如何導入資料科學?

實現data-driven轉型，並提倡data-culture

吳沛燊 Pei-shen Wu, MD

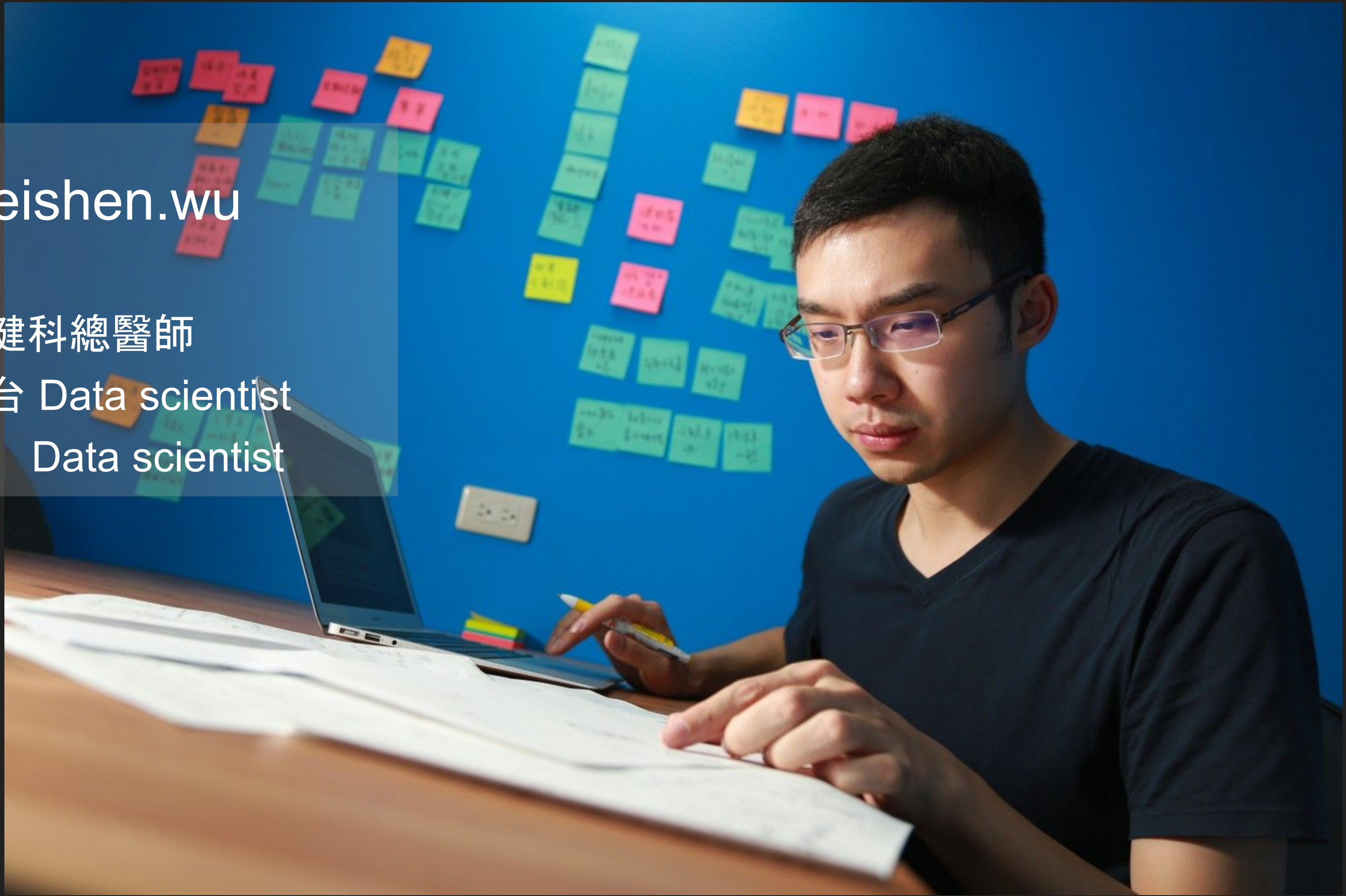
2018-2-24 @ 台灣人工智慧學校

吳沛燊 peishen.wu

台大醫院復健科總醫師

均一教育平台 Data scientist

Deepinsight　Data scientist



故事 1

當時的大問題：

回答速度趕不上

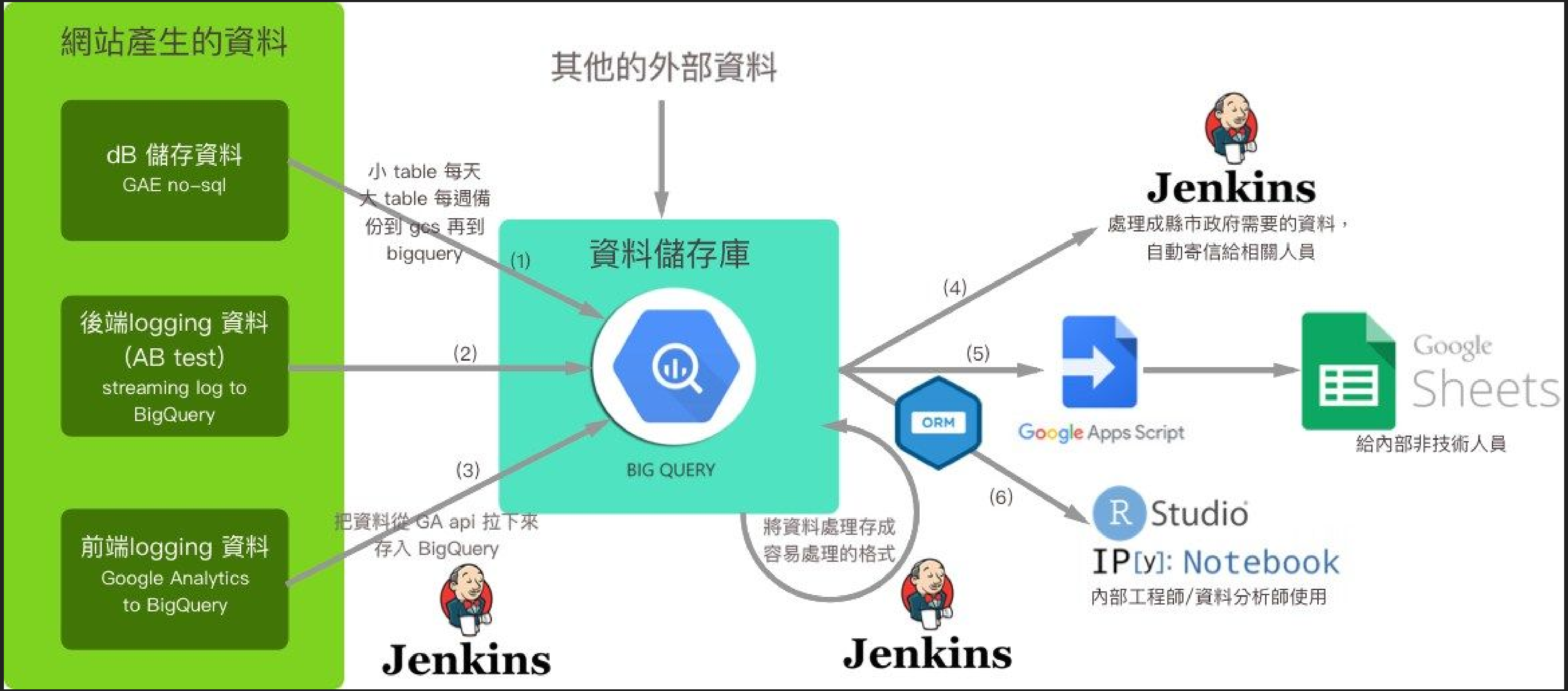
問題的產生

跟複雜度

回答單一問題所需的SQL code

試圖以三週努力
滿足三秒的好奇

均一的 資料pipeline 架構

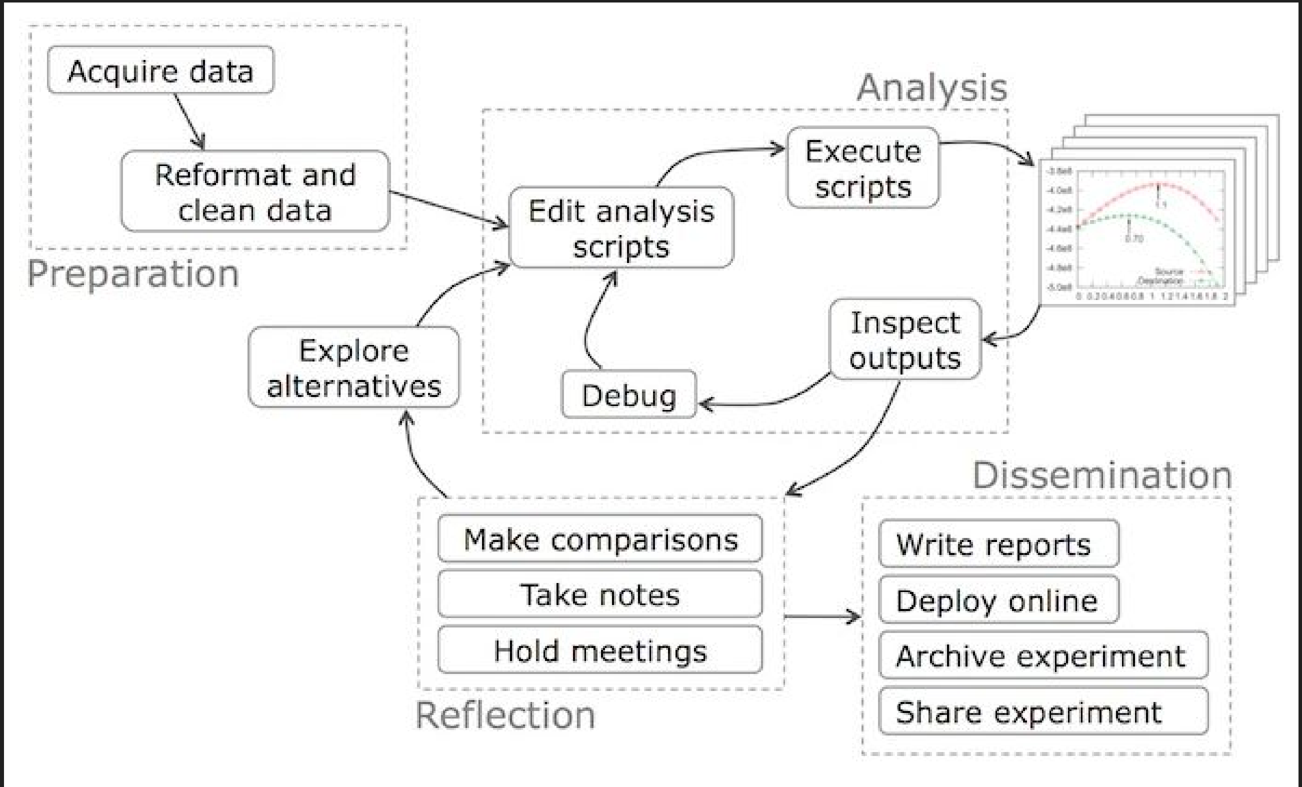


但 資料 =/= 知識

Data 架構 = 資料怎麼被收集、儲存、處理 跟散佈

Information 架構 = 把資料轉換成有用的資訊(知識), 所需要的過程跟practice

Data workflow 才是把資料轉成知識的架構

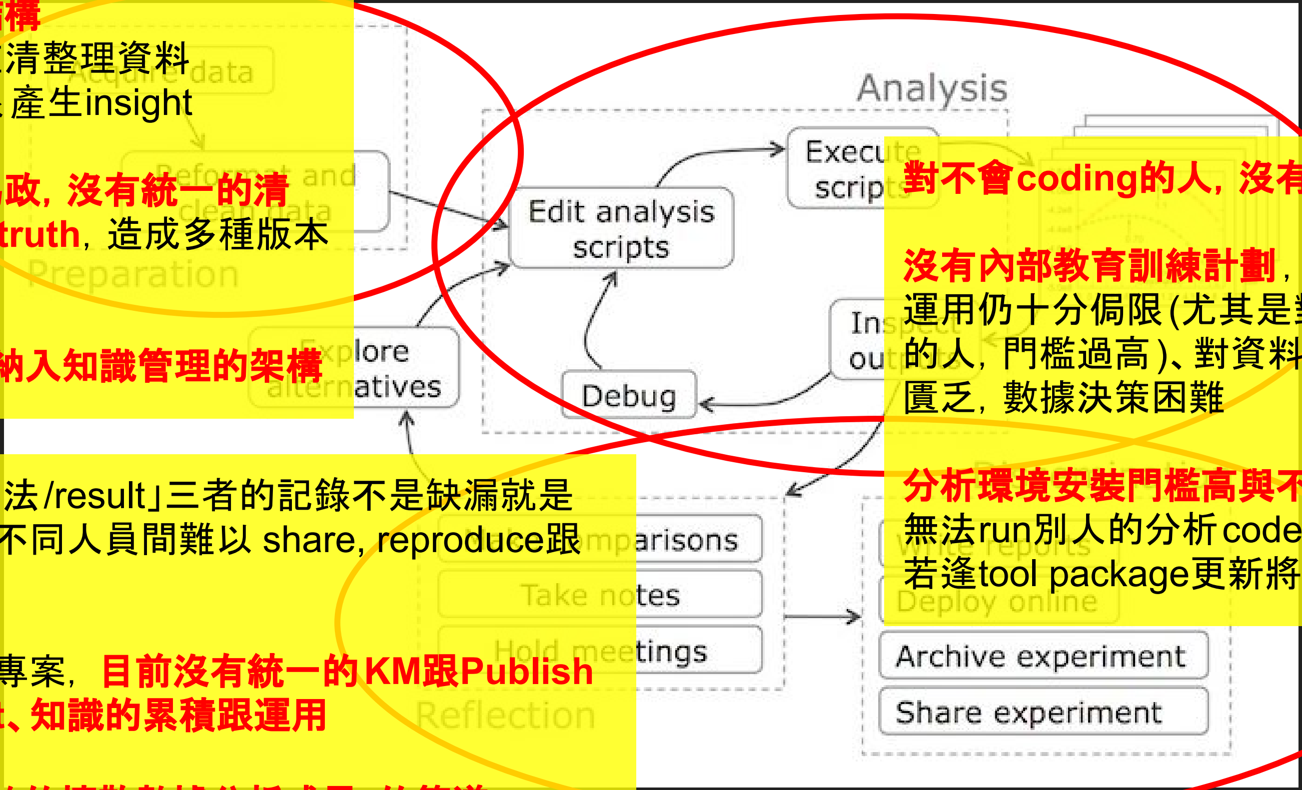


但問題叢生，處處是斷點，阻礙資料發展

不適任的資料庫結構
造成花太多時間在清整理資料
而不是進行分析跟 產生insight

資料表創建各自為政，沒有統一的清理流程 與 single truth，造成多種版本的數據結果

A/B testing 尚未納入知識管理的架構



對不會coding的人，沒有合適工具

沒有內部教育訓練計劃，造成資料的運用仍十分侷限（尤其是對不會coding的人，門檻過高）、對資料表的觀念仍匱乏，數據決策困難

分析環境安裝門檻高與不一致，造成無法run別人的分析code
若逢tool package更新將會是場災難

「code/分析時的想法/result」三者的記錄不是缺漏就是四散在各處，造成不同人員間難以 share, reproduce跟verify彼此的結果

此外，對已結案的專案，**目前沒有統一的KM跟Publish機制，不利Insight、知識的累積跟運用**

目前缺乏「持續有效的擴散數據分析成果」的管道

資料能否產生價值，還是要回歸到架構本身

一個系統的價值能否隨著時間增長的關鍵

= 人員從資料學習的容易度 + 將所得的insight自動化/系統化

(enable to learn from incoming data + rapidly operationalize those learnings)

數據成果得以
持續擴散的管道

內部教育訓練

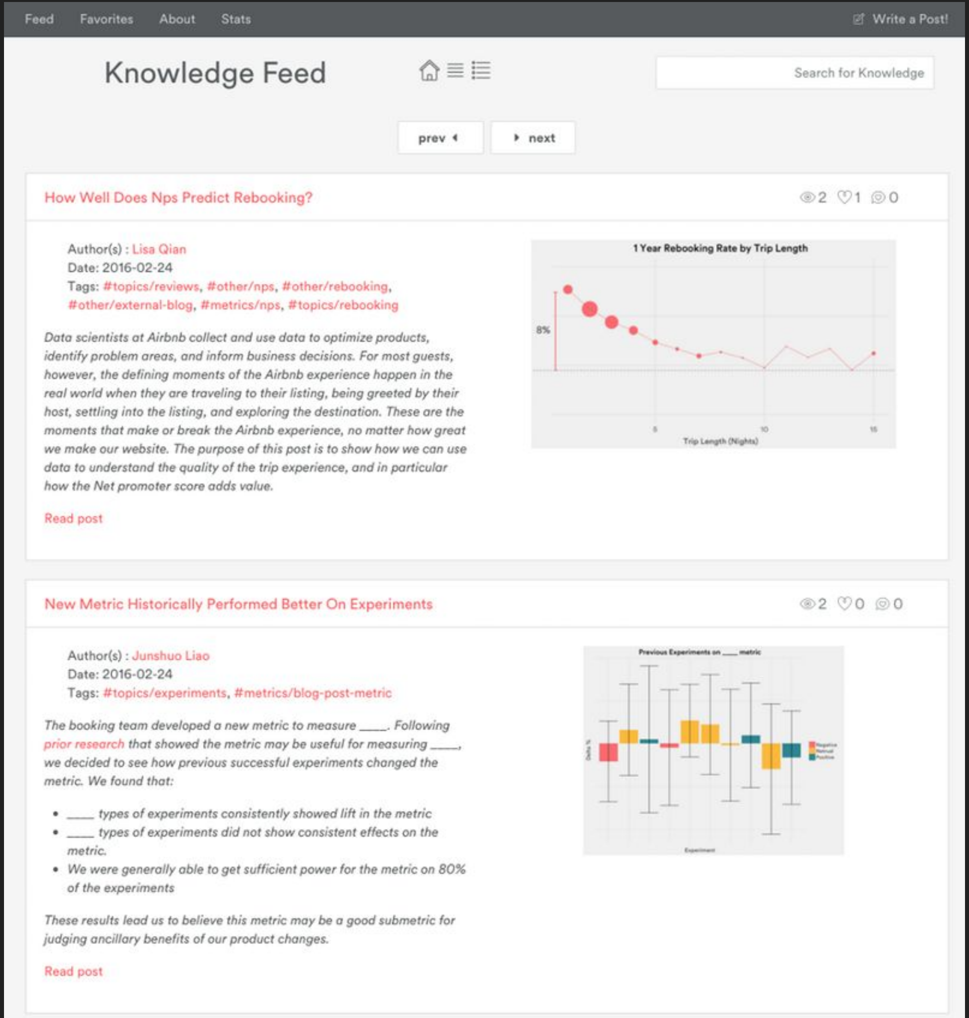
專門例會討論
資料運用議題

唯有 Full-stack solution
才能徹底解決問題

分析工具
與共用的分析環境

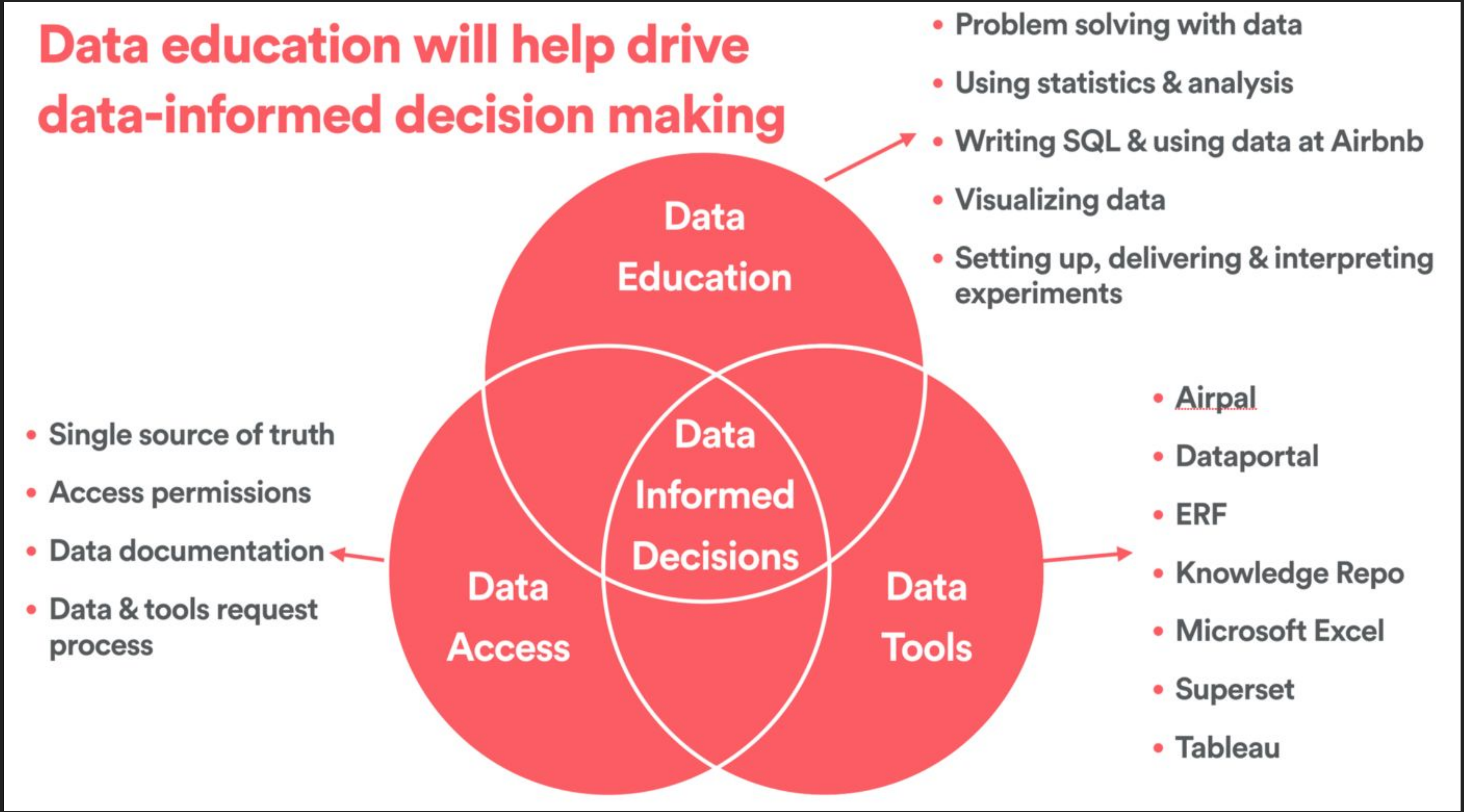
Best practice與
Data standards

好的資料庫架構



Airbnb knowledge repo

https://github.com/airbnb/knowledge-repo



https://medium.com/airbnb-engineering/how-airbnb-democratizes-data-science-with-data-university-3eccc71e073a

故事 2

https://www.junyiacademy.org/

影片

題目

影片

影片 題目

題目

課綱分類

任務

影片

影片

題目

題目

指派作業的UI設計

具有讓使用者**自我揭露**的作用

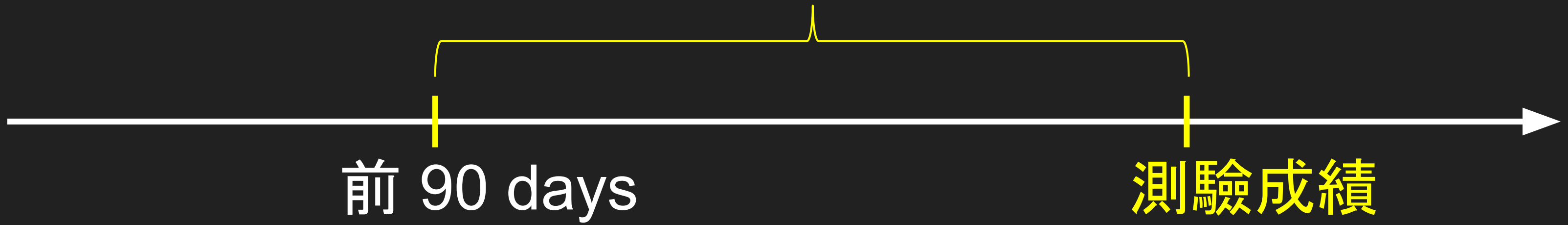
揭露：哪些物件具有學習上的關聯性？

由使用者指派的任務裡，許多(題目/影片)組合是在現行課綱找不到的

問題：

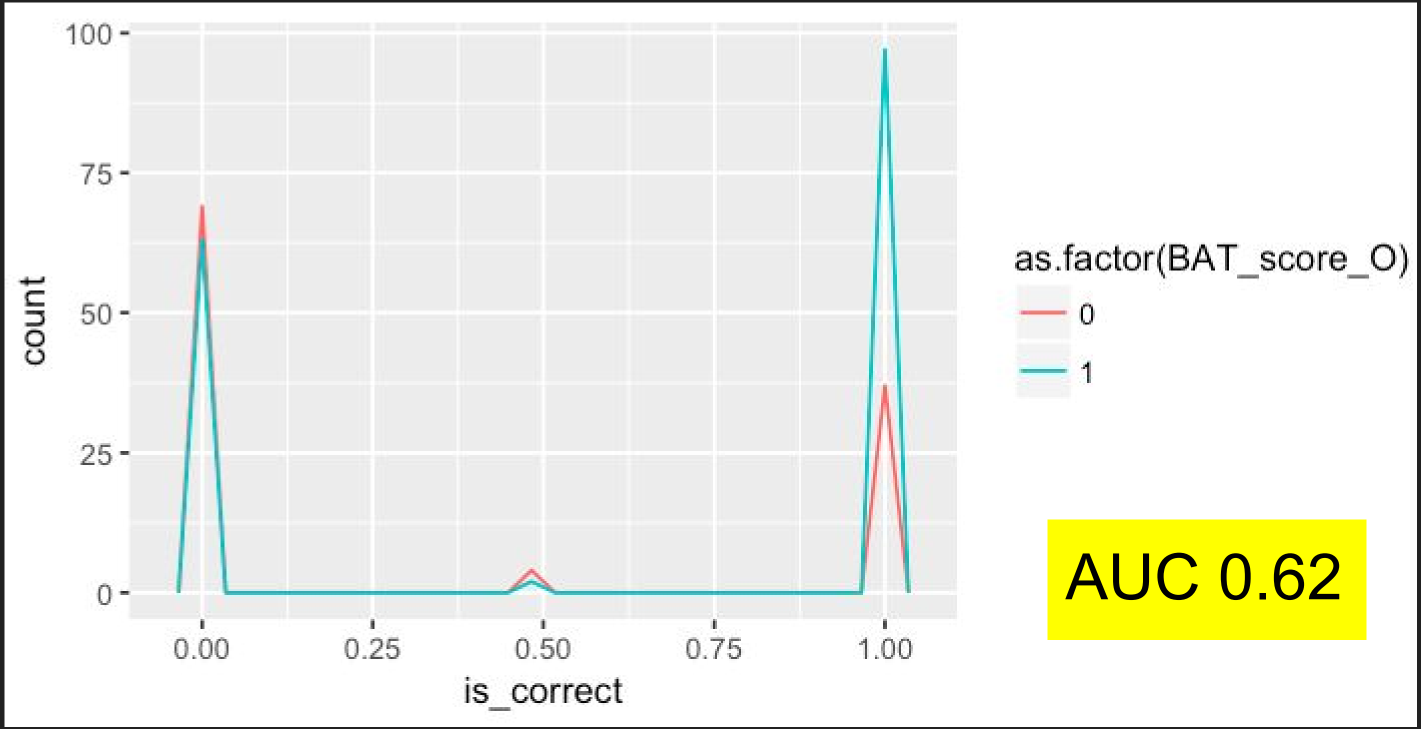
這個觀察有什麼重要性？

以某知識點的對錯去預測90天後對應的成績



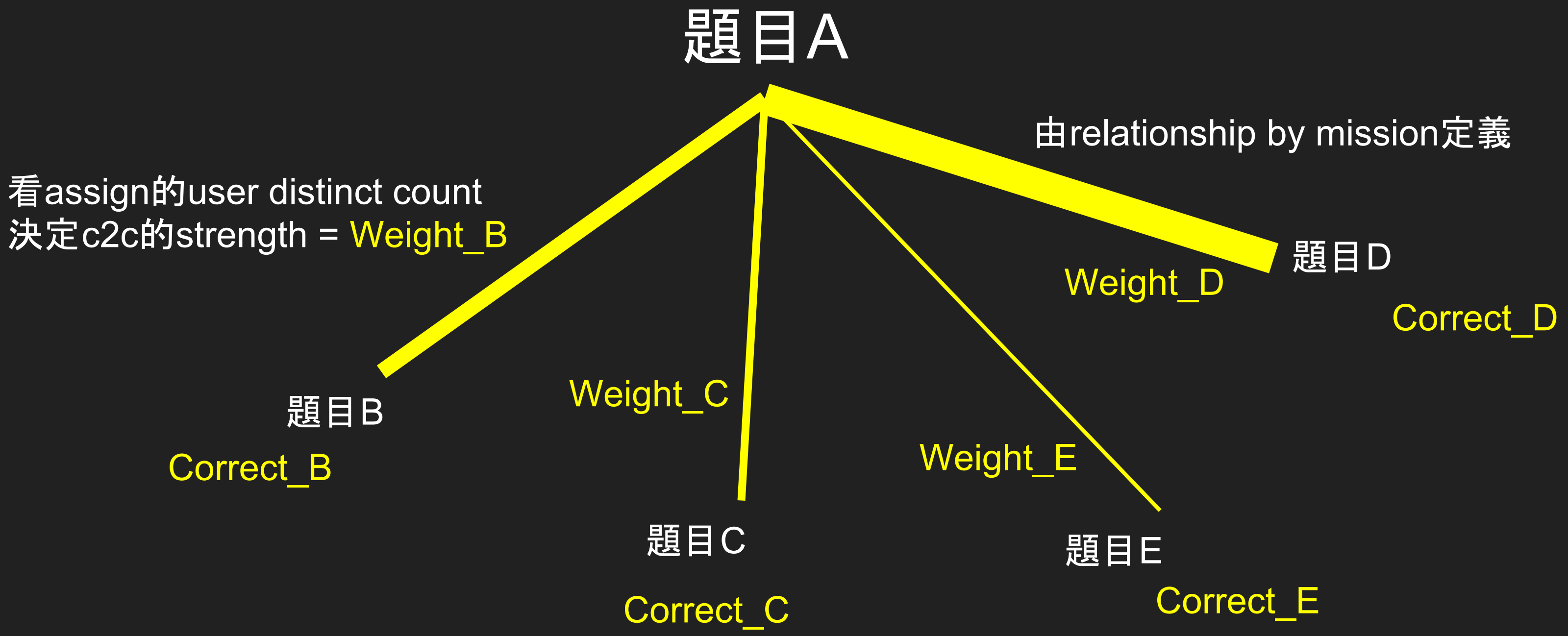
此為「能力」的
True north metric

發現：單點預測效果差

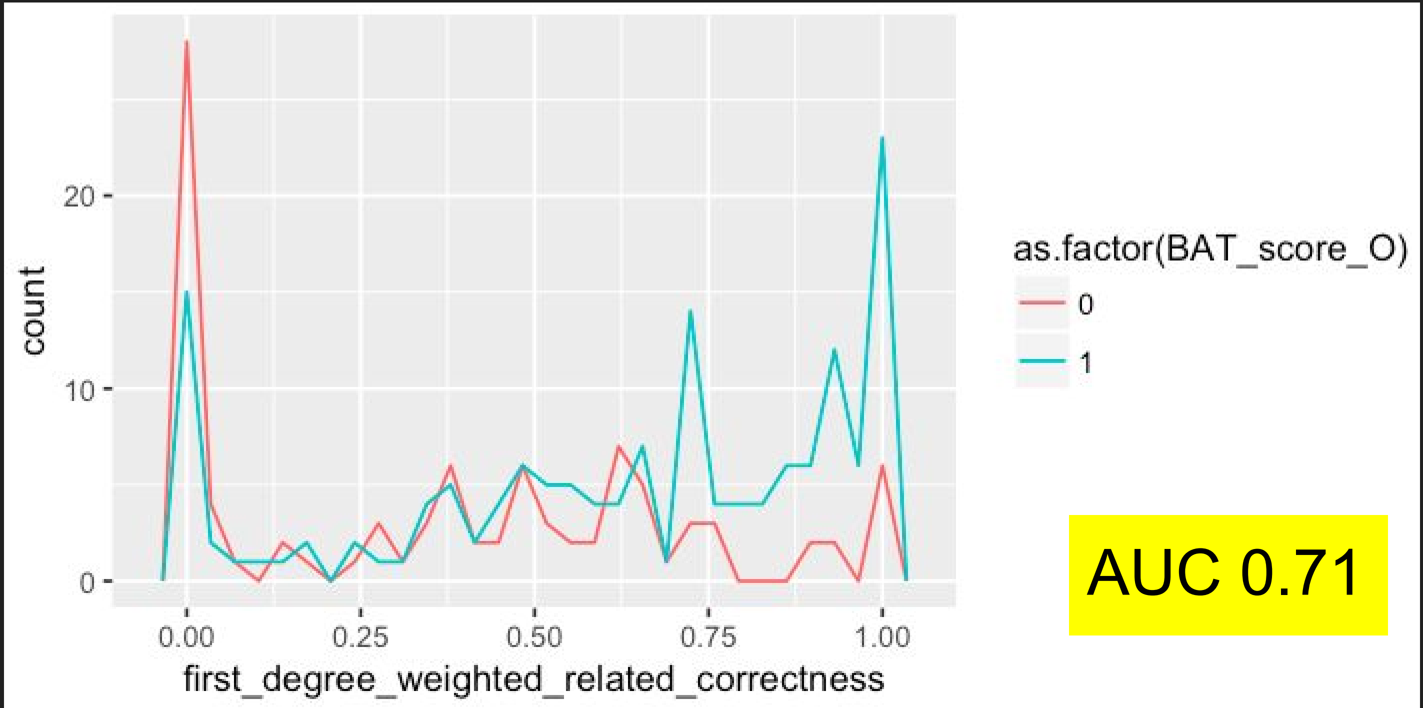


100
75
50
25
0
count
0.00
0.25
0.50
0.75
1.00
is_correct
as.factor(BAT_score_O)
0
1
AUC 0.62

$$\text{Correct\_A} = \frac{\text{SUM(Correct\_B*Weight\_B + … + Correct\_E*Weight\_E)}}{\text{SUM(Weight\_B + … + Weight\_E)}}$$



發現:網絡對單點的預測準確度提高



AUC 0.71

啓發1:

要答對能力測驗不能僅靠單點的能力

(不然不能解釋為何彼此相關的知識點的集體答對狀況，較能預測日後能力測驗成績)

啓發2：

現行課綱的侷限？
存在更好的學習方式？

(可以作為推薦系統的基礎)

啓發3：

網絡/關聯性的資料具有戰略意義

藥物A

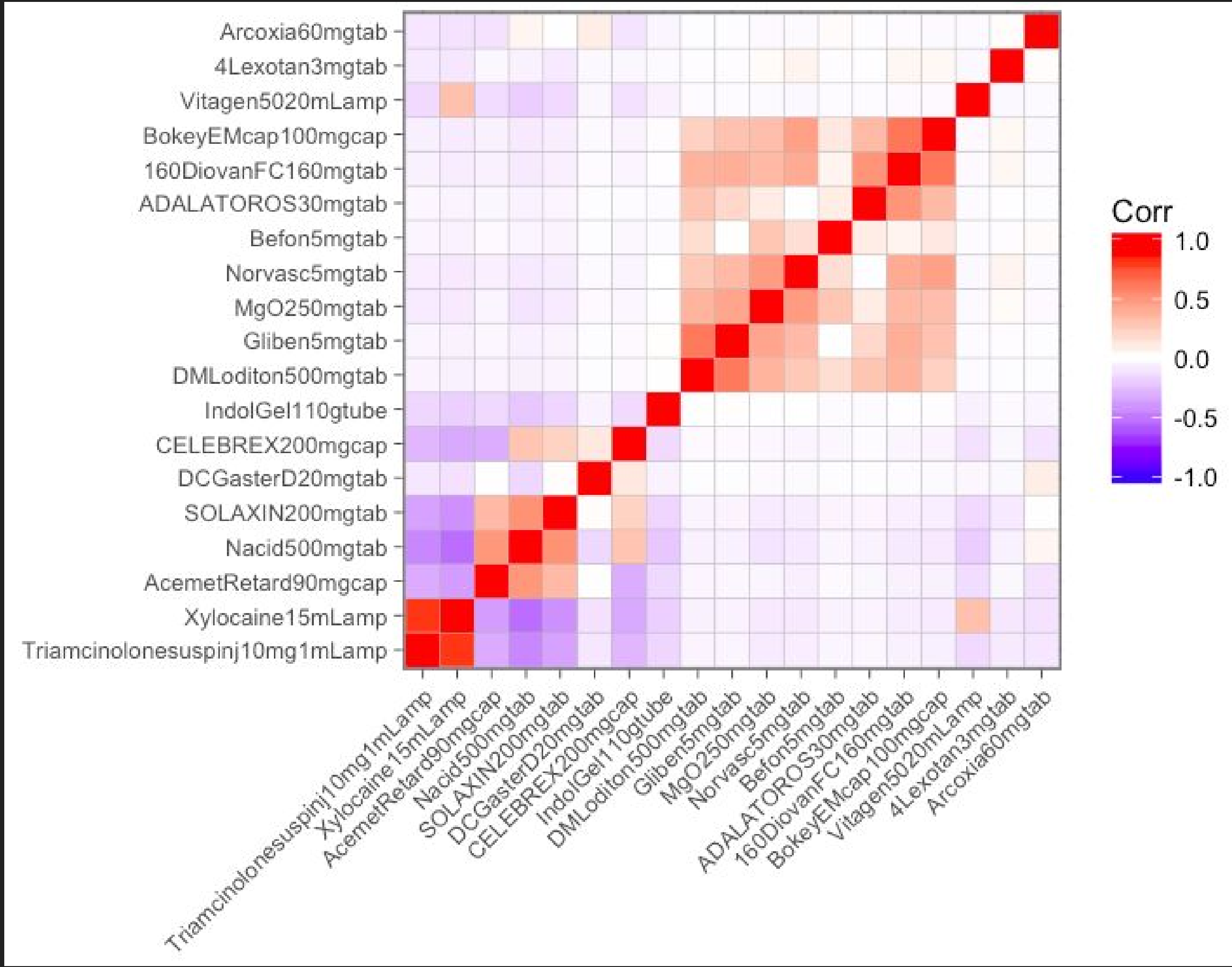
處方B

針劑D

藥物C

某病人
的處方

某教授的
用藥習慣


Arcoxia60mgtab
4Lexotan3mgtab
Vitagen5020mLamp
BokeyEMcap100mgcap
160DiovanFC160mgtab
ADALATOROS30mgtab
Befon5mgtab
Norvasc5mgtab
MgO250mgtab
Gliben5mgtab
DMLoditon500mgtab
IndolGel110gtube
CELEBREX200mgcap
DCGasterD20mgtab
SOLAXIN200mgtab
Nacid500mgtab
AcemetRetard90mgcap
Xylocaine15mLamp
Triamcinolonesuspinj10mg1mLamp
Corr
1.0
0.5
0.0
-0.5
-1.0

後續:

寫成SOP

讓服務水平不因人員經驗而有差異

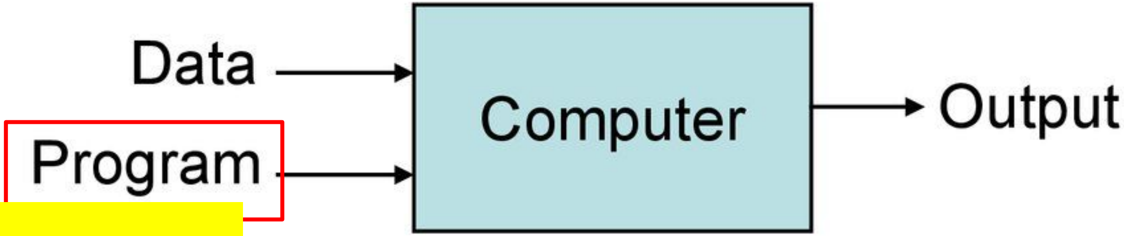
故事 3

機會在哪裡？

Machine learning

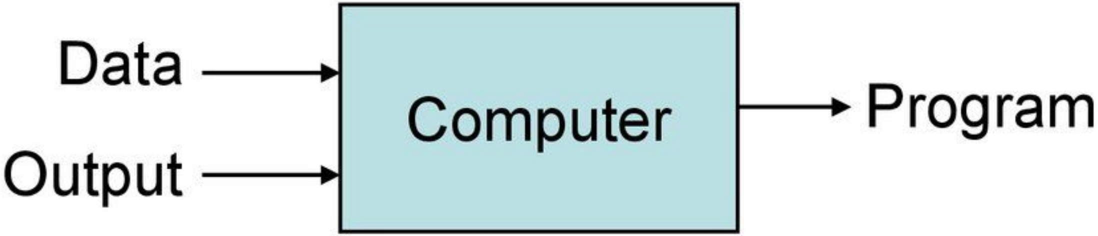
命令電腦做事情的paradigm shift

Traditional Programming

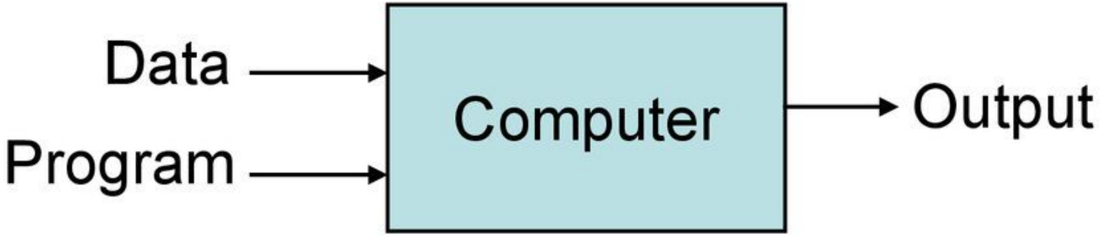


過去這一步最傷腦筋

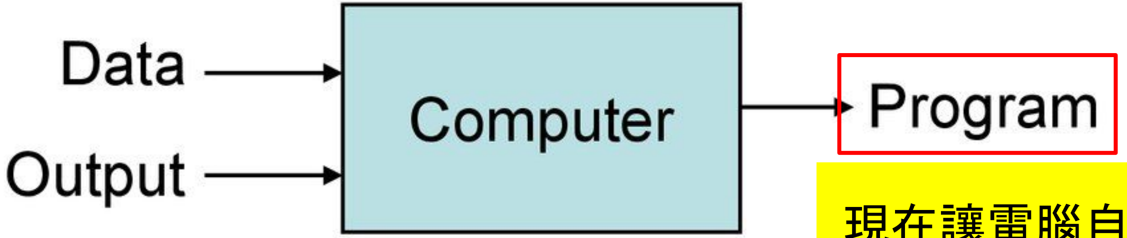
Machine Learning



Traditional Programming



Machine Learning

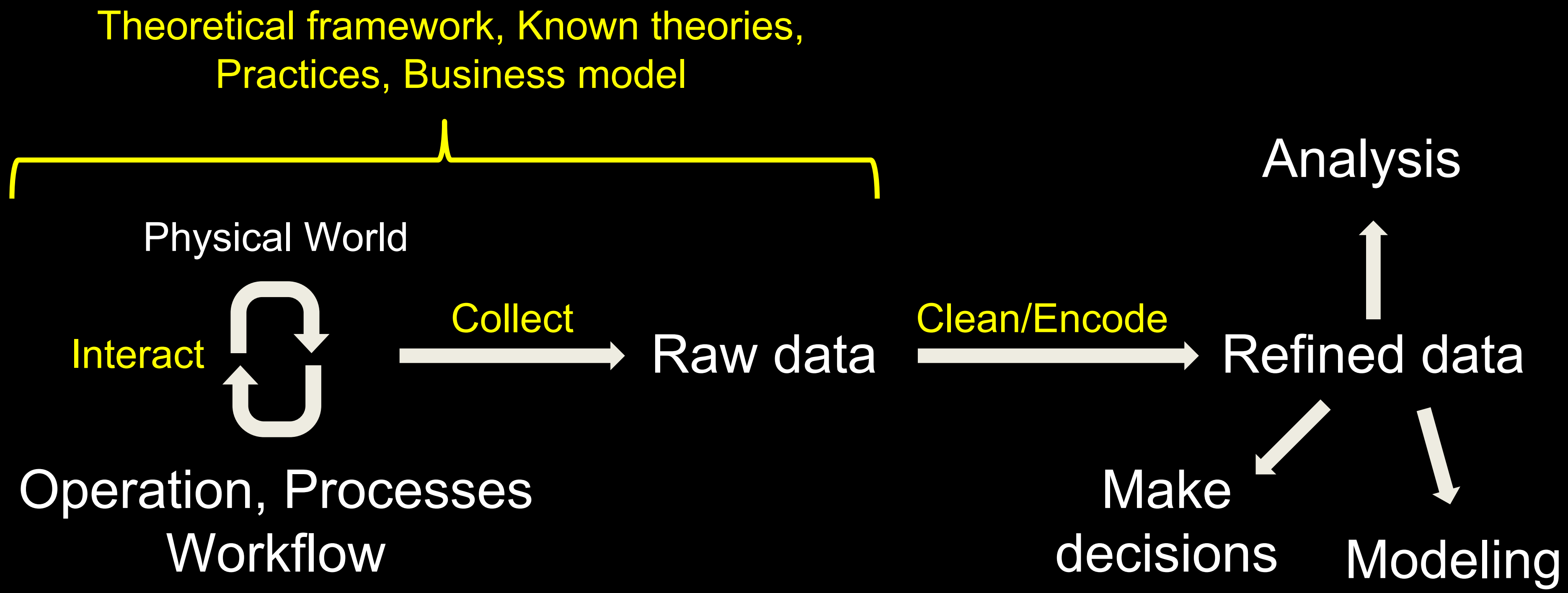


現在讓電腦自己想辦法

AI is not good or bad, it is even not neutral.
It is just as good as the data fed to it.

- Lokke Moerel

TEDx AmsterdamWomen
x = independently organized TED event

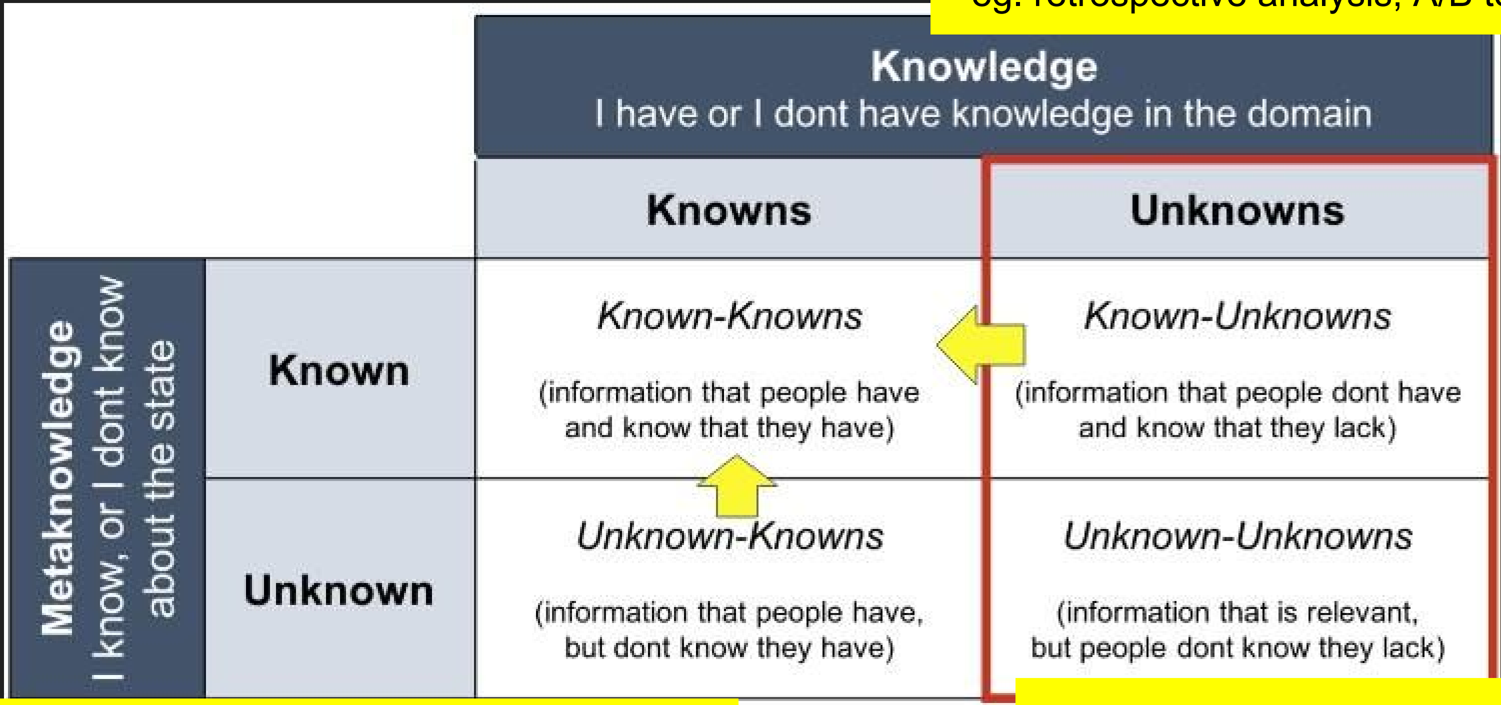

Theoretical framework, Known theories,
Practices, Business model
Physical World
Interact
Operation, Processes
Workflow
Collect
Raw data
Clean/Encode
Refined data
Analysis
Make
decisions
Modeling

資料的價值有其上限

所有的data都有前提

受限於當時collect方式跟real world互動的方法

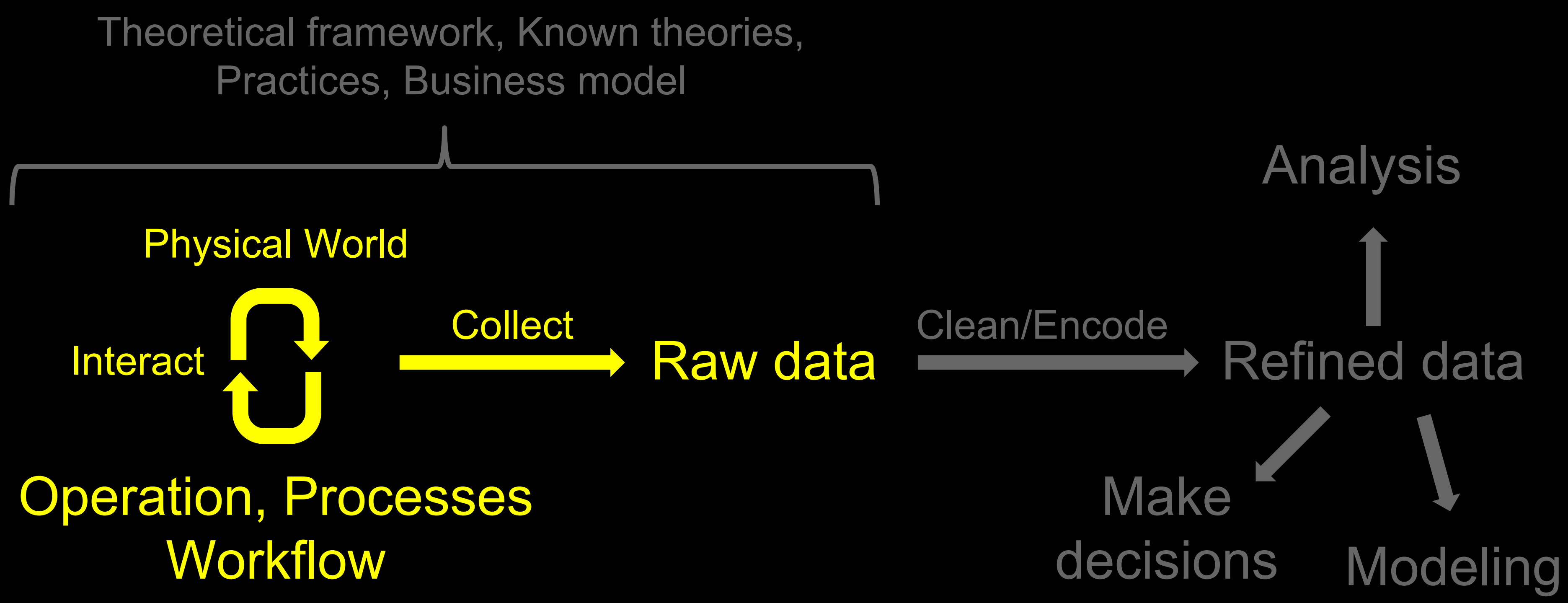
把過去aware但不了解的事情，變成深入了解
- eg. retrospective analysis, A/B tests...



		Knowledge I have or I dont have knowledge in the domain	
		Knowns	**Unknowns**
Metaknowledge I know, or I dont know about the state	**Known**	*Known-Knowns* (information that people have and know that they have)	*Known-Unknowns* (information that people dont have and know that they lack)
	Unknown	*Unknown-Knowns* (information that people have, but dont know they have)	*Unknown-Unknowns* (information that is relevant, but people dont know they lack)

把過去已知但 unaware 的事變成 aware
- eg. dissemination, notification, surveillance

無法從Unknown-unknowns得到價值


Theoretical framework, Known theories,
Practices, Business model
Physical World
Interact
Collect
Raw data
Operation, Processes
Workflow
Clean/Encode
Refined data
Analysis
Make
decisions
Modeling

" People who are really serious about software should make their own hardware "

\- Alan Kay

對於資料:

" Those who are really serious about analytics should devise ways to collect their own data "

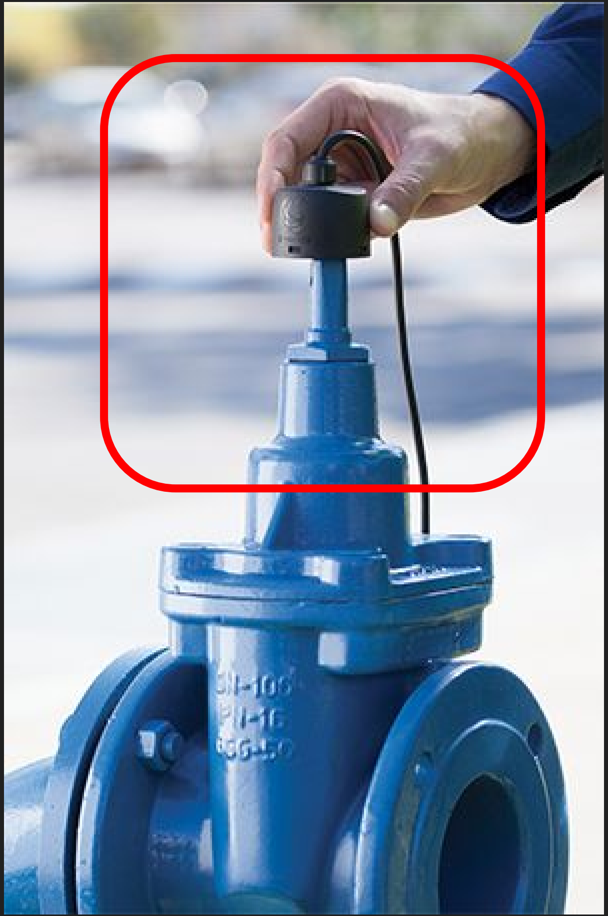
指派作業的UI設計

具有讓使用者自我揭露的作用

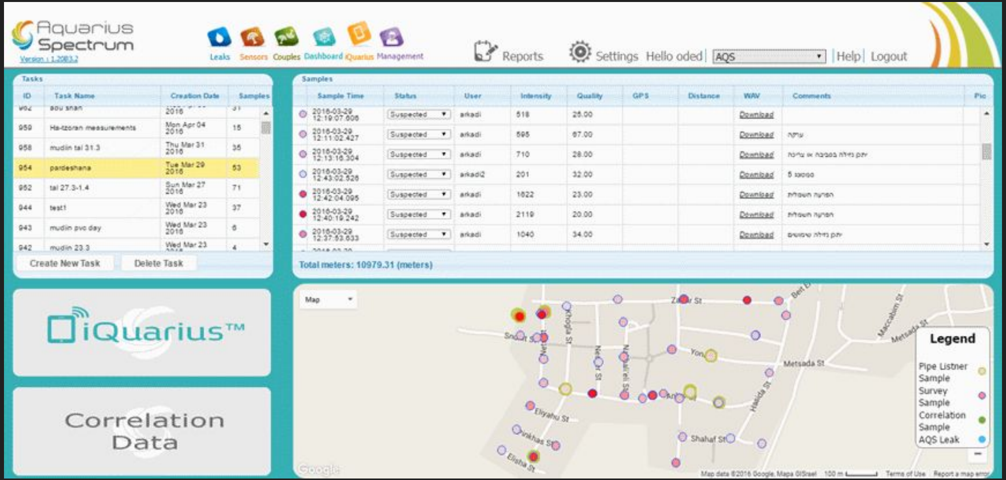
揭露：哪些物件具有學習上的關聯性？

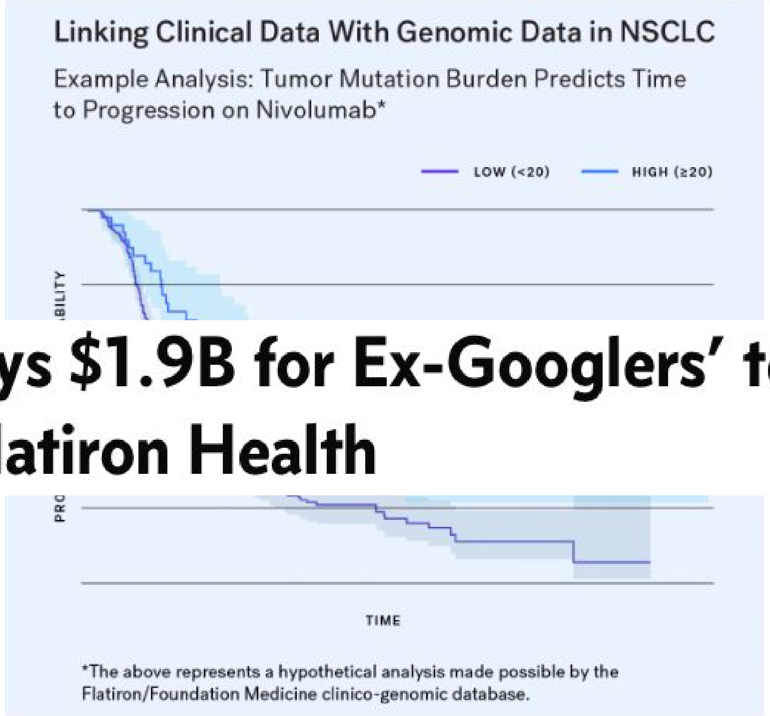

Quantity: 1
Yes, I want FREE One-Day Delivery with a free trial of Amazon Prime
Add to Basket
Sign in to turn on 1-Click ordering.
Add to Wish List

amazon



Aquarius Spectrum





Roche pays $1.9B for Ex-Googlers' tech startup Flatiron Health

flatiron

Integrated at the source. Expanded with linked data sets.

– Derived from the EHRs of over 265 community clinics and academic institutions at over 800 unique sites of care.

– The largest and highest quality source for real-world evidence in oncology – includes both structured and unstructured data.

– Access longitudinal clinical data, with the ability to link to external data sources like genomics, mortality and closed claims.

Our OncoCloud™ Suite

OncoEMR®

OncoBilling®

OncoAnalytics®

OncoTrials®

flatiron

“When we saw what OncoEMR could do, we were thrilled to discover how it thinks like an oncologist. Everything we needed was right at our fingertips.”

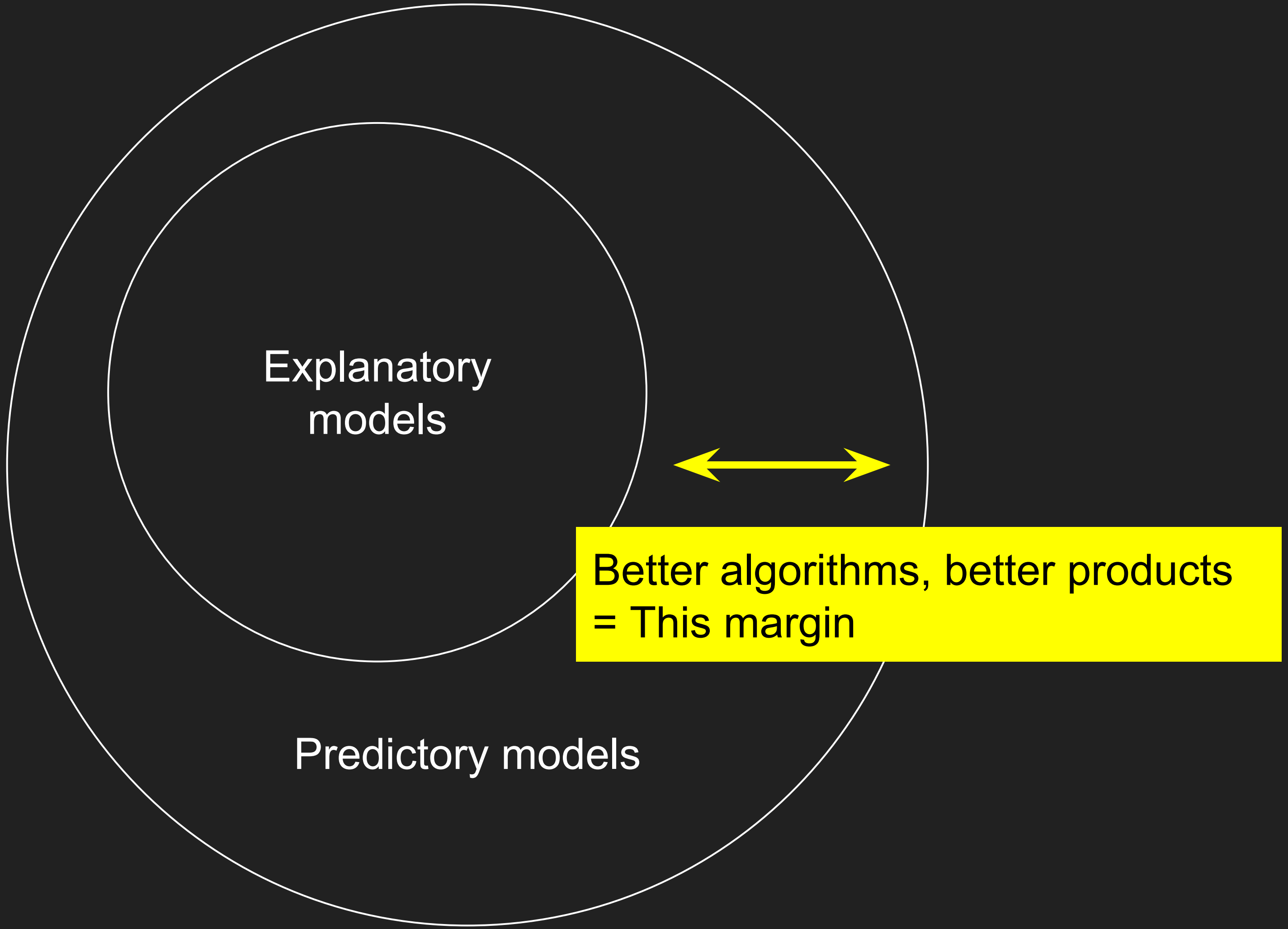
Fred Kudrik, MD
President, South Carolina Oncology Associates

Loss of model explanatory power
→ Increased predictive power

It is possible to reduce variance by increasing bias and still resulting in reduced overall error

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

https://arxiv.org/pdf/1101.0891.pdf

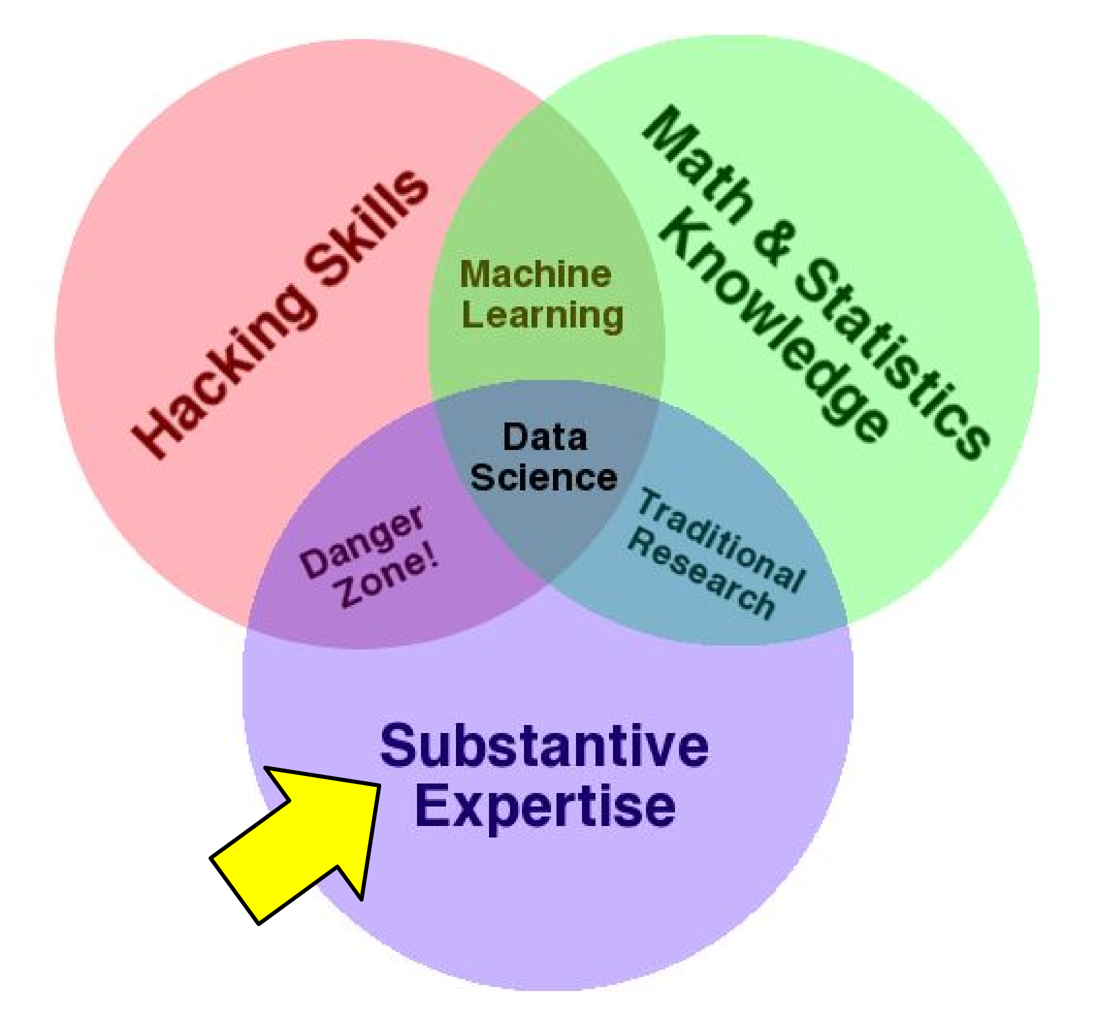

Explanatory
models
Better algorithms, better products
= This margin
Predictory models

小結

DS的重點不在Data
在Science

不光是data，還有

domain-relevant questions



Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

DS的產出是**軟體**嗎?

models, dashboard, database, pipelines ...

哪些可能的 future products?

潛在的marketing strategies
跟user needs尚未被發現?

哪些資料的收集會帶來的優勢？

DS產出的 Actionable insights 具有策略性質

協助組織運用資料加速成長

才是發展DS最重要的目的

Data workflow 才是把資料轉成知識的架構

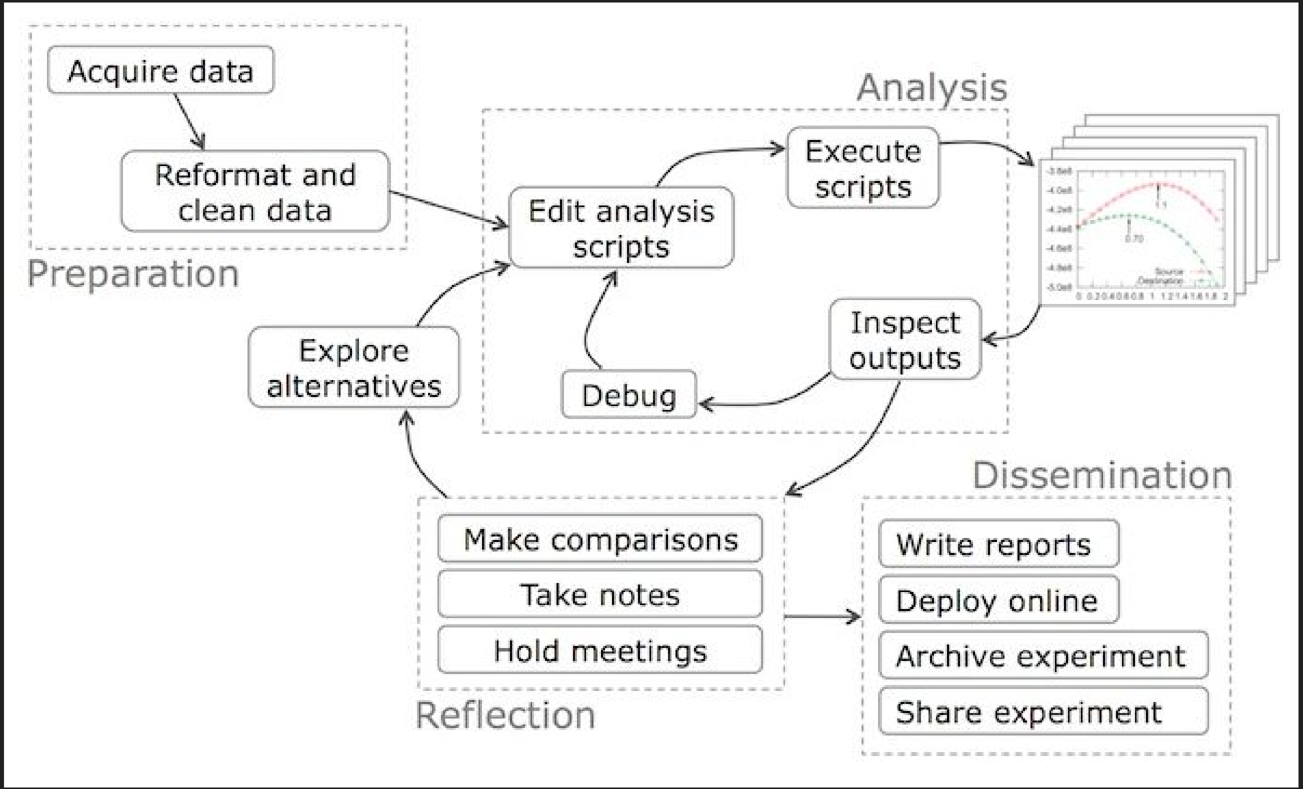
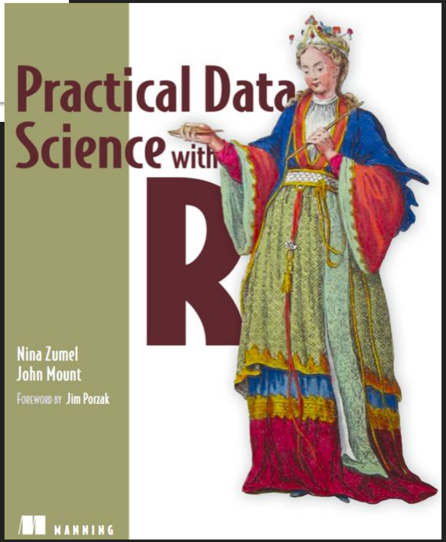


Table 1.1 Data science project roles and responsibilities

Role	Responsibilities
Project sponsor	Represents the business interests; champions the project
Client	Represents end users' interests; domain expert
Data scientist	Sets and executes analytic strategy; communicates with sponsor and client
Data architect	Manages data and data storage; sometimes manages data collection
Operations	Manages infrastructure; deploys final project results



從各端實際需要，去發現資料可以幫忙的地方 = 收集好的question跟痛點

業務端
產品端
客服端
…

Data engineer

Data Product Manager

Data scientist

協助後續追蹤、發展、測試及整合進產品

Take home message

資料端的產出是策略性的，應獨立於軟體產品，而直屬於決策單位

組織的資料力可以由 Data workflow 上的障礙來衡量

- Relevant questions → Data preparation → Analysis/Reflection → Dissemination

所有的data都有前提：上限就是當時的operation與collection process

" Those who are really serious about analytics should devise ways to collect their own data "