Bisser Raytchev　ライチェフ　ビセル
ビジュアル情報学研究室
bisser@hiroshima-u.ac.jp

ビッグデータって何?
How big is BIG DATA?

Megabytes (MB)

- **Audio files**—High-quality MP3s range from 1 to 2.4 MB per minute

- **Photos**—JPEG format photos taken on a digital camera can require about 8 to 10 MB per photo

- **Video**—Smartphone cameras can record video at various resolutions. Each minute of video can require many megabytes of storage. For example, on iPhones the **Camera** settings app reports that 1080p video at 30 frames-per-second (FPS) requires 130 MB/minute and 4K video at 30 FPS requires 350 MB/minute

ビッグデータって何?
How big is BIG DATA?

Gigabytes (GB)

A dual-layer DVD can store up to 8.5 GB, which translates to:

- as much as 141 hours of MP3 audio
- approximately 1000 photos from a 16-megapixel camera
- approximately 7.7 minutes of 1080p video at 30 FPS
- approximately 2.85 minutes of 4K video at 30 FPS

The current highest-capacity Ultra HD Blu-ray discs can store up to 100 GB of video. Streaming a 4K movie can use between 7 and 10 GB per hour (highly compressed).

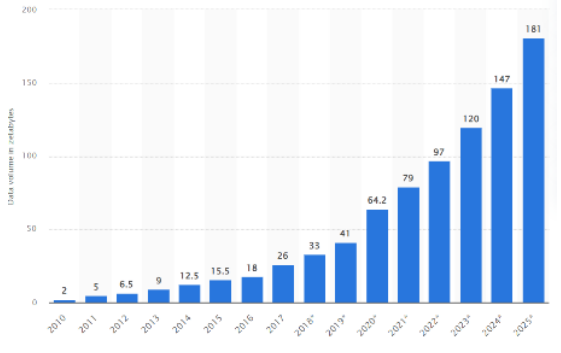
ビッグデータって何?
How big is BIG DATA?

Terabytes (TB)

Recent disk drives for desktop computers come in sizes up to 15 TB (as of 2019), which is equivalent to:

- approximately 28 years of MP3 audio
- approximately 1.68 million photos from a 16-megapixel camera
- approximately 226 hours of 1080p video at 30 FPS
- approximately 84 hours of 4K video at 30 FPS.

ビッグデータって何?
How big is BIG DATA?



2010 2025

Petabytes, Exabytes and Zettabytes

There are nearly 4 billion people online creating about 2.5 quintillion bytes of data each day—that's 2500 petabytes (petabyte is about 1000 terabytes) or 2.5 exabytes (exabyte is about 1000 petabytes). According to a March 2016 *AnalyticsWeek* article, within 5 years there will be over 50 billion devices connected to the Internet and by 2020 we'll be producing 1.7 megabytes of new data every second *for every person on the planet*. At today's numbers (approximately 7.7 billion people), that's about

- 13 petabytes of new data per second
- 46,800 petabytes (46.8 exabytes) per hour
- 1,123 exabytes per day—that's **1.123 zettabytes (ZB) per day** (zettabyte is about 1000 exabytes)

That's the equivalent of over 5.5 million hours (over 600 years) of 4K video every day or approximately 116 billion photos every day!

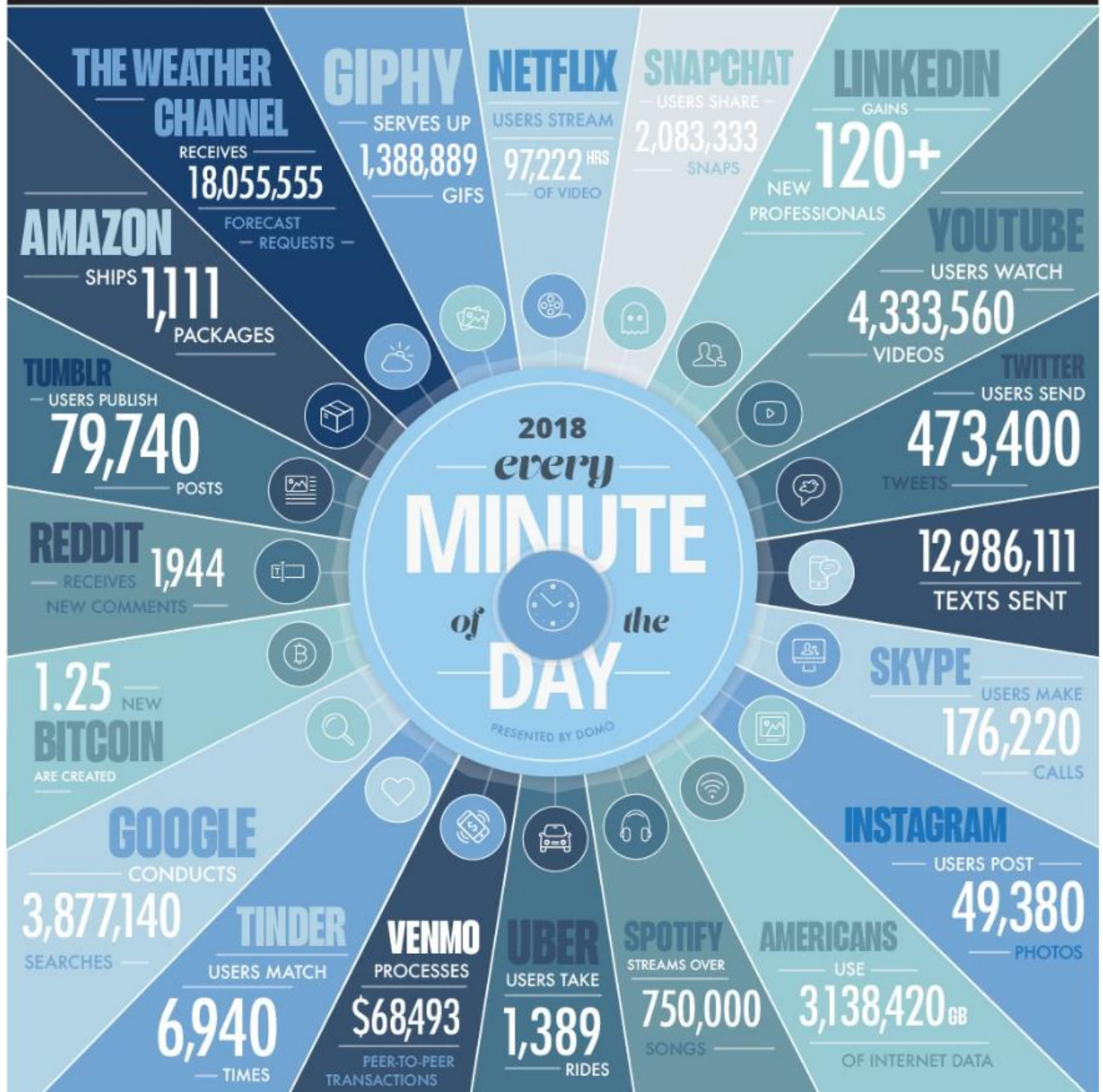
ビッグデータって何?
How big is BIG DATA?

Some other interesting big-data facts:

• Every hour YouTube users upload 24,000 hours of video, and almost 1 billion hours of video are watched on YouTube every day

• Every second, there are 51.773 TBs of Internet traffic, 7894 tweets sent, 64,332 Google searches and 72,029 YouTube videos viewed

• In June 2017, Will Marshall, CEO of Planet, said the company has 142 satellites that image the whole planet's land mass once per day. They add one million images and seven TBs of new data each day. Together with their partners, they're using machine learning on that data to improve crop yields, see how many ships are in a given port and track deforestation. With respect to Amazon deforestation, he said: "Used to be we'd wake up after a few years and there's a big hole in the Amazon. Now we can literally count every tree on the planet every day."



https://www.domo.com/learn/data-never-sleeps-6

ビッグデータって何?

- 米国の調査会社ガートナー（Gartner）の定義

（1）ペタバイトやエクサバイト級の巨大なデータ量

Volume

（2）発信，更新が頻繁に繰り返される発生頻度

Velocity

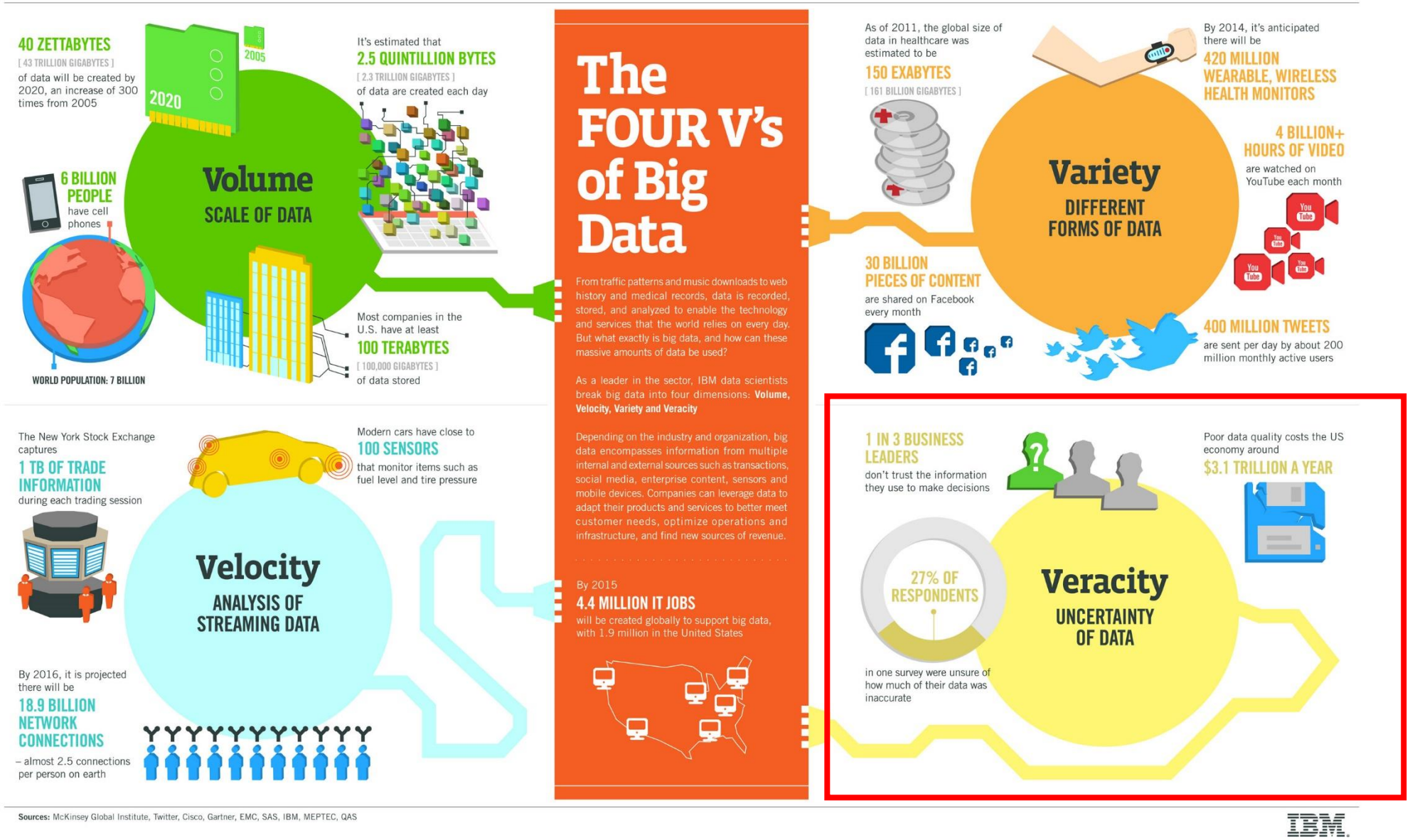
（3）文字に限らずあらゆる種類のデータが，SNSやセンサなど様々な場所から発生する多様性

Variety

企業や研究者のそれぞれの立場によって，
多少ビッグデータのとらえ方に違いがあります

「事業に役立つ知見を導出するためのデータ」とし、ビッグデータは、どの程度のデータ規模かという量的側面だけでなく、どのようなデータから構成されるか、あるいはそのデータがどのように利用されるかという質的側面において、従来のシステムとは違いがある.

鈴木良介著 「ビッグデータビジネスの時代」



+ **Veracity（正確さ）**：データの有効性．データは完全かつ正確だろうか？
重大な決断を下すとき，これらのデータを信頼してよいのだろうか？
データは本物か？

2011年のデータ

+ Value（価値）　様々なデータから「新たな価値」を創造しビジネス競争力を高める

The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.






ICONS: ALEXIDNDZ/ADOBE STOCK

Big Data Use Cases (Big Data is making a difference!)

anomaly detection
assisting people with disabilities
auto-insurance risk prediction
automated closed captioning
automated image captions
automated investing
autonomous ships
brain mapping
caller identification
cancer diagnosis/treatment
carbon emissions reduction
classifying handwriting
computer vision
credit scoring
crime: predicting locations
crime: predicting recidivism
crime: predictive policing
crime: prevention
CRISPR gene editing
crop-yield improvement
customer churn
customer experience
customer retention
customer satisfaction
customer service
customer service agents
customized diets
cybersecurity
data mining
data visualization
detecting new viruses
diagnosing breast cancer
diagnosing heart disease
diagnostic medicine
disaster-victim identification
drones
dynamic driving routes
dynamic pricing
electronic health records
emotion detection
energy-consumption reduction
facial recognition
fitness tracking
fraud detection
game playing
genomics and healthcare
Geographic Information Systems (GIS)
GPS Systems
health outcome improvement
hospital readmission reduction
human genome sequencing
identity-theft prevention
immunotherapy
insurance pricing
intelligent assistants
Internet of Things (IoT) and medical device monitoring
Internet of Things and weather forecasting
inventory control
language translation
location-based services
loyalty programs
malware detection
mapping
marketing
marketing analytics
music generation
natural-language translation
new pharmaceuticals
opioid abuse prevention
personal assistants
personalized medicine
personalized shopping
phishing elimination
pollution reduction
precision medicine
predicting cancer survival
predicting disease outbreaks
predicting health outcomes
predicting student enrollments
predicting weather-sensitive product sales
predictive analytics
preventative medicine
preventing disease outbreaks
reading sign language
real-estate valuation
recommendation systems
reducing overbooking
ride sharing
risk minimization
robo financial advisors
security enhancements
self-driving cars
sentiment analysis
sharing economy
similarity detection
smart cities
smart homes
smart meters
smart thermostats
smart traffic control
social analytics
social graph analysis
spam detection
spatial data analysis
sports recruiting and coaching
stock market forecasting
student performance assessment
summarizing text
telemedicine
terrorist attack prevention
theft prevention
travel recommendations
trend spotting
visual product search
voice recognition
voice search
weather forecasting

データを利用する目的

- ビジネスインテリジェンス（business intelligence ）

　　企業の業績などを集計して，経営上の意思決定に役立てようとするもの

- AI・データマイニング（artificial intelligence, data mining ）

　　統計解析や機械学習などのアルゴリズムを駆使して，データから価値ある情報を見つけ出そうとするもの

これからの講義・演習のスケジュール

講義 （対面）

12月3日（火）・12月10（火）

12月17（火）・1月7（火）

資料はMoodle
でアップロード

演習 （オンライン）

1月21（火）・1月28（火） Python, Jupyter Notebook （Colab）

資料はTEAMS
でアップロード

1月14（火）？

期末試験

2月4日（火） （対面）

評価について

- 期末試験（60%）
- 課題（40%）

人気プログラミング言語のランキング

2023年

・TIOBE Index（TIOBE）　　https://www.tiobe.com/tiobe-index/

TIOBE
the software quality company

About us　Knowledge　News　Coding Standards　TIOBE Index
Products　Quality Models　Markets　Schedule

The index can be used to check whether your programming skills are still up to date or to make a strategic decision about what programming language should be
when starting to build a new software system. The definition of the TIOBE index can be found here.

Nov 2023	Nov 2022	Change	Programming Language	Ratings	Change
1	1		Python	14.16%	-3.02%
2	2		C	11.77%	-3.31%
3	4	↑	C++	10.36%	-0.39%
4	3	↓	Java	8.35%	-3.63%
5	5		C#	7.65%	+3.40%
6	7	↑	JavaScript	3.21%	+0.47%
7	10	↑	PHP	2.30%	+0.61%
8	6	↓	Visual Basic	2.10%	-2.01%
9	9		SQL	1.88%	+0.07%

人気プログラミング言語のランキング

2024年

・TIOBE Index（TIOBE）　　https://www.tiobe.com/tiobe-index/

TIOBE
the software quality company

About us　Knowledge　News　Coding Standards　TIOBE Index

Products　Quality Models　Markets　Schedul

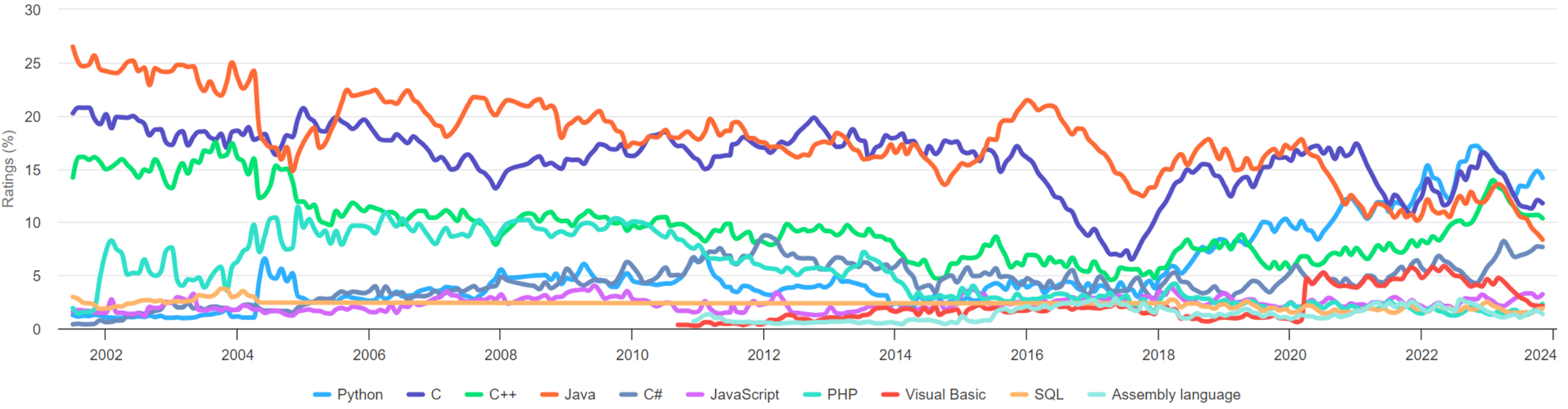
The index can be used to check whether your programming skills are still up to date or to make a strategic decision about what programming language should be when starting to build a new software system. The definition of the TIOBE index can be found here.

Nov 2024	Nov 2023	Change	Programming Language	Ratings	Change
1	1		Python	22.85%	+8.69%
2	**3**	↑	**C++**	**10.64%**	**+0.29%**
3	4	↑	Java	9.60%	+1.26%
4	2	↓	C	9.01%	-2.76%
5	5		C#	4.98%	-2.67%
6	6		JavaScript	3.71%	+0.50%
7	13	⇈	Go	2.35%	+1.16%
8	12	⇈	Fortran	1.97%	+0.67%
9	8	↓	Visual Basic	1.95%	-0.15%

人気プログラミング言語のランキング

2023年

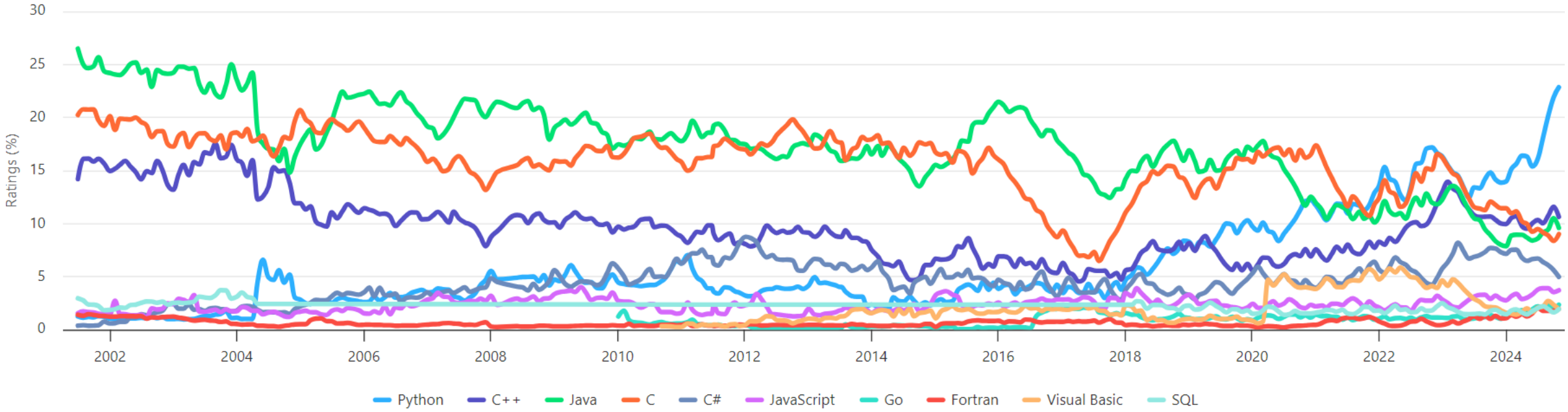
・TIOBE Index（TIOBE）　　https://www.tiobe.com/tiobe-index/



人気プログラミング言語のランキング

2024年

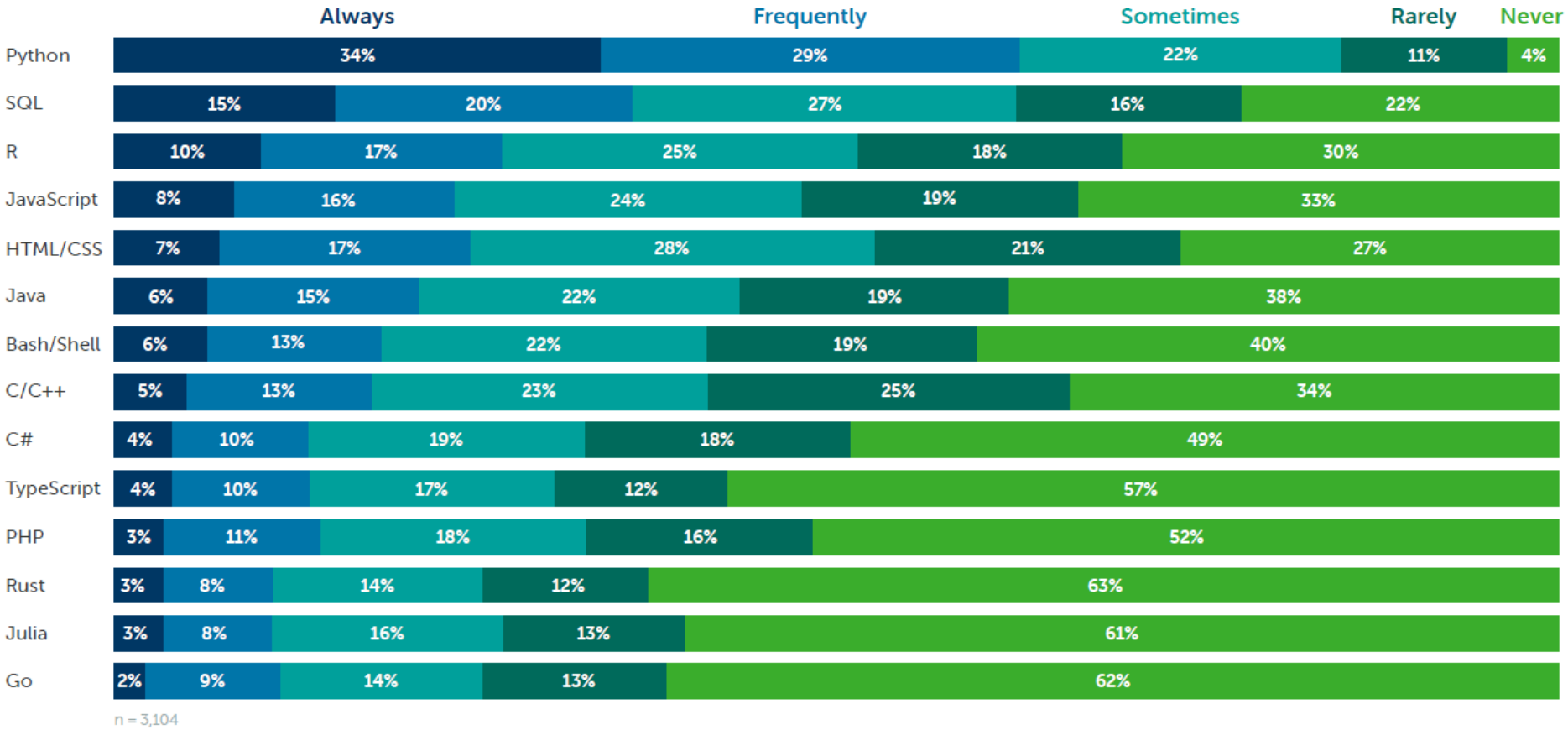
・TIOBE Index（TIOBE）　　https://www.tiobe.com/tiobe-index/



PYTHONについて

POPULARITY OF PYTHON

How often do you use the following languages?

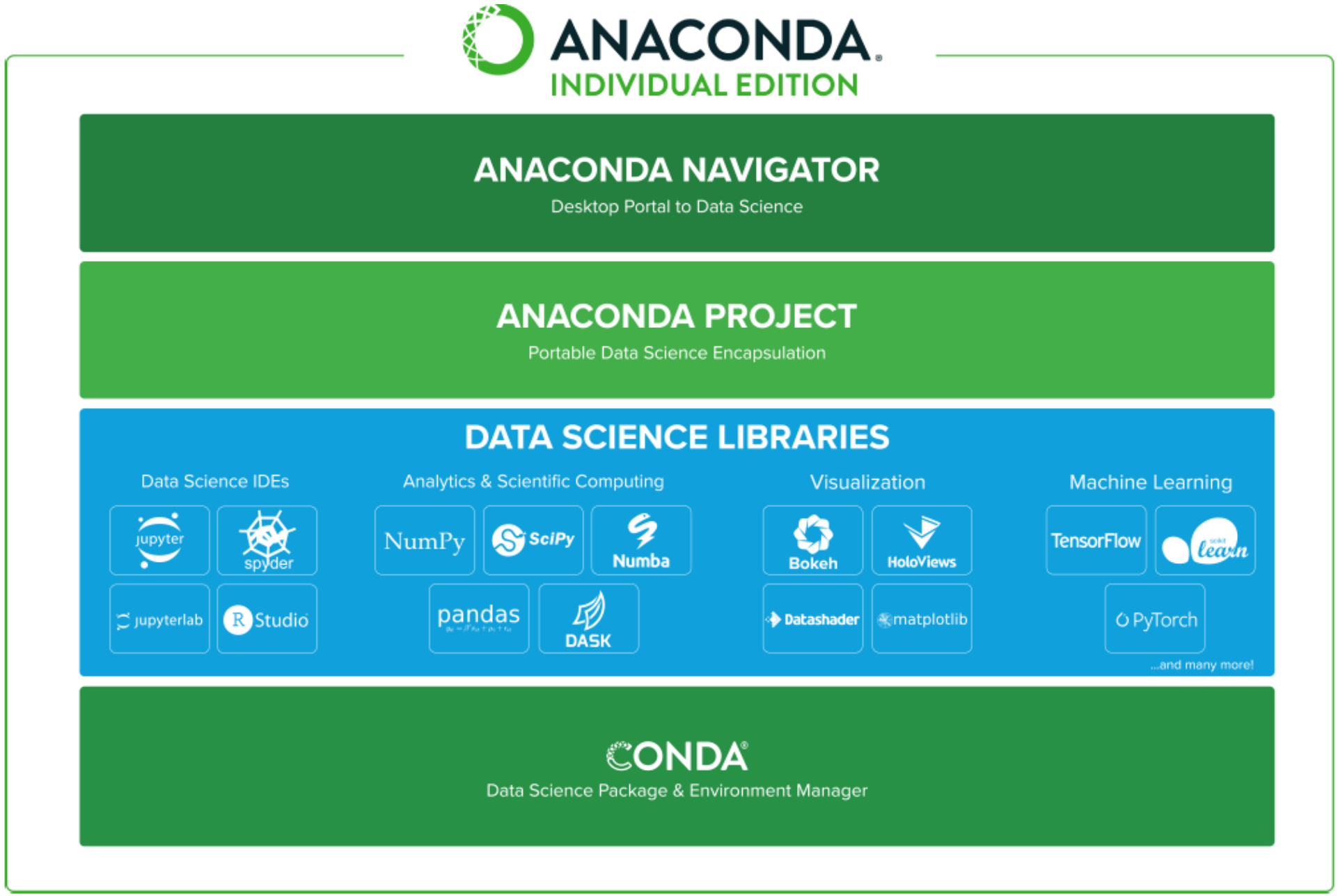


n = 3,104

How often do you use the following languages?

Python appears poised to continue its dominance in the field. 63% of respondents said they always or frequently use Python, making it the most popular language included in this year's survey. In addition, 71% of educators are teaching Python, and 88% of students reported being taught Python in preparation to enter the data science/ML field. Even in our own Anaconda usage data, we've seen impressive growth in Python. Between March 2020 to February 2021, the pandemic economic period, we saw 4.6 billion package downloads, a 48% increase from the previous year. We believe some of this increase could be related to workers transitioning to work from home and more individuals having free time during the pandemic to learn, improve their skills, and pursue their interest in Python.

Beyond being a top language used in commercial environments and taught at universities, Python's popularity can also be demonstrated by various other factors, such as its ease of use, libraries, and community. 20% of students said the biggest obstacle to obtaining the experience required for a career in data science is learning a new language. With most educators teaching Python and Python's continued popularity in the data science community, there is an opportunity for the Python language to become an industry standard. Standardization could help solve re-coding pain points associated with deploying models into production.


ANACONDA
INDIVIDUAL EDITION
ANACONDA NAVIGATOR
Desktop Portal to Data Science
ANACONDA PROJECT
Portable Data Science Encapsulation
DATA SCIENCE LIBRARIES
Data Science IDEs
Analytics & Scientific Computing
Visualization
Machine Learning
jupyter
spyder
jupyterlab
RStudio
NumPy
SciPy
Numba
pandas
DASK
Bokeh
HoloViews
Datashader
matplotlib
TensorFlow
scikit-learn
PyTorch
...and many more!
CONDA
Data Science Package & Environment Manager

https://ai-inter1.com/python-basic/

Python3で学ぶデータ分析・AI・機械学習(Python入門編)

Pythonの基本的なトピックについて、チュートリアル形式で初心者向けに解説した記事です。プログラミング未経験者や初心者でもわかりやすいよう、丁寧に解説しています。

Pythonでデータ分析・AI・機械学習を学ぶ上で欠かせない基礎となる重要な事項を取り上げています。

BIG DATA STATE-OF-THE-ART APPLICATIONS

NVIDIA GTC2024 (watch the video)

https://www.nvidia.com/ja-jp/gtc/keynote/