



Big Data: Paradigm shifts in science and industry

Dr. Sándor Laki

Assistant professor

Dept. Of Information Systems

Eötvös Loránd University

Email: lakis@inf.elte.hu

Web: <http://lakis.web.elte.hu>

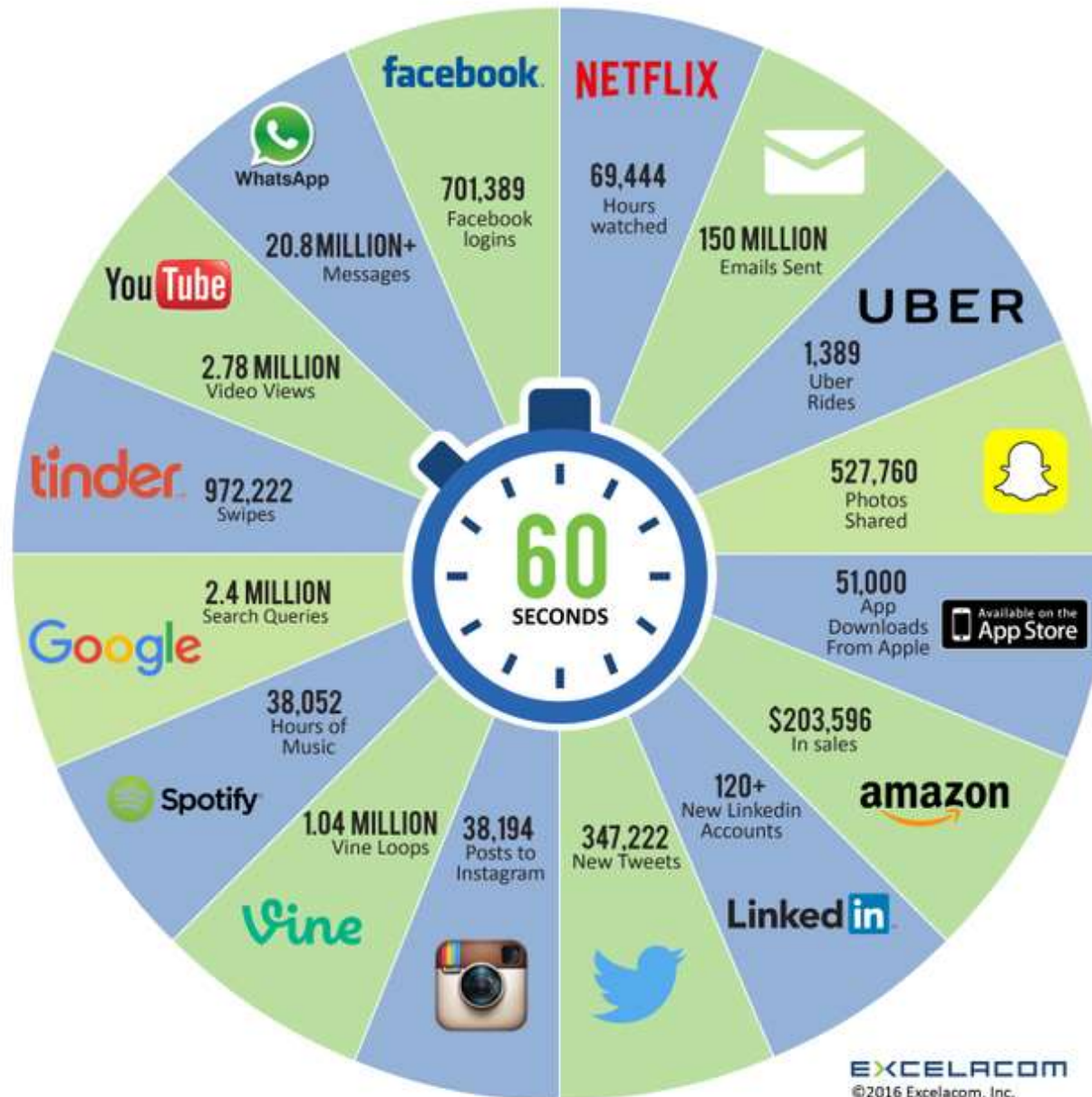
What is Big Data?

Large amount of data

Really?

Does size only matter?

2016 What happens in an INTERNET MINUTE?

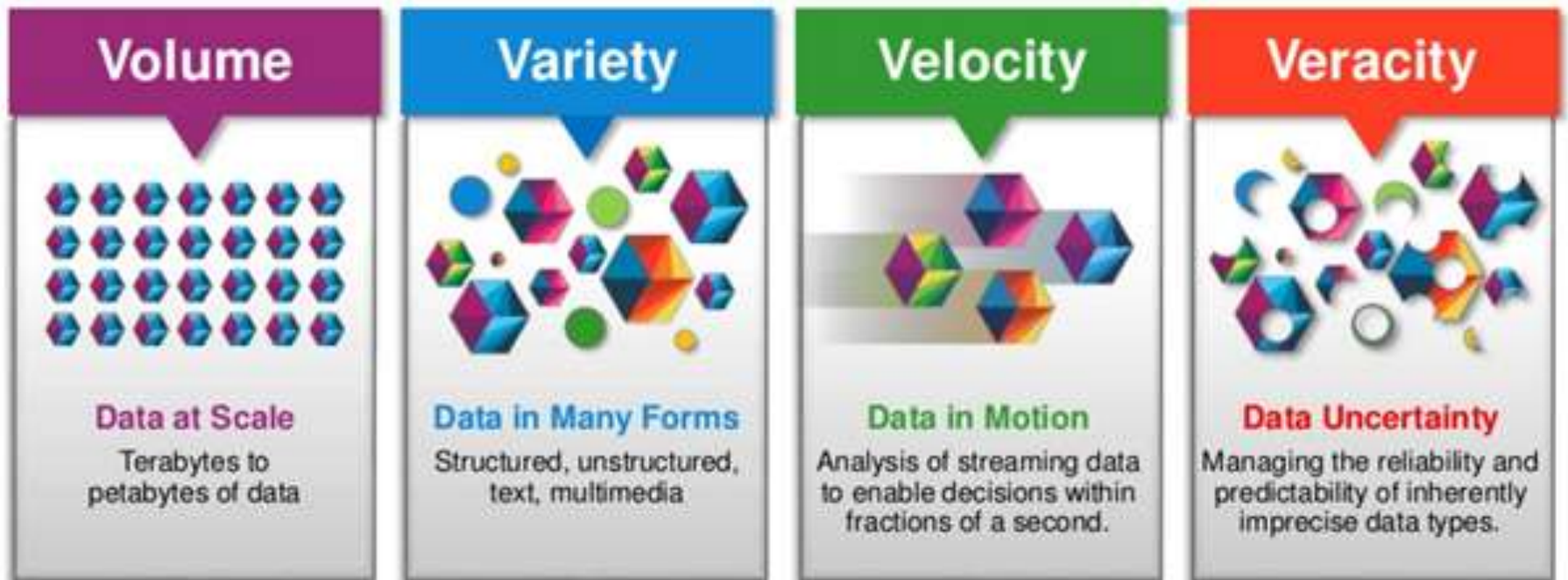


What is Big Data?

One of the definitions

- "Big Data are **high-volume**, **high-velocity**, and/or **high-variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Gartner 2012)
- Complicated (intelligent) analysis of data may make a small data "appear" to be "big"
- Bottom line: Any data that exceeds our current capability of processing can be regarded as "big"

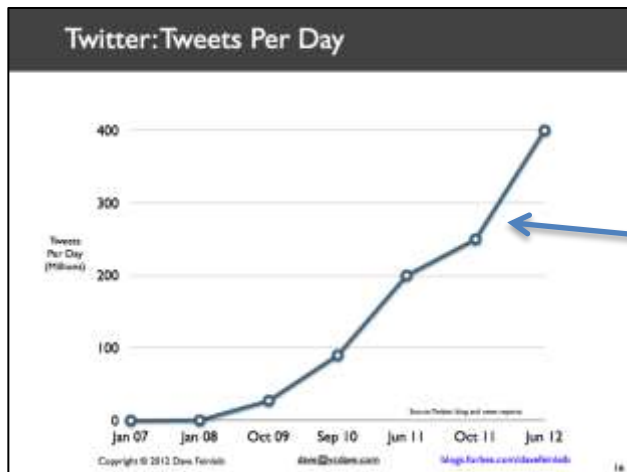
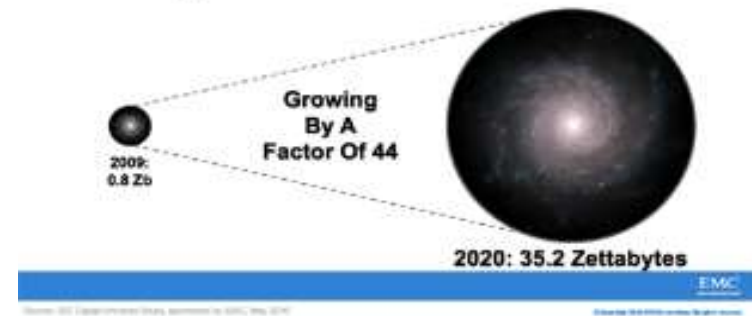
4 Vs of Big Data



1st V: Volume rule them all

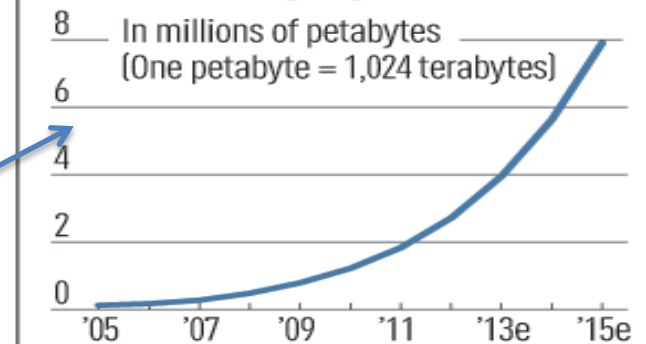
- 30-40 Zettabytes created by 2020
- 2.5 Quintillion bytes of data per day
- Most US companies store 100 Terabytes of data
- 6 billion cell phones worldwide

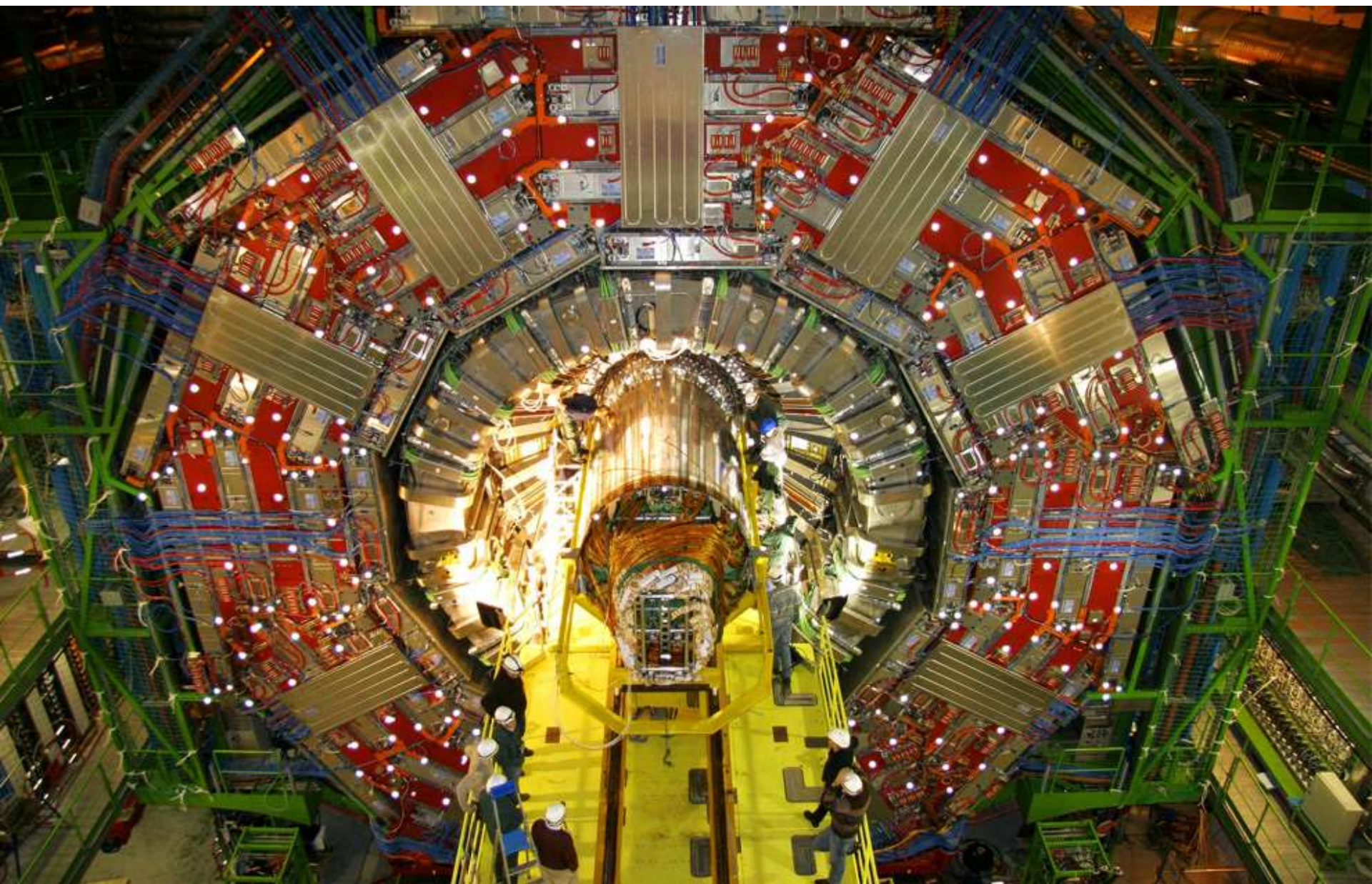
The Digital Universe 2009-2020



Exponential increase in collected/generated data

Data storage growth





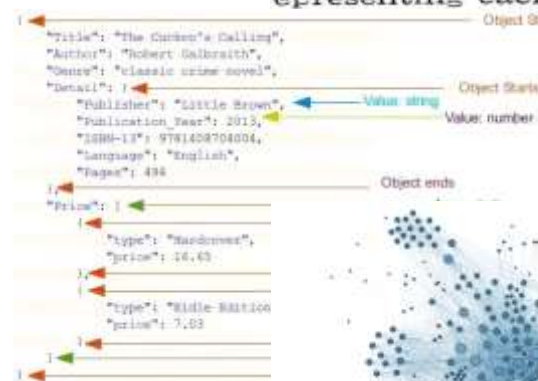
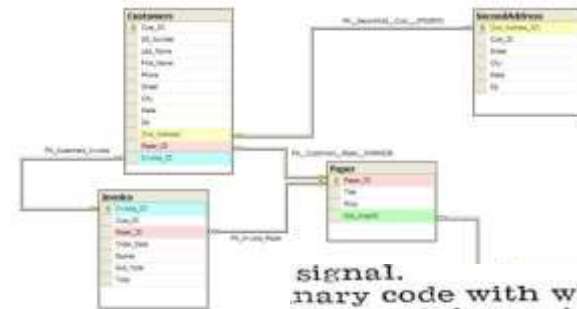
CERN's Large Hadron Collider (LHC) generates 15 PB a year

640K ought to be enough
for anybody.



2nd V: Variety

- **AKA Data Complexity**
- Data in various formats
 - Relational Data
 - Tables, Views, Transactions, etc.
 - Text as unstructured data
 - e.g. web
 - Semi-structured data
 - e.g. XML, JSON
 - Graph data
 - e.g. Social networks, semantic data
 - Streaming data
 - Data can only be scanned once!



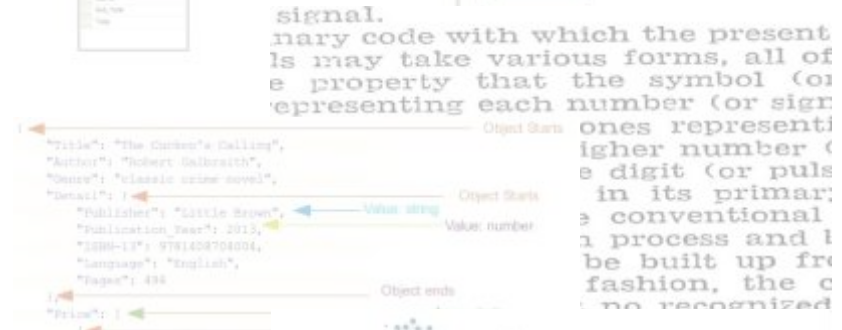
signal.
nary code with which the present
ls may take various forms, all of
e property that the symbol (or
representing each number (or signi
ones representi
higher number (o
e digit (or puls
in its primary
a conventional
a process and l
be built up fro
fashion, the c
no recognized



2nd V: Variety

A single application can be generating/collecting many types of data

- Data in various formats
 - Relational Data
 - Tables, Views, Transactions, etc.
 - Text as unstructured data
 - e.g. web

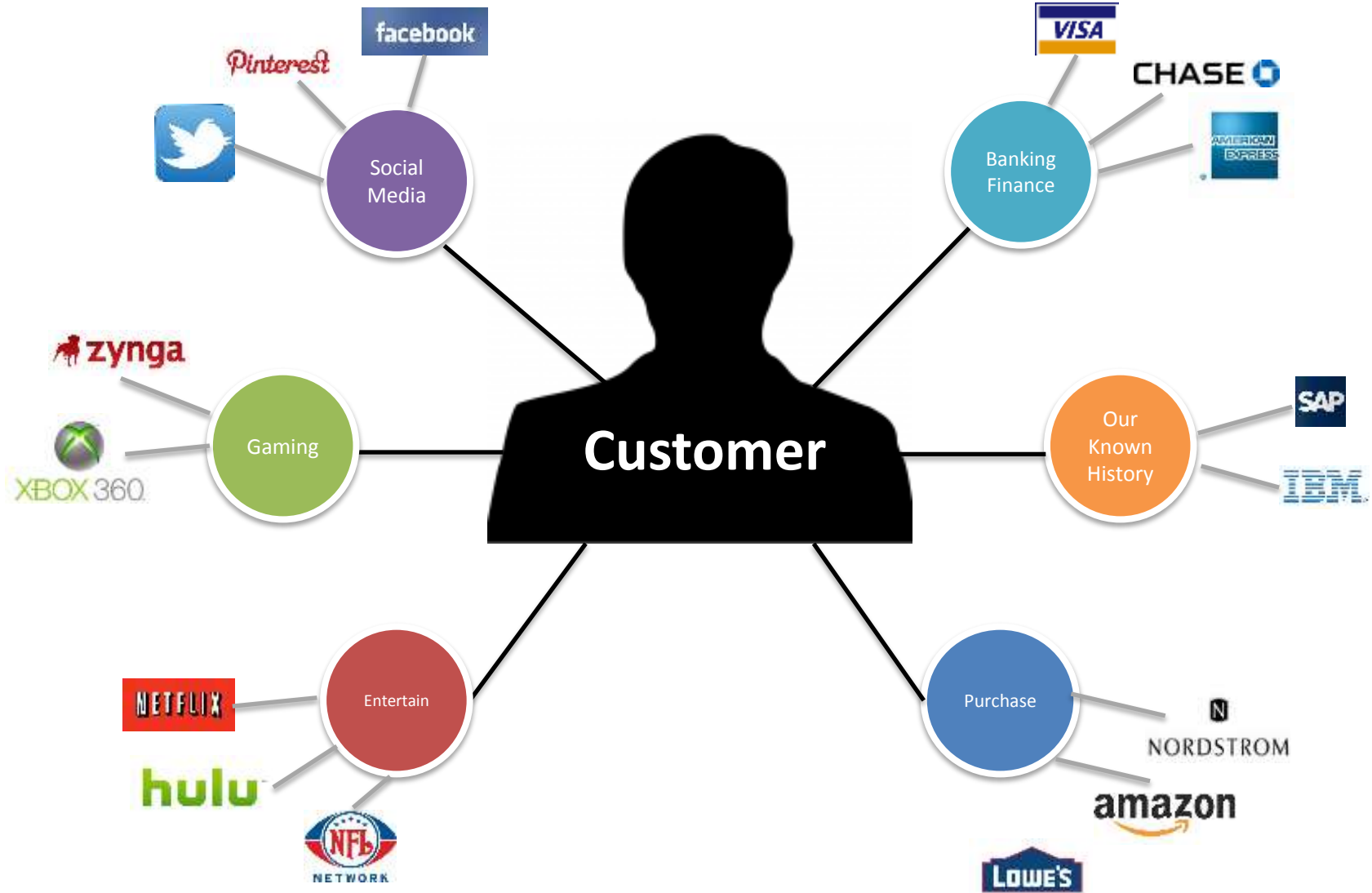


All these types of data need to be linked together to extract knowledge/real value

- Streaming data
 - Data can only be scanned once!



A Single View to the Customer



3rd V: Velocity

- Data is generated fast
 - Modern cars have close to 100 sensors monitoring items with high time resolution
 - NY Stock Exchange captures 1 TB of trade information during each trading session
 - 18.9 billion Internet connection worldwide, subscribers continuously generate data (Twitter, ...)
- (Near) real-time processing is needed
 - Late decision often = missing opportunities
 - Examples
 - **sensors monitoring your activities and body:** any abnormal measurements require immediate reaction
 - Based on your current location, your purchase history, what you like: send **promotions right now for store next to you**
 - In factories lots of sensors may help to **detect faulty items** before they are built in e.g. a car or other product



Real-time/Fast Data



Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



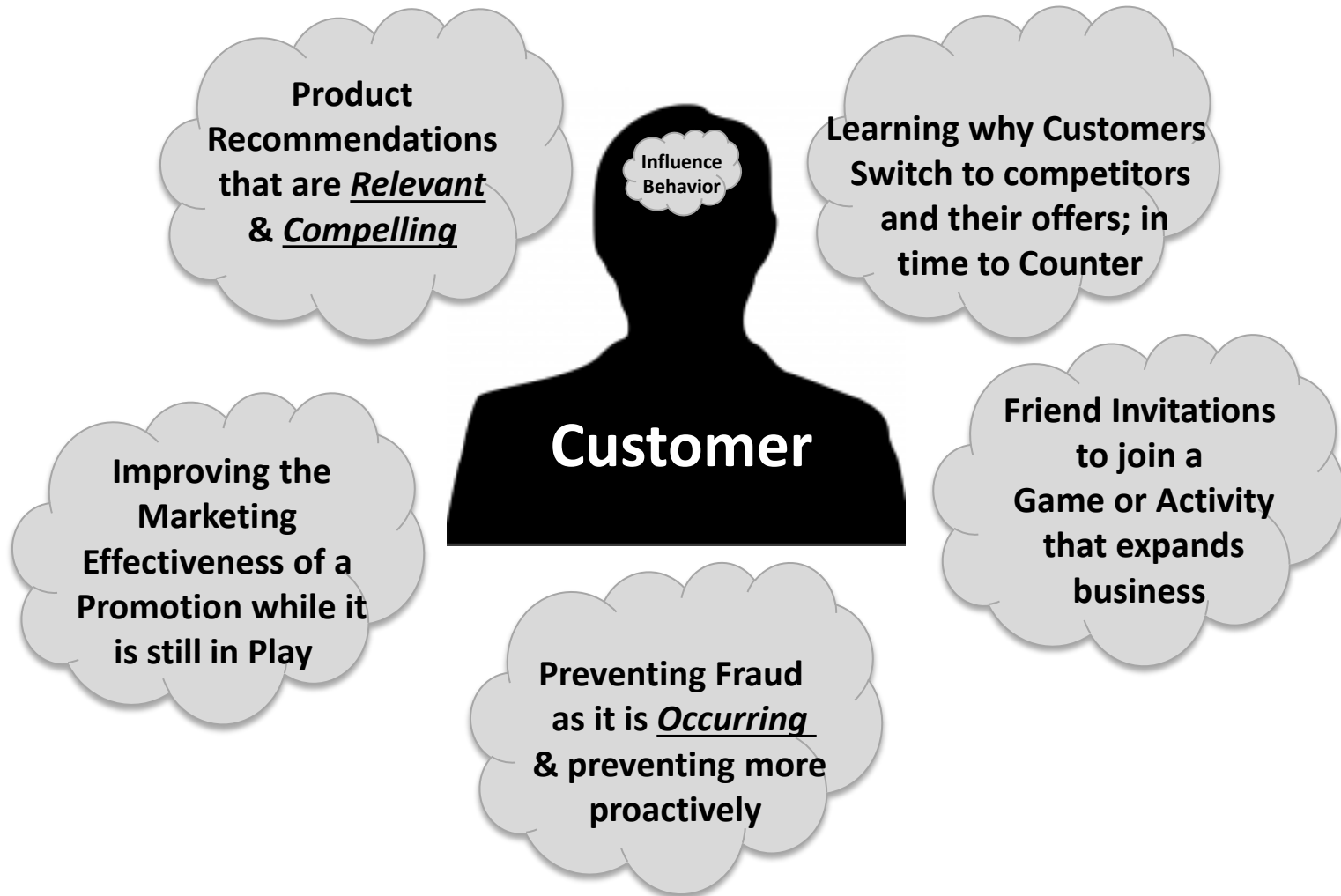
Mobile devices
(tracking all objects all the time)



Sensor technology and networks
(measuring all kinds of data)

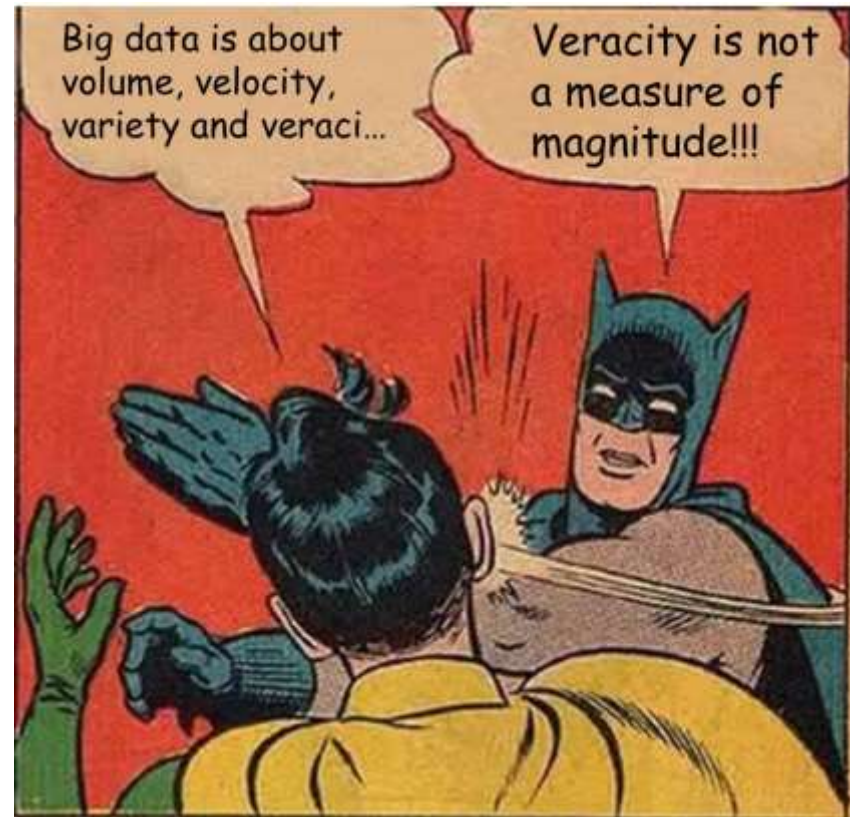
- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Real-Time Analytics/Decision Requirement

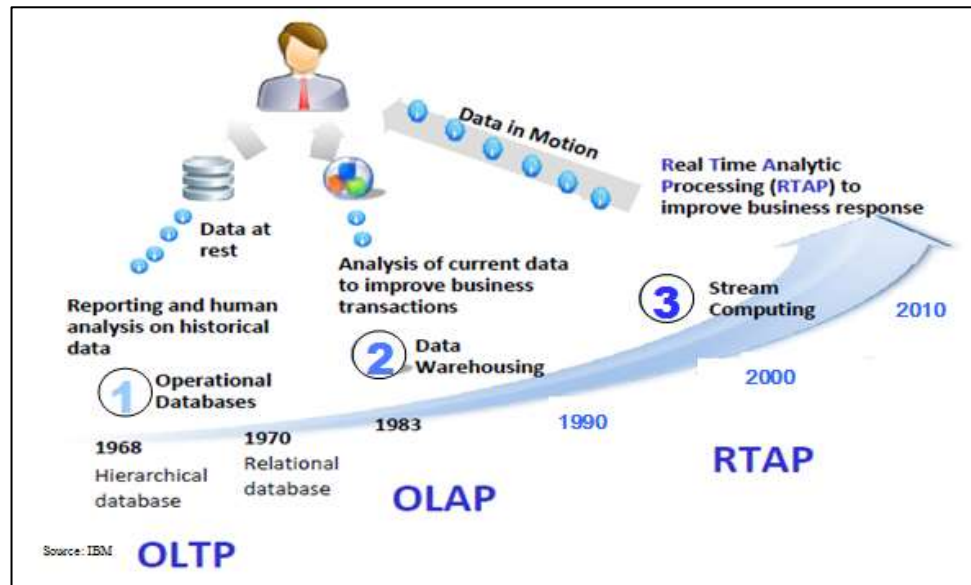


The last V: Veracity

- All about data quality and reliability (IBM 2013)
- 1 in 3 business leaders **do not trust the information** they use to make decision
- In a survey it is always **unsure how many** of their data is **inaccurate**
- **Poor data quality** costs a lot in many sectors



Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

The Model Has Changed...

- **The Model of Generating/Consuming Data has Changed**

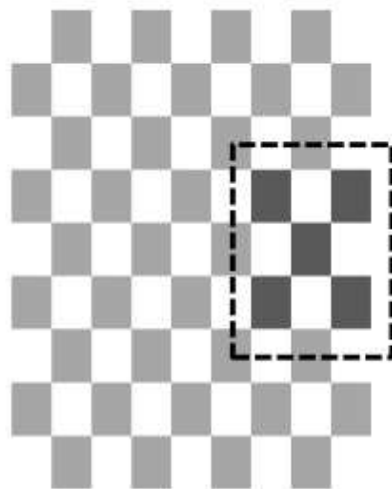
Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



Traditional Approach

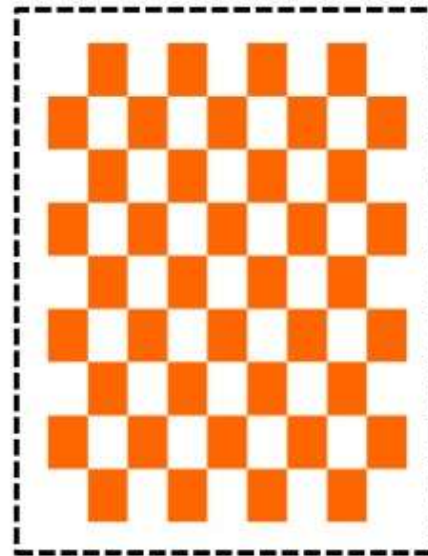


All available
information

Analyzed
information

**Analyze small
subsets of data**

Big Data Approach



All available
information
analyzed

Analyze *all* data



STORM

Big Data Landscape 2016 (Version 3.0)

Infrastructure

Hadoop On-Premise
cloudera, Hadoop, Pivotal, IBM InfoSphere, bluedata, Jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Doble

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon, docker, MESOSPHERE, CoreOS, StackIQ

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, Digital Reasoning, ONSTAIN

Analytics Platforms
Microsoft, guavus, Datasmeer, Bottlenose, Interana

Data Science Platforms
continuum, DataRobot, Alpine, MODE, dataiku, DOMINO, what, ALGORITHMIA

Visualization
tableau, Qlik, looker, Ficox, d3.js, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, hush, machetes, ENGADID

Customer Service
MEDALLIA, ATTENTIVITY, CLARABridge, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, eSour, WISEO

Human Capital
gild, Connect, textile, entelo, hiQ

Legal
RAVEL, JURIMATRA, Everlaw, Brevia, FREEMATION

NoSQL Databases
amazon, Google Cloud Platform, Microsoft Azure, mongoDB, <EROSPIKE>, Couchbase, Sequoia, redislabs, influxdata

NewsSQL Databases
SAP, Clustrix, Pivotal, paradigm, nuodb, splice, MarioDB, VOLTD, citusdata, deepdb, Trafalgar, Cockroach LABS

BI Platforms
Power BI, Amazon, Water Analytics, GoodData, platform, atscale, Qlik

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, cloud physics, loggly

Social Analytics
Hootsuite, NETBASE, DATASIFT, track, bitty, synthasio, similarreach

Ad Optimization
AppNexus, MediaMath, critico, OpenX, Integral, theTradeDesk, Adform, distillery, Livestorm, DMO, DataXu, Applier, MOAT

Security
CYCLONE, CounterTack, AREA1 SECURITY, Recorded Future, FORTSCALE, sift science, Yagham, feedzai, SCNFYS

Vertical AI Applications
facebook, X, Clara, KASIST, lumina

Graph Databases
neo4j, OrientDB, IntelLig

MPP Databases
TERADATA, VERTICA, Netezza, C3tion, cognito, SQL, aremo

Cloud EDW
amazon, Microsoft Azure, Pivotal, Snowflake, Oracle, Infoworks

Data Transformation
alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

Real-Time
amazon, HEMARRETS, striim, confluent, Dandelion, dataArtisans

Machine Learning
H2O, Databricks, SKYYTREE, deep learning, PredictIO, gubler

Speech & NLP
NarrativeScience, NUANCE, semantic, ARRIA, nora, darliff, IDIBON, vici

Horizontal AI
IBM Watson, Cortana, VIV, sentiment, nora, Numera, darliff, IDIBON, vici

Publisher Tools
Outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

Govt / Regulation
Socrata, OPENGOV, EN, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, LendUp, Kabbage, tidemark, INSIGHT, UORO, Datamint, Lenddo, KENSHC, AIDYA, ISENTIUM, Quantopian, sentiment

Management / Monitoring
New Relic, AFDYNAMICS, actifio, Numerify, splunk, Pivotal, Troceno, Tivoli, Arcus

Security
Tanium, Illumio, CODE42, DataGravity, CipherCloud, VECTRA, Splunk, BlueTape

Storage
amazon, Microsoft Azure, Pivotal, NetScout, COHO, Qumulo

App Dev
apigee, CRK, Typesafe, DRIVEN

Crowd-sourcing
amazon, Amazon Mechanical Turk, CrowdFlower, WorkFusion

Search
hp, ORACLE, ELK, Lucidworks, elastic, ThoughtWorks, MAANA, swifttype, Algolia, RINGUA

Data Services
LICO, OPERA, EXL, kaggle, DataKind

For Business Analysts
OriginalLogic, ClearStory, CIRRO, Import IO

Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPUS, granify, sumail, Airtake, retention, custora

Education / Learning
KNEWTON, Clever, Geclara, PANORAMA, knowre

Life Sciences
23andMe, Coursera, Recombine, KYRUS, oregonmyer, HealthTap, METABOTA, ZEPHYRUS, Gingerio, transcript, Glow, enlitic, AICure, Accuweather

Industries
OPower, eHarmony, duetto, STITCH FIX, Workfusion, TACHYON, Seeq, FarmLogs, select, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon, Google, Microsoft, IBM, SAP, sas, H2O, hp, Amazon, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
Hadoop, YARN, Spark, MESOS, TEZ, Flink, COAP

Query / Data Flow
SLAMDATA, DRILL, Google Cloud Dataflow

Data Access
HARISE, mongoDB, cassandra, CouchDB, riak, ScyllaDB, nifi

Coordination
talend, Apache Zookeeper, Apache Airflow

Real-Time
STORM, Spark, Flink, TACHYON, druid

Stat Tools
R, ScalaLab, SciPy

Machine Learning
mlilb, Aerosolve, Caffe, FeatureFu, DIMSUM, DLAI

Search
elasticsearch, Solr

Security
Apache Ranger, Visualization

Data Sources & APIs

Health
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, samsara

Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, Y-DEE, PREMISE, S&P CAPITAL IQ, Quandl, xignite, CB Insights, StockTwits, Glimster, PLATO

Air / Space / Sea
PLANET, spire, AIRWARE, DroneDeploy

Location / People / Entities
axiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, CIRCULATE, placemeter, BASIS, Sense

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, PLURALIGHT, DataCamp, DataElite, the Data Incubator, METIS

Science



The LSST Survey

8.4 meter, wide-field, f/1.2 telescope.

3.2 Gigapixel, 189 4k x 4k CCD camera, with 2-second readout.

PetaFLOPS of computing power, hundreds of PB of storage, gigabit long-haul networks.

Turning the sky into a database.

~18TBytes/night



The LSST Survey

8.4 meter, wide-field, f/1.2 telescope.

3.2 Gigapixel, 189 4k x 4k CCD camera, with 2-second readout.

PetaFLOPS of computing power, hundreds of PB of storage, gigabit long-haul networks.

Turning the sky into a database.

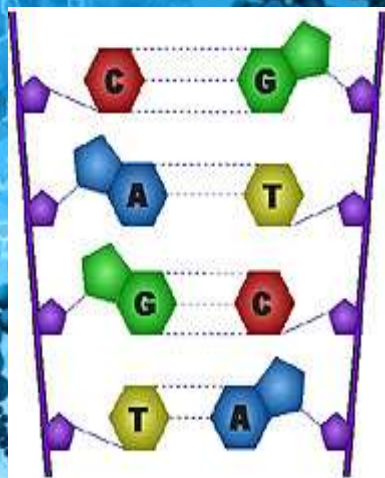
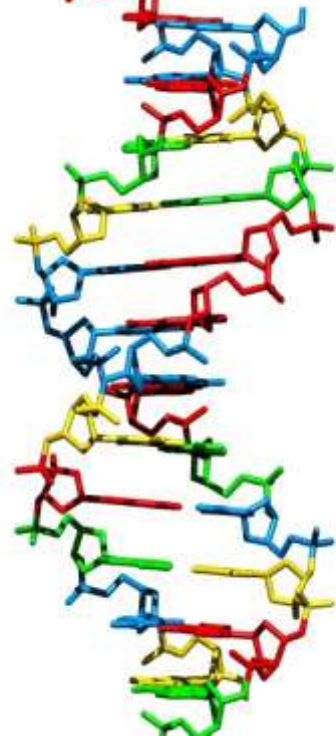
~18TBytes/night



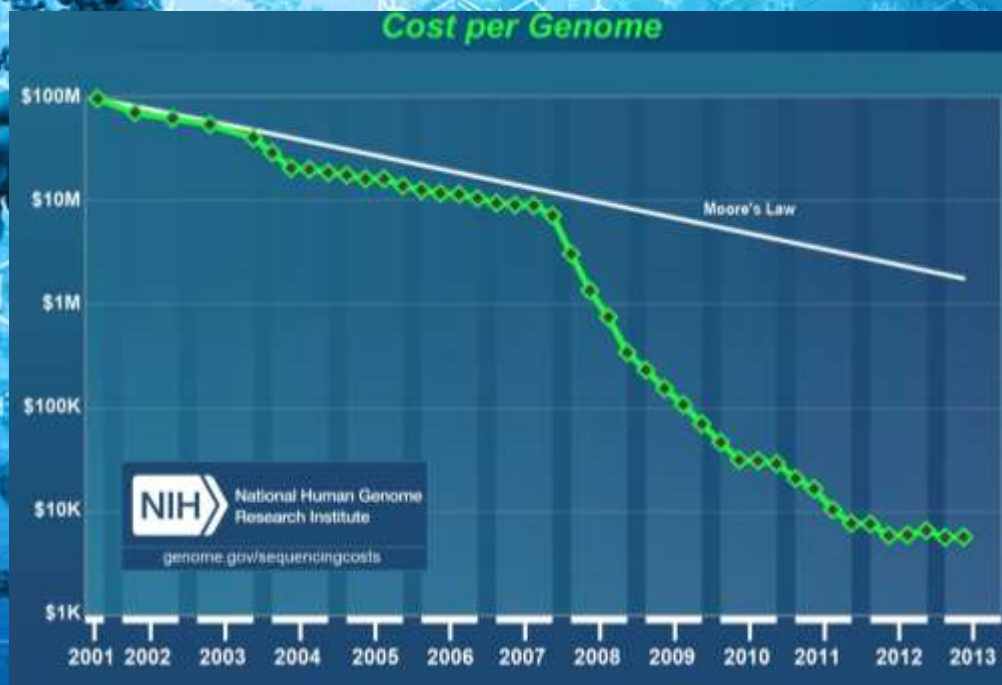
SKA Project





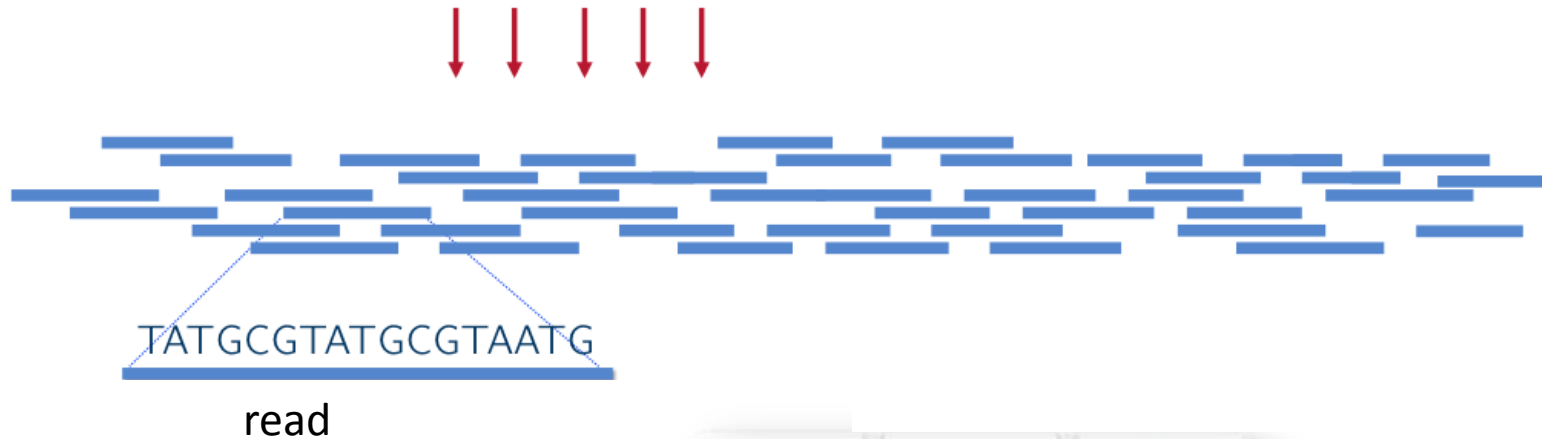


...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAATATT...

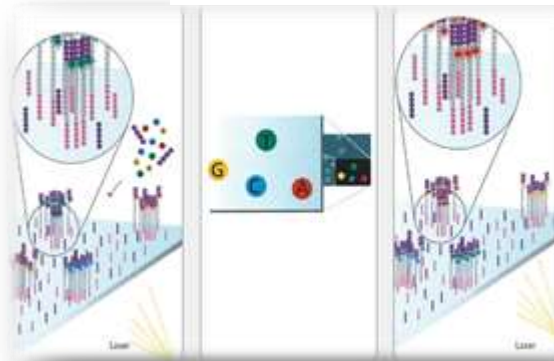


Shotgun sequencing

ACGTCCTATGCGTATGCGTAATGCCACATATTGCTATGCGTAATGCGTACC

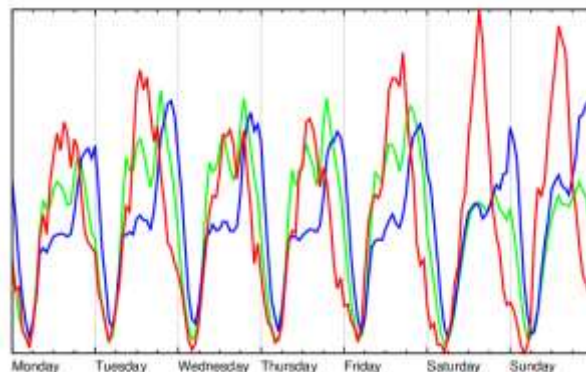
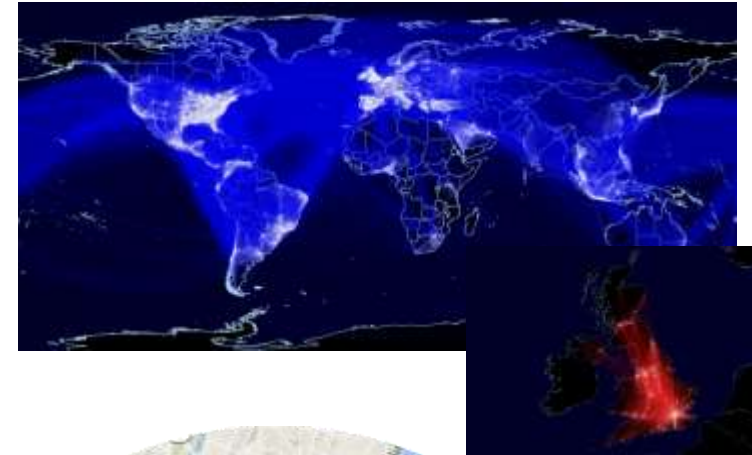


~ 100-1000 M reads / run
~ 1GBytes -1TBytes / run



Analysis of social networks

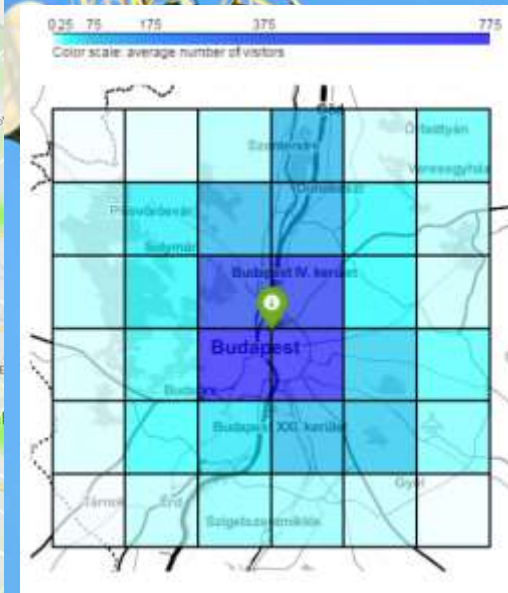
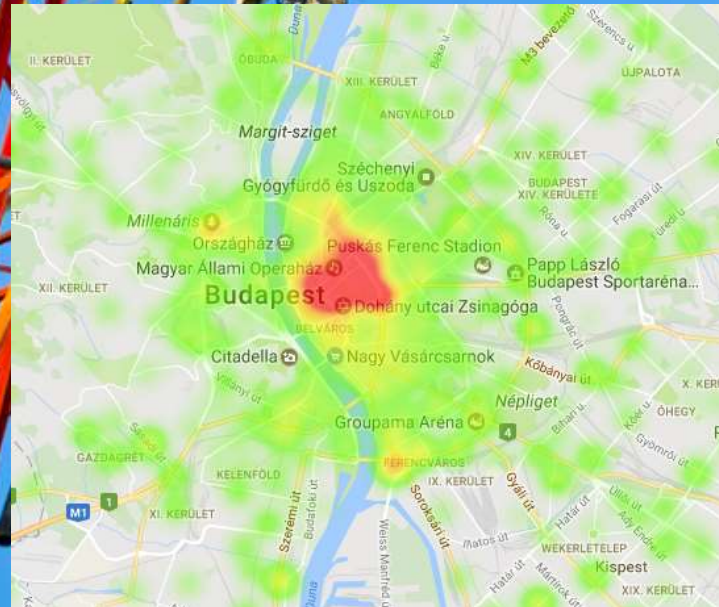
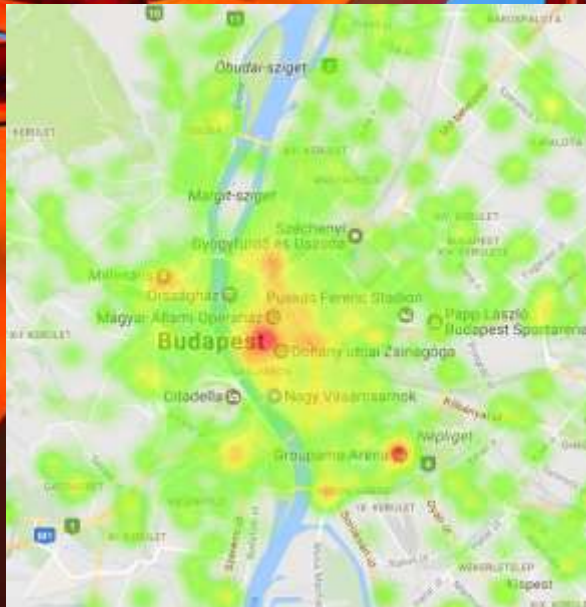
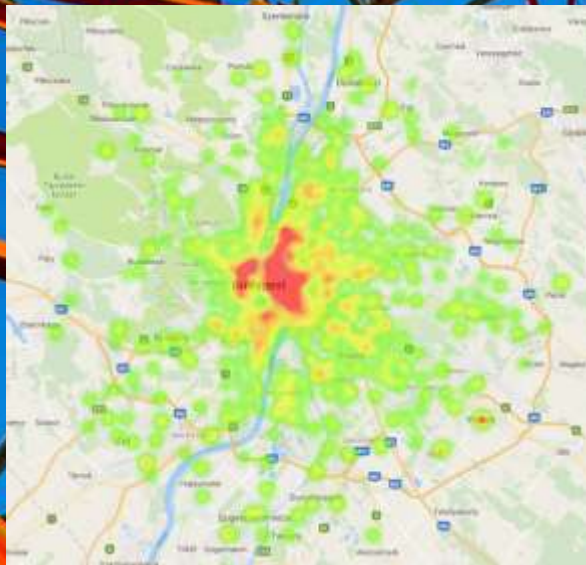
- Sentiment analysis of Twitter data, e.g. for marketing purposes
- Real-time event detection in social data
- Guessing residence and movement patterns based on tweets and friendships
- Information propagation in social and mobile networks
- Human behaviour analysis based on aggregated cellular network data
- Twitter data collection infrastructure



Industry



6 Billion CDRs per day at a large telco



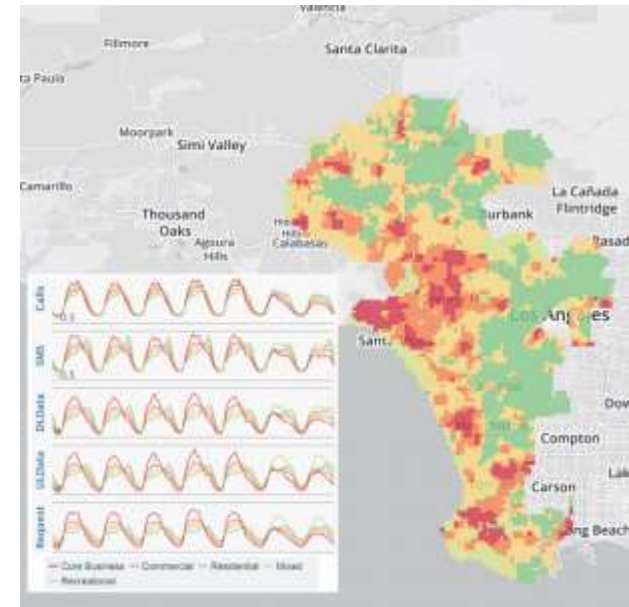
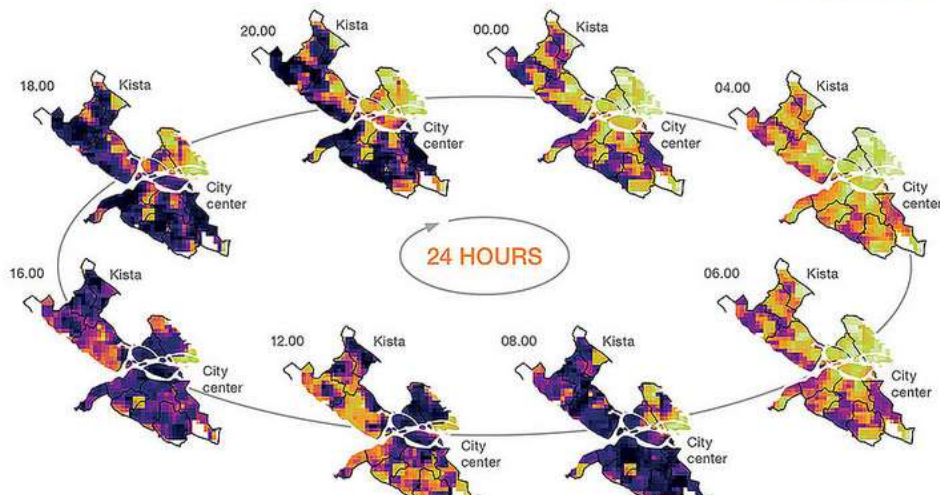
Analysis of social & mobile networks

- Mobile phone activity
- Activity patterns of spatial pixels
 - Revealing the functional spatial clusters of city life
 - Identification of outliers
- Calls, SMS or data traffic + Twitter data
- On going joint research with **Ericsson** and **MIT SENSEable City Lab** (Carlo Ratti)



Daily commuting reflected in mobile traffic density in Stockholm

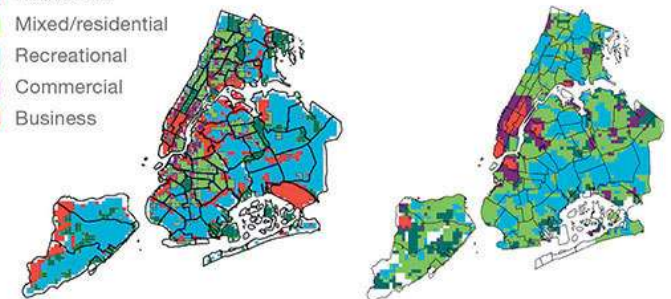
Low High



Comparing land use and mobile traffic patterns

New York

- Residential
- Mixed/residential
- Recreational
- Commercial
- Business



Official census-based land use map²

Traffic-based land use map

Smart Factories

- Root cause analysis
 - In old factories -> manual
 - Let's do it smarter -> automatic data mining solutions
- Really big data
 - Sensors monitoring the whole production chain
 - Enterprise data about external vendors and suppliers
- Industry 4.0



BIG DATA in banking

THE BANKING BUSINESS FINDS A NEW ASSET



Typical banking sources of **BIG DATA** include



Customer bank visits



Call logs



Web interactions



Credit card histories



Social media



Transaction types



Banking volumes

How banks put **BIG DATA** to work



Customer risk assessment



Anti-money laundering procedures and fraud detection



Compliance and regulatory reporting



Customer relationship management



Stock trade surveillance and pattern analysis

BIG DATA problem solving for financial institutions

Insights for Product Development



The costs of developing new products and services can be staggering

SOLUTION

Customer transactional data such as timing of visits and duration of teller transactions can be analyzed to find gaps in product offerings

Scoring Credit Risks



Banks need to cut lending risks while improving customer marketing

SOLUTION

Consumer payment patterns and law enforcement databases supply the data

credit bureaus store over
800 billion records
to be sliced, diced and analyzed for more
accurate credit risk scores.

For comparison, the FBI's Investigative Data Warehouse has only 1.5 billion documents.



www.alacergroup.com

The applications of Big Data (Network Architecture) for organisations and businesses

Proximity to end-users: Front-end to Back-end

Front-end:
Usage of the information



Monitoring : control, home automation, security



Wearable technologies: virtual or augmented reality



Social network: Facebook, Twitter, Amazon, etc.

Middle-end:
Optimisation of the existing solutions



Process: « Smart Cities », default and failure detection, etc.



Improvement of the experience: customer (marketing), sport or health, etc.



Targeting: Marketing, Risk, Fraud, etc.

Back-end:
Conception of new solutions



Flow analysis, energy consumption, etc.



Behavior Analysis (geolocation, body indicators, etc.)



Emerging **needs** detection via text-mining

Copyright Nicolas Gladys

Machine to Machine

Human to Machine

Human to Human

The Network: from the Machine to the Human interactions



What is Data Science?

1935: "The Design of Experiments"

R.A. Fisher



1939: "Quality Control"

W.E. Deming



1958: "A Business Intelligence System"



Peter Luhn

1997: "Machine Learning"

1977: "Exploratory Data Analysis"



1989: "Business Intelligence"

Howard
Dresner



2010: "The Data Deluge"

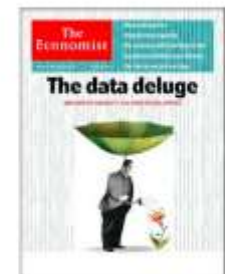
1996: Google



2007: "The Fourth Paradigm"



2009: "The Unreasonable Effectiveness of Data"

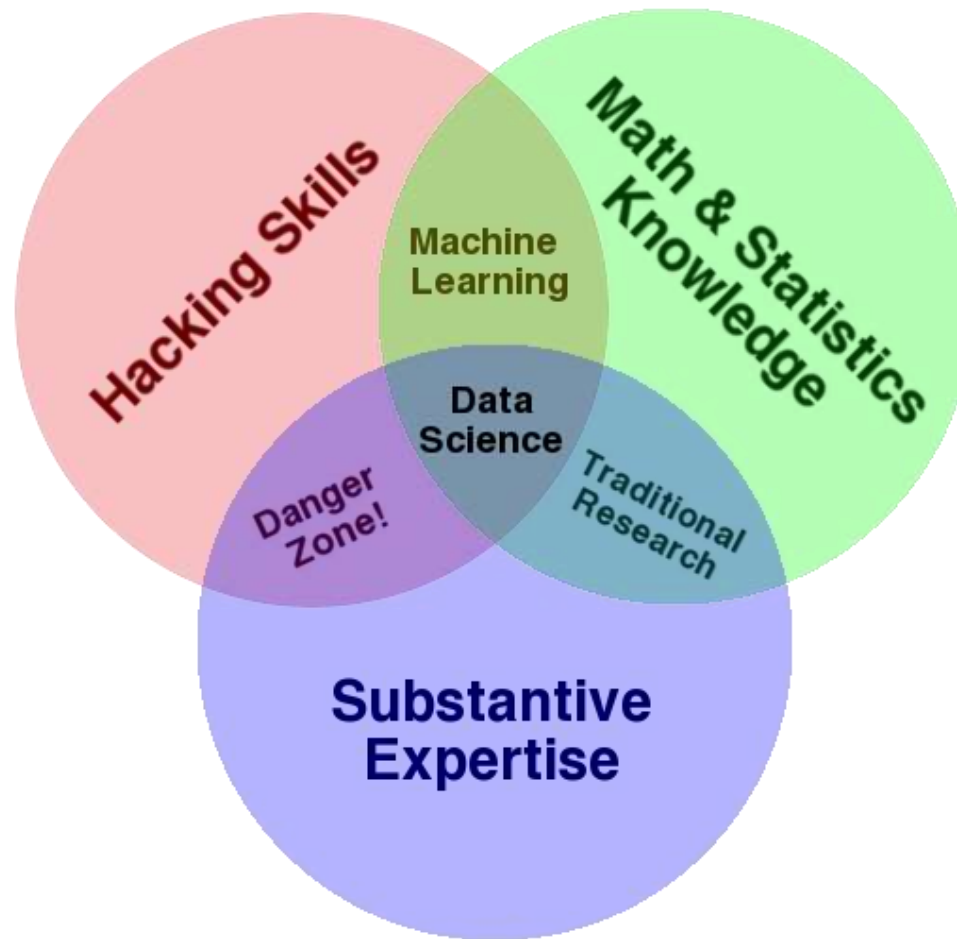


Why do we call it data science?

Data Science is

- an **emerging field**, belonging to applied research.
- the science which uses **computer science, statistics and machine learning, visualization and human-computer interactions** to **collect, clean, integrate, analyze, visualize, interact** with data to **create data products**.
- Models for/based on **large-amount of data**

It is a mixture of various skills



Thank you!



Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.



Note: Data do not include managers Source: Burtch Works

The Wall Street Journal