Big Data Course

Chap 1- Introduction

Electrical Engineering department of

Amirkabir University of technology

Dr. Mohammadreza Pourfard

August 2025

صفحات درس و استاد

اتاق: طبقه سوم ساختمان ابوریحان، دانشکده مهندسی برق، آزمایشگاه سیستمهای هوشمند دیجیتال

@Electronic_daneshbonyan

pourfardm@gmail.com

+982164543373

http://ssail.aut.ac.ir

ssail.aut.ac.ir

DIGITAL
SMART SYSTEM AND DATA MINING LAB

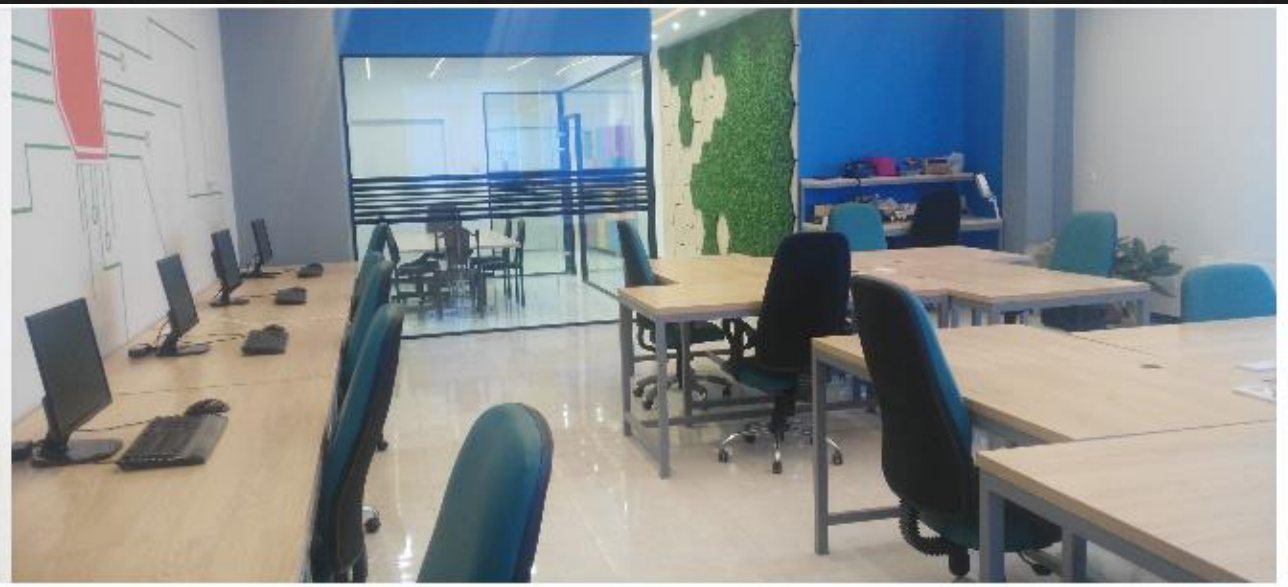
English | Search

Research Groups | Research Projects | People | News | About Us | Admission

Research Groups

In today's rapidly evolving technological landscape, numerous research groups are at the forefront of innovation, each focusing on transformative areas that have the potential to reshape industries and enhance our daily lives. Among these, the Internet of Things (IoT) research group is dedicated to exploring the interconnectedness of devices and sensors, striving to create smart environments that foster efficiency and convenience. Meanwhile, our Artificial Intelligence (AI) team is delving into advanced algorithms and machine learning techniques to develop systems that can learn, adapt, and make informed decisions, pushing the boundaries of automation and intelligence. Additionally, the Big Data research group is

Laboratory page



Profile Contents

Home | Courses

- About
- Theses
- Research Interests
- Research Groups
- Research Projects
- Current Students
- Employment Records
- Contact
- Publications
- News

About

Adjunct Professor of Amirkabir University of Technology

Theses

B.S Degree:
Simulating Multi-Level Cache Memory

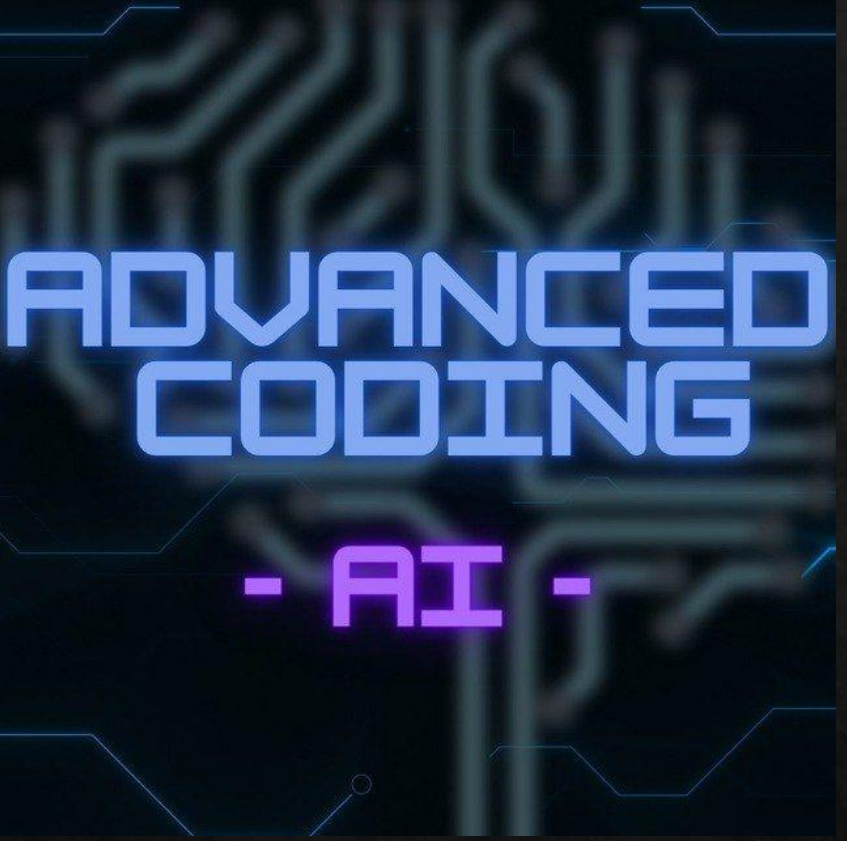
M.S Degree:
Detection and Tracking of a Human in different positions of its body

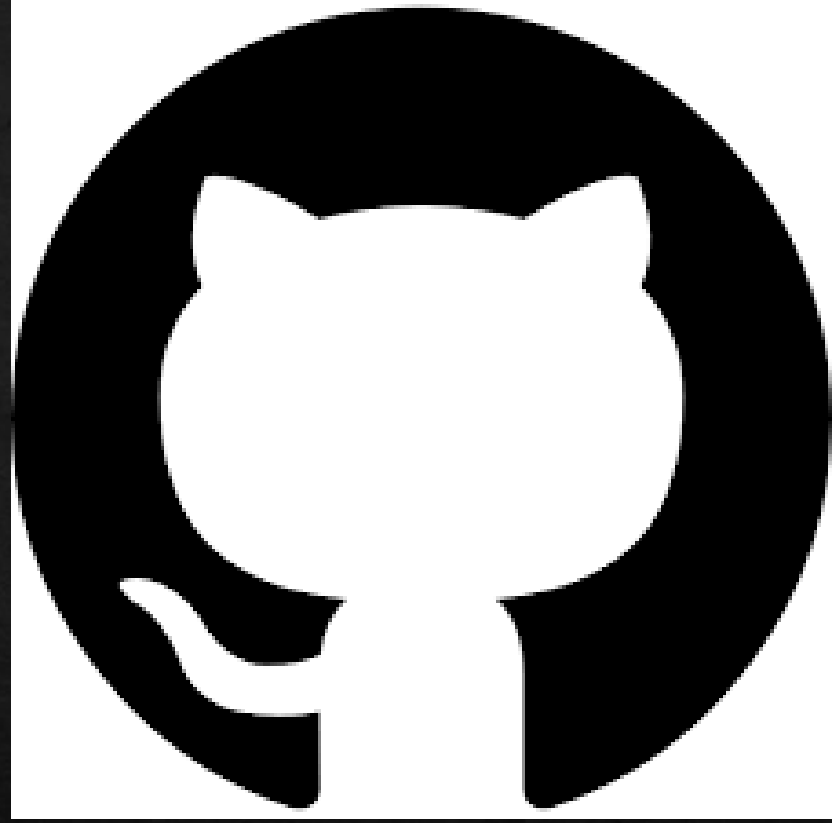
PhD Degree:
Texture Analysis and Separation for Characterization of Material's Structure through

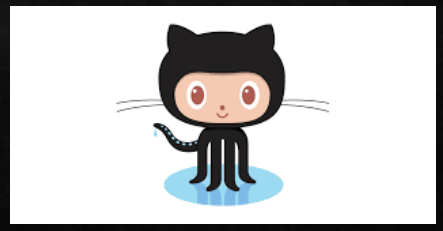
Research Interests

- Genetics data processing,

کانالها و گروههای مرتبط با درس



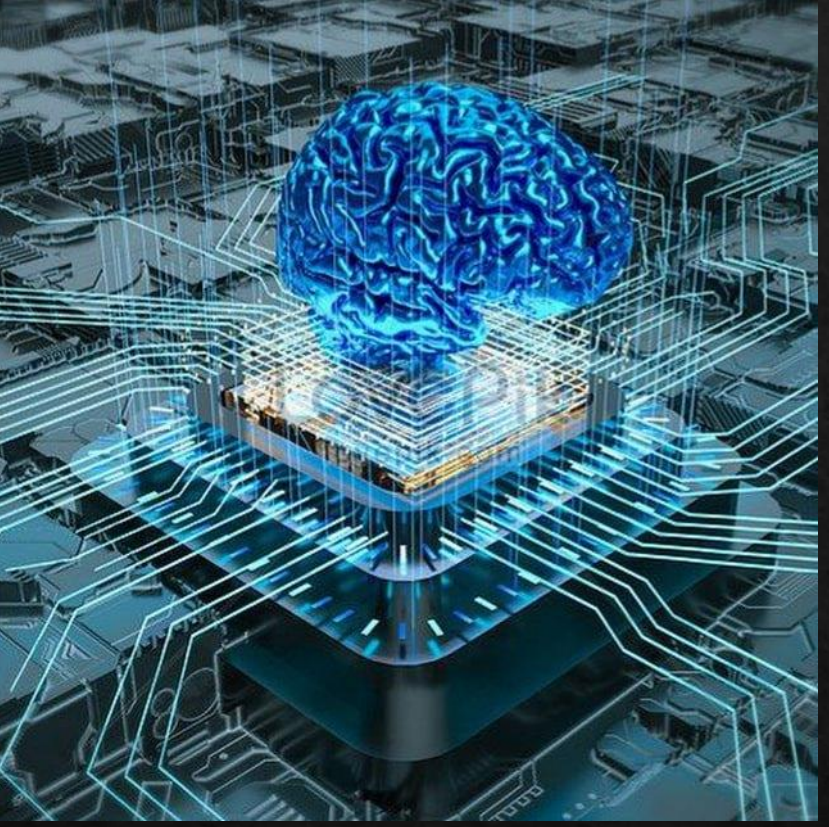


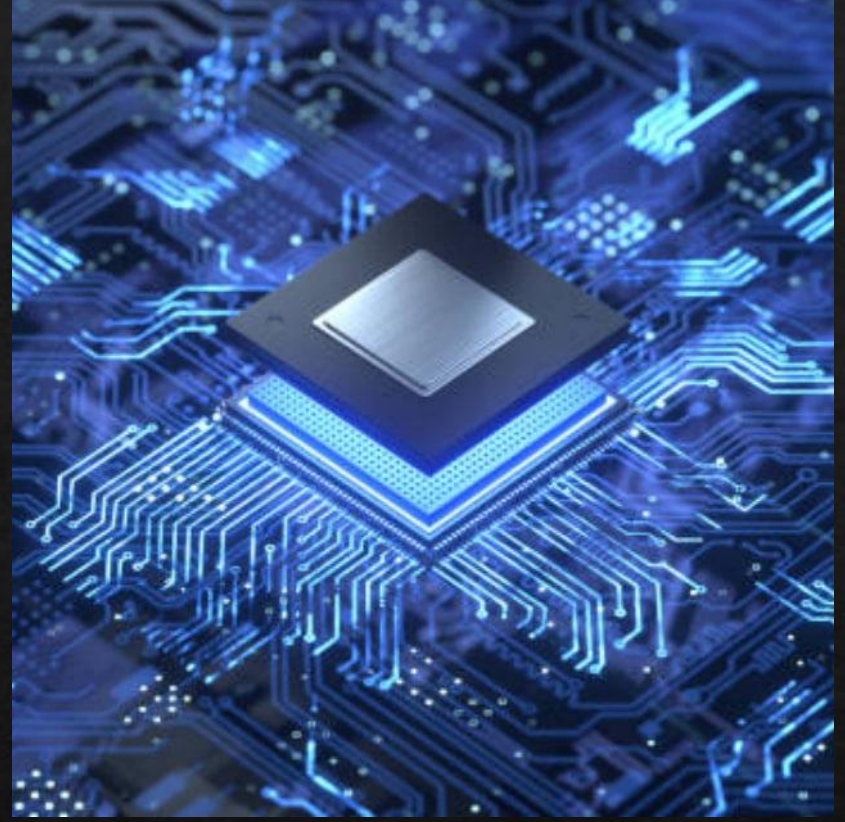


Telegram Channel:
https://t.me/Advanced_programming_algorithm_c

Telegram Group:
https://t.me/Advanced_Programming_Algorithm

GitHub Page:
https://github.com/Pourfardm/

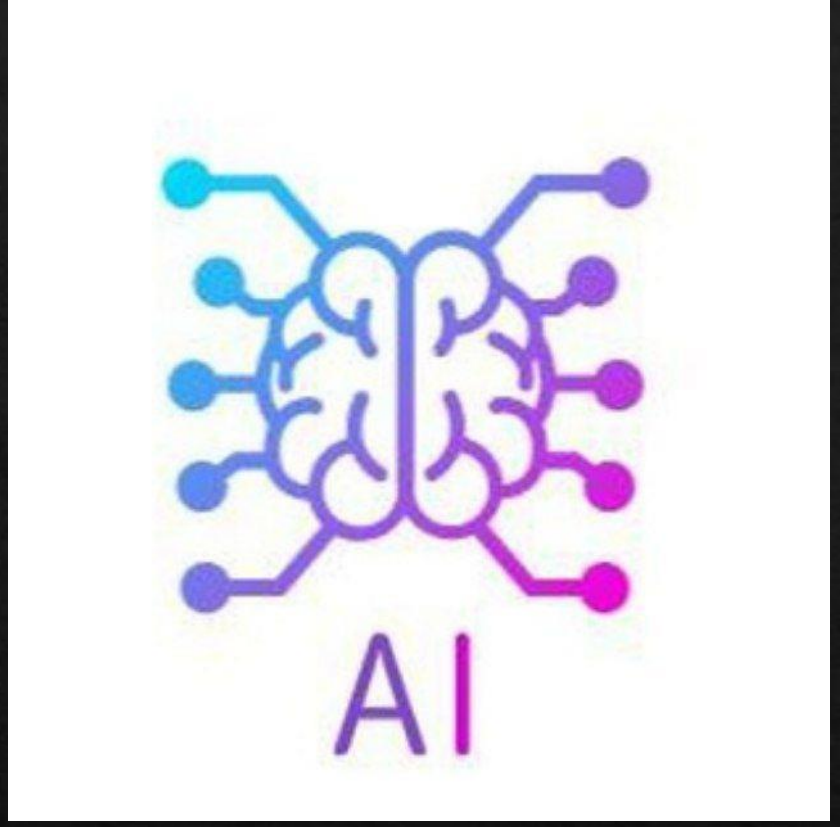




Telegram Channel:
https://t.me/fpga_digital_logic_design

Telegram Group:
https://t.me/Advanced_digital_analog_design

YouTube Channel:
https://www.youtube.com/channel/UCa8ZFUxp37Vy-KEbx7SbbHQ



Telegram group:
https://t.me/AI_Bigdata_amirkabir_group

Telegram Channel:
https://t.me/Deeplearning_BigData_AI

Private Telegram Group:
For class student

Telegram Channel:
https://t.me/AI_Nuroscience_Genomic_Data

YouTube Channel:
https://www.youtube.com/channel/UCa8ZFUxp37Vy-KEbx7SbbHQ

اهداف درس

اهداف کلیدی درس

- آشنایی با کاربردهای مفید پردازش کلان داده به منظور راه اندازی کسب و کار
- آشنایی با ابزارهای مورد استفاده در شرکتهای دانش بنیان برای پردازش کلان داده
- آشنایی با الگوریتم های جدید و پیشرفته پردازش کلان داده
- برقراری ارتباط بین مهارت آشنایی با سخت افزار همانند GPU

مروری بر نسل دانشگاهها در جهان

اهمیت ورود به دانشگاه نسل سوم و

ارایه درس متناسب با دانشگاه نسل سوم

مقایسه نسل های مختلف دانشگاهها در جهان

نسل دانشگاه	دوره تاریخی	مأموریت اصلی	ویژگی‌های کلیدی	نمونه‌ها	نقش اجتماعی
نسل اول – آموزش‌محور	قرون وسطی تا قرن ۱۹	**آموزش و انتقال دانش**	- تمرکز بر علوم کلاسیک (فقه، فلسفه، پزشکی) - نبود پژوهش سیستماتیک - متون سنتی و اقتدار دینی	بولونیا، پاریس، آکسفورد	تربیت نخبگان و مشروعیت‌بخشی به دین و حکومت
نسل دوم – پژوهش‌محور	قرن ۱۹ تا اواسط قرن ۲۰	**آموزش + پژوهش**	- پژوهش بنیادی - آزادی آکادمیک - شکل‌گیری رشته‌های علمی مدرن	برلین (هومبولت)، کمبریج، MIT (مدرن اولیه)	تولید علم و فناوری، پیشرفت علمی جوامع
نسل سوم – کارآفرین	اواخر قرن ۲۰	**آموزش + پژوهش + نوآوری/ کارآفرینی**	- تجاری‌سازی دانش - مراکز رشد و پارک فناوری - ارتباط نزدیک با صنعت	استنفورد، MIT، کمبریج (Silicon Fen)	موتور اقتصاد دانش‌بنیان
نسل چهارم – اجتماعی/ مسئولیت‌پذیر	اوایل قرن ۲۱	**آموزش + پژوهش + کارآفرینی + مسئولیت اجتماعی**	- توسعه پایدار - عدالت اجتماعی - توجه به محیط زیست و حکمرانی داده	دانشگاه‌های اسکاندیناوی، تورنتو، برخی دانشگاه‌های آسیایی	حل مسائل کلان بشری (اقلیم، نابرابری، سلامت)
نسل پنجم – هوشمند/شبکه‌ای	دهه‌های اخیر (در حال شکل‌گیری)	**یکپارچه‌سازی آموزش + پژوهش + نوآوری با فناوری دیجیتال**	- هوش مصنوعی، big data، آموزش هوشمند - دانشگاه شبکه‌ای جهانی - یادگیری شخصی‌سازی‌شده	پروژه‌های دانشگاه‌های مجازی جهانی، edX، دانشگاه‌های آسیای شرقی	هدایت جامعه به سوی اقتصاد دیجیتال و تمدن AI

نکته مهم این درس

به دلیل رویکرد دانشگاه نسل سوم

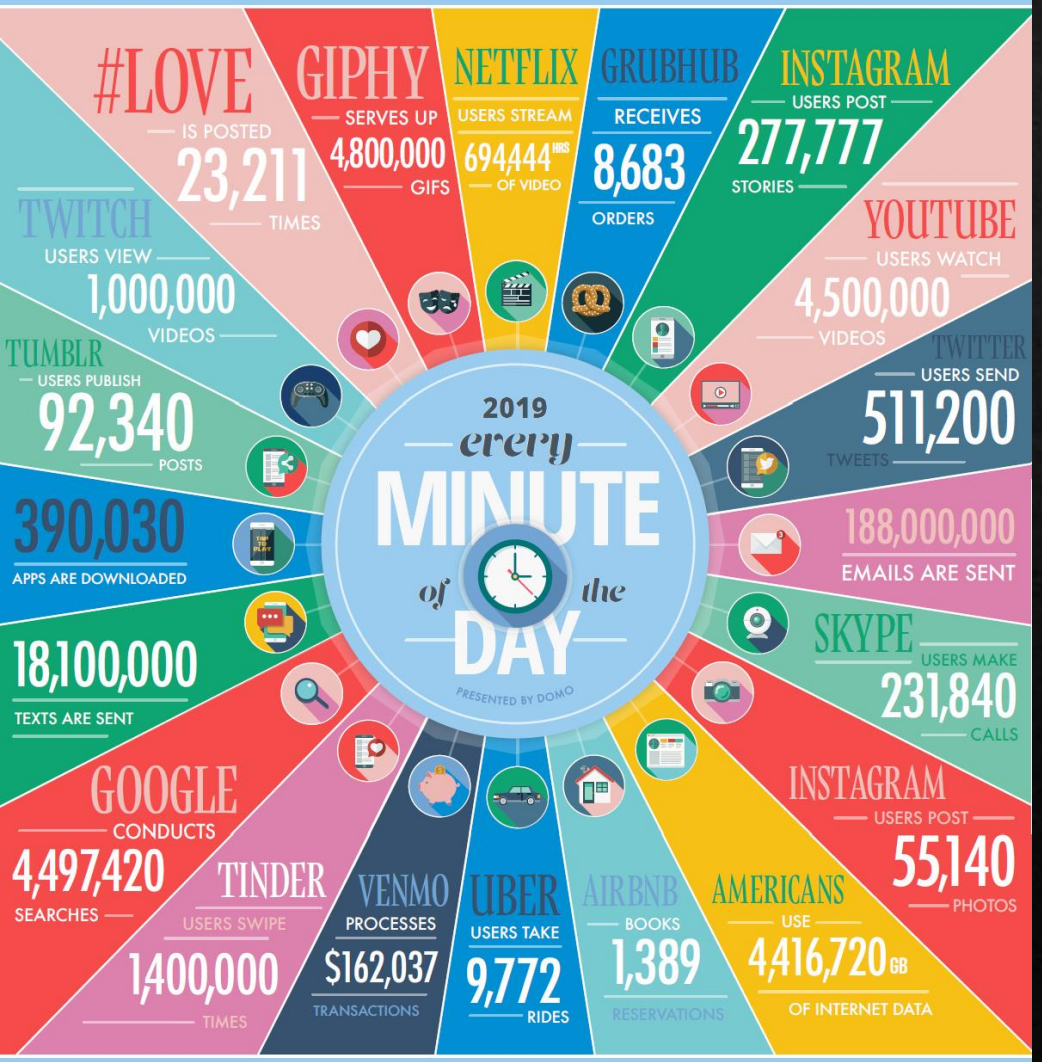
تمرکز این درس فقط روی الگوریتم نیست

حجم وحشتناک و خیره کننده تولید داده در جهان

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.

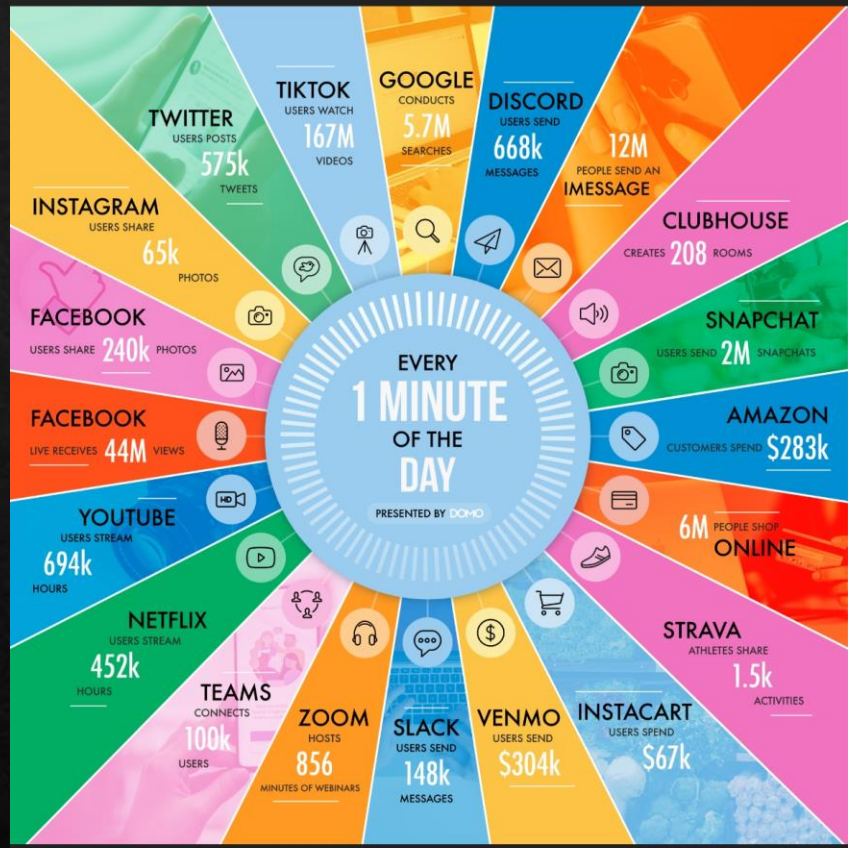


SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRED

Data Never Sleeps 9.0

How much data is generated *every minute*?

The 2020 pandemic upended everything, from how we engage with each other to how we engage with brands and the digital world. At the same time, it transformed how we eat, how we work and how we entertain ourselves. Data never sleeps and it shows no signs of slowing down. In our 9th edition of the "Data Never Sleeps" infographic, we bring you a glimpse of how much data is created every digital minute in our increasingly data-driven world.



As of July 2021, the internet reaches 65% of the world's population and now represents 5.17 billion people—a 10% increase from January 2021. Of this total, 92.6 percent accessed the internet via mobile devices. According to Statista, the total amount of data consumed globally in 2021 was 79 zettabytes, an annual number projected to grow to over 180 zettabytes by 2025.

Global Internet Population Growth

(IN BILLIONS)

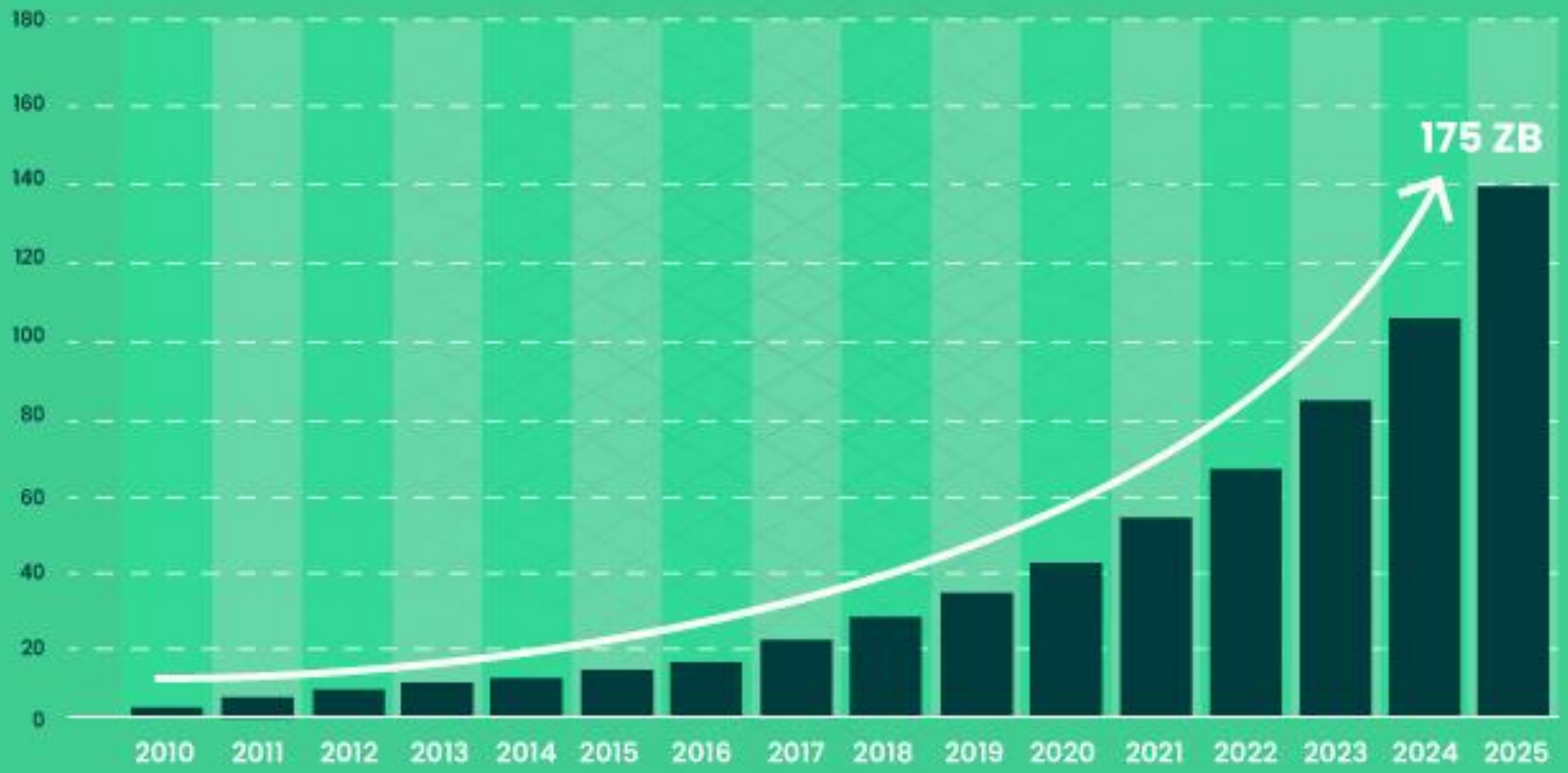
As the world changes, businesses need to change too—and that requires data. Domo gives you the power to make data-driven decisions at any moment, on any device, so that you can make smart choices in a rapidly changing world. Every click, swipe, share, or like tells you something about your customers and what they want, and Domo is here to help you and your business make sense of all of it.

Learn more at domo.com

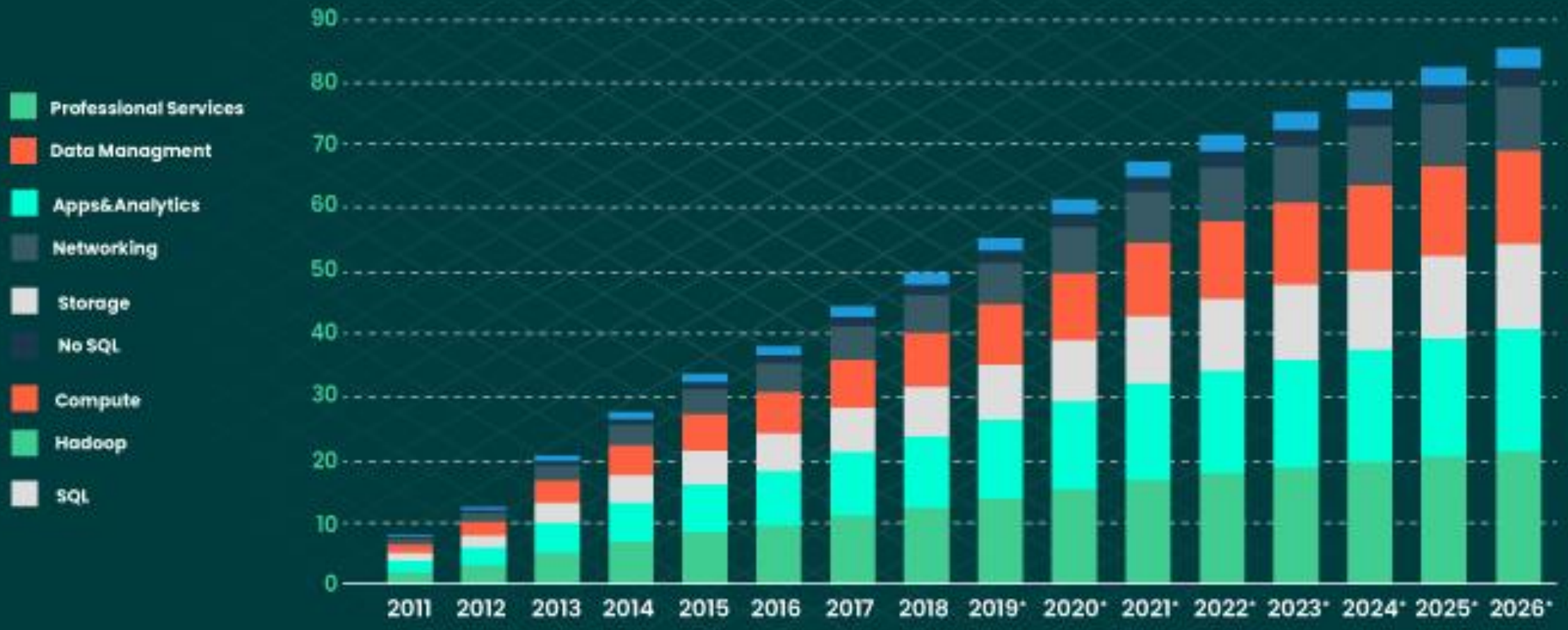
SOURCES: LOCAL IQ, BUSINESS OF APPS, DUSTIN STOUT, HOOTSUITE, EXPANDED RAMBLINGS, INTERNET WORLD STATS, STATISTA, CNBC, BRANDWATCH, KILL THE CABLE BILL, YOUTUBE, KINSTA, THE VERGE, MANAGEMENT COMMUNICATION: A CASE ANALYSIS APPROACH, INTERNET LIVE STATS, SODA, STATISTA

THE INTERNET IN 2023 EVERY MINUTE

Created by: eDiscovery Today & LTMG


180
160
140
120
100
80
60
40
20
0
175 ZB
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

Big Data Market Forecast Worldwide from 2011 to 2026, by segment (in billion U.S. dollars)



اہمیت BigData


Cost reduction
BIG DATA ANALYTICS
Faster, better decision making
New products & services

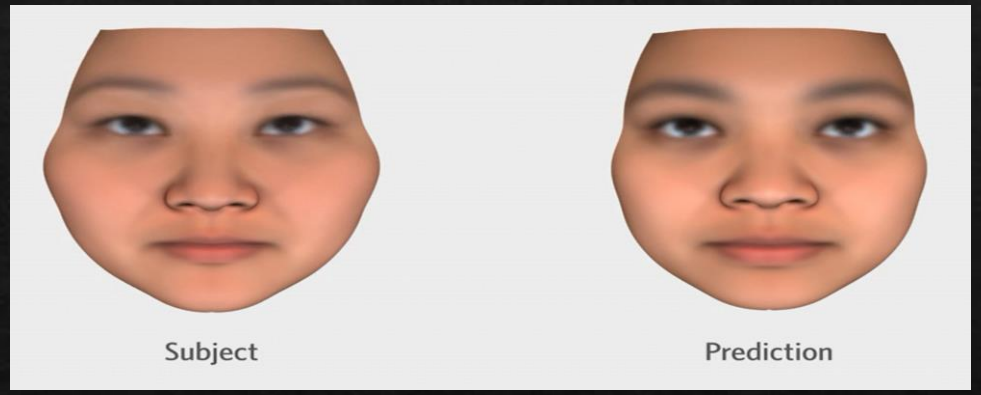
Importance of Big Data for Business

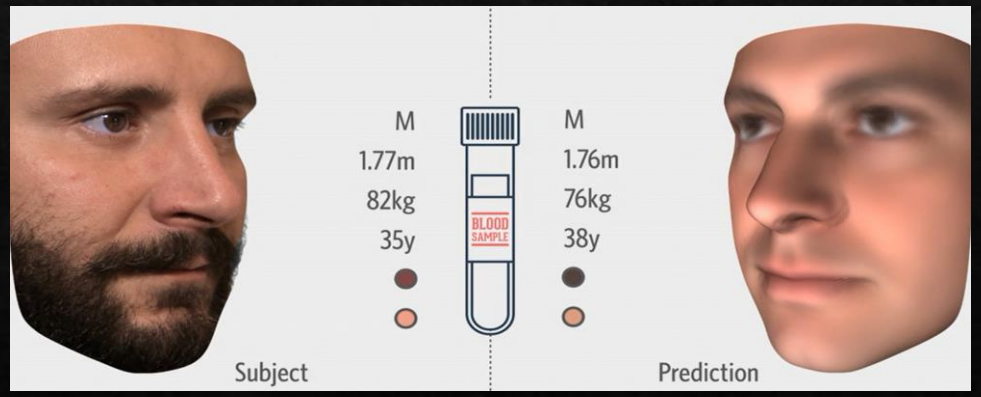
کاربردهای BigData

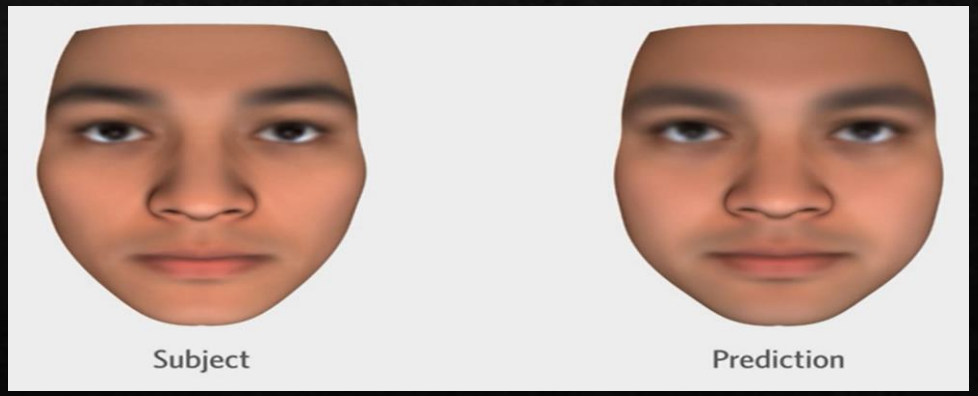
BIG DATA APPLICATIONS

Big Data Contributions to Healthcare

شرح مساله - پیش بینی چهره در سال۲۰۱۶



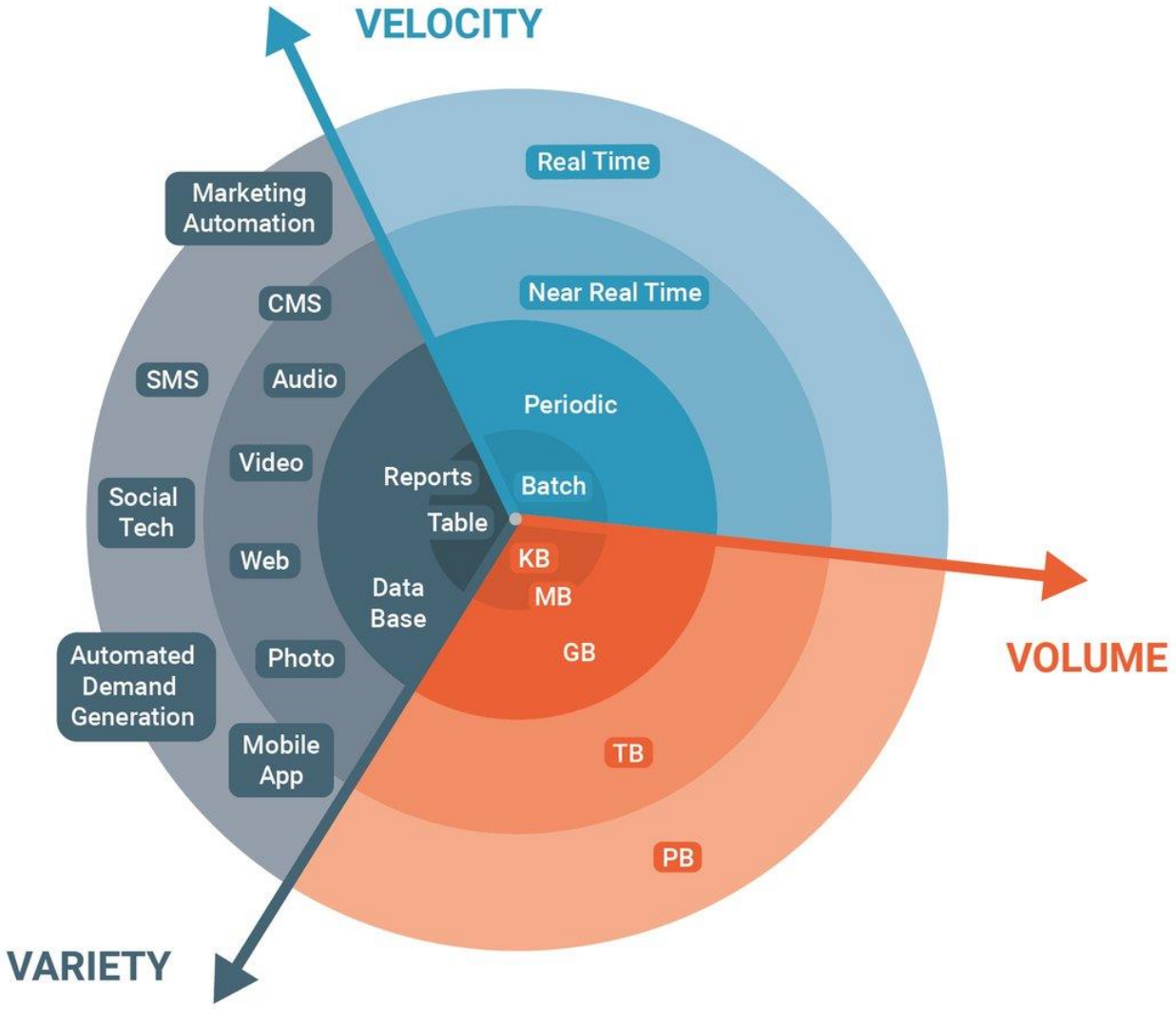


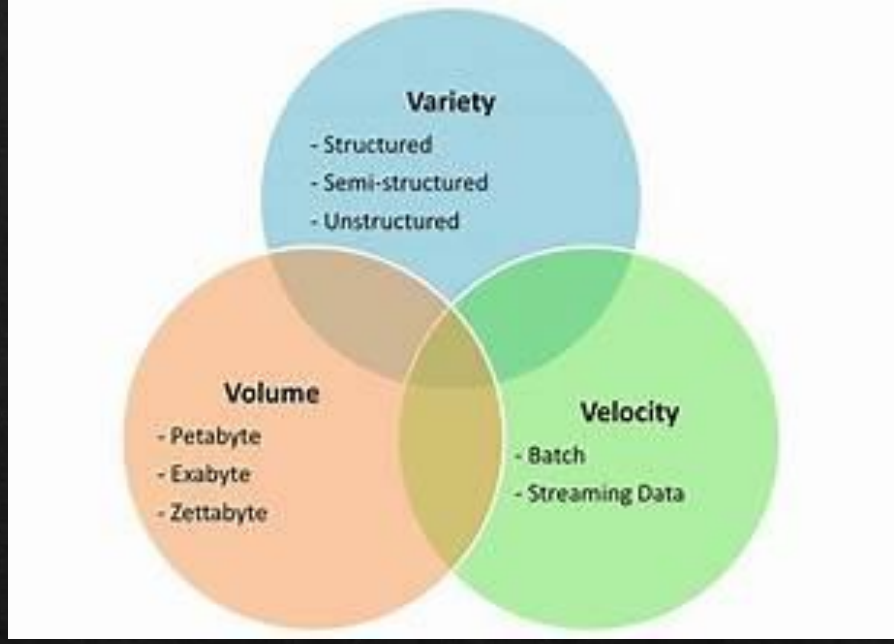


چالش های BigData

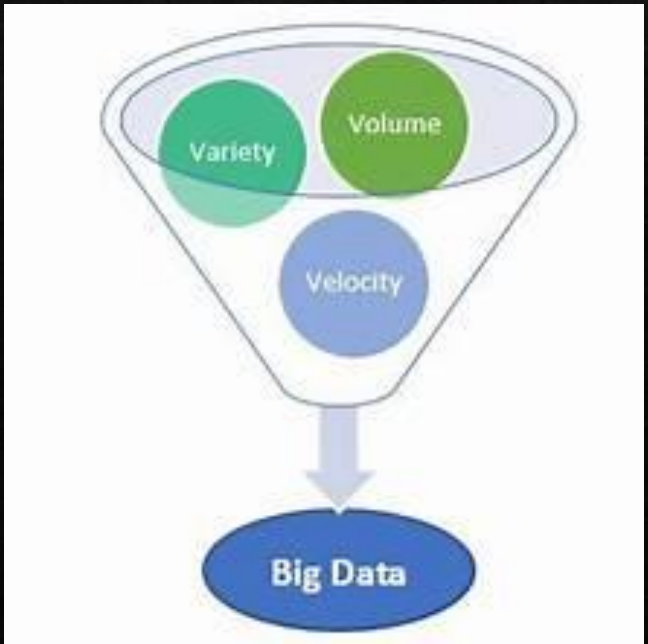
The 3 V's of Big Data

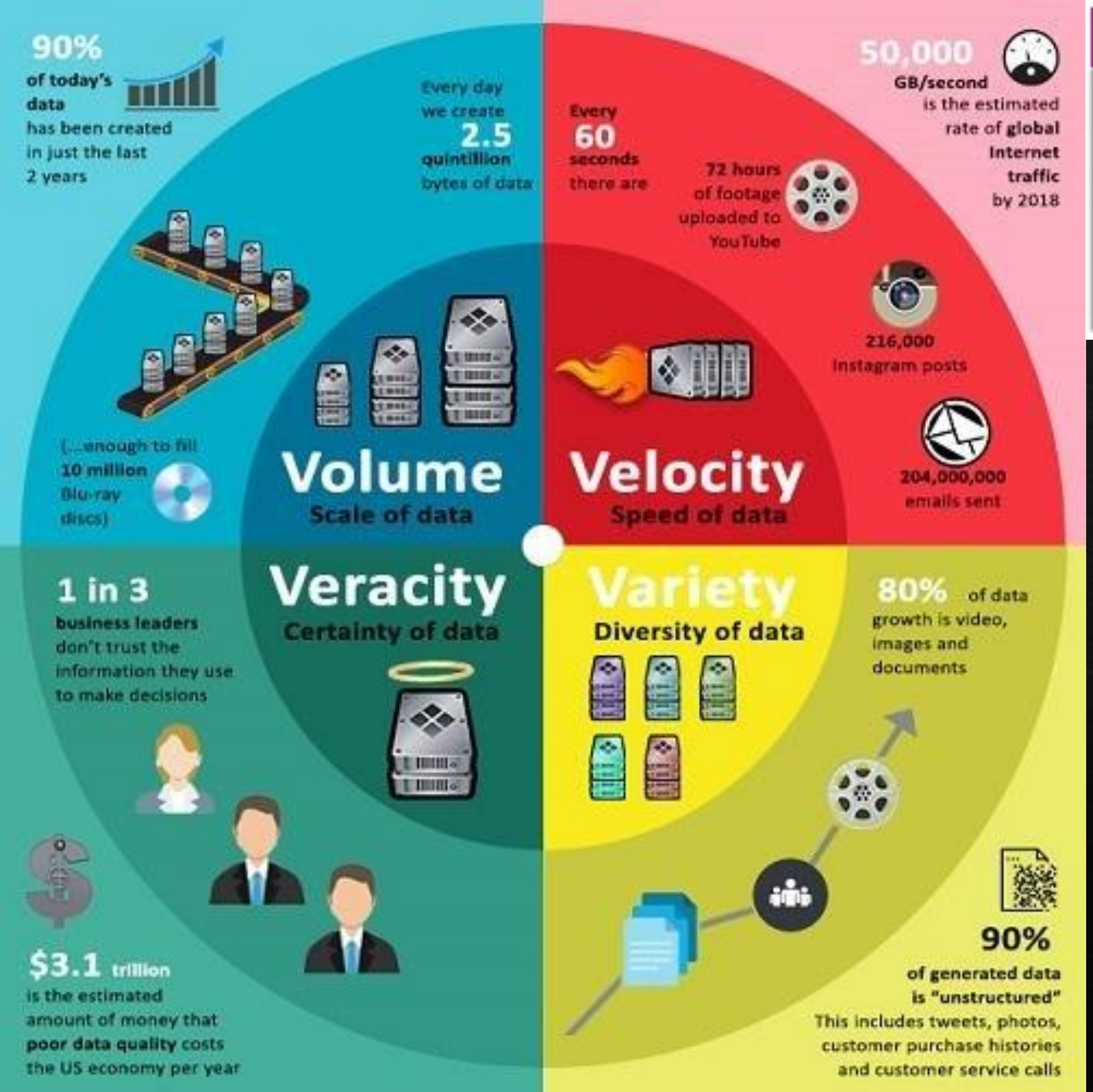
G2 CROWD

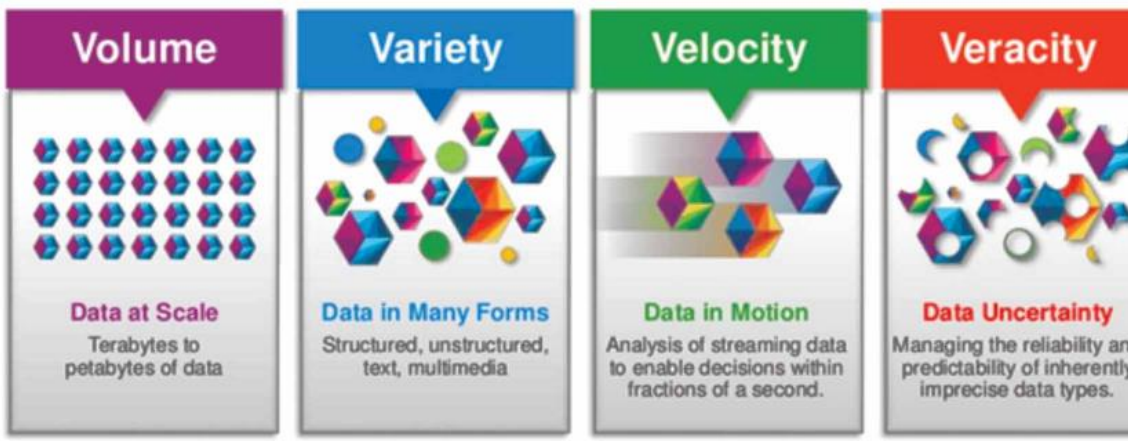




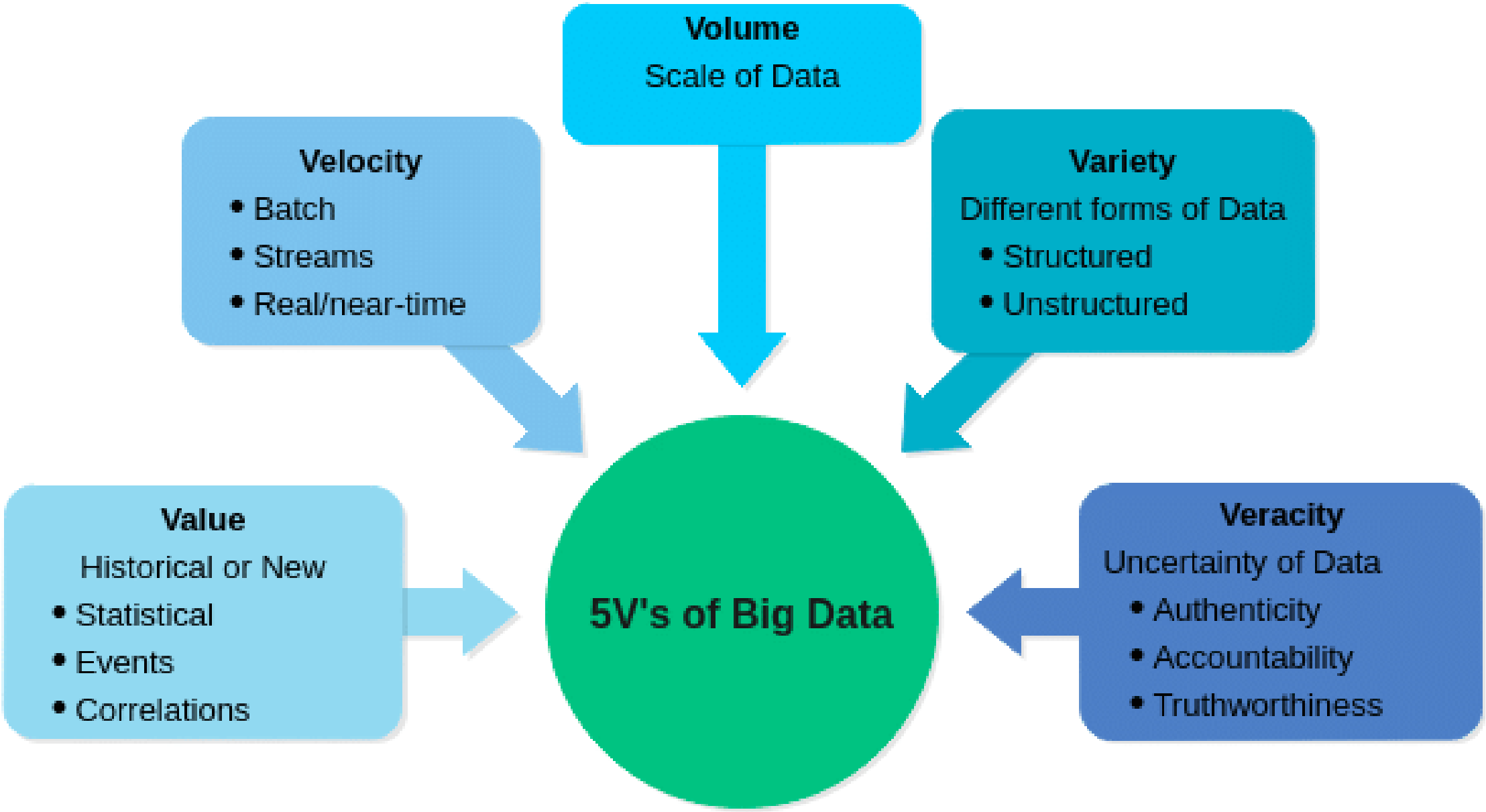
اصل 3V در Big Data





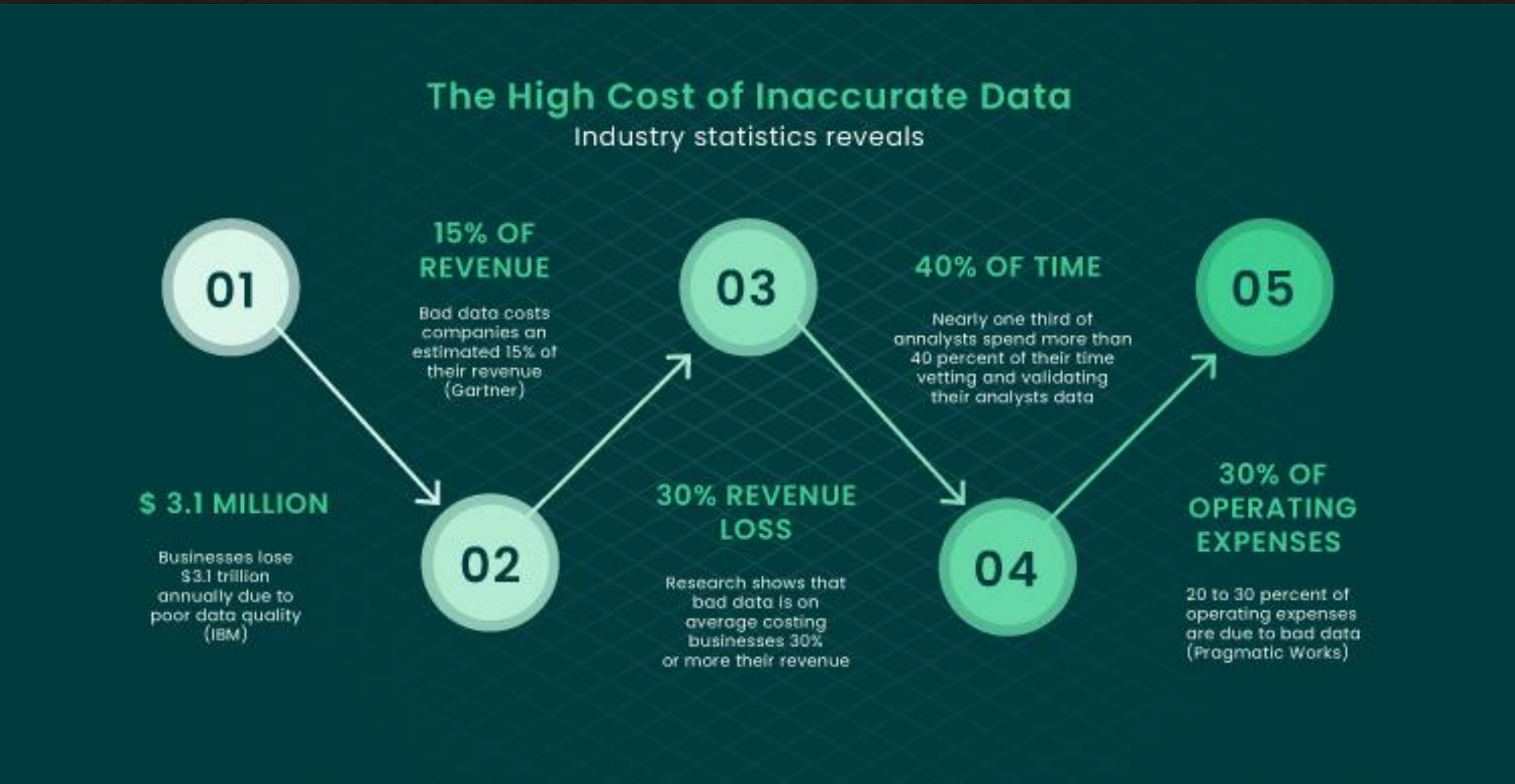


اصل 4V در Big Data


Volume
Scale of Data
Velocity
• Batch
• Streams
• Real/near-time
Variety
Different forms of Data
• Structured
• Unstructured
5V's of Big Data
Value
Historical or New
• Statistical
• Events
• Correlations
Veracity
Uncertainty of Data
• Authenticity
• Accountability
• Truthworthiness


BIG DATA
with 8 V's
01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

چالشهای کلان داده

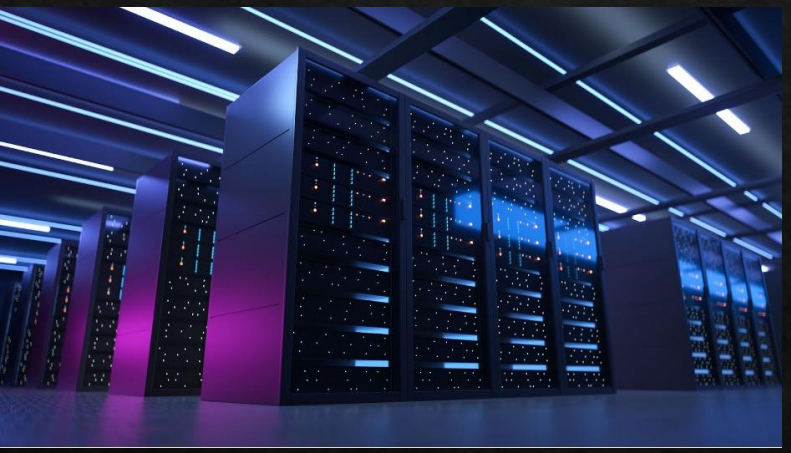


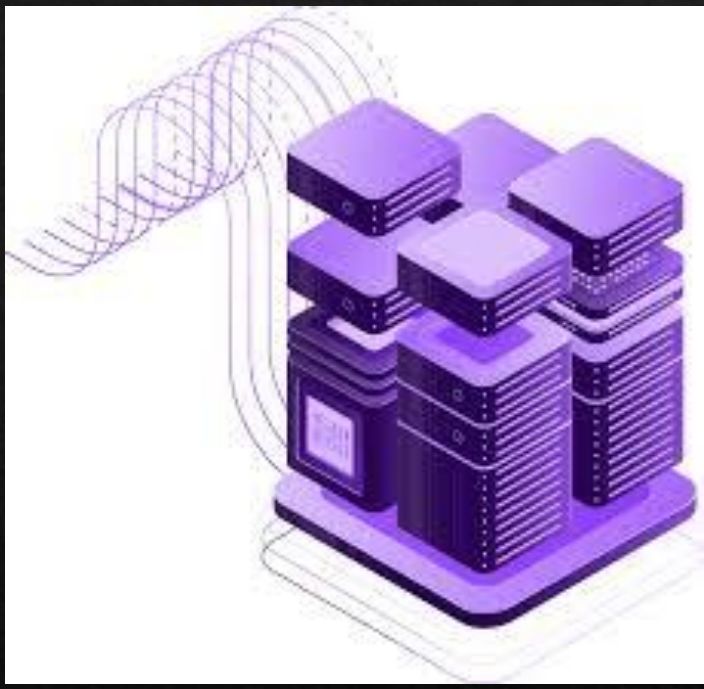
تحولات جهانی بعد از حضور کلان داده و هوش مصنوعی

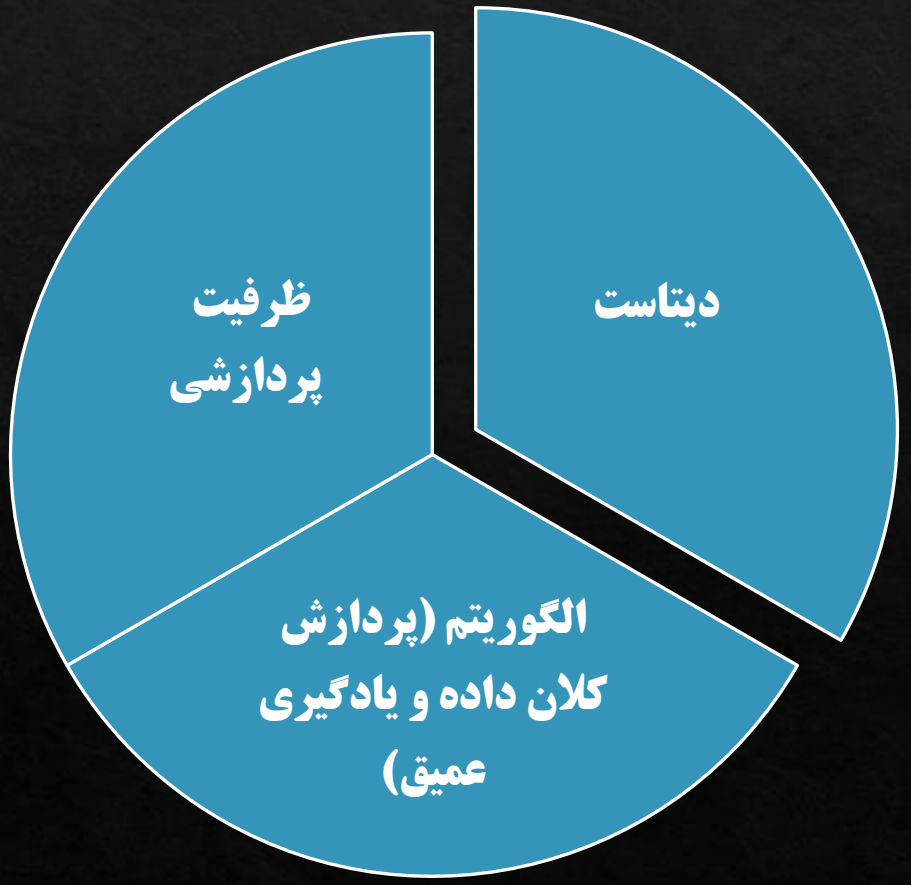
تحولات کلان داده در علوم مختلف در جهان

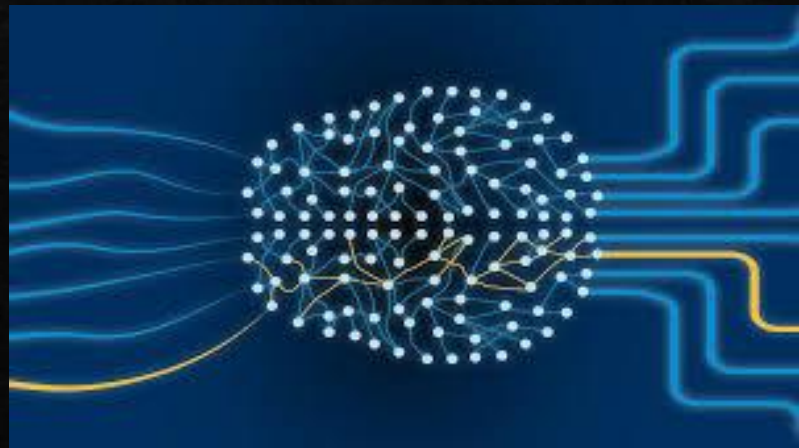
- حذف پزشکی سنتی و جایگزینی آن با پزشکی شخصی (Personalized medicine)
- حذف داروسازی سنتی و ایجاد داروی ویژه هر فرد با استفاده از پردازش داده ژنتیک
- از بین رفتن دموکراسی
- حذف رانندگی سنتی و ایجاد خودرو هوشمند و جاده هوشمند و نقش مهم رگولاتوری داده در آینده
- حذف وکالت در آینده
- نقش مهم مهندسان داده در آینده

سه راس مهم هوش مصنوعی



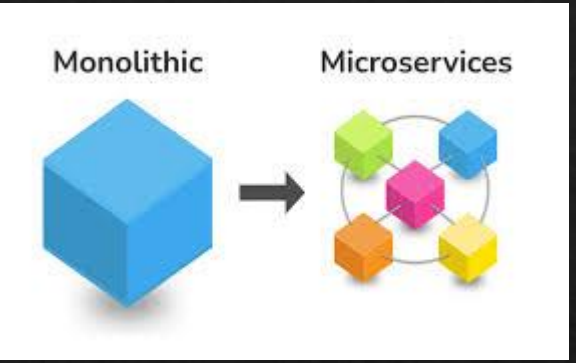






تکنولوژی های کلیدی

Microservice Architecture



Kubernetes

kubernetes

Distributed Storage

Pipelining with Advanced GPU

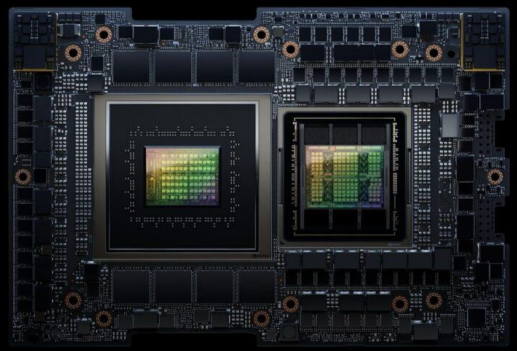
استفاده از ۳۷ گیگابایت رم از ۴۰ گیگابایت رم A100 GPU فقط برای پردازش اطلاعات یک فرد

NVIDIA GH200

Built for the new era of AI supercomputing

CPU to GPU Bandwidth	Memory Bandwidth
900 GB/s NVLink-C2C	4.9 TB/s HBM3e per GPU

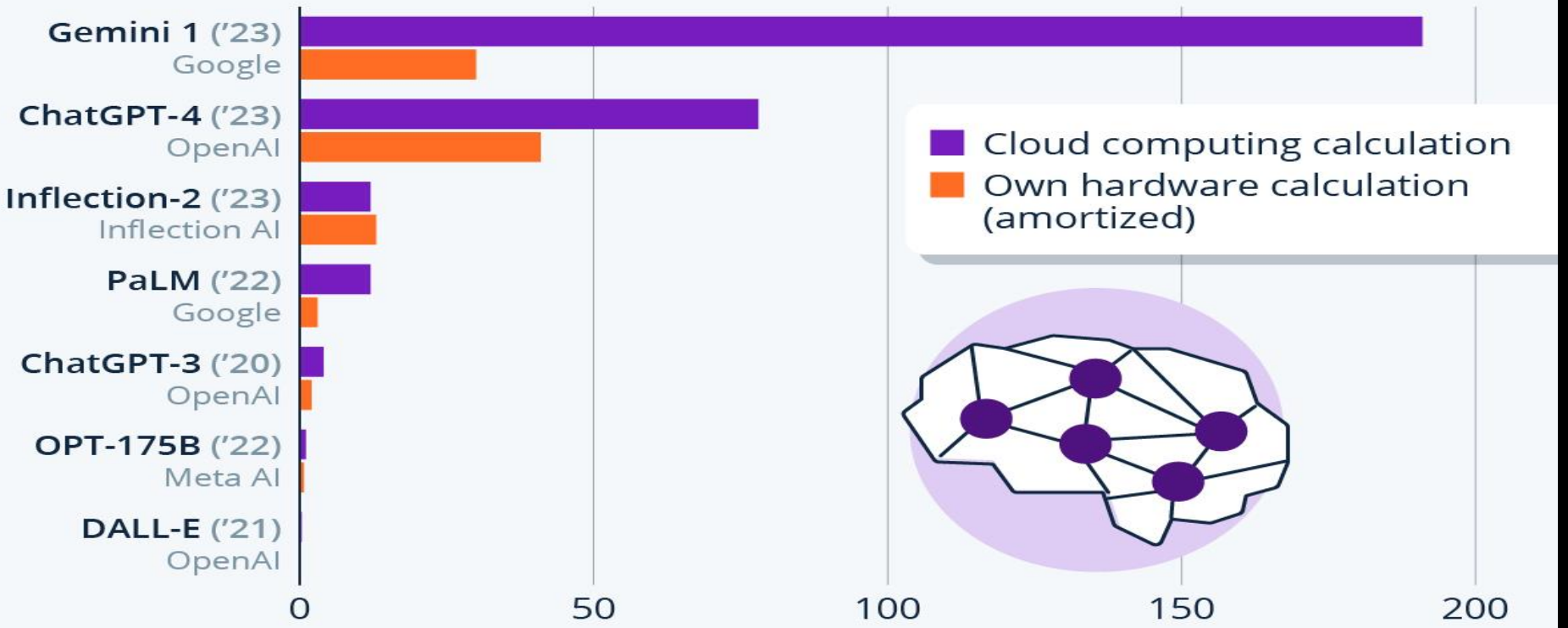
Energy Efficiency	QFT Quantum Simulation	RAG LLM Inference
1.9X Performance vs H100	90X Performance vs dual x86 CPU	100X Performance vs dual x86 CPU



624 GB High-Speed Memory | 4.9 TB/s | 4 PF AI Perf | 72 Arm Cores

The Extreme Cost of Training AI Models

Estimated cost of training selected AI models (in million U.S. dollars), by different calculation models



Rounded numbers. Excludes staff salaries that can make up 29 - 49% of final cost (including equity)

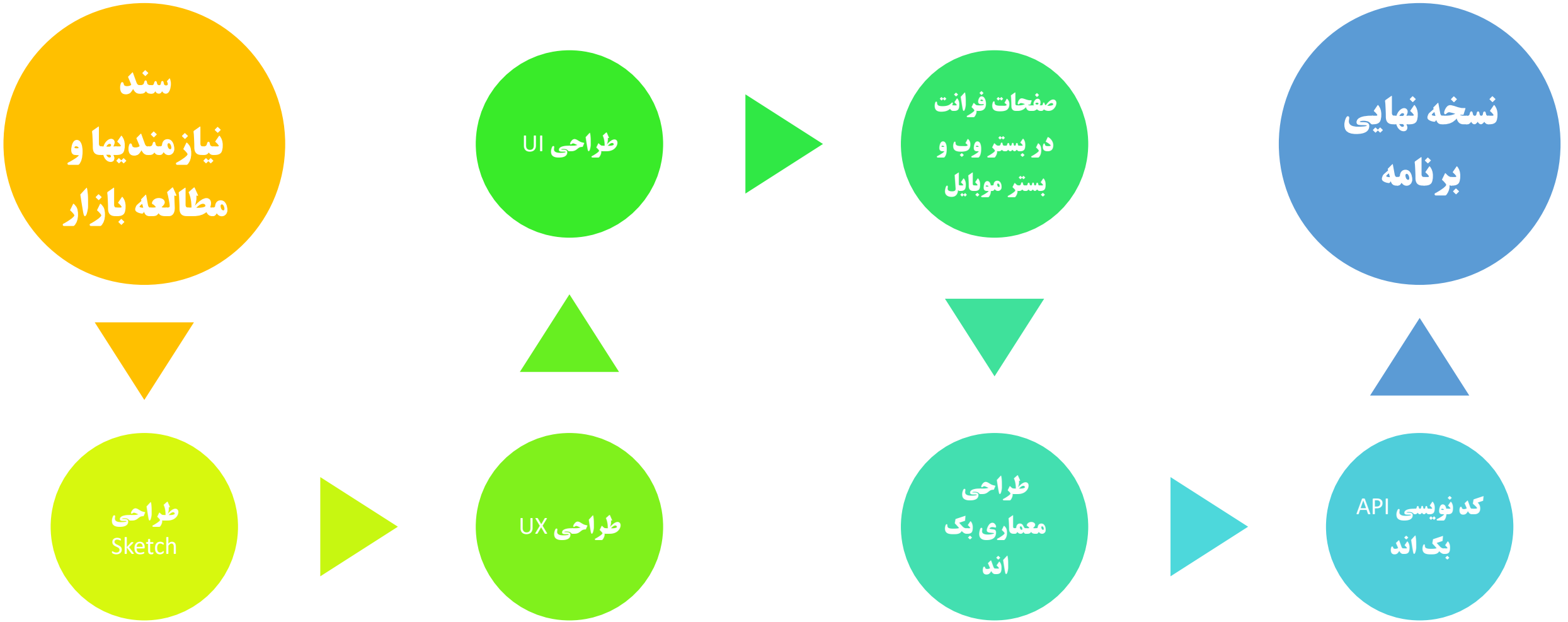
Source: Epoch AI

statista

چالش‌های هزینه آموزش مدل

مراحل توسعه یک نرم افزار

مراحل توسعه نرم افزار



نمونه ای از صفحات App در بستر Figma

MedicMetaVerse

File Assets

Pages

Final

Page 3

Draft

Layers

Frame 1437254888

Frame 1437254887

Anatomy

image 48

UI Kit

Risks

Assistant

buy pack

Incomplete Profile

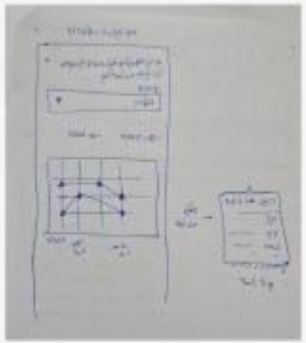
buy pack

Incomplete Profile

buy pack

Frame 1437254793

Sketch



Design Prototype 16%

Page

F6F6F6 100 %

Variables

Styles

Text styles

عنوان 1 · 31/Auto

عنوان 2 · 25/Auto

عنوان 3 · 20/Auto

بدنه 1 · 16/Auto

بدنه 2 · 13/Auto

بدنه 3 · 10/Auto

Export

Share

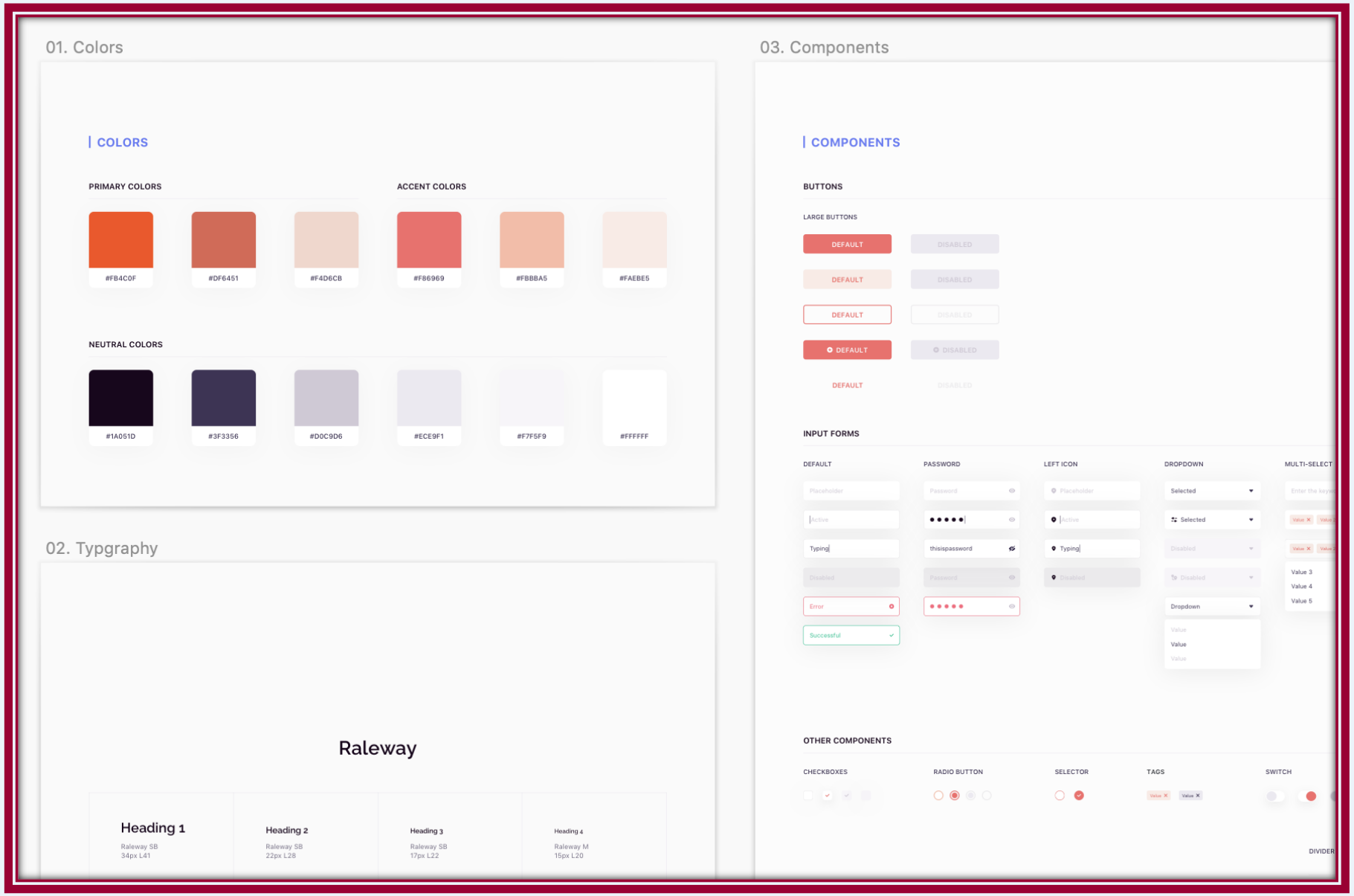
https://www.instagram.com/aspiradesign/

Miller's Law

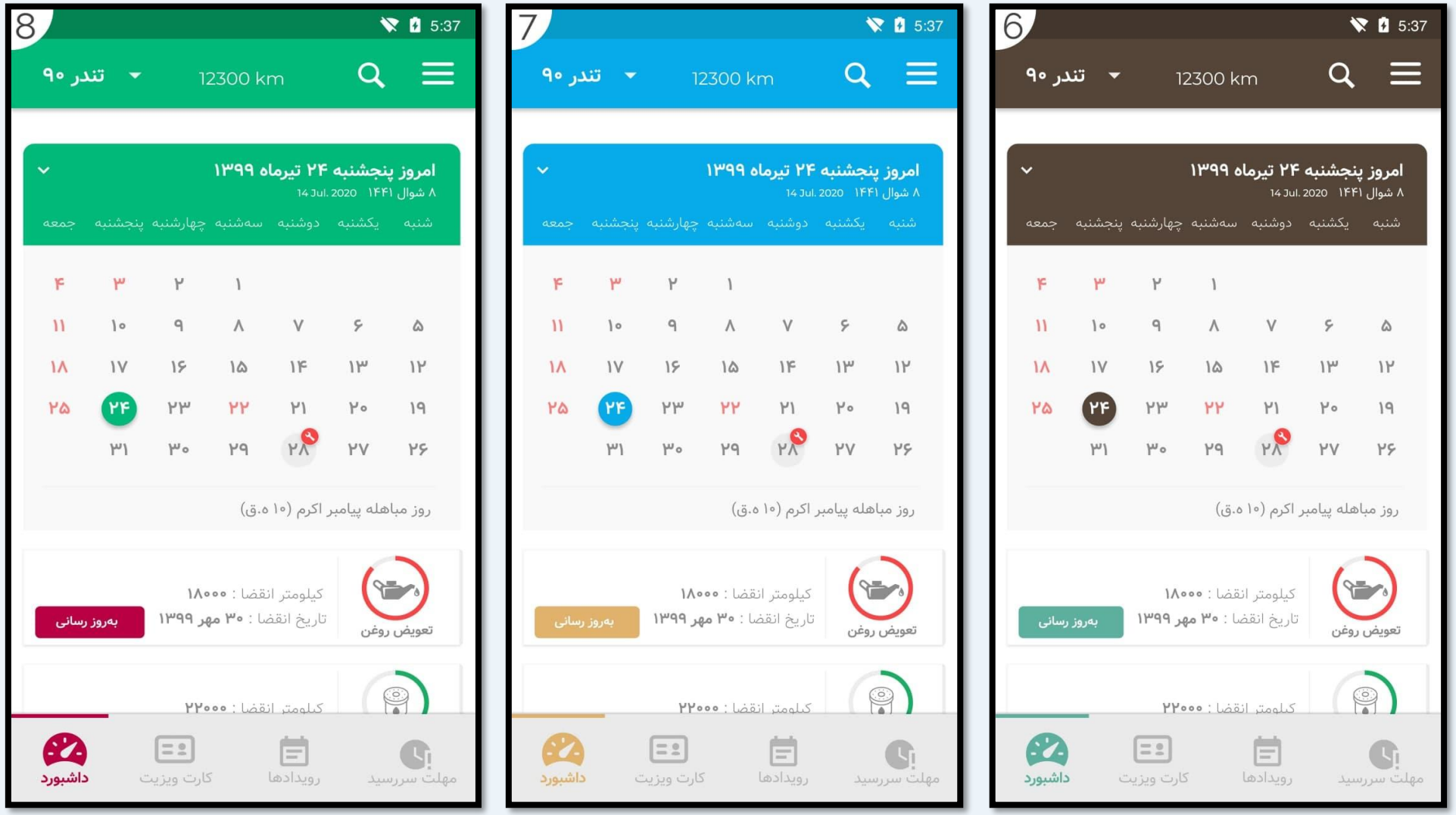
The average person can only keep 7
(plus or minus 2) items in their working memory.

https://www.instagram.com/geekshocom/

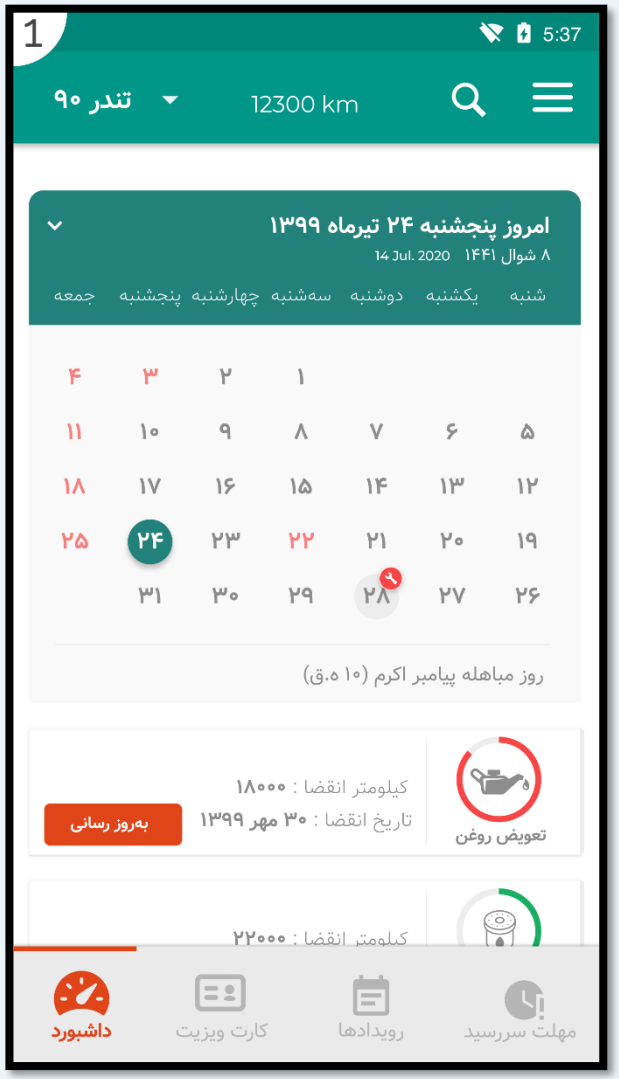
https://www.instagram.com/wwuiuxdesign/

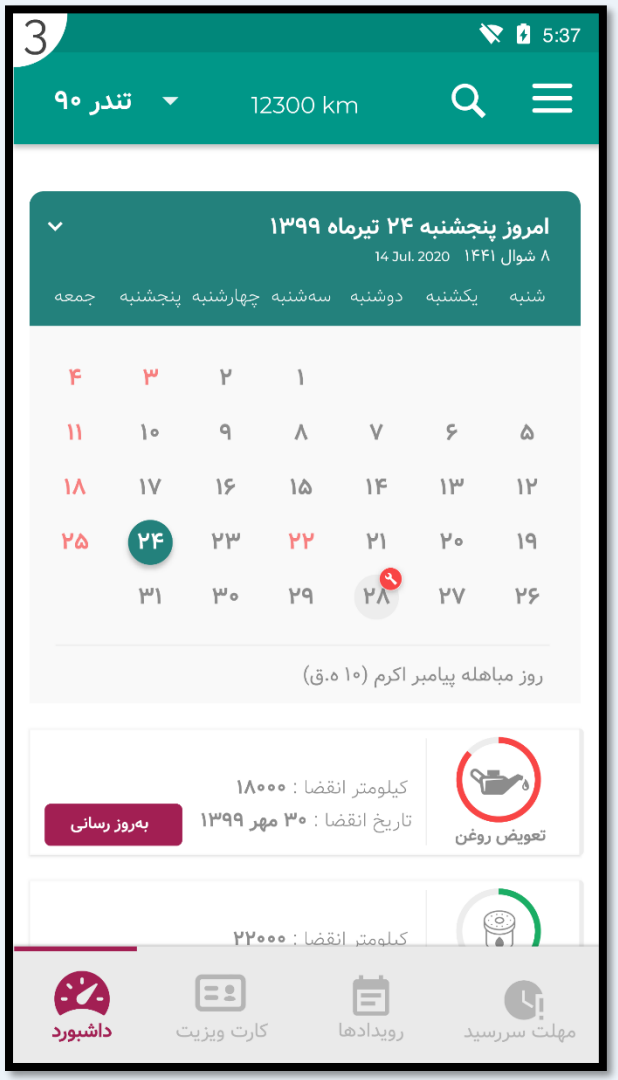


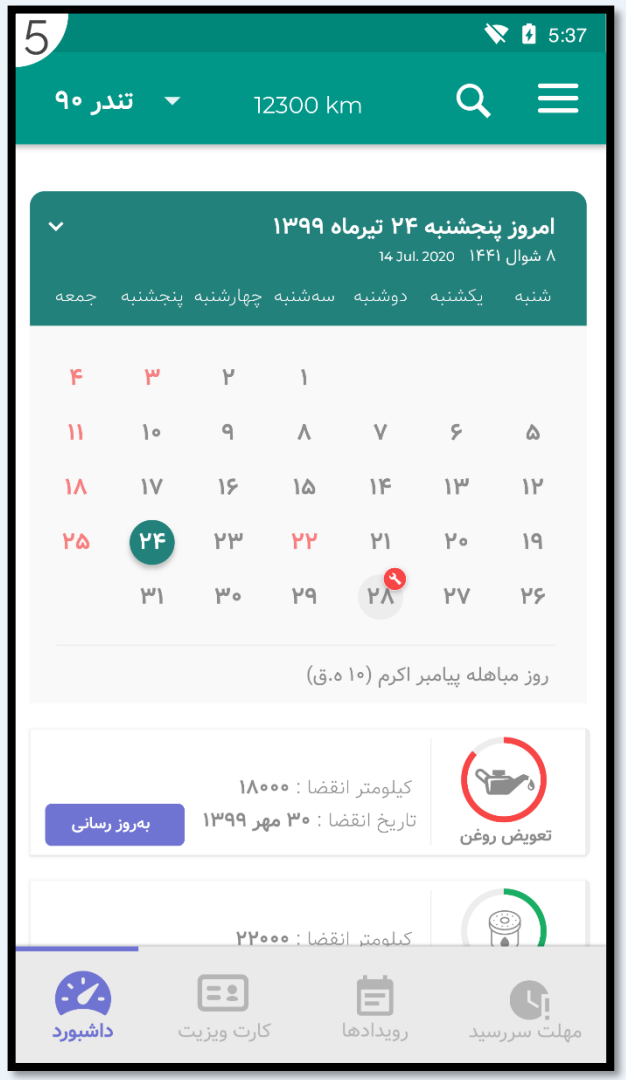
Primary Color

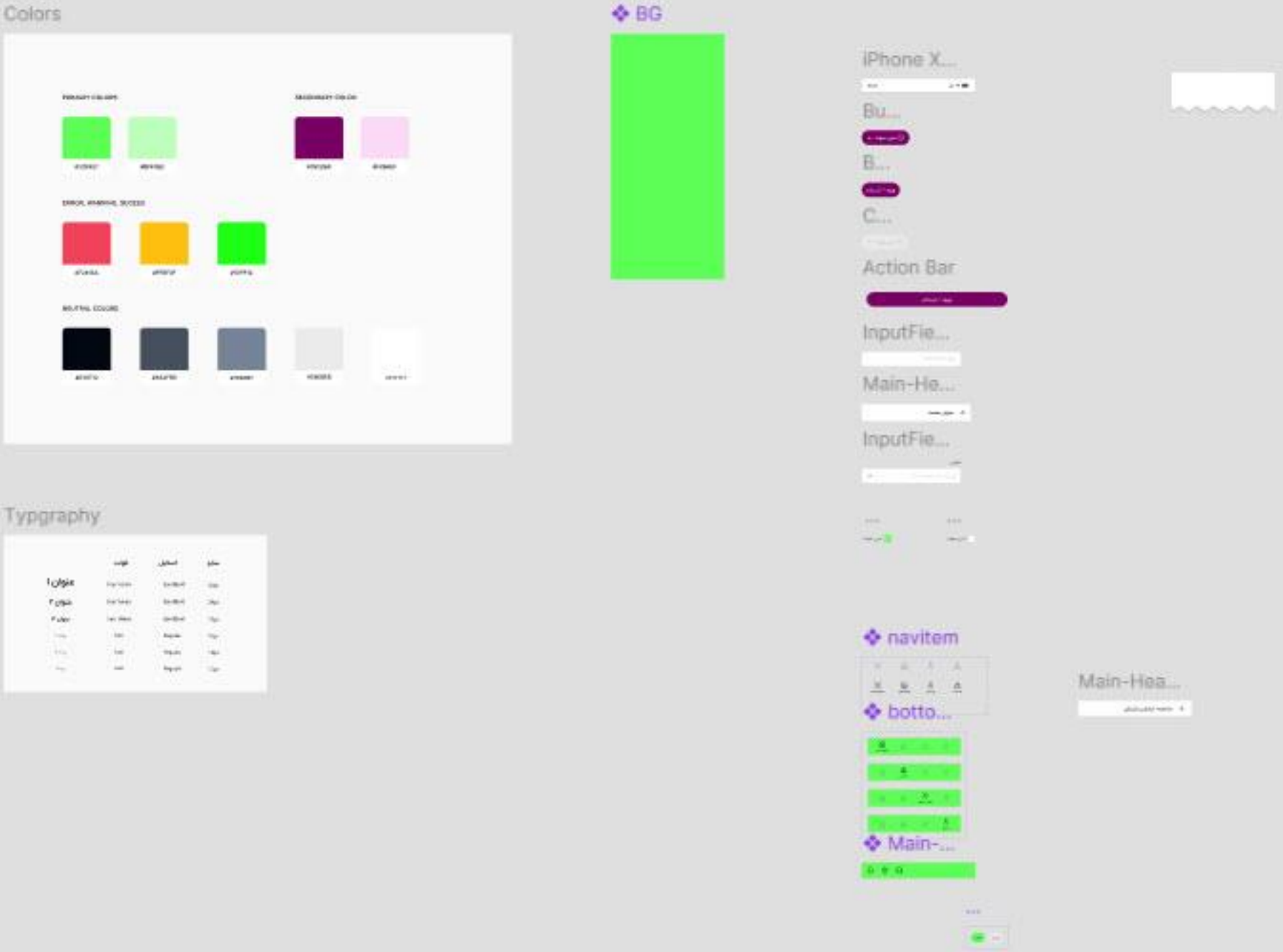


Secondary Color








Colors
BG
iPhone X...
Bu...
B...
C...
Action Bar
InputFie...
Main-He...
InputFie...
navitem
botto...
Main-...
Main-Hea...
Typgraphy

UI Kit

توصیه‌های دستیار پزشک

بر اساس وضعیت کلی سلامت بهتر است موارد زیر را رعایت کنید و در صورت لزوم به پزشک مراجعه کنید

- مصرف چربی‌های ترانس را به حداقل برسانید
- هر شش ماه یک‌بار چکاپ آنزیم‌های کبدی را تکرار کنید
- مصرف سبزیجات روزانه را در رژیم خود بگنجانید
- سریع‌تر آزمایش تری‌گلیسرید را انجام دهید

مشاهده همه توصیه‌ها ←

بیشتر بخوانید

مقالات زیر کمک می‌کنند اطلاعات بیشتری در مورد سلامت خود داشته باشید

از آزمایش خون چه می‌دانیم؟

فرق کلسترول خوب و کلسترول بد چیست؟

آیا برای آزمایش ادرار باید ناشتا باشیم؟

کارنامه سلامت

وضعیت ژنتیکی

ژن	شرایط	تاثیر
BRCA1	Breast Cancer	High
MGF-3	Muscle Degradation	Low
BRCA1	Breast Cancer	Medium
IL-13RA	Asthma	Low
BRCA1	Breast Cancer	High
BRCA1	Breast Cancer	High

مشاهده همه جهش‌ها ←

ریسک بیماری‌ها

نقشه ژنتیک شما نشان‌دهنده ریسک ابتلا به برخی بیماری‌ها است. لیست زیر بر اساس میزان ریسک بیماری‌ها مرتب شده است

آسم Asthma ٪۳۰

هایپوتروفی عضلانی Muscular Hypotrophy ٪۱۲

پنومونی Pneumonia ٪۳

مشاهده همه ریسک‌ها ←

وضعیت کلی سلامت

آزمایش خون و ادرار

بستهٔ فعال ندارید

آزمایش ژنتیک

اعتبار بسته تا ۱۲ اسفند ۱۴۰۴

مشاهده همهٔ امکانات ←

وبلاگ

نقشه‌برداری ژنی در قرن ۲۱ چه پیشرفت‌هایی داشته؟

۲ مرداد ۱۴۰۲ . ژنتیک

هوش مصنوعی به کمک مهندسین ژنتیک آمده و نقشه‌برداری ژنتیکی بسیار دق...

تاثیر ژنتیک بر ورزش

۲۰ شهریور ۱۴۰۴ . عمومی - ژنتیک

ورزشکاران نخبه‌ترین تمرینات قرون گاهی در خصار ژن‌های خود اسیر هستند و بیش...

روش جدید نمونه‌برداری خونی

۱ فروردین ۱۴۰۴ . علوم آزمایشگاهی

بدون درد هوش مصنوعی به کمک مهندسین ژنتیک آمده و نقشه‌برداری ژنتی...

مشاهده همهٔ مطالب ←

صفحه اصلی

آزمایش خون و ادرار

بررسی و تحلیل آزمایش

آزمایش خون ۰۰۱

آزمایشگاه سلامت فردا
دکتر علی محمدی
۲۳ آذر ۱۴۰۳

افزودن آزمایش

آزمایش با موفقیت ثبت شد

مشاهده اطلاعات آزمایش

نوع آزمایش: خون
تاریخ انجام آزمایش: ۲۳ آذر ۱۴۰۳
نام پزشک: دکتر علی محمدی
آزمایشگاه: سلامت فردا
آدرس آزمایشگاه: قلهک، کوچه اسفندیاری
شماره آزمایش: ۱۲۸۰۱۰۳
فایل آپلود شده: 20250312221234.PDF

پارامترهای استخراج شده

با کلیک روی هر سطر راهنمای مربوط به پارامتر را ببینید

پارامتر	واحد	مقدار	حد
WBC	#/ml	12000	00
RBC	#/ml	100000	00
Plaquette	#/ml	100000	00
pH	-	6.9	2.
Triglyceride	mg/l	100000	00
RBC	#/ml	100000	50
Cholesterol	mg/l	12000	00

تمام صفحه

اشتراک‌گذاری
ذخیره

تغییرات در طول زمان

چند پارامتر را انتخاب کنید

پارامترهای آزمایش را انتخاب کنید و تغییرات زمانی را در چارت زیر مشاهده کنید

نمایش اعداد روی چارت

RBC
WBC

افزودن پارامتر

اشتراک‌گذاری
ذخیره چارت

معرفی به دیگران

با دعوت دوستان خود بسته رایگان هدیه بگیرید

با دعوت ۱۰ نفر دیگر به برنامه می‌توانید یک بسته رایگان آزمایش خون یا ژنتیک به انتخاب خود هدیه بگیرید.

۷ نفر دیگر تا دریافت هدیه

- علیرضا محمدی ثبت‌نام کرد
- بهرام سعیدی بسته برنامه را خریداری کرد
- محمد علوی بسته برنامه را خریداری کرد

لینک زیر را برای دوستان خود بفرستید

لینک زیر را کپی کنید و برای دیگران بفرستید یا این‌که از آن‌ها بخواهید کد QR زیر را اسکن کنند

لینک اختصاصی شما

Medic.com/referral=12243908

کپی لینک

نژاد	درصد
آسیایی	٪۴۵
قفقازی	٪۱۲
ایرانی	٪۱۱
ترکیه‌ای	<٪۱
افغان	<٪۱
اروپایی	٪۳۵
ایتالیایی	٪۲۵
اسپانیایی	٪۷
فرانسوی	٪۳
افریقایی	٪۲۰
مصری	٪۱۸
الجزایری	٪۲

مشاهده تاریابی ژنتیکی

نوع فایل: Whole Genome

تاریخ انجام آزمایش: ۲۳ آذر ۱۴۰۳

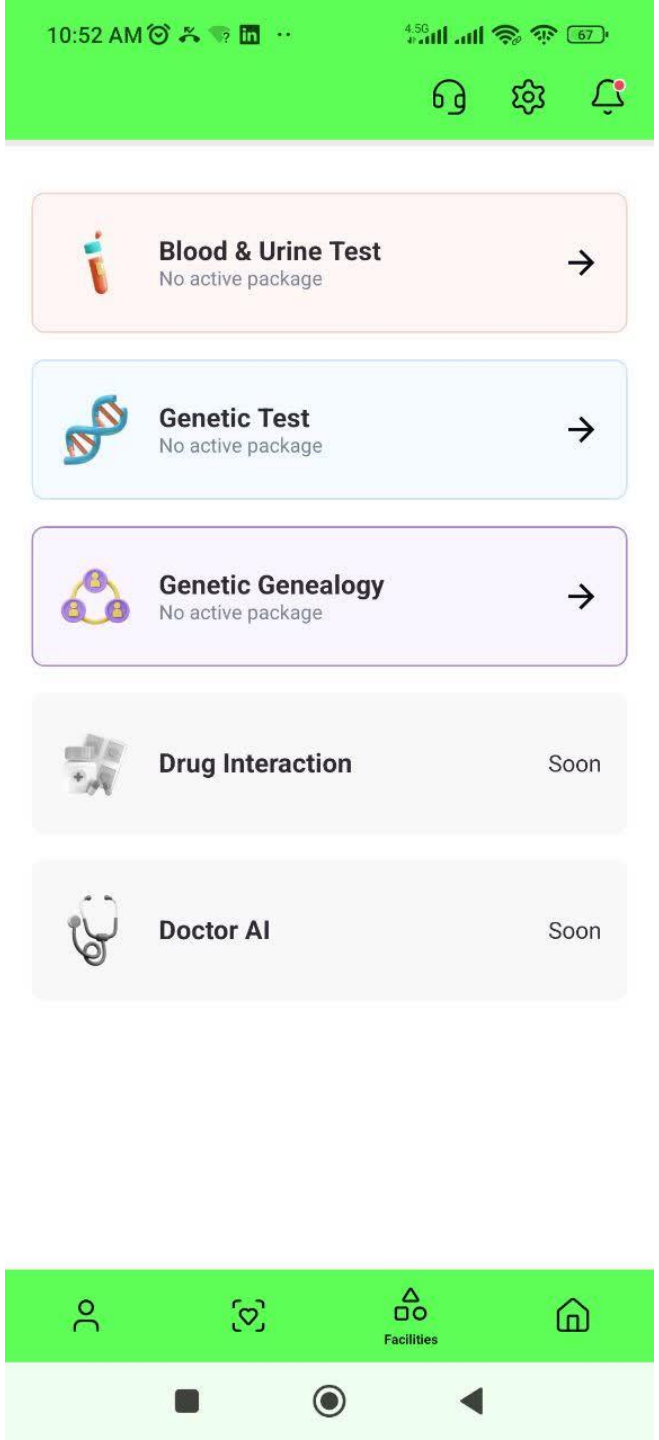
آزمایشگاه: سلامت فردا

شماره آزمایش: ۱۲۸۰۱۰۳

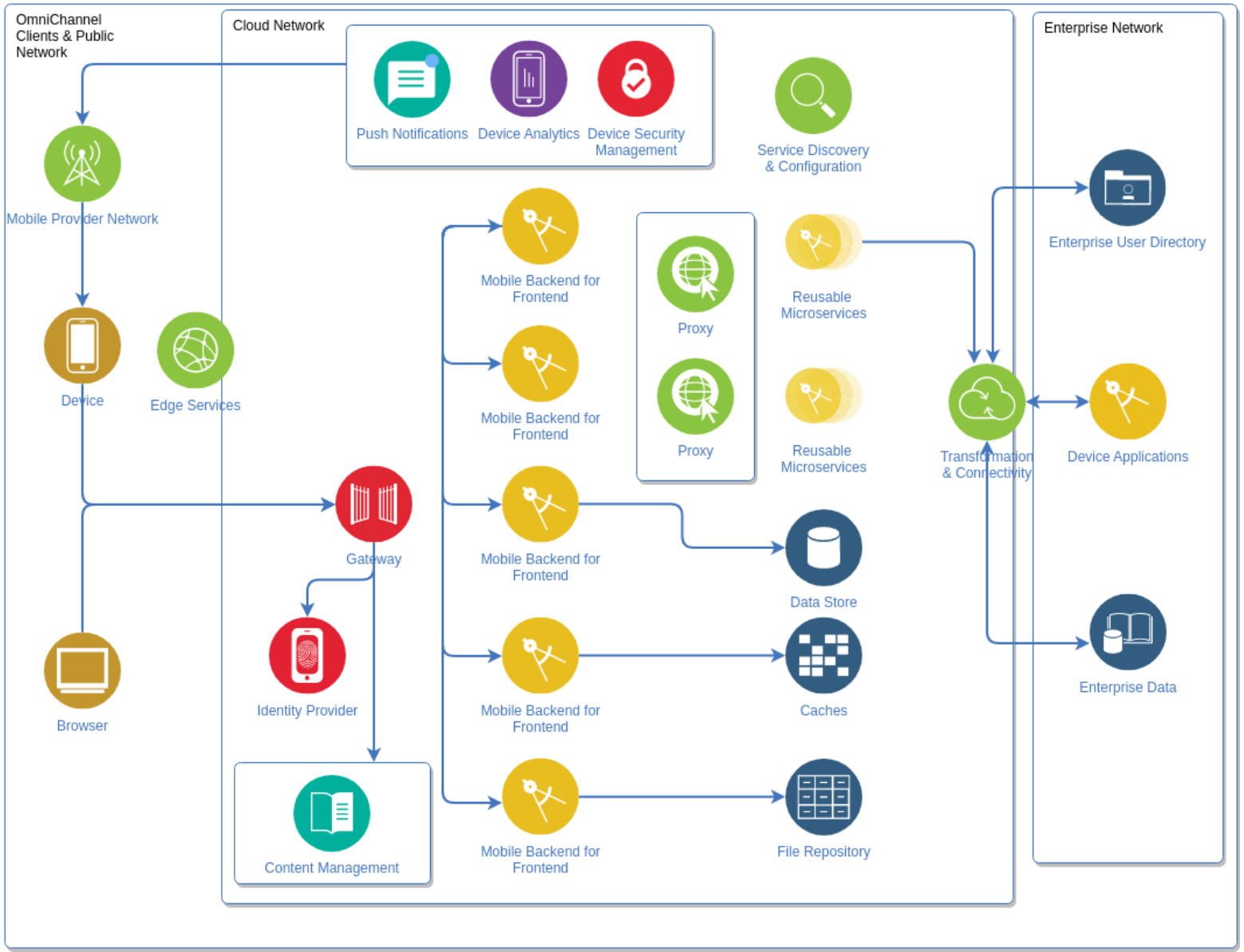
با خرید بسته جمعیت اجدادی خود تاریابی ژنتیکی دقیق‌تر خود را در کشورهای مختلف روی نقشه ببینید

مشاهده نتایج کامل

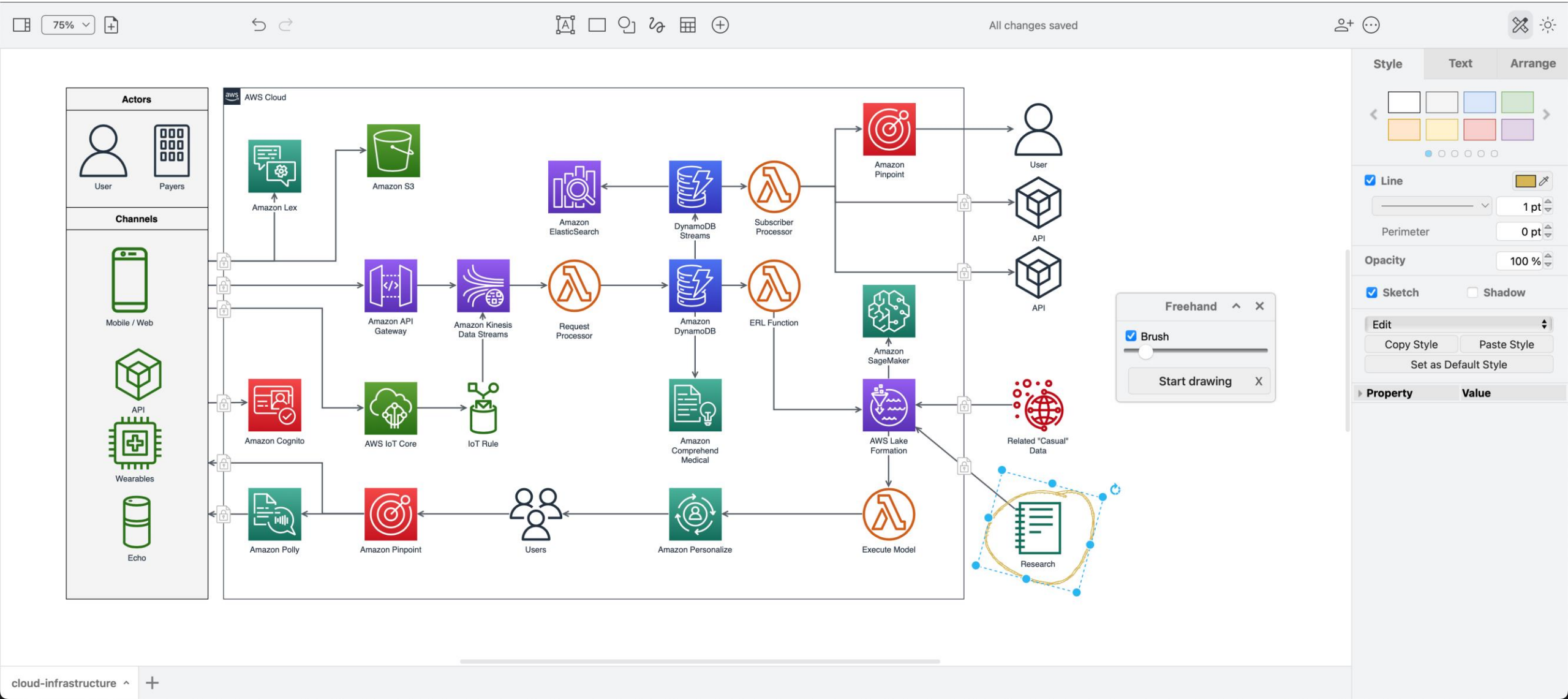
زبان های
فارسی،
عربی،
انگلیسی



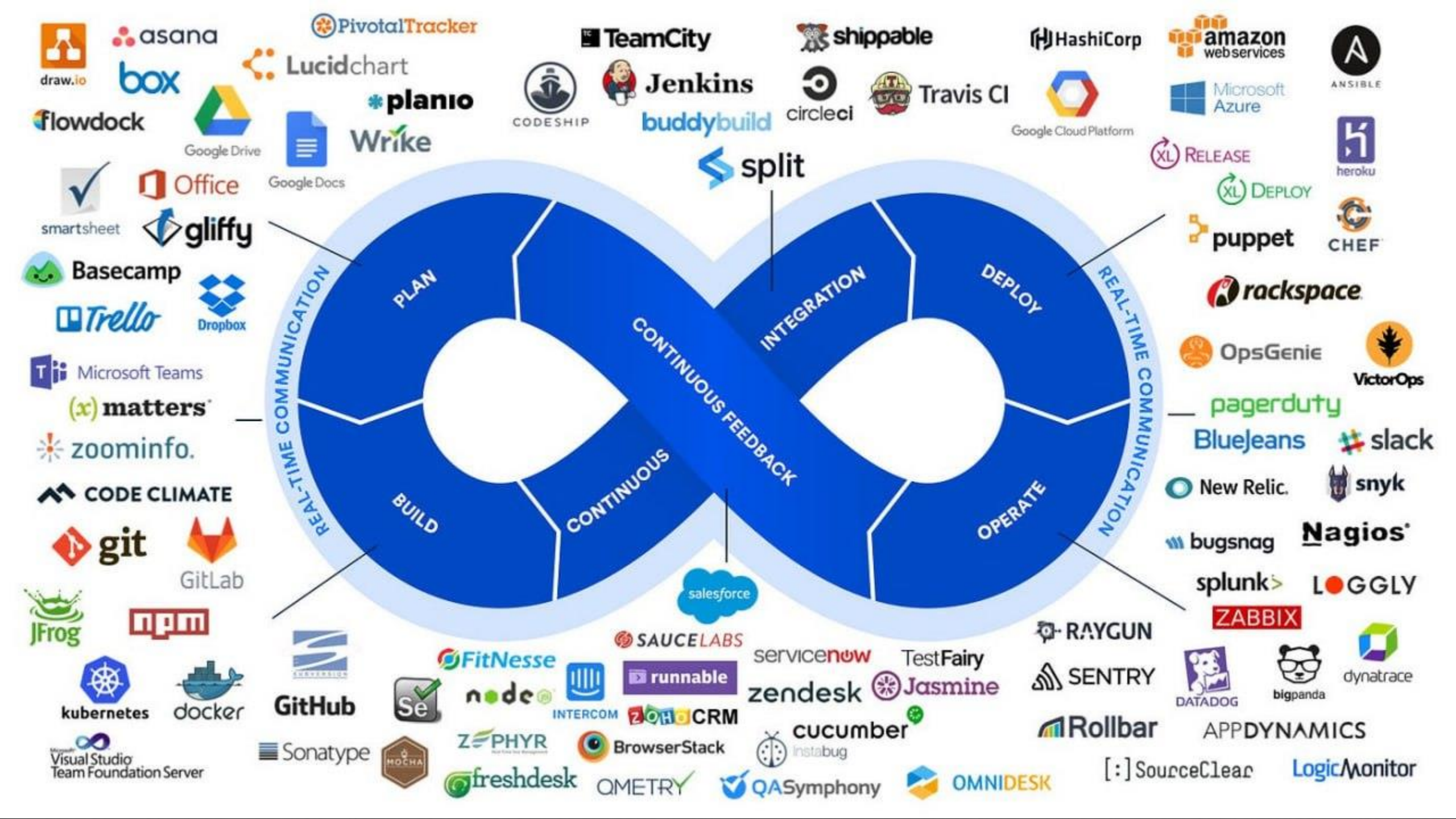
معماری میکروسرویس برنامه در بخش بک اند



Devops **معماری**



نقش مهم ابزارها در حوزه برنامه نویسی و Devops


PLAN
CONTINUOUS
BUILD
CONTINUOUS FEEDBACK
INTEGRATION
DEPLOY
OPERATE
REAL-TIME COMMUNICATION
REAL-TIME COMMUNICATION
draw.io
asana
PivotalTracker
Lucidchart
box
planio
flowdock
Google Drive
Google Docs
Wrike
Office
smartsheet
gliffy
Basecamp
Trello
Dropbox
Microsoft Teams
(x) matters
zoominfo.
CODE CLIMATE
git
GitLab
JFrog
npm
kubernetes
docker
Visual Studio Team Foundation Server
GitHub
Sonatype
TeamCity
CODESHIP
Jenkins
buddybuild
shippable
circleci
Travis CI
split
FitNesse
node
INTERCOM
runnable
ZOHO CRM
ZEPHYR
BrowserStack
freshdesk
QMETRY
SAUCELABS
salesforce
servicenow
TestFairy
zendesk
Jasmine
cucumber
Instabug
QASymphony
OMNIDESK
HashiCorp
amazon web services
Google Cloud Platform
Microsoft Azure
ANSIBLE
RELEASE
heroku
DEPLOY
puppet
CHEF
rackspace
OpsGenie
VictorOps
pagerduty
BlueJeans
slack
New Relic.
snyk
bugsnag
Nagios
splunk>
LOGGLY
ZABBIX
RAYGUN
SENTRY
DATADOG
bigpanda
dynatrace
Rollbar
APPDYNAMICS
[:] SourceClear
LogicMonitor


Big Data Analytics Tools
APACHE HBASE
No SQL Database
APACHE Spark
Realtime Processing
talend
Software Integration Platform
splunk>
Log Analysis Platform
HIVE
Data Warehousing
kafka
Messaging System
hadoop
Data storage and processing Platform
Data Analysis Platform

Big Data Analytics Tools.

MongoDB

Open-source cross-platform source that utilizes a document oriented program.

Apache Hadoop

An open-source platform that helps in the distribution and storage of large data.

Tableau Public

Offers a larger insight into the hypothesis generated.

Knime

Helps in analyzing and manipulating the information through the use of visual programming.

NodeXL

A free open-source network analysis and visualization software tool.

Microsoft HDInsight

Microsoft HDInsight is a big data solution powered by Apache Hadoop.

NoSQL

NoSQL databases are used to store unstructured data which have no particular scheme.

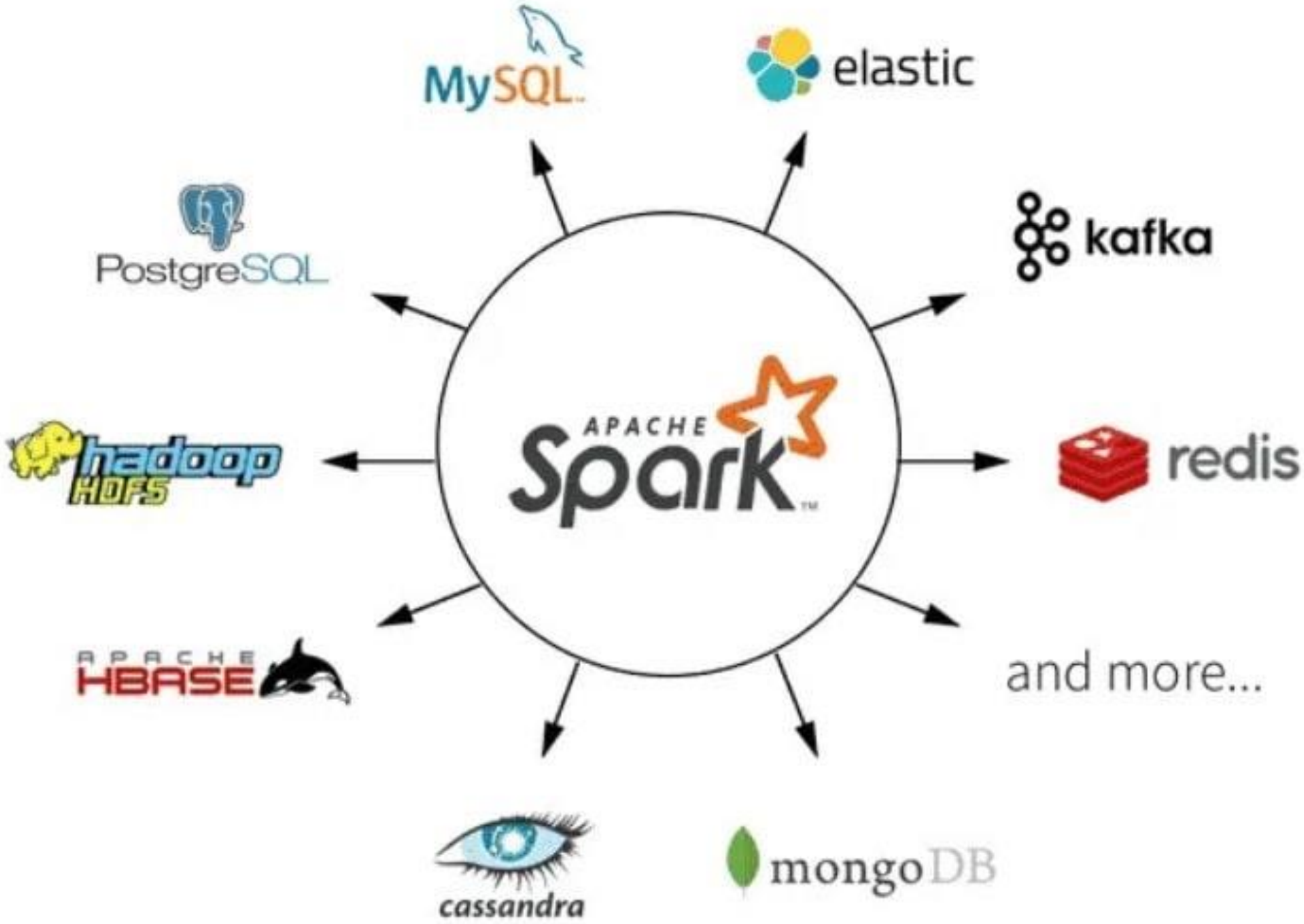
Hive

A distributed data management for Hadoop used for Data mining purpose.

Sqoop

A tool that is used to connect Hadoop with various relational databases that are used to transfer data.

Five Great Big Data Tools


MySQL
elastic
PostgreSQL
kafka
hadoop
HDFS
APACHE
Spark
redis
APACHE
HBASE
and more...
cassandra
mongoDB

مهارت های مورد نیاز

Machine Learning

- Classification
- Regression
- Reinforcement Learning
- Deep Learning
- Clustering
- Dimensionally reduction

Programming Language

- Python
- R
- Java

Data Visualization

- Tableau
- Power BI
- Matplotlib
- GG Plot
- Seaborn

Data Science

Data Analysis

- Feature Engineering
- Data Wrangling
- EDA

IDE

- Pycharm
- Jupyter
- Colaboratory
- Spyder
- R-Studio

Math

- Statistics
- Linear Algebra
- Differential Calculas

Deploy

- AWS
- AZURE

Web Scraping

- Beautiful Soup
- Scrapy
- URLLIB

DATA SCIENCE

By: @pythoncodess

Statistics = Maths

Statistics + Python = Data Analytics

Statistics + Python + Model = Machine Learning

Statistics + Python + Model + Domain Knowledge = Data Science

نمایش داده ها

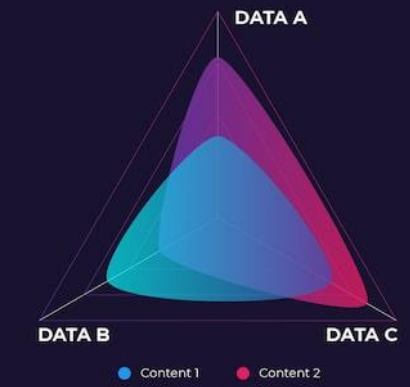
Visualization

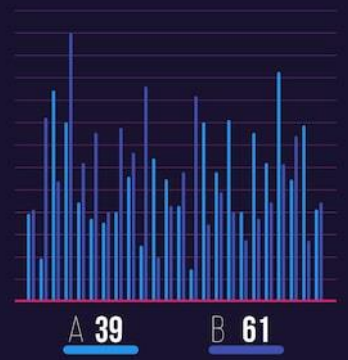
Visualization of data

designed by freepik

designed by freepik

Visualization of data





کتب مرتبط با درس

بیشترین کتب مورد استفاده دانشگاههای برتر جهان در درس کلان داده

- *Database Systems: The Complete Book* — Garcia-Molina, Ullman, Widom web.cs.wpi.edu
- *Modern Database Management* — Hoffer, Ramesh, Topi web.cs.wpi.edu
- *Mining of Massive Datasets* — Leskovec, Rajaraman, Ullman library.buid.ac.ae
- *Hadoop: The Definitive Guide* — Tom White cs.sjtu.edu.cn
- *Learning Spark* — Holden Karau et al.

کتب ۱۰ دانشگاه رتبه برتر دنیا در کلان داده

🔍 مثال‌هایی از دانشگاه‌ها و منابعی که استفاده می‌کنند

دانشگاه / دوره	کتاب‌ها و منابع مورد استفاده
Stanford University دوره *Mining Massive Datasets (CS246)*	از Mining of Massive Datasets — Jure Leskovec, Anand Rajaraman, Jeffrey Ullman استفاده می‌شود. mmds.org
Worcester Polytechnic Institute (WPI) دوره Big Data Management	کتب پیشنهادی شامل: • *Database Systems: The Complete Book* — Garcia-Molina, Ullman & Widom web.cs.wpi.edu • *Modern Database Management* — Hoffer, Ramesh, Topi web.cs.wpi.edu • *Principles of Distributed Database Systems* — Tamer Ozsu & Patrick Valduriez web.cs.wpi.edu
York University دوره Big Data Systems	کتاب اصلی: *Mining of Massive Datasets, 2nd Edition* eecs.yorku.ca
MIT	کتاب *Mathematics of Big Data: Spreadsheets, Databases, Matrices, and Graphs* — Jeremy Kepner & Hayden Jananthan در برخی دوره‌ها استفاده می‌شود. 2+ ...te of Technology

🔍 Examples of Universities and the References They Use in Big Data Courses

University / Course	Books and References Used
Stanford University *Mining Massive Datasets (CS246)*	Uses **Mining of Massive Datasets** — Jure Leskovec, Anand Rajaraman, Jeffrey Ullman (mmds.org ↗)
Worcester Polytechnic Institute (WPI) *Big Data Management*	Suggested books include: • *Database Systems: The Complete Book* — Garcia-Molina, Ullman & Widom (web.cs.wpi.edu ↗) • *Modern Database Management* — Hoffer, Ramesh, Topi (web.cs.wpi.edu ↗) • *Principles of Distributed Database Systems* — Tamer Ozsu & Patrick Valduriez (web.cs.wpi.edu ↗)
York University *Big Data Systems*	Main textbook: *Mining of Massive Datasets, 2nd Edition* (eecs.yorku.ca ↗)
MIT	In some courses, they use *Mathematics of Big Data: Spreadsheets, Databases, Matrices, and Graphs* — Jeremy Kepner & Hayden Jananthan (mit.edu ↗)

کتب مفید درس Big data
برای شروع و دید کلی

- برای شروع و دید کلی
- **Big Data: A Revolution That Will Transform How We Live, Work, and Think –**
- نوشته Viktor Mayer-Schönbergerو Kenneth Cukier
  کتابی غیر فنی که بیشتر درباره اهمیت داده‌های عظیم و تأثیرشان بر اقتصاد و جامعه است.

کتب مفید درس Big data
برای یادگیری مفاهیم فنی و ابزارها

- **Hadoop: The Definitive Guide –**
  - نوشته Tom White
    مرجع کلاسیک برای یادگیری Hadoopزیرساخت پایه‌ای Big Data.
- **Big Data: Principles and Best Practices of Scalable Real-Time Data Systems –**
  - نوشته Nathan Marzو James Warren
    کتابی عالی برای درک اصول طراحی سیستم‌های داده عظیم (از دید معماری).
- **Streaming Systems –**
  - نوشته Tyler Akidauو همکاران (مهندسان گوگل)
    مرجع بسیار خوب برای پردازش داده‌های جریانی stream processing

کتب مفید درس Big data
برای سطح پیشرفته و کاربردی

- **Designing Data-Intensive Applications** – نوشته Martin Kleppmann
  شاید محبوب‌ترین کتاب حال حاضر در جامعه داده؛ مباحثی مثل معماری داده، سیستم‌های توزیع‌شده، پایگاه داده‌های NoSQL، پردازش real-time و batch را پوشش می‌دهد.
- **Mining of Massive Datasets** – نوشته Jure Leskovec, Anand Rajaraman, Jeffrey Ullman
  کتابی دانشگاهی با الگوریتم‌ها و روش‌های داده‌کاوی در مقیاس بزرگ (بسیار مفید برای پژوهش و مبانی نظری).
- خلاصه تفاوت:
- **Kleppmann** → "چطور یک سیستم پایدار و مقیاس‌پذیر برای مدیریت داده طراحی کنیم؟" (تمرکز روی **سیستم‌ها و معماری**)
- **Leskovec et al.** → "چطور از داده‌های عظیم، الگو و دانش استخراج کنیم؟" (تمرکز روی **الگوریتم‌ها و تحلیل داده**)

Designing Data-Intensive Applications (Martin Kleppmann, 2017)

📌 تمرکز اصلی:

- معماری سیستم‌های داده‌ای مدرن
- نحوه طراحی اپلیکیشن‌هایی که باید حجم زیاد داده را با **پایداری، مقیاس‌پذیری و کارایی** مدیریت کنند
- 🔑موضوعات کلیدی:
- مقایسه **SQL vs NoSQL**و انتخاب پایگاه داده مناسب
- معماری سیستم‌های **batch processing**و **stream processing**
- replication, partitioning, consistency, CAP theorem
- طراحی سیستم‌های fault-tolerantو distributed
- 🎯مناسب برای: مهندسان داده، معماران نرم‌افزار، کسانی که می‌خواهند **زیرساخت** بسازند یا انتخاب کنند (مثل Kafka, Cassandra, Spark).

Mining of Massive Datasets (Leskovec, Rajaraman, Ullman, 2020)

- 📌 تمرکز اصلی:
- الگوریتم‌ها و تکنیک‌های **تحلیل و استخراج دانش** از داده‌های بسیار بزرگ
- 🔑موضوعات کلیدی:
- MapReduceو مدل‌های محاسباتی داده‌های عظیم
- الگوریتم‌های خوشه‌بندی (clustering)و طبقه‌بندی (classificationدر مقیاس بزرگ
- PageRankو تحلیل گراف‌های عظیم (شبکه‌های اجتماعی، وب)
- الگوریتم‌های data miningبرای streamو big data

محتوی دور فرنس اصلی درس

Book: Mining of Massive Datasets

موضوعات اصلی کتاب

1. **Introduction to Big Data and MapReduce**
   - مدل محاسباتی MapReduce
   - پیاده‌سازی الگوریتم‌ها در محیط‌های توزیع‌شده
2. **Similarity Search and Locality-Sensitive Hashing (LSH)**
   - تعریف شباهت (Cosine, Jaccard)
   - روش‌های یافتن اقلام مشابه در داده‌های عظیم
3. **Data Stream Mining**
   - مدل داده‌های جریانی (streams)
   - الگوریتم‌های تقریبی برای شمارش، انتخاب نمونه، فرکانس بالا
4. **Link Analysis**
   - الگوریتم **PageRank** (مبنای موتور جستجوی گوگل)
   - تحلیل شبکه‌ها و گراف‌های بزرگ
5. **Clustering**
   - الگوریتم‌های k-means, hierarchical clustering
   - خوشه‌بندی در مقیاس بزرگ
6. **Recommendation Systems**
   - فیلترینگ مشارکتی (collaborative filtering)
   - مدل‌سازی ماتریسی (matrix factorization) برای سیستم‌های توصیه‌گر
7. **Dimensionality Reduction**
   - SVD (Singular Value Decomposition)
   - Principal Component Analysis (PCA)
   - کاربرد در داده‌های با ابعاد بالا
8. **Frequent Itemsets and Association Rules**
   - الگوریتم Apriori
   - تحلیل سبد خرید (market-basket analysis)
9. **Mining Web Advertising and Social Networks**
   - مدل‌های تبلیغات آنلاین
   - تحلیل رفتار کاربران و شبکه‌های اجتماعی
10. **Large-Scale Machine Learning Basics**
    - در نسخه‌های جدید، بخش‌هایی از یادگیری ماشین در مقیاس بزرگ (مثلاً regression, classification در محیط توزیع‌شده) هم اضافه شده است.

Book: Designing Data-Intensive Applications

موضوعات اصلی کتاب

1. **Foundations of Data Systems**
   - چگونه داده‌ها در سیستم‌های مختلف ذخیره و پردازش می‌شوند
   - بررسی trade-offها بین consistency، availability، partition-tolerance (CAP theorem)
   - مفاهیم **data models** و **query languages**
2. **Data Storage and Retrieval**
   - Storage engines: Log-structured storage vs B-trees
   - Indexing, replication, compaction
   - پایگاه داده‌های relation و NoSQL (key-value, document, column-family)
3. **Data Encoding and Evolution**
   - فرمت‌های داده (JSON, Avro, Protocol Buffers)
   - مدیریت schema evolution در سیستم‌های توزیع‌شده
4. **Replication**
   - انواع replication: single-leader, multi-leader, leaderless
   - consistency models و conflict resolution
5. **Partitioning (Sharding)**
   - تقسیم داده‌ها بین سرورها برای افزایش مقیاس‌پذیری
   - طراحی partition key و مسائل مربوط به rebalancing
6. **Transactions**
   - مفهوم ACID و انواع تراکنش‌ها در سیستم‌های توزیع‌شده
   - Consensus protocols: Paxos, Raft
7. **The Trouble with Distributed Systems**
   - چالش‌های سیستم‌های توزیع‌شده: failures, latency, network partitions
   - Idempotence و retry strategies
8. **Batch and Stream Processing**
   - پردازش داده در دسته (batch) و جریان (stream)
   - مدل‌های Lambda و Kappa
   - ابزارها: Hadoop, Spark, Kafka Streams
9. **Derived Data and Materialized Views**
   - View materialization، caching، data pipelines
   - ETL و event sourcing
10. **Consistency and Consensus**
    - Strong vs weak consistency
    - Distributed consensus برای هماهنگی و reliability

ویژگی	Designing Data-Intensive Applications (Kleppmann)	Mining of Massive Datasets (Leskovec, Rajaraman, Ullman)
تمرکز اصلی	معماری و طراحی سیستم‌های داده‌ای در مقیاس بزرگ	الگوریتم‌ها و تکنیک‌های داده‌کاوی و تحلیل داده‌های عظیم
سطح	پیشرفته (معماری نرم‌افزار و مهندسی سیستم‌ها)	دانشگاهی – پیشرفته (علوم داده و الگوریتم‌ها)
موضوعات کلیدی	- SQL vs NoSQL - Replication, Sharding - CAP theorem و consistency models - Batch vs Stream processing - طراحی fault-tolerant systems	- MapReduce و مدل محاسباتی - الگوریتم‌های خوشه‌بندی و طبقه‌بندی - PageRank و تحلیل گراف - الگوریتم‌های mining برای streams - کاهش ابعاد (dimensionality reduction)
نوع نگاه	**System-oriented** (چطور داده را ذخیره و پردازش کنیم در سطح زیرساخت)	**Algorithm-oriented** (چطور داده را تحلیل و از آن دانش استخراج کنیم)
مخاطب اصلی	مهندس داده، معمار نرم‌افزار، توسعه‌دهنده سیستم‌های توزیع‌شده	پژوهشگر، دانشجوی علوم داده/یادگیری ماشین، متخصص تحلیل داده
کاربردها	طراحی و پیاده‌سازی سیستم‌هایی مثل Kafka، Spark، Cassandra، ElasticSearch	تحلیل شبکه‌های اجتماعی، موتور جستجو، توصیه‌گرها، داده‌کاوی در وب
سبک کتاب	توضیح مفهومی + مثال‌های معماری + مقایسه ابزارها	رویکرد ریاضی/الگوریتمی + تمرینات دانشگاهی + اثبات‌ها
پیش‌نیازها	آشنایی با پایگاه داده‌ها و مفاهیم سیستم‌های توزیع‌شده	ریاضی (احتمال، آمار)، الگوریتم‌ها، برنامه‌نویسی سطح دانشگاهی
مناسب برای	کسی که می‌خواهد بداند **چطور سیستم big data طراحی کند**	کسی که می‌خواهد بداند **چطور داده‌های عظیم را تحلیل کند**

سیلابس درس

طرح درس پردازش داده های حجیم–دکتر پورفرد

1-**Basic Concepts**
- Importance of Big data
- Applications of Big data
- 8v concept in Big data
- Challenges of Big data
- Tools of Big data
- Visualizations of Data
- GPU architectures in Big data
- Kappa and Lambda architectures
- MLOPS-machine learning model life cycle
- Curse of Dimensionality
- Supervised vs unsupervised learning
- TEST, TRAIN, VAL, KFOLD

2. **Foundations of Data Systems**
   - Data types and models (relational, document, graph)
   - OLTP vs OLAP systems
   - consistency, availability, and durability
   - Characteristics of reliable, scalable, and maintainable system
3. **Unsupervised ML**
   - Clustering
   - Clustering Problems
   - select clustering algorithm for bigdata
   - k-means and mean shift
   - Hierarchical clustering
   - Defining "nearness"
   - Merging and stopping policy
4. **Data Models and Query Languages**
   - relational vs document vs graph databases
   - SQL و NoSQL
   - query processing basics
5. **Storage and Retrieval**
   - storage engines: log-structured storage, B-trees, LSM-trees
   - indexing techniques
   - data compression and encoding
6. **Encoding and Evolution**
   - serialization formats (JSON, Avro, Protocol Buffers)
   - schema evolution
   - backward/forward compatibility
7. **The Trouble with Distributed Systems**
   - clock synchronization and causality
   - fault tolerance, network partitions, timeouts
   - consistency vs availability trade-offs (CAP theorem)
8. **Batch and Stream Processing**
   - batch processing (Hadoop, MapReduce)
   - stream processing (Spark Streaming, Kafka Streams)
   - exactly-once vs at-least-once semantics
9. **Derived Data**
   - materialized views
   - caches, indexes, search engines
   - change-data-capture and event sourcing
10. **Similarity Search and Locality-Sensitive Hashing (LSH)**
    - Near-duplicate detection (e.g., web pages, documents)
    - Minhashing and shingling techniques
    - LSH for approximate nearest neighbors
11. **Data Stream Mining**
    - Stream models and challenges
    - Sampling techniques
    - Counting distinct elements, frequency moments, heavy hitters
12. **Link Analysis**
    - Graphs and networks basics
    - PageRank algorithm
    - HITS and other ranking methods
13. **Dimensionality Reduction**
    - SVD (Singular Value Decomposition)
    - Principal Component Analysis (PCA)
    - Graphical model (MPPCA, MFA)
14. **Supervised Learning at Scale**
    - Decision trees, SVMs, and logistic regression
    - Stochastic gradient descent (SGD)
    - Handling imbalanced datasets
15. **Partitioning (Sharding)**
    - partitioning strategies: range, hash, composite
    - rebalancing shards
    - handling hotspots
16. **Deep learning at scale**
    - Serving a TensorFlow Model, TFLite, Mobile or Embedded Device
    - Serving through the gRPC and REST API
    - Scaling up TF Serving and Parallelized execution
    - Using GPUs to Speed Up Computations- Managing the GPU RAM
    - Training Models Across Multiple Devices (Model Parallelism vs Data parallelism)
    - TensorFlow Cluster
    - Transformers
    - Mamba

بارم بندی

بارم بندی درس (۲۴ نمره)

- نمرات اصلی
  - پایان ترم ۹ نمره
  - تکالیف ۷ نمره
  - پروژه پایانی ۲ نمره
  - ارایه مقاله ۲ نمره
- نمرات امتیازی
  - تکالیف اختیاری ۲ نمره
  - ارزیابی استاد از دانشجو ۱ نمره اضافی (به صورت استثنا)
  - تولید محتوی ۱ نمره اضافی

موضوعات تکالیف

- ۱- پیاده سازی برخی از الگوریتم ها با زبان برنامه نویسی Python
- ۲- کار با بسترهایی مثل Kafka
- ۳-کار با بسترهایی مثل Spark و Hadoop
- ۴- پیاده سازی یک الگوریتم بر روی بستر GPU با استفاده از CUDA
- ۵- انجام Fine tunning یک شبکه LLM متن باز
- ۶- اجرای یک کد با استفاده از شبکه Mamba