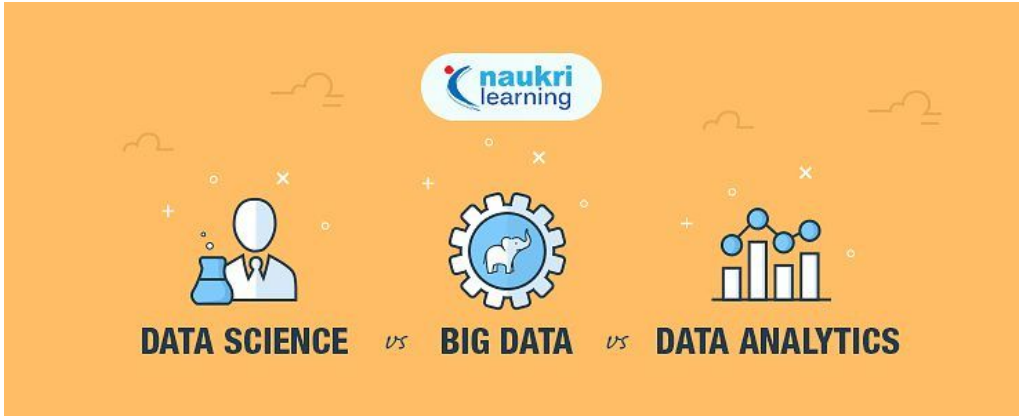
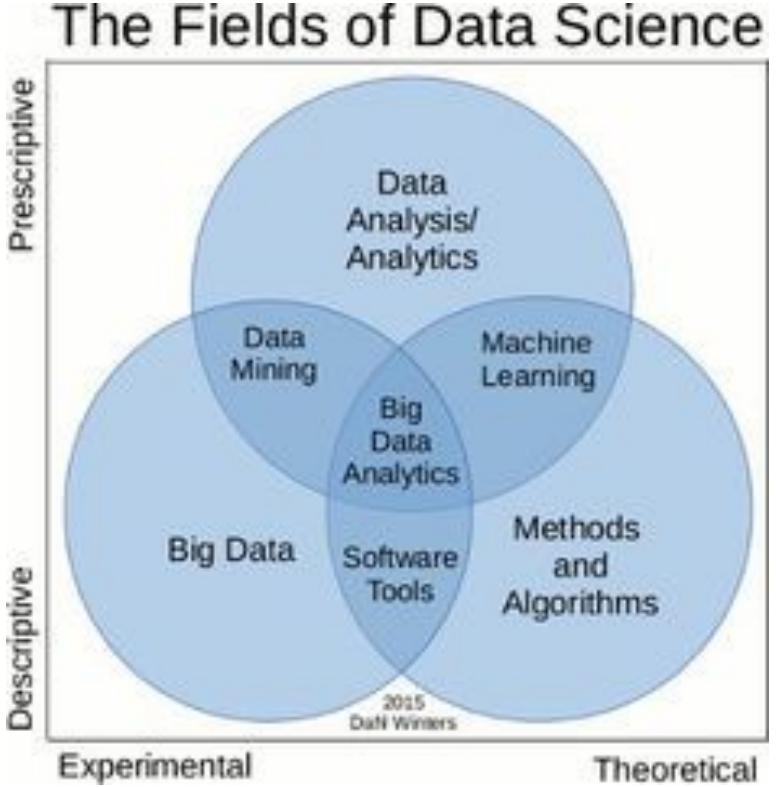
Tóm tắt

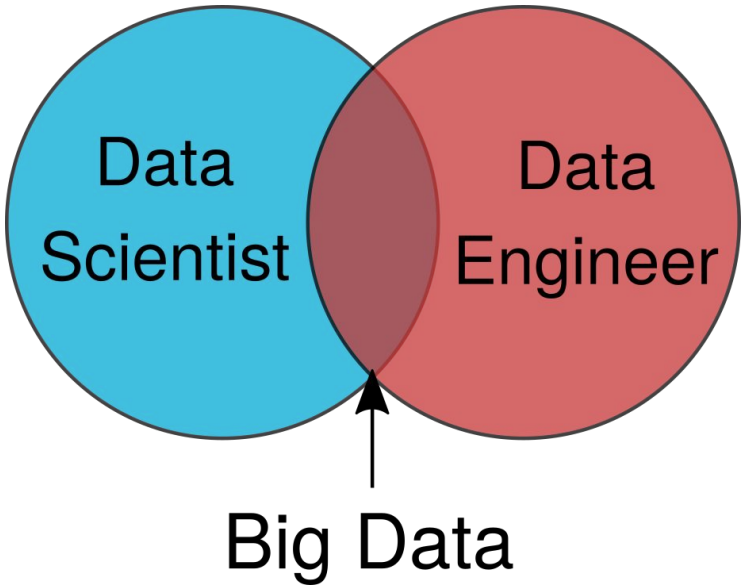
1. DLL là những tập rất lớn vượt quá khả năng lưu trữ, xử lý, tính toán của những công nghệ truyền thống.
2. Dữ liệu lớn đến từ nhiều nguồn khác nhau và không ngừng biến đổi.
3. Đặc trưng chính: Volume (Kích thước), Velocity (Tốc độ), Variety (Đa dạng).
4. DLL đem lại nhiều cơ hội: quốc gia, doanh nghiệp, cá nhân, khoa học.
5. Nhưng DLL có không ít thách thức

Big Data, Data Science, Data Analytics





Data Engineer and Data Scientist



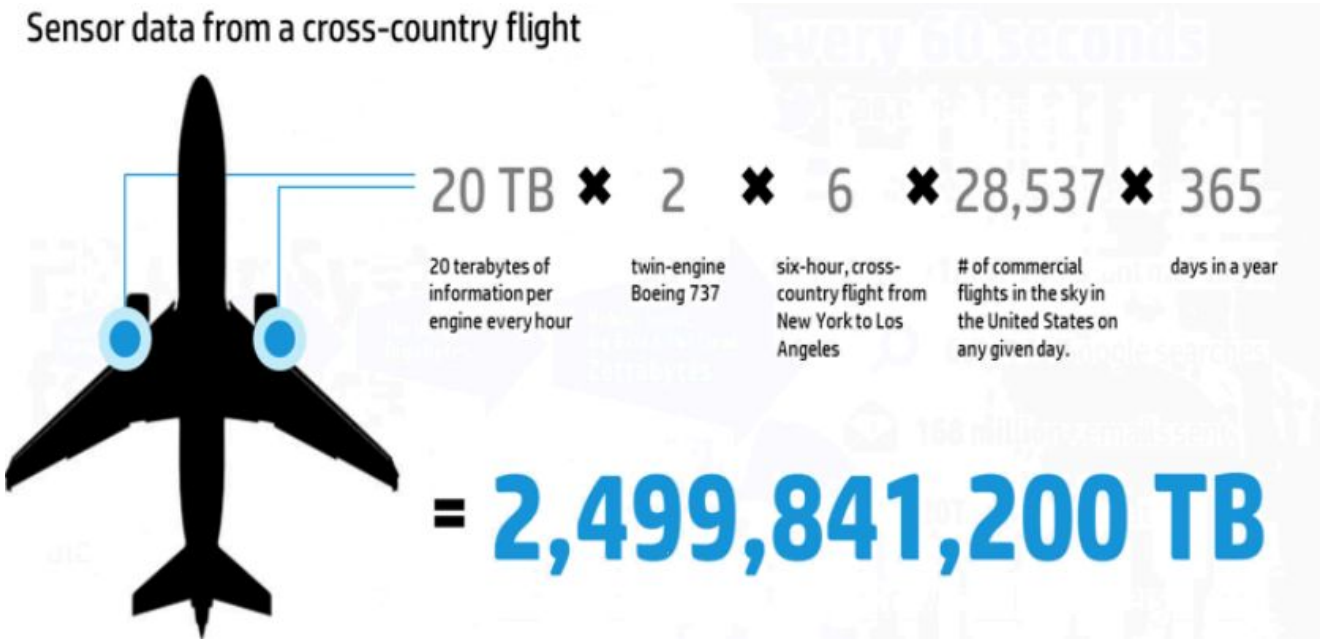
Đặc điểm DLL - Volume

- **Mỗi phút**
  - **Email**: 204 million emails
  - **FB**: 1.8 million likes; 200,000 photo
  - **Youtube**: 1.3 million video views; 72 hours video upload

Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Volume

More data = Better safety



Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Volume

- **Khó khăn là gì?**
  - Lưu trữ (Storage)
  - Truy cập và xử lý (Access, Processing)

Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Velocity

Velocity == Speed

- Tốc độ sinh dữ liệu
- Tốc độ lưu trữ
- Tốc độ phân tích, tính toán

Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Velocity

- Big Data → **Real time Action**.
- Late Decision → Missing Opportunities.

Đặc điểm DLL - Velocity

Ví dụ: dự báo thời tiết hôm nay

- Dùng dữ liệu, thông tin của năm trước → **“nắng, không mưa"**
- Dùng dữ liệu của tháng rồi → **“nắng, không mưa"**
- Dùng dữ liệu của tuần rồi và trạng thái diễn biến của ngày hôm nay → **“Mưa lớn, Bão”**

Đặc điểm DLL - Velocity

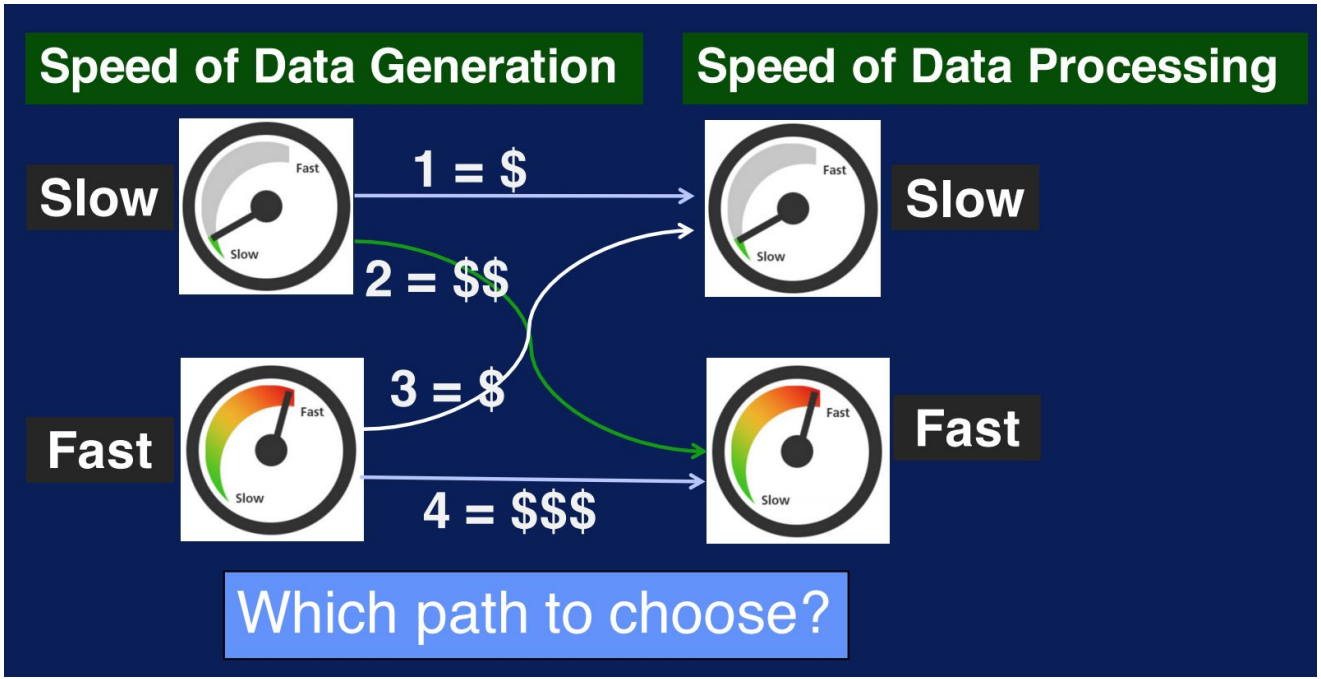
- Xử lý theo lô, đợt (**Batch Processing**)

Thu thập DL → Dọn dẹp DL → Chia khối, Xử lý → Đợi → Hành động.

- Xử lý theo thời gian thực (**Real time Processing**)

Ghi nhận DL tức thời → Lưu trữ thời gian thực về các máy → Xử lý thời gian thực → Hành động

Đặc điểm DLL - Velocity



Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Velocity

- **Khó khăn**
  - Lưu trữ, xử lý, tính toán: **thời gian thực**

Đặc điểm DLL -

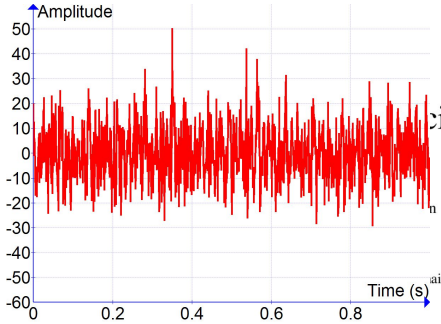
~~Variety~~

- **Variety == Complexity**
- **Trước đây**, dữ liệu thông thường chỉ giới hạn trong các **tables**.

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Đặc điểm DLL - Variety

- Ngày nay dữ liệu không đồng nhất xuất hiện nhiều hơn: phức tạp về cấu trúc, định dạng, ngữ nghĩa.



cientific Publication Recommendations Based on Collaborative Citation Networks

n Huynh, Kiem Hoang, Loc Do, Huong Tran
Department of Computer Science
University of Information Technology
Ho Chi Minh City, VietNam
ail: {tinhn, kiemhv, locdo, huongtran}@uit.edu.vn

Hiep Luong, Susan Gauch
Department of Computer Science & Computer Engineering
University of Arkansas
Fayetteville, Arkansas, USA
Email: {hluong, sgauch}@uark.edu

***Abstract*—To learn about the state of the art for a research project, researchers must conduct a literature survey by searching for, collecting, and reading related scientific articles. Popular search systems, online digital libraries, and Web of Science (WoS) sources such as IEEE Explorer, ACM, SpringerLink, and Google Scholar typically return results or articles that are similar to keywords in the user's query. Some digital libraries also include content-based recommenders that suggest papers similar to one the user likes based on the contents of paper, i.e., the keywords it contains. In this work, we present a recommender module that suggests papers to users based on the seed paper's Citation Network. This work takes into account the combination of the *co-citation* and *co-reference* factors to improve algorithm's effectiveness. We applied and improved the the CCIDF (Common Citation Inverse Document Frequency) algorithm used by the CiteSeer digital library. This improved algorithm, called CCIDF+, was evaluated using data collected from Microsoft Academic Search (MAS). Experimental results show that CCIDF+ outperforms CCIDF.**

(2) collaborative; and (3) hybrid recommendation systems. Content-based recommender systems use features extracted from the actual items to suggest new items to the user whereas collaborative recommenders suggest items based on the user's past preferences compared with other users' ratings, assuming that like-minded people tend to have similar choices [1]. As the name suggests, hybrid recommenders use a combination of content-based and collaborative techniques.

Recommendation systems are a widely used, popular tool for e-business. They can help customers to find and select a suitable product. However, their potential impact on scientific research is largely unexplored. Thus, our goal is to develop and evaluate a recommender system for academic research papers to allow experience researchers to do their research efficiently. In this paper we present the results of the improvement of CCIDF algorithm used in CiteSeer system and we also build a module that can recommend relevant articles. CiteSeer uses

Đặc điểm DLL - Variety

- Người nhận, người gởi, ngày giờ: cấu trúc
- Nội dung: phi cấu trúc
- Đính kèm: multi-media
- Ai gởi ai: Network
- Nội dung email này có liên quan email nào khác trong quá khứ: ngữ nghĩa

Kính chào quý thầy, cô!

Nhờ quý thầy, cô kiểm tra và đôn đốc những GV trong file đính kèm chấm và nhập điểm các thành phần để P.ĐTĐH, VPCCTĐB tiến hành duyệt ĐKHP bổ sung, duyệt hủy/mở lớp, xét tốt nghiệp, xét xử lý học vụ HK 1.

Nếu đã hết thời hạn GV nhập không được đề nghị GV gửi file điểm trực tiếp cho chuyên viên phụ trách của P.ĐTĐH, VPCCTĐB.

Trân trọng cảm ơn.

Sincerely,

Trần Bá Nhiệm, MSc.

Deputy Head, Office of Academic Affairs.
University of Information Technology - VNU-HCMC.
Phone: (028) 37251993 - Ext: 114.

4 Attachments

Đặc điểm DLL - Variety

Khó khăn, thách thức:

- Creating a common storage
- Integration
- Comparison, matching

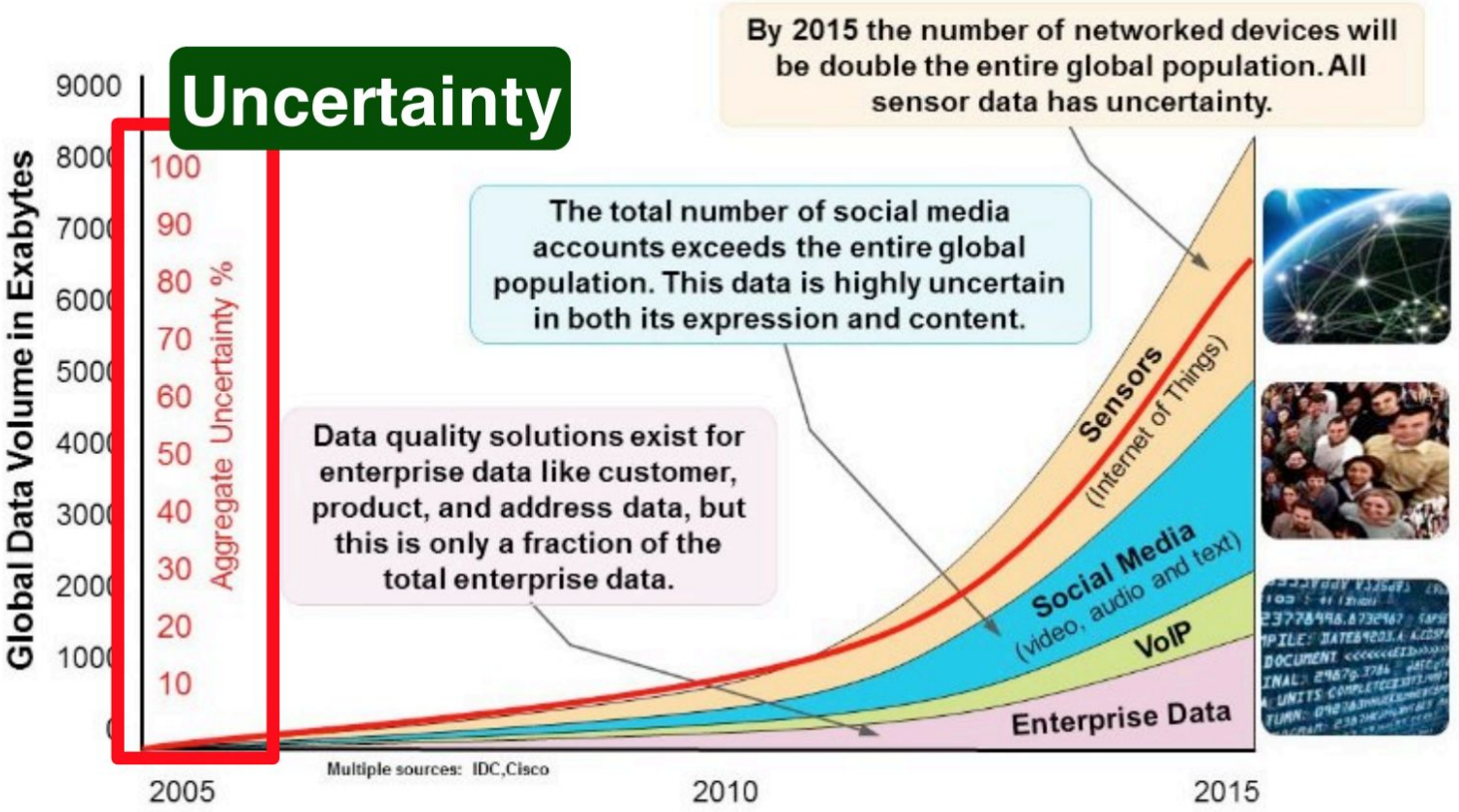
Đặc điểm DLL - Veracity (độ chính xác)

Veracity == Quality

- Độ chính xác của dữ liệu
- Độ tin cậy của nguồn dữ liệu
- Ngữ cảnh phân tích

→ **Lập luận, chứng minh không chắc chắn**

Đặc điểm DLL - Veracity



Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Đặc điểm DLL - Veracity (độ chính xác)

Khó khăn

- Ảnh hưởng độ tin cậy, độ chính xác của mô hình.

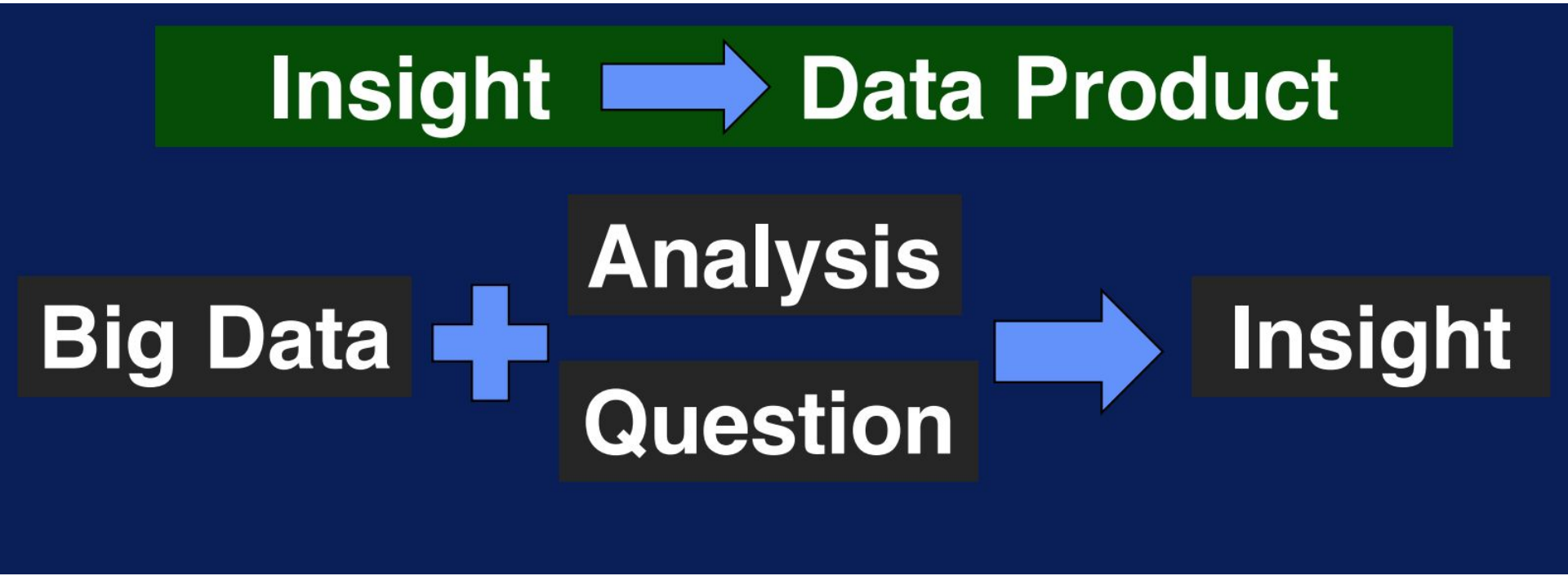
Đặc điểm DLL - Value

- Giá trị mà dữ liệu lớn mang lại?
  - lịch sử mua hàng, đọc tin, …
  - Dữ liệu từ các camera, sensors, …

Làm thế nào để lấy được Value từ BigData?

Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Làm thế nào để lấy được Value từ BigData?

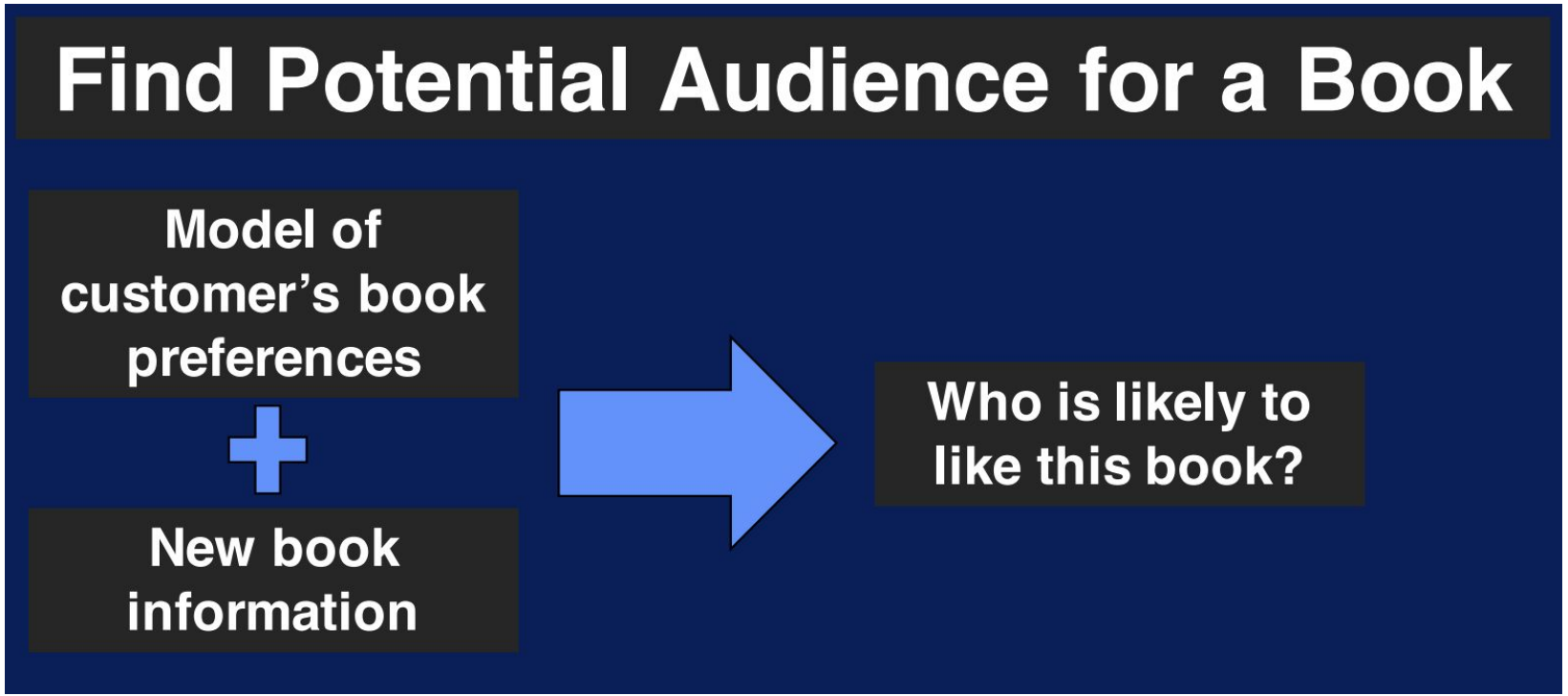


Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Làm thế nào để lấy được Value từ BigData?

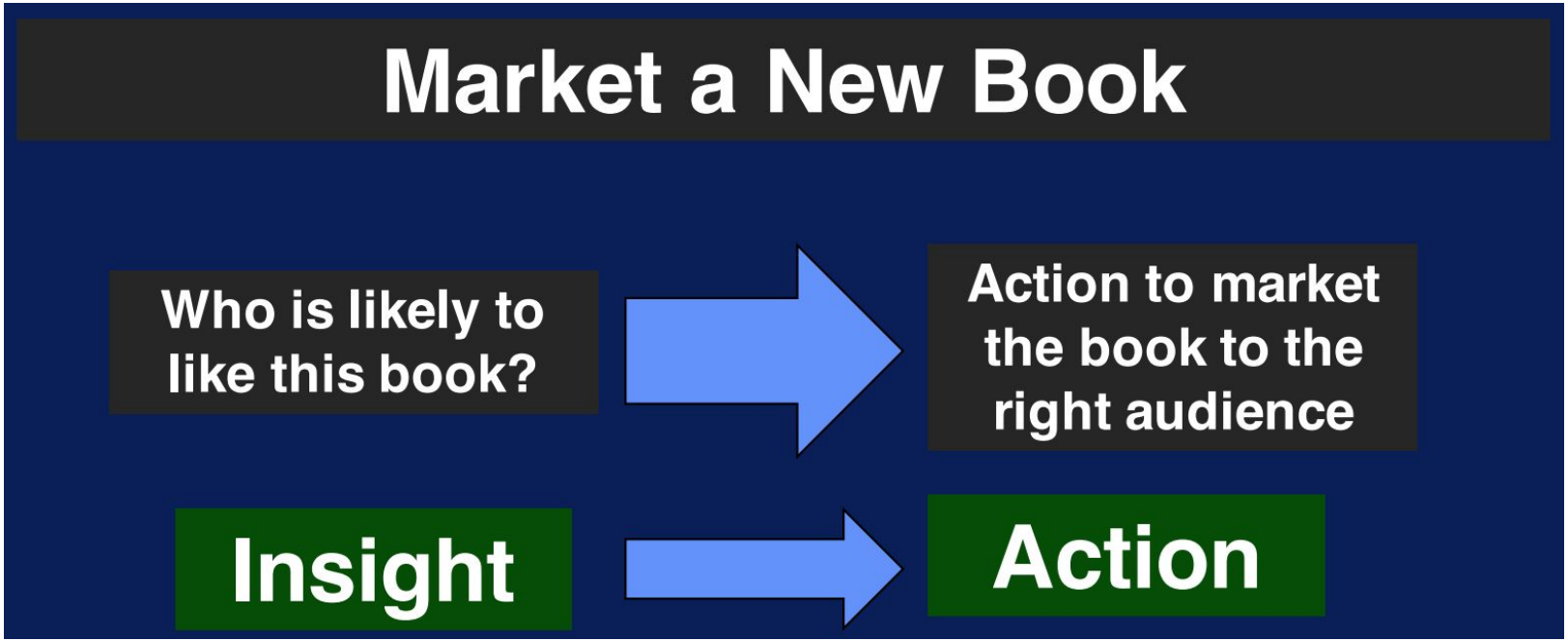
Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Làm thế nào để lấy được Value từ BigData?



Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Làm thế nào để lấy được Value từ BigData?



Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

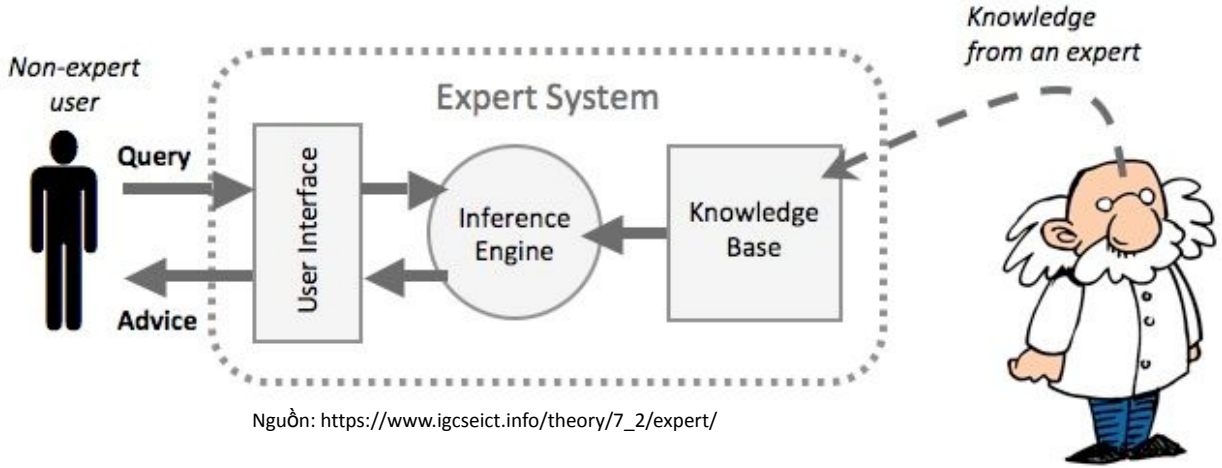
MỘT SỐ BÀI TOÁN DLL

DỮ LIỆU LỚN TRONG Y TẾ

Tham khảo:
GS Hồ Tú Bảo, Electronic Medical Records, Mini-Course Data Science, VIASM, 2017
http://www.jaist.ac.jp/~bao/DS2017/

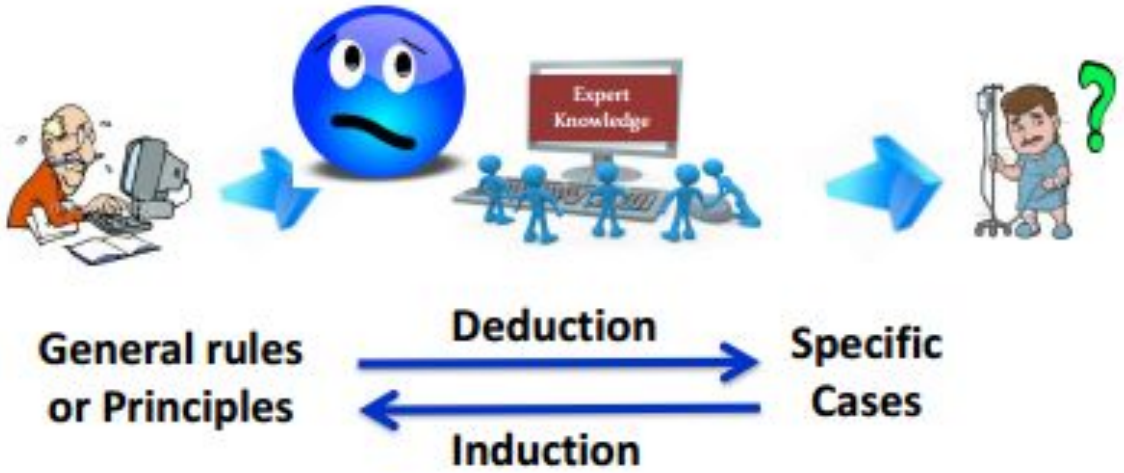
HỆ CHẨN ĐOÁN Y KHOA MYCIN (TIẾP CẬN SUY DIỄN)

MYCIN, là một trong những hệ chuyên gia đầu tiên do Edward Shortliffe phát triển vào giữa thập niên 70 ở trường Y khoa Stanford, viết bằng LISP. (A Rule-Based System)



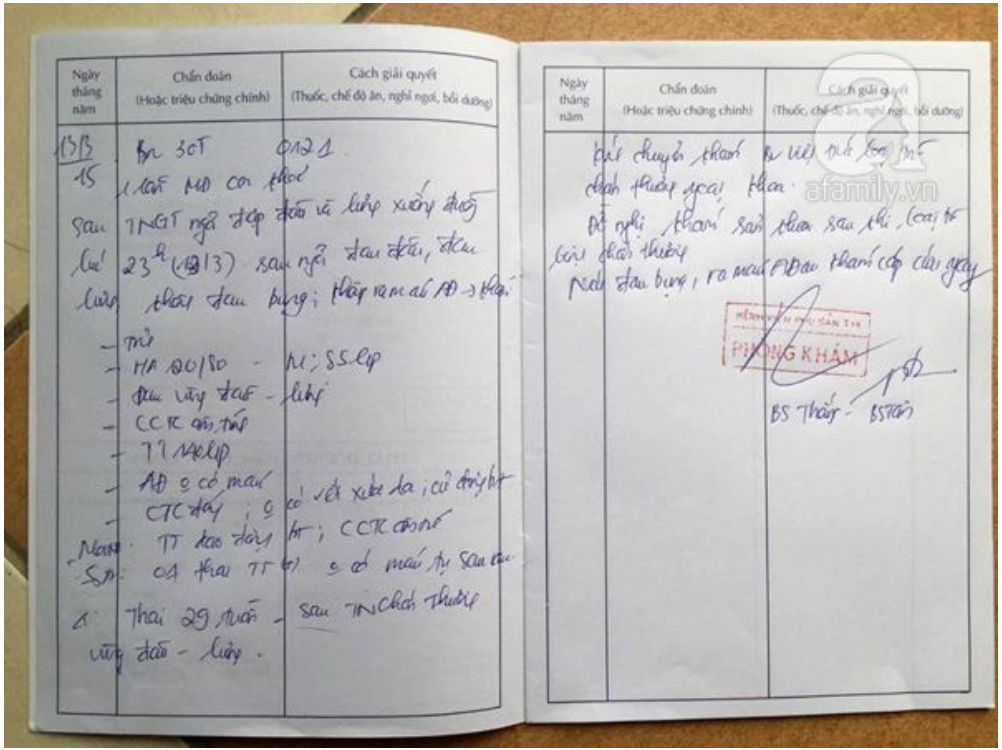
Nguồn: https://www.igcseict.info/theory/7_2/expert/

BỆNH ÁN ĐIỆN TỬ
(TIẾP CẬN QUI NẠP - INDUCTION)



Nguồn: GS Hồ Tú Bảo, Electronic Medical Records, Mini-Course Data Science, VIASM, 2017.

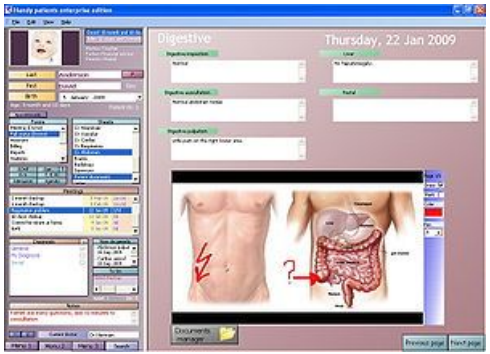
BỆNH ÁN TRUYỀN THỐNG



Giúp theo dõi quá trình bệnh lý của từng bệnh nhân riêng biệt mỗi lần đi khám chữa bệnh hay mỗi lần nhập viện.

BỆNH ÁN ĐIỆN TỬ (ELECTRONIC MEDICAL RECORDS)

- **EMRs**: phiên bản số hóa của bệnh án truyền thống, lưu giữ dữ liệu khám chữa bệnh của bệnh nhân trong một lần khám chữa bệnh. EMRs được thu thập và dùng trong các hệ thống thông tin quản lý trong bệnh viện.
- **EHRs** (Electronic Health Records - hồ sơ sức khỏe điện tử): lưu trữ thông tin nhiều lần khám chữa bệnh của người bệnh, được chia sẻ giữa các bệnh viện, tổ chức y tế.



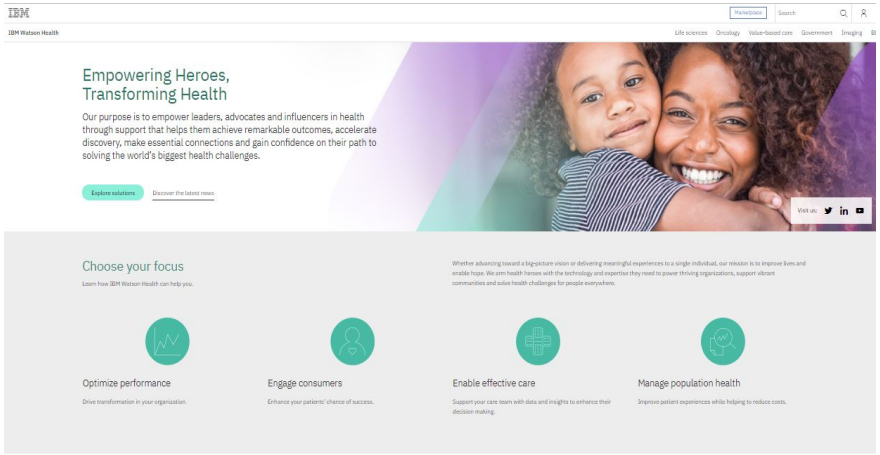
Nguồn: https://en.wikipedia.org/wiki/Electronic_health_record

Nguồn:
http://www.medicalbillingcodings.org/2015/07/emrvsehr-phrcprep rpcr-electronicmedicalrecords.html

Vài trò ý nghĩa EMRs, EHRs

- Số hóa thông tin khám chữa bệnh phục vụ quản lý, xử lý, tìm kiếm, chia sẻ thông tin.
- Cung cấp một lượng dữ liệu rất lớn hỗ trợ khám chữa bệnh, nghiên cứu khoa học trong y học.

IBM Watson Health



→ IBM Watson Health đại diện cho "quan hệ đối tác giữa nhân loại và công nghệ". Một kỷ nguyên chăm sóc sức khỏe mới.

→ Tạo cơ hội cho những người lãnh đạo, những người có ảnh hưởng trong y tế, giúp họ đạt được những kết quả đáng ghi nhận, tạo ra những kết nối cần thiết và tự tin trên con đường giải quyết những thách thức về sức khỏe lớn nhất trên thế giới.

IBM Watson Health

Nguồn: https://www.ibm.com/blogs/watson-health/watson-medical/

"Watson đã giúp tôi chống lại bệnh ung thư phổi như thế nào"

BÀI TOÁN TRONG THƯƠNG MẠI ĐIỆN TỬ

Báo cáo doanh số quý 2, 2012 của Amazon tăng 29%, từ 9.9 tỷ đô lên 12.83 tỷ đô so với cùng kỳ năm trước. Sự tăng trưởng đó được cho là do cách Amazon đã tích hợp các chức năng recommendation vào mọi nơi của quá trình bán hàng.

Nguồn: http://fortune.com/2012/07/30/amazons-recommendation-secret/

ỨNG DỤNG TRONG MARKETING

Sentiment Analysis

What is Sentiment Analysis?

Bài toán phân tích cảm xúc, quan điểm cộng đồng về các thương hiệu, nhãn hàng, sản phẩm ...

ỨNG DỤNG TRONG AD

- Hiểu hành vi người dùng, → phân phối quảng cáo phù hợp.
- Hiểu nội dung bài viết (content Insight)

KHUYẾN NGHỊ TIN TỨC (NEWS RECOMMENDATION

MYCAFEBIZ

- HOT Nằm viện 72 ngày, bán cả 2 căn biệt thự ở trung tâm thành phố, tôi mới nhận ra nếu còn khỏe mạnh, dù chỉ với hai bàn tay trắng, tôi đã là tỷ phú triệu đô!
- HOT Bài học xương máu của "Vua bánh mì" Sài thành: Để vợ giữ két sắt suốt 20 năm gây dựng công ty, sau ly hôn phải ra đi với 2 bàn tay trắng, vợ nói 1 câu làm ông tỉnh ngộ!
- Tập đoàn Thiếu Lâm Tự và đế chế kinh doanh triệu USD ít người biết tới
- Tài sản ông Phạm Nhật Vượng tăng 1 tỷ USD chỉ trong 10 ngày, lần đầu lọt top 200 người giàu nhất hành tinh
- Đại gia ăn chay Xuân Trường: Doanh nhân kín tiếng rước xá lợi Phật, mua thiên thạch mặt trăng, xây ngôi chùa lớn nhất thế giới
- Một gia đình Nhật tán gia bại sản vì trò đùa dại dột của con trai tại chuỗi sushi băng chuyền

Xem tiếp tin MyCafeBiz »

TIN NỔI BẬT KENH 14



ĐỌC THÊM

Cảnh đìu hiu trước trận U22 Việt Nam đấu U22 Philippines tại giải Đông Nam Á

22 giờ trước

U22 Việt Nam chạm trán U22 Philippines là trận đấu mở màn U22 AFF Cup 2019. Trận đấu diễn ra vào khung giờ khá nắng nóng và không thu hút nhiều người xem.

ỨNG DỤNG TRONG GIAO THÔNG

Nguồn: https://www.slideshare.net/mok/smart-city-and-innovation-in-hong-kong-priorities-and-challenges

- Cải thiện tình trạng giao thông công cộng?

Các bước chính trong khoa học dữ liệu lớn

Nguồn: https://www.coursera.org/learn/big-data-introduction/, week 2

Bàn luận

- Bạn quan tâm đến DLL không?
- Bạn quan tâm đến vấn đề/bài toán nào?

Tài liệu tham khảo

1. Mini-Course Data Science, VIASM, 2017, http://www.jaist.ac.jp/~bao/DS2017/
2. http://www.jaist.ac.jp/~bao/DS2017/
3. http://vnexpress.net/tin-tuc/khoa-hoc/hieu-ve-cach-mang-cong-nghiep-lan-thu-4-3574624.html
4. http://vnexpress.net/tin-tuc/khoa-hoc/viet-nam-di-trong-cach-mang-cong-nghiep-lan-thu-4-the-nao-3575368.html
5. https://www.coursera.org/learn/big-data-introduction/, week 2