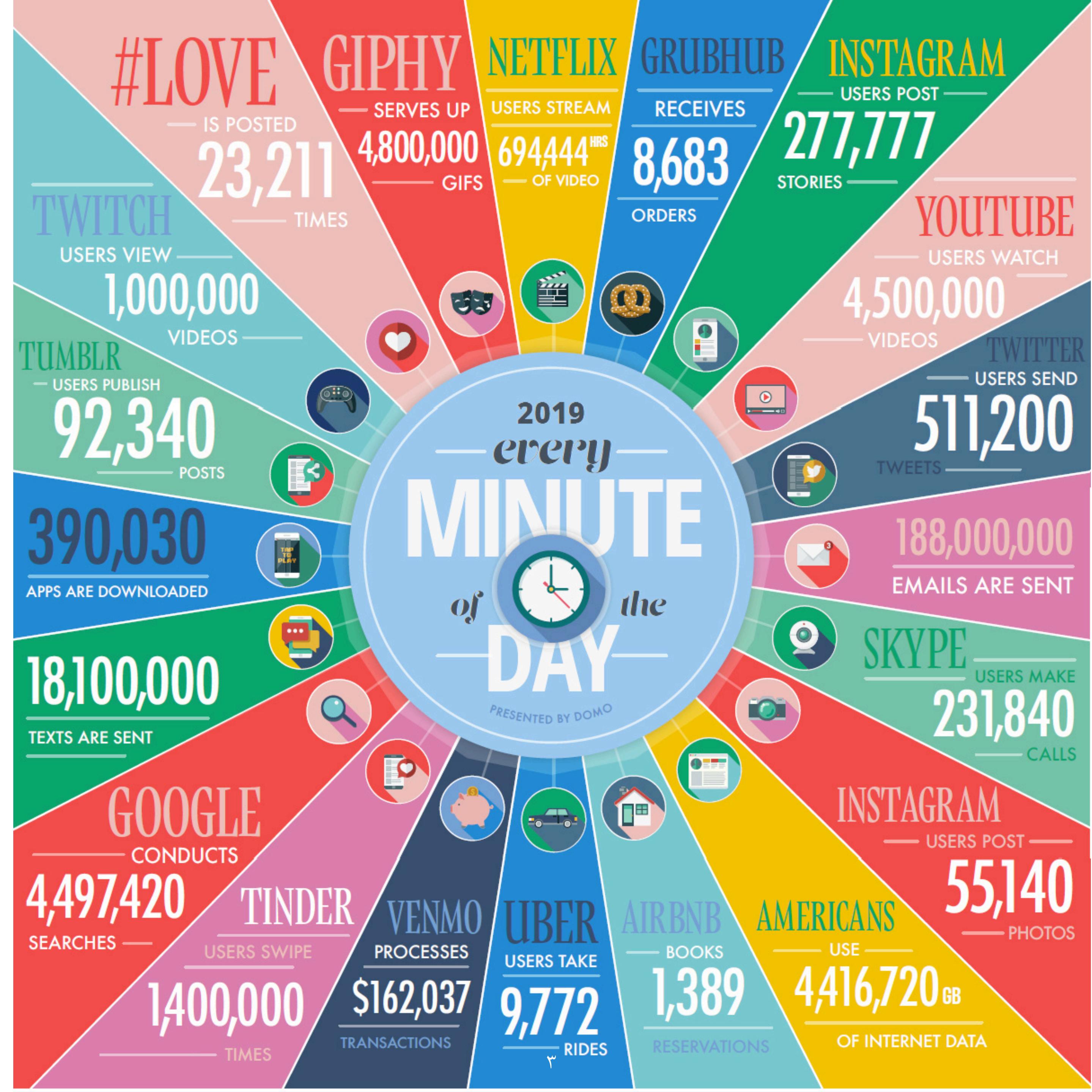


Introduction to Big Data and Its Applications

Hamed Malek – ٢٠٢٢

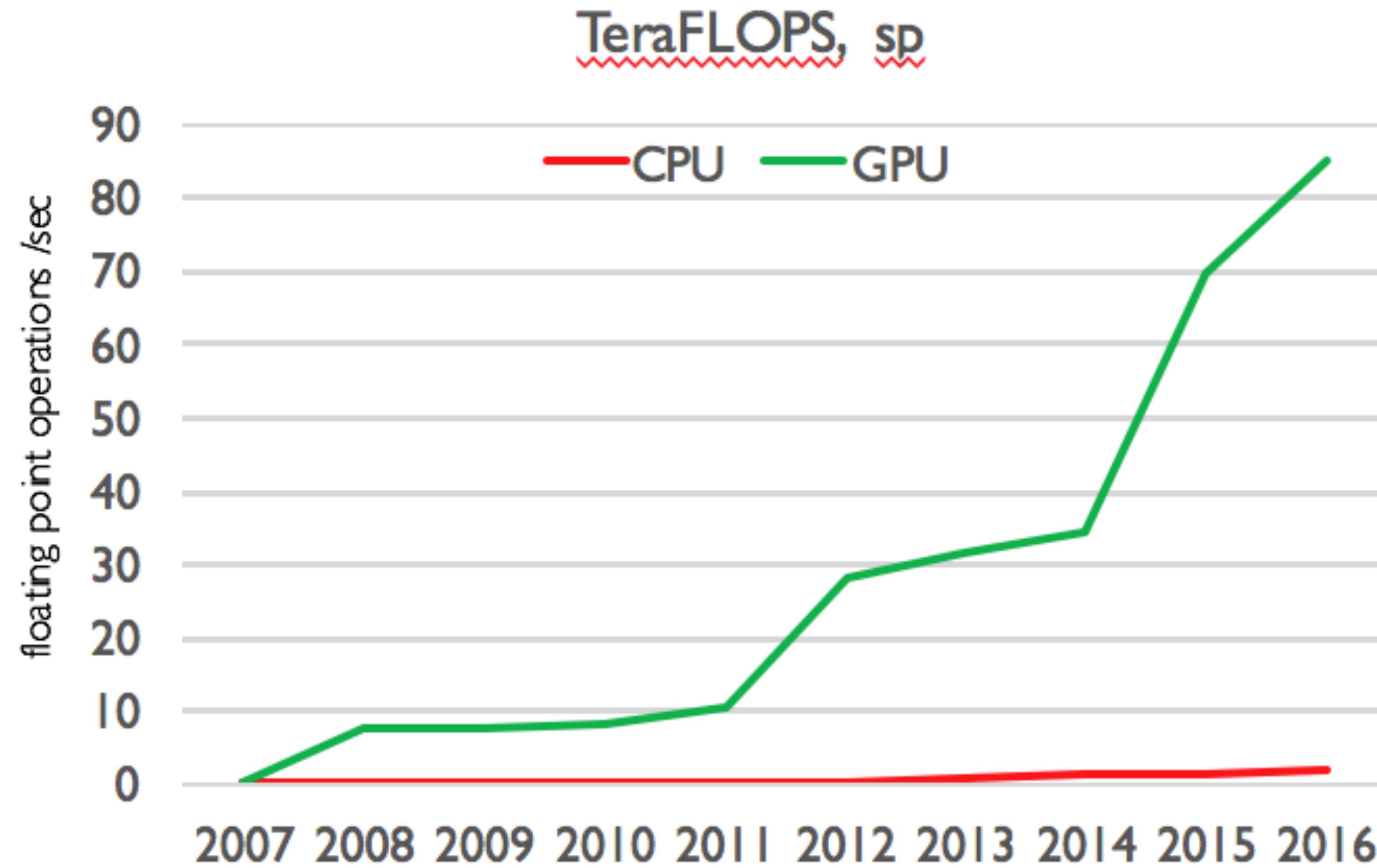
Big Data = مهادهه‌ها

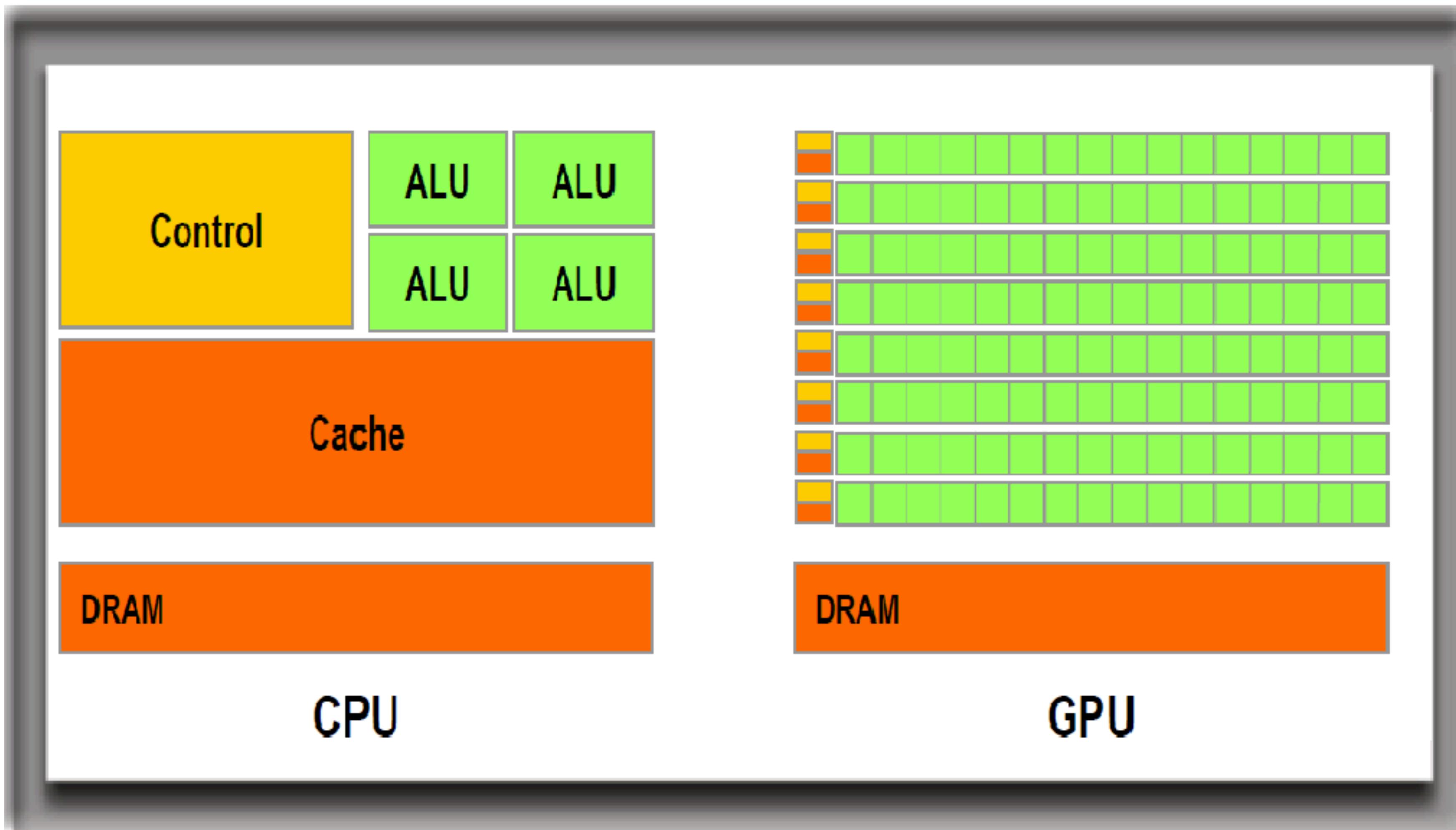
فرهنگستان زبان ✓ مصوب



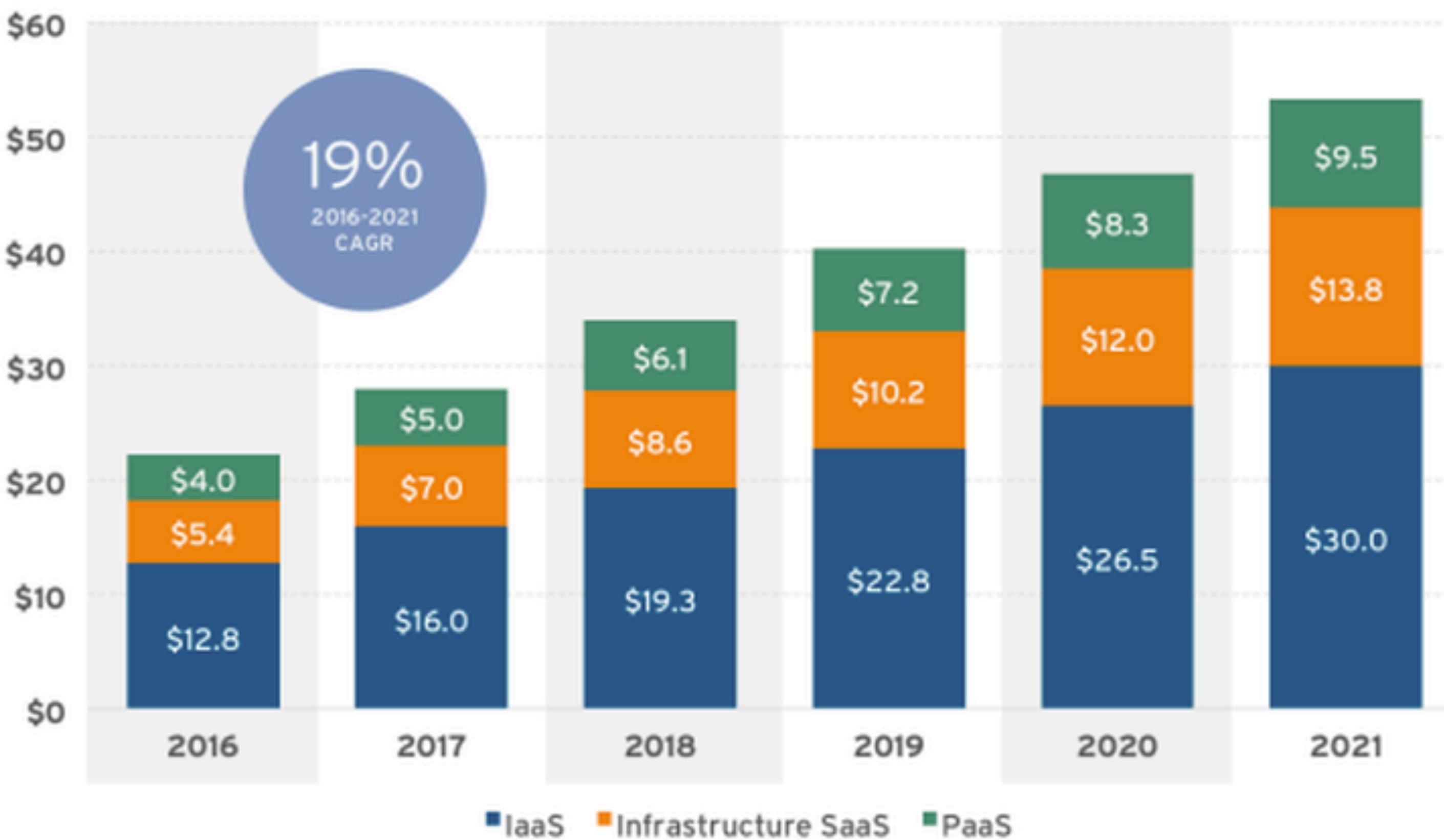
چرا مهاددها؟

- دو عامل اصلی منجر به توجه روز افزون به مهاددها شد:
 - افزایش شدید حجم دادهها
 - افزایش قدرت پردازش از طریق GPU و سرویس‌های ابری





Cloud Computing 'as a Service' Revenue (\$bn)



کاربردهای مددادهای

Banking And Securities



Media and Entertainment



Insurance



HealthCare



Big data Applications



Transportation

Energy and Utilities

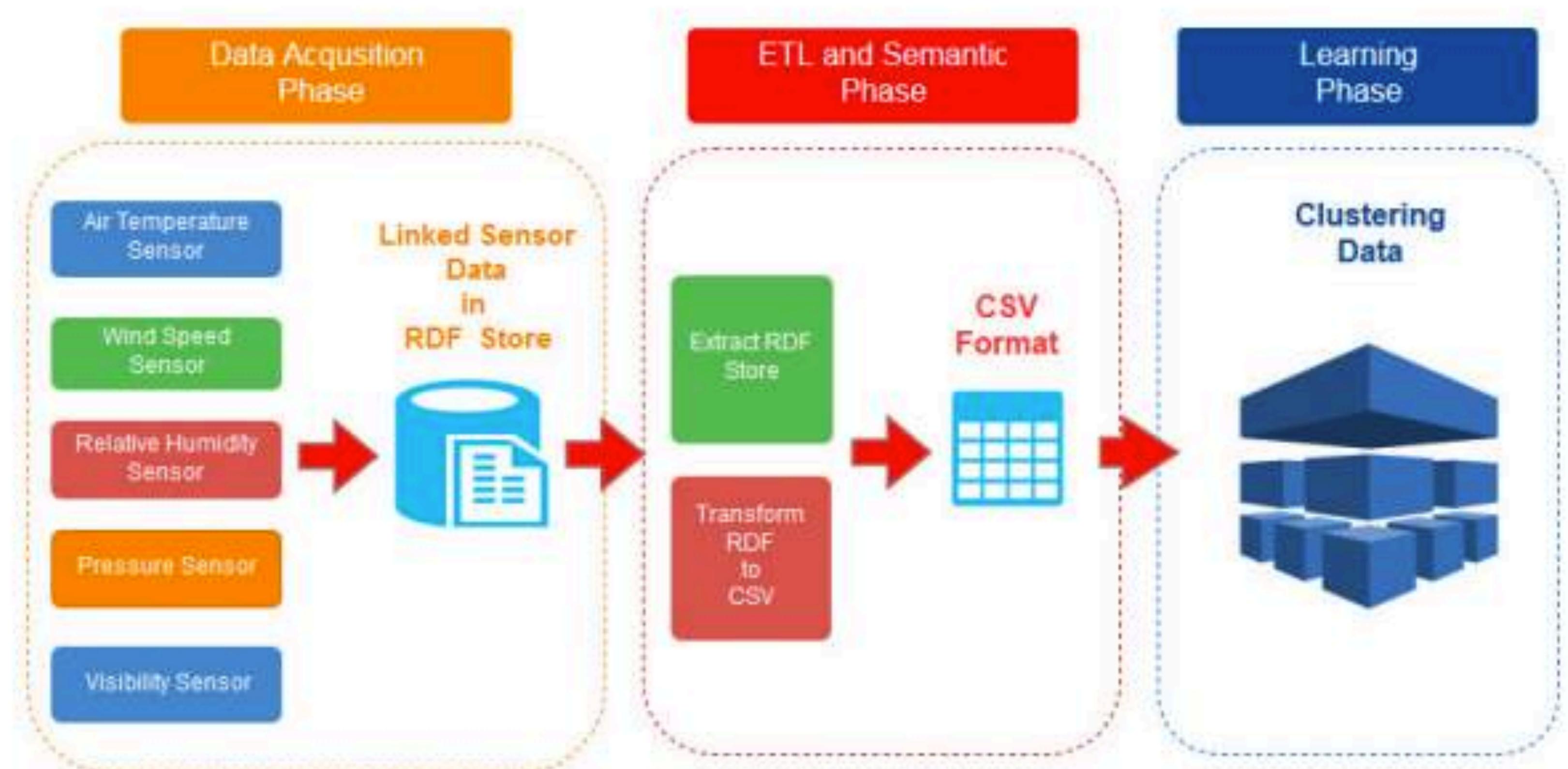


Education



Manufacturing

هواشناسی



Use-case Scenario of our proposed IoT Framework

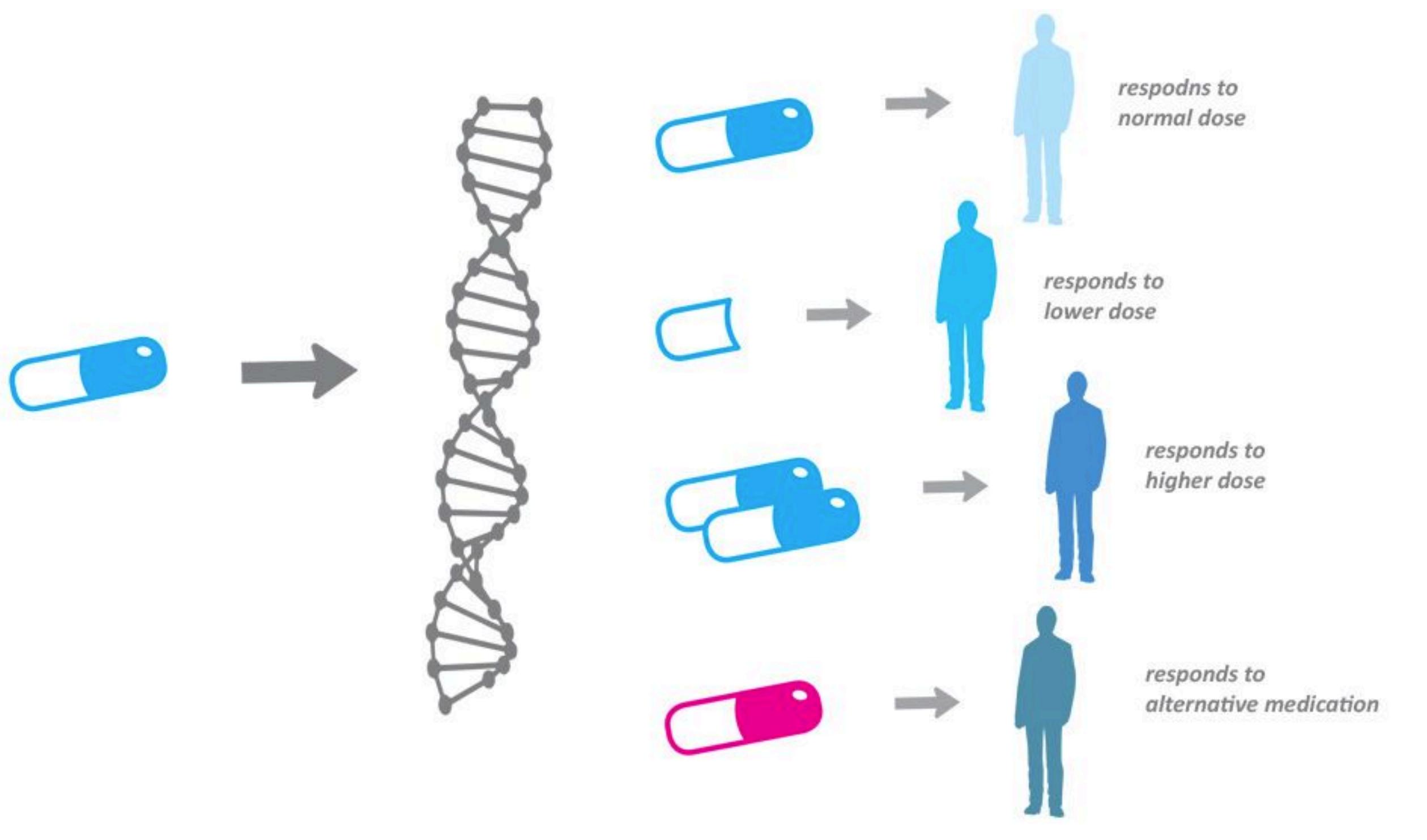
- پیش‌بینی:
- دما
- فشار
- بارندگی
- باد

تصاویر ماهواره‌ای



- کاربردهای:
- مخابراتی
- نظامی
- حمل و نقل
- صوت و تصویر
- اینترنت

پزشکی دقیق



- شخصی سازی پزشکی
- درمان موثر تر
- استفاده از حجم بالای داده:
 - داده های ژنتیکی
 - سوابق قبلی پزشکی
 - تصاویر پزشکی
 - سوابق خانواده

تشخیص آتش سوزی



- استفاده از منابع مختلف داده:
 - سنسورها
 - دوربین‌های نظارتی
 - تماس‌های تلفنی
 - پیام‌ها در توئیتر
 - تصاویر ماهواره‌ای

مخابرات

Role of Big Data in Telecom Industry



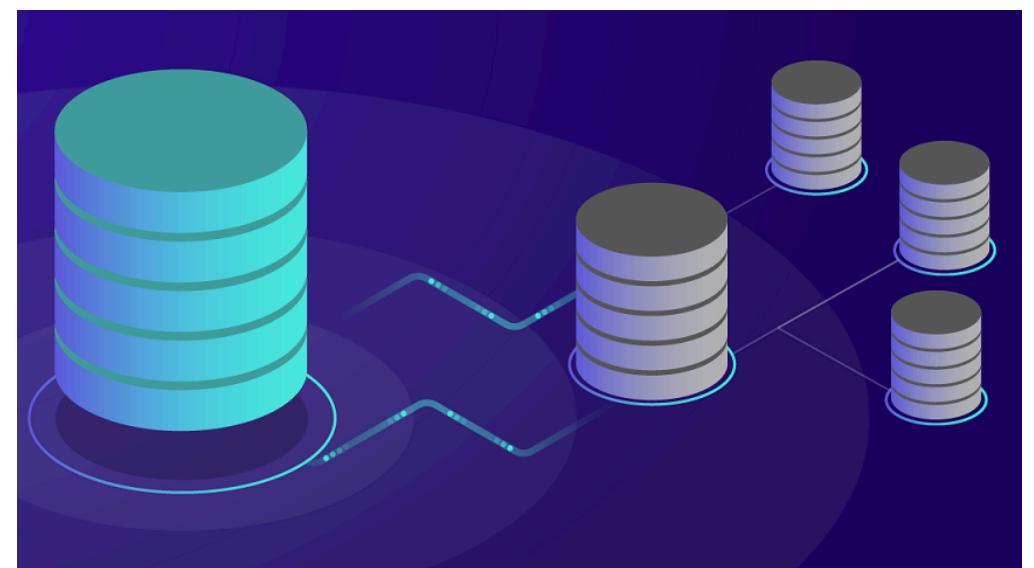
- بیش از ۳ میلیون تراکنش در ثانیه
- تعداد بالای کاربران
- لزوم ارائه سرویس در لحظه
- لزوم ارائه تحلیل‌های پیچیده

کاربردهای دیگر

- بانکداری
- دوربین‌های نظارتی
- سیستم‌های شبیه‌ساز
- تحلیل شبکه‌های اجتماعی
- خدمات حوزه سلامت

موارد دیگر؟

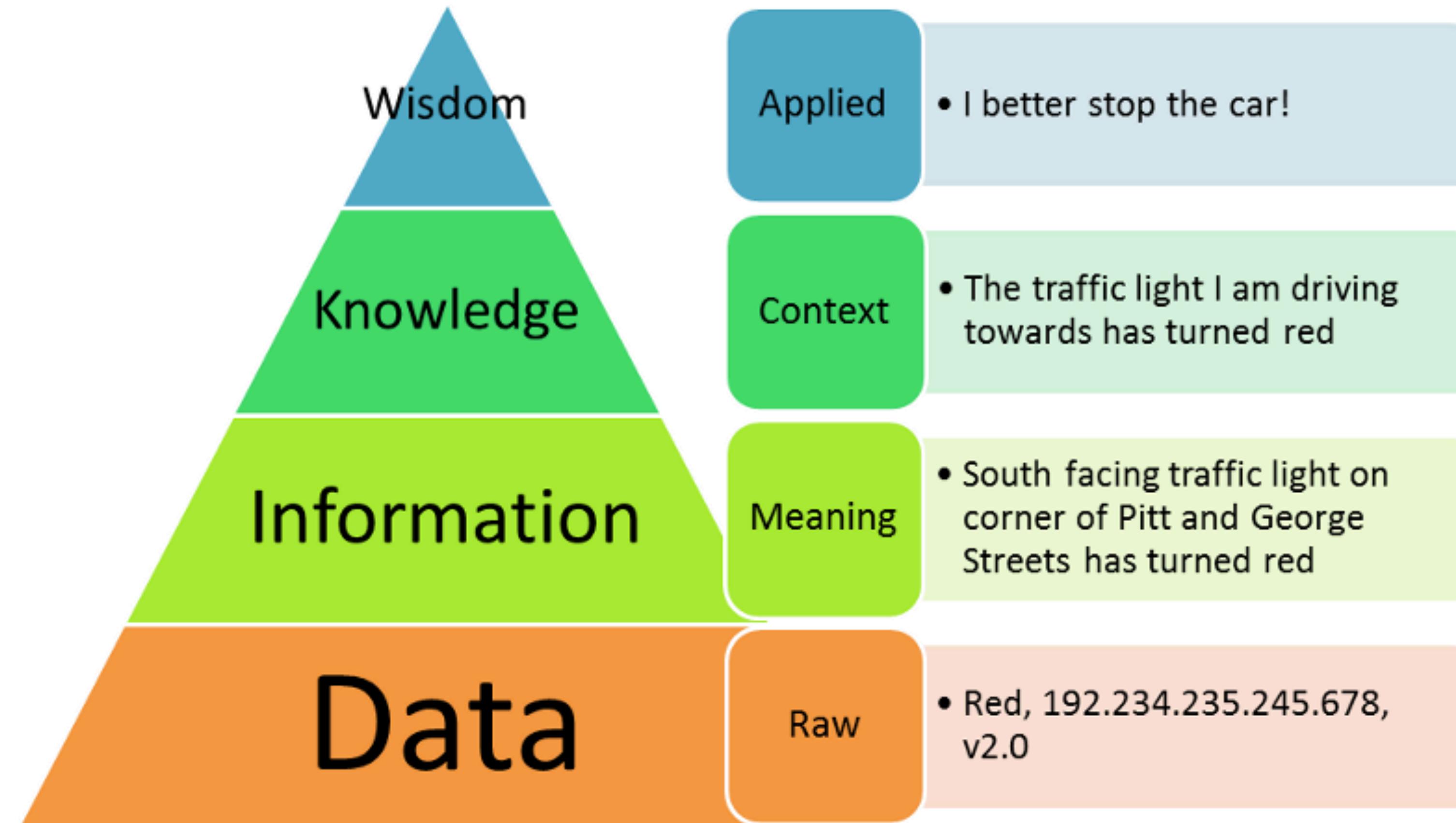
منابع داده



IOT
Internet of Things



- افراد
 - مثال: شبکه‌های اجتماعی
- ماشین‌ها
 - مثال: سنسورها
- سازمان‌ها
 - داده‌های ساختاریافته سازمانی



یکپارچگی

- ارزش اصلی زمانی اتفاق می‌افتد که ما یکپارچگی منابع مختلف داده ایجاد کرده و سپس به تحلیل آنها در سطحی بالاتر اقدام کنیم.
- یکپارچگی به معنی جمع‌آوری اطلاعات از منابع مختلف و تبدیل آنها به اطلاعاتی با ارزش بیشتر است.
- به این اطلاعات با ارزش بیشتر «دانش» می‌گوییم.

پکارچگی



- مثال: پیش‌بینی وقوع آتش‌سوزی و جنگل‌ها
- منابع مختلف اطلاعاتی:
 - اطلاعات جغرافیایی،
 - آب و هوا،
 - سنسورهای محیطی،
 - اطلاعات جدولی مرتبط،
 - تصاویر و فیلم‌های آتش‌سوزی
 - مدل‌سازی و شبیه‌سازی
- نتیجه: پیش‌بینی نحوه و شدت گسترش آتش‌سوزی

V's of Big Data

V's of Big Data

- تعریف مهاده‌ها:
- به مجموعه‌ای از داده‌های بزرگ و پیچیده گفته می‌شود که سیستم‌های مدیریت داده و تکنیک‌های معمول امکان مدیریت آنها را نداشته باشند
- پنج ویژگی اصلی در مهاده‌ها که آنها را V های مهاده‌ها می‌نامیم:
 - Volume
 - Velocity
 - Variety
 - Veracity
 - Valence

V's of Big Data

Volume .۱

- به حجم بالای اطلاعاتی که در حال تولید است اشاره می‌کند.

Velocity .۲

- به سرعت جابجایی و تولید داده اشاره می‌کند

Variety .۳

- به تنوع فرمتهای داده اشاره می‌کند.

Veracity .۴

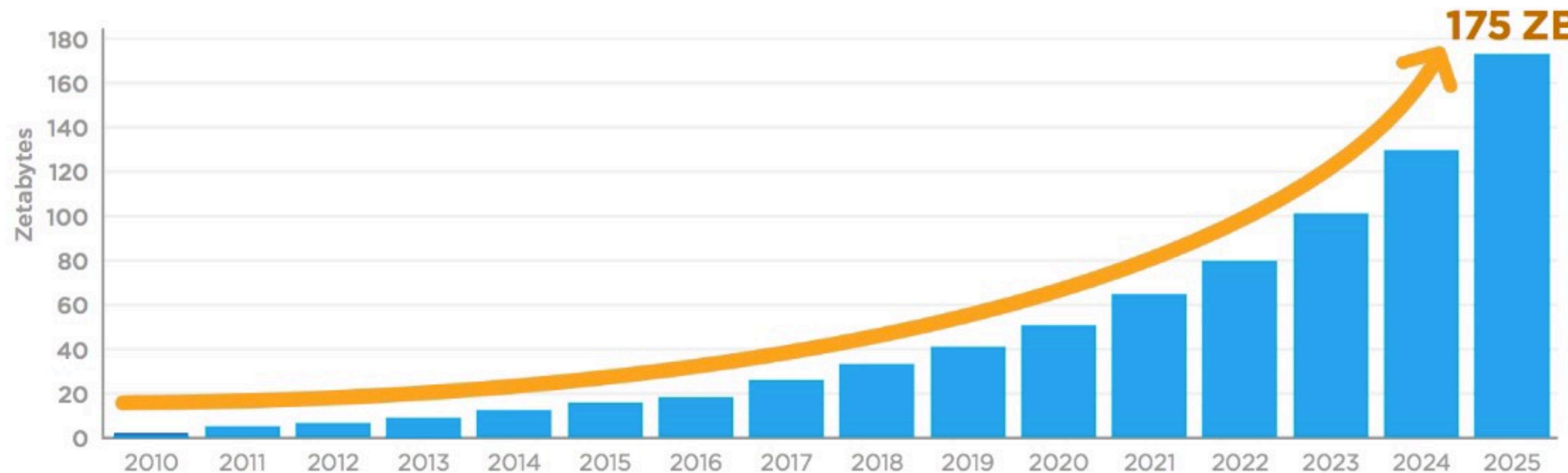
- به وجود بایاس، نویز و داده‌های غیرعادی اشاره دارد

Valence .۵

- به وجود ارتباط بین داده‌ها مشابه حالت گراف‌ها اشاره می‌کند

Volume

- حجم داده‌ها به شکل نمایی در حال رشد است.



Source: *IDC Data Age 2025*

Volume

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

Volume

3 Remarkable YouTube Statistics You Probably Didn't Know



1 YouTube impact on the Internet landscape is impressive



1.9
BILLION PEOPLE

YouTube users are almost a third of the internet.



91+
COUNTRIES

Number of countries with local versions of YouTube.



1
BILLION HOURS

Daily time spent watching YouTube in 2017.

Source: YouTube

2 YouTube is actually helping us to be smarter

Source: Google



86%

viewers who use YouTube to learn new things in 2017



46%

people who feel more prepared for new tasks after watching YouTube videos



41%

people who say they feel smarter watching videos on YouTube

3 YouTube lets people make massive amounts of money

Source: YouTube



\$2
BILLION

The amount paid to partners for YouTube content in the last 5 years.



50%
YEAR-ON-YEAR

The growth rate of channels earning 5 figures on YouTube.



75%
YEAR-ON-YEAR

The growth rate of channels with 1 million+ subscribers.

Volume

3 Key Facebook Statistics You Should Know



1 Facebook is an absolute social media leader



2.2 BILLION

Facebook active users, the highest among social media channels.



400

Facebook users who sign up every minute.



500,000

New Facebook users added daily.

Source: wearesocial, FB Newsroom, HubSpot, socialmediatoday.com

2 How people use Facebook?

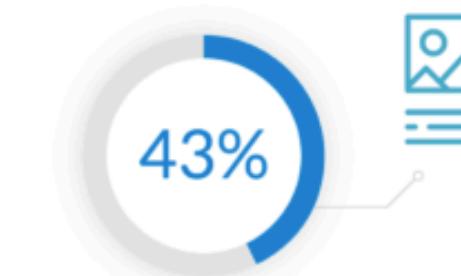
Source: HootSuite, s21.q4cdn.com, journalism.org



Facebook users who access the platform via mobile devices.



Facebook users who use the platform daily.



US users who get their news from Facebook.

3 Facebook is a huge player in the online advertisement market

Source: Statista, investor.fb.com



77%

The share of Facebook in social network ad revenues.



\$6,42

Facebook's average revenue per user.



\$16.9 MILLION

Facebook's average global revenue.

Volume

- چالش‌های حجم بالای داده:
- ذخیره‌سازی
- چالش‌هایی را در جمع‌آوری، نگهداری و استخراج داده به همراه دارد.
- مشکلاتی را در بخش‌های سخت‌افزار، شبکه، سرویس‌های ابری و غیره ایجاد می‌کند.
- پردازش داده‌ها
- بسیاری برنامه‌ها ممکن است برای پردازش داده نیاز به قراردادن آن در حافظه داشته باشند. در حجم بالا اینکار شدنی نیست.
- در واقع با افزایش حجم داده، کارایی سیستم‌ها کاهش و هزینه آنها افزایش می‌یابد.

Velocity

- منظور سرعت ارسال، ثبت و تحلیل داده است.
- در بسیاری کاربردها، پردازش در لحظه اطلاعات (realtime processing) اهمیت دارد.
- مثال:
 - ارائه تبلیغات مناسب و در لحظه به کاربران یوتیوب،
 - تصمیمات در لحظه برای معامله سهام،
 - تحلیل در لحظه اطلاعات حیاتی بیمار و اتخاذ تصمیمات درست،
 - تشخیص در لحظه رفتار ناهنجار در تراکنش‌های بانکی.
- در مقایسه با batch processing مفهوم realtime processing قرار می‌گیرد.
- سرعت پردازش وابسته است به سرعت تحلیل‌هایی که نیاز است.

Variety

- افزایش تنوع داده‌ها منجر به پیچیدگی سیستم‌ها می‌شود.
- تنوع می‌تواند در ابعاد مختلف باشد.
 ١. تنوع در ساختار و فرمت داده
 - مثال: داده‌های جغرافیایی با داده‌های یک تؤییت متفاوت است.
 ٢. تنوع در رسانه انتقال داده
 - مثال: یک سخنرانی می‌تواند به صورت صدا و یا متن و یا حتی ویدئوی همراه با صدا و متن منتقل شود

Variety

۳. تنوع معنایی: چگونه داده‌ها را تفسیر کرده و روی آنها اقدام کنیم.

- مثال: سن را می‌توان به صورت یک عدد و یا گروه سنی تعریف کرد.
- و یا پیشفرض‌های پشت سر بعضی داده‌ها، امکان مقایسه آنها را سلب می‌کند. برای مثال نمی‌توان بدون دانستن بعضی پیشفرض‌ها، بین دو دسته داده از دو جمعیت مختلف افراد مقایسه انجام داد.

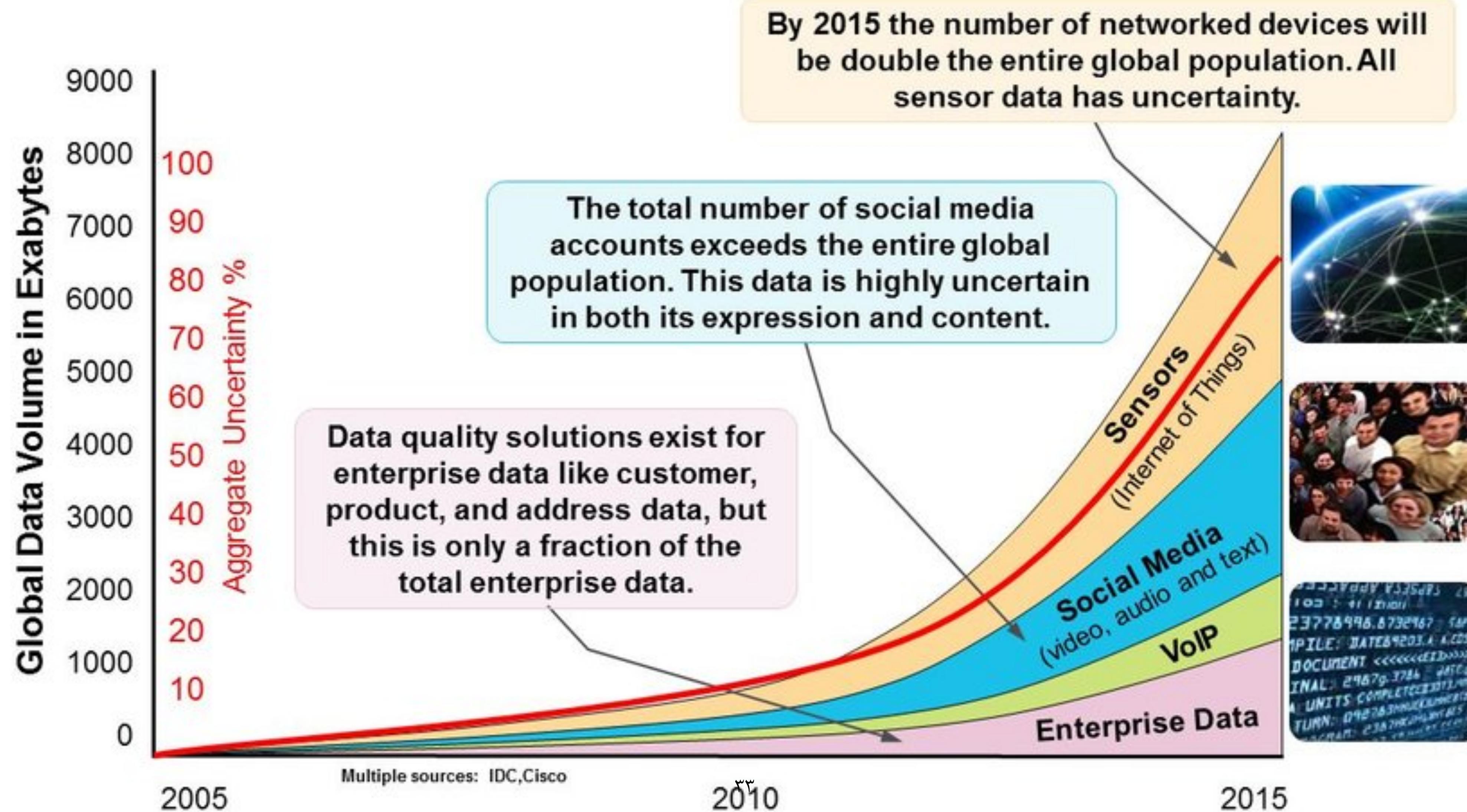
۴. تنوع دسترسی: داده می‌تواند به صورت آنلاین تولید شود و یا قبلاً ذخیره شده باشد.

- همچنین می‌تواند به صورت مستمر باشد و یا در بازه‌های زمانی متفاوت. مثلاً مقایسه تصاویر دوربین‌های شهری و تصاویر ماهواره‌ای که هر چند وقت یک بار عبور می‌کند.

Veracity

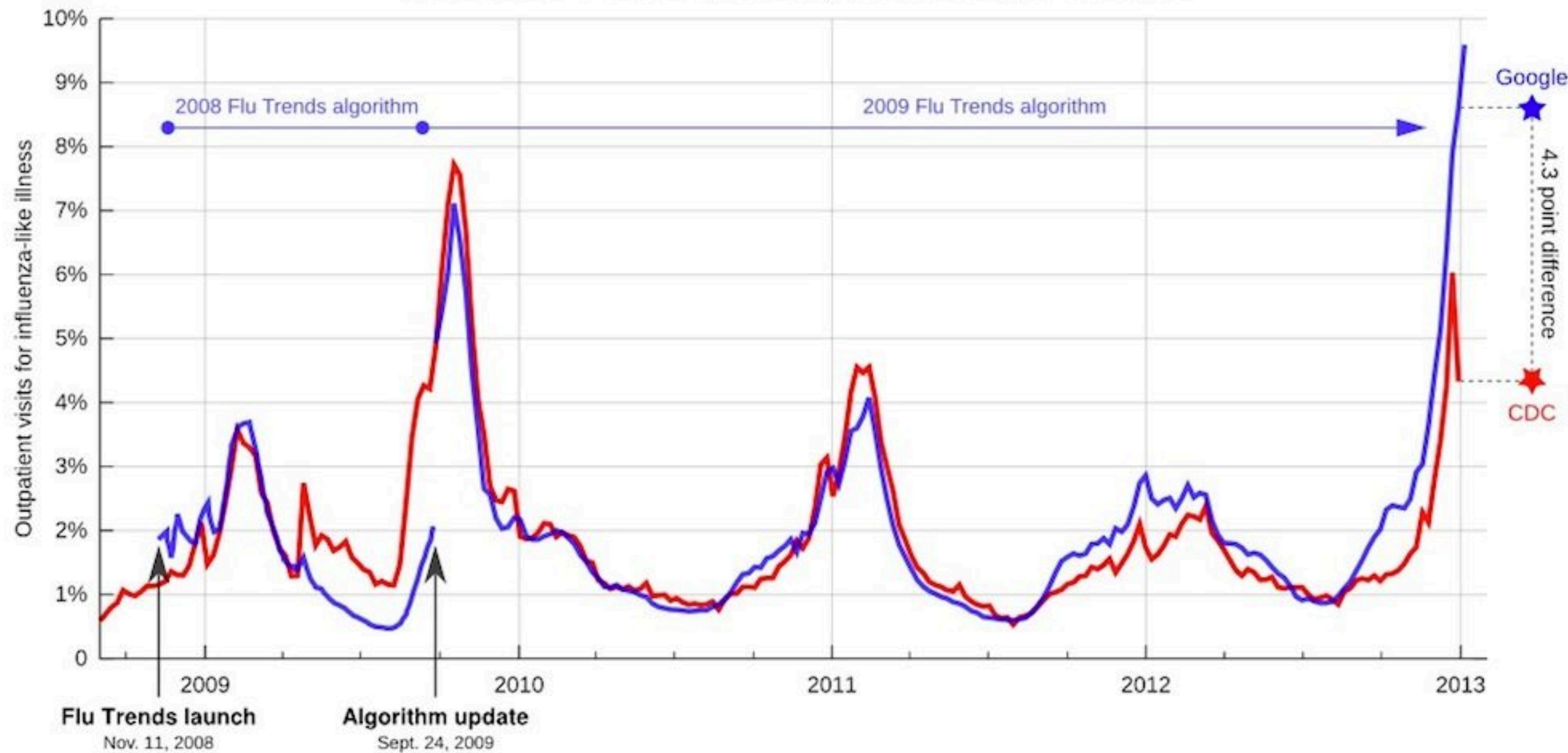
- این ویژگی بیشتر درباره کیفیت داده است. عنوان دیگر: Validity و Volatility
- به طور مشخص این ویژگی درباره وجود بایاس و داده‌های غیرنرمال صحبت می‌کند.
- فاکتورهای مختلفی مانند دقیق داده، قابل اطمینان بودن منبع داده و غیره می‌تواند نقش داشته باشد.

Veracity



Veracity

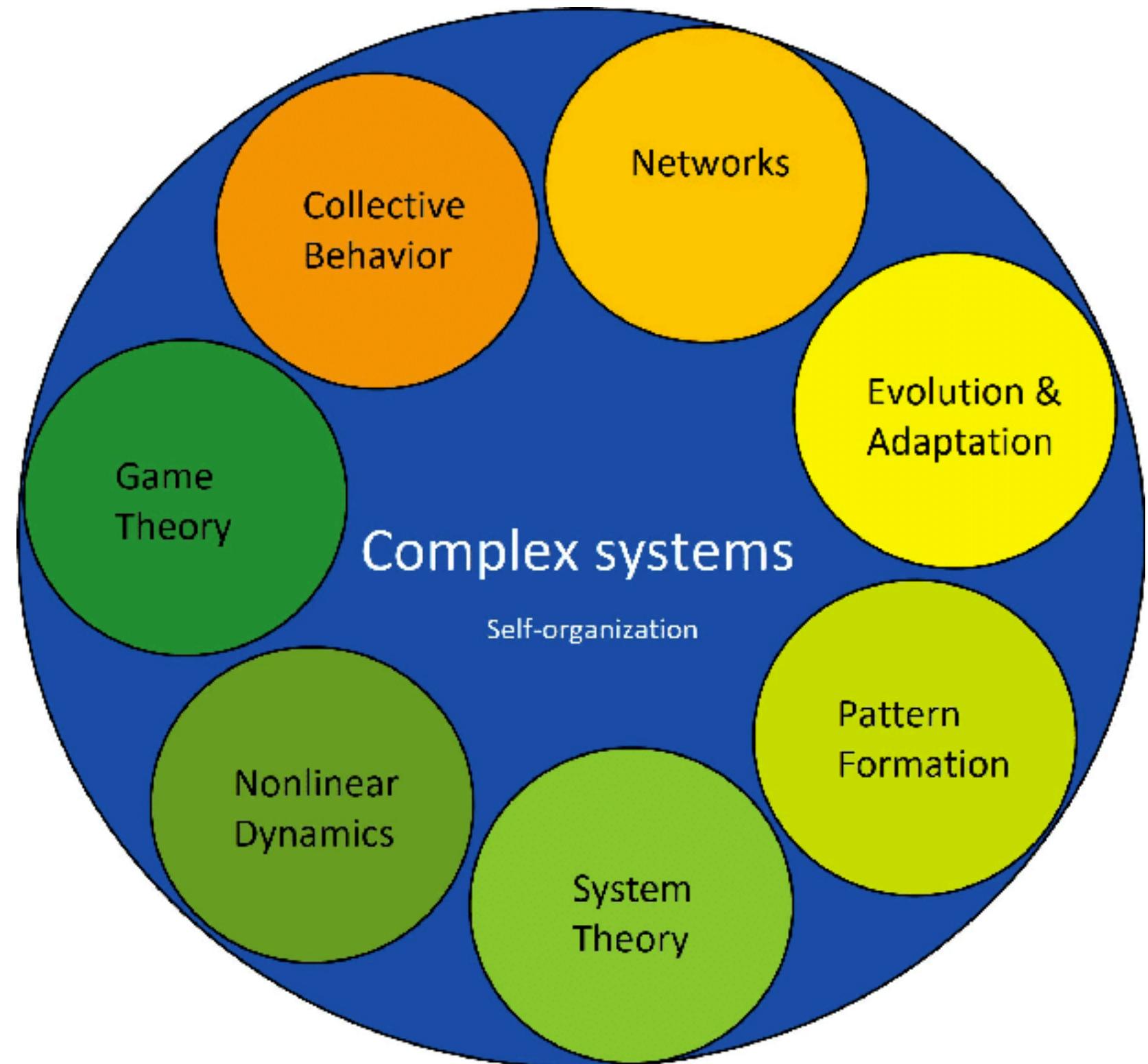
Google Flu Trends U.S. may have diverged again from the CDC data it predicts, but too early to be sure.



Sources: http://www.google.org/flu_trends/us, CDC ILinet data from <http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE 6(8): e23610. doi:10.1371/journal.pone.0023610,

Data as of Jan. 12, 2013. Keith Winstein (keithw@mit.edu)

Valence



- این ویژگی عموما درباره میزان ارتباط داده‌ها است.
- مثال: ارتباط معنایی/ارتباط شبکه‌های اجتماعی.
- افزایش ارتباط بین اجزا می‌تواند منجر به بروز رفتارهای جمعی پیچیده شود.
- این موضوع به خصوص در حوزه سیستم‌های پیچیده اهمیت زیادی دارد.

V's of Big Data

- هر کدام از پنج V اشاره شده می‌تواند منجر به بروز چالشی در سیستم‌ها شود.
 - اندازه
 - سرعت
 - پیچیدگی
 - کیفیت
 - ارتباط
- مدیریت همه این موارد می‌بایست در جهت ایجاد ارزش (Value) در سازمان باشند.

V's of Big Data

- یک مثال: پیش‌بینی قیمت سهام در بازار بورس
- منابع مختلف داده:
 - داده‌های ماشینی: رفتار کاربران در خرید و فروش‌ها
 - داده‌های انسانی: صحبت‌های افراد در شبکه‌های اجتماعی
 - داده‌های سازمانی: اطلاعات مالی از شرکت‌ها
- چالش‌ها:
 - اندازه بزرگ داده‌ها
 - سرعت بالای تولید داده و نیاز به تحلیل سریع آنها
 - منابع مختلف داده و تنوع آنها
 - عدم دقیقیت بعضی داده‌ها
 - ارتباط بین سهامداران حقیقی و حقوقی و غیره

راهبرد در مهاددها

- راهبرد: یک برنامه اقدام یا سیاست است که برای رسیدن به یک هدف تعریف می‌شود.
- قدمهای اصلی:
 - قدم اول: تعریف اهداف اصلی
 - اهداف بلند مدت
 - اهداف کوتاه مدت
 - لازم است ابتدا سوالات به شکل درستی تعریف شوند.
 - مثلا: چطور داده‌های فروش و بازاریابی می‌تواند منجر به بهبود فروش شود.
 - چطور داده‌های سنسورها می‌تواند منجر به تشخیص خرابی دستگاه‌های خط تولید شود.
 - همچنین لازم است ریسک‌ها و مزایا، منابع در دسترس، نیازمندی‌های پروژه و قوانین حاکم بر سازمان به دقت بررسی شوند.
 - هرچه اهداف دقیق‌تر و به صورت عددی تعریف شود، امکان ارزیابی آن در طول اجرا بهتر خواهد بود.

راهبرد در مهاددها

- قدم دوم: جلب حمایت مدیران ارشد سازمان
- بدون حمایت مدیران ارشد، امکان جلب همراهی همه اجزای سازمان و رسیدن به اهداف مدنظر وجود ندارد.
- قدم سوم: تشکیل تیم اصلی
 - آموزش و تمرین ابزارها و ایجاد تخصص لازم در تیم
- قدم چهارم: شروع با یک پروژه کوچک
 - با گرفتن نتایج مطلوب می‌توان تیم را بزرگتر کرد.
- قدم پنجم: ایجاد امکان دسترسی مناسب به داده.
- لزوم توجه به اشتراک‌گذاری حداکثری داده‌ها
- قدم ششم: توجه به جنبه‌های مختلف دیگر
- مانند سیاست حریم خصوصی، مدیریت دسترسی، سیاست مدیریت عمر داده، سیاست تمیز نگه داشتن داده،