

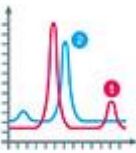
# NHẬP MÔN LẬP TRÌNH KHOA HỌC DỮ LIỆU

---

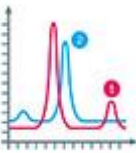
## Bài 1: Giới Thiệu Môn Học

# Nội dung

---

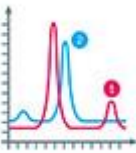


1. Thông tin chung về môn học
2. Data science (khoa học dữ liệu) là gì?
  1. Khoa học dữ liệu khác các khoa học khác ở điểm nào?
  2. Một số vấn đề khoa học dữ liệu xung quanh chúng ta
  3. Nghề làm khoa học dữ liệu có ưu thế gì?
3. Data scientist (nhà khoa học dữ liệu) làm gì?
  1. Data scientist workflow
  2. Data scientist cần gì?



Phần 1

# Thông tin chung về môn học



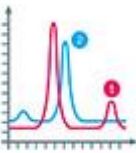
# Giới thiệu môn học

---

- Tên môn: Nhập môn Lập trình Khoa học Dữ liệu (Introduction to Programming for Data Science)
- Số tín chỉ: 3 (24 tiết lý thuyết + 21 tiết bài tập)
- Nội dung chính:
  - Ngôn ngữ python (cơ bản)
  - Một số thư viện xử lý dữ liệu của python
  - Trực quan hóa dữ liệu
  - Học từ dữ liệu như thế nào
- Giảng viên: Trương Xuân Nam, khoa CNTT
- Email: [truongxuannam@gmail.com](mailto:truongxuannam@gmail.com)

# Tài liệu môn học

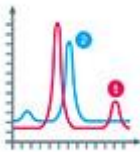
---



- Scipy Lecture Notes, [www.scipy-lectures.org](http://www.scipy-lectures.org)
- Các tài liệu tham khảo nên đọc:
  - “Think Python: How to think like a computer scientist”
  - “Learning Python”
  - “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython”
  - “Python Crash Course, A Hands-On, Project-Based Introduction to Programming”
- Bài giảng, bài tập, mã nguồn, điểm số,... sẽ được đưa lên site <https://txnam.net> mục **BÀI GIẢNG**

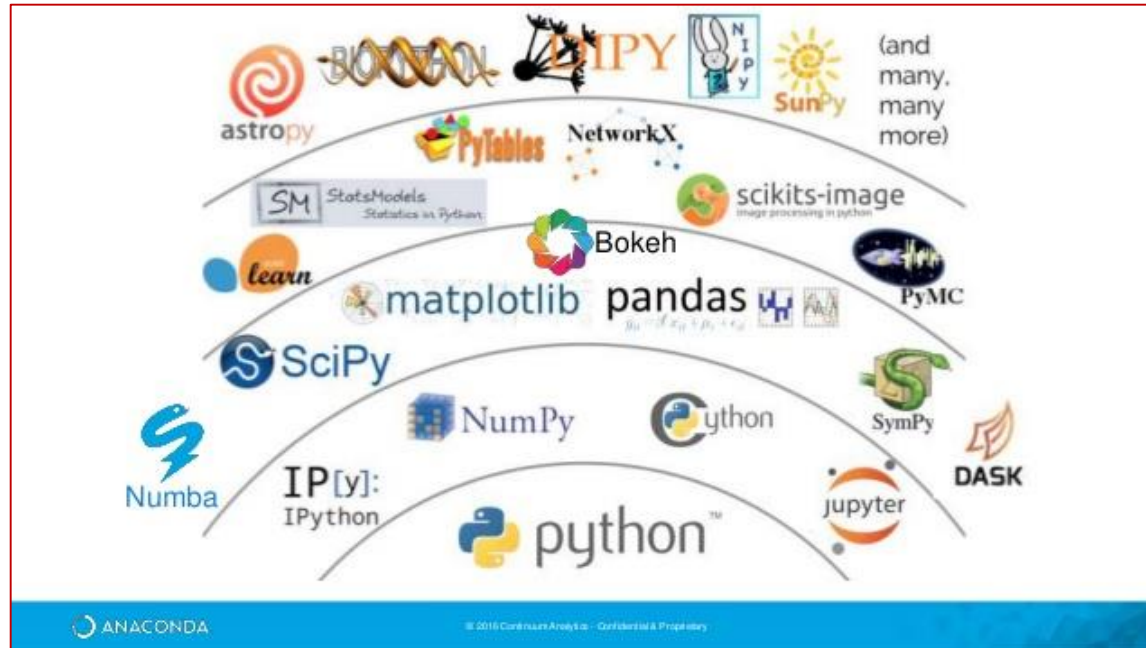
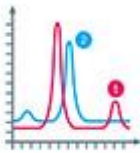
# Kiến thức yêu cầu

---

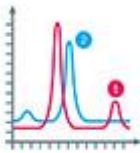


- Đã biết và sử dụng tạm ổn một ngôn ngữ lập trình nào đó (C/C++, C#, Java,...) – vì chúng ta sẽ học khá nhanh phần ngôn ngữ python
- Cấu trúc dữ liệu: mảng, danh sách, cây,... – đặc biệt là mảng nhiều chiều và các phép xử lý trên nó
- Hiểu cách làm việc của hệ thống file, đọc ghi dữ liệu dạng văn bản từ file – hầu hết dữ liệu của môn học và ngành học này đều ở dạng text
- Có kiến thức về các định dạng dữ liệu thường dùng trong cuộc sống (văn bản, ảnh, âm thanh, phim,...)

# Phần mềm học tập



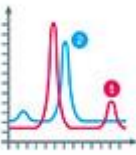
# Đánh giá kết quả



- Điểm môn học = ĐQT x **50%** + ĐTCK x **50%**
- Điểm quá trình:
  - Điểm danh
  - Bài làm trên lớp, trong phòng lab
  - Bài tập về nhà (nộp qua email)
  - Thi giữa kỳ
- Điểm thi cuối kỳ:
  - Thi thực hành trên máy
  - Được sử dụng tài liệu tham khảo
  - Chi thi những gì học, không có giới hạn nội dung thi



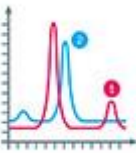




# Tại sao phải học môn này?

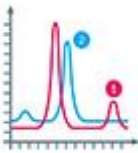
---

- Để có kiến thức về khoa học dữ liệu
- Để có kỹ năng viết chương trình phục vụ cho các bài toán thuộc ngành khoa học dữ liệu
- Để có hiểu biết về công việc của người làm khoa học dữ liệu và các bài toán liên quan
- Để có hiểu biết về cách ứng dụng khoa học dữ liệu vào các vấn đề trong thực tế
- Có thêm lựa chọn cho đề tài làm tốt nghiệp
- Có điểm môn học và được ra trường



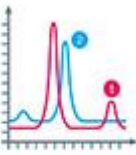
Phần 2

# Data science (khoa học dữ liệu) là gì?



Phần 2.1

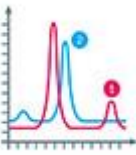
Khoa học dữ liệu khác các  
khoa học khác ở điểm nào?



# Khoa học dữ liệu là gì?

---

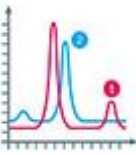
- Hầu hết các ngành khoa học từ xưa đến nay đều giải quyết vấn đề dựa trên **lập luận** và **tri thức**
  - Ngành toán: dựa trên các mệnh đề, công thức, lập luận... để chứng minh bài toán
  - Ngành vật lý: dựa trên các quan sát, thực nghiệm, tính toán,... kiểm chứng các giả thiết
  - Ngành hóa học:...
  - ...
  - Ta gọi các ngành khoa học này là “knowledge-driven” (dẫn dắt bởi tri thức)
- Có ngành có chút ngoại lệ, ví dụ: **ngành xác suất**



# Khoa học dữ liệu là gì?

---

- Với quan điểm như vậy, tất cả những quan sát mà không được chứng minh chặt chẽ thường được cho là “không khoa học”
  - Chẳng hạn: chuồn chuồn bay thấp thì mưa
- Khoa học dữ liệu  $\neq$  Khoa học thông thường ở quan điểm: **tìm tri thức từ dữ liệu** (dẫn dắt bởi dữ liệu – “data-driven”)
  - Chúng ta rút ra tri thức bằng việc tìm tòi từ dữ liệu chứ không nhất thiết phải chứng minh nó
  - Tất nhiên tri thức tìm ra phải có tính ổn định (luôn có cùng kết quả nếu sử dụng cùng một phương pháp)

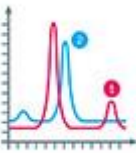


Phần 2.2

# Một số vấn đề khoa học dữ liệu xung quanh chúng ta

# Vấn đề quanh ta

---



## ■ Các bài toán dự báo:

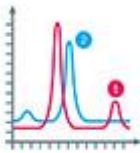
- Dự báo thị trường nhà đất: ngôi nhà ở mảnh đất A liệu có giá bao nhiêu vào năm 2020?
- Dự báo thời tiết: đi nghỉ giỗ tổ và 30/4-1/5 ở Hạ Long có cần mang áo mưa hay không?
- Dự báo hành vi mua hàng: có thích món hàng này hay không? Mức độ thích như thế nào?
- ...

## ■ Các bài toán ra quyết định:

- Lái xe tự động
- Đặt mua, đặt bán cổ phiếu theo tin tức

# Vấn đề quanh ta

---

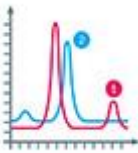


- Các bài toán ra quyết định:
  - Điều chỉnh nhiệt độ điều hòa tối ưu cho hoạt động của người trong phòng
  - Điều hành xe để đáp ứng nhu cầu của khách gọi taxi
  - ...
- Các hệ thống phân tích thời gian thực:
  - Xu hướng của truyền thông về doanh nghiệp hoặc nhân vật nào đó
  - Cảnh báo cháy qua camera
  - Cảnh báo nguy hiểm với trẻ con, người già
  - ...

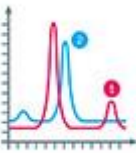


# Thảo luận

---



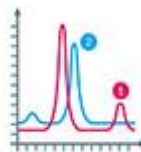
1. Hãy nêu một vài vấn đề liên quan đến địa phương (quê) của bạn, mà bạn cho rằng có thể giải quyết bằng khoa học dữ liệu.
2. Theo bạn có những vấn đề nào của trường ta có thể là đối tượng nghiên cứu của khoa học dữ liệu?
3. Gần đây Facebook có vụ bê bối vì lộ thông tin cá nhân của khách hàng, bạn có cho rằng các thông tin mà bạn đưa lên facebook là quan trọng?
4. (vui) Đánh số đề có phải là bài toán của ngành khoa học dữ liệu?



Phần 2.3

# Nghề làm khoa học dữ liệu có ưu thế gì?

# Nghề hấp dẫn của thế kỉ 21!



**Harvard Business Review**

THE MAGAZINE | BLOGS | VIDEO | BOOKS | CASES | WEBINARS | COURSES

Guest | Subscribe today and get access to all current articles and HBR online archive.

**THE MAGAZINE**  
October 2012

ARTICLE PREVIEW To read the full article, [sign-in or register](#). HBR subscribers, click [here to register for FREE access](#).

**Data Scientist: The Sexiest Job of the 21st Century**

by Thomas H. Davenport and D.J. Patil

**Forbes**

CEO Keep this: The World's Highest Paid Top Execs

Strategic Moves: An Top-Six Tech Exec Is The World's Most Valuable CEO

EMC<sup>2</sup> The Hottest Jobs In IT: Training Tomorrow's Data Scientists

**25 CNBC** | Enter Symbols GO | Enter Keywords GO

HOME U.S. | NEWS | MARKETS | INVESTING | TECH | SMALL BIZ | VIDEO | SHOWS | PRIM

NEW SHOW **SQUAWK** | The Intersection of Wall St. & Tech

**BIG DATA** | A CNBC SPECIAL REPORT

## Why your kids will want to be data scientists

John Phillips | @J\_Phillips\_IV  
Tuesday, 3 Jun 2014 | 7:05 PM ET



**TechRepublic / U.S.** | All Topics | Newsletters | Photos | Forums | Resource Library | Rese

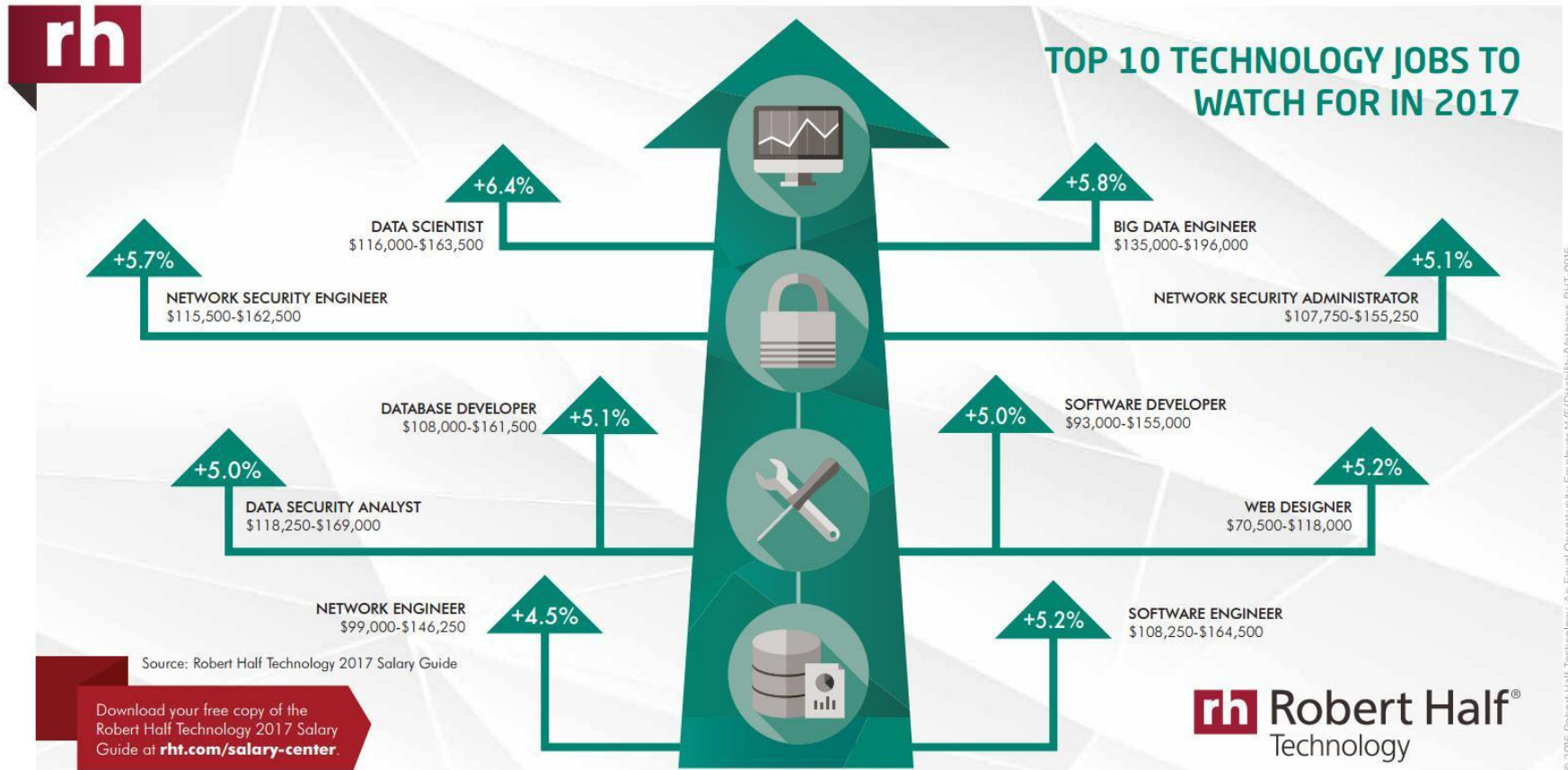
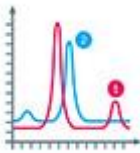
CXO | Software | Startups | Cloud | Data Center | Mobile | Microsoft | Apple | Google | Search TechRepu

**SAP** Is your business making the most out of today's technologies?

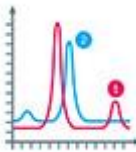
BIG DATA

## Big data skills: Should data scientist be your next job?

# Nhu cầu tăng cao



# Nhu cầu tăng cao...



## WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A \$37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS

BY NEAL UNGERLEIDER

It's official: America needs more data scientists. This week, a \$37.8 million project

Berkeley Research  
UNIVERSITY OF CALIFORNIA

CONTACT US | HOME

RESEARCH HIGHLIGHTS NEWS ABOUT US RESEARCH LIMITS FACILITY CAPABILITY RESEARCH POLICIES & ADMINISTRATION TECH TRENDS FIND YOUR RESEARCH

HOME > DATA SCIENCE

### Data Science

DATA SCIENCE  
DISCUSSION  
INSTITUTE FOR DATA SCIENCE  
Topics & Releases  
Press Events  
PEOPLE  
CAREER OPPORTUNITIES  
2013-14 LECTURE SERIES  
CAMPUS EVENTS  
Articles  
NEWS  
INSTITUTES AND PROGRAMS



Data Science at UC Ber

SCIENTIFIC AMERICAN™

Sign In | Register

Search Science@ScientificAmerican.com

Subscribe News & Features Topics Blogs Videos & Podcasts Education

More Science » Scientific American Volume 309, Issue 2



### How Big Data Can Transform Society for the Better

The digital traces we leave behind each day reveal more about us than we know. This could become a privacy nightmare—as it could be the foundation of a healthier, more prosperous world.

By Neal Ungerleider



### RESEARCH CENTERS IN THE FIELD OF DATA SCIENCE

#### Center for Data Science (CDS)

The NYU Center for Data Science (CDS) is a focal point for New York University's university-wide initiative in data science. It was established to help advance NYU's goal of creating the country's leading data science training and research facilities, among researchers and professionals with tools to harness the power of big data.

LEARN MORE

#### Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM)

The Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM) is a new center dedicated to improving the caliber of research in quantitative social, educational, behavioral, allied health and policy science.

# 500k

The world's 500,000+ data centers are large enough to fill 5,000 football fields. (Source: Statista)

# 75%

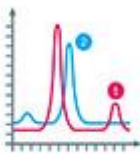
75% of digital information is generated by individuals, without enterprise-level budgets for 80% of digital data at some point in its life. (Source: Statista)



## New Ph.D. Tracks in "Big Data"



# Cầu vượt cung



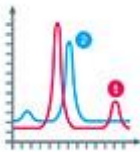
Over 2/3 believe demand for talent will outpace the supply of data scientists

**OVER THE NEXT FIVE YEARS, DEMAND FOR DATA SCIENTISTS WILL:**



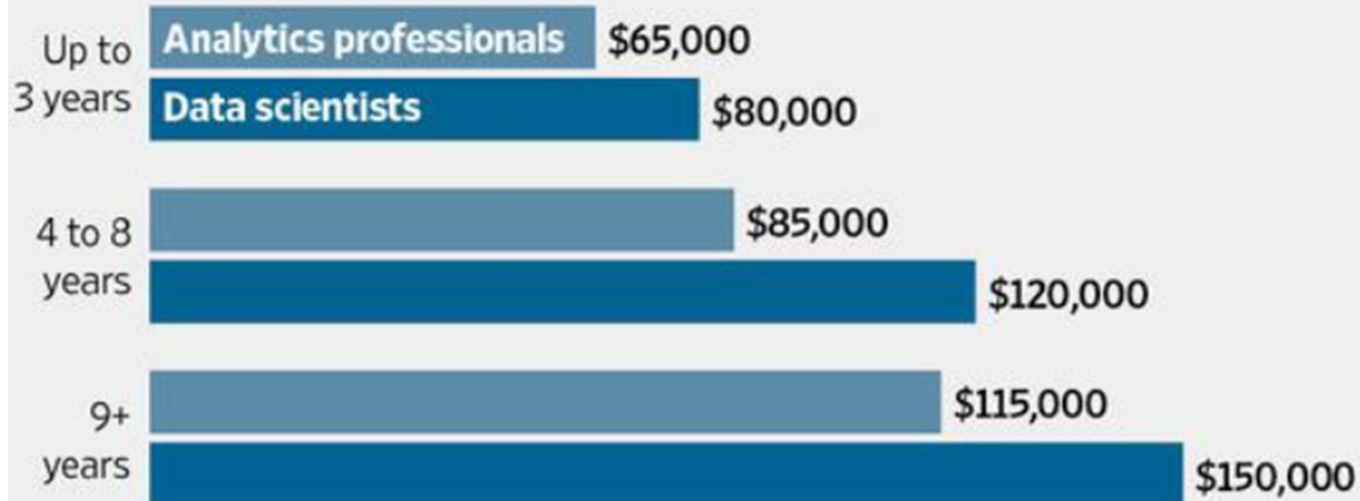
Only 12% see today's BI professional as the best source for new data scientists

# Lương cao



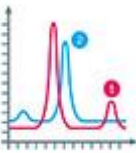
## Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.



Note: Data do not include managers Source: Burtch Works

The Wall Street Journal

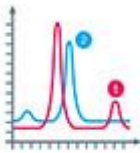


Phần 3

# Data scientist (nhà khoa học dữ liệu) làm gì?




# Data scientist làm gì?



- Với skillset chuyên sâu và trải dài trên nhiều lĩnh vực
  - Math and Statistics
  - Programming and Database
  - Communication and Visualization
  - Domain Knowledge and Soft Skills

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



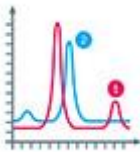
- MATH & STATISTICS**
  - ☆ Machine learning
  - ☆ Statistical modeling
  - ☆ Experiment design
  - ☆ Bayesian inference
  - ☆ Supervised learning: decision trees, random forests, logistic regression
  - ☆ Unsupervised learning: clustering, dimensionality reduction
  - ☆ Optimization: gradient descent and variants
- PROGRAMMING & DATABASE**
  - ☆ Computer science fundamentals
  - ☆ Scripting language e.g. Python
  - ☆ Statistical computing package e.g. R
  - ☆ Databases SQL and NoSQL
  - ☆ Relational algebra
  - ☆ Parallel databases and parallel query processing
  - ☆ MapReduce concepts
  - ☆ Hadoop and Hive/Pig
  - ☆ Custom reducers
  - ☆ Experience with xaaS like AWS
- DOMAIN KNOWLEDGE & SOFT SKILLS**
  - ☆ Passionate about the business
  - ☆ Curious about data
  - ☆ Influence without authority
  - ☆ Hacker mindset
  - ☆ Problem solver
  - ☆ Strategic, proactive, creative, innovative and collaborative
- COMMUNICATION & VISUALIZATION**
  - ☆ Able to engage with senior management
  - ☆ Story telling skills
  - ☆ Translate data-driven insights into decisions and actions
  - ☆ Visual art design
  - ☆ R packages like ggplot or lattice
  - ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing  
DISTILLERY

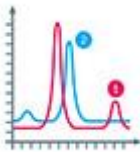
# Data scientist làm gì?

---



- Thu thập và xử lý dữ liệu để tìm ra những “insight” (giá trị bên trong)
  - Ví dụ: dựa trên các thông tin thu thập được từ các post/comment/status trên mạng xã hội, Data Scientist có thể tìm ra được: cứ gần đến ngày valentine thì tần suất xuất hiện các thương hiệu ABC cao hơn hẳn
- Giải thích, trình bày những insight đó cho các bên liên quan, để chuyển hóa insight thành hành động
  - Ví dụ: khi tìm ra được insight giá trị từ data, bạn cần làm report/presentation hay visualization để biểu diễn, giải thích cho các bên liên quan hiểu được

# Data analyst và Data scientist



## Data Scientist

also known as Data Managers, statisticians.



A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

**Skills:** Mathematics, Programming, Communication



*Will use programmes such as:*  
SQL, Python, R

## Data Engineers

also known as database administrators and data architects.



They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

**Skills:** Programming, Mathematics, Big data



*Will use programmes such as:*  
Hadoop, NoSQL, and Python

## Data Analysts

also known as business Analysts.



They typically help people from across the company understand specific queries with charts.

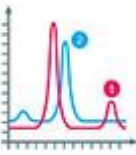
**Skills:** Statistics, Communication, Business knowledge



*Will use programmes such as:*  
Excel, Tableau, SQL

# Sản phẩm data là gì?

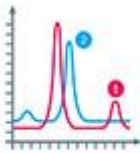
---



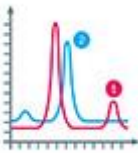
- Sản phẩm data được xây dựng dựa trên dữ liệu
  - Tính năng recommendation của Amazon được xây dựng dựa trên dữ liệu của nó: người dùng muốn mua món đồ gì? Những món đồ nào nên mua kèm?
- Sản phẩm data có thể là một sản phẩm riêng biệt hoặc một phần trong sản phẩm lớn
  - Facebook có thể tự tag ảnh bạn bè của bạn
- Sản phẩm data bao gồm nhiều thành phần nhưng mô hình dữ liệu là cốt lõi của nó và được xây dựng bằng các thuật toán học máy

# Mô hình dữ liệu là gì?

---



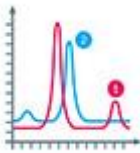
- **Ví dụ:** bạn muốn dùng một chiếc hộp đen để nhận diện loài vật
  - B1: Bạn phải tìm rất nhiều hình ảnh con chó và con mèo
  - B2: Cho hộp đen đọc những hình ảnh này
  - B3: Dạy cho hộp đen biết đặc điểm nào trên bức hình là của con chó, đặc điểm nào là của con mèo
  - B4: Bạn đưa ra 2 hình ảnh mới, hộp đen sẽ trả lời đâu là hình ảnh con chó, hình ảnh con mèo
- Toàn bộ quá trình này gọi là học máy (machine learning) và cái hộp đen chính là mô hình dữ liệu



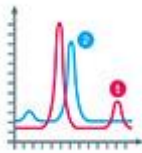
Phần 3.1

# Data scientist workflow

# Data scientist workflow



# Data scientist workflow – Bước 1

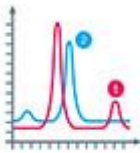


## ■ Input

- Workflow bắt đầu từ một yêu cầu hoặc nhiệm vụ: “Nhu cầu tìm kiếm hình ảnh của Google: đưa cho máy 1 bức ảnh, trả về những bức ảnh tương tự”
- Nhu cầu này có thể bắt nguồn từ:
  - Do bộ phận business thu thập phản hồi từ người dùng và đề nghị có thêm tính năng ABC
  - Hoặc, do chính Data Scientist khi làm việc với dữ liệu, nghiên cứu đặc tính của sản phẩm/ công ty cũng như kiểu/ lượng data hiện có... thì nảy sinh thêm sáng kiến phát minh tính năng XYZ



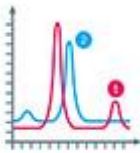
# Data scientist workflow – Bước 2



## ■ Lên kế hoạch

- Làm tính năng này có khả thi hay ko?
- Sẽ cần loại dữ liệu gì? Ở đâu? Bao nhiêu là đủ? Lấy dữ liệu như thế nào?
- Cần bao nhiêu resource (nhân lực, thời gian)
- Tính năng này sẽ được gắn vào đâu trong sản phẩm cuối cùng và sẽ giúp ích được gì cho người dùng

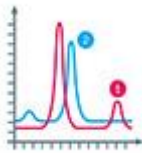
# Data scientist workflow – Bước 3



## ■ Thu thập và làm sạch dữ liệu

- Để dạy cho máy biết phân biệt chó/mèo, nó càng phải học nhiều hình ảnh càng tốt. Nên phải đi “gom dữ liệu”
- Dữ liệu gom xong sẽ còn lộn xộn và nhiều rác thì phải “làm sạch dữ liệu”.
  - Hình ảnh ko cần thì loại bỏ; Hình mờ thì làm cho rõ ...
  - Đồng bộ hóa dữ liệu
  - Hình ảnh mang về có kích thước khác nhau, phải đưa hết về cùng kích thước, định dạng theo mô hình dữ liệu đã chọn
- Nếu dữ liệu chưa đủ phải thu thập thêm

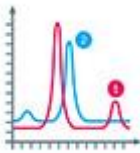
# Data scientist workflow – Bước 4



## ■ Chọn giải pháp

- Nếu vấn đề đã có sẵn giải pháp
  - Lựa chọn / kết hợp các giải pháp lại, chạy thử nghiệm, kiểm tra thử nghiệm nào tốt nhất và vì sao, chọn giải pháp để phát triển thêm
- Nếu vấn đề chưa có sẵn giải pháp
  - Cần làm nghiên cứu: tìm hiểu xem trước mình đã có ai từng làm về vấn đề này hay chưa
  - Sau đó, chọn ra một hoặc một loạt các phương pháp để thử nghiệm

# Data scientist workflow – Bước 5

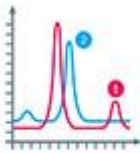


## ■ Máy học

- Chạy thử mô hình và đánh giá hiệu năng
  - Tưởng tượng bạn điều khiển bảng điều khiển với nhiều nút. Bạn thử chỉnh nút này 1 chút, thấy kết quả ra tốt hơn chút xíu thì giữ lại và chỉnh thử nút khác
- Nhận diện các yếu tố ảnh hưởng đến kết quả. Điều chỉnh dấu hiệu ưu tiên để ra được kết quả tốt nhất.

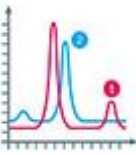
# Data scientist workflow – Bước 6

---



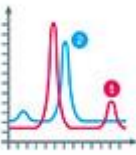
## ■ Output

- Kết quả gắn vào một sản phẩm lớn có tính ứng dụng
- Viết bài báo
- Tổ chức hội thảo





Phần 3.2

# Data scientist cần gì?

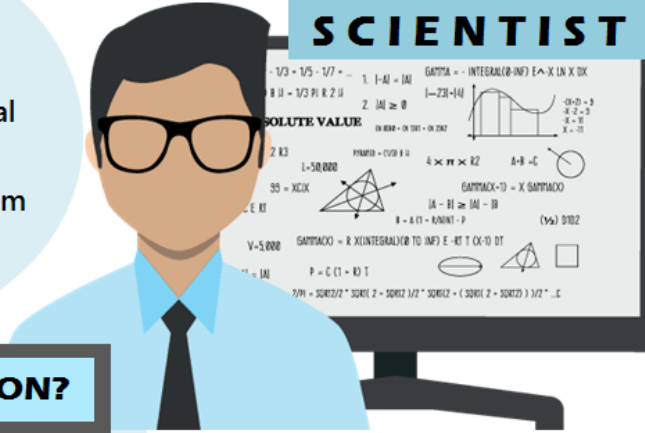


# Data scientist: tố chất cần có?

D A T A

SCIENTIST



**WHO AM I?**

I am a part analyst & part artist. I use my analytical and technical abilities to extract meaning / insights from massive data sets.

**WHAT DO I DO?**

1. I cleanse existing raw data & build models to predict future data.
2. I go beyond merely collecting and reporting data, to look at data from multiple angles & give meaning to it.
3. I identify the correct business problem(s) & offer solutions (via visualizations, reports or blogs) by best applying the data.

**WHAT DO I RELY ON?**

1. Analytics
2. Predictive Models
3. Statistical Analysis & Modeling
4. Data Mining
5. Sentiment Analysis
6. What-if Analysis

**THE PROCESS I FOLLOW**

Define Problem

Structure Data

Use Programming Language

**WHAT DO I EARN?**

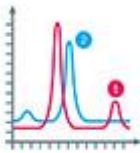
After oil & gas geologists, mine is the 2<sup>nd</sup> highest paid job in the world!

\$
100,000  
to  
150,000

**HOW DO I HELP ORGANIZATIONS TODAY?**

- Increase data accuracy
- Develop strategies
- Improve operational efficiency
- Reduce costs
- Mitigate risks
- Offer personalized products/services

# Data scientist: tổ chất cần có?



## ■ Kiên nhẫn

- Tổ chất này cực kì quan trọng vì DS phải dành phần lớn thời gian để thu thập và làm sạch dữ liệu

*Ví dụ, bạn muốn làm một model dự đoán giá nhà.*

*Bạn sẽ phải thu thập dữ liệu về nhà từ nhiều nguồn khác nhau.*

*Mỗi nguồn này lại lưu dữ liệu theo một cấu trúc riêng. Vậy thì bạn phải quy chúng về một cấu trúc chung.*

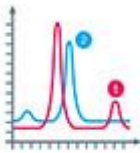
*Sau đó, bạn làm sạch bằng cách loại bỏ các dữ liệu không phù hợp, như:*

- *Dữ liệu thiếu: có số lượng phòng mà không có diện tích.*
- *Dữ liệu rác: diện tích 10m<sup>2</sup> mà giá 200 tỷ.*



# Data scientist: tổ chất cần có?

---

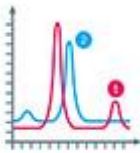


## ■ Giao tiếp tốt

- Với Team Business: để hiểu rõ hơn về sản phẩm cũng như requirements, từ đó tìm ra các insights có giá trị
- Với Team Engineer: để áp dụng mô hình của mình vào hệ thống hoặc đề nghị họ tổ chức/hệ thống data cho mình sử dụng
- Trình bày, giải thích insights cho các bên liên quan hiểu

# Data scientist: tố chất cần có?

---

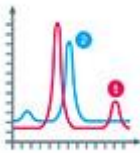


## ■ Thích tìm hiểu và thử cái mới

- Nghề Data Scientist còn mới mẻ và sử dụng nhiều kiến thức liên ngành.
  - Mỗi ngành riêng lại có bước tiến và công nghệ mới: Bạn cần cập nhật kiến thức liên tục

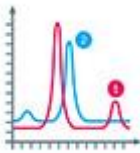
# Data scientist cần kỹ năng gì?

---



- Nghề Data Scientist đòi hỏi khá nhiều kiến thức và kỹ năng tổng hợp
  - **Machine Learning**: để học từ dữ liệu, từ đó tạo ra các mô hình dự đoán
  - **Database**: giúp lưu trữ, truy xuất dữ liệu cũng như thực hiện tính toán
  - **Programming language**: viết code để áp dụng các mô hình đã học được nói trên vào sản phẩm cụ thể hoặc để thao tác với database
  - **Visualization**: giúp hiểu hơn về dữ liệu hoặc trình bày kết quả phân tích

# Data scientist cần kỹ năng gì?



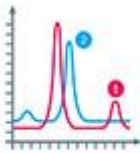
## ■ Kiến thức toán học: yếu tố quan trọng số 1

- Nghề data science sử dụng nhiều kiến thức liên ngành.
  - Machine learning là sự kết hợp của các mô hình toán học chạy bên dưới
  - Khi xử lý / làm việc với dữ liệu, bạn sẽ cần sử dụng rất nhiều kiến thức về toán, xác suất thống kê,...
  - Tư duy toán học sẽ giúp bạn dễ tiếp thu và học các kỹ năng khác nhau

*Ví dụ, khi cho máy học một bức ảnh để phân biệt con chó với con mèo. Thì bức ảnh đó sẽ được chia làm nhiều vùng tương ứng với 100 ô vuông chẳng hạn.*

*Rồi bạn dạy cho cái máy rằng, trong bức ảnh, ô ở cụm phía bên góc trái có nhiều màu đen, kết hợp với ô ở cụm phía bên góc phải có nhiều màu trắng, thì đó là đặc điểm nhận biết con chó.*

# Data Scientist: Các kỹ năng cần thiết?



## ■ Khả năng Lập trình phần mềm

- Công việc của Data Scientist rất gần với Software Engineer. Vì vậy, code cứng là một yêu cầu quan trọng

## ■ Sự nhạy bén

- Khi nhìn vào dữ liệu, bạn cần đủ nhạy để suy đoán: đối với loại dữ liệu này thì nên làm gì với nó, nên estimate như thế nào?
- Sự nhạy bén là tố chất song cũng tích lũy dần theo kinh nghiệm và thời gian