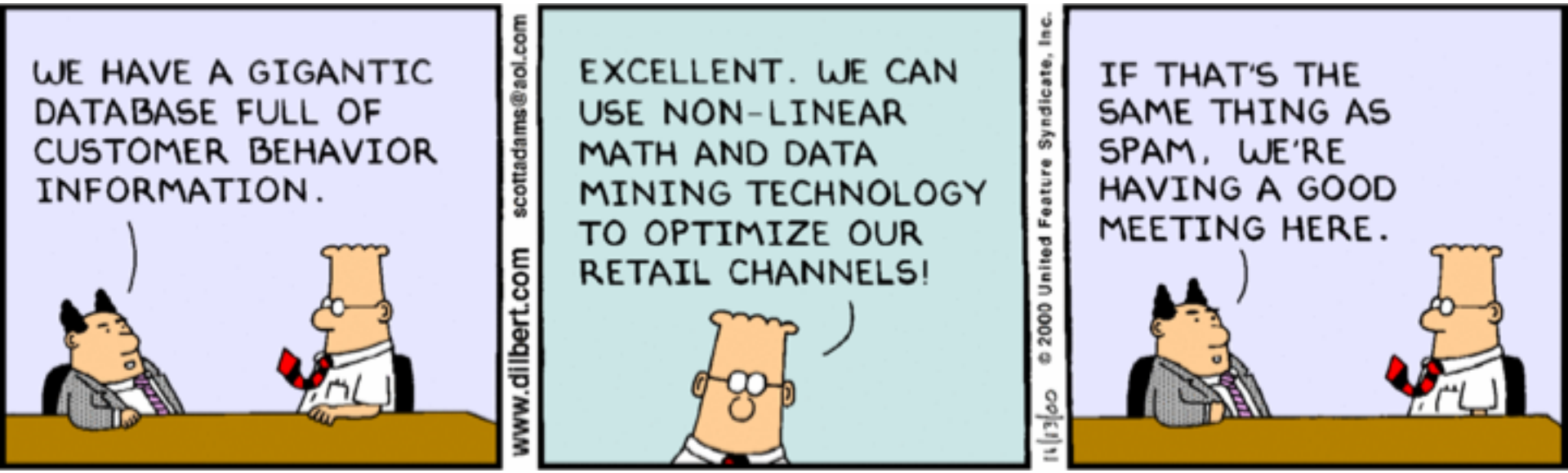

WE HAVE A GIGANTIC DATABASE FULL OF CUSTOMER BEHAVIOR INFORMATION.
www.dilbert.com scottadams@aol.com
EXCELLENT. WE CAN USE NON-LINEAR MATH AND DATA MINING TECHNOLOGY TO OPTIMIZE OUR RETAIL CHANNELS!
© 2000 United Feature Syndicate, Inc.
IF THAT'S THE SAME THING AS SPAM, WE'RE HAVING A GOOD MEETING HERE.

November 2011

New World!

- Largest retailer does not have an inventory
  - Alibaba
- Largest hotel chain does not own a hotel
  - Airbnb
- Largest media company does not generate any original content
  - Facebook
- Largest taxi company does not own a single car
  - Uber

2008: 38 B 2004: 541B 2009: 55B 1999: 352B

1892: 230 B 1903: 36 B 1902: 96 B 1955: 164 B

Data in Business

- Completely data-driven: Organizations that compete based on transforming information into monetizable assets
  - Airbnb, eBay, Facebook
- Data-infused: Businesses that manage and sell products/services but use information to drive marketing, sales and business process optimization
  - Amazon, Netflix, Capital One
- Data-informed: More conventional businesses that are trying to adapt new information and data management technologies to fit their existing business models to improve their overall competitiveness
  - John Deere, General Electric.

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?

A Case Study: Alibaba

Based on a Bloomberg's report
October 29, 2018

Courtsey: Leen-Kiat Soh

Alibaba

Alibaba = Google + Netflix + Amazon?

- Operates the world's biggest e-commerce platform with 600 million monthly active users
- Operates China's biggest online ads business
- Also controls the Chinese versions of YouTube and Netflix (Youku)
- Also controls a supermarket chain and department store franchise
- Hosts a financial transactions platform called Alipay
  - Alipay's dominance in mobile payment systems + Alibaba's retail-management software (Ling Shou Tong)
  - Track consumer behavior offline in brick-and-mortar retail locations and ferret out insights

Courtsey: Leen-Kiat Soh

Alibaba

"Nobody else has this ecosystem where one player has all the pieces together and can put together a single profile of you. Alibaba has the ability to use this to get their seller base to create their product, which is a holy grail in e-commerce."

- E-commerce industry expert Ken Leaver

Courtsey: Leen-Kiat Soh

Alibaba Successes

- Helped Mars Inc create a candy bar
- Gave Unilever NV valuable data for a new line of pollution-fighting cosmetics
- Then advised both companies how to market the products

Courtsey: Leen-Kiat Soh

Unilever

- Earlier in 2018, Alibaba data researchers noticed growing demand from urbanites for pollution-fighting, "deep-cleansing" personal care products

- Some premium brands already sold cleansers and shampoos designed to strip off pollutants, but there weren't many mid-priced options

Courtsey: Leen-Kiat Soh

Unilever

- Product Development
  - Unilever acted on this insight and came up with a line of affordable anti-pollution products, starting with a skin cleanser
  - Developed 48 different prototypes of the cleaners at different price ranges

- Customer Market Testing
  - Prototypes were shown to users, such as young mothers, on Alibaba's online malls Taobao and Tmall
  - When someone tried to buy a prototype, a pop-up message informed them that they were participating in a consumer testing exercise and offered them a voucher for taking part

Courtsey: Leen-Kiat Soh

Unilever

- New Line Launch
  - In September 2018, Unilever launched the Purifi line, starting with a skin cleanser based on the purchasing decisions of tens of thousands of those young mothers

- Faster Process
  - The entire process of conception, design and testing took Unilever just 6 months with Alibaba's help, down from the usual 18 months to two years for a new product

Courtsey: Leen-Kiat Soh

"Alibaba gives us a real environment to test new products. Because consumers have no idea that they are taking part in a survey or study, their reactions and purchasing decisions are real. It makes the feedback real, which is a huge advantage in an industry where product innovation is essential, but costly and risky."

– Susan Ren, Director of data and digital development, Unilever

Courtsey: Leen-Kiat Soh

Mars

- The same people who buy a lot of chocolate also like spicy snacks
- That prompted the creation of the Spicy Snickers candy bar
  - Incorporates the Sichuan peppercorn, the source of China's famous "mala" (numb and spicy) taste
- Faster Process
  - Typically Mars spends two to three years developing a new product; the Spicy Snickers came together in less than one
  - Alibaba's cross-platform harvesting of data reduces guesswork in marketing

Courtsey: Leen-Kiat Soh

"The age-old tradition is that 90% of innovation fails. This helps us bring the rate of failure down."

– Ian Burton, China president of Mars Wrigley Confectionary

Courtsey: Leen-Kiat Soh

"I believe that within 12 months we will be able to see not just by consumer, but by store type and location, what is the perfect product mix for any one store to stock. This is not a level of consumer insight we can get anywhere else because it does not exist anywhere else."

– Ian Burton, China president of Mars Wrigley Confectionary

Courtsey: Leen-Kiat Soh

Potential Issues

- Alibaba is able to collect user data with relative impunity because privacy is less of an issue in China than elsewhere

- While the data is anonymous, users can't opt out if they want to use the company's platforms and agree to terms and conditions
  - much the way people using Facebook or Amazon do

- Still, Chinese consumers are starting to wake up to – and even resent – Alibaba's omnipresence

Courtsey: Leen-Kiat Soh

Potential Issues

- Alibaba's dominance is also giving some consumer products companies pause

- Associated Press reported complaints from five major brands that Alibaba had made it harder to find their online storefronts after they refused to sign exclusive partnerships
  - Unilever and Mars both say Alibaba hasn't insisted that they cease partnerships with rival JD.com
  - A spokesperson for Alibaba says the company gives brands "full autonomy" to choose their distribution platform

Courtsey: Leen-Kiat Soh

Potential Issues

- All the consumer insights in the world do *not* guarantee a blockbuster
- Consumer products that have changed the world are often the result of intuition

Courtsey: Leen-Kiat Soh

"No one told Steve Jobs they need an iPhone. Consumers can only tell you their problems and needs, but you still need creativity."

– Pedro Yip, a retail and consumer goods partner at consultancy Oliver Wyman

Courtsey: Leen-Kiat Soh

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

Looking Ahead

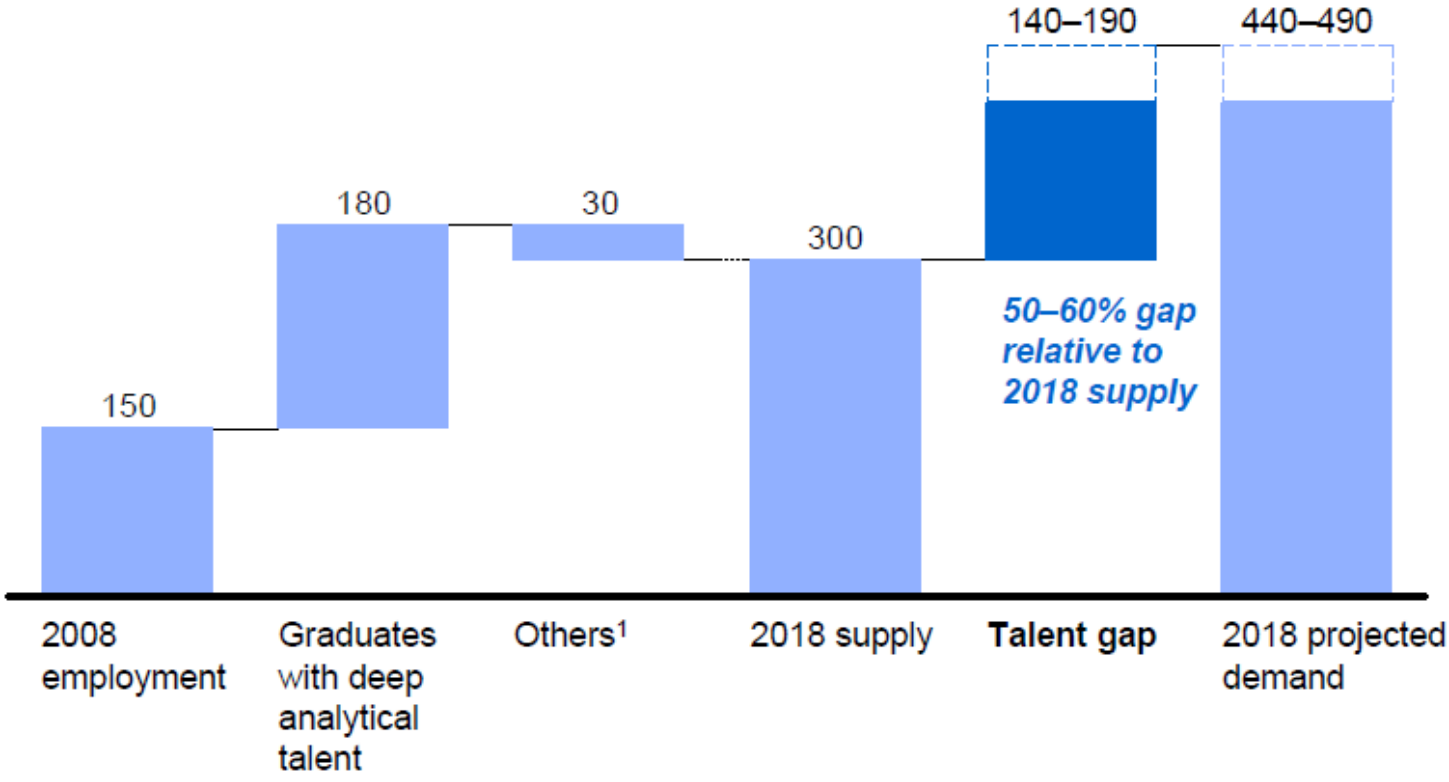
- 163 Zettabytes of data generated per year by 2025 (IDC)

- 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

DATA NEVER SLEEPS 2.0

How Much Data is Generated **Every Minute?**

Data is being created every minute of every day without us even noticing it. Given how much information is floating around these days, it's tempting to talk about big data only in terms of size. Big data describes the massive avalanche of digital activity pulsating through cables and airwaves, but it also describes all the things we were never able to measure before. With every status we share, every article we read or every photo we upload, we are creating a digital trail that tells a story. Below, we explore how much data is generated in one minute.

EVERY MINUTE OF THE DAY

PINTEREST USERS PIN 3,472 images

YOUTUBE USERS UPLOAD 72 HRS. OF NEW VIDEO.

EMAIL USERS SEND 204,000,000 MESSAGES.

Google RECEIVES OVER 4,000,000 SEARCH QUERIES.

FACEBOOK USERS SHARE 2,460,000 PIECES OF CONTENT.

TINDER USERS SWIPE 416,667 TIMES.

WHATSAPP USERS SHARE 347,222 PHOTOS.

TWITTER USERS TWEET 277,000 TIMES.

INSTAGRAM USERS POST 216,000 NEW PHOTOS.

AMAZON MAKES $83,000 IN ONLINE SALES.

PANDORA USERS LISTEN TO 61,141 HOURS OF music.

APPLE USERS DOWNLOAD 48,000 apps.

YELP USERS POST 26,380 REVIEWS.

SKYPE USERS CONNECT FOR 23,300 HOURS.

VINE USERS SHARE 8,333 VIDEOS.

THE GLOBAL INTERNET POPULATION GREW 14.3% FROM 2011 - 2013 AND NOW REPRESENTS **2.4 BILLION PEOPLE.**

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. Learn more at www.domo.com.

DOMO

SOURCES:

BITS.BLOGS.NYTIMES.COM, INTEL.COM, APPLE.COM, TIME.COM, DAILYMAIL.CO.UK, SKYPE.COM, STATISTICBRAIN.COM

DATA NEVER SLEEPS 3.0

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

Data is being created all the time without us even noticing it. Much of what we do every day now happens in the digital realm, leaving an ever-increasing digital trail that can be measured and analyzed. Just how much data do our tweets, likes and photo uploads really generate? For the third time, Domo has the answer—and the numbers are staggering.

every MINUTE of the DAY

Skype USERS MAKE 110,040 CALLS

UBER PASSENGERS TAKE 694 RIDES

Facebook USERS LIKE 4,166,667 POSTS

TWITTER USERS SEND 347,222 TWEETS

YOUTUBE USERS UPLOAD 300 HOURS OF NEW VIDEO

Buzzfeed USERS VIEW 34,150 VIDEOS

SNAPCHAT USERS SHARE 284,722 SNAPS

Instagram USERS LIKE 1,736,111 PHOTOS

Tinder USERS SWIPE 590,278 TIMES

PINTEREST USERS PIN 9,722 IMAGES

VINE USERS PLAY 1,041,666 VIDEOS

AMAZON RECEIVES 4,310 UNIQUE VISITORS

reddit USERS CAST 18,327 VOTES

Netflix SUBSCRIBERS STREAM 77,160 HOURS OF VIDEO

APPLE USERS DOWNLOAD 31,000 APPS

THE GLOBAL INTERNET POPULATION GREW 10.5% FROM 2013–2015 AND NOW REPRESENTS **3.2 BILLION PEOPLE.**

With each click, share and like, the world's data pool is expanding faster than we can comprehend. Businesses today are paying attention to scores of data sources to make crucial decisions about the future. The team at Domo can help your business make sense of this endless stream of data by providing executives with all their critical information in one intuitive platform. Domo delivers the insights you need to transform the way you run your business. **Learn more at www.domo.com.**

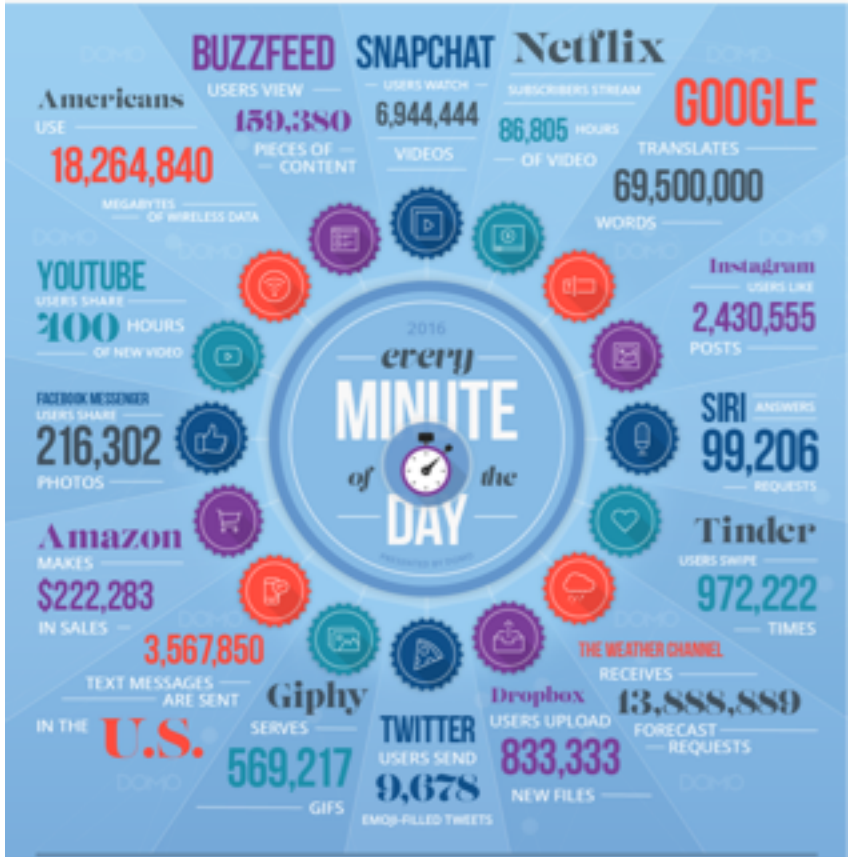
DOMO

SOURCES:

FACEBOOK, TWITTER, YOUTUBE, INSTAGRAM, PINTEREST, APPLE, NETFLIX, REDDIT, AMAZON, TINDER, BUZZFEED, STATISTA, INTERNET LIVE STATS, STATISTICBRAIN.COM

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.

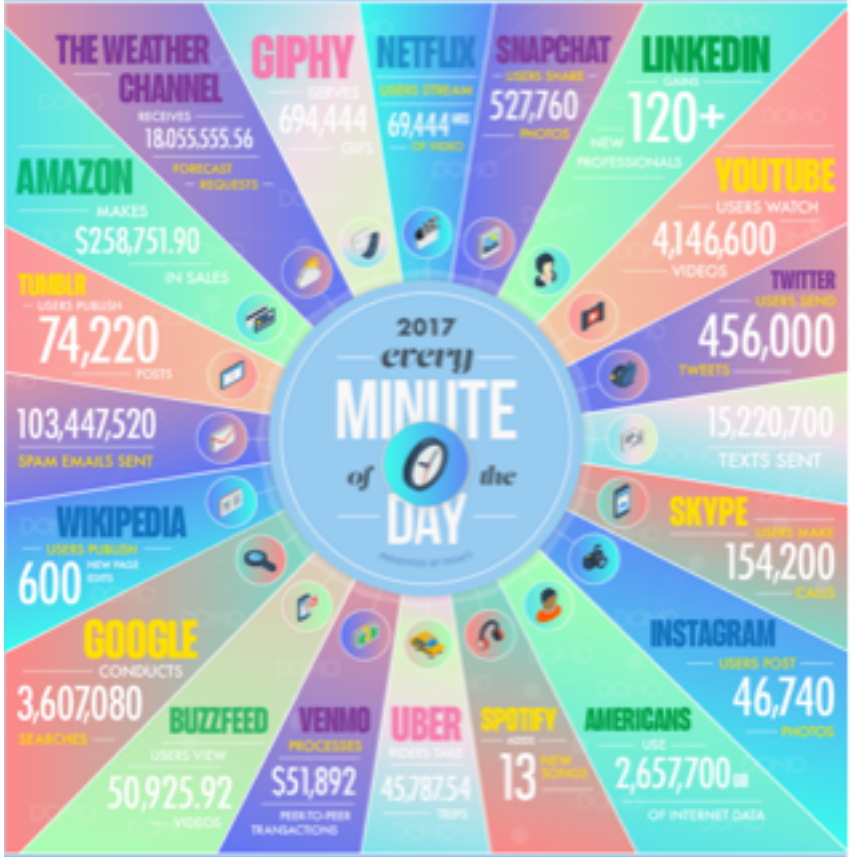


SOURCES: SNAPCHAT, NETFLIX, GOOGLE, INSTAGRAM, TINDER, THE WEATHER COMPANY, DROPBOX, GITHUB, GIPHY, YOUTUBE, BUZZFEED, AMAZON, CTIA, MARY MEEKER'S 2016 INTERNET TRENDS REPORT, USA TODAY, GLOBAL WEB INDEX

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.

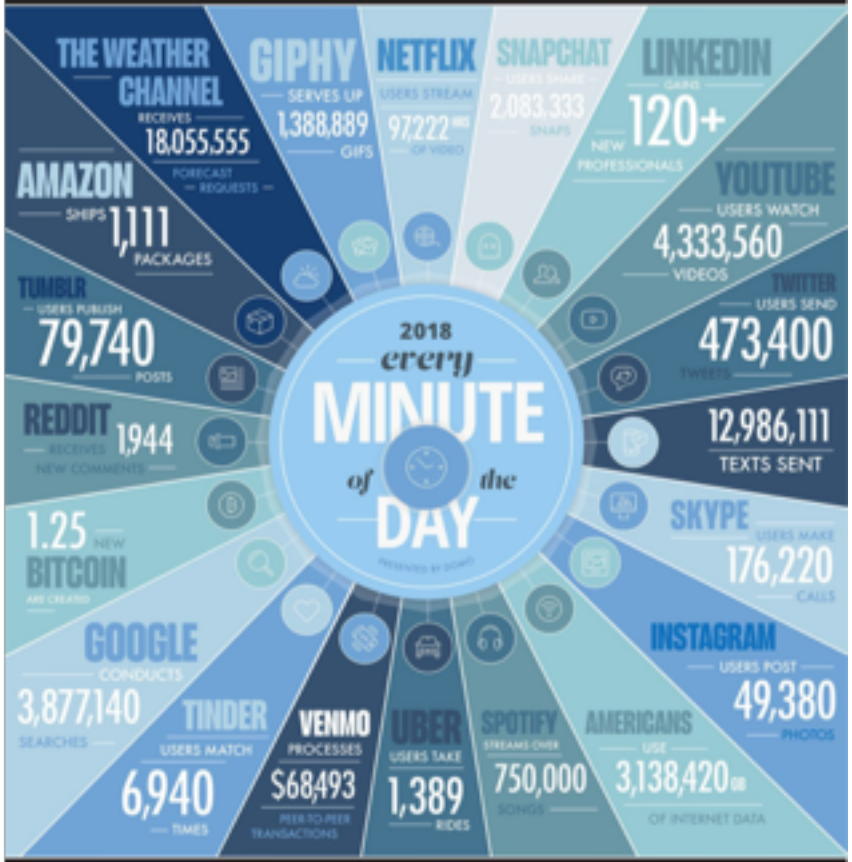


SOURCES: EXPANDEDRAMBLINGS.COM, MEMEDROID.COM, WIKIPEDIA, FORBES, ADWEEK.COM, FORTUNE.COM, BLOOMBERG.COM, ONEREACH.COM, IBM, BUZZFEED, INTERNET LIVE STATS, INTERNET WORLD STATS, BBC

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 world's 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.

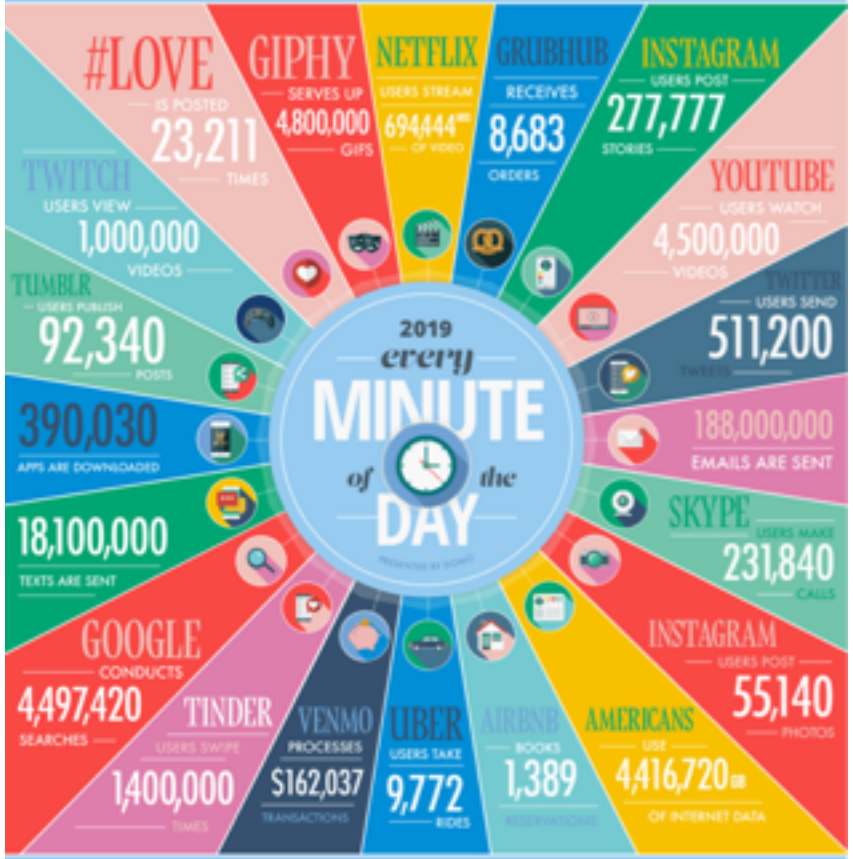
Learn more at domo.com

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

DATA NEVER SLEEPS 7.0

How much data is generated *every minute?*

There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute.



The world's internet population is growing significantly every year. As of January 2019, the internet reached 56.1% of the world's population and now represents 4.39 billion people — a 9% increase from January 2018.

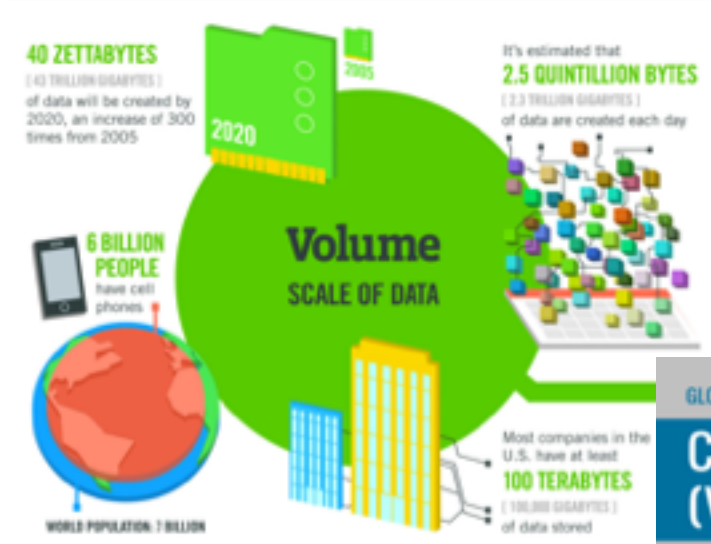
GLOBAL INTERNET POPULATION GROWTH 2012-2018
(IN BILLIONS)

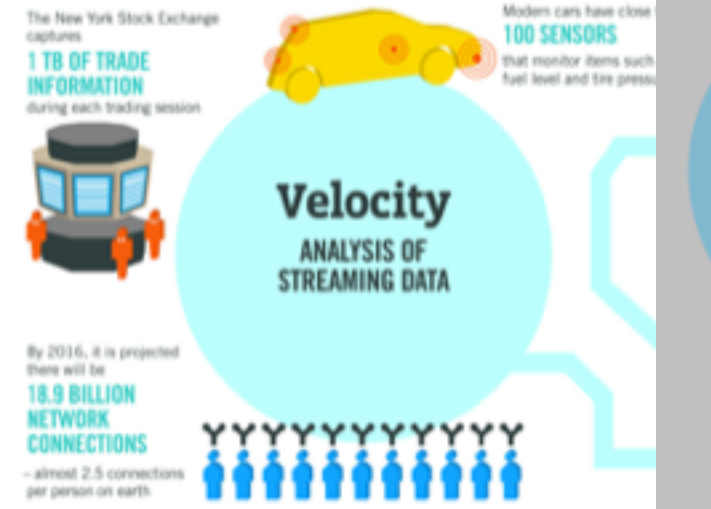
The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device,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

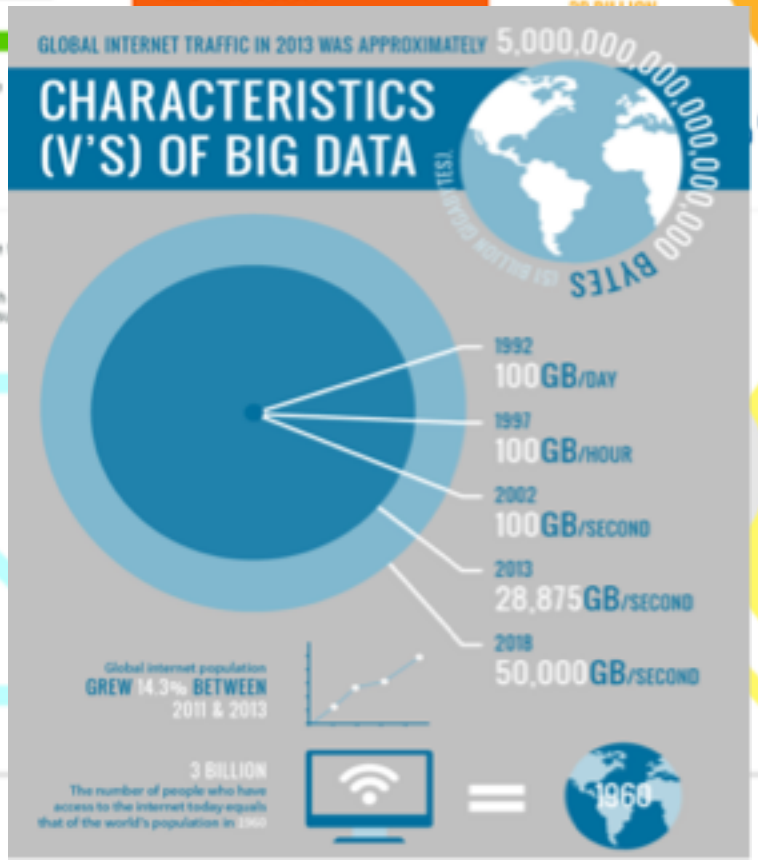
Learn more at domo.com

SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED

The FOUR V's of Big Data







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

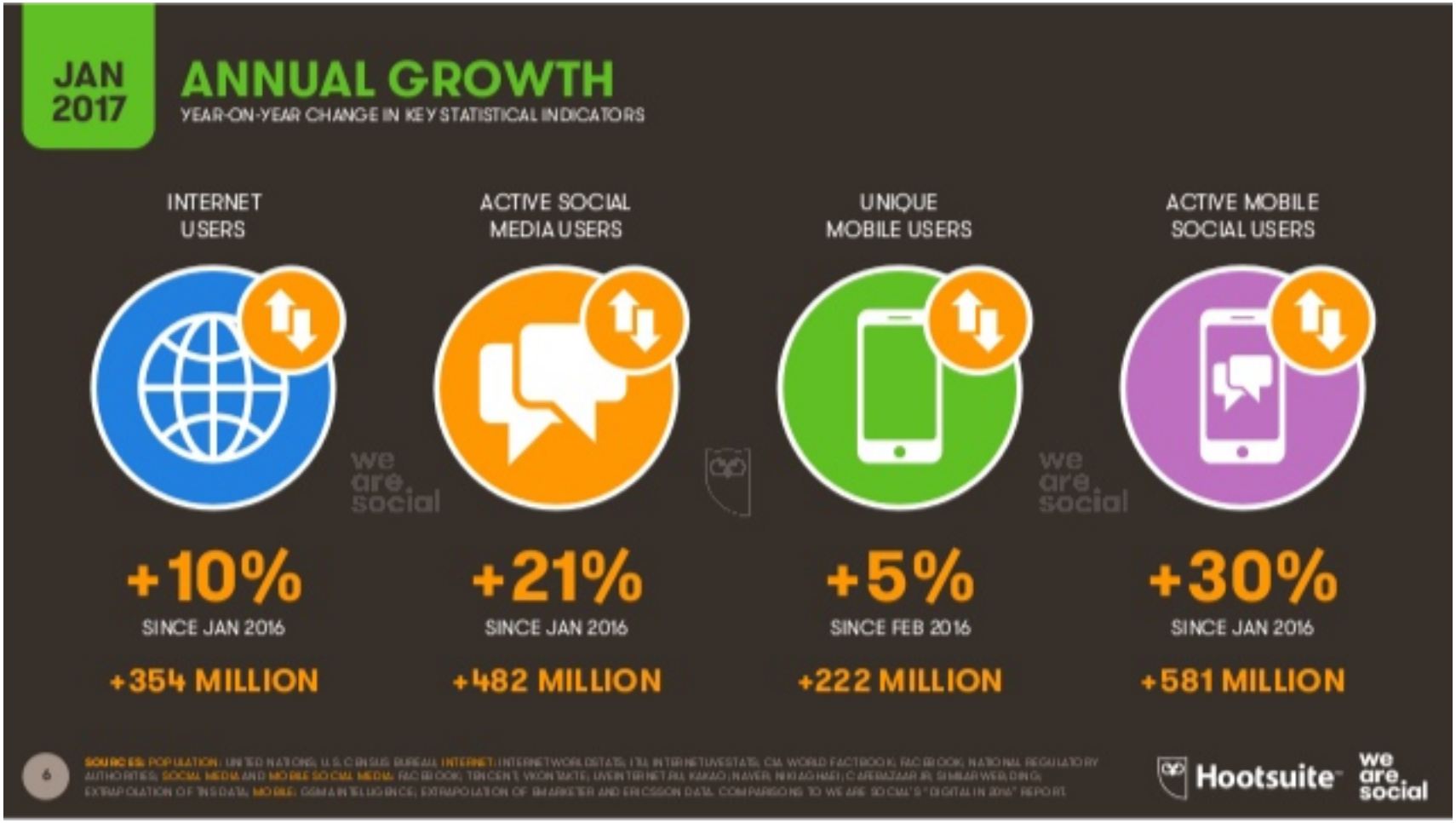
IBM

Looking Ahead

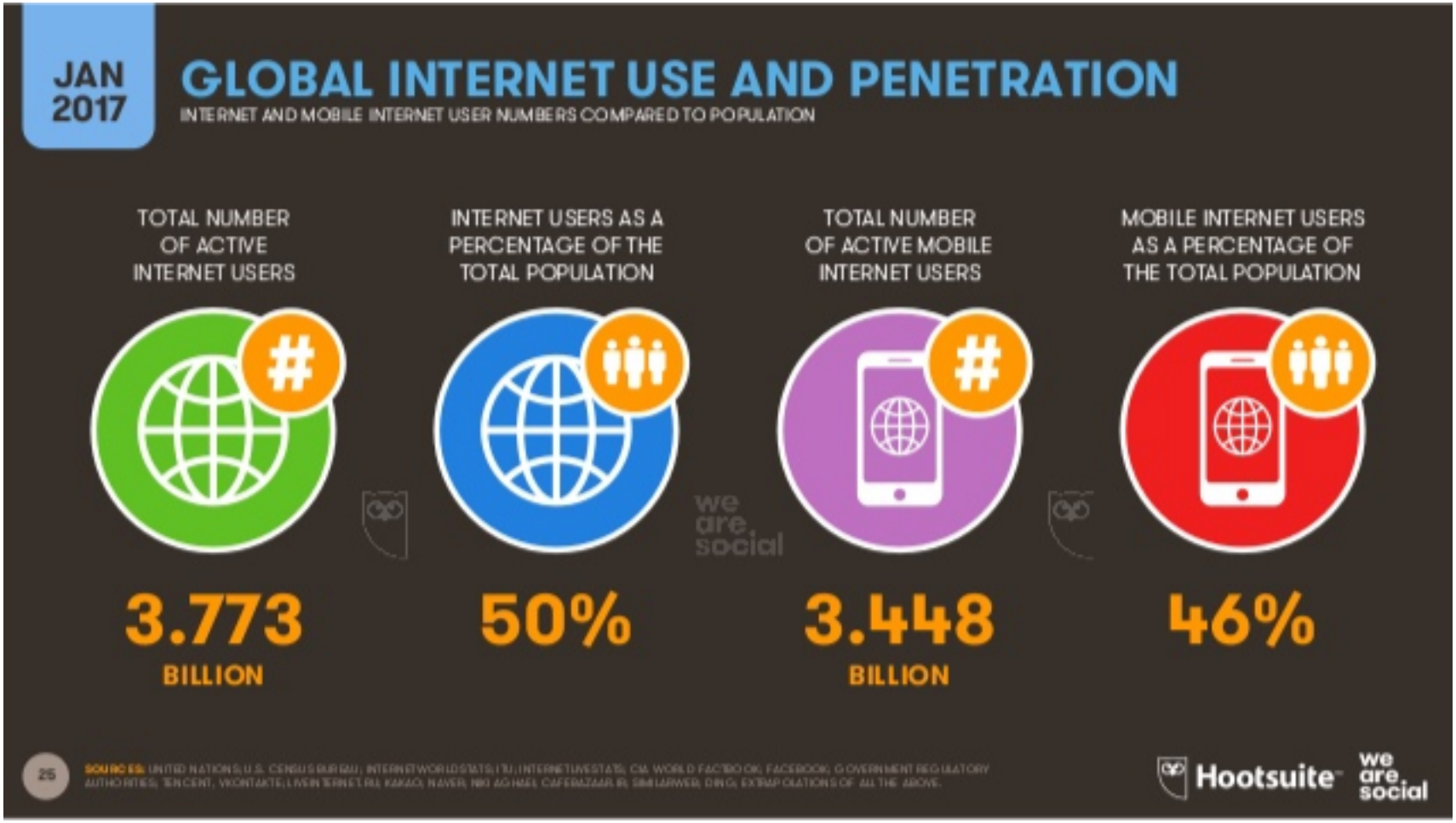
- 163 Zettabytes of data generated per year by 2025 (IDC)
- Revenues for big data and business analytics (BDA) will grow from $130B billion in 2016 to $203B in 2020 (IDC)

Internet Trends

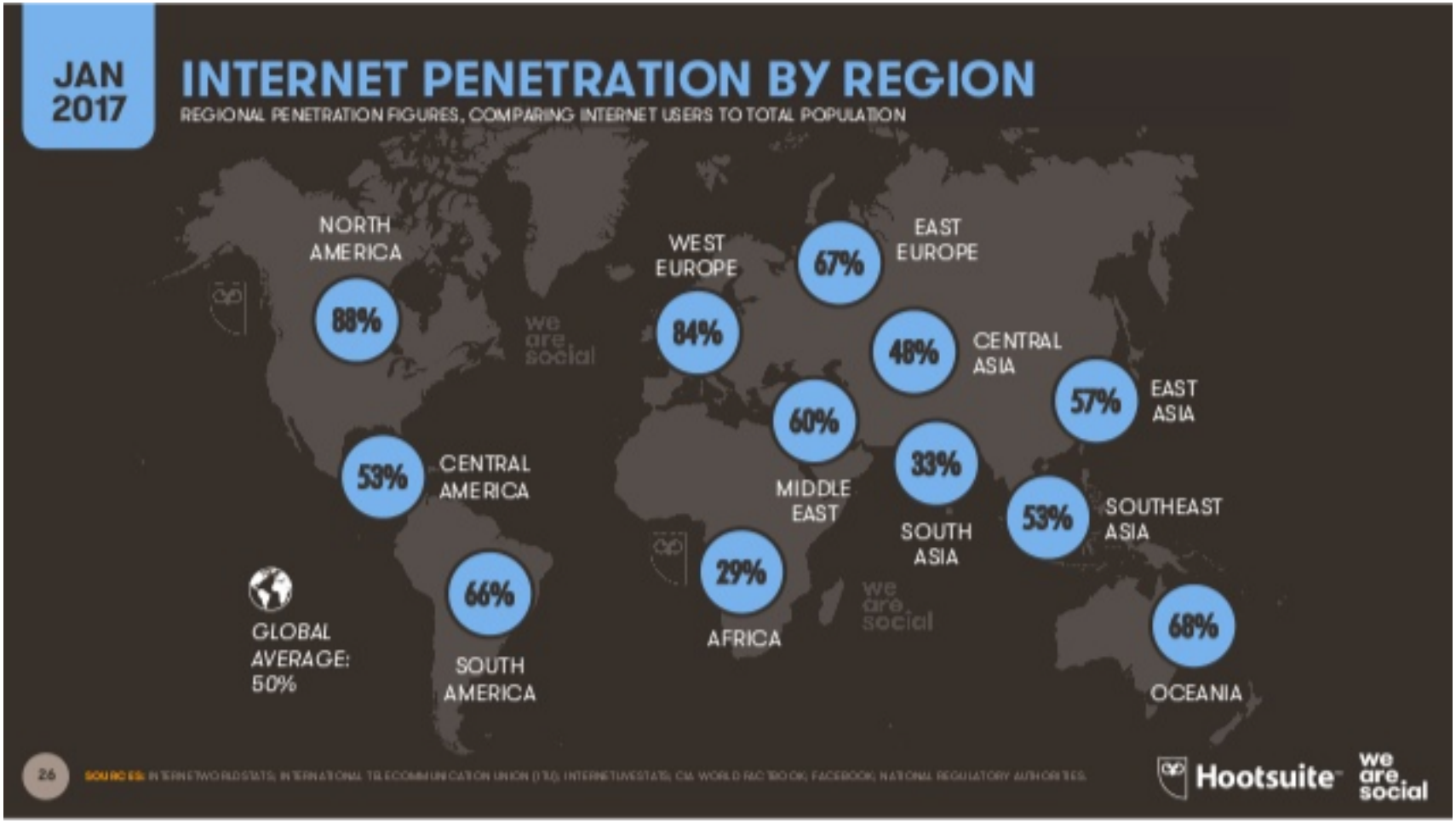
Internet Trends



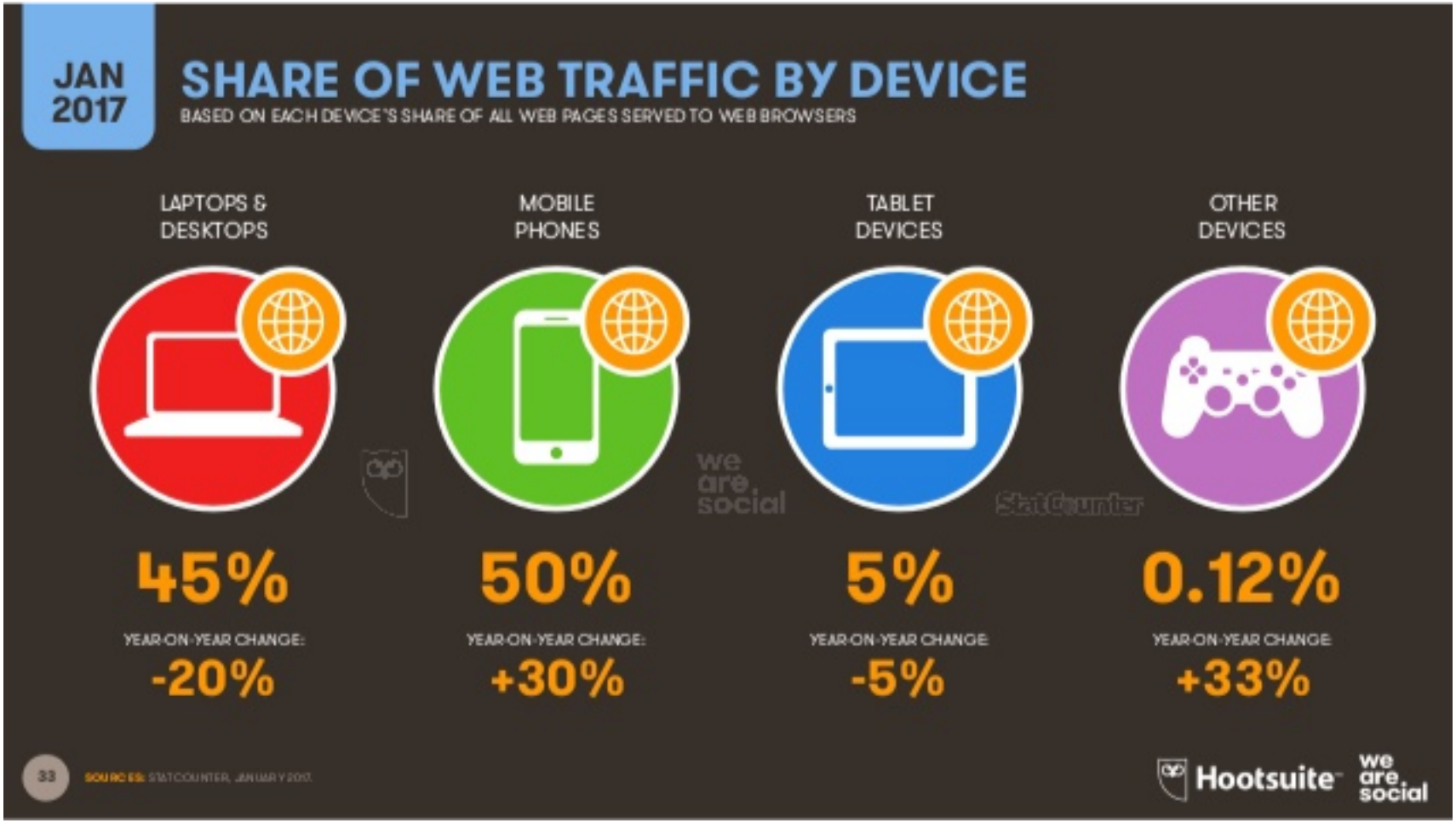
Internet Trends



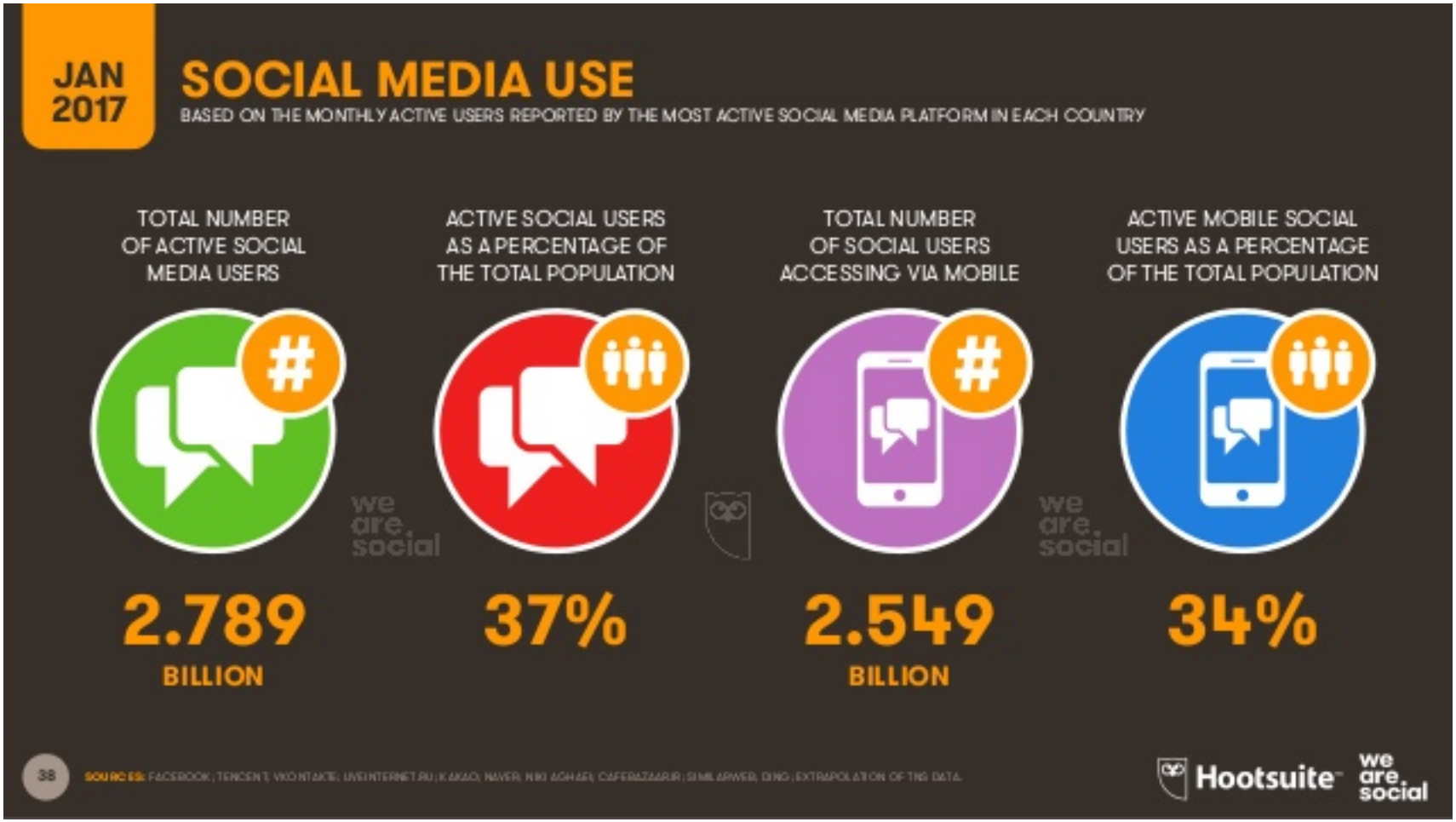
Internet Trends



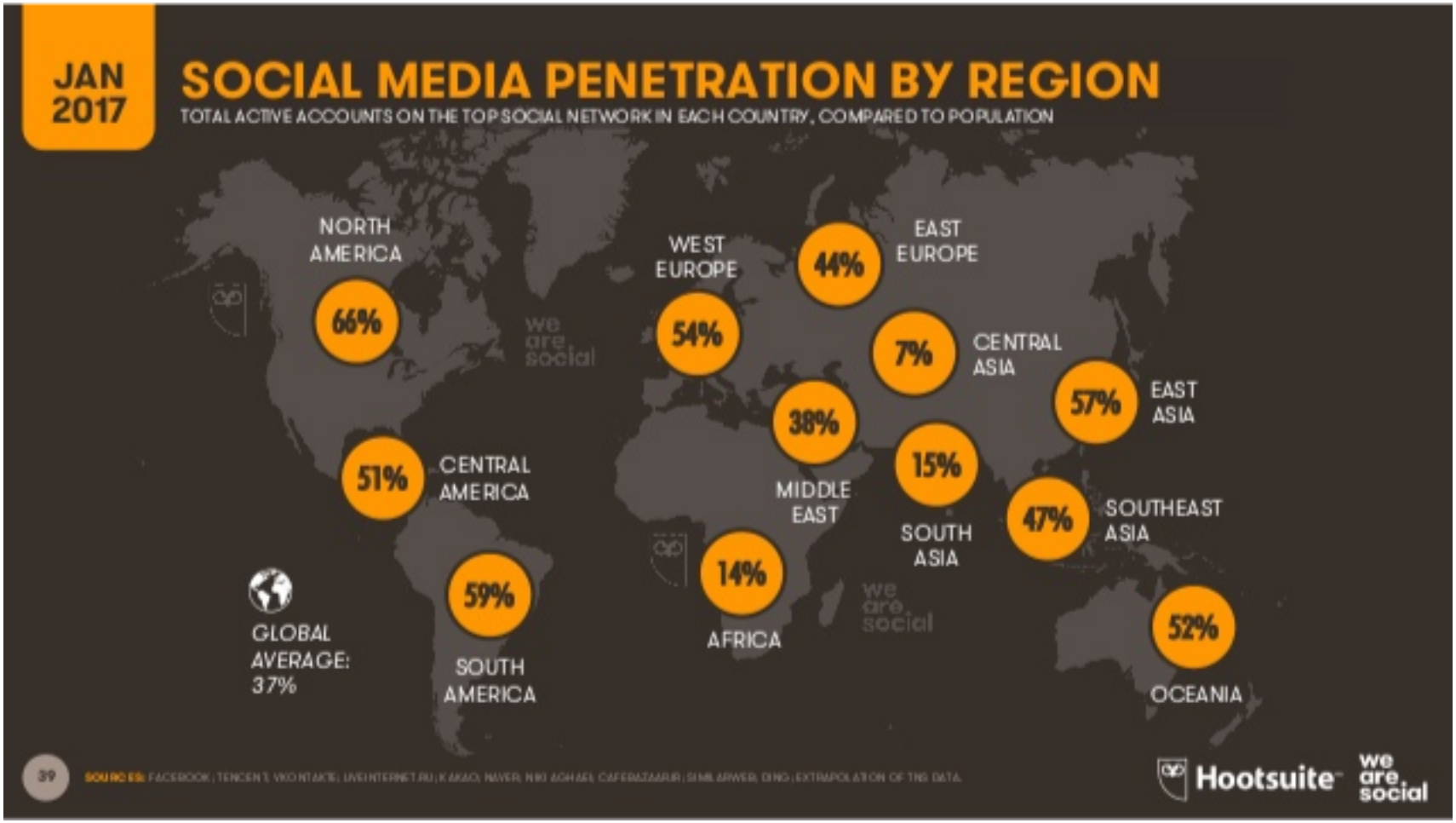
Internet Trends



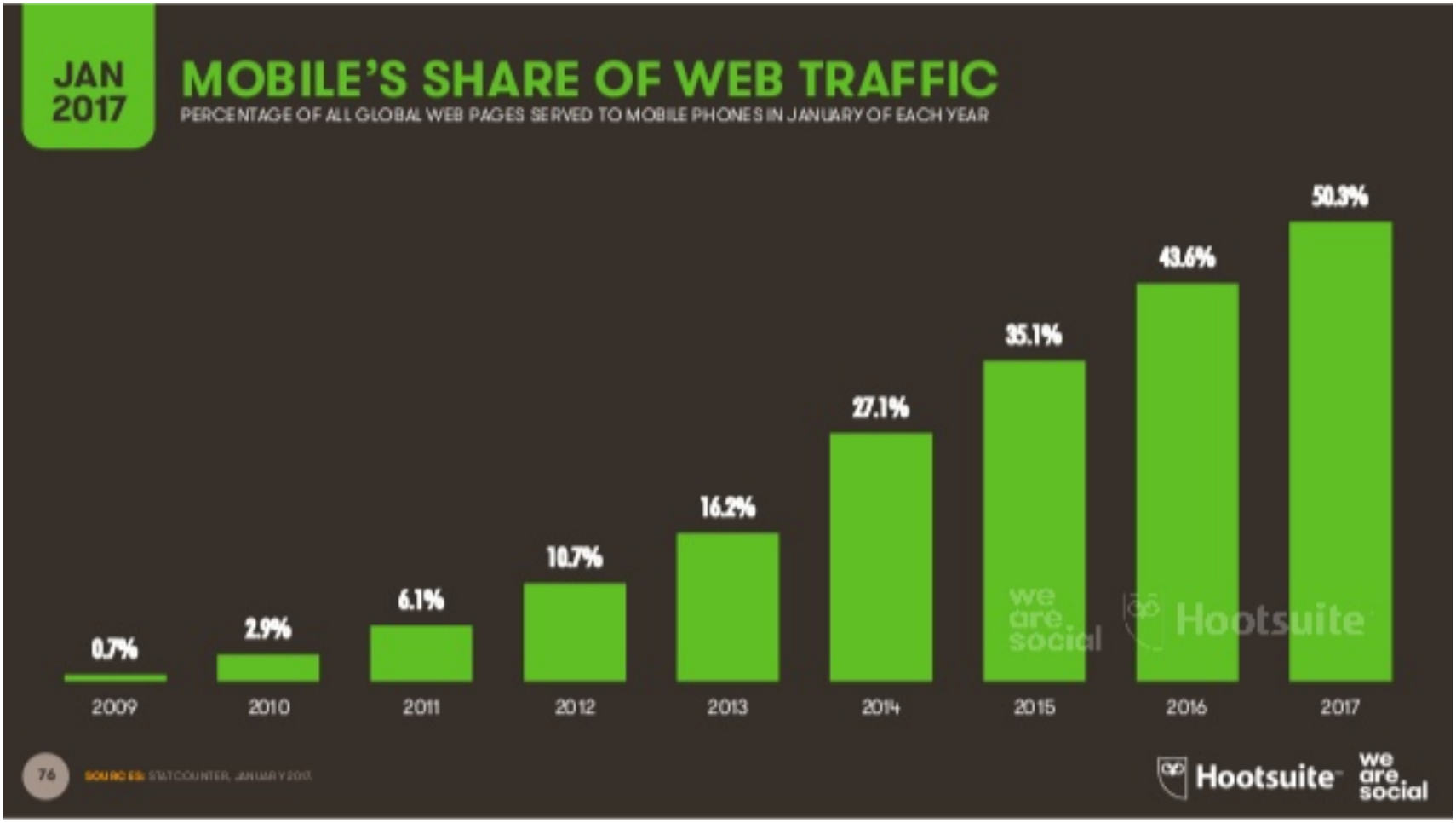
Internet Trends



Internet Trends



Internet Trends



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

The Economist

Reasons for Growth in Data

- Many areas of science have mature theories whose validation requires probing extreme phenomena.
  - These probes often generate very large data sets.
  - Particle physics: the Large Hadron Collider generates petabytes per year
- Many areas of science and engineering have become increasingly exploratory
  - Large data sets gathered in the hope that new phenomena will emerge.
  - Example: Genome sequencing projects

Reasons for Growth in Data

- Much human activity now takes place on the Internet
  - Generates data with substantial commercial and scientific value.
  - Commercial enterprises aim to provide personalized services
- Significant growth in the deployment of sensor networks that record biological, physical, and social phenomena at ever-increasing scale,
  - Sensor networks are increasingly interconnected.

Potential…

- If massive data could be exploited effectively
  - science would extend its reach
  - technology would become more adaptive, personalized, and robust
- A health-care system with detailed individual data
  - genomic, cellular, and environmental data
  - data from other individuals
  - results from biological and medical research
  - optimized treatments for each individual.

Potential…

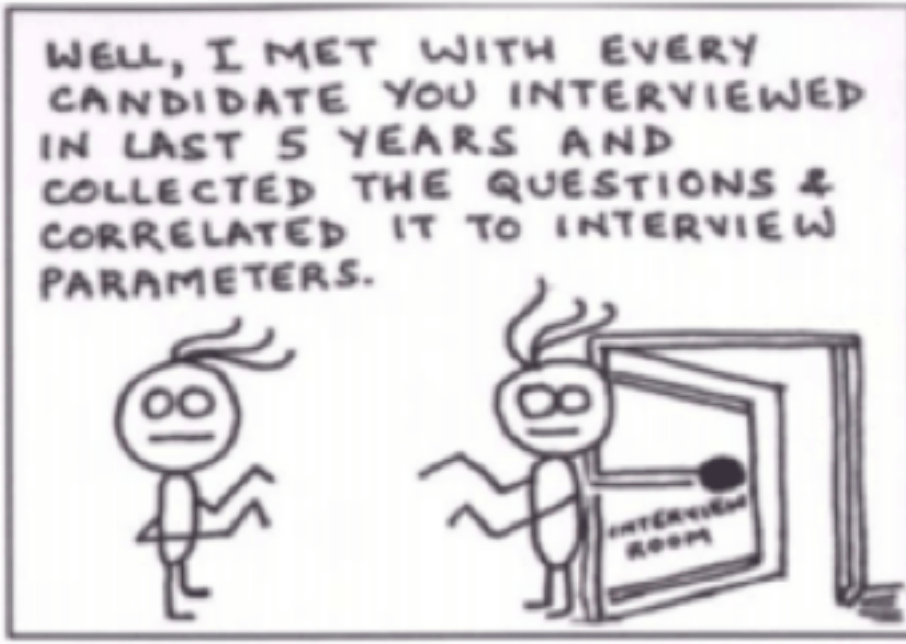
- If massive data could be exploited effectively
  - science would extend its reach
  - technology would become more adaptive, personalized, and robust
- A health-care system with detailed individual data
  - genomic, cellular, and environmental data
  - data from other individuals
  - results from biological and medical research
  - optimized treatments for each individual.

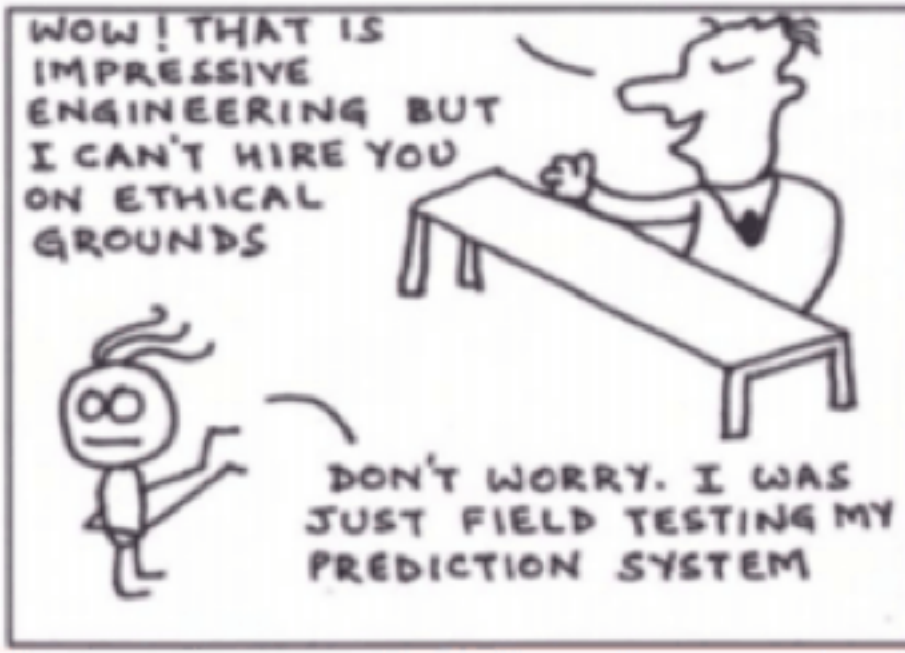
Potential…

- Microeconomics
  - Preferences and needs at the level of single individuals
  - Fine-grained descriptions of goods, skills, and services to
  - May help create new markets.

"what is particularly notable about the recent rise in the prevalence of "big data" is not merely the size of modern data sets, but rather that their fine-grained nature permits inferences and decisions at the level of single individuals." National Academy of Sciences

When you interview a data scientist...





Driving Factor

- Data centric technology
  - Databases, search
- Computing technology
  - Cloud, parallel and distributed computing
- Data Analysis methods
  - Machine learning
  - Data Mining
  - Statistics
  - Algorithms

Challenges

- Inference
  - Data to knowledge
- Data privacy
  - Is it OK for a cell-phone company to give the tracking data of customers for academic research without restrictions or controls?
  - Would it be acceptable for law-enforcement purposes?
  - What if the phone data were correlated with the owners' patterns of activities?
- Data ownership
  - If Google were to go out of business, who owns all the stored email data?
  - Many transit agencies now track their buses in real time. Does the public own that data?

Impact on Science

TABLE 1.1 Scientific and Engineering Fields Impacted by Massive Data

Area Affected in 1995	Area Affected in 2012	Noteworthy Use Cases
Physical sciences	Physical sciences	Astronomy, particle physics
Climatology	Climatology	
Signal processing	Signal processing	
Medicine	Medicine	Imaging, medical records
Artificial intelligence	Artificial intelligence	Natural language processing, computer vision
Marketing	Marketing	Internet advertising, corporate loyalty programs
N/A	Political science	Agent-based modeling of regime change
N/A	Forensics	Fraud detection, drug/human/CBRNe trafficking
N/A	Cultural studies	Human terrain assessment, land use, cultural geography
N/A	Sociology	Comparative sociology, social networks, demography, belief and information diffusion
N/A	Biology	Genomics, proteomics, ecology
N/A	Neuroscience	fMRI, multi-electrode recordings
N/A	Psychology	Social psychology

NOTE: CBRNe, chemical, biological, radiological, nuclear, enhanced improvised explosive devices; fMRI, functional magnetic resonance imaging; N/A, not applicable.

Data Age 2025

Data Age 2025:

DATA AGE 2025

The Evolution of Data to Life-Critical

Don't Focus on Big Data; Focus on the Data That's Big

IDC

Evolution of Computing

Before 1980

- Data sits almost exclusively in datacenters
- Data and compute centralized
- Business-focused

1980—2000

- Data and compute are distributed
- Datacenters expand role in managing data
- Quick expansion in entertainment

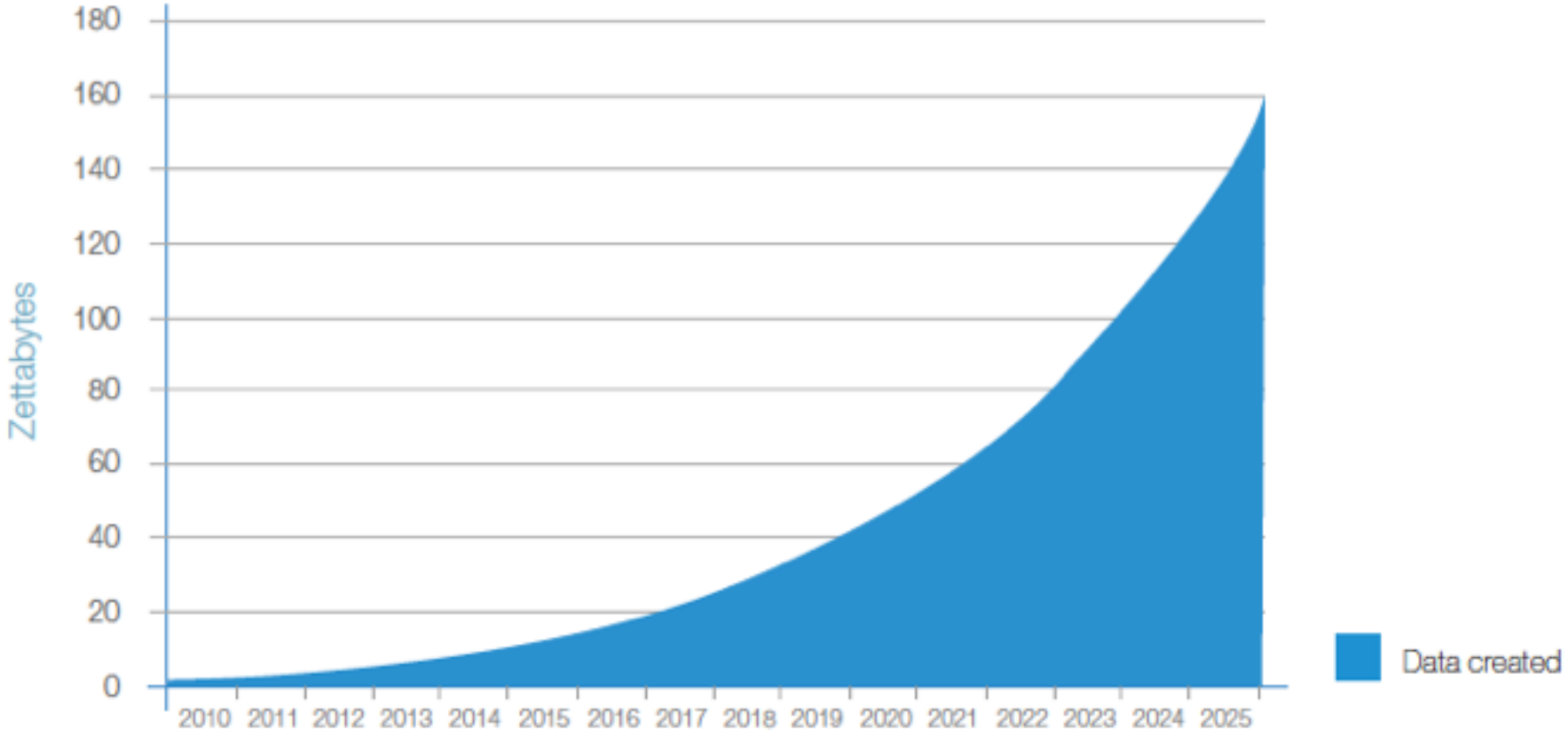
2000 to Today

- Datacenters expand to cloud infrastructures
- Compute continues to be distributed; data begins to contract
- Add social to the mix

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017

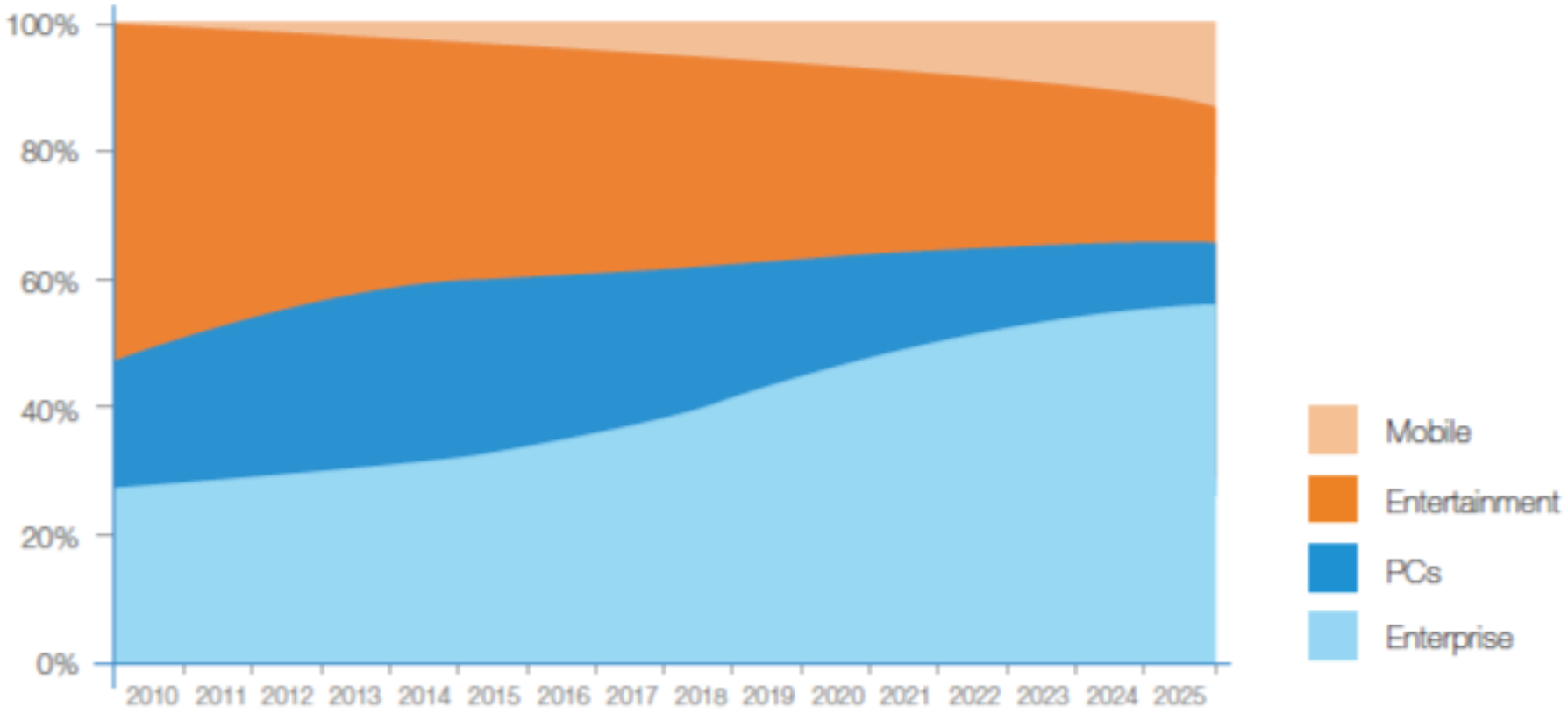
IDC
Analyze the Future

Evolution of Computing



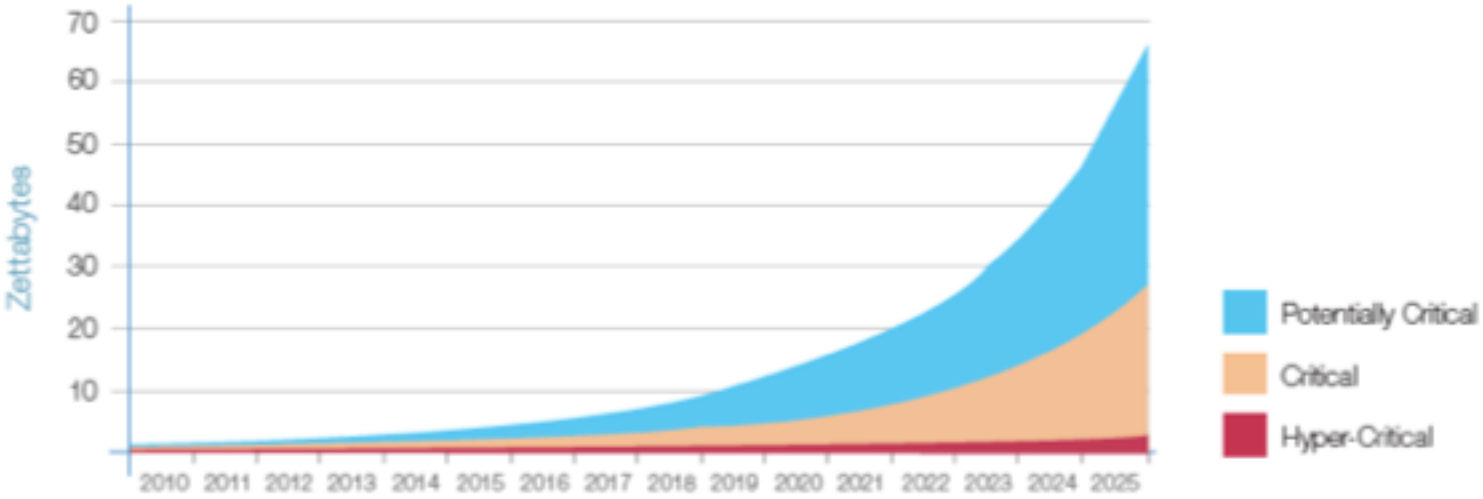
IDC

Where data is stored



IDC
Analyze the Future

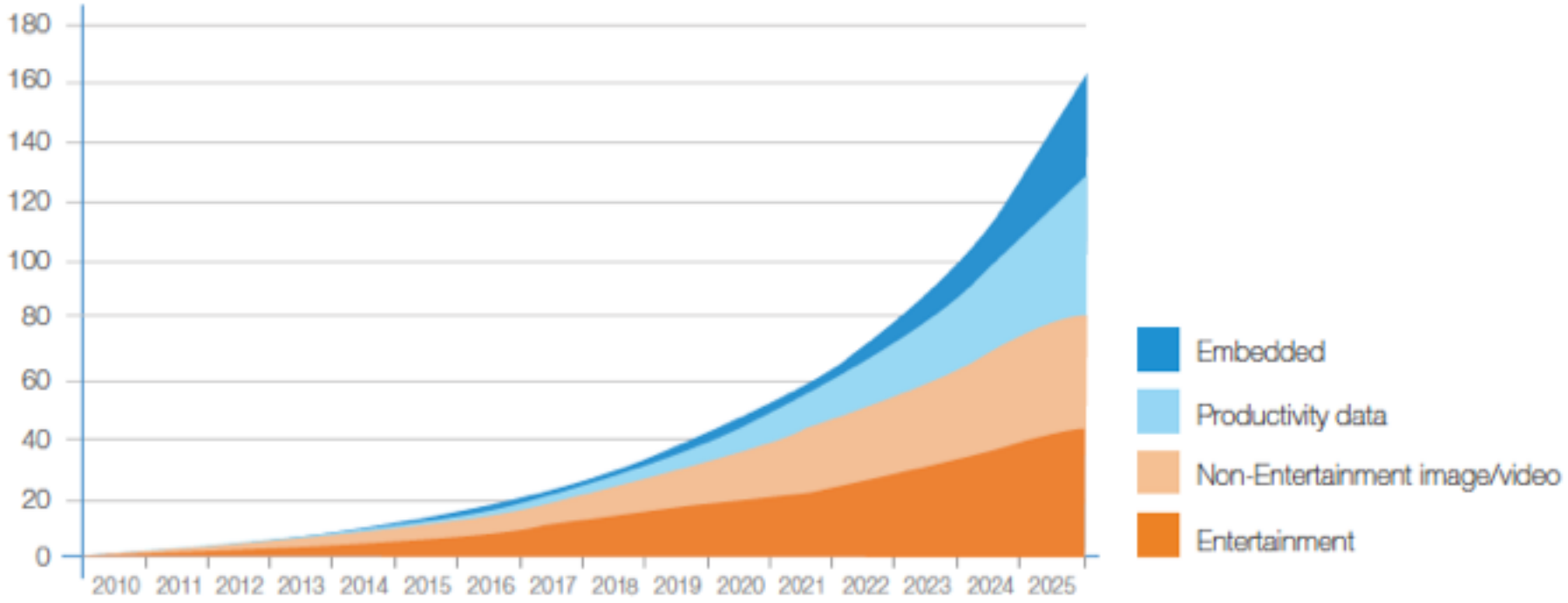
Data Criticality



Data Type	CAGR 2015 to 2025
All Data. Includes all data in the global datasphere.	30%
Potentially critical. Data that may be necessary for the continued, convenient operation of users' daily lives	37%
Critical. Data known to be necessary for the expected continuity of users' daily lives.	39%
Hypercritical. Data with direct and immediate impact on the health and well-being of users. (Examples include commercial air travel, medical applications, control systems, and telemetry. This category is heavy in metadata and data from embedded systems.)	54%

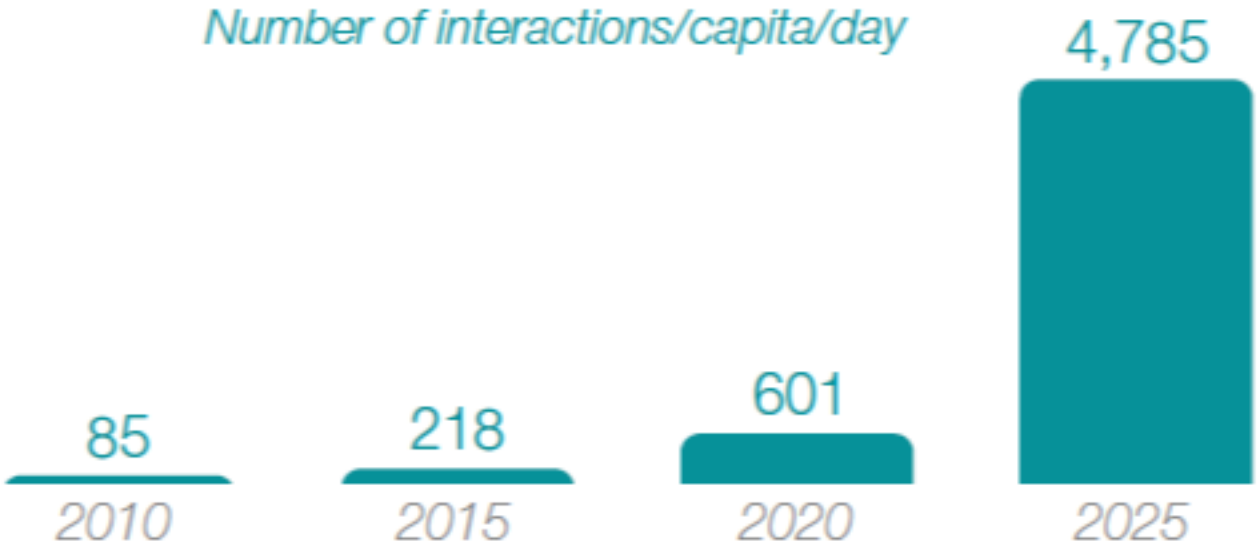
IDC
Analyze the Future

Data Creation



IDC
Analyze the Future

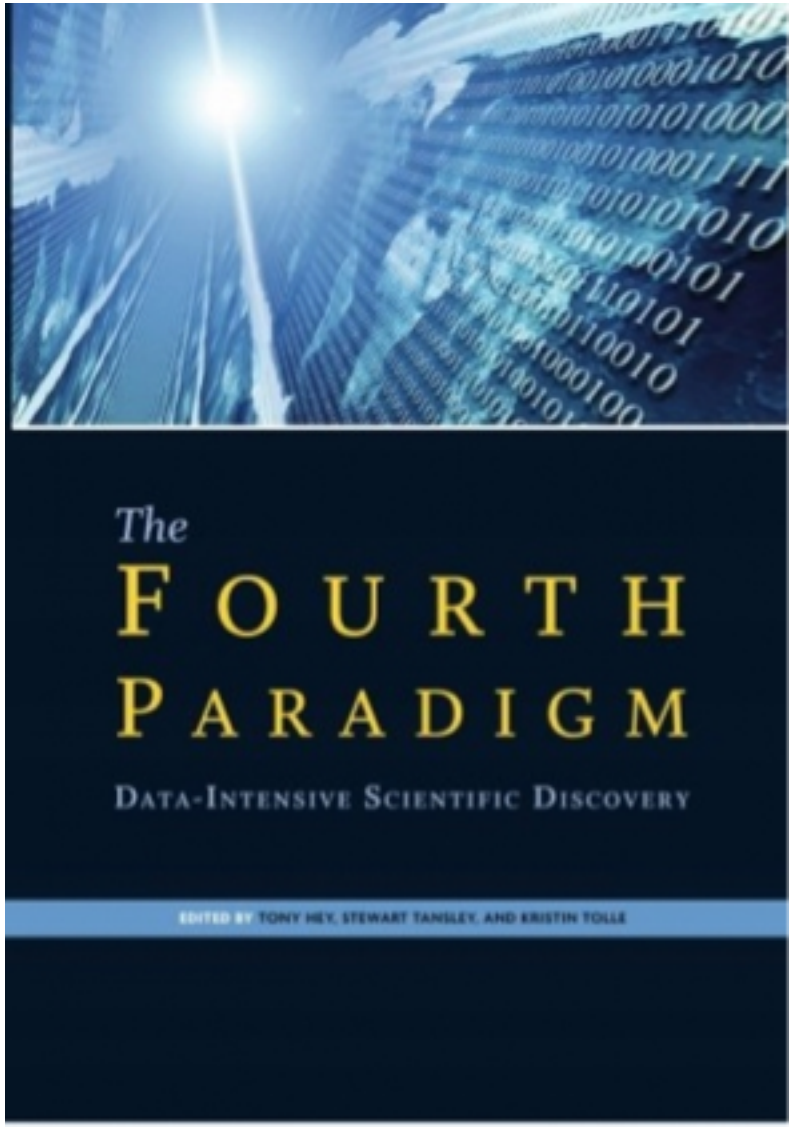
Interaction w. Embedded Devices



IDC

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.

Models

- "All models are wrong, but some are useful". statistician George Box

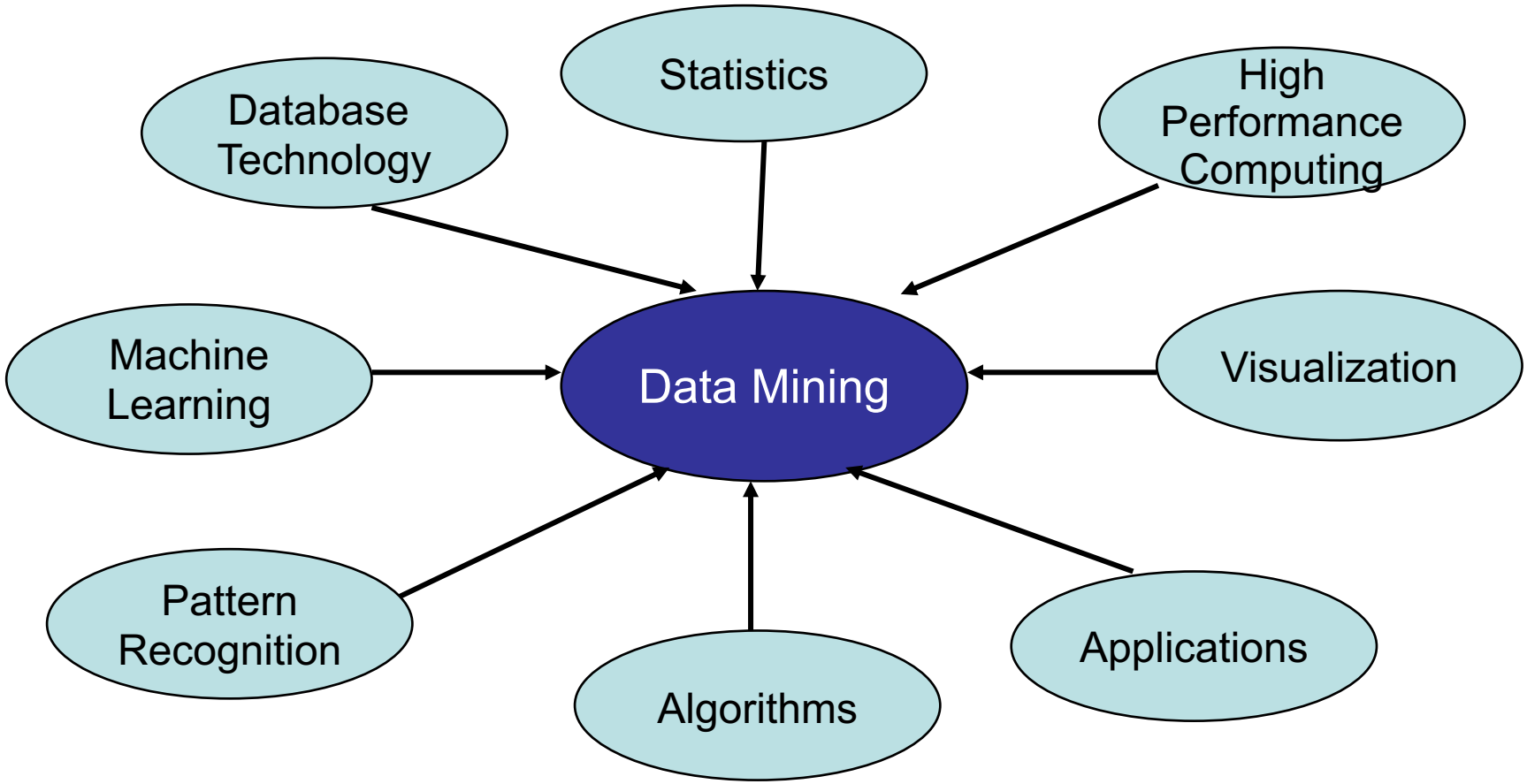
- Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

Models: Peter Norvig

- In complex, messy domains, particularly game-theoretic domains involving unpredictable agents such as human beings, there are no general theories that can be expressed in simple equations like $F = m\,a$ or $E = m\,c^2$.
- But if you have a dense distribution of data points, it may be appropriate to employ non-parametric density approximation models such as nearest-neighbors or kernel methods rather than parametric models such as low-dimensional linear regression.

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?

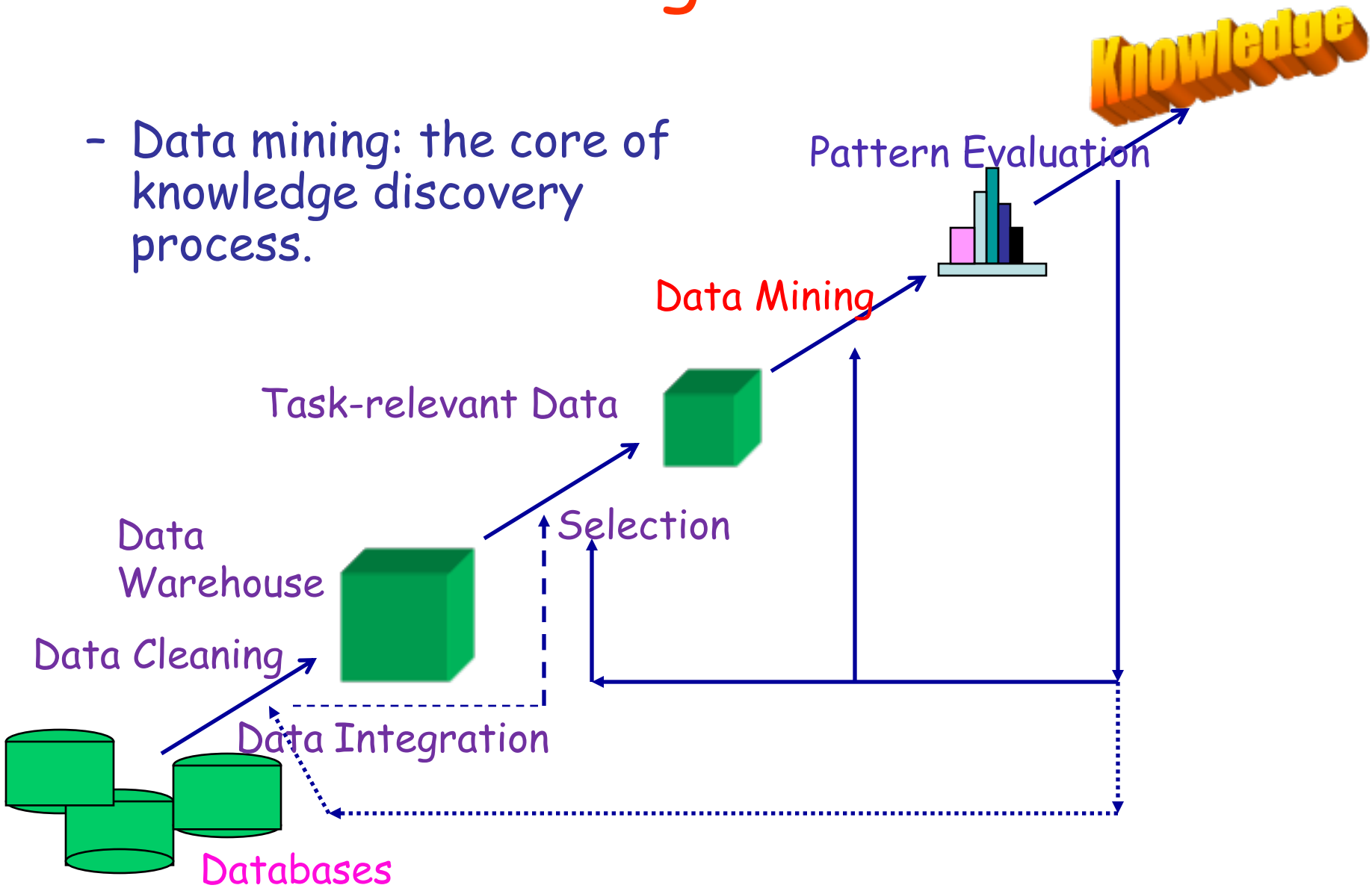
- Novel: previously unknown, not obvious
- Valid: broadly applicable (on new data) with some certainty
- Meaningful: humans should be able to understand
- Useful: should be possible to act on the result (actionable)

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

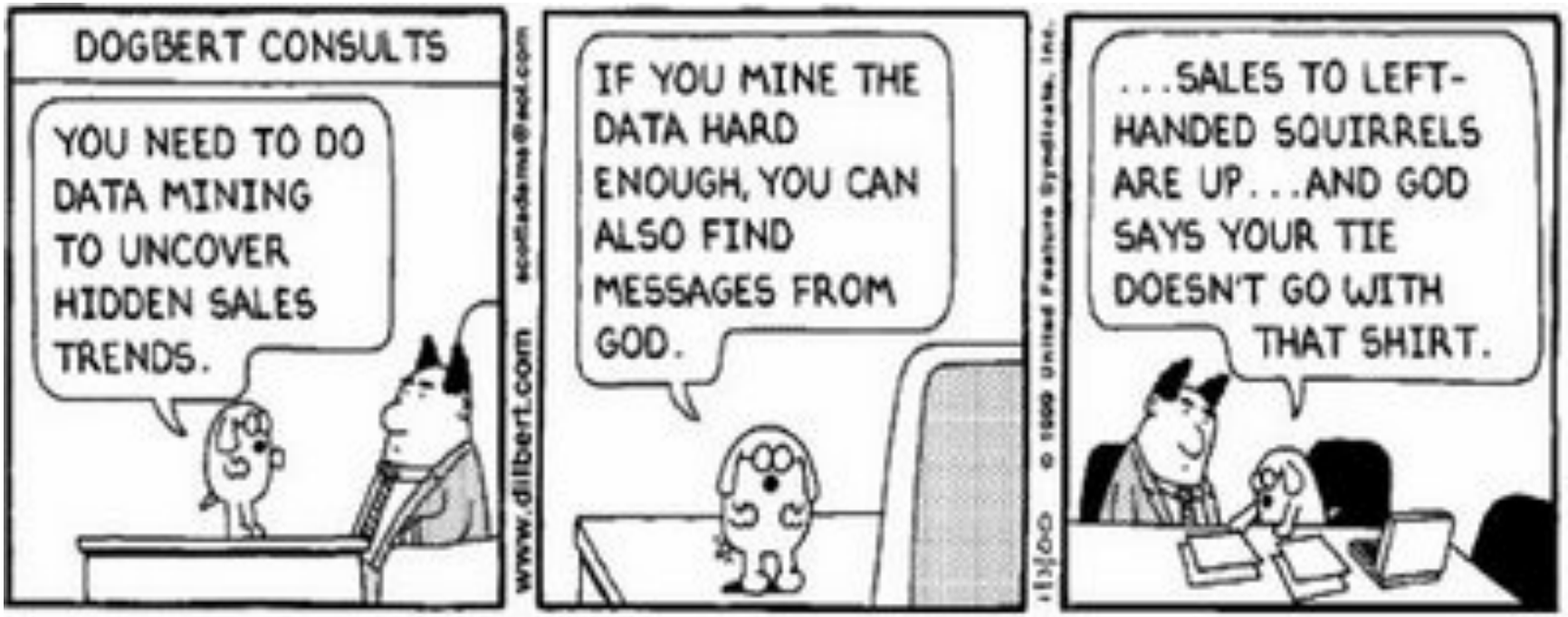
Data Mining: A KDD Process

- 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

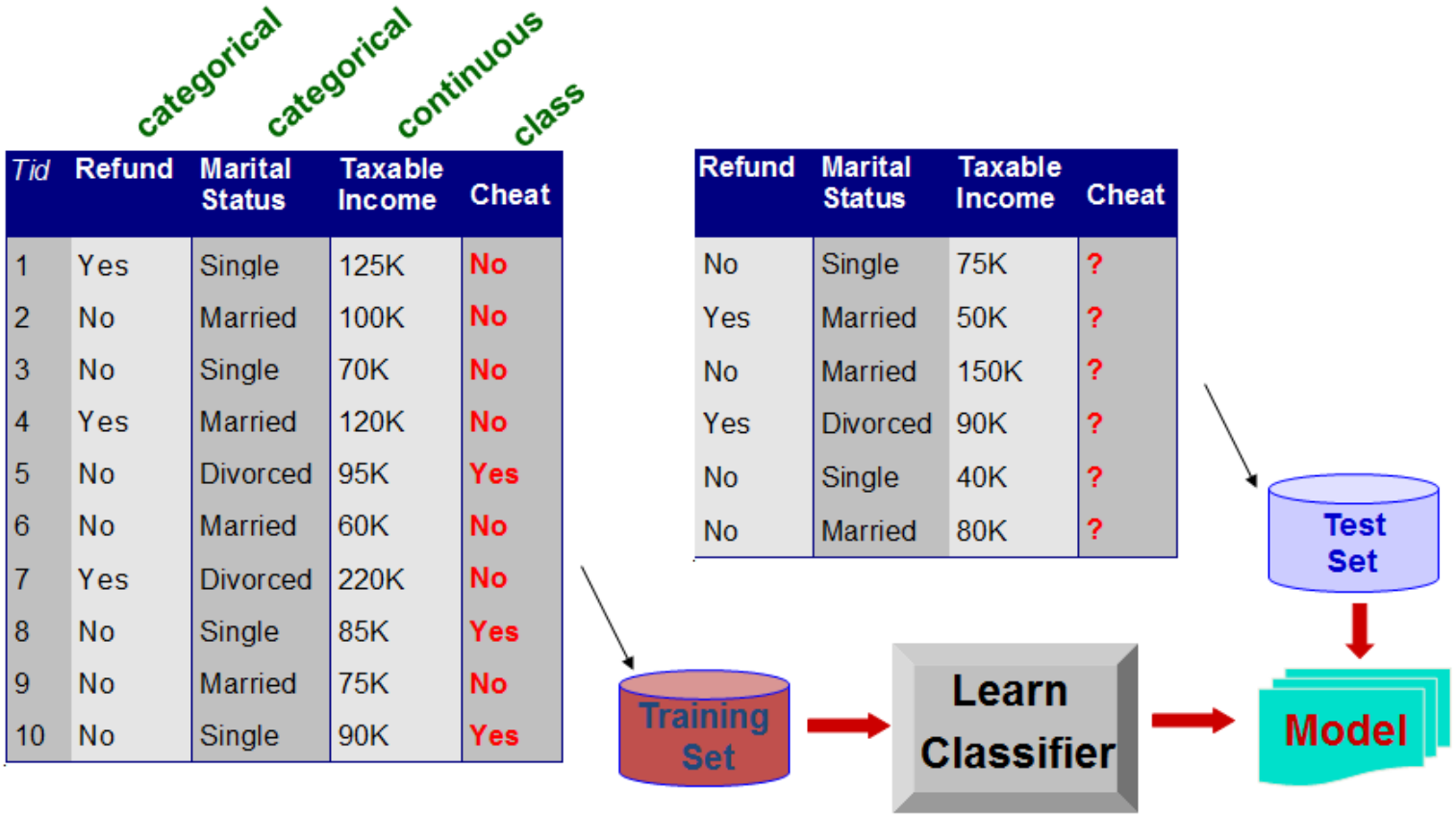
- Descriptive methods
  - Find human-interpretable patterns that describe the data
    - Example: Clustering
- Predictive methods
  - Use some variables to predict unknown or future values of other variables
    - Example: Recommender systems

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.

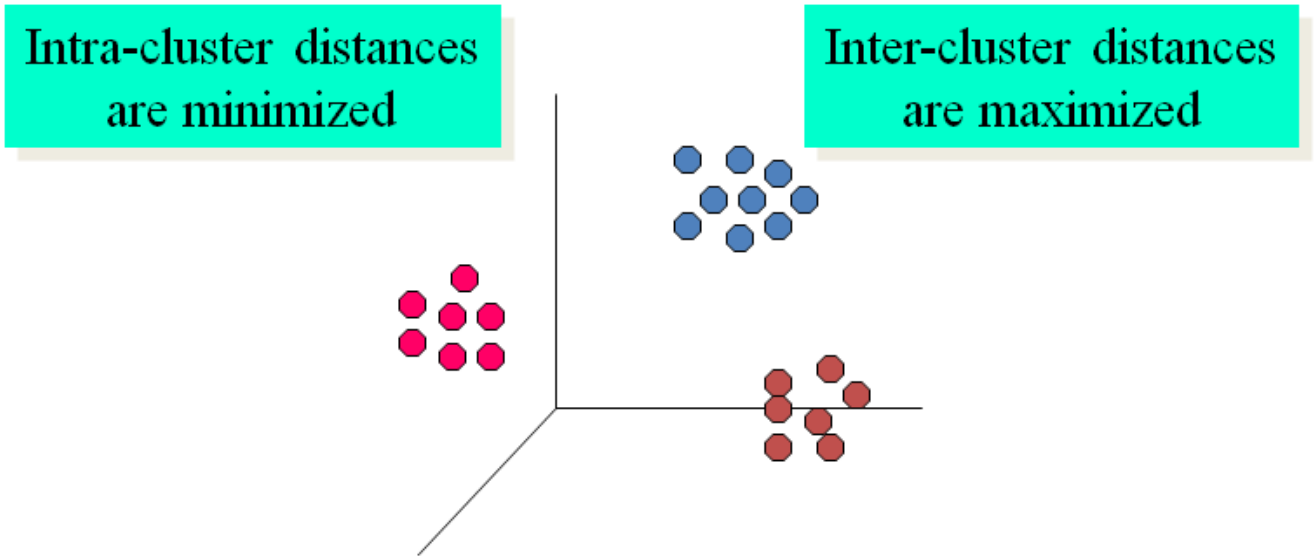


Classification

- 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, …

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

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

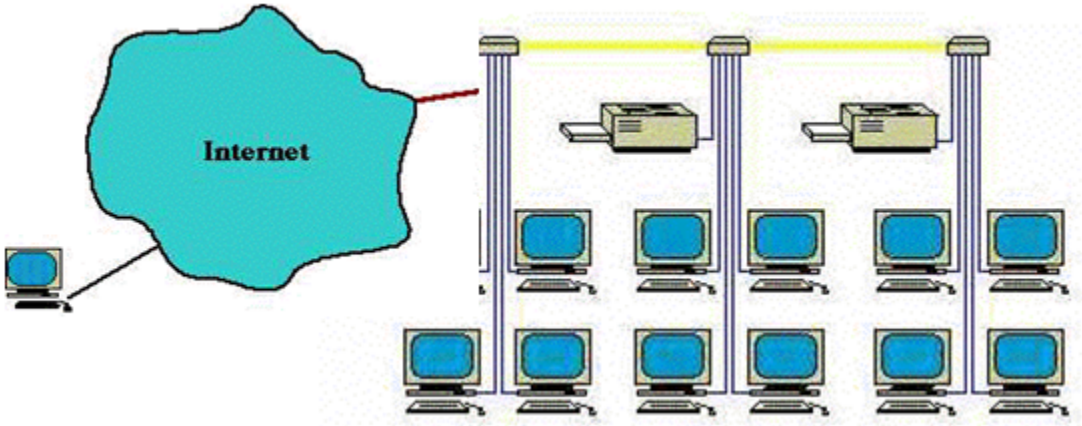
- Applications: marketing and sales promotion (cross-selling)

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

Spatial Data Mining

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

COMMUNITY COLLABORATIVE RAIN, HAIL & SNOW NETWORK

"Because every drop counts"

Home | States | View Data | Maps | My Data Entry | Login

Welcome to CoCoRaHS! "Volunteers working together to measure precipitation across the nation."

Main Menu

- Home
- About Us
- Join CoCoRaHS
- Contact Us
- Donate

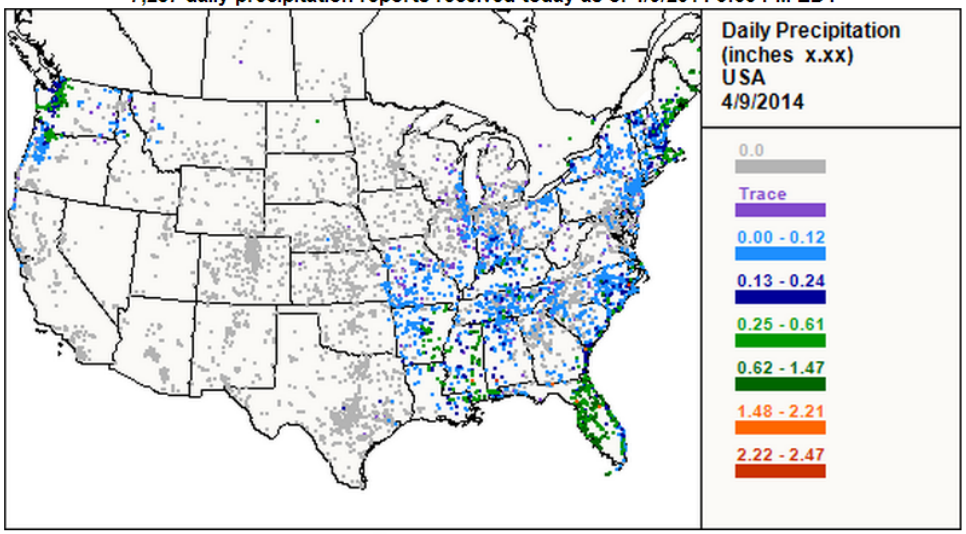
Resources

- FAQ / Help
- Education
- Training Slide-Shows
- Videos
- Drought Impacts
- Evapotranspiration

- Volunteer Coordinators
- Hail Pad Distribution/Drop-off
- Help Needed
- Printable Forms

- *The Catch*
- Message of the Day

7,287 daily precipitation reports received today as of 4/9/2014 3:55 PM EDT



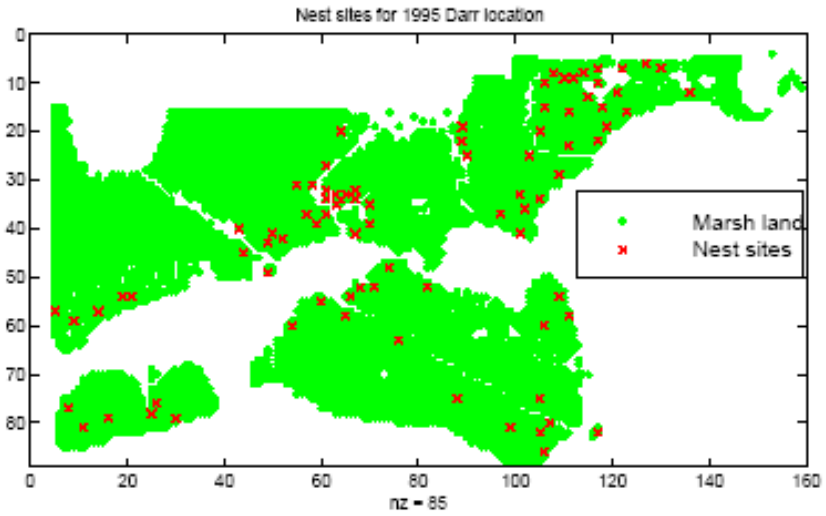
JOIN CoCoRaHS

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

Spatial Predictive Models

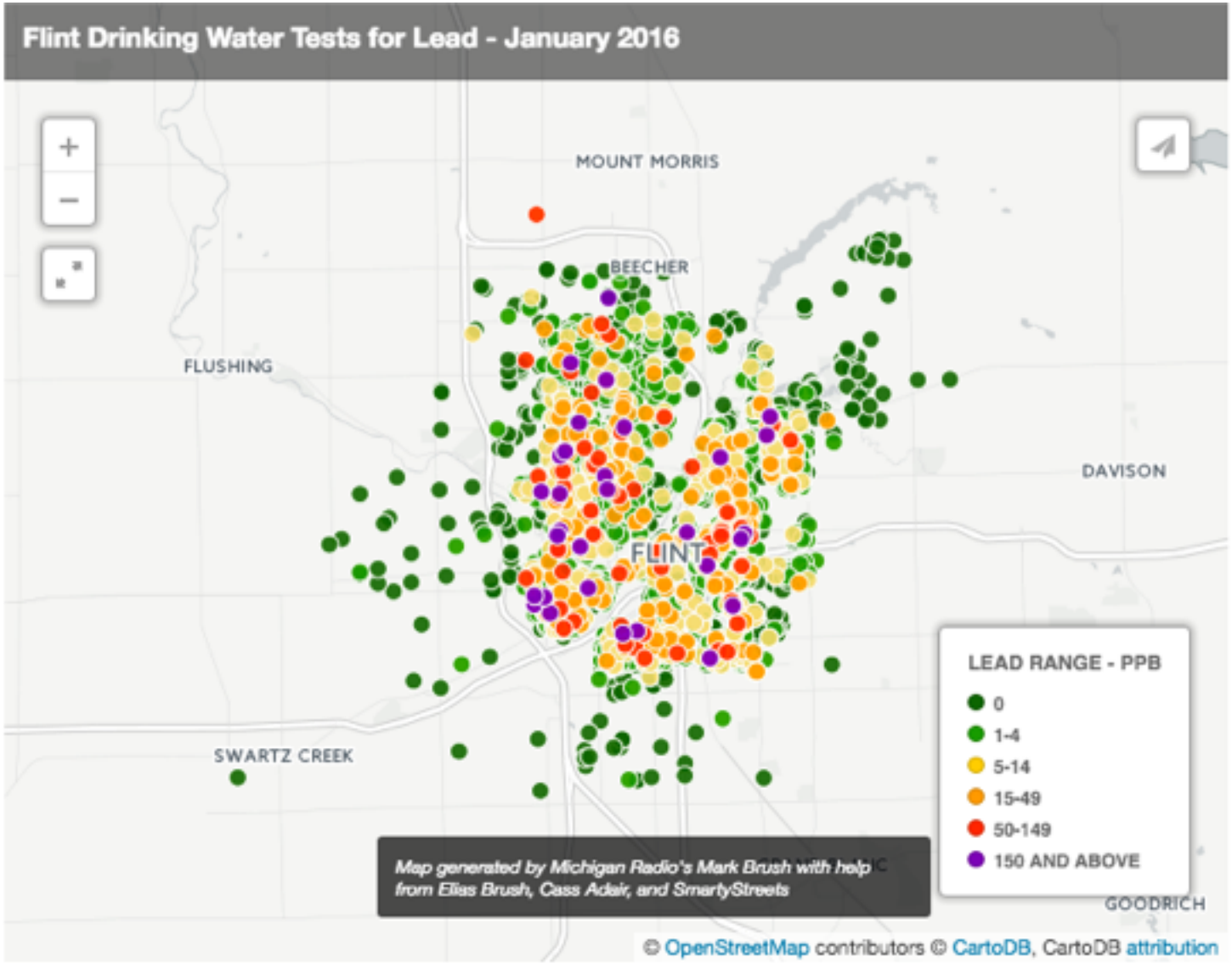
- Location Prediction: Bird Habitat Prediction
  - Given training data
  - Predictive model building
  - Predict new data



Spatial Clustering

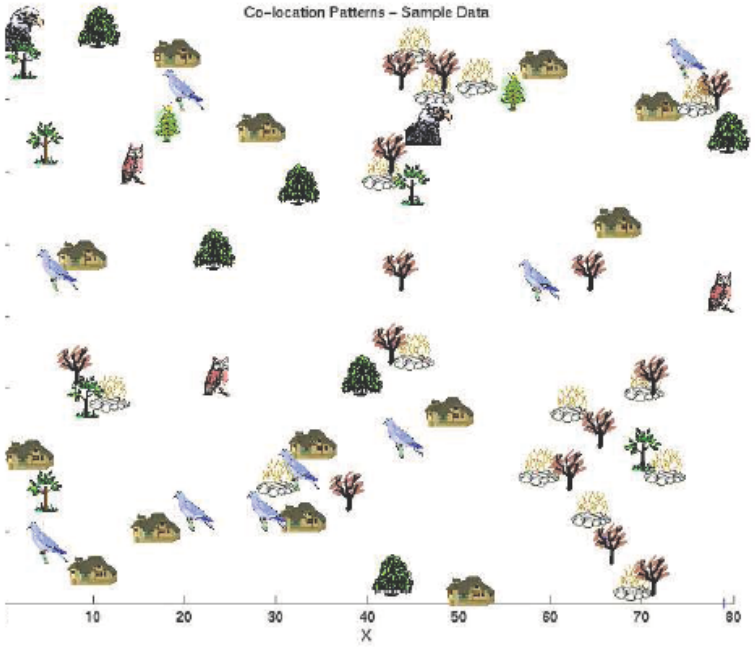
- The 1854 Asiatic Cholera in London

Spatial Clustering



Spatial Co-location Patterns

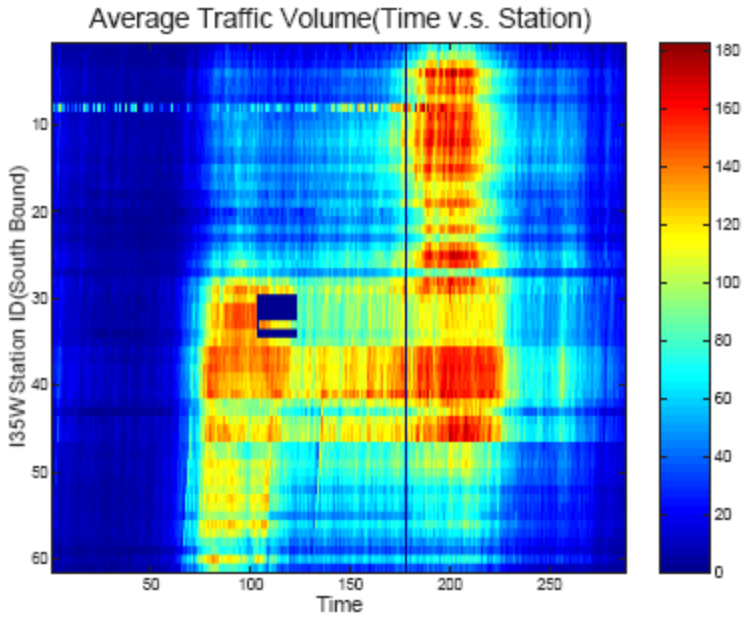
- Given:
  - A collection of different types of spatial events
- Find: Co-located subsets of event types



Answers:   and  

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

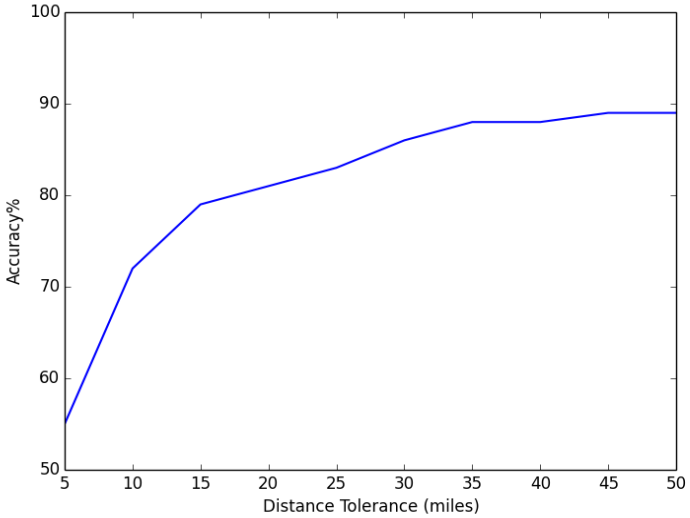
"All it is, is the number pairs, when those calls took place, how long they took place.

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

Photo: CMU Machine Learning Department Protests G20, 2009

Slides: James Hays, Isabelle Guyon, Erik Sudderth, Mark Johnson, Derek Hoiem

Photo: CMU Machine Learning Department Protests G20

Slides: James Hays, Isabelle Guyon, Erik Sudderth, Mark Johnson, Derek Hoiem

National Academy of Sciences Recommendations

- Academic institutions should encourage the development of a basic understanding of data science in all undergraduates.
- Academic institutions should embrace data science as a vital new field that requires specifically tailored instruction delivered through majors and minors in data science as well as the development of a cadre of faculty equipped to teach in this new field.
- As data science programs develop, they should focus on attracting students with varied backgrounds and degrees of preparation and preparing them for success in a variety of careers.

National Academy of Sciences Recommendations

- Ethics is a topic that, given the nature of data science, students should learn and practice throughout their education. Academic institutions should ensure that ethics is woven into the data science curriculum from the beginning and throughout.
- The data science community should adopt a code of ethics; such a code should be affirmed by members of professional societies, included in professional development programs and curricula, and conveyed through educational programs. The code should be reevaluated often in light of new developments.


COUGH! COUGH! YEARS OF DATA MINING HAVE LEFT ME WITH DATA LUNG. DON'T BE LIKE YOUR OLD MAN - GO INTO MODELING OR VISUALIZATION!
CARTOONSTOCK .com
Search ID: jcen1586
© Jon Carter • cartertoons