



Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

PYTHON FOR MACHINE LEARNING, DATA SCIENCE & DATA VISUALIZATION

Bài 1: Tổng quan Data Science



Phòng LT & Mạng

2020



Nội dung



1. Giới thiệu
2. Quy trình của Data Science
3. Lý do chọn Python

Giới thiệu Data Science

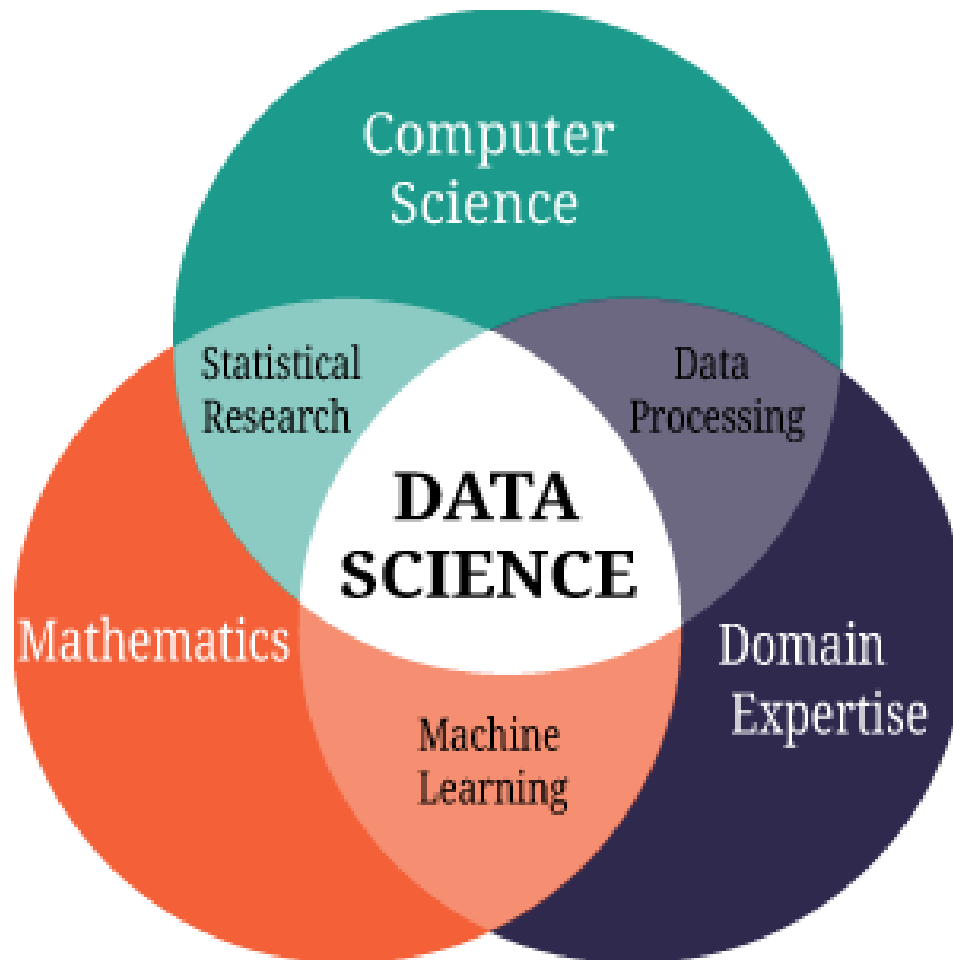


□ Data Science là gì?

- Data science (Khoa học dữ liệu) là một lĩnh vực liên ngành sử dụng các phương pháp, quy trình, thuật toán, và hệ thống khoa học để rút trích kiến thức và thông tin từ dữ liệu dưới nhiều dạng khác nhau, cả cấu trúc và phi cấu trúc tương tự như data mining (khai thác dữ liệu)
- Data science là một “khái niệm hợp nhất giữa thống kê (statistic), phân tích dữ liệu (data analysis), học máy (machine learning) và các phương pháp liên quan” để “hiểu và phân tích hiện tượng thực tế” với dữ liệu
- Data science sử dụng các kỹ thuật và lý thuyết được rút ra từ nhiều lĩnh vực như toán, thống kê, khoa học thông tin và khoa học máy tính.

(Theo https://en.wikipedia.org/wiki/Data_science)













Giới thiệu Data Science



**Data Scientist: The Sexiest
Job of the 21st Century**
(Harvard Business Review, October
2012)



"Chỉ Thượng đế là đáng tin. Mọi
thứ khác đều phải dựa vào dữ liệu"

	AI & Data Scientist (Up to \$4500!!!) Navigos Search via VietnamWorks 19 days ago  ₫69M–₫104M a month	
	Data Scientist - Data Platform Công ty Cổ Phần TIKI via CareerBuilder 15 days ago  Full-time	
	Senior Data Scientist - Search Platform Công ty Cổ Phần TIKI via Tuyển Dụng - Tiki Over 1 month ago  Full-time	
	Data Scientist (1000\$ - 3000\$) Navigos Search' Client via VietnamWorks 18 days ago  ₫23M–₫69M a month	

Data Scientist – cơ hội nghề nghiệp rộng mở với nhiều cơ hội thăng tiến

Giới thiệu Data Science



Giới thiệu Data Science



2019 *This Is What Happens In An Internet Minute*



2020 *This Is What Happens In An Internet Minute*



Giới thiệu Data Science



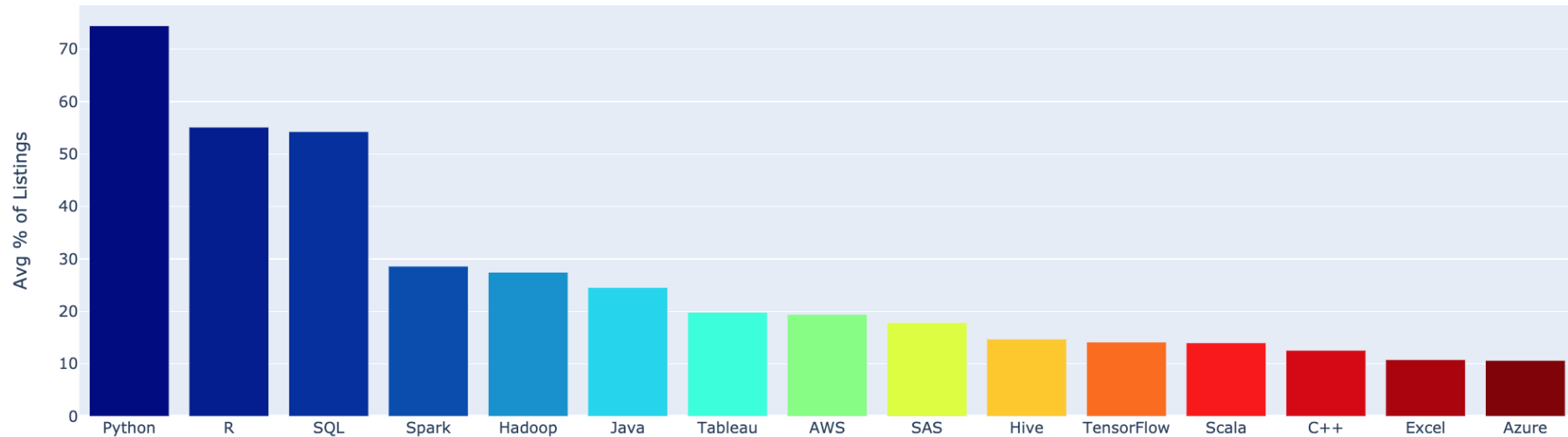
	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	●	●	●
Data Intuition	●	●	●	●
Statistics	●	●	●	●
Data Wrangling	●	●	●	●
Machine Learning	●	●	●	●
Software Engineering	●	●	●	●
Multivariable Calculus and Linear Algebra	●	●	●	●

● Biết thì tốt ● Quan trọng ● Rất quan trọng

Giới thiệu Data Science



Technologies in Data Scientist Job Listings 2019



Trích từ <https://towardsdatascience.com/the-most-in-demand-tech-skills-for-data-scientists-d716d10c191d>



❑ Các kỹ năng khoa học dữ liệu hiện đại

- Lập trình Python (Python programming)
- Phân tích thống kê (Statistical Analysis)
- Học máy (Machine Learning)
- Phân tích dữ liệu lớn (Scalable Big Data Analysis)

Nội dung



1. Giới thiệu
2. Quy trình của Data Science
3. Lý do chọn Python

Quy trình của Data Science



□ Đặt câu hỏi

- Xác định vấn đề

- Ví dụ:

- Dữ liệu marketing + dữ liệu khách hàng => mục tiêu marketing tốt hơn
 - Dữ liệu khách hàng thân thiết + dữ liệu khách hàng tiềm năng => mục tiêu ra mắt sản phẩm

**“A problem well defined
is a problem half
solved.”**

Charles F. Kettering

Quy trình của Data Science



- Đánh giá tình huống:
 - Dựa vào các yếu tố như: rủi ro, thuận lợi, dự phòng, quy định, tài nguyên, yêu cầu



Quy trình của Data Science



□ Xác định mục tiêu

- Các mục tiêu
- Điều kiện

Quy trình của Data Science



Xác định vấn đề

Đánh giá tình huống

Xác định mục tiêu

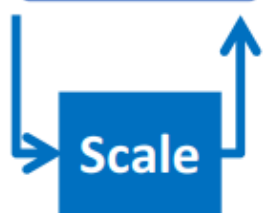
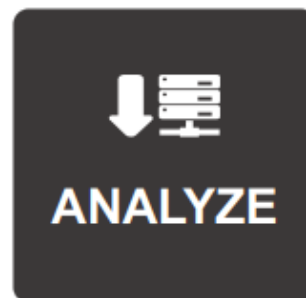
Xây dựng câu hỏi

Quy trình của Data Science



Data Engineering

Computational Data Science



Programmability

Quy trình của Data Science



□ Bước 1: Thu thập dữ liệu

- Xác định các bộ dữ liệu
- Truy vấn dữ liệu

Quy trình của Data Science



❑ Bước 1: Thu thập dữ liệu

- Dữ liệu có từ đâu? – Dữ liệu có thể đến từ nhiều nguồn với nhiều cách truy xuất chúng (từ CSDL truyền thống, các loại tập tin, remote data, dữ liệu phi cấu trúc)
 - Xác định dữ liệu phù hợp
 - Thu thập từ tất cả các dữ liệu có sẵn

Quy trình của Data Science



❑ Bước 2: Chuẩn bị dữ liệu

- Khám phá dữ liệu
 - Tìm hiểu về bản chất của dữ liệu
 - Phân tích sơ bộ
- Tiền xử lý
 - Làm sạch
 - Tích hợp
 - Đóng gói

Data preparation is
very important for
meaningful analysis!

Quy trình của Data Science



❑ Bước 2: Chuẩn bị dữ liệu

- Khám phá dữ liệu – tại sao phải khám phá?
 - Tìm hiểu về bản chất của dữ liệu: tương quan, xu hướng chung, các ngoại lai (outliers)
 - Phân tích sơ bộ: mô tả dữ liệu (tìm mean, mode, median, range) và quan sát dữ liệu (dùng histogram, line graph, scatter plot, boxplot)

Khám phá dữ liệu



Hiểu dữ liệu



Phân tích thông tin

Quy trình của Data Science

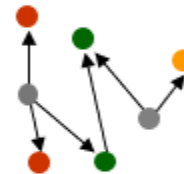


❑ Dữ liệu (*data*)



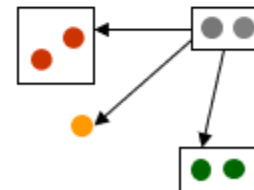
(rời rạc)

❑ Thông tin (*information*)



(mối liên hệ với nhau)

❑ Tri thức (*knowledge*)



(quy luật chung /
khuôn mẫu của các đối tượng)

Quy trình của Data Science



❑ Bước 2: Chuẩn bị dữ liệu

● Tiền xử lý

■ Làm sạch: vì dữ liệu thu được rất lộn xộn

- Các giá trị không nhất quán
- Mẫu tin trùng lặp
- Giá trị bị mất
- Dữ liệu không hợp lệ
- Giá trị ngoại lệ (mẫu ngoại lai - outlier)

=> Giải quyết vấn đề về chất lượng dữ liệu dựa trên Domain Knowledge

- Bỏ/thay thế các dữ liệu bị thiếu giá trị
- Gộp các mẫu tin bị trùng lặp
- Tạo ước tính tốt nhất cho các giá trị không hợp lệ
- Bỏ các mẫu ngoại lai

Quy trình của Data Science



- Tiền xử lý

- Lấy dữ liệu theo Shape:

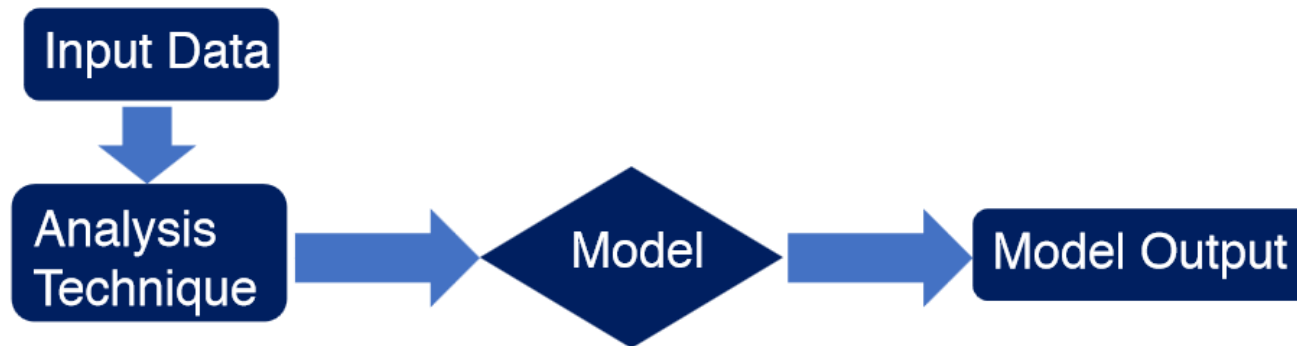
- Giảm dữ liệu: giảm kích thước, thao tác dữ liệu, chuyển đổi dữ liệu, chọn lựa theo đặc điểm (loại bỏ, kết hợp, thêm), chia tỷ lệ
 - Sắp xếp dữ liệu
 - Tiền xử lý dữ liệu

Quy trình của Data Science



❑ Bước 3: Phân tích dữ liệu

- Lựa chọn các kỹ thuật phân tích
- Xây dựng các mô hình



Quy trình của Data Science



- Xây dựng các mô hình

- Phân loại các kỹ thuật phân tích

- Classification (Phân nhóm gán nhãn): mục tiêu là dự đoán Category
 - Clustering (Phân nhóm tương tự): mục tiêu là sắp xếp các item tương tự vào các nhóm
 - Regression (Hồi quy): mục tiêu là dự đoán giá trị số
 - Graph Analytics (Phân tích biểu đồ): mục tiêu là sử dụng các cấu trúc biểu đồ để tìm mối liên hệ giữa các thực thể
 - Association Analytics (Phân tích kết hợp): tìm quy tắc để nắm bắt các liên kết giữa các item

Chọn kỹ thuật



Xây dựng model



Xác thực model

Quy trình của Data Science



❑ Bước 4: Báo cáo

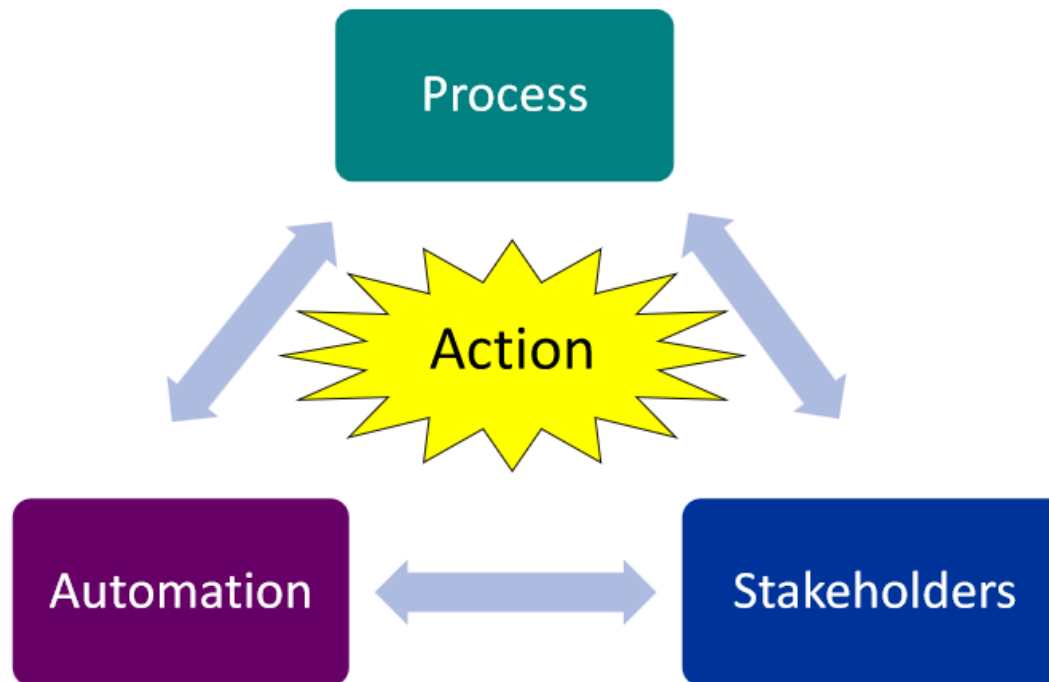
- Trao đổi về các kết quả
 - Báo cáo các nội dung gì?
 - Báo cáo như thế nào?
 - Sử dụng công cụ trực quan nào? (R, Python, google Developers Charts...)

Quy trình của Data Science



❑ Bước 5: Thực hiện

- Chuyển thông tin chi tiết từ báo cáo thành hành động

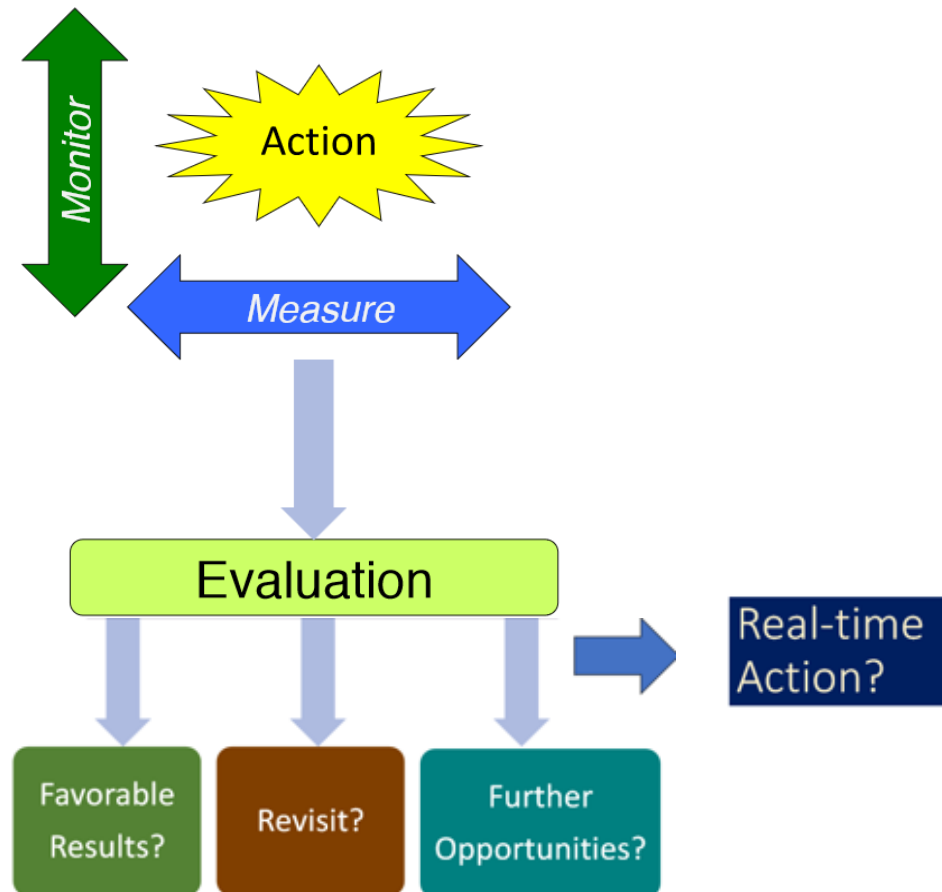




Quy trình của Data Science

❑ Bước 5: Thực hiện

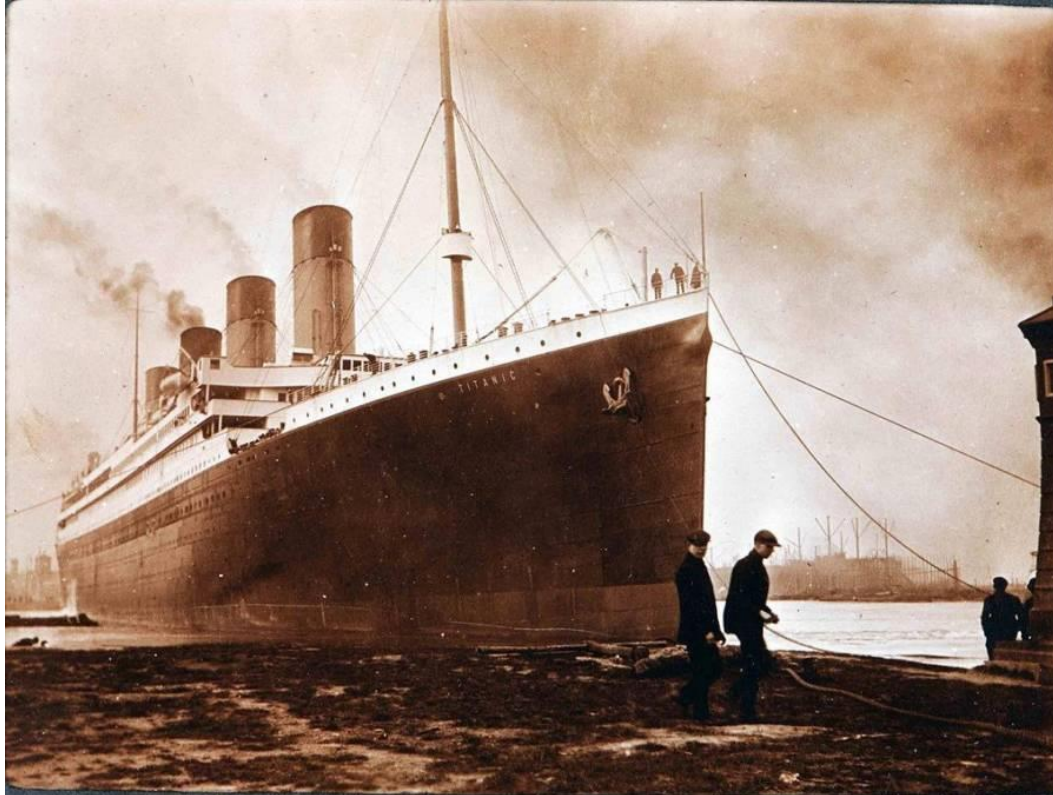
- Đánh giá tác động



Quy trình của Data Science



Quy trình của Data Science



Câu hỏi: Những loại hành khách nào có khả năng sống sót cao hơn ?

Quy trình của Data Science



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

- **train.csv**: dữ liệu của hành khách, bao gồm thông tin còn sống hay không
- **test.csv**: dữ liệu của hành khách, không có thông tin còn sống hay không.

Quy trình của Data Science



Chuẩn bị dữ liệu

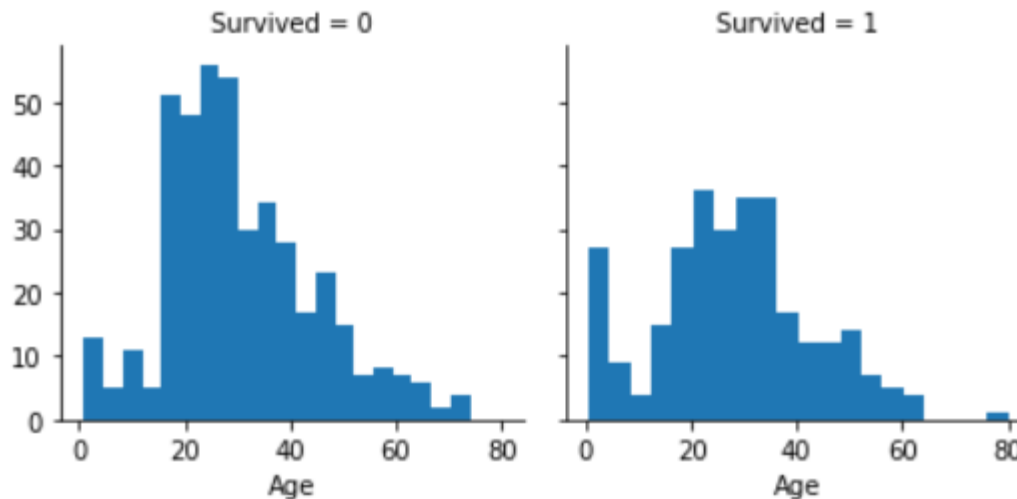
- Khám phá dữ liệu
- Tiền xử lý

Quy trình của Data Science

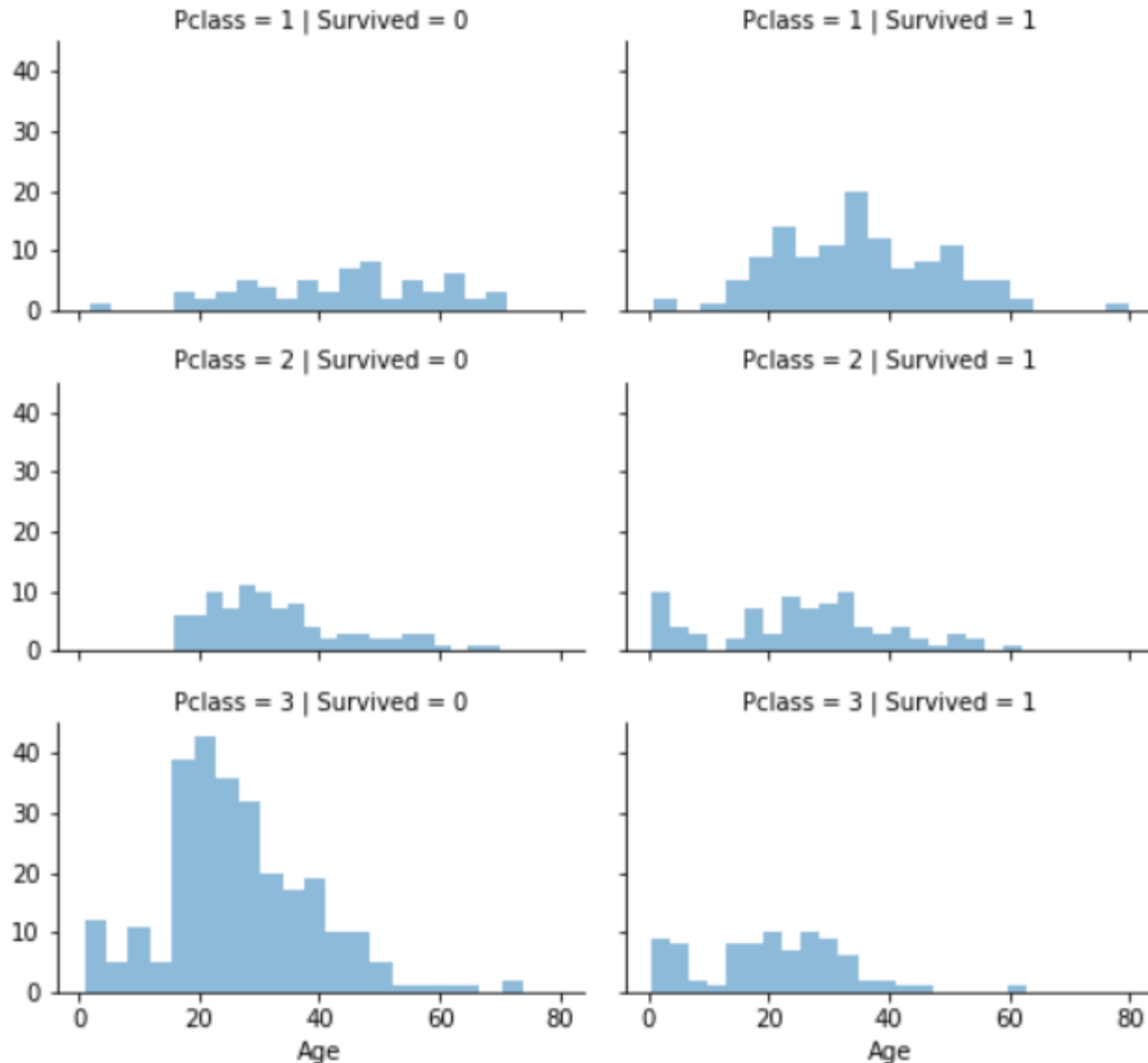


Phân tích dữ liệu

- Lựa chọn các kỹ thuật phân tích
- Xây dựng các mô hình



Quy trình của Data Science

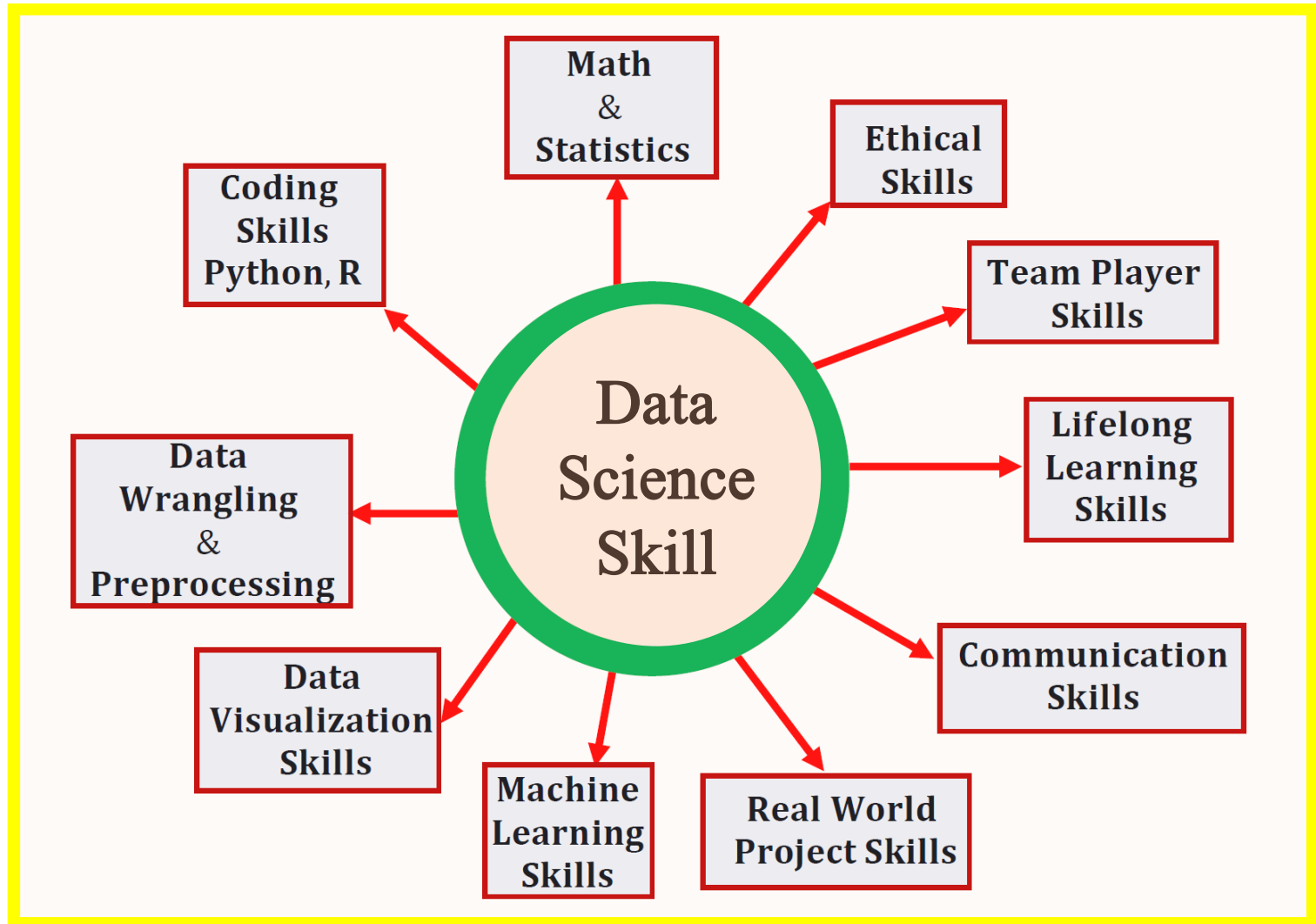


Quy trình của Data Science



	Model	Score
3	Random Forest	86.76
8	Decision Tree	86.76
1	KNN	84.74
0	Support Vector Machines	83.84
2	Logistic Regression	80.36
7	Linear SVC	79.12
6	Stochastic Gradient Decent	78.56

Kỹ năng cần thiết



Nội dung



1. Giới thiệu
2. Quy trình của Data Science
3. Lý do chọn Python



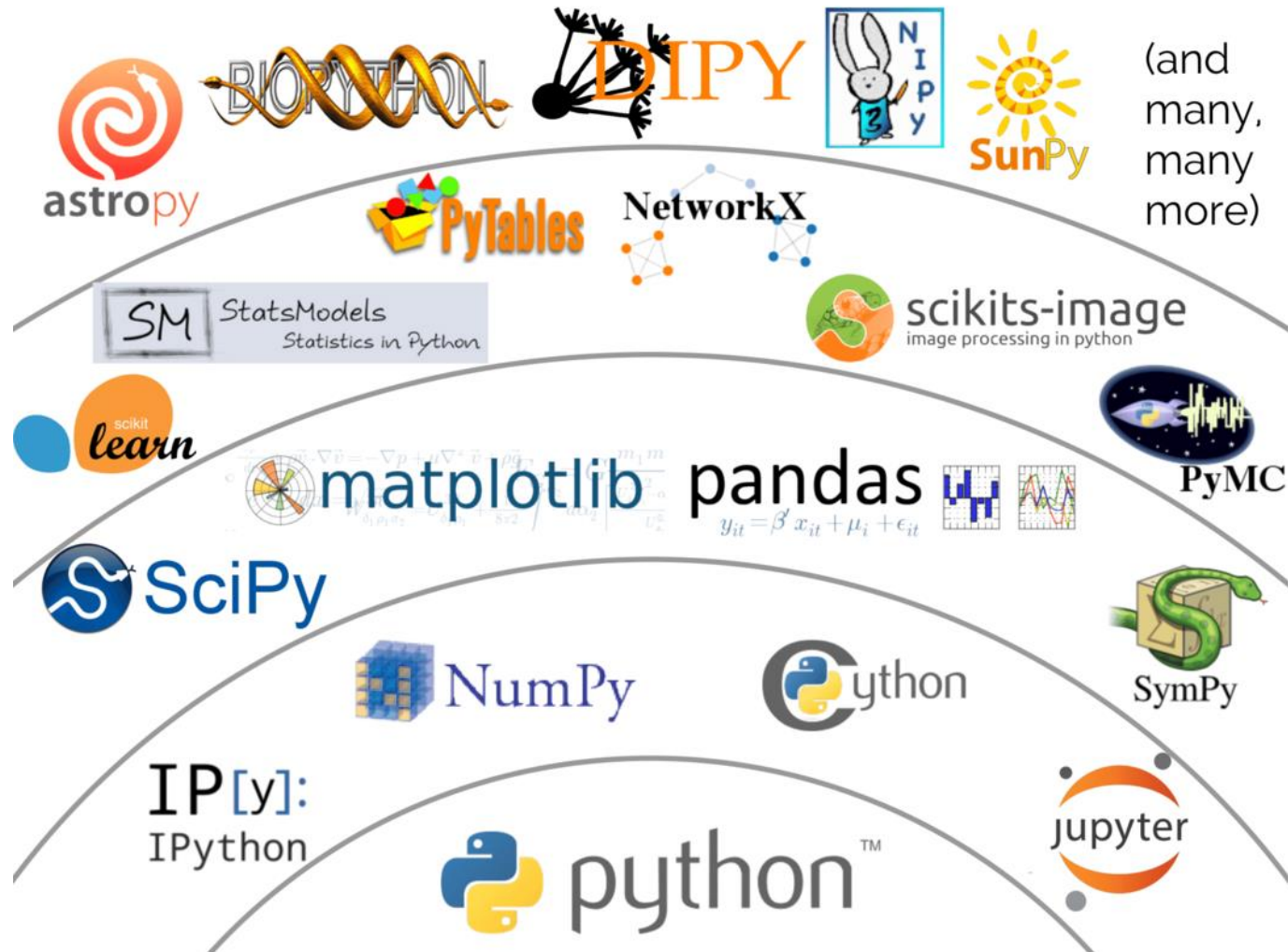
Lý do chọn Python

- ☐ Dễ đọc, dễ học
- ☐ Cộng đồng người dùng lớn
- ☐ Có số lượng thư viện hỗ trợ lớn và luôn luôn phát triển
 - Data management
 - Analytical processing
 - Visualization
- ☐ Có thể sử dụng để triển khai từng bước trong quy trình của data science
- ☐ Notebooks

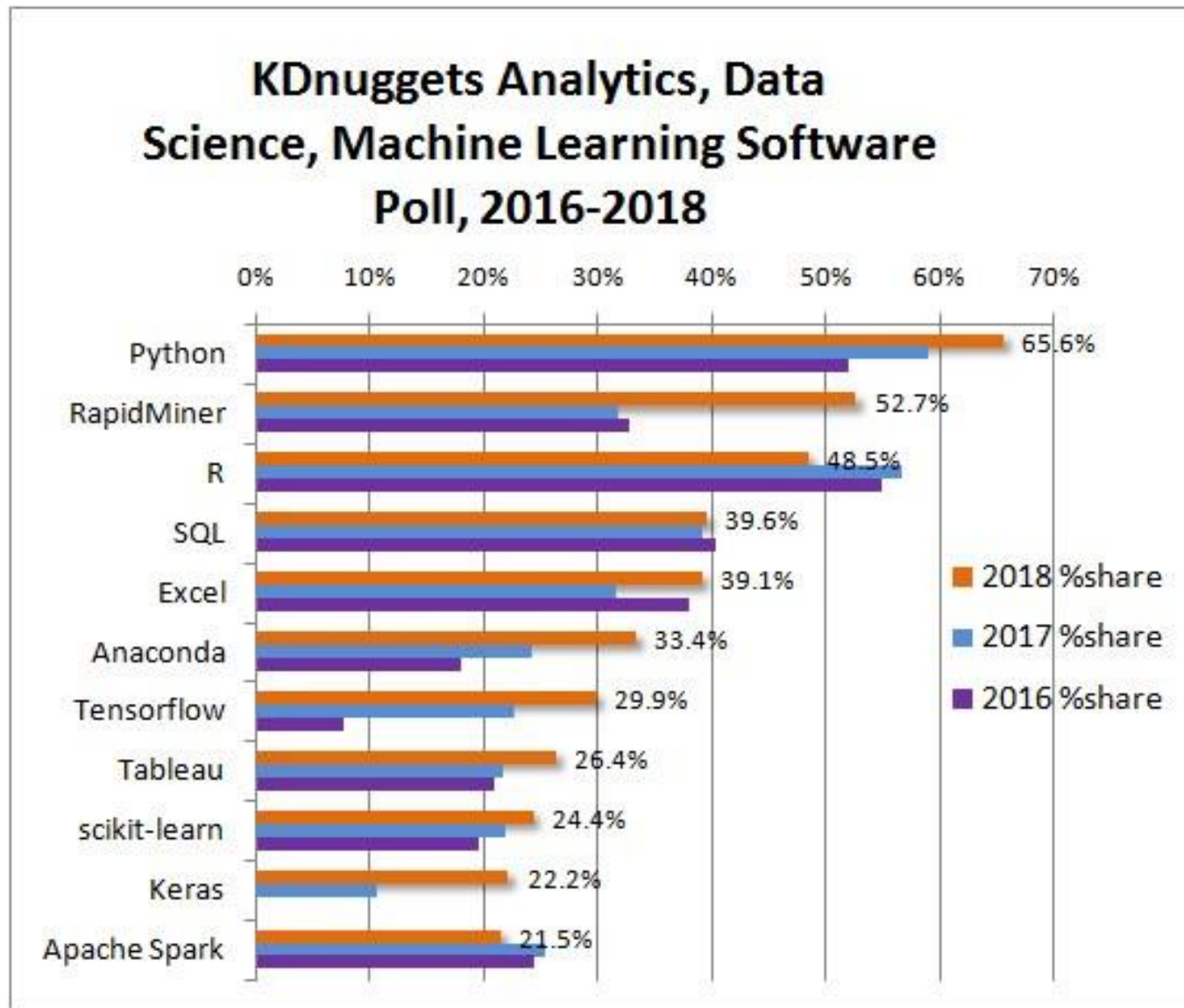
Tại sao là PYTHON



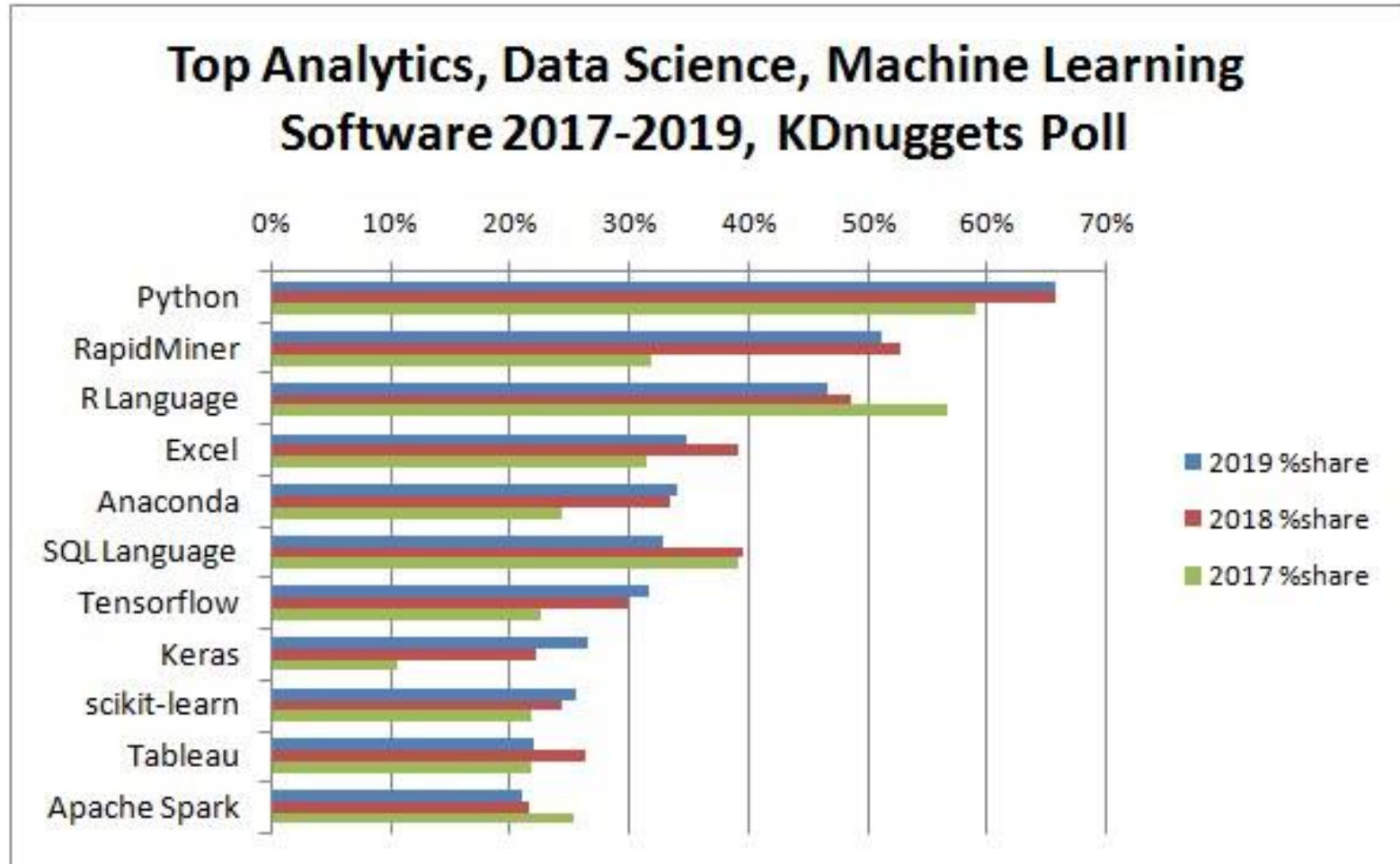
Python – Thư viện hỗ trợ phong phú



Lý do chọn Python

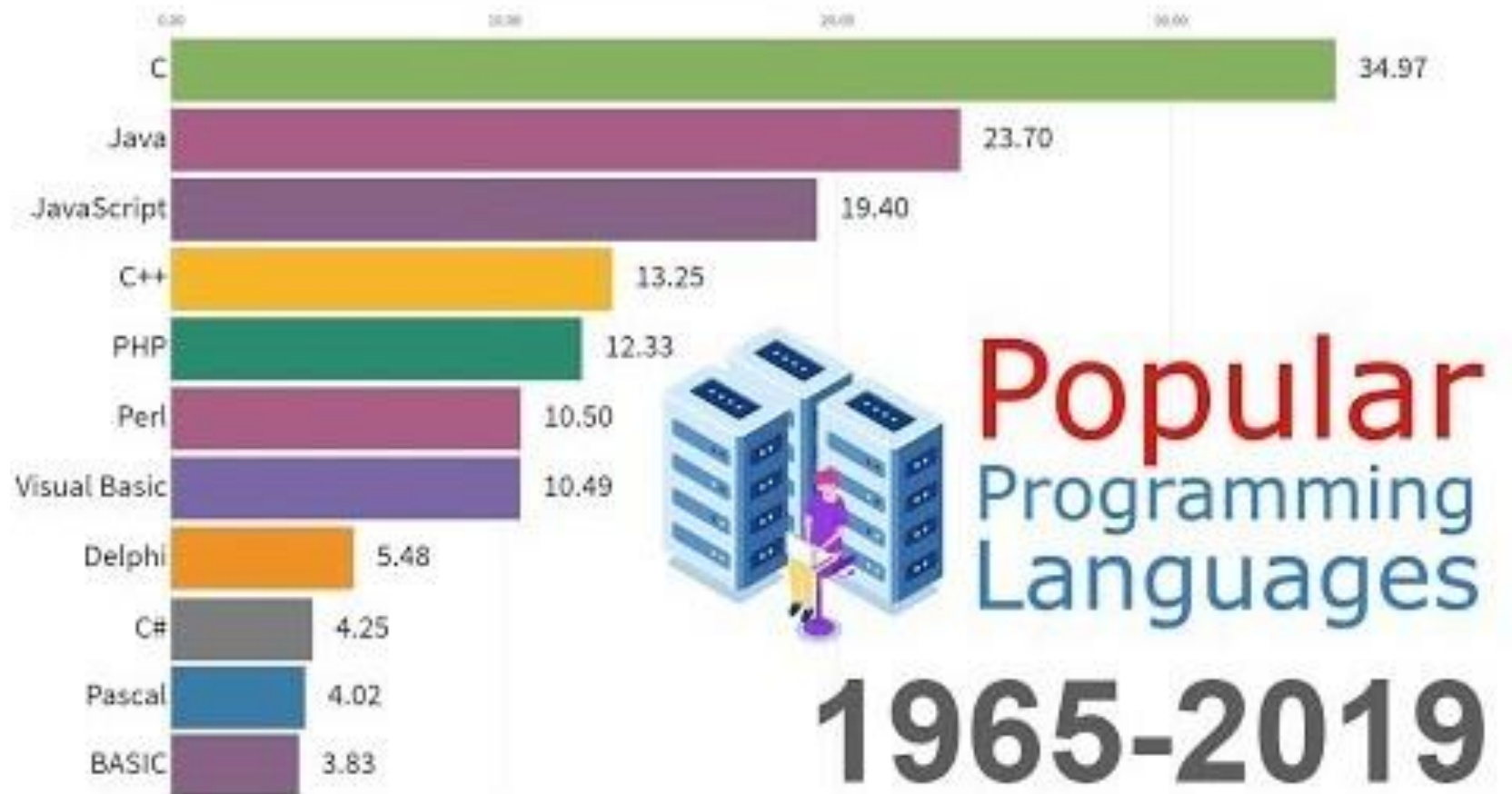


Lý do chọn Python



<https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/2>

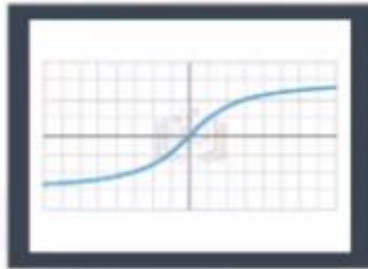
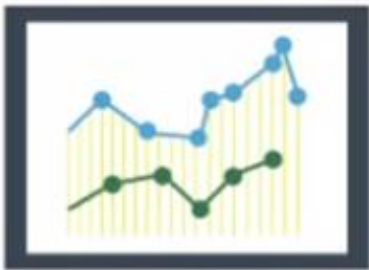
Python – Ngôn ngữ lập trình của tương lai







Pick analytic approach based on type of question



Descriptive

- Current status

Diagnostic (Statistical Analysis)

- What happened?
- Why is this happening?

Predictive (Forecasting)

- What if these trends continue?
- What will happen next?

Prescriptive

- How do we solve it?



What are the types of questions?

If the question is to determine probabilities of an action

- Use a Predictive model

If the question is to show relationships

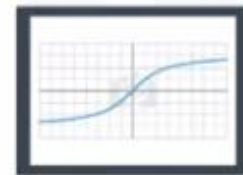
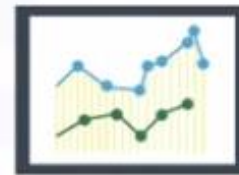
- Use a descriptive model

If the question requires a yes/no answer

- Use a classification model

Analytic approach

- *How can you use data to answer the question?*



- The correct approach depends on business requirements for the model