Introduction to Big Data Algorithms

大数据算法引论

Edison Xin Bi (毕鑫)
Sino-Dutch Biomedical and Information Engineering School
Northeastern University

Course homepage: http://edijason.github.io/courses/course-BDA1718.html

Chapter 0

Preface

Edison Xin Bi

- Ph.D., Computer Science, NEU, 2016.7

- Research Interests
  - Semi-structured data management
  - Data mining
  - Large-scale medical and healthcare data analysis
  - Deep learning

- Background in big data and cloud computing
  - 4/6 NSFC grants
    - 云环境下社交空间关键字查询处理与优化技术研究
    - Spark环境下LBSN大数据管理与分析关键技术研究
    - 云计算环境下海量XML数据管理关键技术研究
    - **面向医疗健康大数据的半结构化数据管理关键技术研究**
  - 4/11 SCI papers
    - Distributed query processing
    - Distributed mining

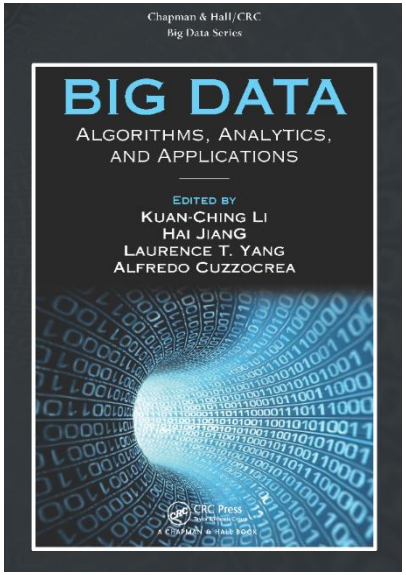
Edison Xin Bi

- Contacts
  - **Office**: B220, Shengming Building, Hunnan Campus, NEU
  - **E-mail**: bixin@bmie.neu.edu.cn
  - **Homepage**
    - Personal: http://edijason.github.io/
    - Google Scholar: https://scholar.google.com/citations?user=LsQXMWAAAAAJ&hl=en
    - NEU Faculty: http://faculty.neu.edu.cn/bmie/bixin/

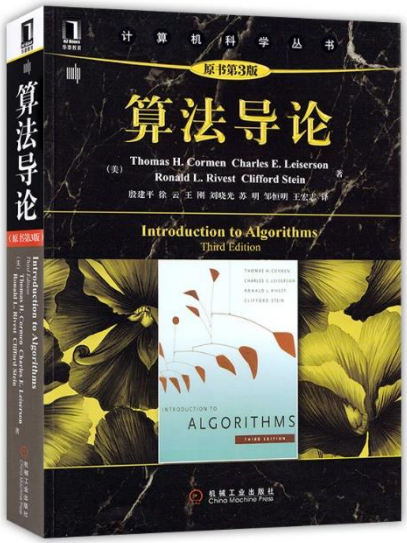
Course Goals

- **Industrial** background of big data and cloud computing
- **Academic** background of data science
- Big data in **biomedical engineering**
- **Distributed algorithms design** in the cloud
- Distributed **query processing** algorithms
- Distributed **data mining and analysis** algorithms

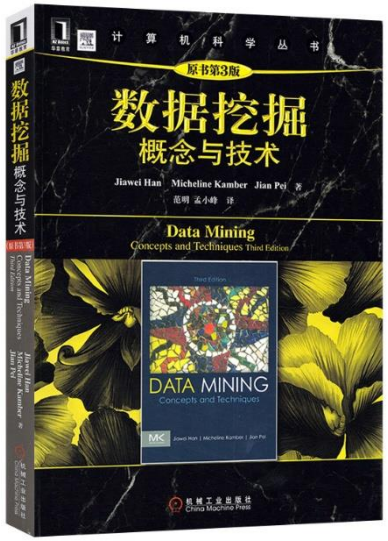
Course References



Big data: algorithms, analytics, and applications.
By Kuan-Ching Li, *et al*.

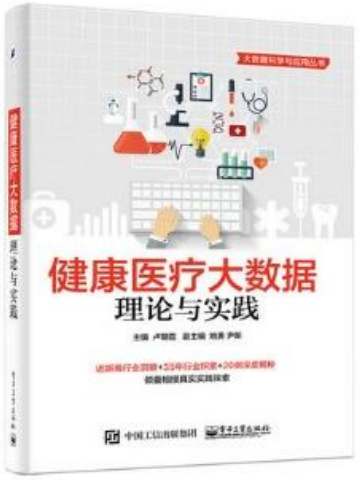


Introduction to Algorithms (3rd Edition)
By Thomas H. Cormen, *et al*.

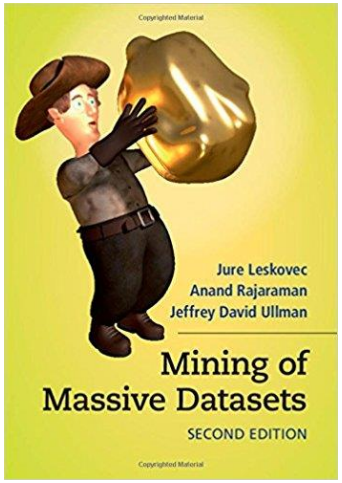


Data Mining: Concepts and Techniques (3rd Edition)
By Jiawei Han, *et al*.

Course References



健康医疗大数据：理论与实践
By 卢朝霞，等



Mining of Massive Datasets
By Jure Leskovec, *et al*.

Academic papers

Course Outline

1. Intro to Big Data
2. Intro to Cloud Computing
3. Distributed Algorithms based on MapReduce
4. Distributed Algorithms based on MapReduce - Applications
5. Distributed Algorithms beyond MapReduce
6. External Memory Structures
7. External Memory Algorithms
8. Tree Algorithms - Keyword Query
9. Tree Algorithms - Twig Query
10. Graph Algorithms
11. Big Data Mining - Intro
12. Big Data Mining - Classification
13. Big Data Mining - Clustering
14. Big Data Mining - Applications
15. Recommendation System
16. Outsourcing Algorithms

Course Info

- Marking Scheme
  - Float grades: 10% (according to the attendance rate, activity, *etc*.)
  - 2 in-class quizzes: 10% each (20% in total)
  - 3 Assignments: 10% each (30% in total)
  - Final Test: 40%
- Assignments (may be updated)
  - MapReduce Algorithms
  - Tree and Graph Algorithms
  - Mining Algorithms

Course Requirements

- No classroom disruptions.
- Think independently, and do your assignments **independently**.
- ***PLAGIARISM*** is equal to **FAIL** grade, no excuses, no exceptions.
- The more **active** you are during the class, the higher floating score you will get.
- All the assignments are handed in via *Blackboard Learn* platform.
- **Late submission** leads to grade penalty.

Academic foundation: Conferences

- Associations
  - **ACM**: Association for Computing Machinery
  - **IEEE**: Institute of Electrical and Electronics Engineers
  - **CCF**: China Computer Federation
- Top confs in Database
  - **SIGMOD**: ACM Special Interest Group on Management of Data
  - **VLDB**: Very Large Data Base
  - **ICDE**: IEEE International Conference on Data Engineering
  - **CIKM**: ACM International Conference on Information and Knowledge Management
  - **DASFAA**: International Conference on Database Systems for Advanced Applications
- Confs in other fields
  - **SIGIR**: ACM Special Interest Group on Information Retrieval
  - **SIGKDD**: ACM Special Interest Group on Knowledge Discovery and Data Mining
  - **NIPS**: Conference on Neural Information Processing Systems
  - **ICML**: International Conference on Machine Learning
  - **AAAI**: Association for the Advancement of Artificial Intelligence
  - **OSDI**: USENIX Symposium on Operating Systems Design and Implementation

Academic Foundation: Journals

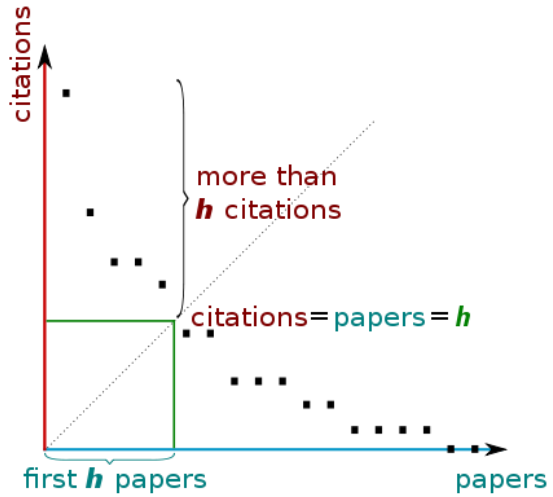
- Top journals
  - **TOCS**: ACM Transactions on Computer Systems (ACM)
  - **TOC**: IEEE Transactions on Computers (IEEE)
  - **TODS**: ACM Transactions on Database Systems (ACM)
  - **TOIS**: ACM Transactions on Information and Systems (ACM)
  - **TKDE**: IEEE Transactions on Knowledge and Data Engineering (IEEE)
  - **VLDBJ**: VLDB Journal (Springer-Verlag)
  - **TKDD**: ACM Transactions on Knowledge Discovery from Data (ACM)
  - **DKE**: Data and Knowledge Engineering (Elsevier)
  - **AI**: Artificial Intelligence (Elsevier)

Academic Foundation: Evaluation

- Criteria
  - **Impact factor** (for journals)

$$\mathrm{IF}_y = \frac{\mathrm{Citations}_{y-1} + \mathrm{Citations}_{y-2}}{\mathrm{Publications}_{y-1} + \mathrm{Publications}_{y-2}}$$

  - **h-index** (for scholars)



- Index
  - **EI**: Engineering Index
  - **SCI/SSCI**: (Social) Science Citation Index
  - **ISTP**: Index to Scientific & Technical Proceedings
  - 中文核心期刊
- Rankings
  - **CCF** Ranks (A, B, C)
  - **JCR** (Journal Citation Reports)
    - Chinese Academy of Sciences
    - Thomson Reuters

Academic Foundation: Methods

- Full text papers
  - Google Scholar
  - DBLP
  - ScienceDirect
  - ACM Digital Library
  - IEEE Xplore
  - ResearchGate
- Ability Improvement
  - Github
  - Slideshare
  - Coursera, edX
  - Kaggle, Topcoder

Acknowledgement

- Special thanks to
  - My Ph.D. supervisor: Professor **Guoren Wang** (CSE)
  - My secondary supervisor: Associate professor **Xiangguo Zhao** (CSE)
  - My colleagues/teaching team:
    - Professor **He Ma** (BMIE)
    - Associate professor **Xiaoyu Cui** (BMIE)

- References
  - Google
  - Wikipedia
  - Related books
  - Web sources footnoted in each slice

Chapter 1

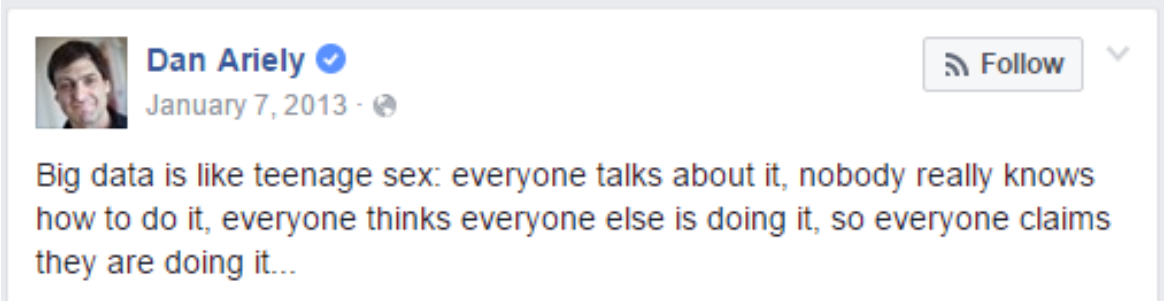
Intro to Big Data

Chapter Outline

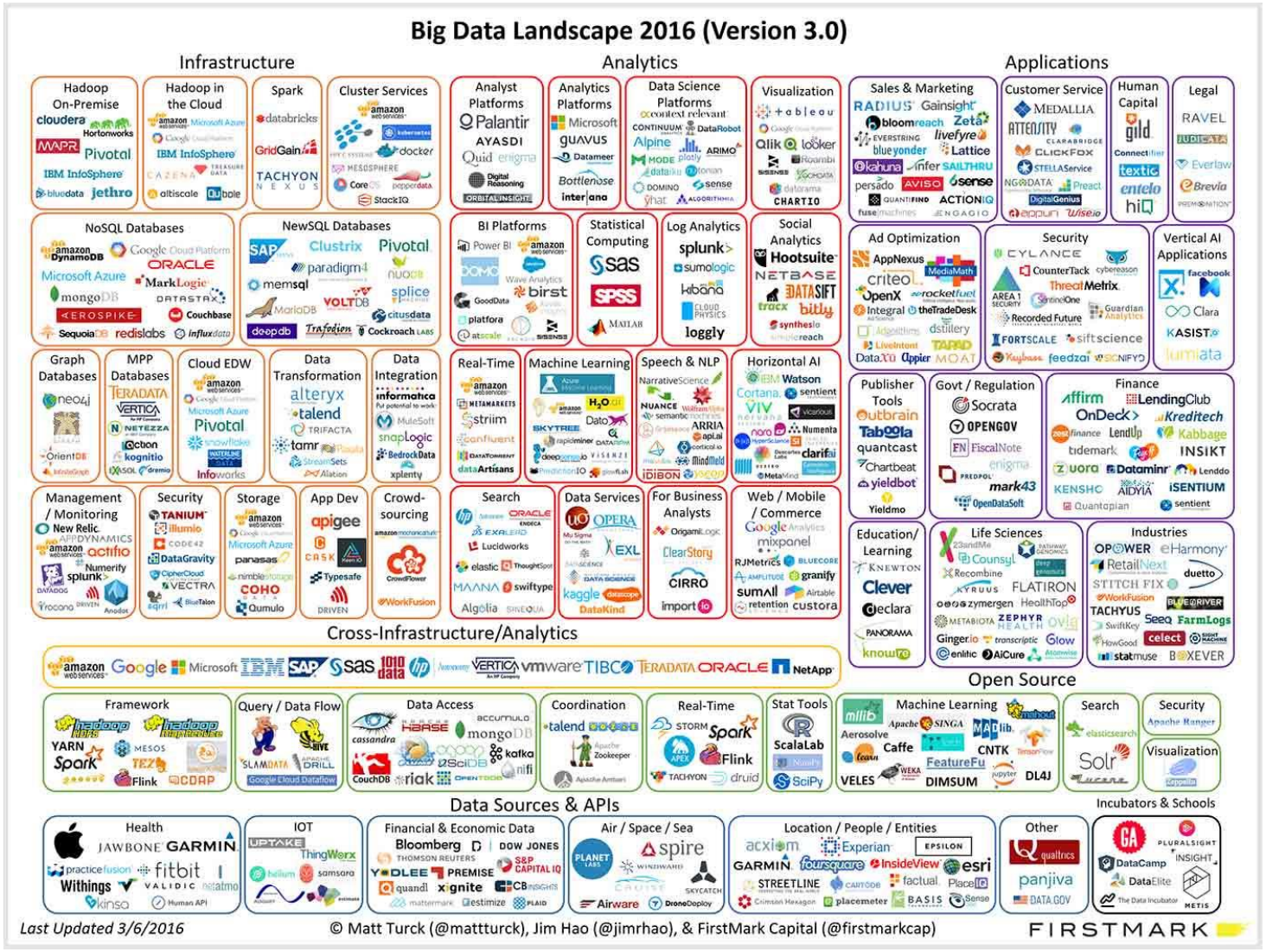
1. What is big data
2. Big data in biomedical engineering
3. Cloud computing

What is big data?

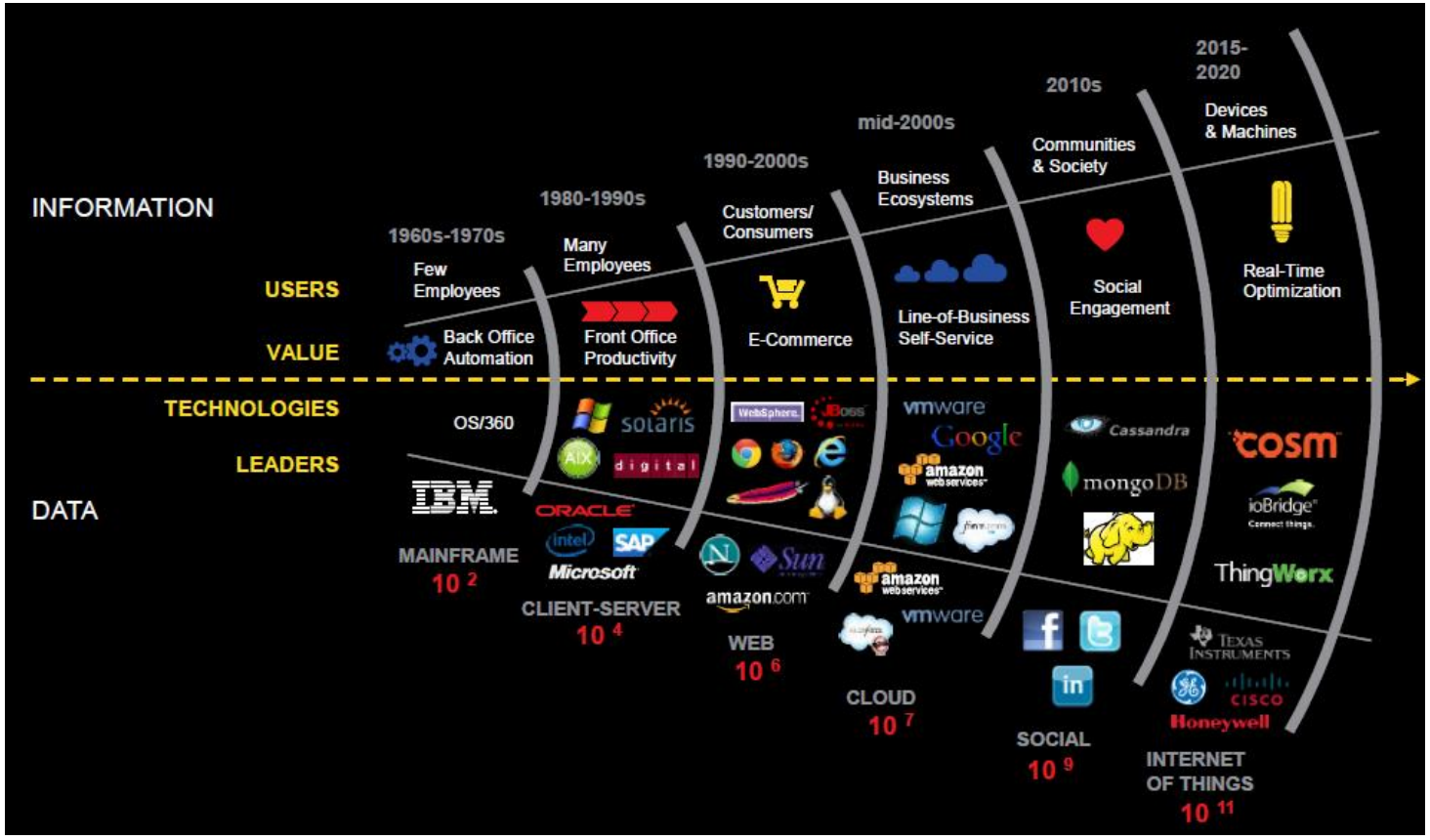
What is big data



What is big data



What is big data



What is big data

- How big is big data
  - 1 Byte =

What is big data

- How big is big data

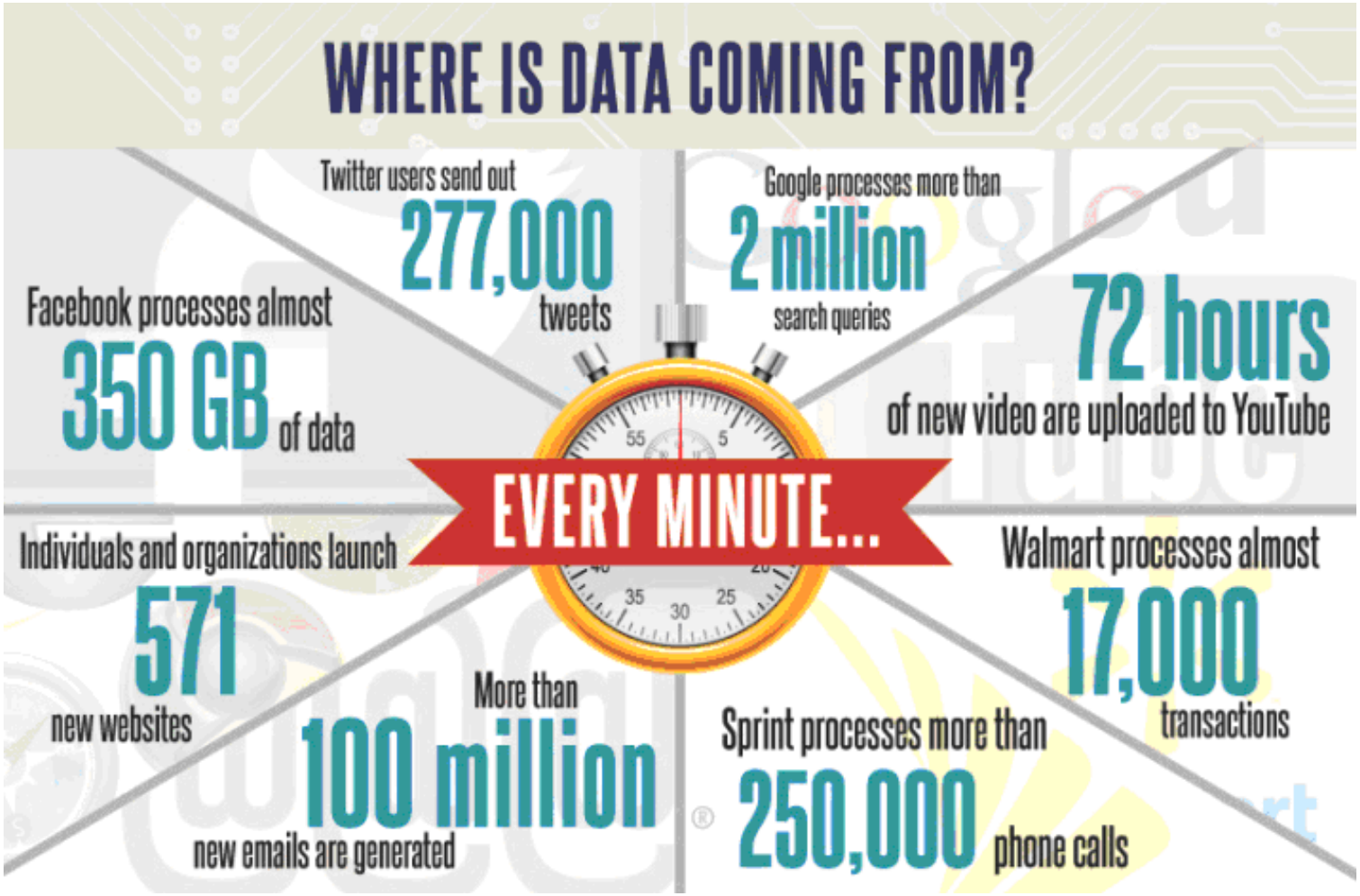
仟 10^{3}
萬 10^{4}
億 10^{8}
兆 10^{12}
京 10^{16}
垓 10^{20}
秭 10^{24}
穰 10^{28}
沟 10^{32}
涧 10^{36}
正 10^{40}
载 10^{44}

俱胝 10^{7}
无我 10^{10}
阿庾多 10^{14}
阿畔多 10^{18}
那由他 10^{28}
青莲华 10^{32}
频波罗 10^{56}
钵头摩 10^{80}
矜羯罗 10^{112}
憍祇 10^{224}
阿伽罗 10^{448}
最胜 10^{896}
摩婆罗 10^{1792}
阿僧祇 10^{3584}
阿僧祇转 10^{7168}
多波罗 10^{14336}
无量 10^{28672}
鼻公 10^{57344}
无量转 10^{114688}
普摩 10^{229376}
无边 10^{458752}
阿婆鈐 10^{917504}
无等 $10^{1835008}$
无等转 $10^{3670016}$
毗摩质 $10^{7340032}$
不可数 $10^{14680064}$
不可数转 $10^{29360128}$
僧揭邪摩 $10^{58720256}$
不可称 $10^{117440512}$
不可称转 $10^{234881024}$
不可思 $10^{469762048}$
不可思转 $10^{939524096}$
毗蒲罗 $10^{1879048192}$
不可量 $10^{3758096384}$
不可量转 $10^{7516192768}$
毗婆伽 $10^{15032385536}$
不可说 $10^{30064771072}$
不可说转 $10^{60129542144}$
毗睹罗 $10^{120259084288}$
最妙 $10^{240518168576}$
泥罗婆 $10^{481036337152}$
河理婆 $10^{962072674304}$
一动 $10^{1924145348608}$
河理薄 $10^{3848290697216}$
河理三 $10^{7696581394432}$
娑鲁伽 $10^{15393162788864}$
达擺步陀 $10^{30786325577728}$
河鲁那 $10^{61572651155456}$
摩鲁陀 $10^{123145302310912}$
忙莫陀 $10^{246290604621824}$
娑擺陀 $10^{492581209243648}$
摩鲁摩 $10^{985162418487296}$

调伏 $10^{1970324836974592}$
离憍慢 $10^{3940649673949184}$
不动 $10^{7881299347898368}$
极量 $10^{15762598695796736}$
阿么怛罗 $10^{31525197391593472}$
勃么怛罗 $10^{63050394783186944}$
伽么怛罗 $10^{126100789566373888}$
那么怛罗 $10^{252201579132747776}$
奚么怛罗 $10^{504403158265495552}$
鞞么怛罗 $10^{1008806316530991104}$
钵頭说转 10^{2016}
不可说不可说 10^{4064}
那由他不可说转 10^{8128}
薄頭说不可说不可说 10^{16256}
谛罗 $10^{32281802128991715328}$
偈罗 $10^{64563604257983430656}$
窣步罗 $10^{129127208515966861312}$
泥罗 $10^{258254417031933722624}$
计罗 $10^{516508834063867445248}$
细罗 $10^{1033017668127734890496}$
睥罗 $10^{2066035336255469780992}$
谜罗 $10^{4132070672510939561984}$
娑擺荼 $10^{8264141345021879123968}$
谜鲁陀 $10^{16528282690043758247936}$

契鲁陀 $10^{33056565380087516495872}$
摩睹罗 $10^{66113130760175032991744}$
娑母罗 $10^{132226261520350065983488}$
阿野娑 $10^{264452523040700131966976}$
迦么罗 $10^{528905046081400263933952}$
摩伽婆 $10^{1057810092162800527867904}$
阿怛罗 $10^{2115620184325601055735808}$
酰鲁耶 $10^{4231240368651202111471616}$
薜鲁婆 $10^{8462480737302404222943232}$
羯罗波 $10^{16924961474604808445886464}$
河婆婆 $10^{33849922949209616891772928}$
毗婆罗 $10^{67698845898419233783545856}$
那婆罗 $10^{135396961796838467567091712}$
娜婆罗 $10^{270793938593676835134183424}$
婆婆罗 $10^{541598767187353870268366848}$
迷擺普 $10^{1083197534374707740536733696}$
者么罗 $10^{2166395068749415481073467392}$
駄么罗 $10^{4332790137498830962146934784}$
鉢擺么陀 $10^{8665580274997661924293869568}$
毗迦摩 $10^{17331160549995323848587739136}$
乌波跋多 $10^{34662321099990647697175478272}$
演說 $10^{69324642199981295394350956544}$
无尽 $10^{138649284399962590788701913088}$
出生 $10^{277298568799925181577403826176}$

What is big data



http://mariajose-ccmc.weebly.com/pbl-vii/march-31st-2015

What is big data

- Google
  - Processes 20 PB a day (2008)
  - Crawls 20B web pages a day (2012)
  - Search index is 100+ PB (5/2014)
  - Bigtable serves 2+ EB, 600M QPS (5/2014)
- Yahoo!
  - Hadoop: 365 PB, 330K nodes (6/2014)
- ebay
  - Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)
- Facebook
  - 300 PB data in Hive + 600 TB/day (4/2014)
- Amazon
  - S3: 2T objects, 1.1M request/second (4/2013)

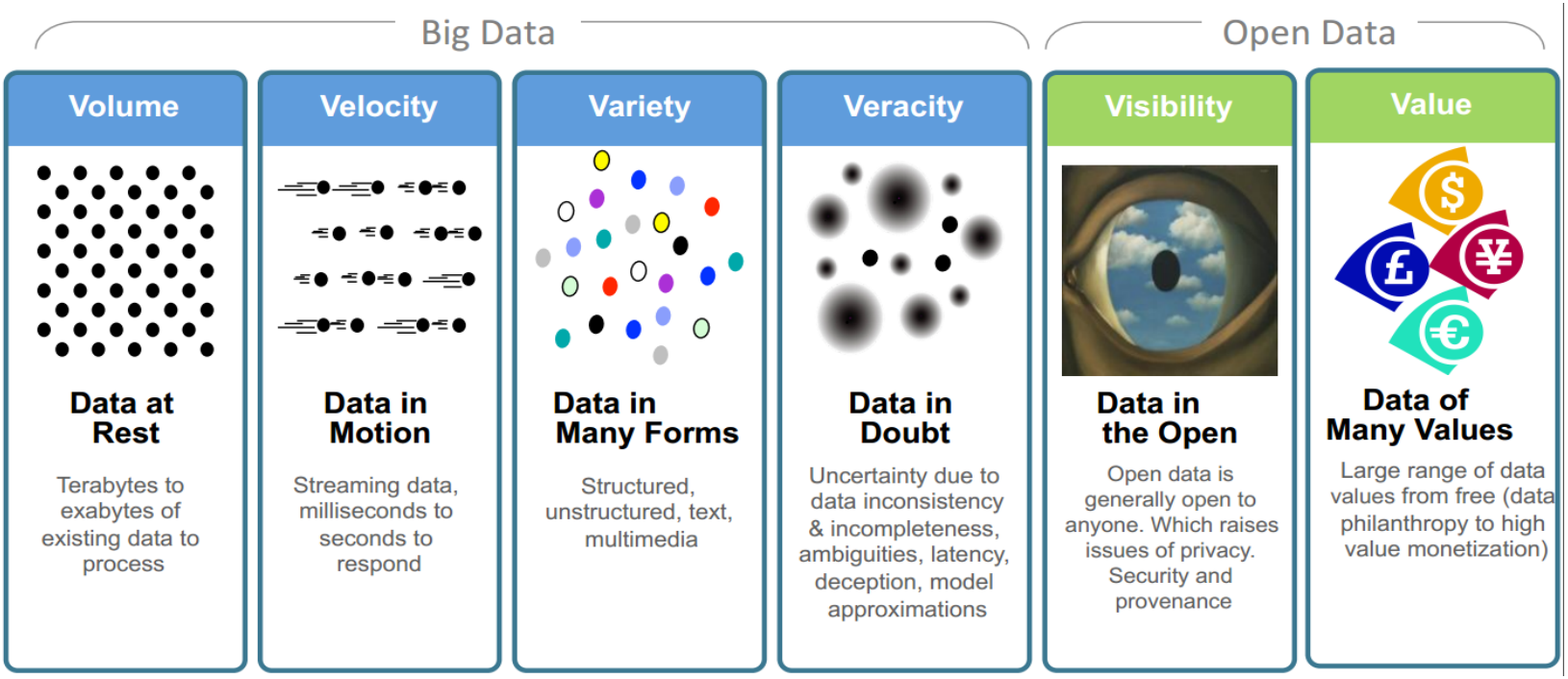
Sources: cs9313, Xin Cao, UNSW

What is big data

- Definitions
  - **Oxford English Dictionary**: data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.
  - **IBM**: Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. Big data is arriving from multiple sources at an alarming velocity, volume and variety. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills.
  - **Wikipedia**: Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
  - **Microsoft**: In essence, big data is data that is valuable but, traditionally, it was not practical to store or analyze it due to limitations of cost or the absence of suitable mechanisms. Big data typically refers to collections of datasets that, due to size and complexity, are difficult to store, query, and manage using existing data management tools or data processing applications.
  - **Gartner**: Big data is 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.
  - ……

What is big data

- Characteristics



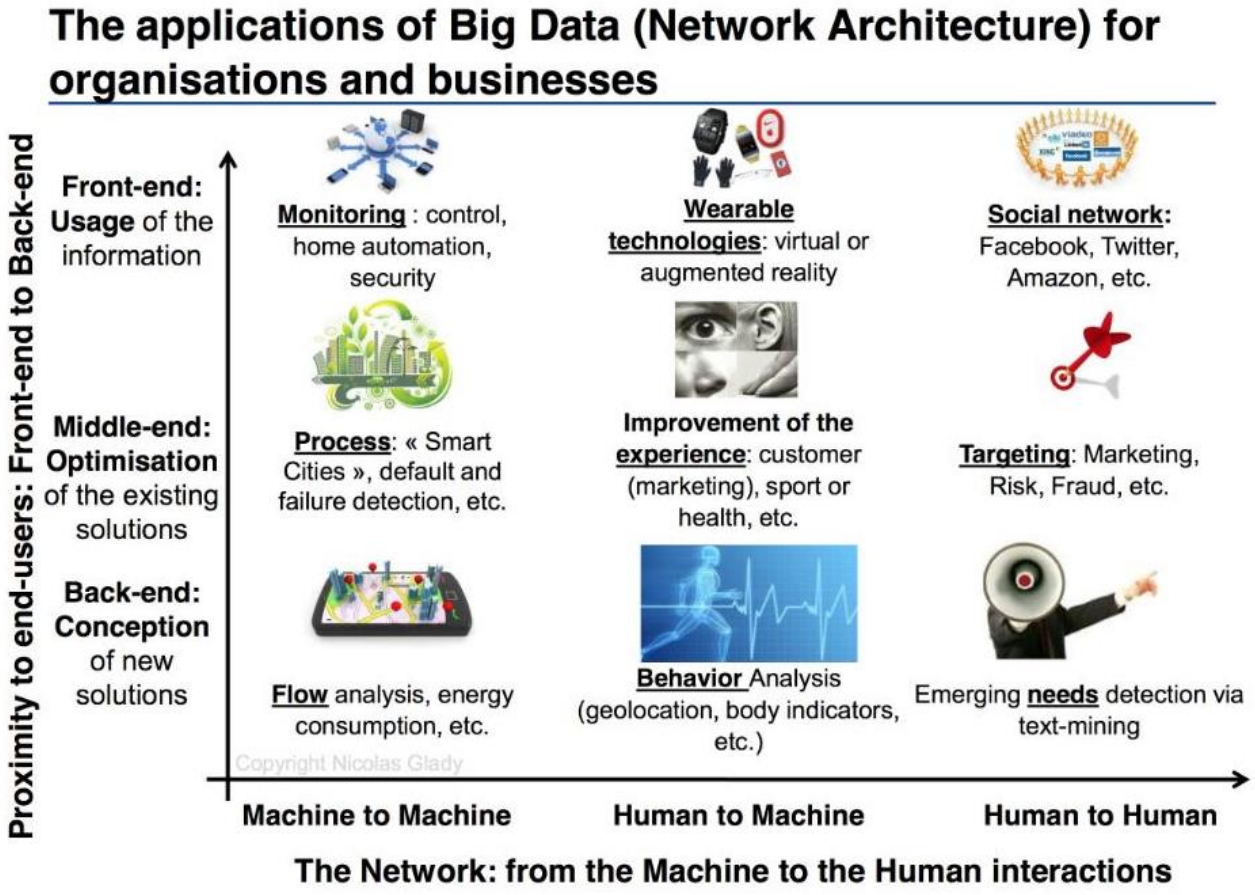
via Anders Quitzau @ IBM

What is big data

- Extended characteristics (via Anders Quitzau @ IBM)
  - **Variability（变异性）**
    - Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
  - **Viscosity（粘性）**
    - This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
  - **Virality （病毒性）**
    - Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.
  - **Volatility（挥发性）**
    - Big data volatility refers to how long is data valid and how long should it be stored. You need to determine at what point is data no longer relevant to the current analysis.
- More Vs in the future …

What is big data

- Applications

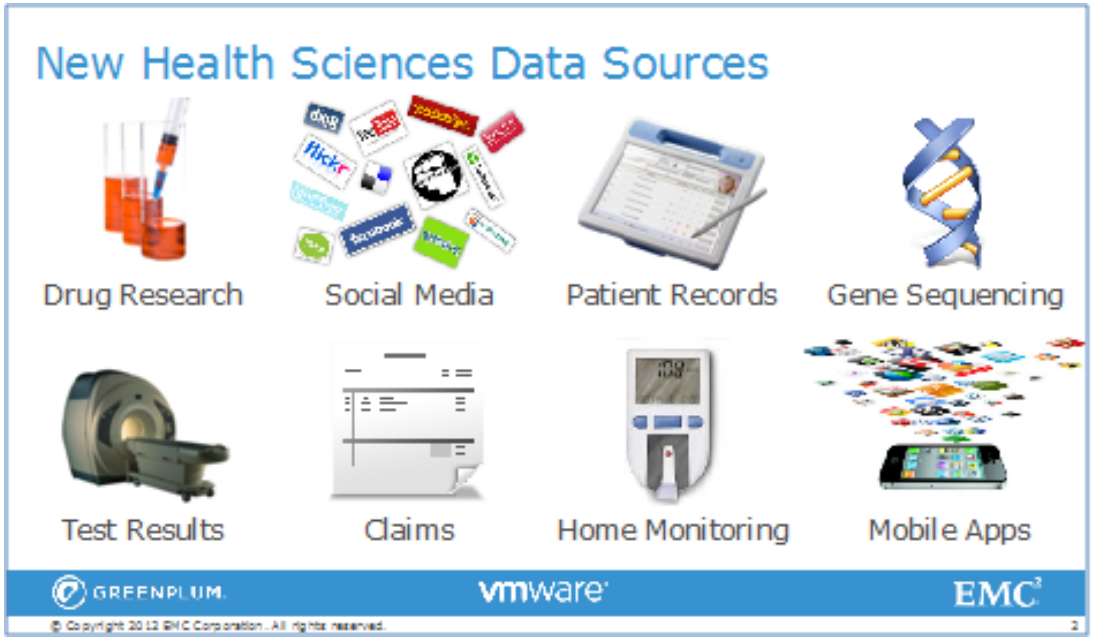


via The Council on Business & Society

Big data in biomedical engineering

Big data in biomedical engineering

- Data sources
  - Medical records
  - Medical devices
  - Clinical scale providers
  - Pharmaceutical companies and research institutions
  - Government surveys
  - Patients



Big data in biomedical engineering

- What'd be good for health care
  - Democratizing analysis
    - One thing for headquarters to calculate quality based on missed opportunities of care; another to push that information out to patients and front-line clinicians
  - Time series for chronic illness
    - Most pop health tools take a years of data to predict a year. Chronic illness runs for decades
  - Patient-supplied data
    - Social work and clinical questionnaire information usually isn't in EMRs
  - Medical device integration
    - Many have computers with audit trails but don't export the data into EMRs
  - Merging administrative and clinical data, especially interorganizationally
    - Would rather base diagnoses on lab readings than diagnostic codes
  - Census integration (the missing denominator problem)

Big data in biomedical engineering

- Intelligent medicine
  - Telemedicine
  - Computer Aided Diagnosing
  - Precision medicine
- Personalized healthcare
  - Wearable devices
  - Chronic disease monitoring
- Medical fraud identification
- Health insurance risk control

Cloud computing

Cloud computing

- What is cloud computing
  - Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) [Mell_2009], [Berkely_2009].
  - It can be rapidly provisioned and released with minimal management effort.
  - It provides high level abstraction of computation and storage model.

Cloud computing: Why?

- It's cheaper and more scalable to tie together large numbers of commodity computers than to make bigger single computers
  - That's what Hadoop and Spark are about
- This is important if you have a data set that's really big
  - Won't fit onto single disk drives
  - Too big to be processed in a reasonable period of time
- Metaphor: To find a needle in a haystack, divide it into 100 smaller haystacks and assign 100 people to check each one at the same time
  - For mass personalization, speed is more important than accuracy

Cloud computing: How cheap?

- Buying Oracle on Exadata for a 100 terabyte data warehouse costs about $6K per terabyte plus 20% annual maintenance
  - Add labor for a system administrator, tech refresh every five years & real estate
- Buying Amazon Redshift (cloud data warehouse service) with a three year commitment is about $1K per terabyte per year
  - Administration, tech refresh, and real estate is included
- You can rent computing by the hour
  - 64 cores/256 RAM $3.83/hour on demand (Amazon)
  - 31% off for one year commitment no upfront
  - 61% off for three year commitment all upfront
  - Could be as low as 43 cents per hour on auction
- Cloud storage is $30/terabyte per month, less for archive

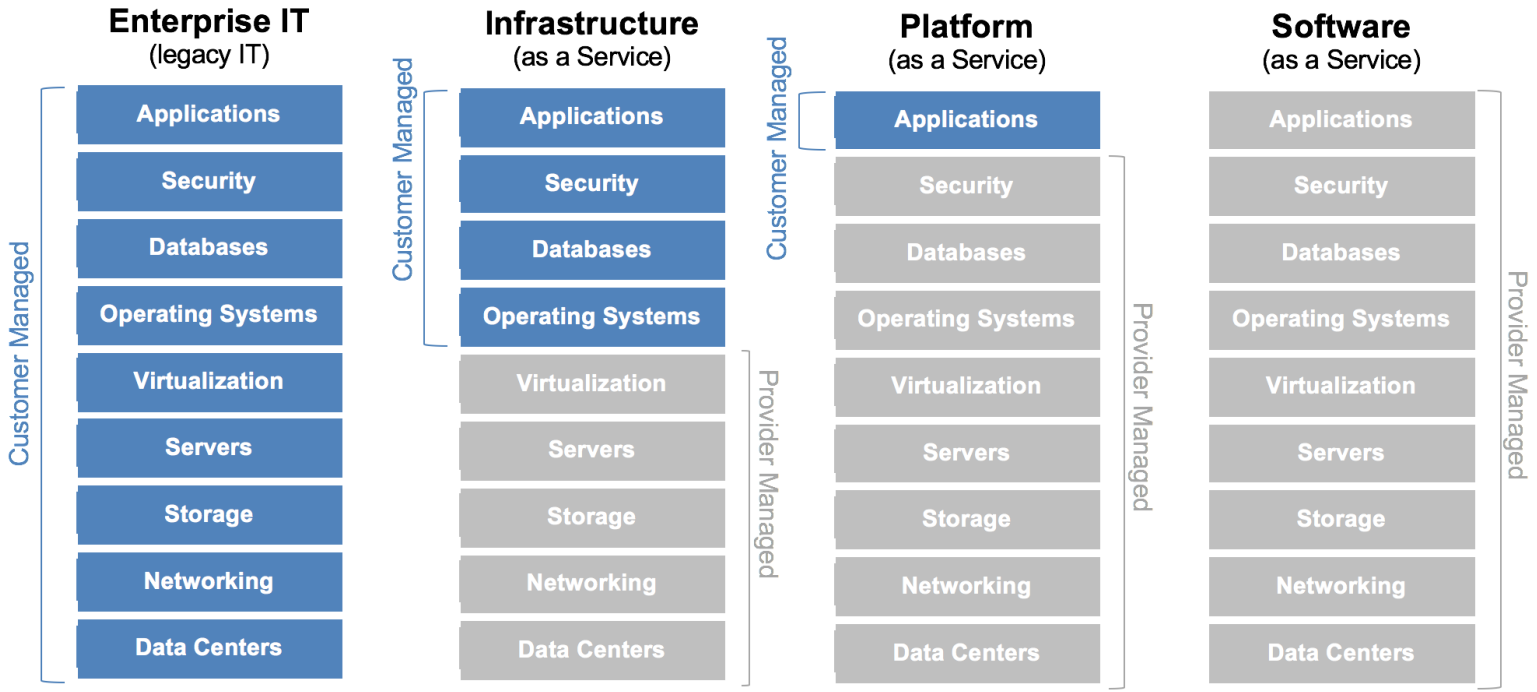
Cloud computing

- Evolution
  - Grid computing
    - Loosely coupled (Decentralization)
    - Diversity and Dynamism
    - Distributed Job Management & scheduling
  - Ubiquitous computing
    - A concept in software engineering
    - Computing is made to appear everywhere and anywhere
    - Always-on platforms embodied by smart devices
  - Distributed computing
    - Tightly coupled systems
    - Single system image
    - Centralized Job management & scheduling system
  - Cloud computing
    - Dynamic computing infrastructure
    - IT service centric approach
    - Self service based usage model
    - Minimally or self managed platform

Cloud computing: Services

- Service models (via TechTarget)
  - Infrastructure as a Service (**IaaS**)
    - Providing the fundamental building blocks of computing resources, such as virtualization, storage, networking, load balancers, system maintenance, backup and resiliency planning.
    - Google Cloud Platforms, Amazon Web Services, VMWare, *etc*.
  - Platform as a Service (**PaaS**)
    - A cloud provider delivers hardware and software tools from its own hosting infrastructure to its users as a service, which frees users from having to install in-house hardware and software to develop or run a new application.
    - Google App Engine, Microsoft Azure, *etc*.
  - Software as a Service (**SaaS**)
    - A software distribution model in which a third-party provider hosts applications and makes them available to customers over the Internet.
    - Google Apps (Gmail, Calendars, Docs, Books), Dropbox, Instagram, GitHub, *etc*.

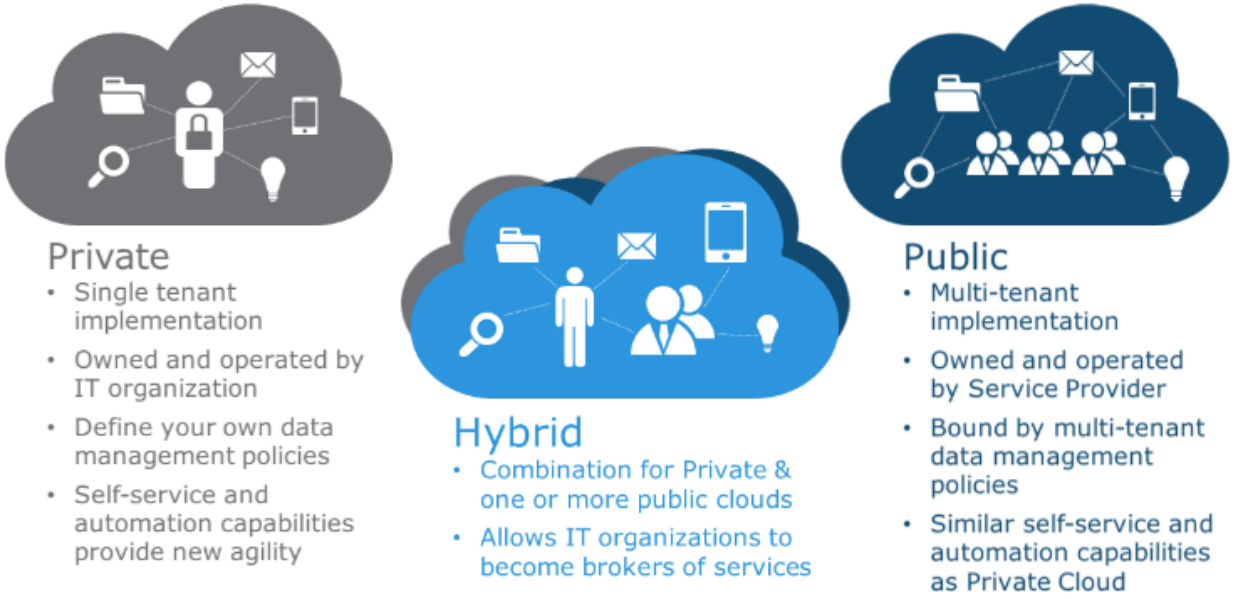
Cloud computing: Services



via LinkedIn: Arsalan Eizadirad

Cloud computing: Deployment

- Deployment models



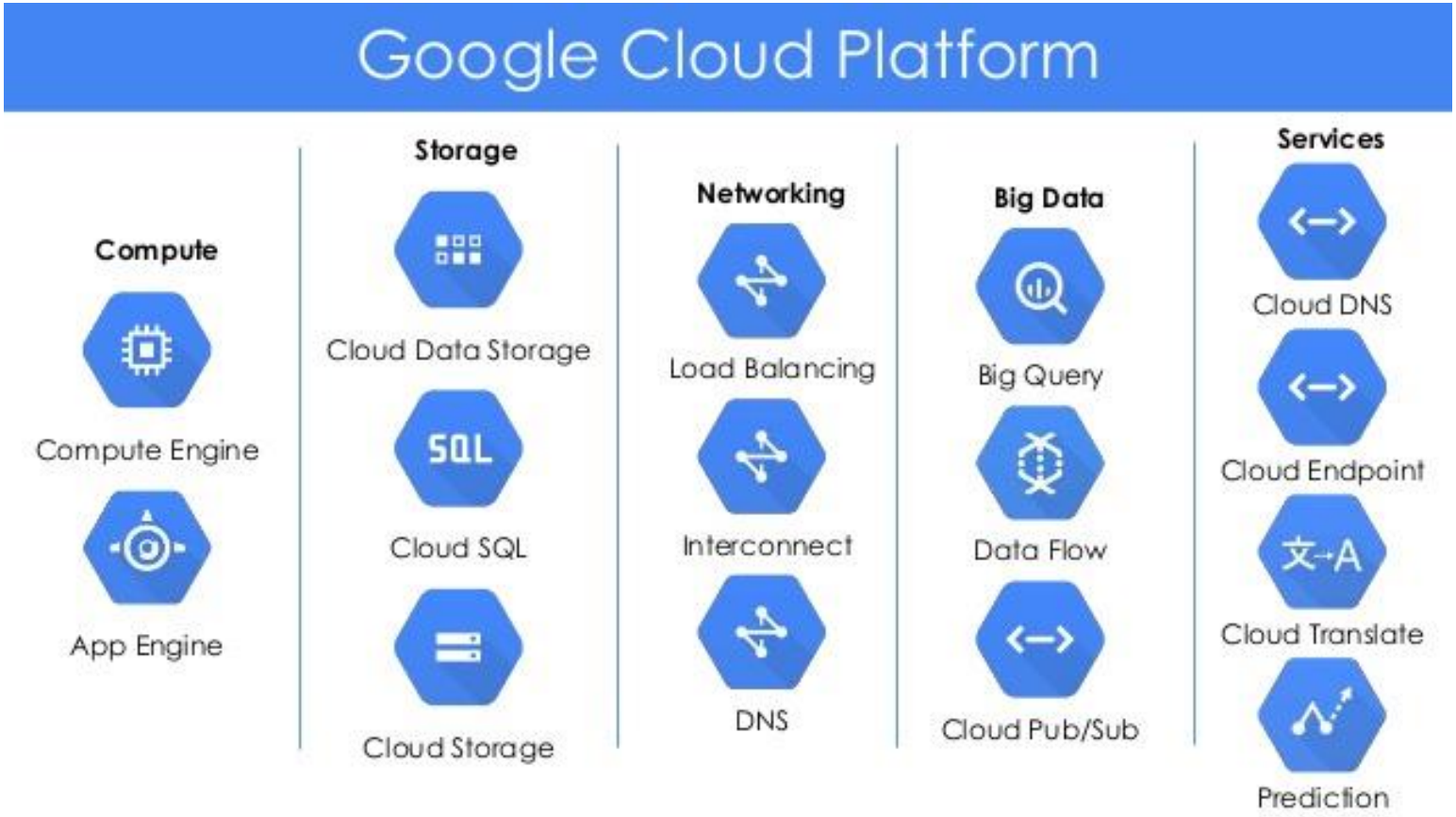
via LinkedIn: Ankur Minotra

Cloud computing

- Big shots
  - Alphabet (Google)
  - Amazon
  - Apple
  - Facebook
  - Microsoft
  - Tesla
  - Twitter

Cloud computing: Google

- Google cloud platform



Cloud computing: Google

- Google apps

Cloud computing: Amazon

- AWS

Cloud computing: Amazon

- Recommendation system

Frequently bought together

Total price: $155.09

Add both to Cart

Add both to List

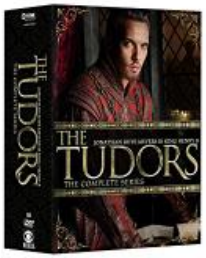
☑ **This item:** Game of Thrones: The Complete Seasons 1-6 DVD $125.45

☑ A Game of Thrones / A Clash of Kings / A Storm of Swords / A Feast of Crows / A Dance with Dragons by George R. R. Martin Mass Market Paperback $29.64

Customers who bought this item also bought

Page 1 of 5

A Game of Thrones / A Clash of Kings / A Storm of Swords / A Feast of...
› George R. R. Martin
★★★★½ 9,255
Mass Market Paperback
$29.64 ✓prime



Tudors: The Complete Series
James Frain
★★★★½ 234
DVD
$31.49 ✓prime

Dexter: The Complete Series
James Remar
★★★★½ 154
DVD
$52.27 ✓prime

Breaking Bad: The Complete Series
Bryan Cranston
★★★★½ 2,031
DVD
$76.52 ✓prime

Better Call Saul: Season 2
Bob Odenkirk
★★★★½ 97
DVD
$16.69 ✓prime

Cloud computing: Amazon

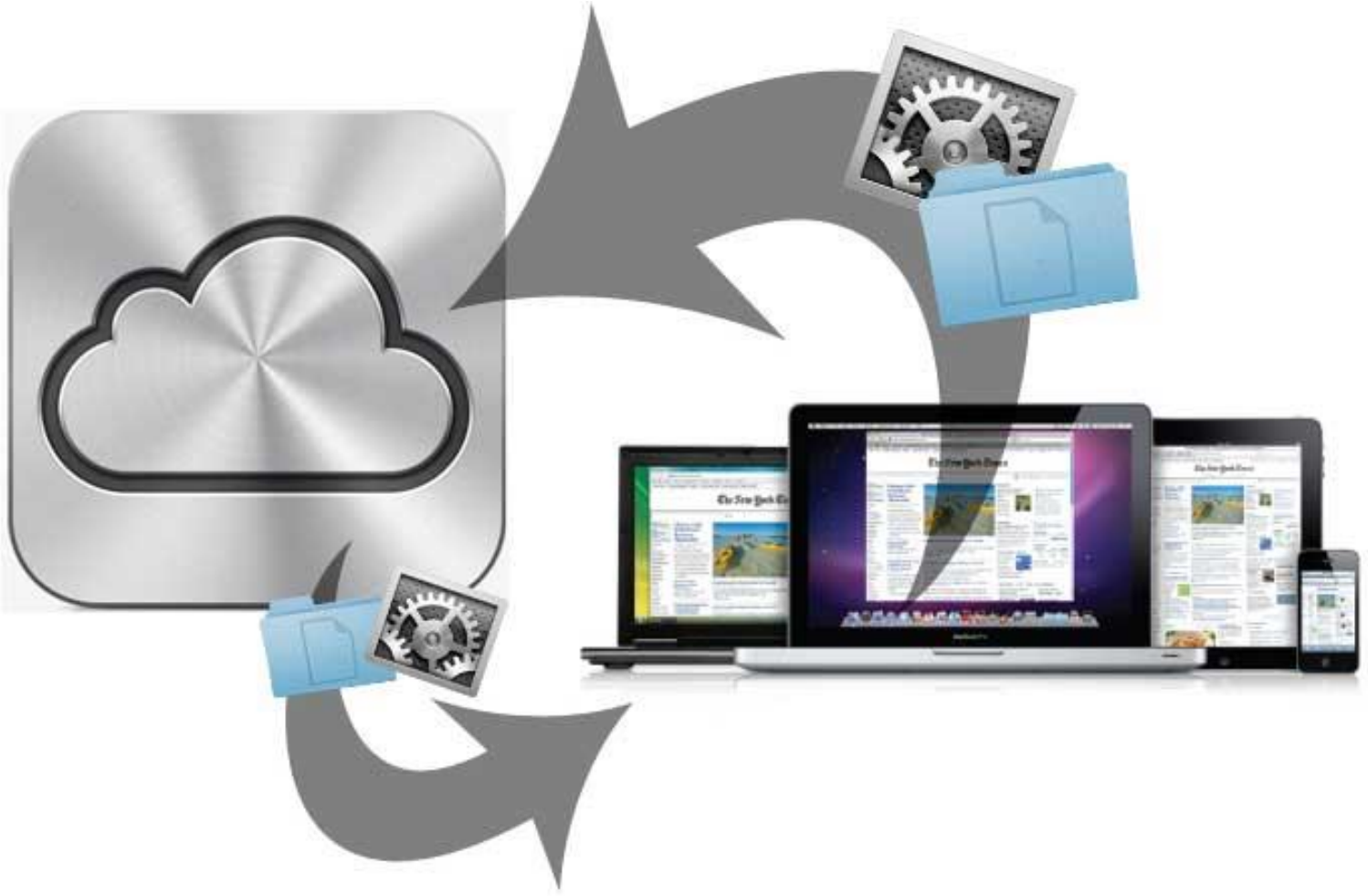
- Echo

Cloud computing: Apple

- Siri

Cloud computing: Apple

- iCloud



Cloud computing: Apple

- Photos of iOS 10

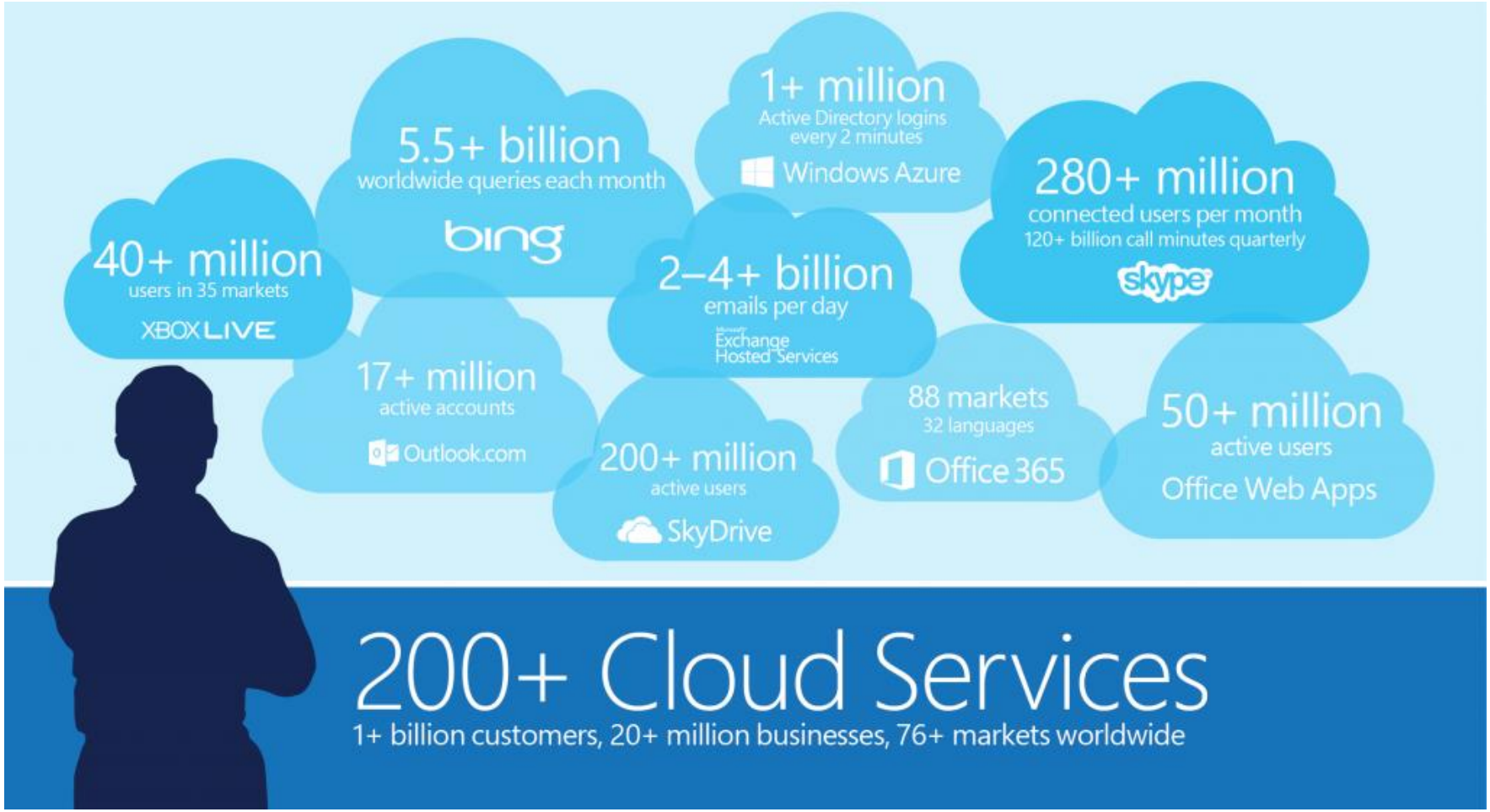
Cloud computing

- Facebook

via MacWorld

Cloud computing

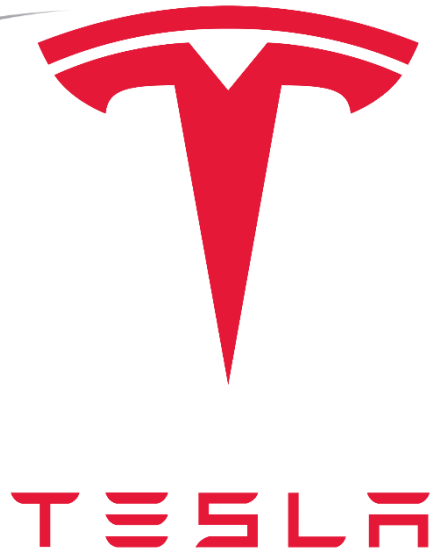
- Microsoft



http://news.filehippo.com/2013/03/analyzing-microsofts-cloud-prospects/

Cloud computing

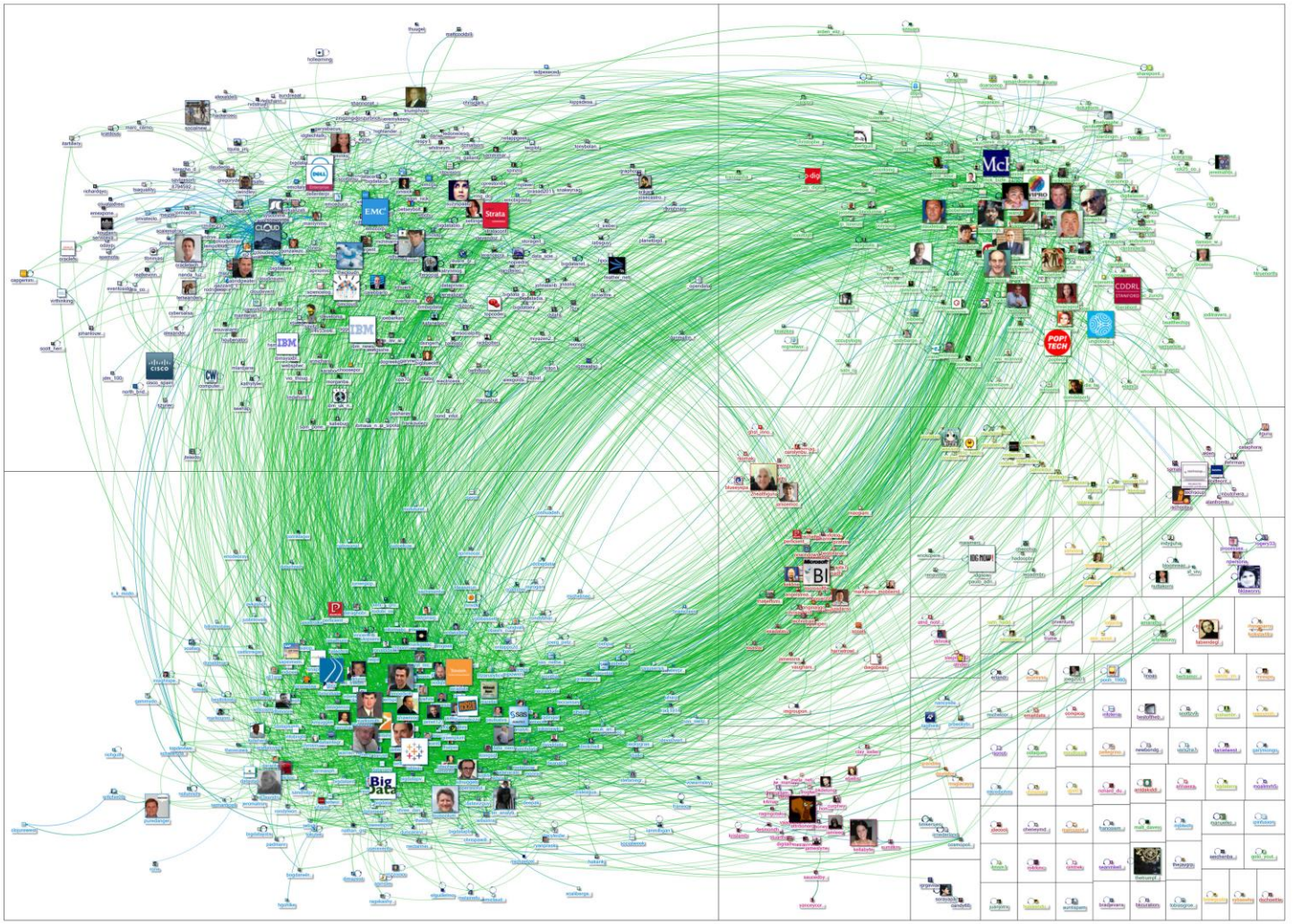
- Elon Musk



Cloud computing

- Twitter

Social media network connections among Twitter users



Created with NodeXL (http://nodexl.codeplex.com) from the Social Media Research Foundation (http://www.smrfoundation.org)

Cloud computing

- Data center (Google)

Data center locations

We own and operate data centers around the world to keep our products running 24 hours a day, 7 days a week. Find out more about our data center locations, community involvement, and job opportunities in our locations around the world.

Americas

Berkeley County, South Carolina
Council Bluffs, Iowa
Douglas County, Georgia
Jackson County, Alabama
Lenoir, North Carolina
Mayes County, Oklahoma
Montgomery County, Tennessee
Quilicura, Chile
The Dalles, Oregon

Asia

Changhua County, Taiwan
Singapore

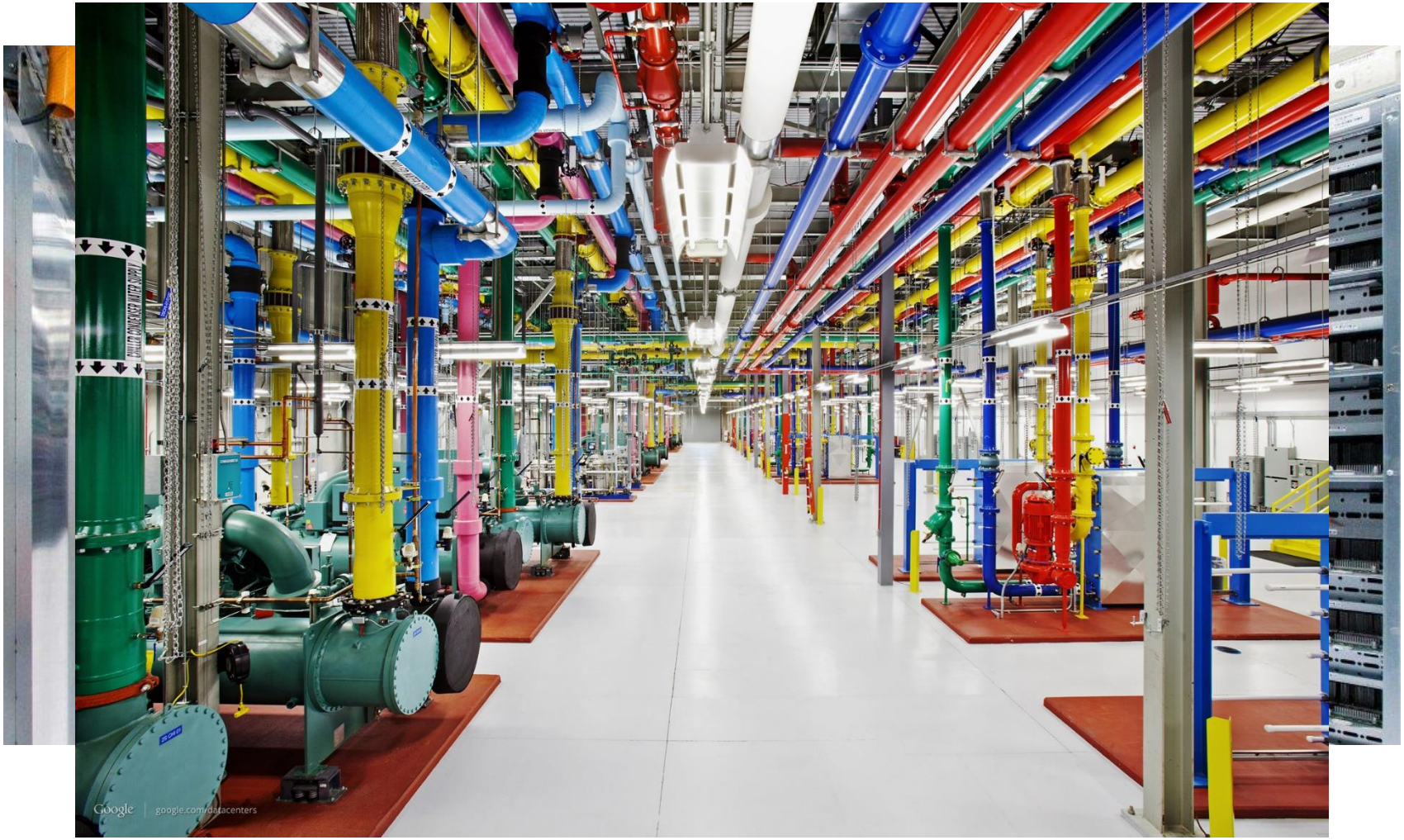
Europe

Dublin, Ireland
Eemshaven, Netherlands
Hamina, Finland
St Ghislain, Belgium

https://www.google.com/about/datacenters/inside/locations/

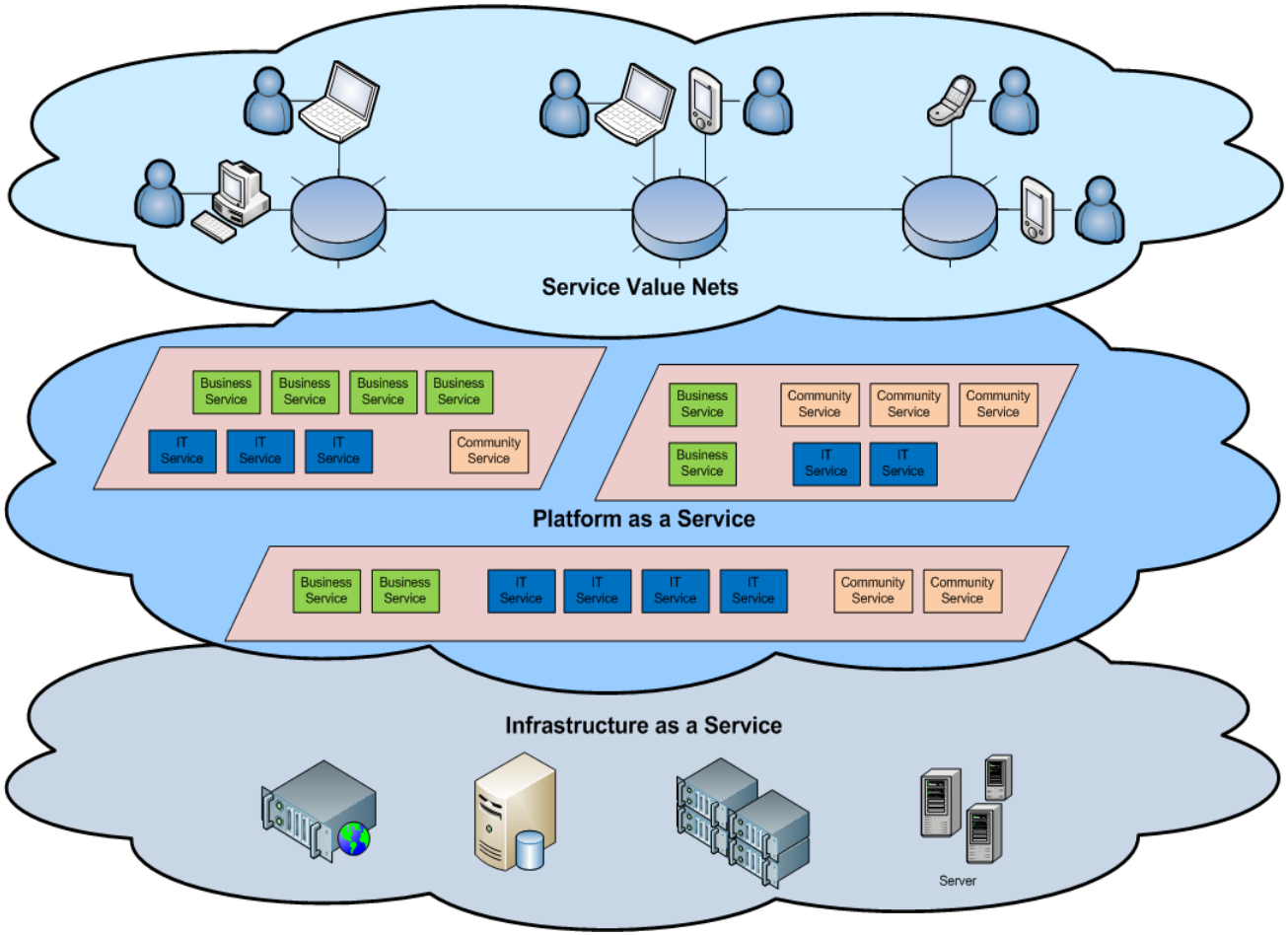
Cloud computing

- Data center (Google @ Douglas County, Geogia)



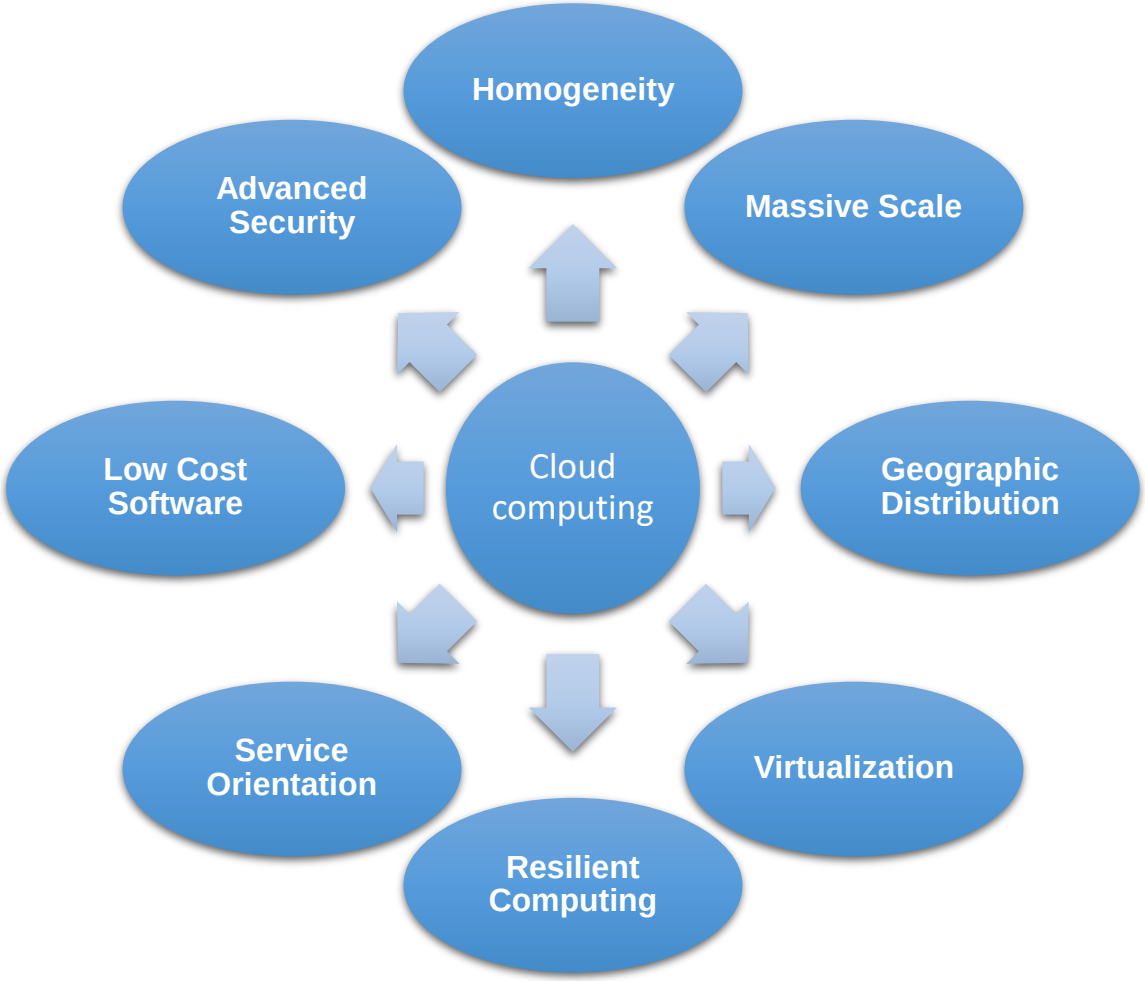
Cloud computing

- Architecture



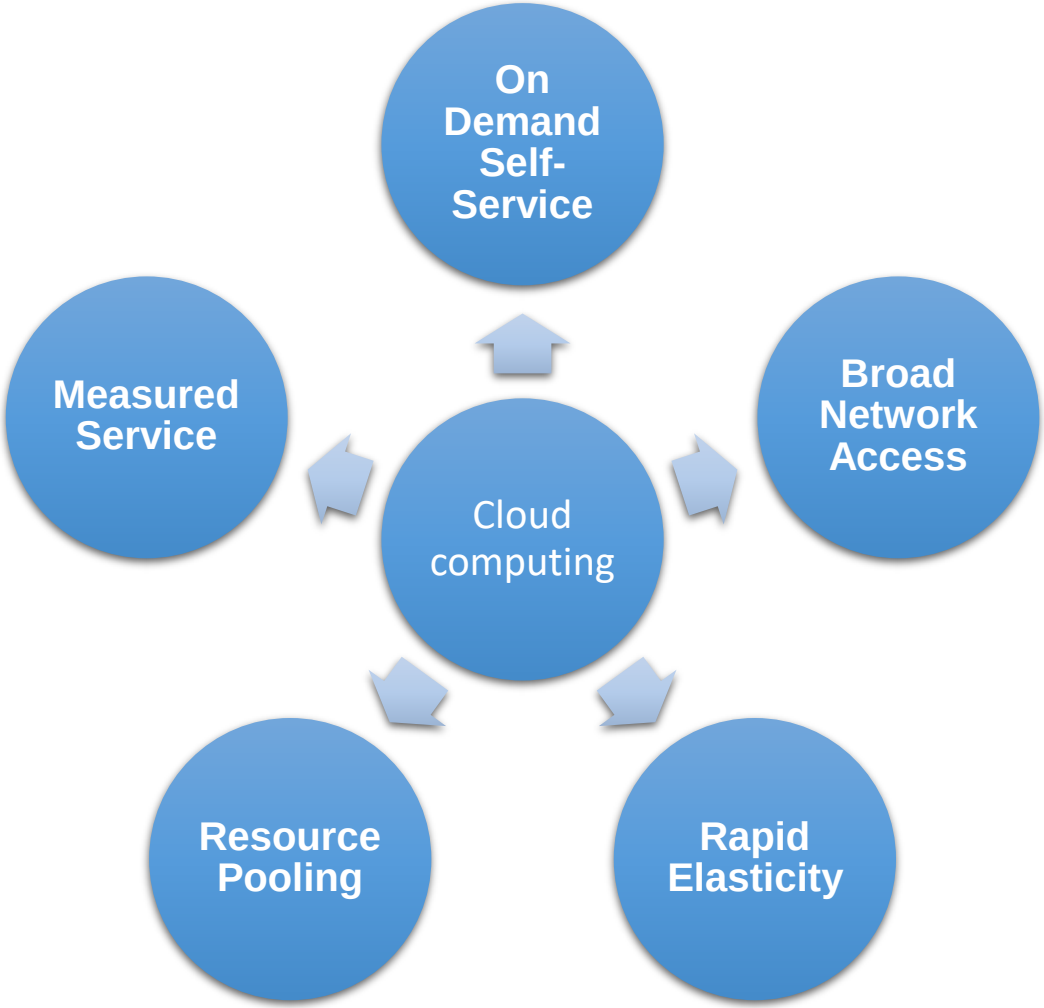
Cloud computing

- Common characteristics



Cloud computing

- Essential characteristics



Cloud computing

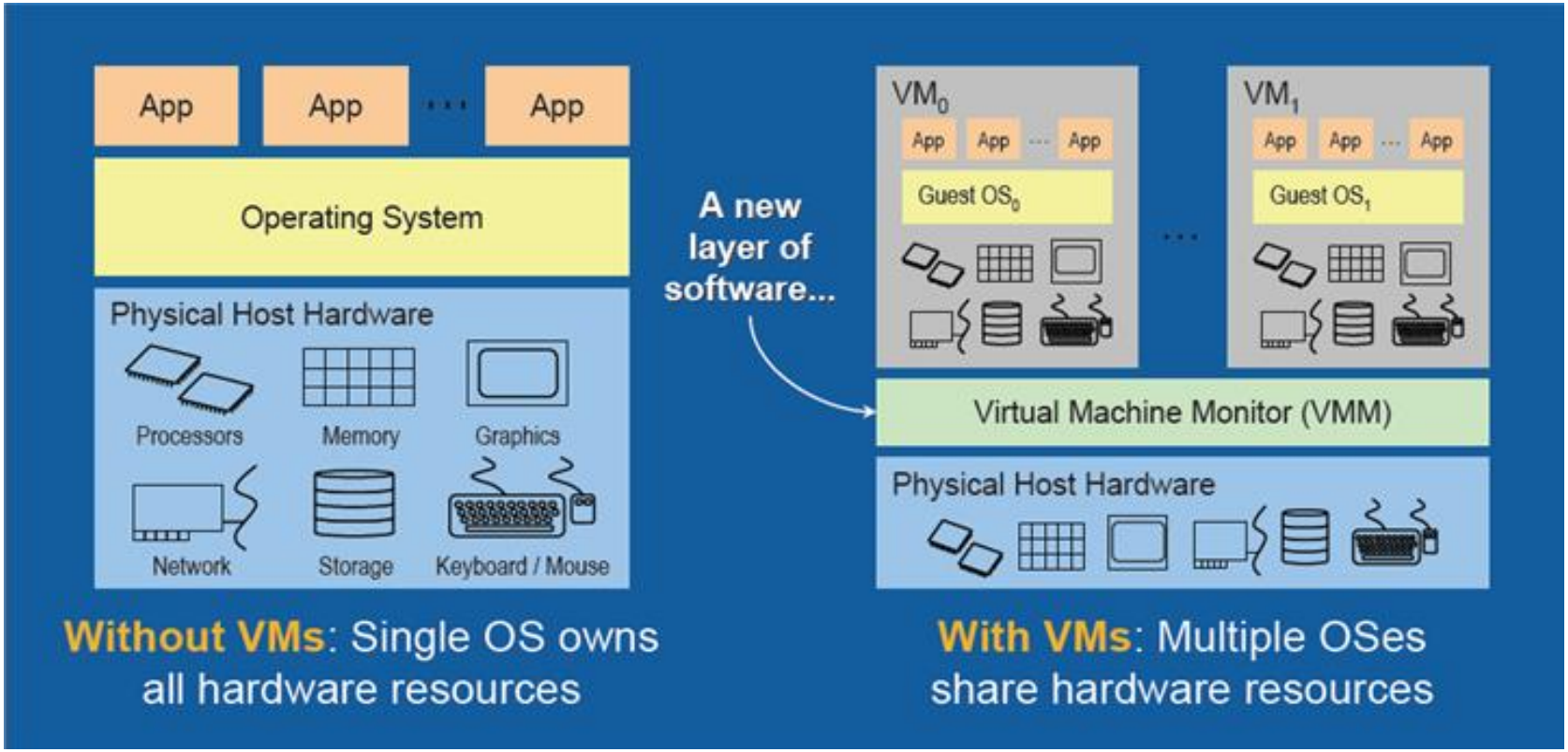
- Essential characteristics
  - On-Demand Self Service:
    - A consumer can unilaterally provision computing capabilities, automatically without requiring human interaction with each service's provider.
  - Heterogeneous Access:
    - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms.
  - Resource Pooling:
    - The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model.
    - Different physical and virtual resources dynamically assigned and reassigned according to consumer demand.
  - Measured Service:
    - Cloud systems automatically control and optimize resources used by leveraging a metering capability at some level of abstraction appropriate to the type of service.
    - It will provide analyzable and predictable computing platform.

Cloud computing

- The “no-need-to-know” in terms of the underlying details of infrastructure, applications interface with the infrastructure via the APIs.
- The “flexibility and elasticity” allows these systems to scale up and down at will.
- The “pay as much as used and needed” type of utility computing and the “always on!, anywhere and any place” type of network-based computing.
- Cloud are transparent to users and applications, they can be built in multiple ways: branded products, proprietary open source, hardware or software, or just off-the-shelf PCs.
- Implement on Virtual Machines (VMs):
  - Abstraction of a physical host machine,
  - Hypervisor intercepts and emulates instructions from VMs, and allows management of VMs,
  - VMWare, Xen, *etc*.
- Provide infrastructure API:
  - Plug-ins to hardware/support structures.

Cloud computing

- Virtualization technology
  - allows multiple virtual machines to run on a single physical machine.



via Intel Software Developer Zone

Cloud computing

- In parallel there has been backlash against cloud computing:
  - Use of cloud computing means dependence on others and that could possibly limit flexibility and innovation:
    - The others are likely become the bigger Internet companies like Google and IBM, who may monopolise the market.
    - Some argue that this use of supercomputers is a return to the time of mainframe computing that the PC was a reaction against.
  - Security could prove to be a big issue:
    - It is still unclear how safe out-sourced data is and when using these services ownership of data is not always clear.
  - There are also issues relating to policy and access:
    - If your data is stored abroad whose policy do you adhere to?
    - What happens if the remote server goes down?
    - How will you then access files?
    - There have been cases of users being locked out of accounts and losing access to data.

Cloud computing

- Advantages (from service aspect)
  - Simplicity – easy to deploy and use
  - Pay as you use – pay only for the services you consume
  - Cost saving – Lesser in-house IT costs
  - Scalability
  - Backup and Recovery
  - Easy to upgrade
  - On demand availability

Cloud computing

- Advantages (from application aspect via Mark Baker)
  - Lower computer costs
  - Improved performance
  - Reduced software costs
  - Instant software updates
  - Improved document format compatibility
  - Unlimited storage capacity
  - Increased data reliability
  - Universal document access
  - Latest version availability
  - Easier group collaboration
  - Device independence

Cloud computing

- Disadvantages
  - Can be slow
    - Requires a constant high-speed Internet connection.
    - Even with a fast connection, web-based applications can be slower than accessing a similar software program on your desktop PC.
  - Stored data might not be secure
    - With cloud computing, all your data is stored on the cloud.
    - Can unauthorised users gain access to your confidential data?
  - Stored data can be lost
    - Theoretically, data stored in the cloud is safe, replicated across multiple machines.
    - But on the off chance that your data goes missing, you have no physical or local backup.
  - HPC Systems
    - Not clear that you can run compute-intensive HPC applications that use MPI/OpenMP!
    - Scheduling is important with this type of application
  - General Concerns
    - Each cloud systems uses different protocols and different APIs. May not be possible to run applications between cloud based systems
    - Your normal applications will have to be adapted to execute on these platforms.

Cloud computing

- Challenges
  - Security
  - Performance
  - Uninterrupted availability
  - Integration with in-house IT
  - Ability to customize to internal needs
  - Migrating back to in-house
  - Regulatory requirements prohibit cloud (data storage abstracted)

End of Chapter 1