UNSW SYDNEY | Australia's Global University

COMP9321 Data Services Engineering

Term 1, 2023

Week 1: Introduction

Data Services

- What? Data services are software services that encapsulate operations on key data entities relevant to the consumer
- Why? Data nowadays is stored in multiple systems and require multiple interfaces or mechanisms to interact with them. There are varying channels (e.g., legacy systems, Online, third-party) and mechanisms (e.g., event driven, on demand, batch process) that need to be served as well adding additional challenges to data services. Without an abstraction layer for data consumers that insulate them from this complexity we will end up with a spaghetti of point to point integrations between data sources and data consumers

Source: * https://www.infoq.com/articles/narayanan-soa-data-services

Let’s Go Deeper

Data Recording… The Beginning

Source: Wikipedia.org

Not that Deep

Data-oriented Services …

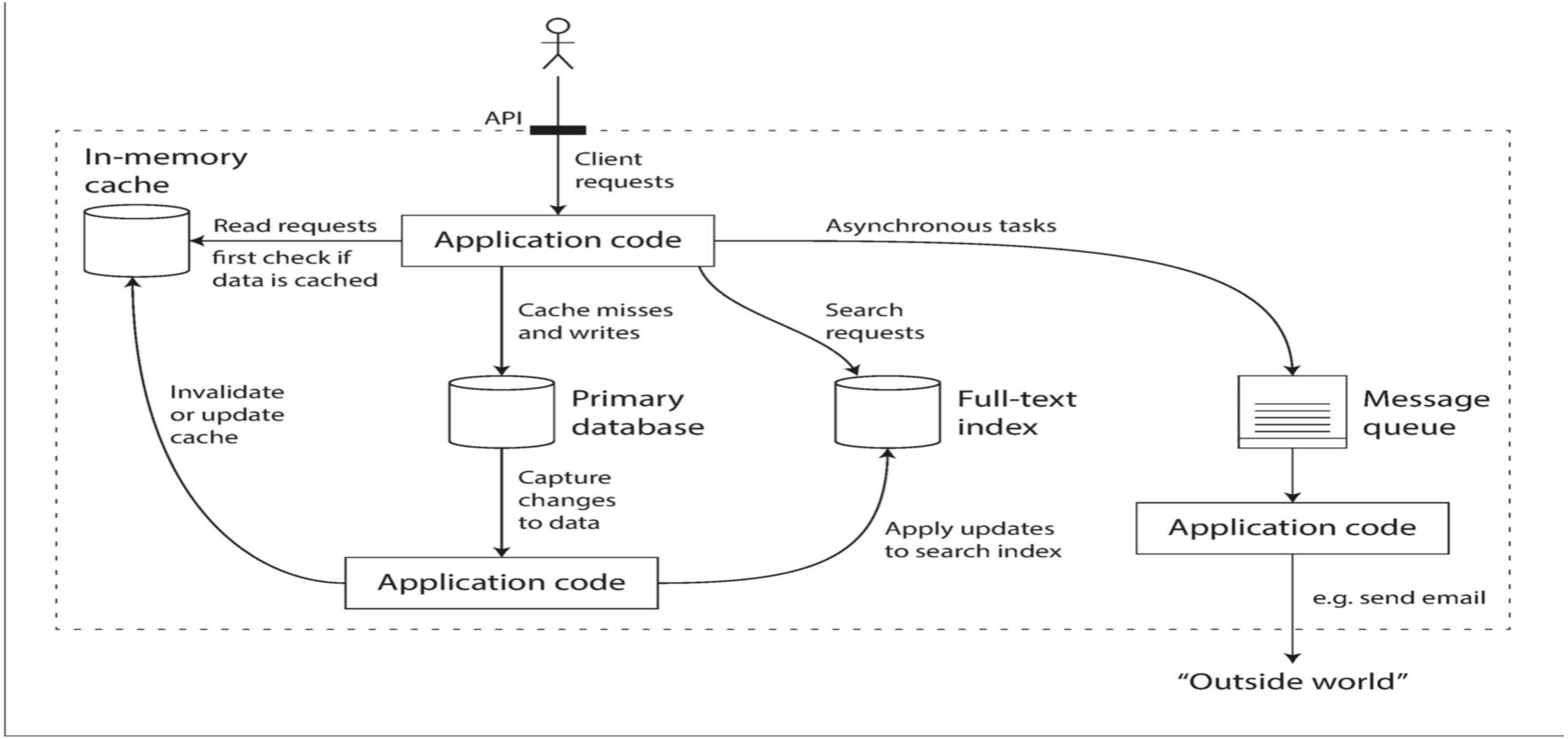


Figure 1-1. One possible architecture for a data system that combines several components.

Information Systems/Applications Integration

A set of services and solutions for bringing together disparate application and business processes as needed to meet the diverse information requirements of your customers, partners, suppliers and employees.

Motivations: Streamlining business operations, globalisation, competition, mergers and acquisition, new business models, technology development, etc.

- e.g., merger of two companies (data + processes)

Problems: systems to be integrated are not homogeneous.

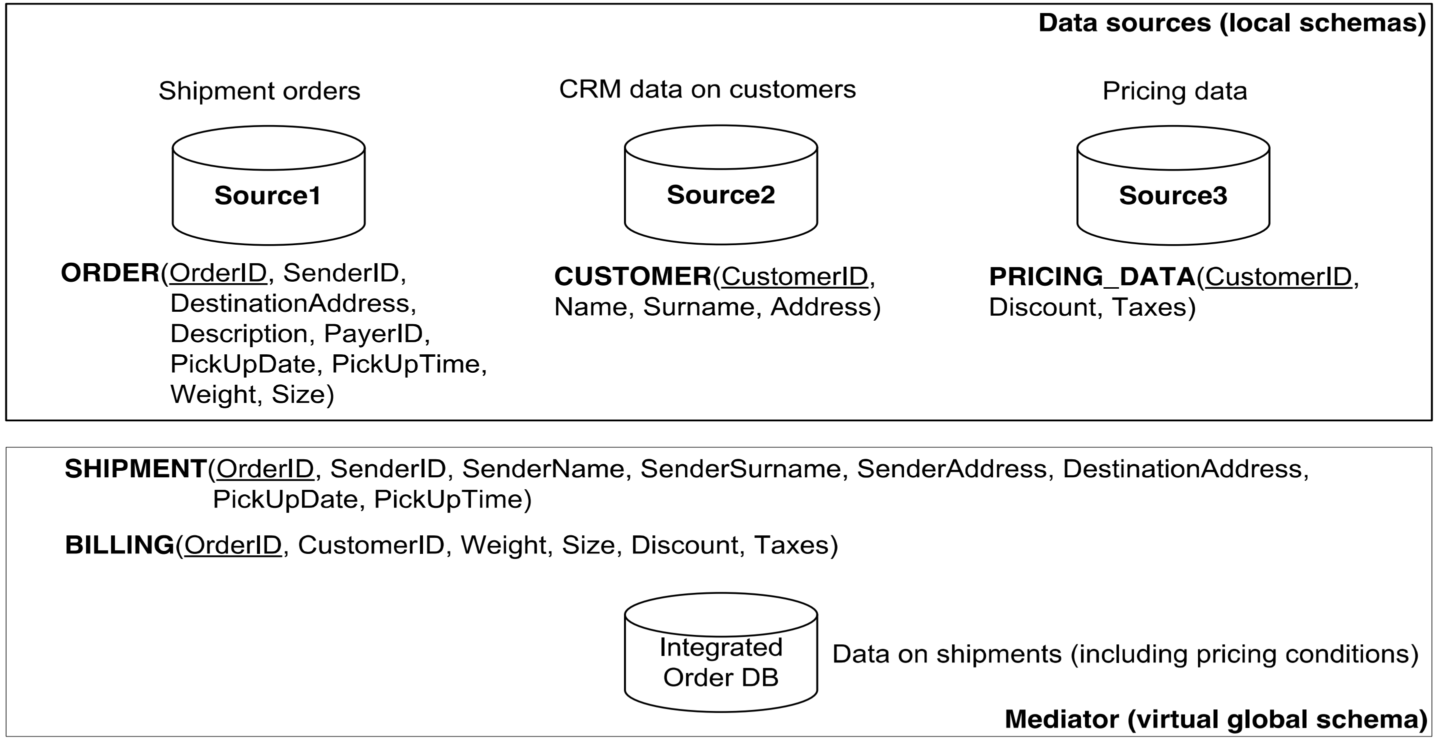
- they are individually developed (ad-hoc) systems overtime
- some are “off-the-shelf” packages
- different execution platforms, technologies and business rules

Heterogeneity at different levels: language, platform, schema (data, process)

- Data integration, Process/Systems integration

Data Level Integration …

Data integration = combining data from different sources and providing users with a unified view over them



Data sources (local schemas)

Shipment orders

Source1

ORDER(OrderID, SenderID, DestinationAddress, Description, PayerID, PickUpDate, PickUpTime, Weight, Size)

CRM data on customers

Source2

CUSTOMER(CustomerID, Name, Surname, Address)

Pricing data

Source3

PRICING_DATA(CustomerID, Discount, Taxes)

SHIPMENT(OrderID, SenderID, SenderName, SenderSurname, SenderAddress, DestinationAddress, PickUpDate, PickUpTime)

BILLING(OrderID, CustomerID, Weight, Size, Discount, Taxes)

Integrated Order DB

Data on shipments (including pricing conditions)

Mediator (virtual global schema)

Fig. 2.2 Example of an integrated database storing shipment data extracted from different data sources. Each data source is characterized by a *local schema*. Data integration is performed according to a *vritual global schema* managed by the mediator.

Picture from Mashups: Concepts, Models and Architectures

Data Level Integration …

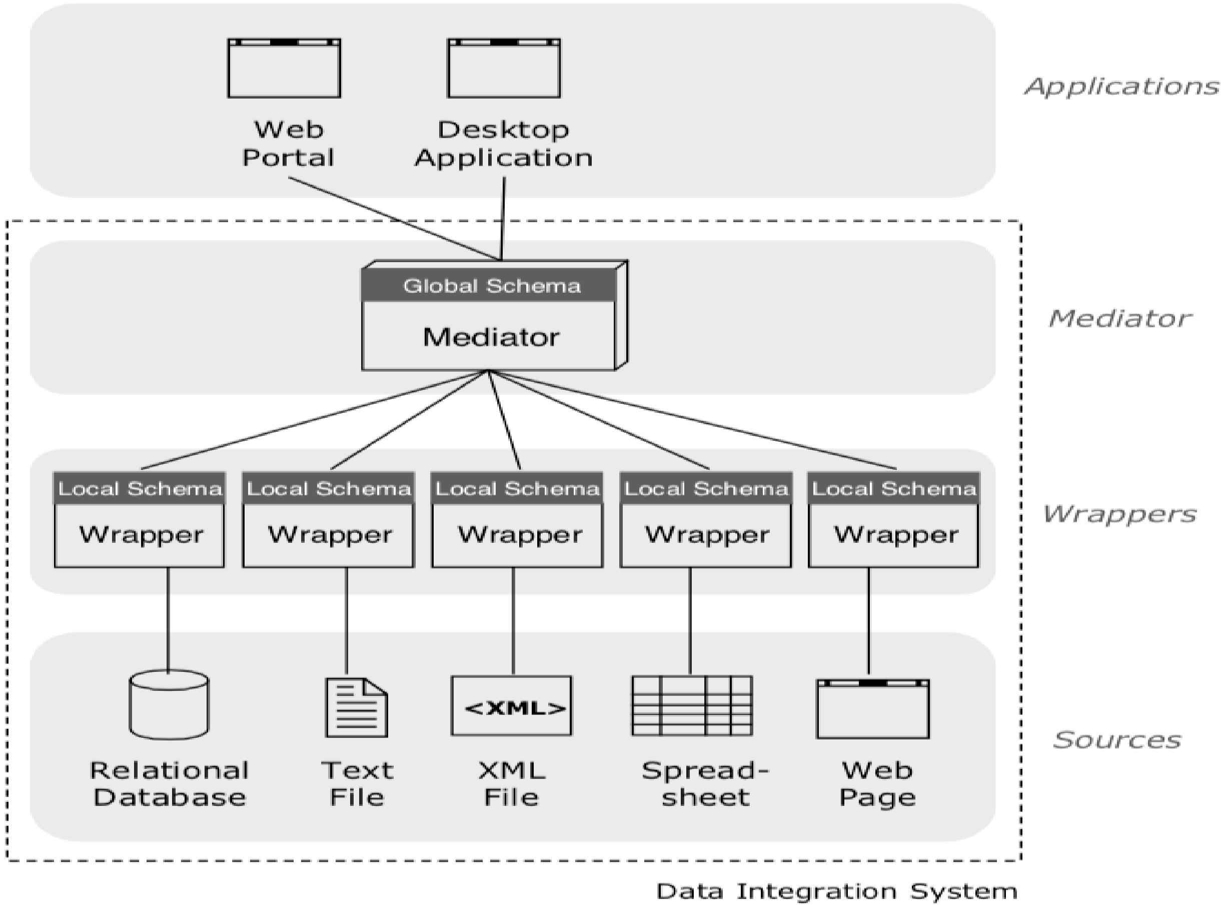
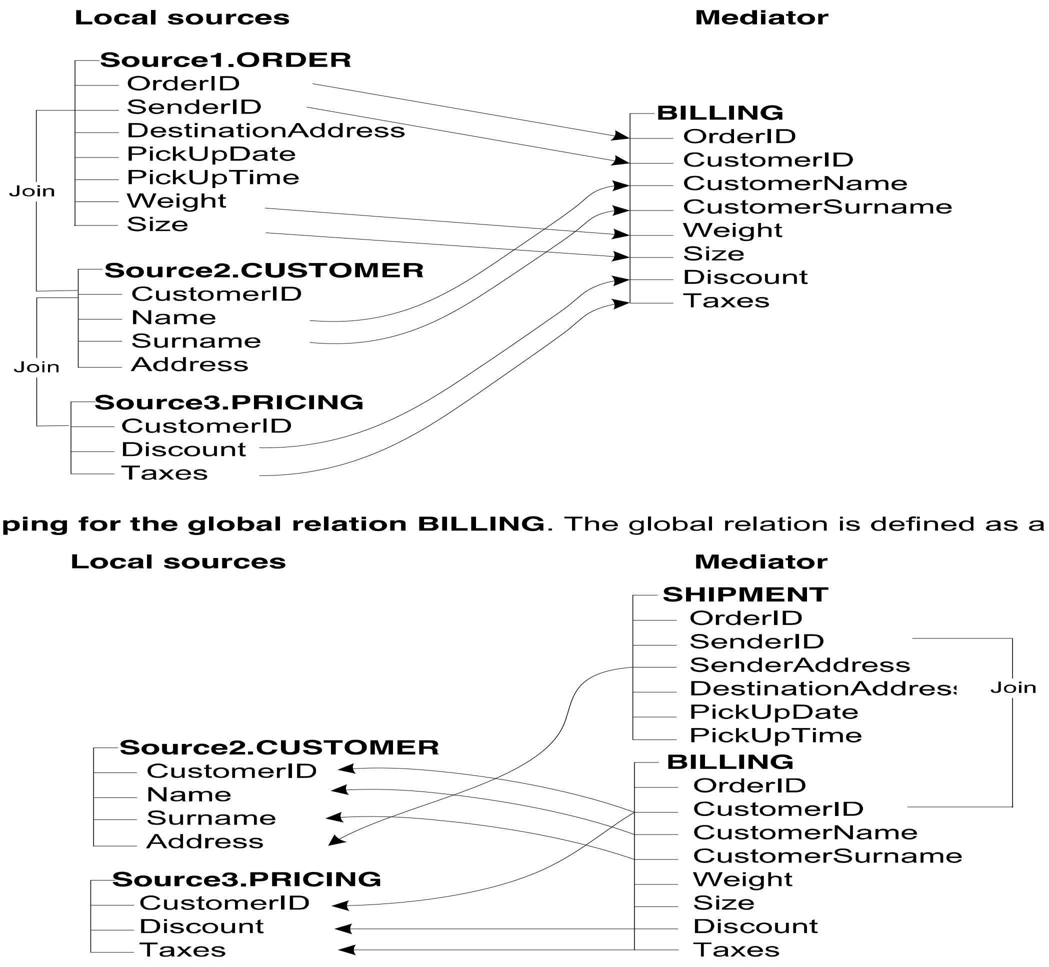


Figure 1: View-based Data Integration System (VDIS) Architecture

Picture from View-based Integration, Yannis Katsis, CSE, UC San Diego

Data Level Integration …



a) **GAV Mapping for the global relation BILLING**. The global relation is defined as a view on the local source relations.

b) **LAV Mapping for Source2 and Source3.** The local source relations are defined as views over the global relations.

Fig. 2.3 Example of GAV and LAV schema mappings for the integrated order DB.

Picture from Mashups: Concepts, Models and Architectures

Data Level Integration …

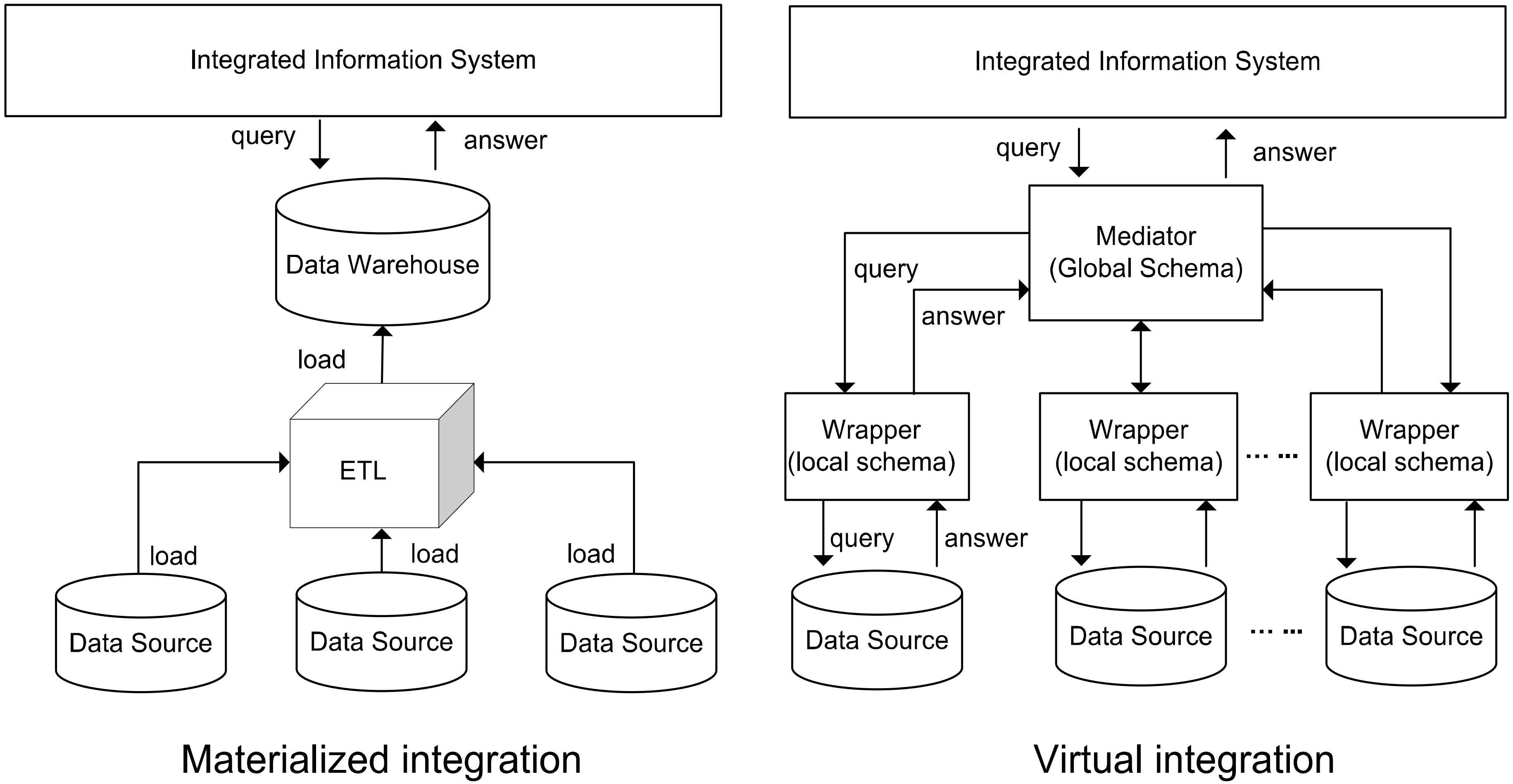


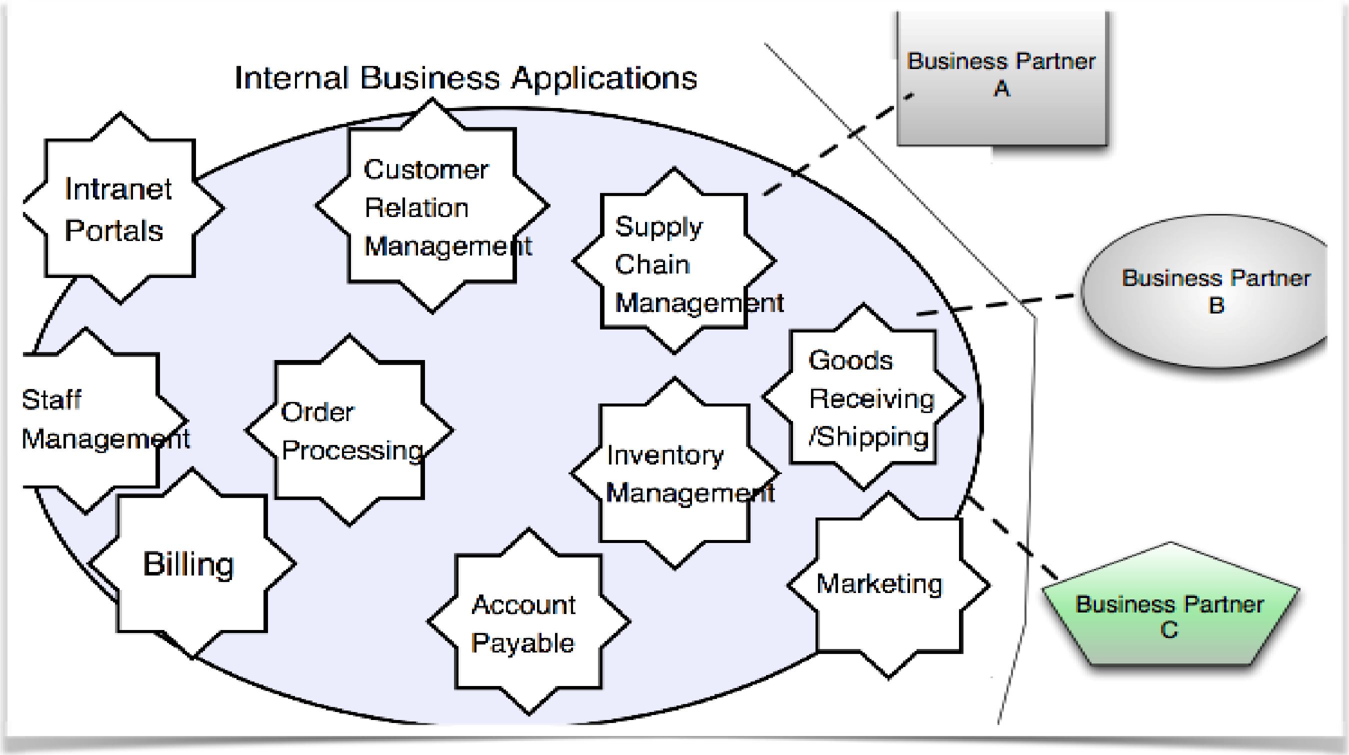
Fig. 2.1 Typical architectures for materialized and virtual data integration [3].

Picture from Mashups: Concepts, Models and Architectures

System Level Integration …

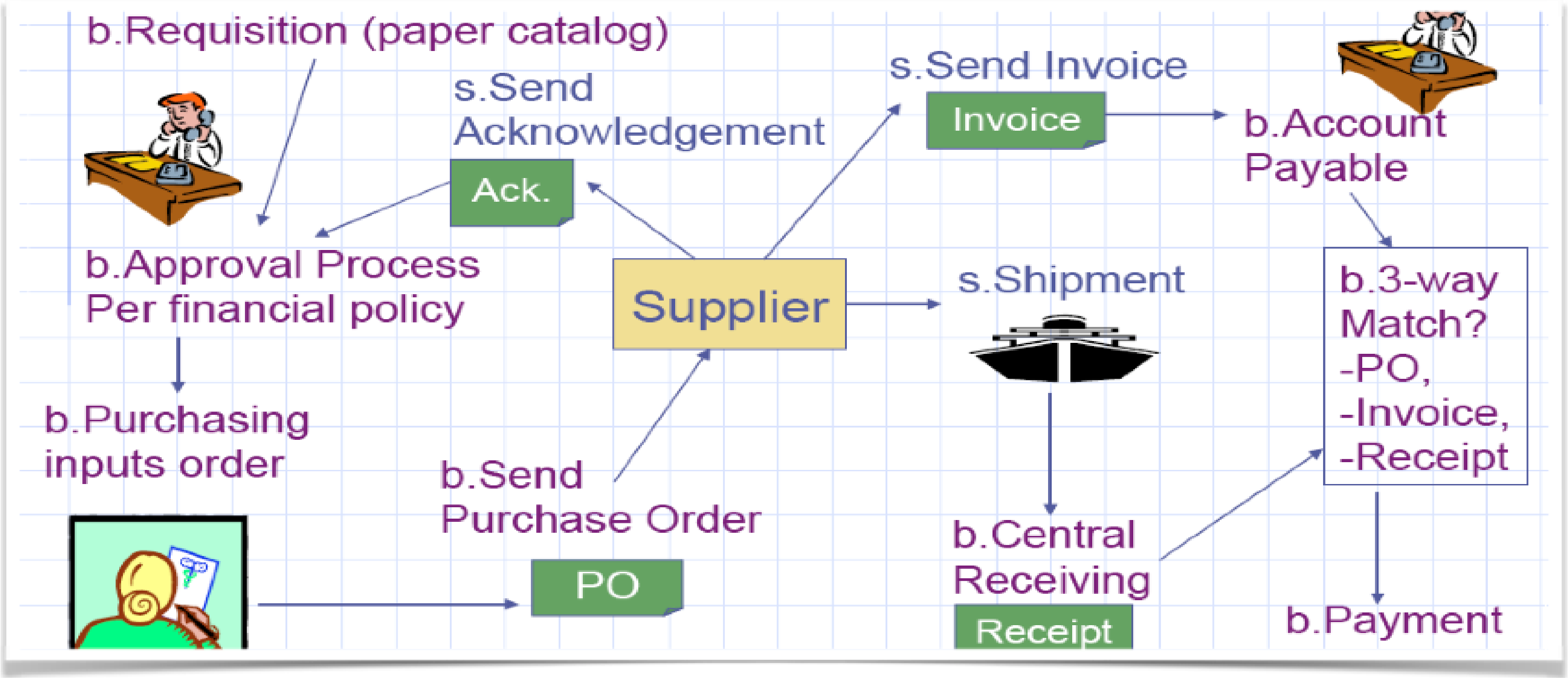
In enterprise environments, pick any sizeable organisation. You will see many departments performing different functionality

In silos, often supported by software systems

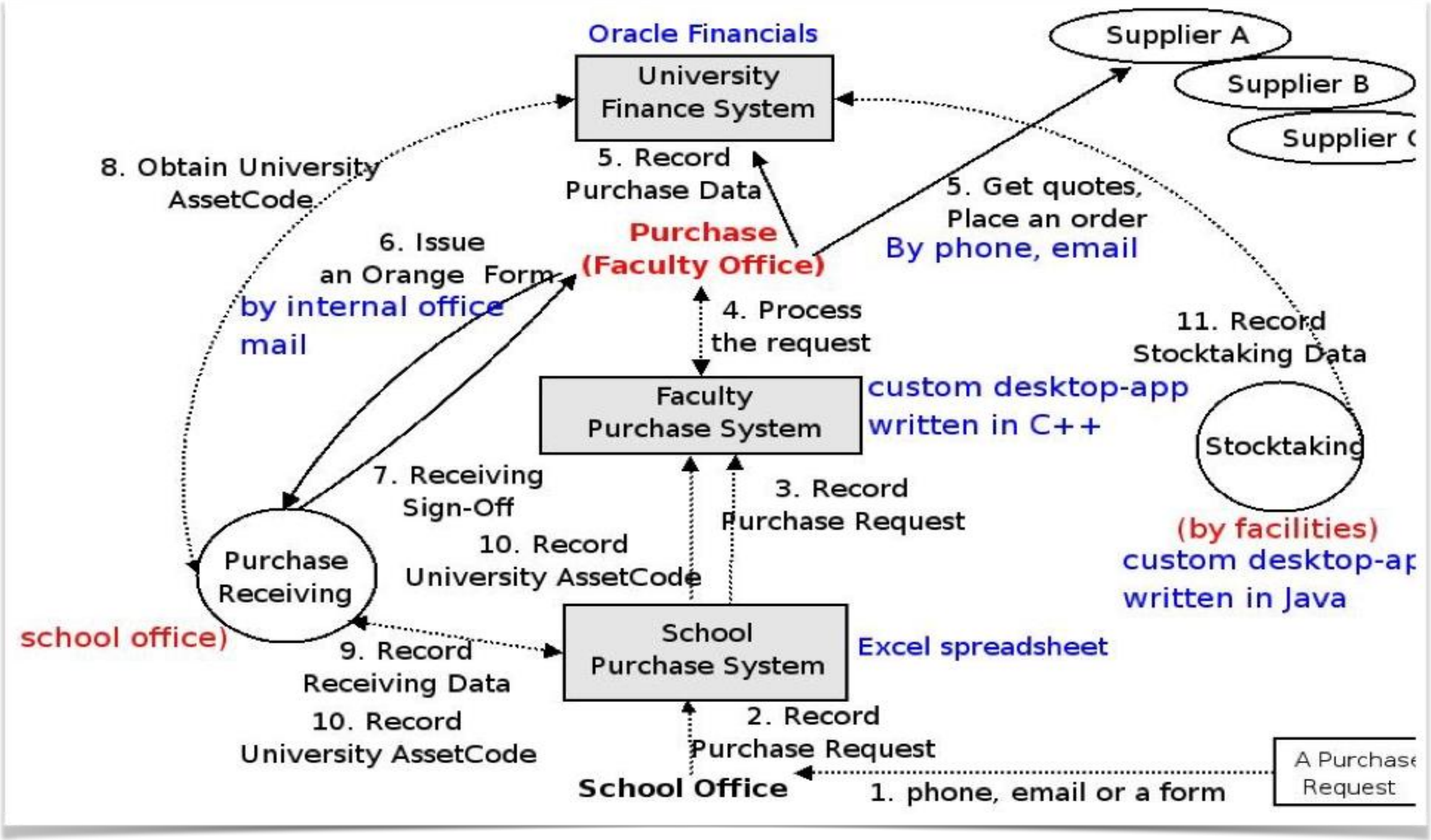


A Typical Purchase Order Process

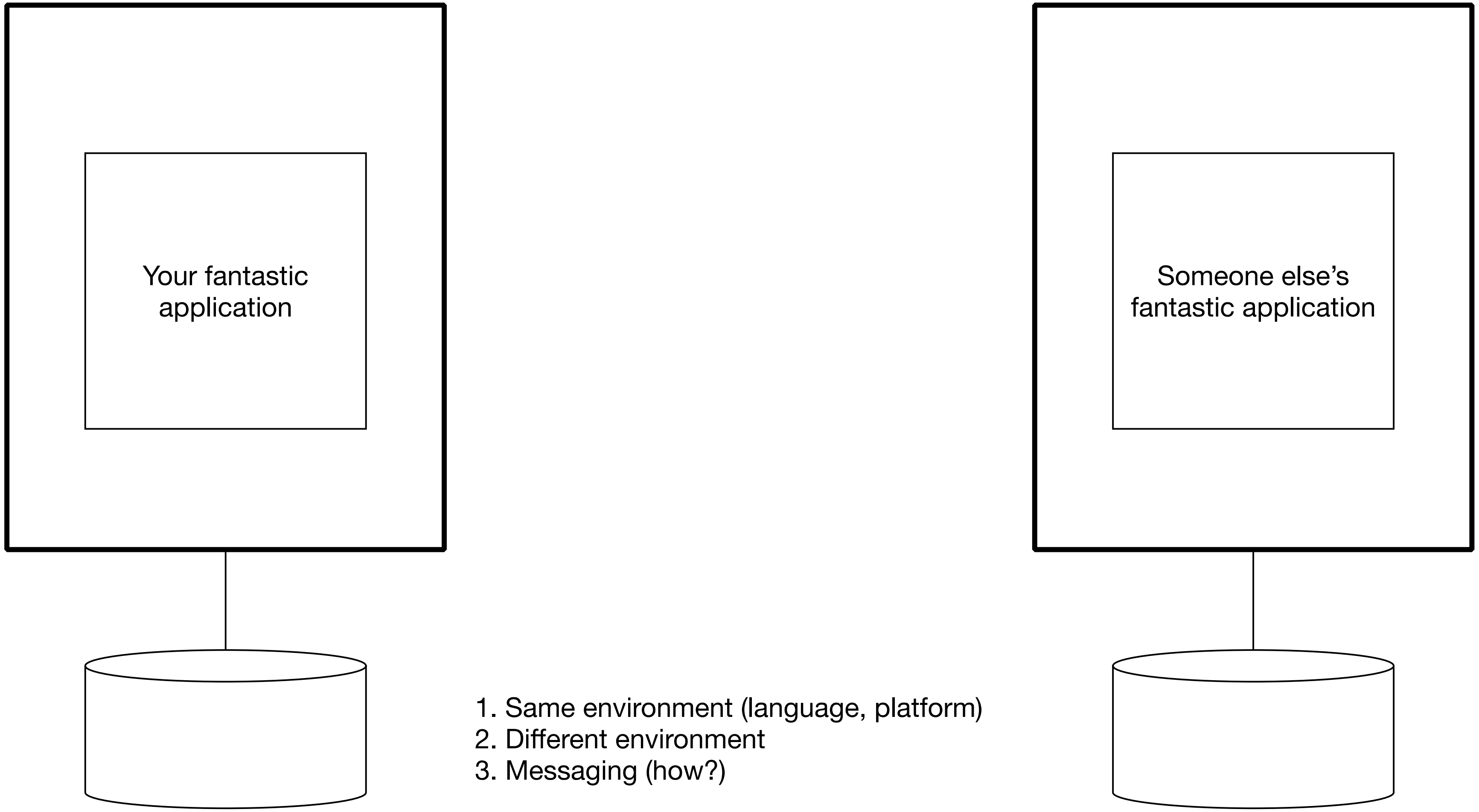
In reality: communication/coordination between the silos needed



An example of (real) Purchase Order Process



Going outside of your system boundary …

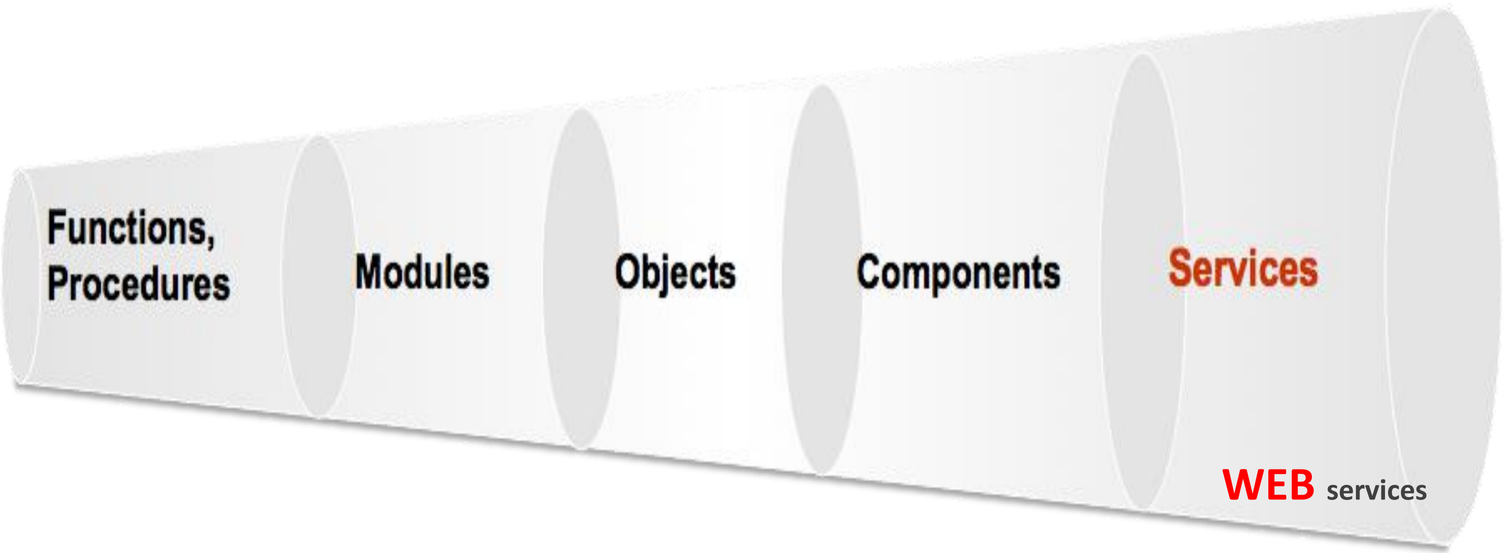

Your fantastic application
Someone else's fantastic application

1. Same environment (language, platform)
2. Different environment
3. Messaging (how?)

The evolution of programming abstractions

Services: “customer” and “service provider”

Lines of code vs. Services - consider software building exercise as ‘building services’, ‘discovering services’ and ‘combining services’



Web = platform/language neutral

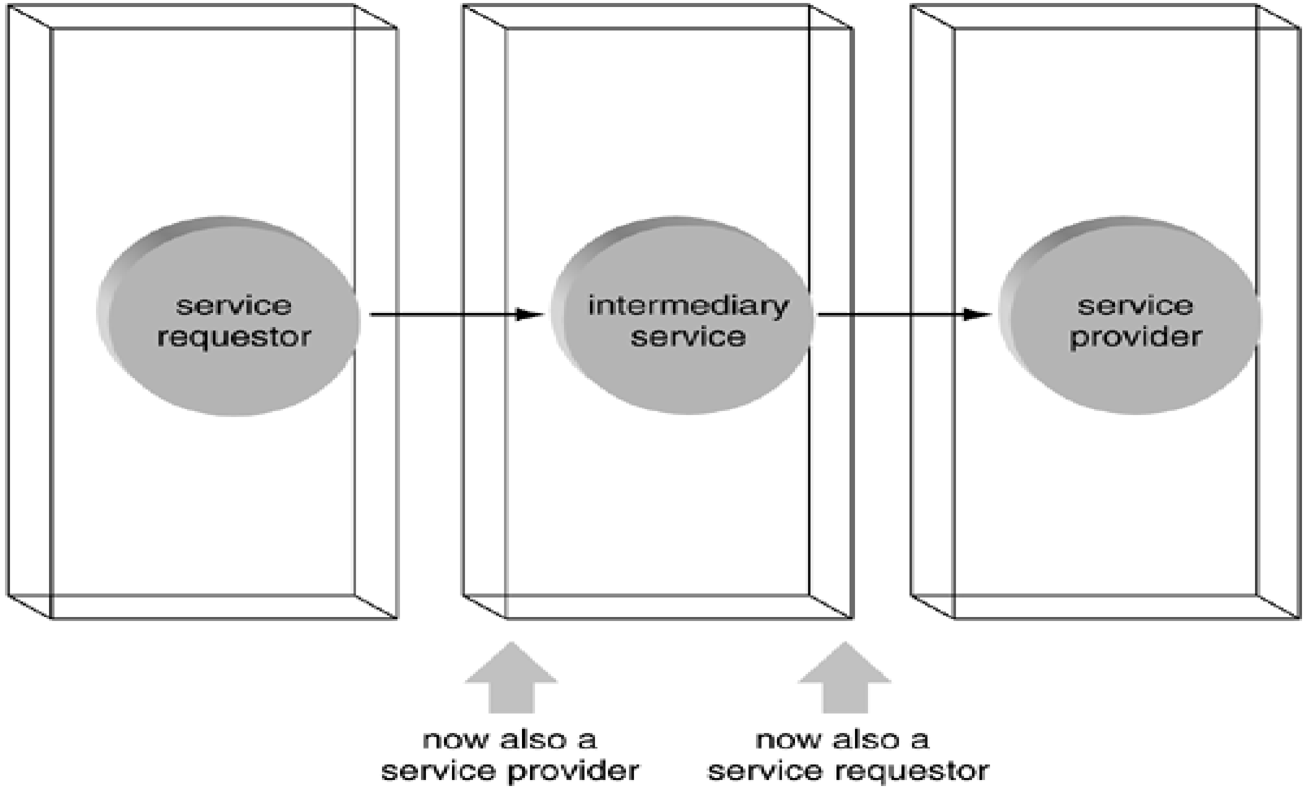
The evolution of programming abstractions

In SOA, we talk about software as a service ... That is, SOA is about building software systems composed of a collection of (software) services

A software service:

- A software asset that is deployed at an endpoint and is continuously maintained by a provider for user by one or multiple clients
- Services have explicit contracts that establish their purpose and how they should be used
- Software services are (supposed to be) reusable (“compose-able”) …
  – like lego blocks
  – “my” (the developer) service could be used in scenarios that I never anticipated

Simplified view of services (or API ?!)



Service-orientation - a way of integrating your applications as a set of linked services. If you can define the services, you can begin to link the services to realize more complicated 'services'

So again…Why Data Services?

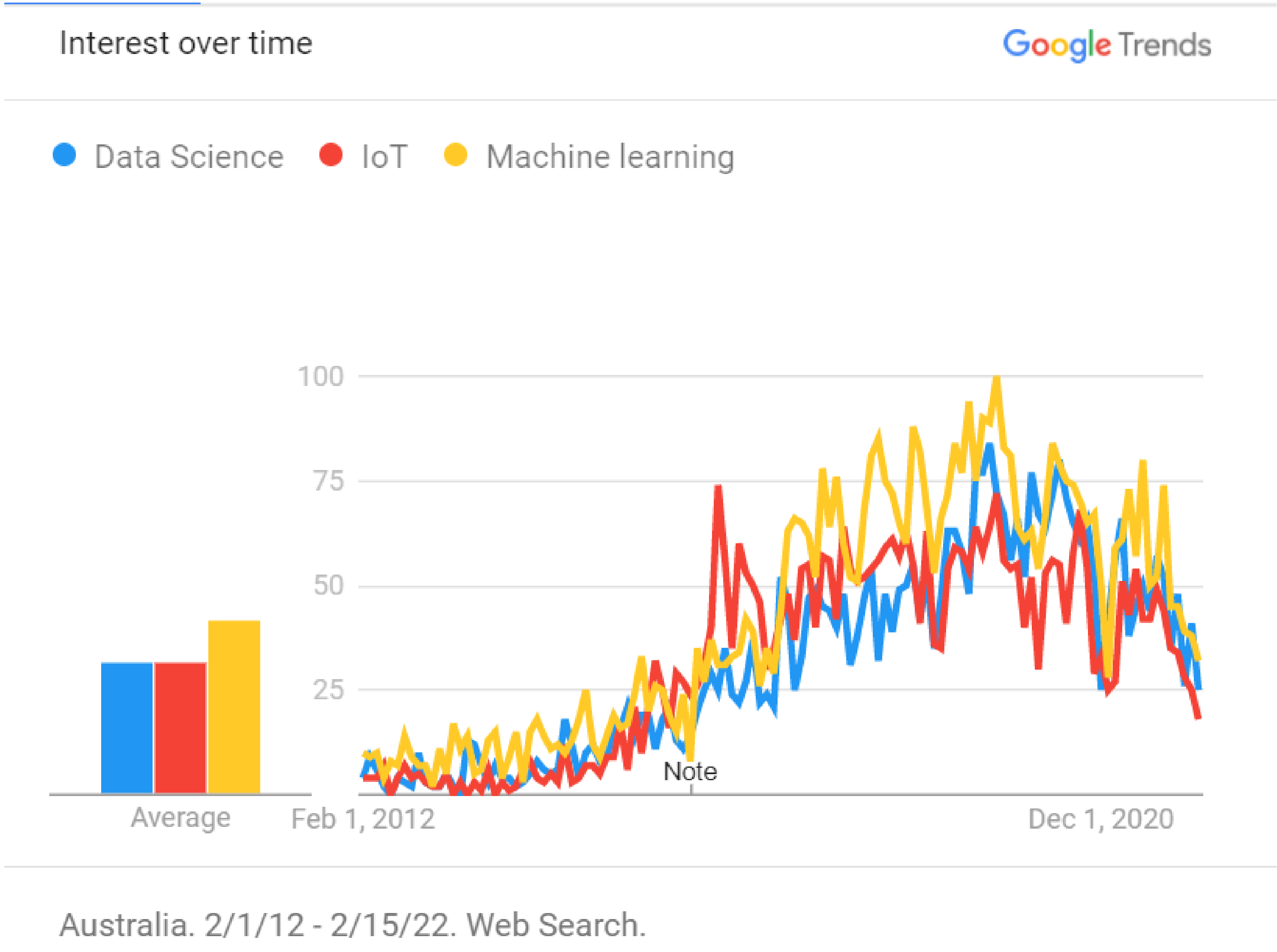
Sexy Job

“I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?”

“The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it’s going to be a at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data.”

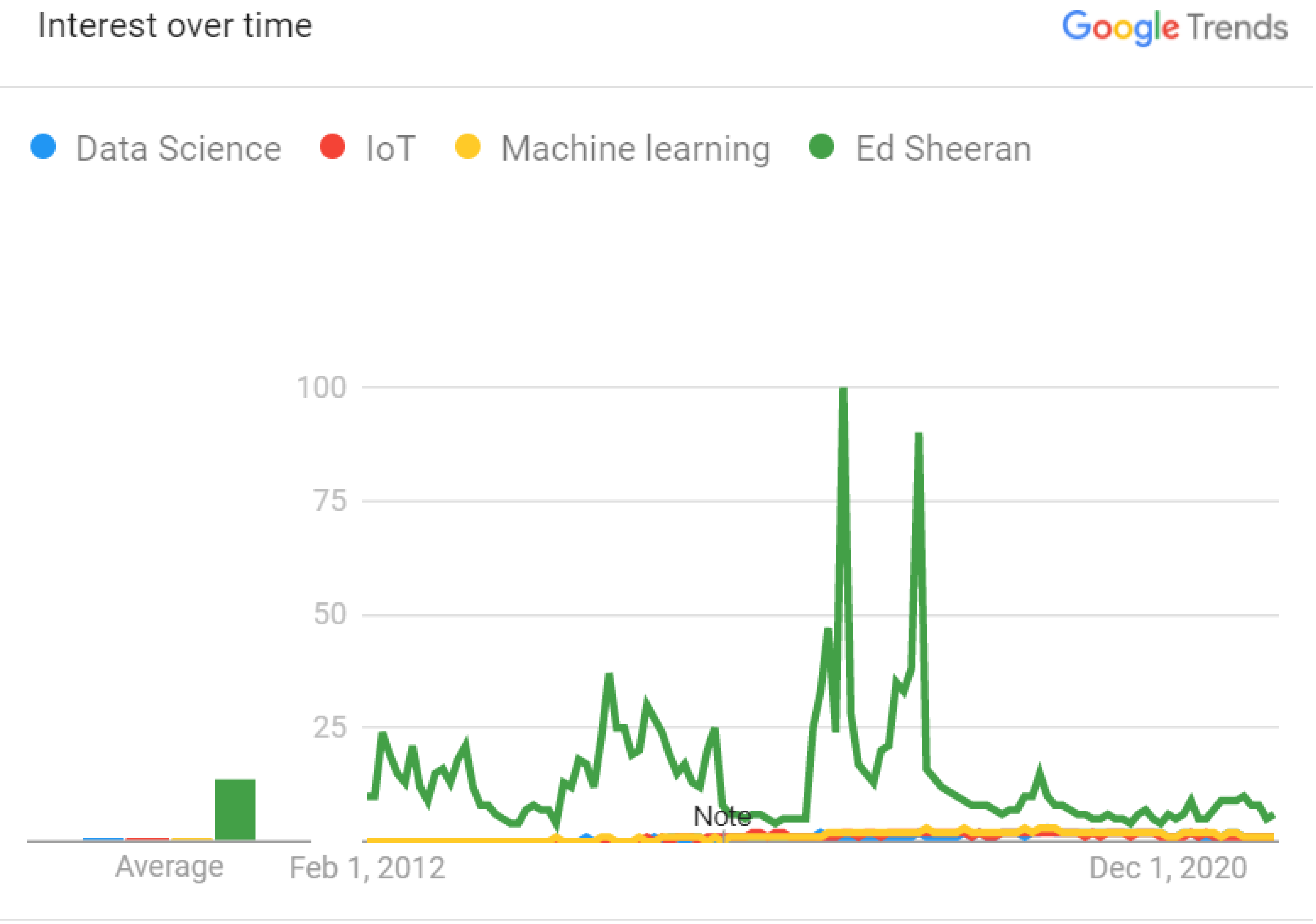
Hal Varian, Google’s Chief Economist

Data is Popular



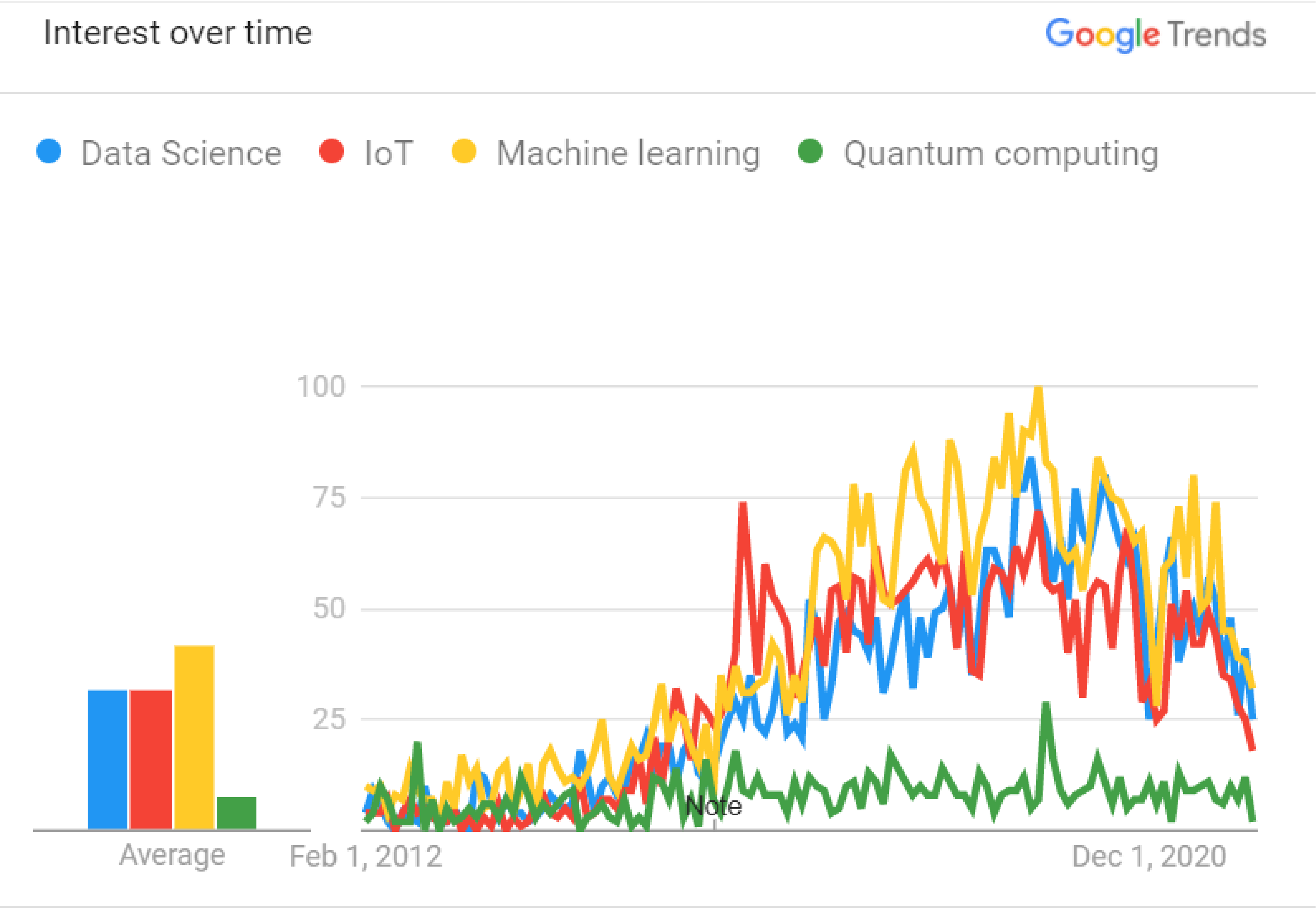
Australia. 2/1/12 - 2/15/22. Web Search.

Data is Popular?



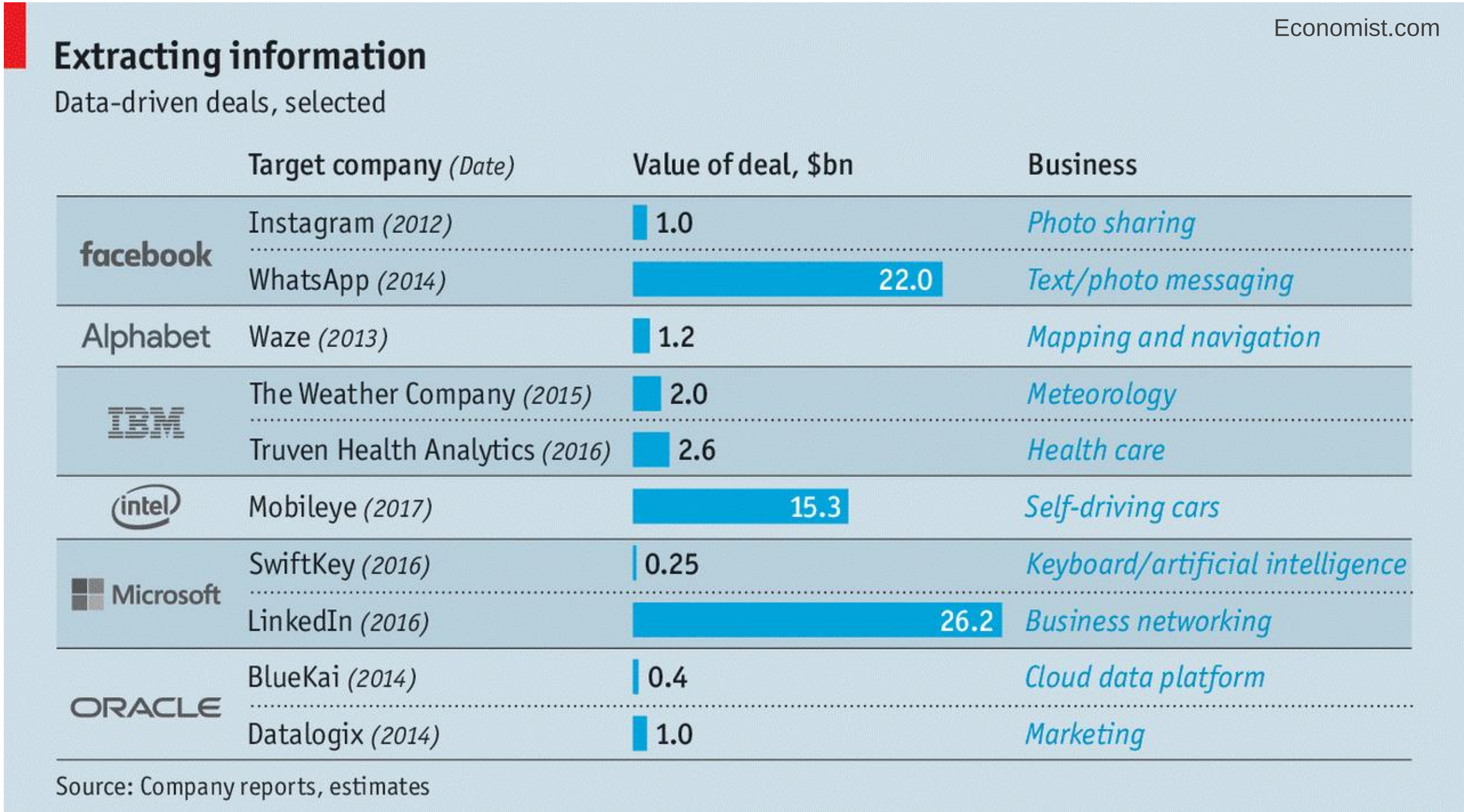
Australia. 2/1/12 - 2/15/22. Web Search.

Data is Popular



Australia. 2/1/12 - 2/15/22. Web Search.

Data is valuable



Economist.com

Extracting information

Data-driven deals, selected

	Target company *(Date)*	Value of deal, $bn	Business
facebook	Instagram *(2012)*	1.0	*Photo sharing*
	WhatsApp *(2014)*	22.0	*Text/photo messaging*
Alphabet	Waze *(2013)*	1.2	*Mapping and navigation*
IBM	The Weather Company *(2015)*	2.0	*Meteorology*
	Truven Health Analytics *(2016)*	2.6	*Health care*
intel	Mobileye *(2017)*	15.3	*Self-driving cars*
Microsoft	SwiftKey *(2016)*	0.25	*Keyboard/artificial intelligence*
	LinkedIn *(2016)*	26.2	*Business networking*
ORACLE	BlueKai *(2014)*	0.4	*Cloud data platform*
	Datalogix *(2014)*	1.0	*Marketing*

Source: Company reports, estimates

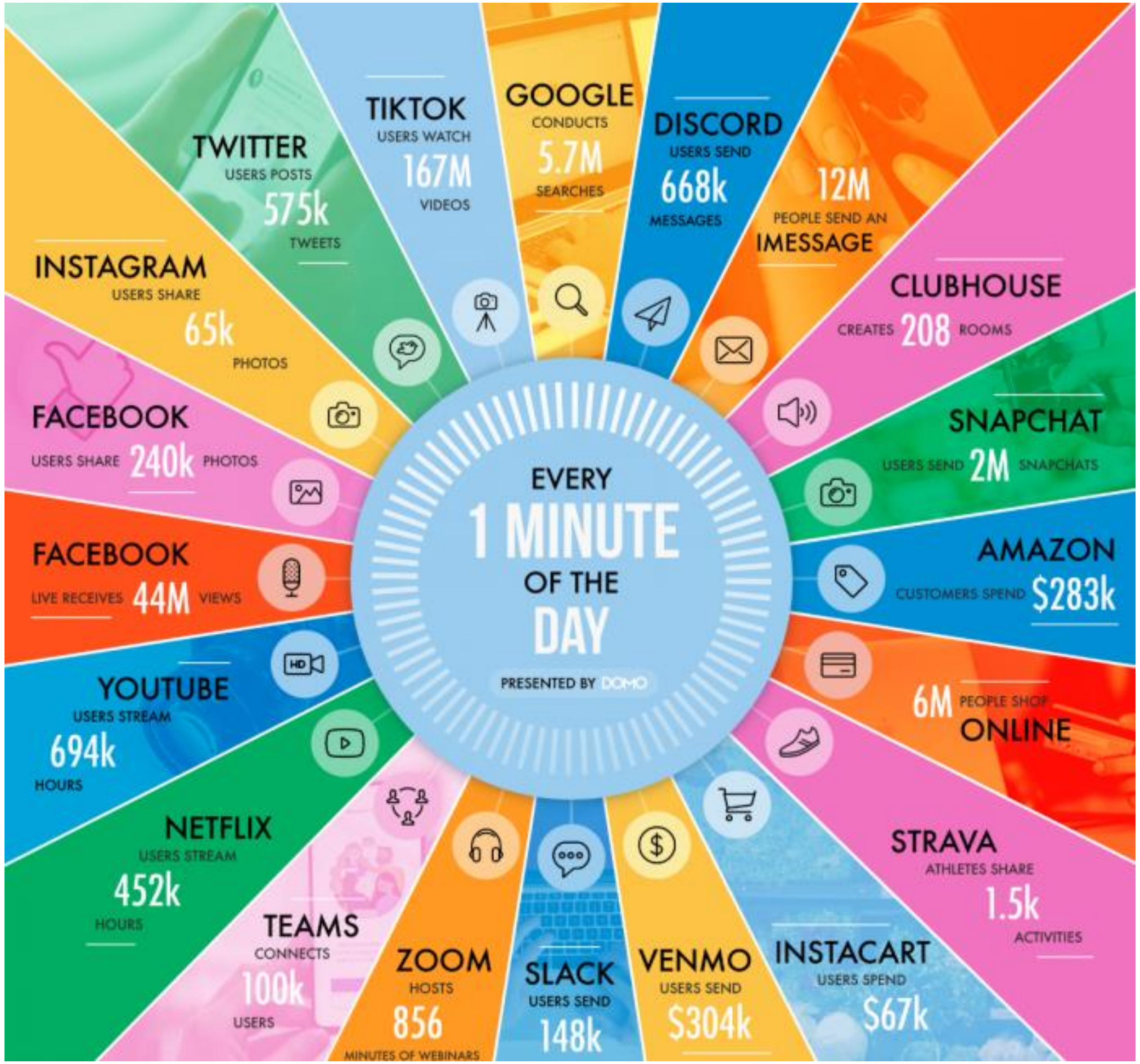
Data is Massive

"There are **2.5 quintillion bytes (EB)** of data created **each day** at our current pace, but that pace is **only accelerating** with the growth of the Internet of Things (IoT)." [1]

- 2.7 Zetabytes of data exist in the digital universe today.[2]
- Facebook stores, accesses, and analyzes 30+ Petabytes of user generated data.[3]
- Akamai analyzes 75 million events per day to better target advertisements.[3]
- Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.[4]
- In 2008, Google was processing 20,000 terabytes of data (20 petabytes) a day.[5]

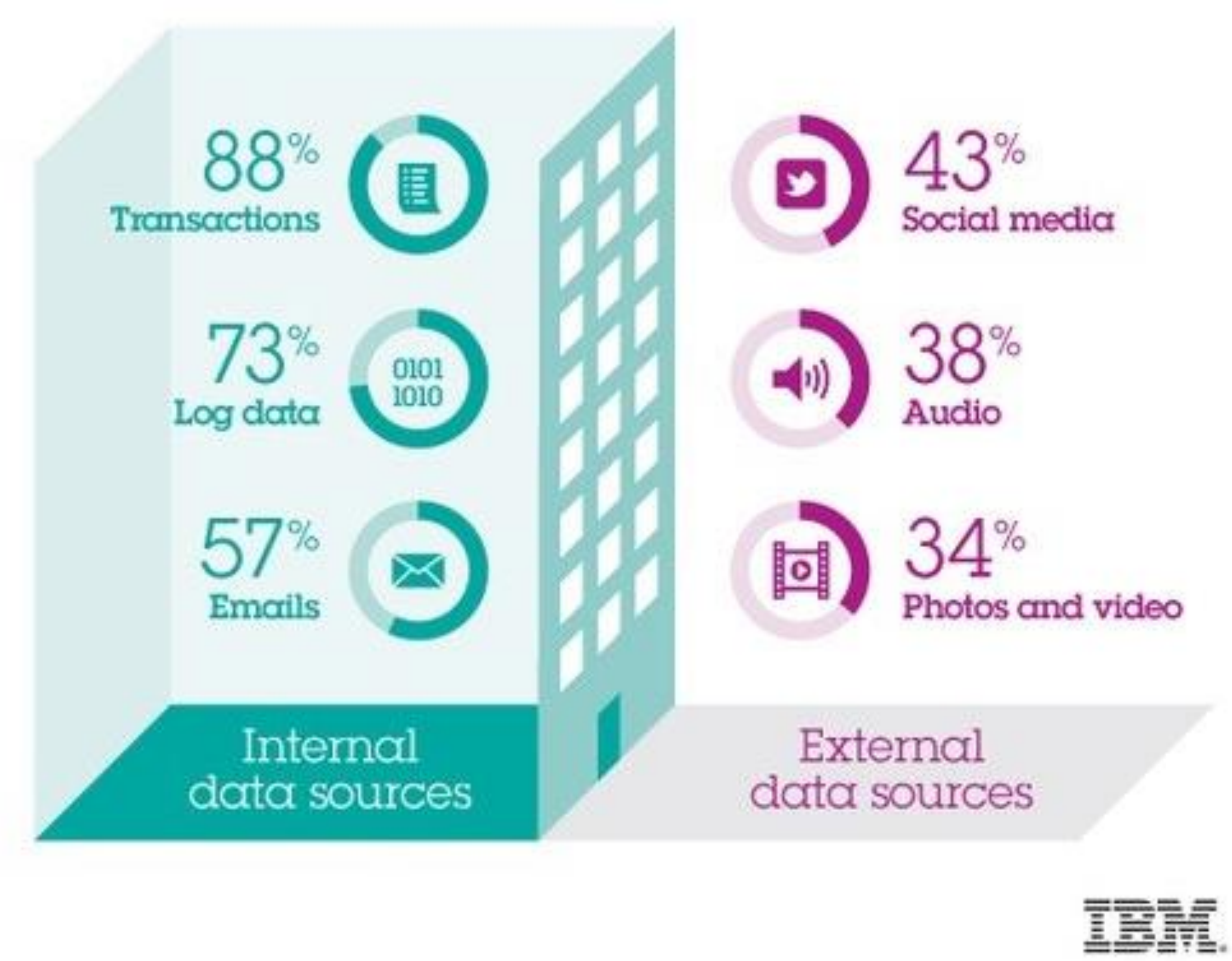
1. Forbes, *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*
2. Wikibon*, The Rapid Growth in Unstructured Data*
3. Wikibon*, Taming Big Data*
4. SAS*, Big Data Meets Big Data Analytics*
5. TechCrunch*, Google Processing 20,000 Terabytes A Day, And Growing*

Data Every Minute

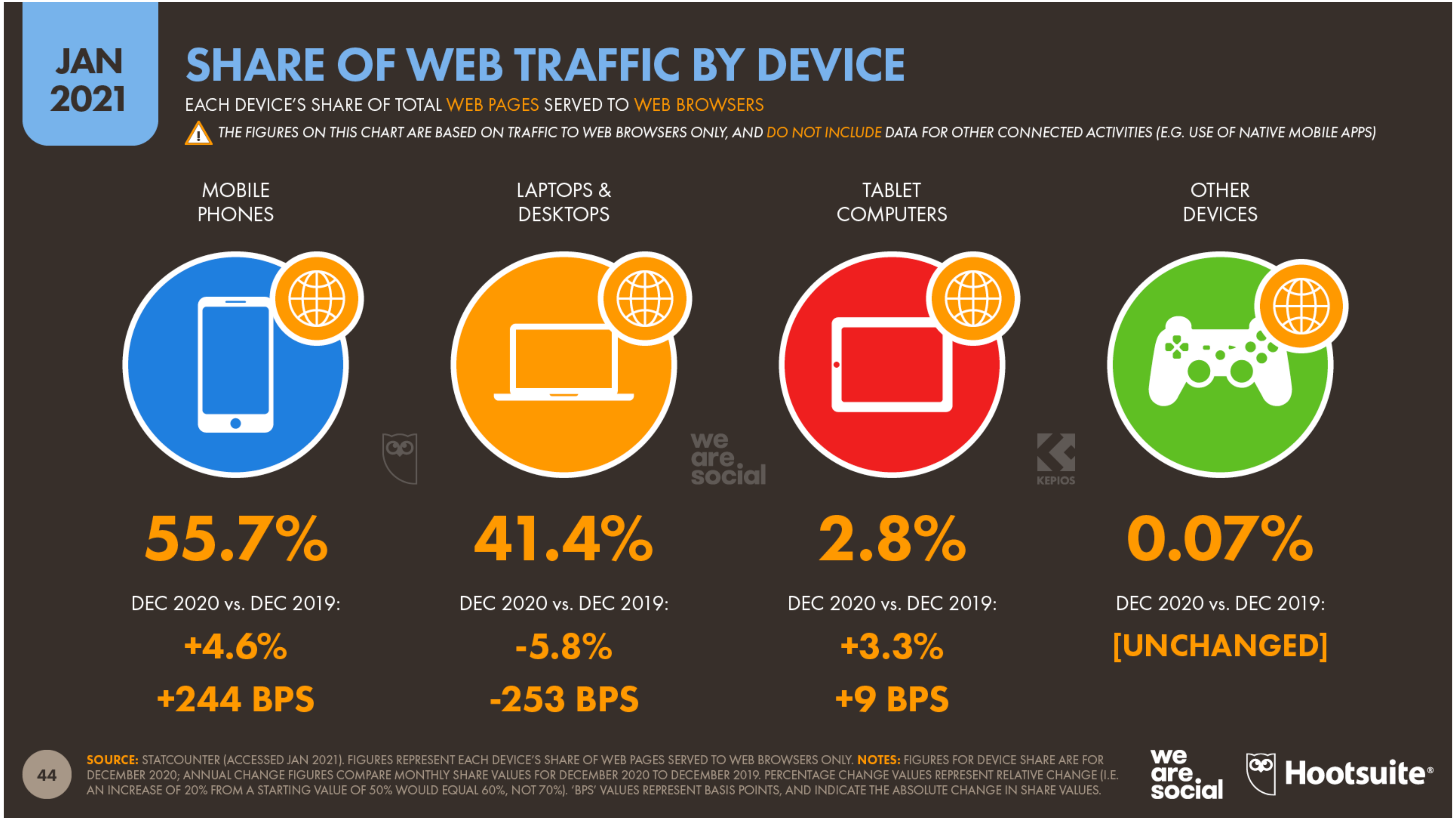


https://dailyinfographic.com/how-much-data-is-generated-every-minute

Where does data come from?



Where does data come from?



JAN 2021

SHARE OF WEB TRAFFIC BY DEVICE

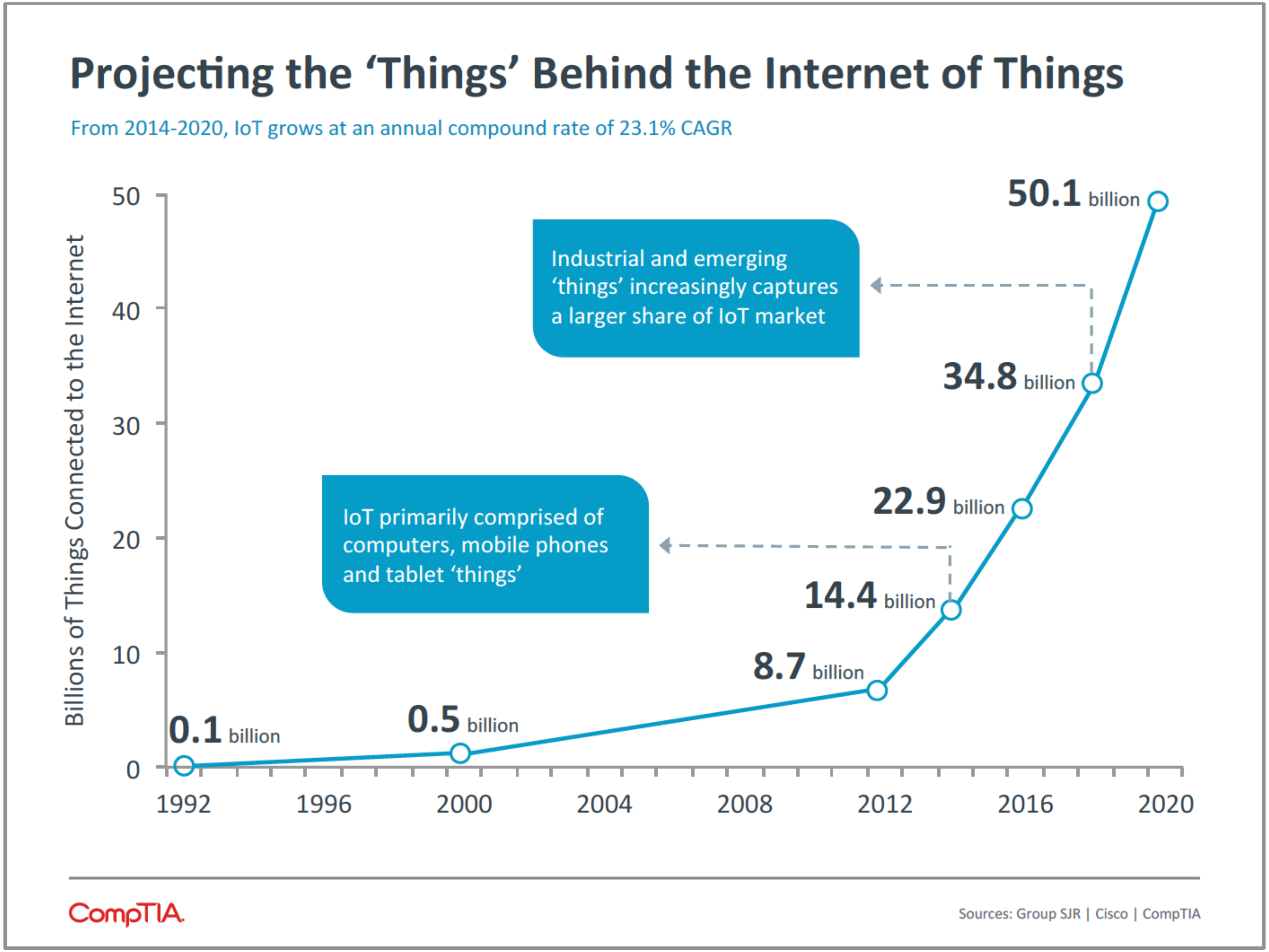
EACH DEVICE'S SHARE OF TOTAL WEB PAGES SERVED TO WEB BROWSERS

⚠ *THE FIGURES ON THIS CHART ARE BASED ON TRAFFIC TO WEB BROWSERS ONLY, AND DO NOT INCLUDE DATA FOR OTHER CONNECTED ACTIVITIES (E.G. USE OF NATIVE MOBILE APPS)*

MOBILE PHONES	LAPTOPS & DESKTOPS	TABLET COMPUTERS	OTHER DEVICES
55.7%	41.4%	2.8%	0.07%
DEC 2020 vs. DEC 2019:	DEC 2020 vs. DEC 2019:	DEC 2020 vs. DEC 2019:	DEC 2020 vs. DEC 2019:
+4.6%	-5.8%	+3.3%	[UNCHANGED]
+244 BPS	-253 BPS	+9 BPS	

44

SOURCE: STATCOUNTER (ACCESSED JAN 2021). FIGURES REPRESENT EACH DEVICE'S SHARE OF WEB PAGES SERVED TO WEB BROWSERS ONLY. **NOTES:** FIGURES FOR DEVICE SHARE ARE FOR DECEMBER 2020; ANNUAL CHANGE FIGURES COMPARE MONTHLY SHARE VALUES FOR DECEMBER 2020 TO DECEMBER 2019. PERCENTAGE CHANGE VALUES REPRESENT RELATIVE CHANGE (I.E. AN INCREASE OF 20% FROM A STARTING VALUE OF 50% WOULD EQUAL 60%, NOT 70%). 'BPS' VALUES REPRESENT BASIS POINTS, AND INDICATE THE ABSOLUTE CHANGE IN SHARE VALUES.



Projecting the ‘Things’ Behind the Internet of Things

From 2014-2020, IoT grows at an annual compound rate of 23.1% CAGR

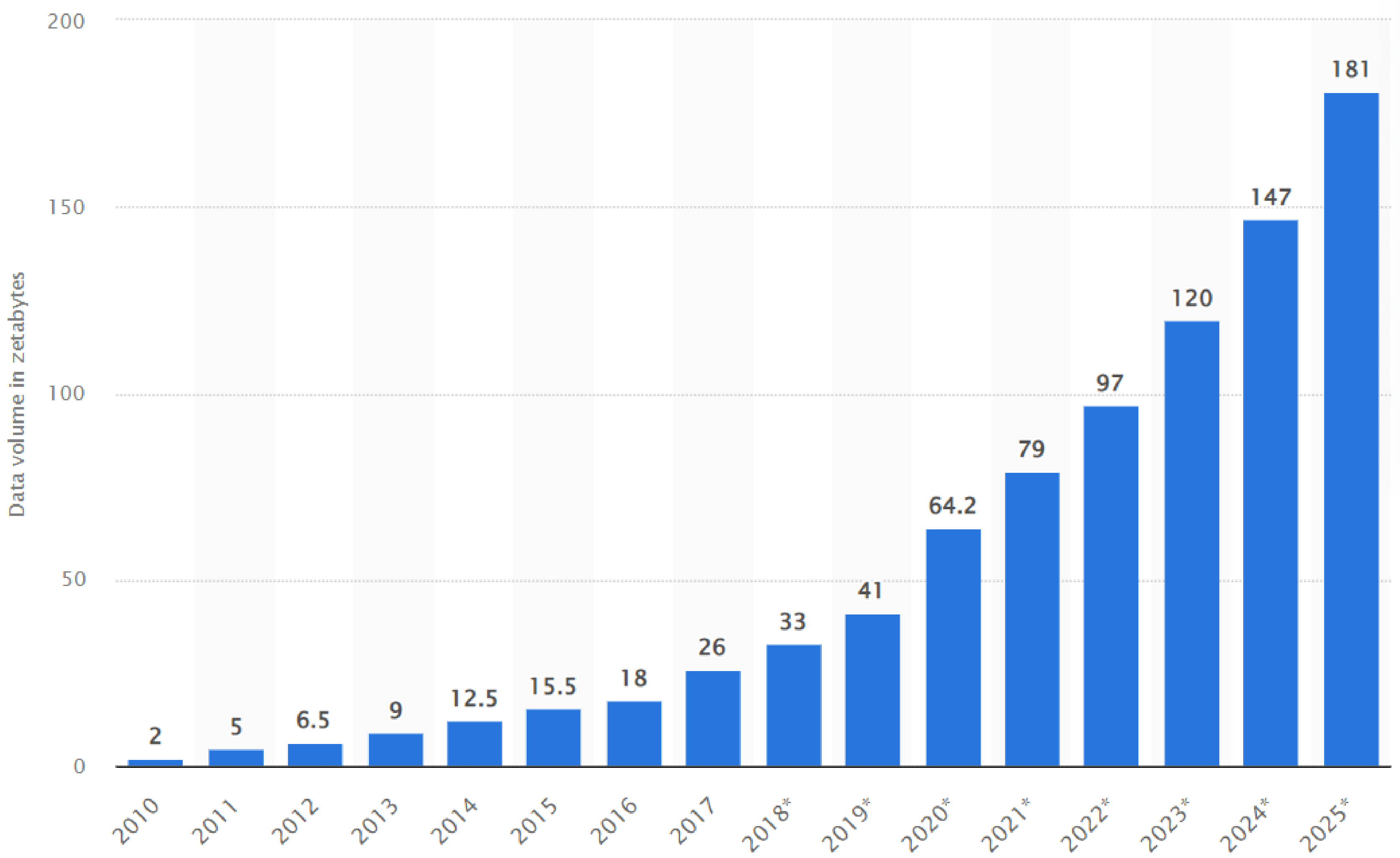
Industrial and emerging ‘things’ increasingly captures a larger share of IoT market

IoT primarily comprised of computers, mobile phones and tablet ‘things’

Sources: Group SJR | Cisco | CompTIA

The connected world

Even more data…



https://www.statista.com/statistics/871513/worldwide-data-created/

What can we do with the data?

Life Sciences

Clinical research is a slow and expensive process, with trials failing for a variety of reasons. Advanced analytics, artificial intelligence (AI) and the Internet of Medical Things (IoMT) unlocks the potential of improving speed and efficiency at every stage of clinical research by delivering more intelligent, automated solutions.

Banking

Financial institutions gather and access analytical insight from large volumes of unstructured data in order to make sound financial decisions. Big data analytics allows them to access the information they need when they need it, by eliminating overlapping, redundant tools and systems.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

What can we do with the data?

Manufacturing

For manufacturers, solving problems is nothing new. They wrestle with difficult problems on a daily basis - from complex supply chains, to motion applications, to labor constraints and equipment breakdowns. That's why big data analytics is essential in the manufacturing industry, as it has allowed competitive organizations to discover new cost saving opportunities and revenue opportunities.

Health Care

Big data is a given in the health care industry. Patient records, health plans, insurance information and other types of information can be difficult to manage – but are full of key insights once analytics are applied. That’s why big data analytics technology is so important to heath care. By analyzing large amounts of information – both structured and unstructured – quickly, health care providers can provide lifesaving diagnoses or treatment options almost immediately.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

What can we do with the data?

Government

Certain government agencies face a big challenge: tighten the budget without compromising quality or productivity. This is particularly troublesome with law enforcement agencies, which are struggling to keep crime rates down with relatively scarce resources. And that's why many agencies use big data analytics; the technology streamlines operations while giving the agency a more holistic view of criminal activity.

Retail

Customer service has evolved in the past several years, as savvier shoppers expect retailers to understand exactly what they need, when they need it. Big data analytics technology helps retailers meet those demands. Armed with endless amounts of data from customer loyalty programs, buying habits and other sources, retailers not only have an in-depth understanding of their customers, they can also predict trends, recommend new products – and boost profitability.

Source: https://www.sas.com/en_au/insights/analytics/big-data-analytics.html

Also

Spam/False Information Detection

Credit card fraud detection

Recommendation systems

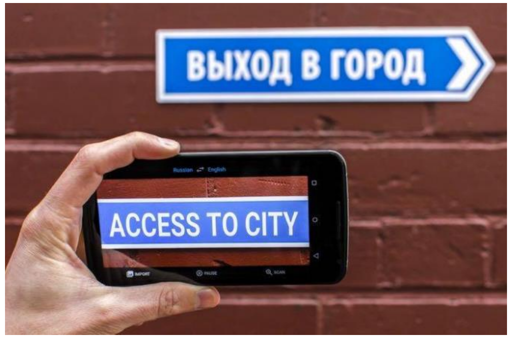
Human activity recognition/prediction

Machine translation

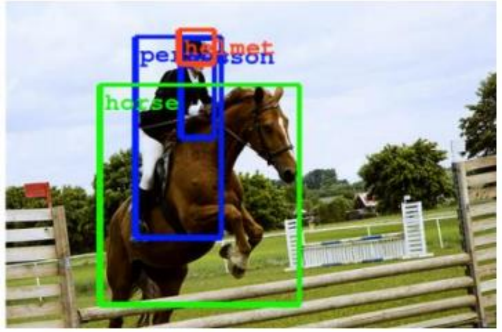
Face/Scene recognition

Image caption

Self-driving cars



a cat is sitting on a toilet seat
logprob: -7.79



Unraveling Power of Deeply Connected World

- Produce a treasure trove of big data
  - » data that can help cities predict accidents and crimes
- Give doctors real-time insight into information from pacemakers or biochips
  - » enable optimized productivity across industries through predictive maintenance on equipment and machinery
- Create true smart homes with connected appliances
- Provide critical communication between self-driving cars
- …

Looks promising…Yet how?

What do you do with all this data?

- » Too much data to search through it manually or processing in traditional ways…

But there is valuable information in the data:

- » How can we use it for fun, profit, and/or the greater good?

Boosting in computing power helps.

- » ***Machine learning*** is key tool we use to make sense of very large datasets.

So What is Next?

- In order to build a data service you need to know how to work with data
  - Accessing the data from multiple sources
  - Exploring the Data
  - Cleansing the data (e.g., removing corrupted or useless data)
  - Manipulating the data (e.g., merging, transformation, normalization)
  - Presenting the data (visualization)

Useful Reading

- **View-based Data Integration**, Yannis Katsis (http://db.ucsd.edu/wp-content/uploads/pdfs/355.pdf)
- **Mashups: Concepts, Models and Architectures,** Daniel, Florian, Matera, Maristella (https://link.springer.com/book/10.1007%2F978-3-642-55049-2)

Q&A