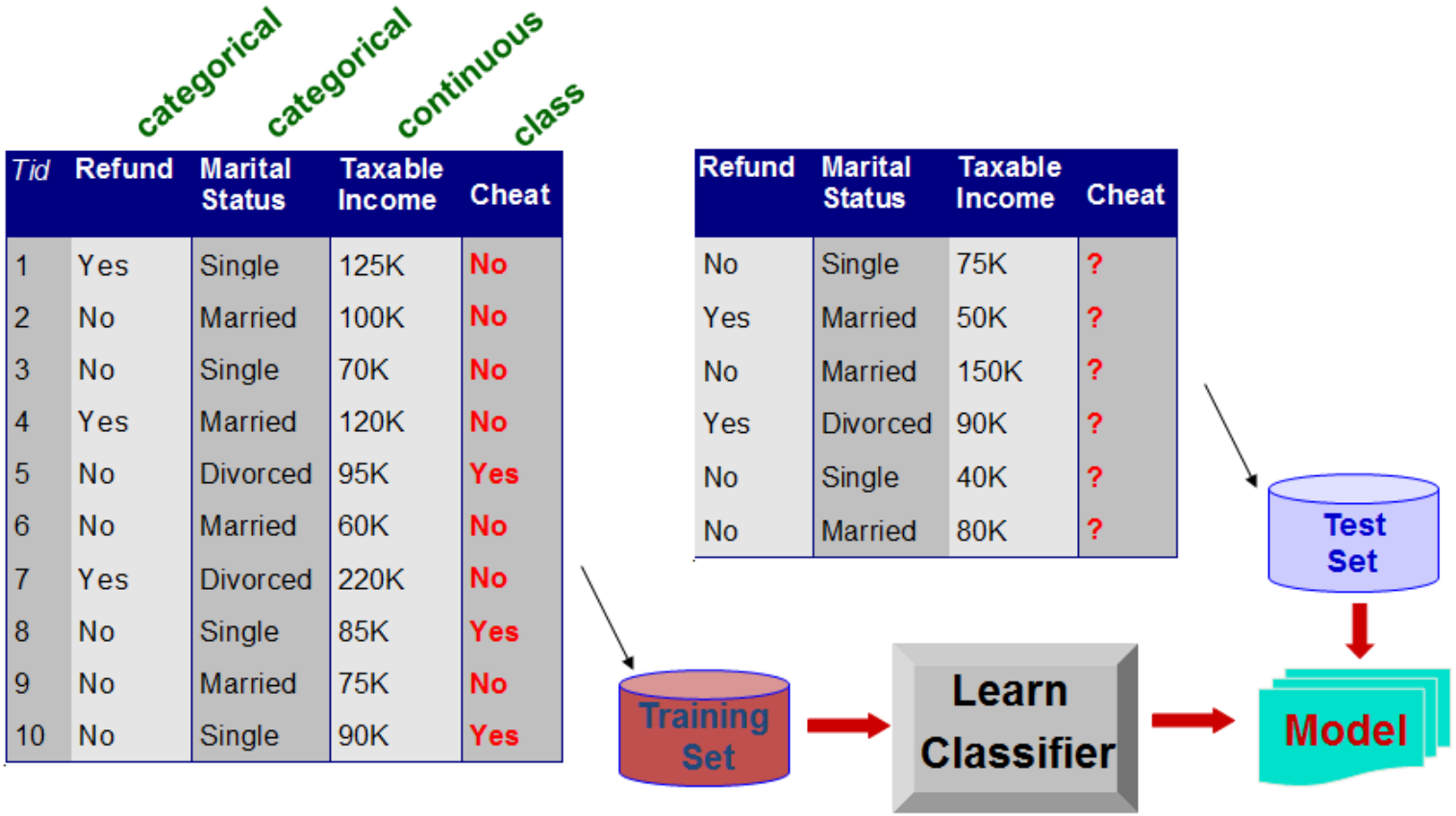
- Descriptive methods
  - Find human-interpretable patterns that describe the data
    - Example: Clustering
- Predictive methods
  - Use some variables to predict unknown or future values of other variables
    - Example: Recommender systems

# Data Mining Tasks

- Classification
- Clustering
- Association Rule Discovery
- Deviation Detection

# Classification

- Given a collection of records, find a model for class attribute as a function of the values of other attributes, so that previously unseen records can be assigned a class as accurately as possible.

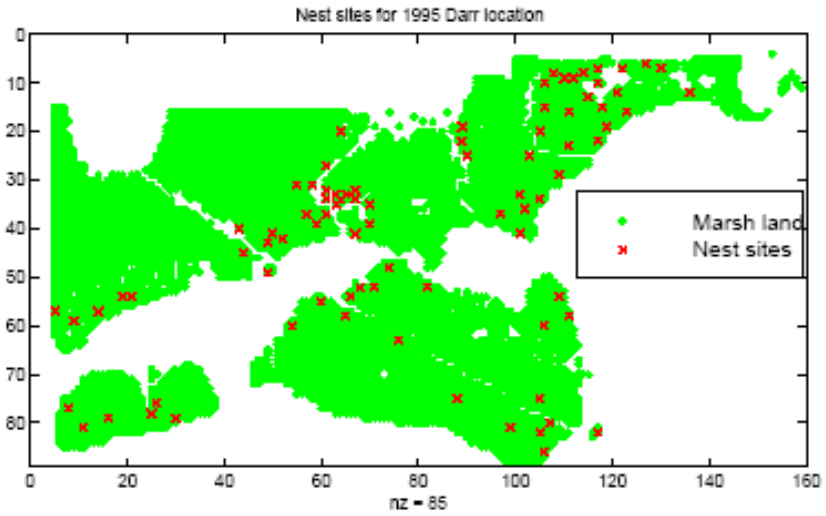


# Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

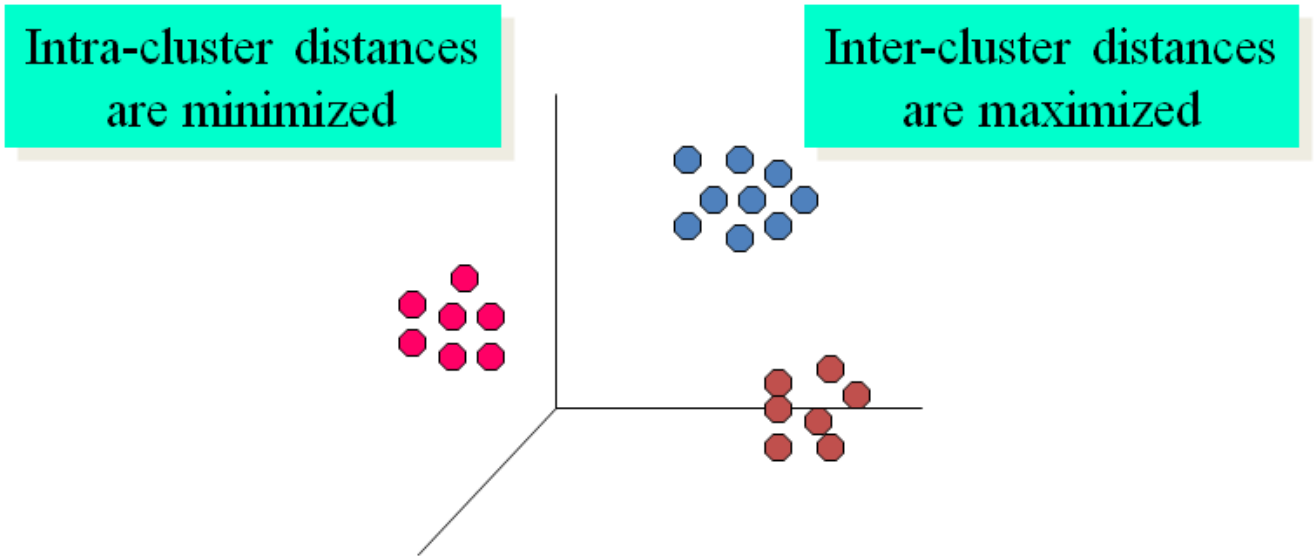
# Spatial Predictive Models

- Location Prediction: Bird Habitat Prediction
  - Given training data
  - Predictive model building
  - Predict new data



# Clustering

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another



# Clustering

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

# Clustering

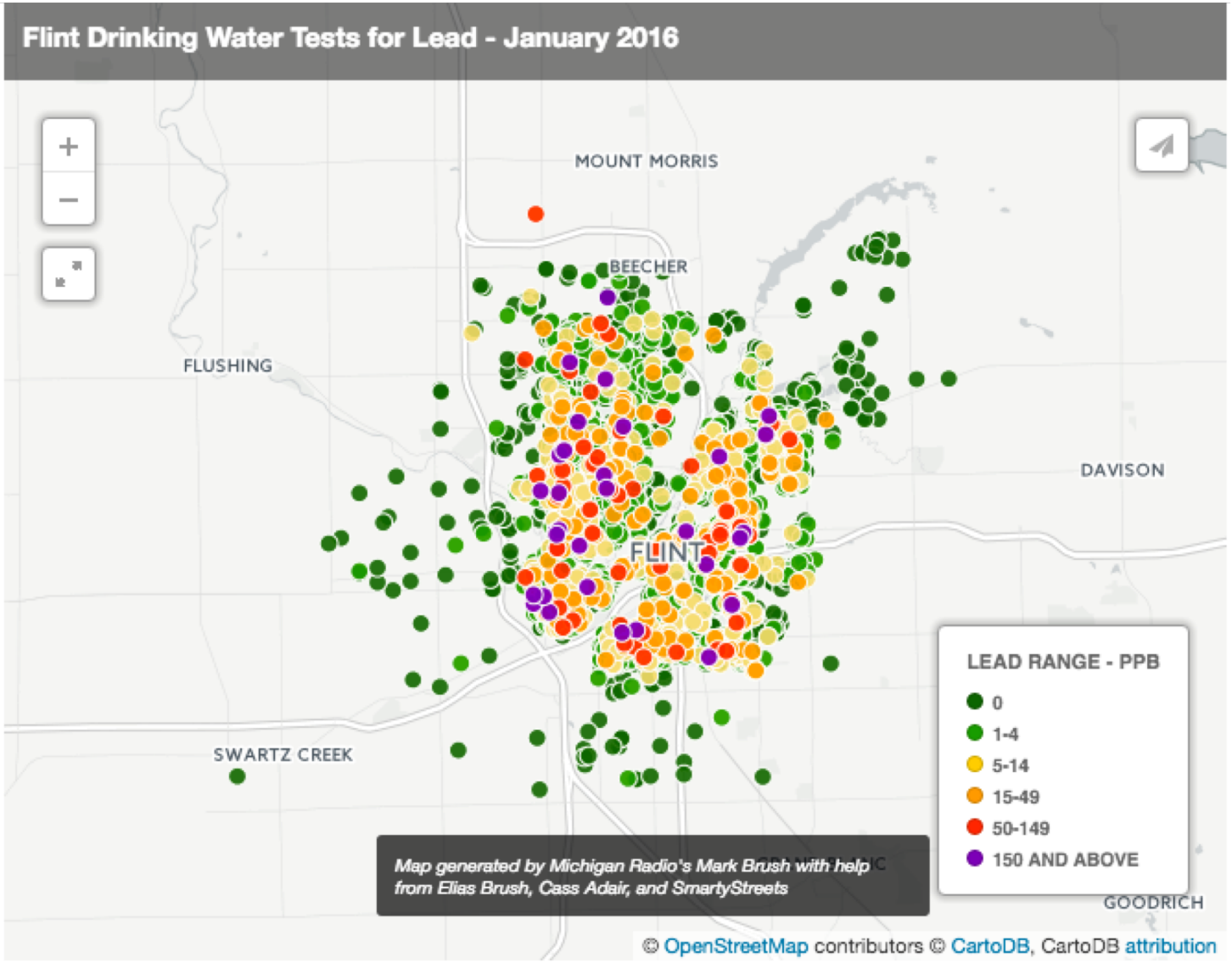
- Application: document clustering

| Category | Total Articles | Correctly Placed |
|---|---|---|
| Financial | 555 | 364 |
| Foreign | 341 | 260 |
| National | 273 | 36 |
| Metro | 943 | 746 |
| Sports | 738 | 573 |
| Entertainment | 354 | 278 |

# Spatial Clustering

- The 1854 Asiatic Cholera in London


# Spatial Clustering



# Association Rule Discovery

- Given a set of records, each of which contain some number of items from a given collection, produce dependency rules which will predict occurrence of an item based on occurrences of other items.

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

Rules Discovered:

**{Milk} --> {Coke}**

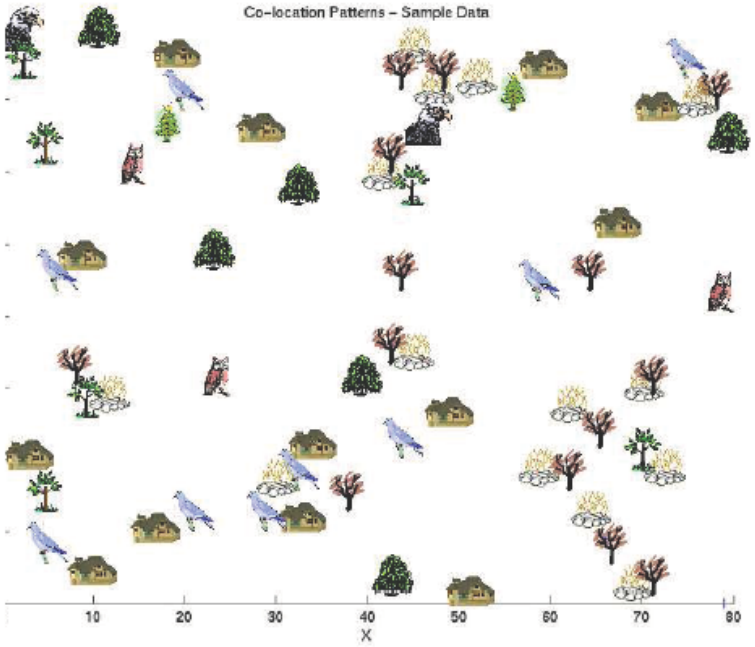
**{Diaper, Milk} --> {Beer}**

# Association Rule Discovery

- Applications: marketing and sales promotion (cross-selling)

# Spatial Co-location Patterns

- Given:
  - A collection of different types of spatial events
- Find: Co-located subsets of event types



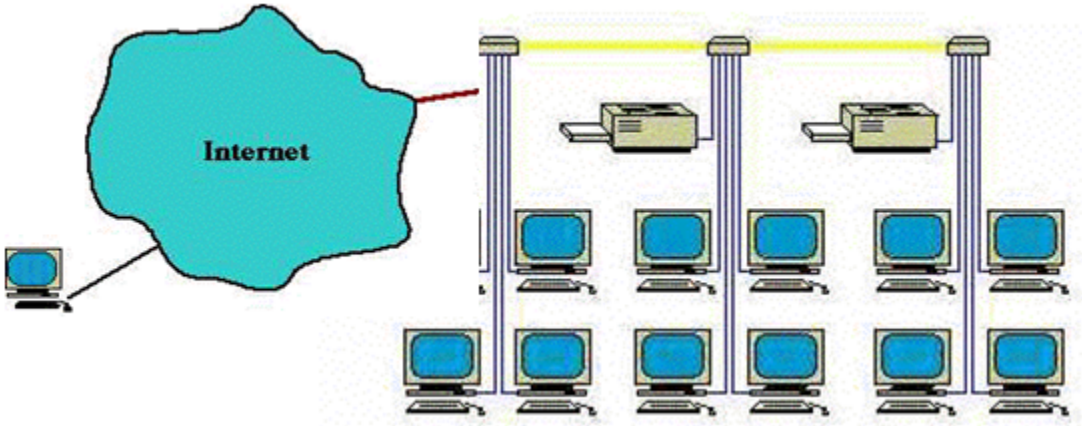

Answers:   and  

# Anomaly Detection

- Detect *significant deviations from normal behaviors*

Credit card fraud detection


Network intrusion detection

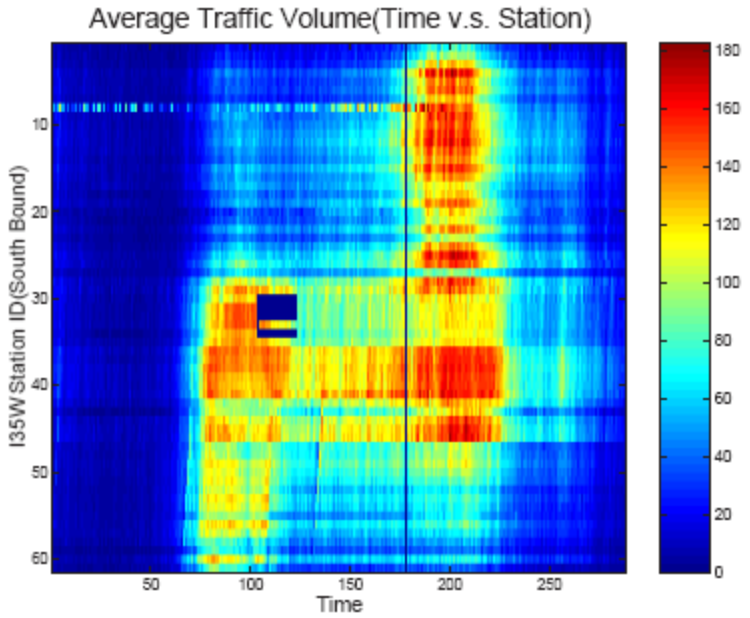


# Anomaly Detection

- Outlier: A data object that does not comply with the general behavior of the data
- Noise or exception? — One person's garbage could be another person's treasure
- Methods: By product of clustering or regression analysis, …
- Useful in fraud detection, rare events analysis

# Example Spatial Pattern: Spatial Outliers

- Spatial Outliers
  - Traffic Data in Twin Cities
  - Abnormal Sensor Detections
  - Spatial and Temporal Outliers




OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
www.dilbert.com
scottadams@aol.com
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...
© 1999 United Feature Syndicate, Inc.

# Data Mining and Privacy

## Privacy Properties of Telephone Metadata

"You have my telephone number,
connecting with your telephone number.

There are no names… in that database."

-President Obama

# Data Mining and Privacy

## Re-Identification

| Lookup Source | % Matched |
|---|---|
| Google Places | 16.6 |
| Yelp | 10.5 |
| Facebook | 13.7 |
| All Automated Sources | 31.9 |

Automated approaches

| Lookup Source | % Matched |
|---|---|
| Intelius | 65 |
| Google Search | 58 |
| All Automated Sources | 26 |
| All Sources | 82 |

Manual and combined approaches.

# Data Mining and Privacy

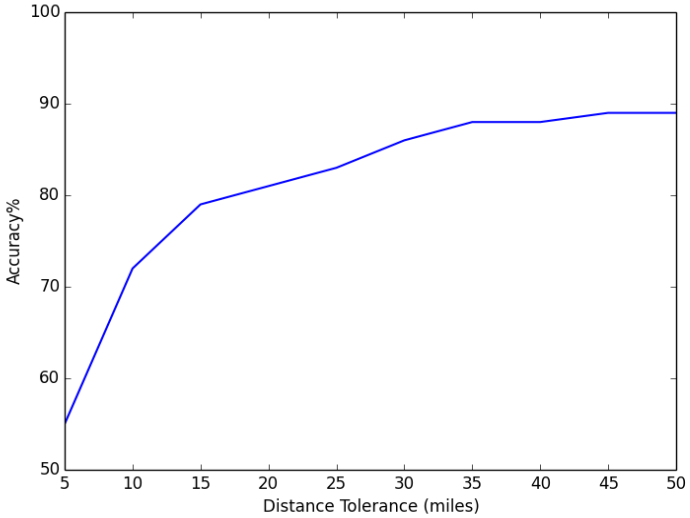
"All it is, is the number pairs, when those calls took place, how long they took place.

So that database is sitting there."

-President Obama

# Data Mining and Privacy

## Home Location Inference




Methodology: re-identify businesses, cluster their locations

## Religion Inference

≈ ¾ accuracy

(naïve heuristic on a small sample)

Methodology: comparison to Facebook data

# Data Mining and Privacy

## Sensitive Trait Inference

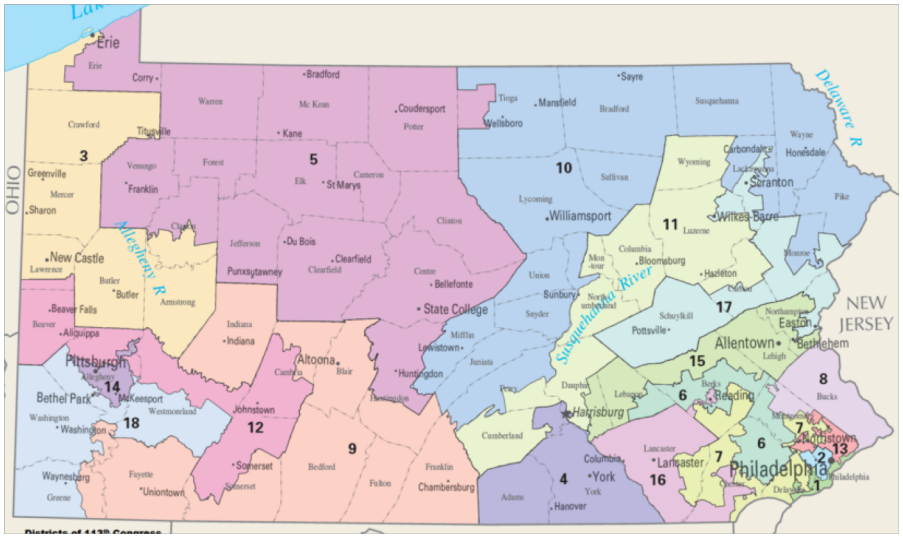
- Relapsing-Remitting Multiple Sclerosis (?)
- Cardiac Arrhythmia (✓)
- Owning an Assault Rifle (✓)
- Building a Grow House (?)
- Seeking an Abortion (?)

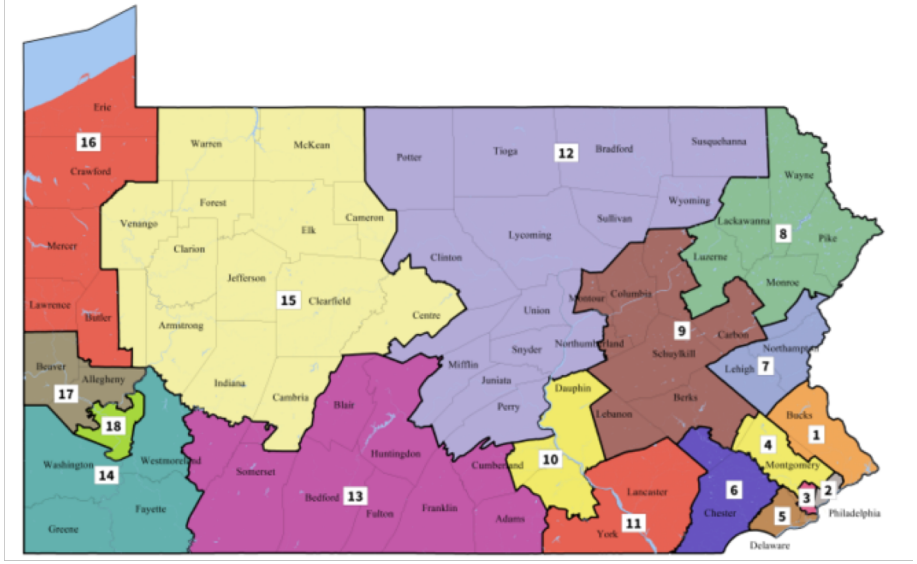
Methodology: automated and manual number re-identification

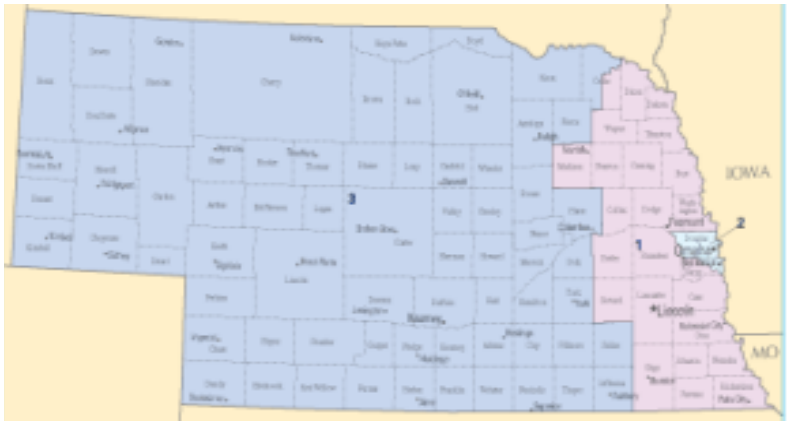
Idea: Intelligence law and policy should be informed by science, not lawyerly intuition

# Data Mining – Spatial Clustering

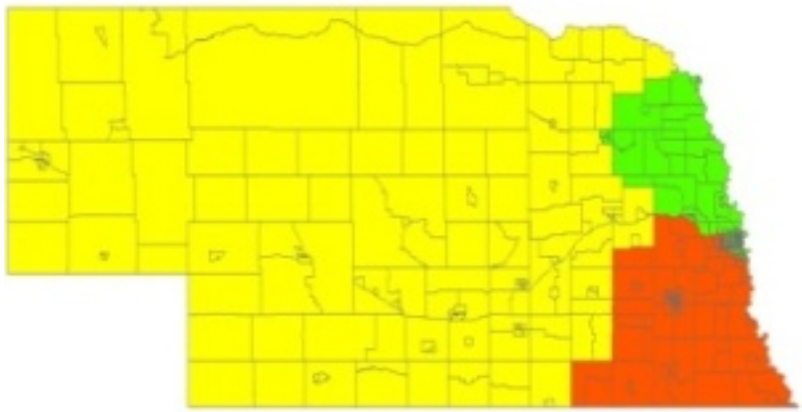
## Gerrymandering



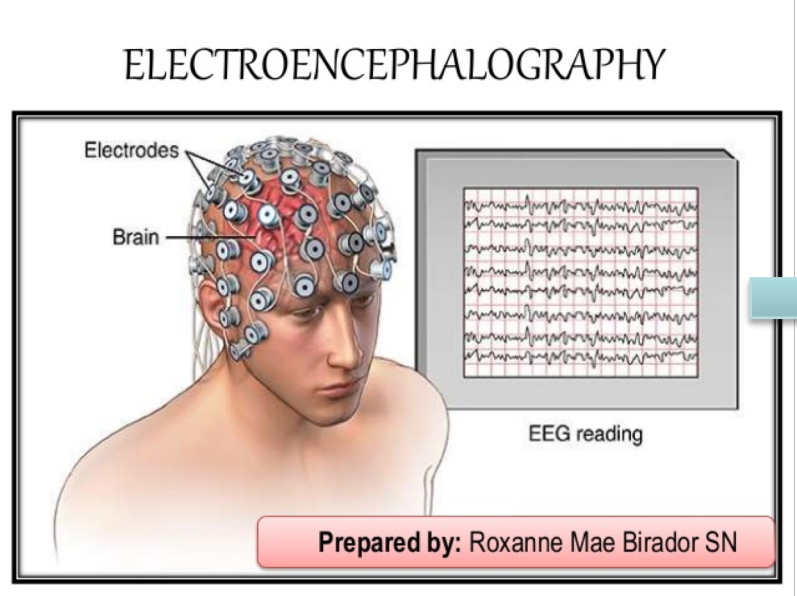

NE Congressional Districts



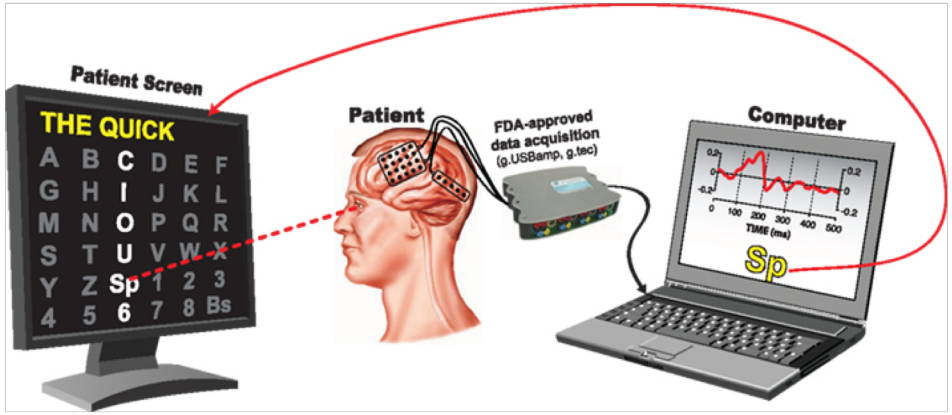
Results of CPSC

Joshi, Samal, Soh, ICDM 2009

# Classification of EEG Data




OR



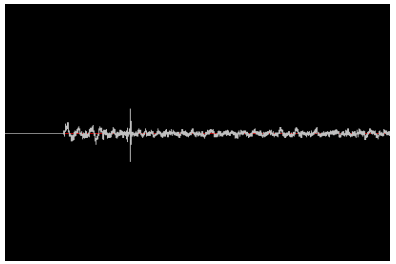

Brain Computer Interfaces

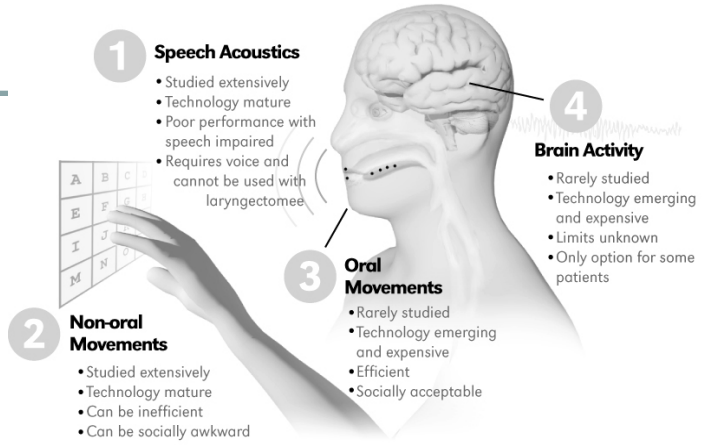
# Addressing Speech Impairments

## Using Oral Movement and Brain Activity for Assessment and Treatments of Speech Motor Impairments

Articulatory Movements → Words / Sentences → Speech Acoustics

Hello!
How are you?





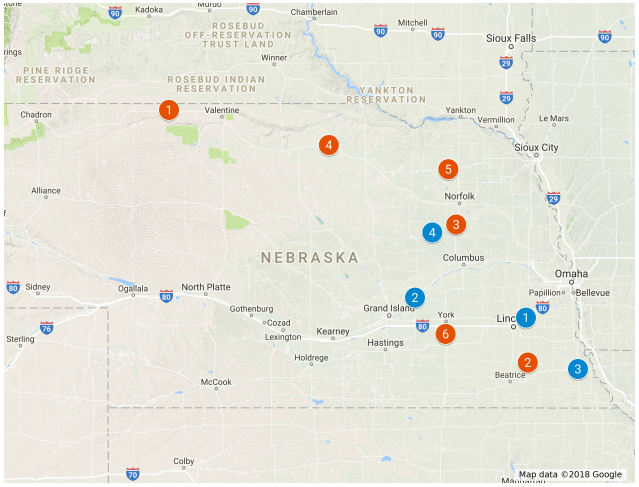
# Plant Phenotyping

UNL Plant Vision Group

# VGI: Water Quality

## 2017-2018

Auburn High School - 20
Newman Grove High School - 19
Waverly High School - 21
Central City High School - 18

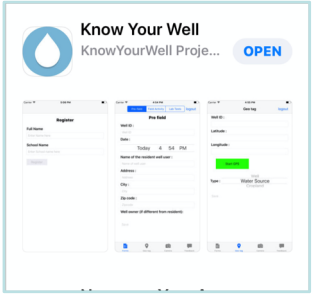



## 2018-2019

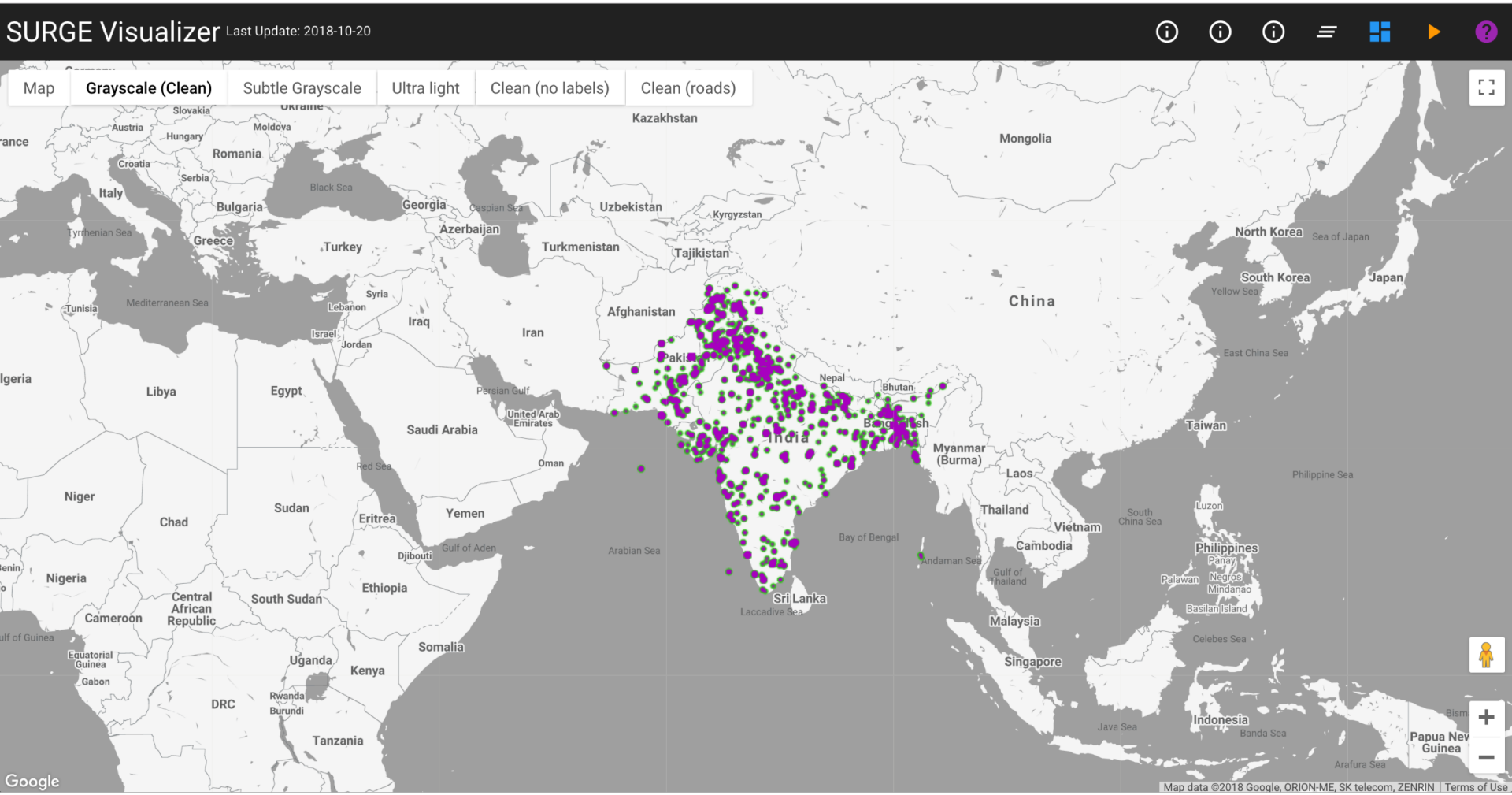
Cody-Kilgore High School
Freeman High School
Madison High School
Stuart High School
Bassett High School
McCool Junction High School

Conductivity
Magnesium
Orthophosphate-P
Ammonium-N
Calcium
Chloride
Sulfate

pH
Arsenic
Uranium
Nitrate
Alachlor
Atrazine



# SURGE


SURGE Visualizer
Last Update: 2018-10-20
Map
Grayscale (Clean)
Subtle Grayscale
Ultra light
Clean (no labels)
Clean (roads)
Kazakhstan
Mongolia
China
North Korea
South Korea
Japan
Sea of Japan
Yellow Sea
East China Sea
Philippine Sea
Taiwan
Uzbekistan
Kyrgyzstan
Tajikistan
Turkmenistan
Afghanistan
Iran
Iraq
Syria
Lebanon
Israel
Jordan
Saudi Arabia
Yemen
Oman
United Arab Emirates
Persian Gulf
Red Sea
Gulf of Aden
Arabian Sea
Bay of Bengal
Andaman Sea
Laccadive Sea
Sri Lanka
India
Nepal
Bhutan
Myanmar (Burma)
Laos
Thailand
Vietnam
Cambodia
Gulf of Thailand
South China Sea
Philippines
Luzon
Palawan
Mindanao
Basilan Island
Malaysia
Singapore
Celebes Sea
Java Sea
Indonesia
Banda Sea
Arafura Sea
Papua New Guinea
Georgia
Azerbaijan
Caspian Sea
Black Sea
Turkey
Greece
Bulgaria
Romania
Moldova
Ukraine
Slovakia
Hungary
Austria
Croatia
Serbia
Italy
Tyrrhenian Sea
Mediterranean Sea
Tunisia
Libya
Egypt
Algeria
Niger
Chad
Sudan
Eritrea
Djibouti
Ethiopia
Nigeria
Cameroon
Central African Republic
South Sudan
Somalia
Uganda
Kenya
DRC
Rwanda
Burundi
Tanzania
Equatorial Guinea
Gabon
Google
Map data ©2018 Google, ORION-ME, SK telecom, ZENRIN
Terms of Use