Introduction to Data Science

Dr. Sethavidh Gertphol

Outline

* What is "Data Science"?
* Modern Data Science
* Data Science Maturity Model
* Examples
* Data Science v.s. X

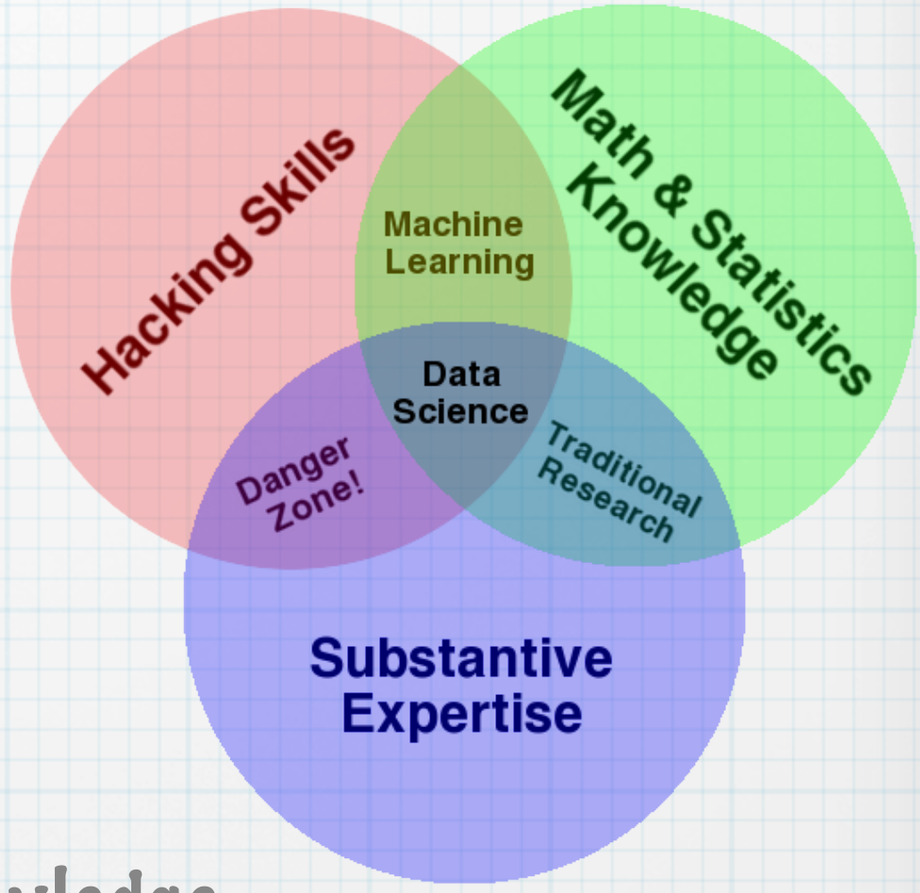
Data Science

* เก็บรวบรวมข้อมูลให้มากที่สุดเท่าที่จะเป็นไปได้
  * คอมพิวเตอร์/แท็บเล็ต
  * เซ็นเซอร์
  * มนุษย์
  * แบบสำรวจ
* ทำอะไรบางอย่างที่มีประโยชน์กับข้อมูลที่เก็บรวบรวมมาได้
  * เข้าใจสถานการณ์จนถึงปัจจุบัน
  * ยืนยันสมมติฐาน
  * สกัดข้อมูลเชิงลึก **(insight)**
  * ทำนายอนาคต และ ตัดสินใจ

Source: https://www.youtube.com/watch?v=yYLjbruwIQs

Data Science Venn Diagram

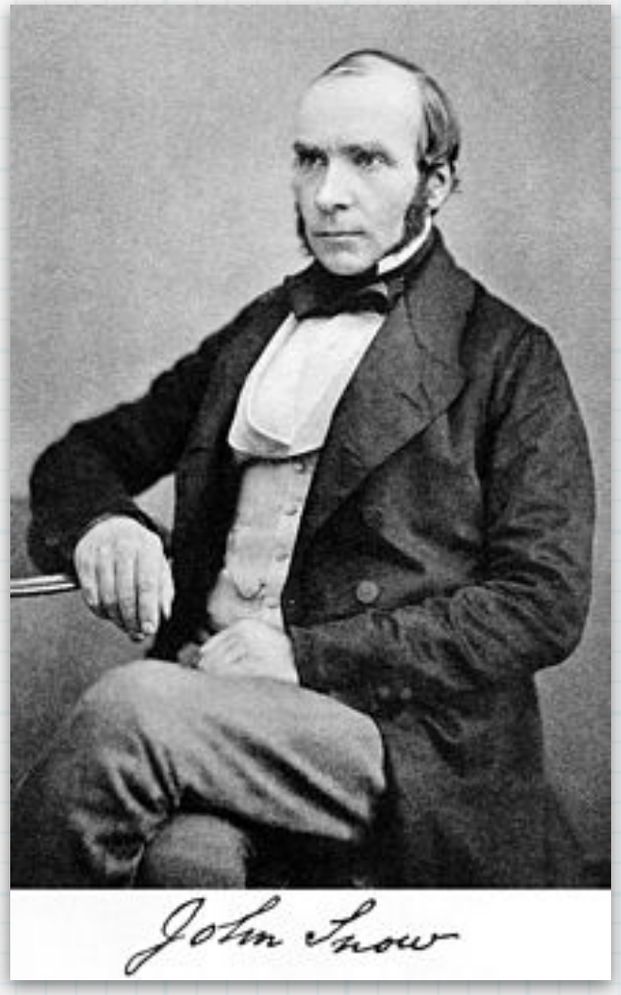
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



* Originally by Drew Conway in 2013
* **Hacking Skills** = Computer Skills
  * Gather, clean, store data
  * Use and tune algorithms
* **Math & Stats** for understanding
  * Apply appropriate models
  * Understanding results
* **Substantive Expertise** = Domain knowledge
  * Ask appropriate questions
  * Setup hypothesis

London 1854 Cholera Outbreak

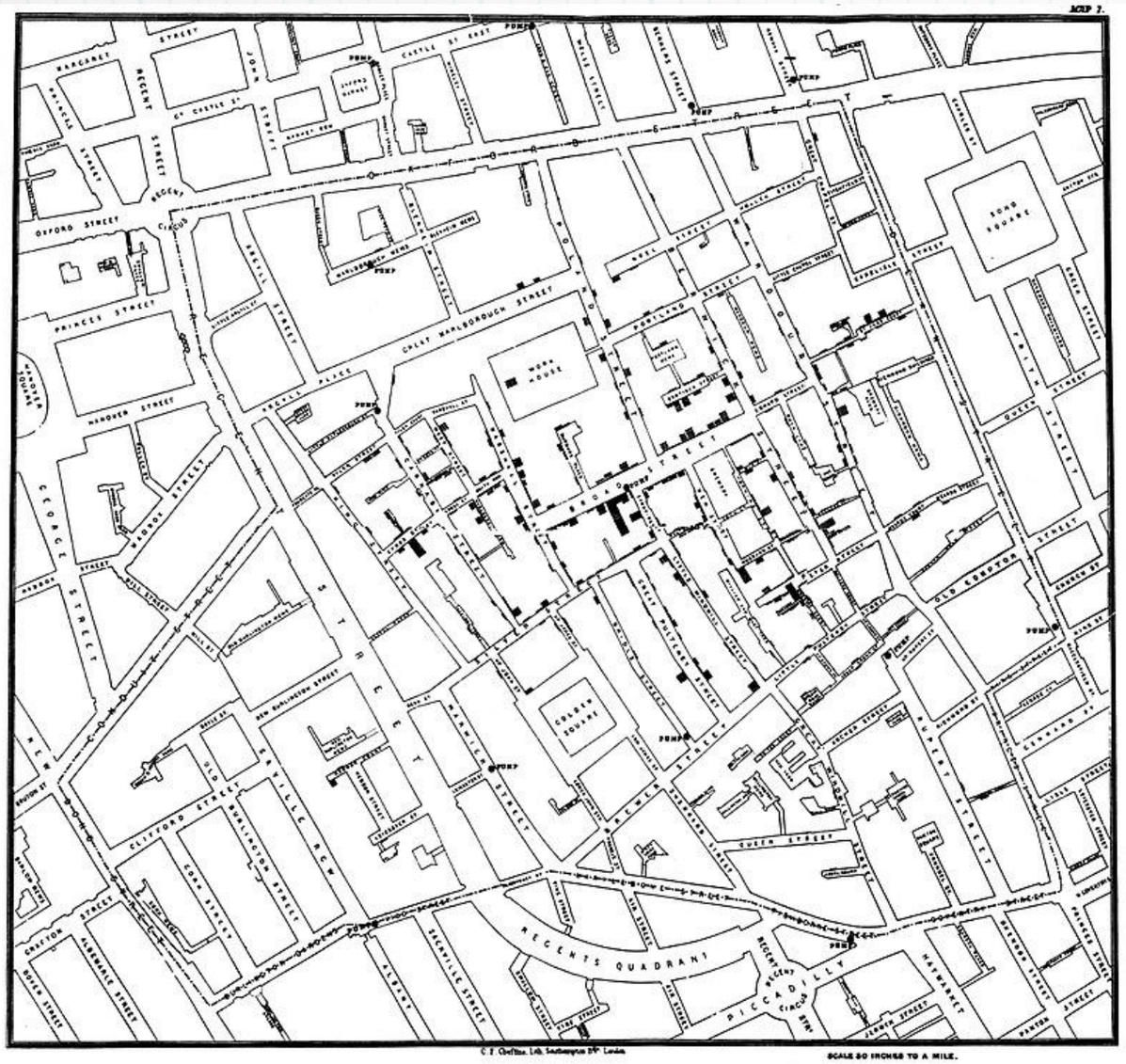
* Example of using data to verify hypotheses and solve problem
* August 1854: Cholera outbreak in Soho area of London
  * 127 deaths near Broad Street in 3 days
  * 616 deaths total
* One doctor named John Snow found and solved the problem



https://en.wikipedia.org/wiki/John_Snow

John Snow's Spot Map

* Show **deaths** as black bars on Soho street map
* Also plot **street water pumps** on the map
* The map showed that **most deaths grouped around water pump on Broad Street**

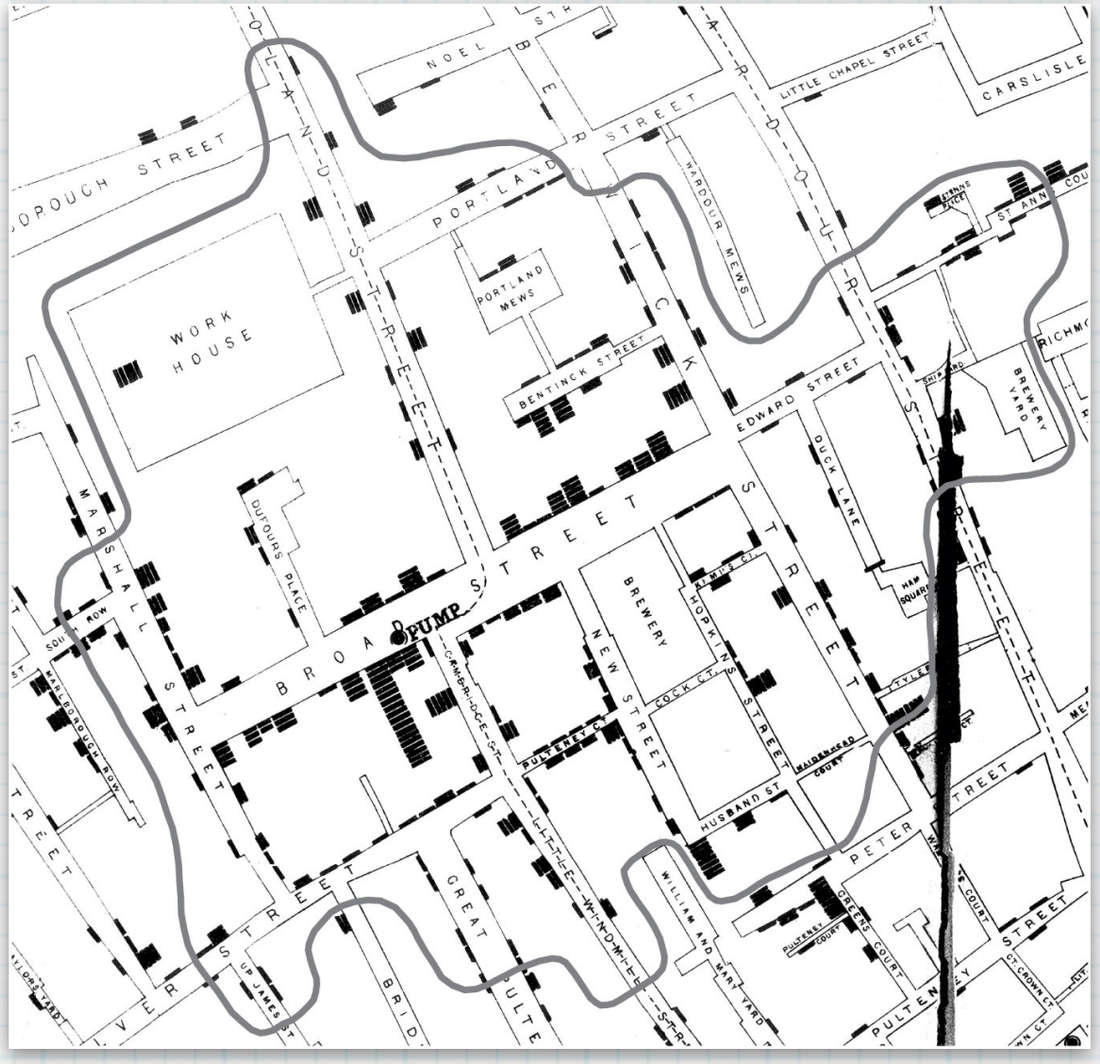


https://en.wikipedia.org/wiki/John_Snow#/media/File:Snow-cholera-map-1.jpg

http://blog.rtwilson.com/john-snows-cholera-data-in-more-formats/

John Snow's Voronoi Map

* Voronoi diagram divides areas according to the closest point in the map.
  * So the area "belongs" to one point
* John Snow plotted Voronoi diagram using **shortest walking distance from each house to each pump**.
* Area that belonged to Broad Street Pump contained most deaths



https://johnsnow.matrix.msu.edu/images/online_companion/chapter_images/fig1 2-6.jpg

Hypothesis and Solution

* He also investigated deaths outside of Broad Street pump voronoi, and found that they had probable causes to drink water from the pump.
  * E.g. walking by the pump when going to market or school
* He also interviewed people living inside the voronoi that did not die.
  * Found that they had their own supply of water.
* Snow presented his findings to London officials, and they **removed the Broad Street pump handle**. The outbreak died down.

Important things to consider...

* At that time, the scientific community **believed that Cholera was spread by miasma ("Bad Air")** from dirty things, e.g., trash or sewage.
* John Snow believed that **Cholera were transmitted by water** and had written a paper about it before. (**Domain Knowledge!!**)
* He tested water from Broad Street pump, but his instruments showed that nothing was wrong.
* London officials did not believe his hypothesis, but took action based on his data analysis.

Epilogue

* In 1884, Robert Koch finally isolated Vibrio Cholerae, a bacteria that causes Cholera, and showed that it is transmitted by water.
* Analysis of John Snow's 1855 publication shows that his model of the disease was substantially complete and correct.
* John Snow is considered the Father of Modern Epidemiology.
* In 1854, Filippo Pacini discovered Vibrio Cholerae, but his publication was ignored due to the strong belief of miasma theory at that time.

Modern Data Science

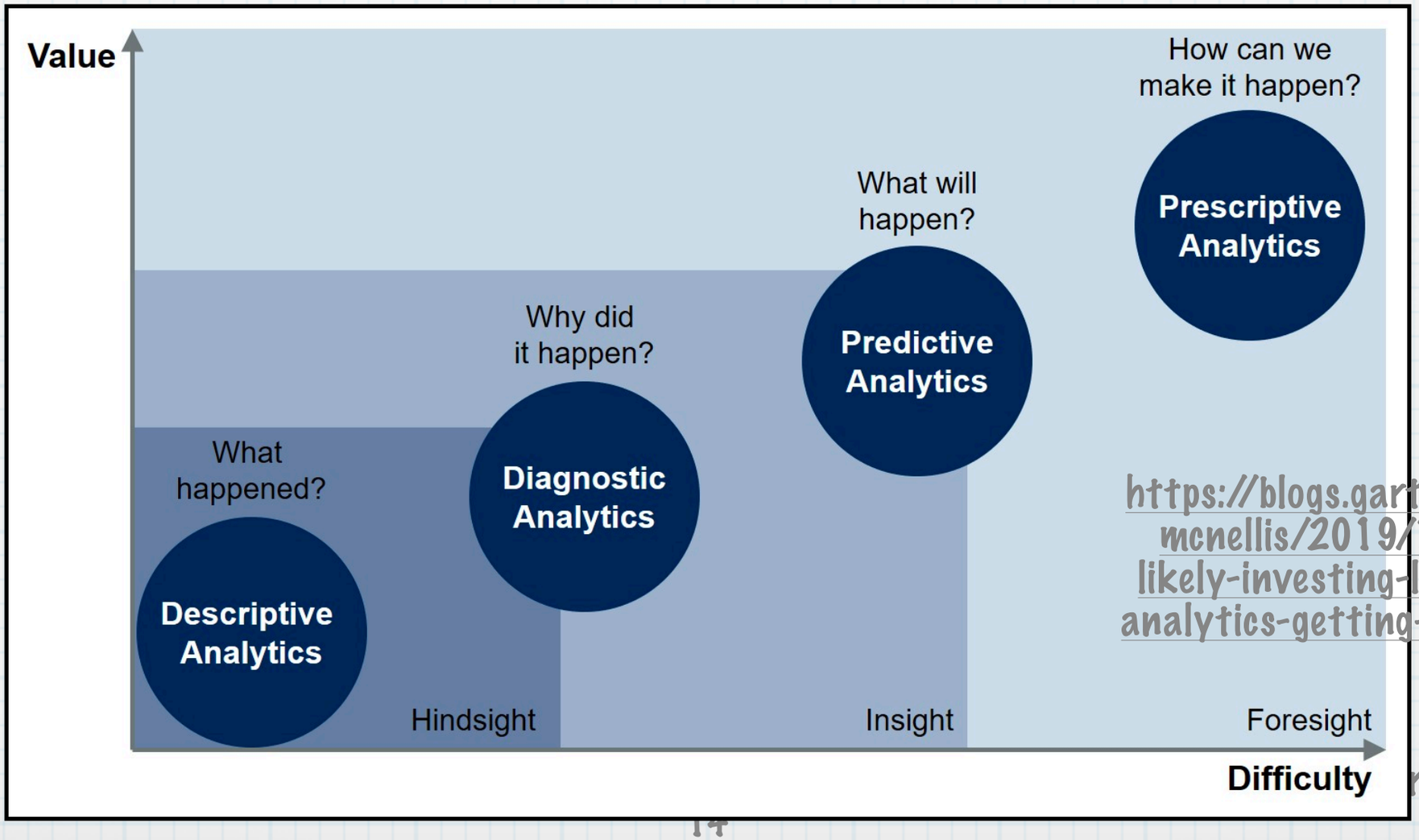
- Use data from various sources
  - Data you collected yourself: manual, using sensors, or logging
  - Data from other departments in the same company
  - Data collected by other companies or organization
  - Data scraped from the Web
  - Data collected using APIs (e.g. Twitter, Google Places)
- Data may not originally collected for your use
  - Not in a format that you can readily use
  - Not intended to answer **your** question

Modern Data Science

* Relies on Statistical techniques to analyze data
  * Required basic-to-intermediate understanding of Statistics and methods
  * Is there bias in data? Which technique is the most appropriate in this case? What are the processes to apply that technique? How to interpret results?
* Utilizes tools (software) to do data science
  * Collect, clean, store data
  * Implement and run Statistical techniques
  * New machine learning techniques for data modeling
  * Visualize data and resulting analysis
  * Handle "Big Data"

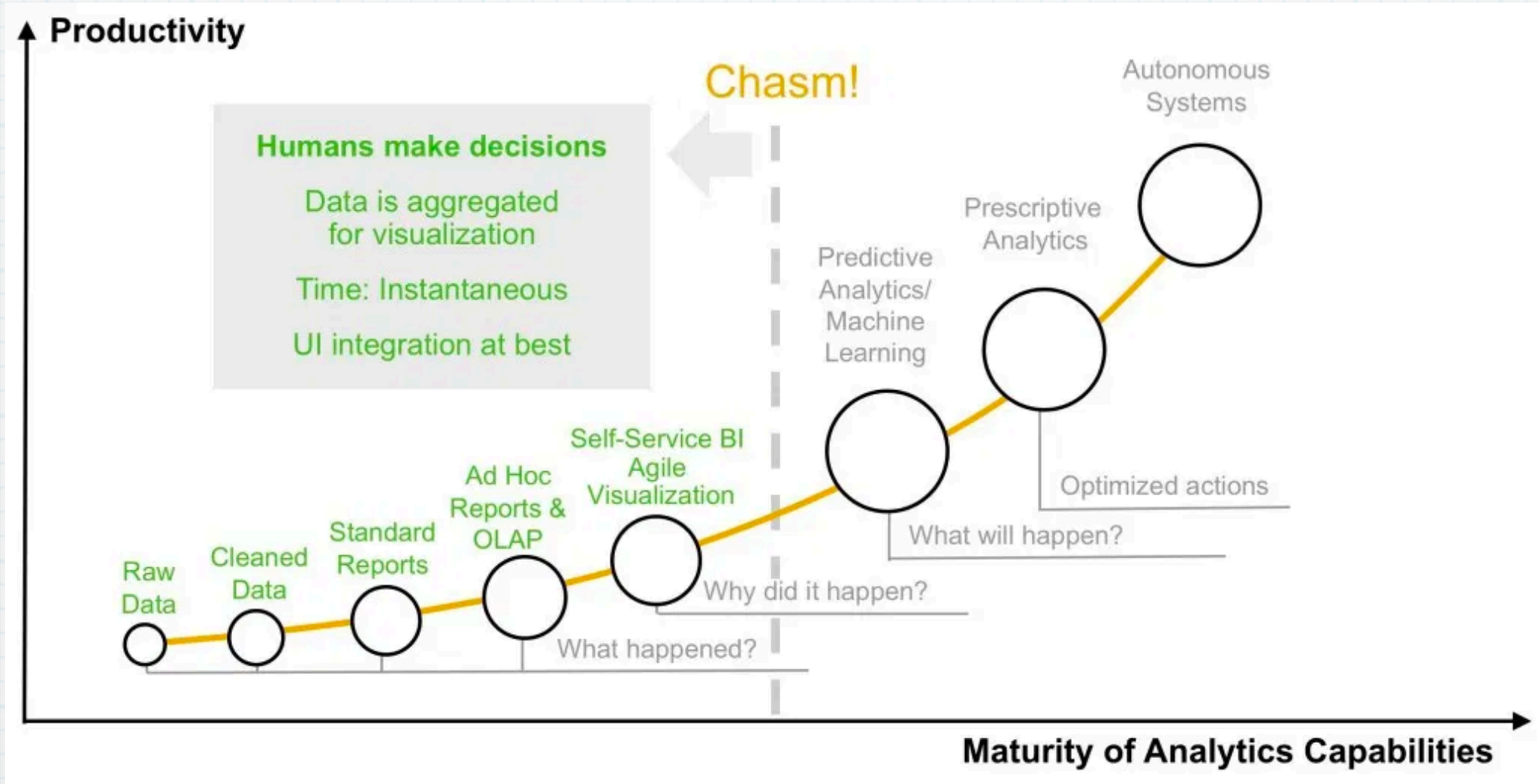
Data Analytic Maturity Model

* How sophisticate are you analyzing and using data?
* Gartner's 4 types of analytics



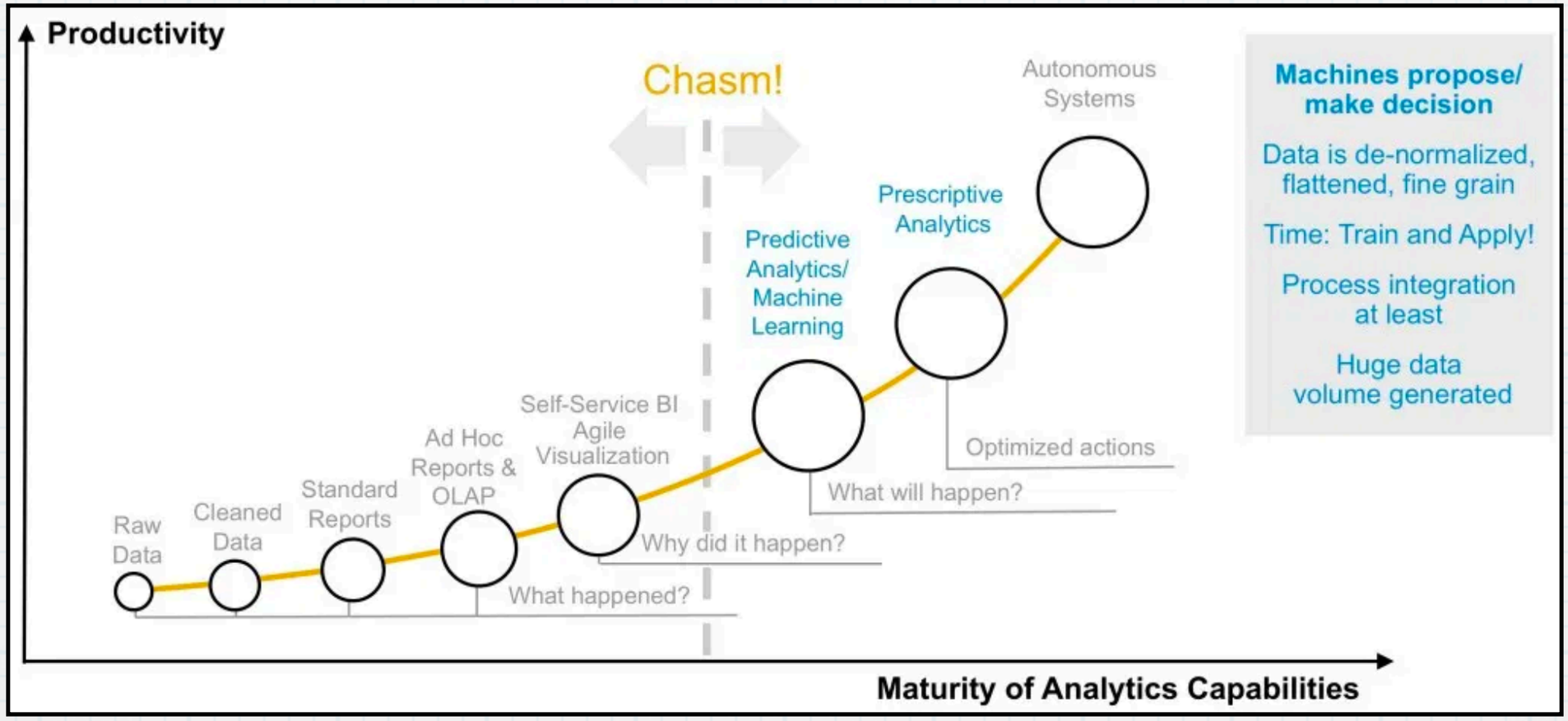
https://blogs.gartner.com/jason-mcnellis/2019/11/05/youre-likely-investing-lot-marketing-analytics-getting-right-insights/

* **Descriptive**: what happened? hindsight
  * total sales year-to-date, number of customers
* **Diagnostic**: why it happened? more insight
  * sales by customer segments and regions



https://timoelliott.com/blog/2018/04/predictive-is-the-next-step-in-analytics-maturity-its-more-complicated-than-that.html

* **Predictive**: what will happen?
  * sales next quarter, number of new customers
* **Prescriptive**: how can we make it happen? optimized action
  * system predicts, humans act
  * best allocation of fund to advertisement channels to increase number of new customers



https://timoelliott.com/blog/2018/04/predictive-is-the-next-step-in-analytics-maturity-its-more-complicated-than-that.html

Example: Descriptive

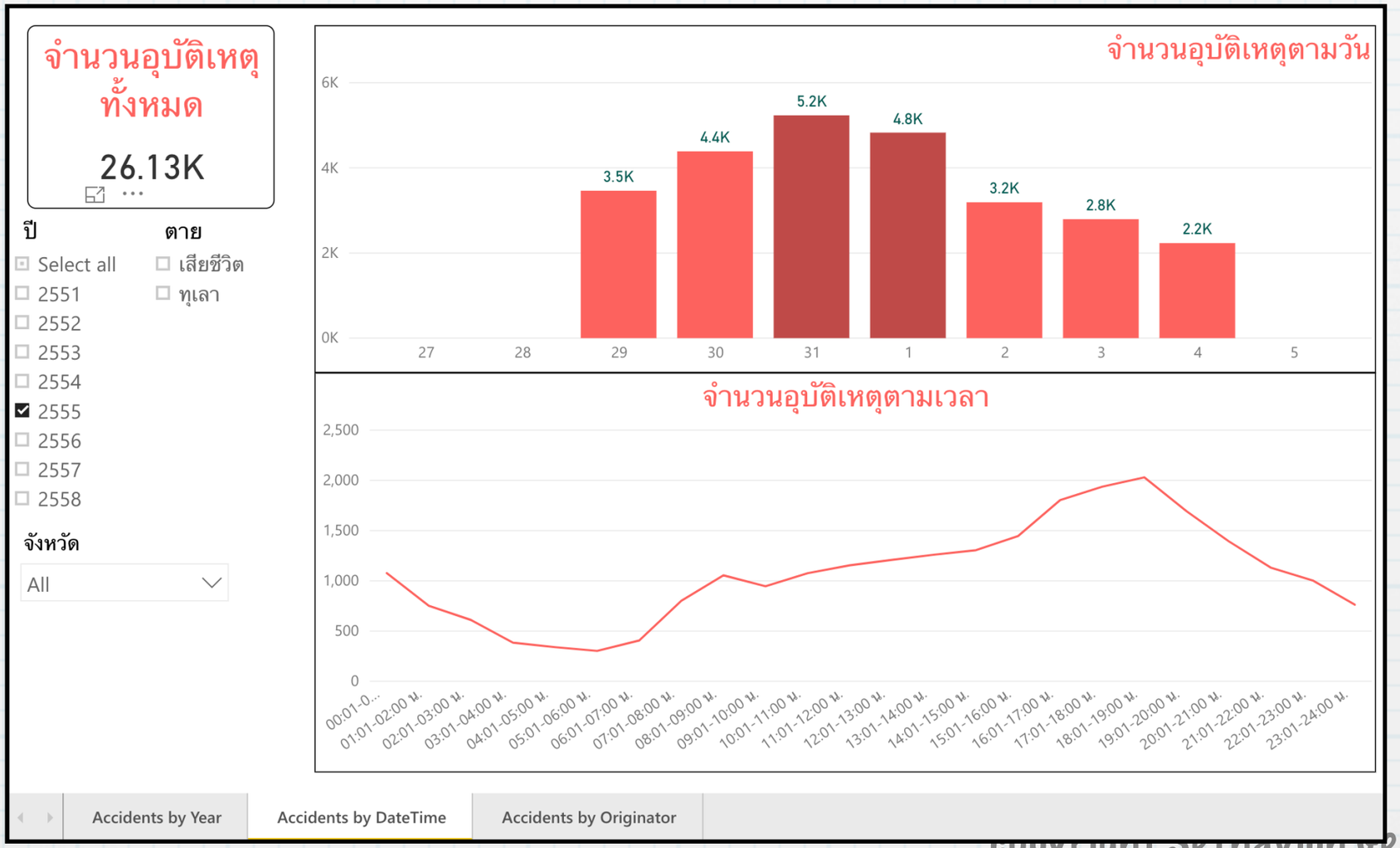
* ad-hoc or planned reports showing KPI of interests
* summarizing and/or segmenting data

รายงานจำนวนผู้ยืนยันสิทธิ์เข้าศึกษาผ่านระบบ TCAS มหาวิทยาลัยเกษตรศาสตร์ ประจำปีการศึกษา 2563 จำแนกตามรอบและสาขาวิชา เทียบกับแผนกลั่นกรอง

วิทยาเขต คณะ และสาขาวิชา	แผน	TCAS 1		TCAS 2			TCAS 3			TCAS 4			รวมทุกรอบ			%ยืนยันจาก
	กลั่นกรอง	แผนรับ	ยืนยัน	แผนรับ	ปรับเพิ่ม	ยืนยัน	แผนรับ	ปรับเพิ่ม	ยืนยัน	แผนรับ	ปรับเพิ่ม	ยืนยัน	แผนรับ	ปรับเพิ่ม	ยืนยัน	แผนกลั่นกรอง
รวม เกษตรศาสตร์	16,887	6,911	3,370	5,951	8,330	2,666	7,205	11,098	5,231	6,533	10,324	5,350	26,571	36,663	16,617	98.4
รวม บางเขน	8,339	1,718	1,008	2,045	2,287	1,135	3,484	4,746	3,104	2,418	3,859	3,201	9,636	12,610	8,448	101.3
เกษตร	455	75	75	182	184	86	150	242	143	48	158	144	426	659	448	98.5
คหกรรมศาสตร์	40	5	0	15	15	4	15	29	21	5	17	16	40	66	41	102.5
เคมีการเกษตร	35	6	4	14	17	9	11	18	15	5	14	13	36	55	41	117.1
วิทยาศาสตร์เกษตร	195	25	33	82	81	33	80	116	75	10	45	42	197	267	183	93.8
เกษตรเขตร้อน	25	4	11	12	12	3	10	12	8	5	16	16	31	44	38	152.0
เกษตรเขตร้อน (นานาชาติ)	40	20	6	0	0	0	2	15	0	3	5	0	25	40	6	15.0
การจัดการศัตรูพืชและสัตว์	30	4	1	11	11	2	10	23	5	5	26	22	30	64	30	100.0
สัตวศาสตร์อุตสาหกรรม	30	3	8	15	15	17	7	7	3	5	15	15	1	40	43	143.3
อาหาร โภชนาการ และการกำหนดอาหาร	35	4	6	16	16	13	10	10	8	5	5	5	35	35	32	91.4
เทคโนโลยีระบบเกษตร	25	4	6	17	17	5	5	12	8	5	15	15	31	48	34	136.0

Example: Diagnostic

* self-service visualization for users to drill and slice
* cross-tabbing data from several reports



Example: Predictive

* **Question**: "Can we identify which customer is in her **2nd trimester of pregnancy without them telling?**"
* To: "send her specialized coupons for things she need before pregnancy"
* Why?: "Studies showed that during the hectic time of baby's birth, parents break their previous shopping habits and **buy everything from one store**."
* If we can capture the clients at this time, they will come back for years.

Target

- Data:
  - Unique ID for each customer
  - Things customer buy tied to that ID
  - Demographic information (bought from other company)
  - Customer who uses Target's own baby registry
- Result: Predictive Model telling which customer is likely pregnant and due date
  - 25 key products including vitamin supplement and unscented lotion
- Target's Revenue: 44B (2002) to 67B (2010)

"Take a fictional Target shopper named Jenny Ward, who is 23, lives in Atlanta and in March bought cocoa-butter lotion, a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug. There's, say, an 87 percent chance that she's pregnant and that her delivery date is sometime in late August"

Example: Prescriptive

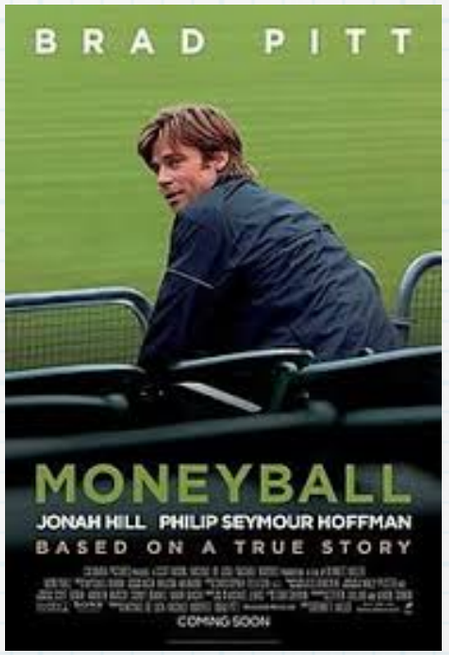
* 99,000 vehicles, 16 millions packages a day
  * Different requirements for each package
* Question: "Can we save fuel cost and reduce CO2 emission?"
* Data: collected from UPS vehicles and handheld devices
* Result: optimization for delivery route
  * System picks the order of delivery and route
  * Save 1.5 million gallons of fuel, $50M (for 10,000 drivers only in 2013)

UPS

* Insight about saving fuel
  * Use right vehicle for right job (use smallest vehicle possible)
  * Consolidate trips (don't go out twice, park-and-walk between nearby places)
  * Try not to make left-hand turn (time-consuming, waste fuel, more risk)
* The system does not take into account (yet)
  * Traffic
  * Weather
  * Problems along the way

Example: Moneyball

* Use Statistics to find undervalued players
* Previous method: gut instinct of scouts
  * Player's posters, handling of bats, etc.
* Statistics used
  * On-base Percentage, Slugging Percentage
  * Earned Run
* Result:
  * 2006 Season: Oakland's A ranking 5/30, spending 24/30
* Statistics used in baseball: https://www.fangraphs.com/



Sports

Athletes to analysts: How big data gave the German football team a leg up

Saheli Roy Choudhury

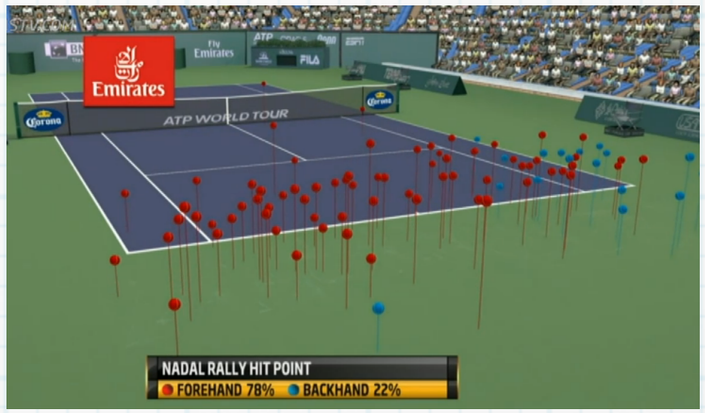
Published 12:39 AM ET Thu, 7 July 2016 | Updated 8:31 PM ET Thu, 7 July 2016

CNBC

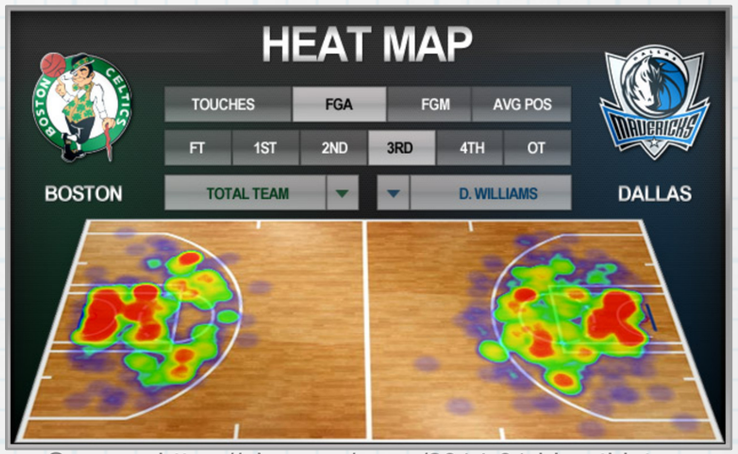
Mehdi Fedouach | AFP | Getty Images

Source: https://www.cnbc.com/2016/07/07/euro-2016-sap-and-german-football-team-worked-to-develop-big-data-analytics.html

Source: http://www.lemonde.fr/coupe-du-monde/article/2014/06/12/big-data-football-club_4432507_1616627.html

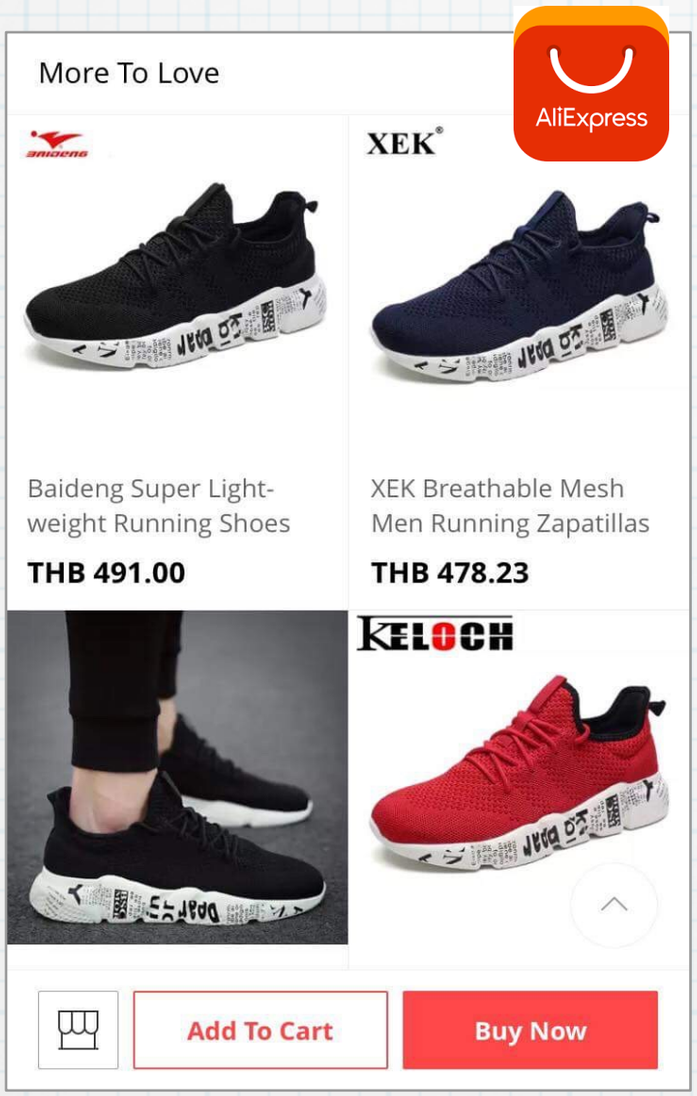


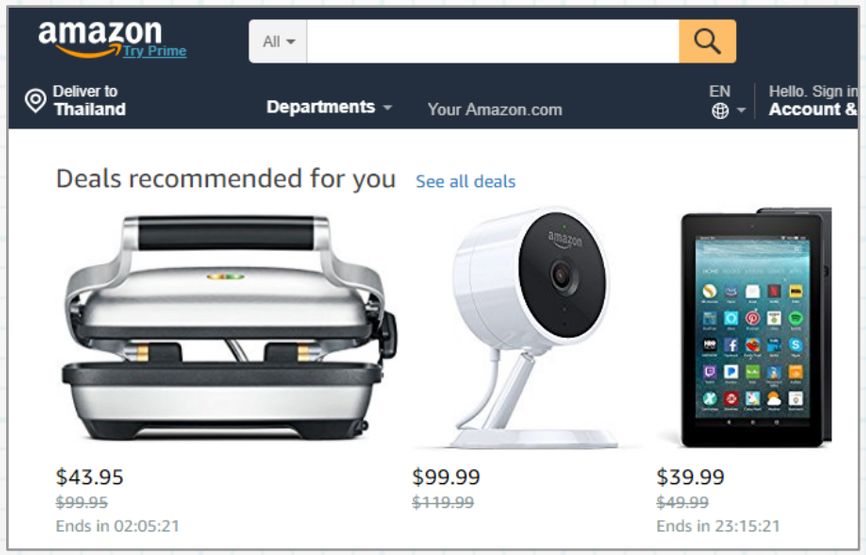
Source: https://www.mikenwjames.com/craig-oshannessy-tennis-strength-is-in-numbers-by-alessandro-mastroluca/

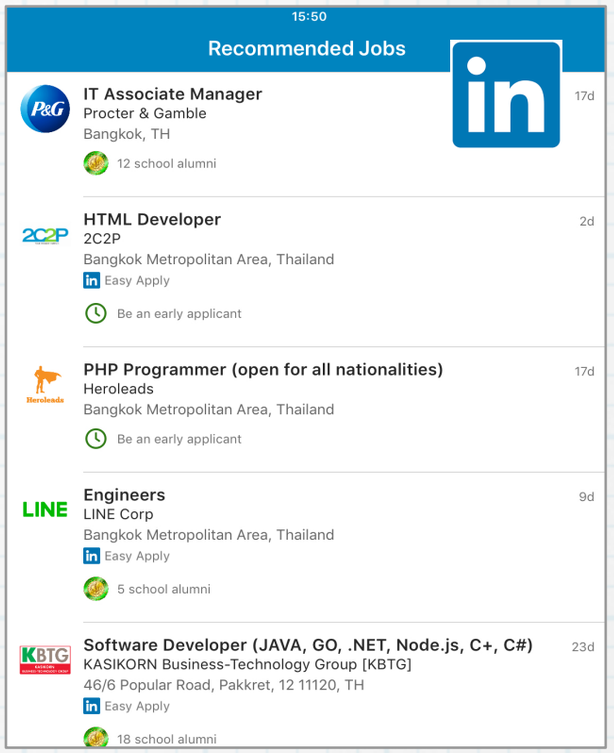


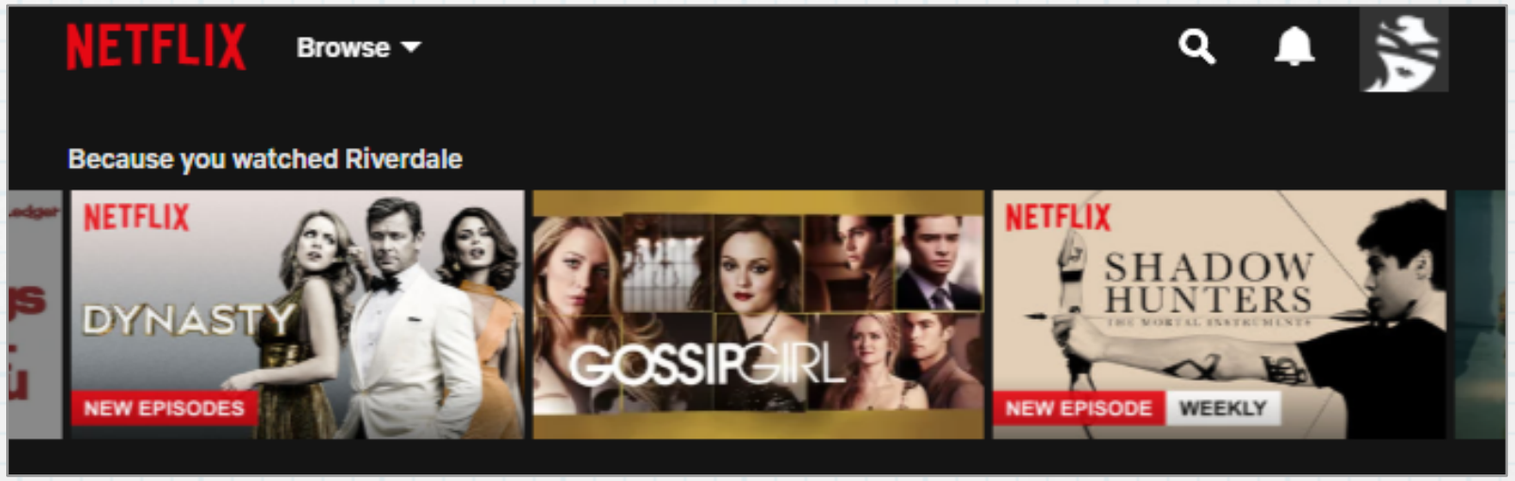
Source: https://phys.org/news/2014-01-big-athletes-edge.html

Recommender Systems



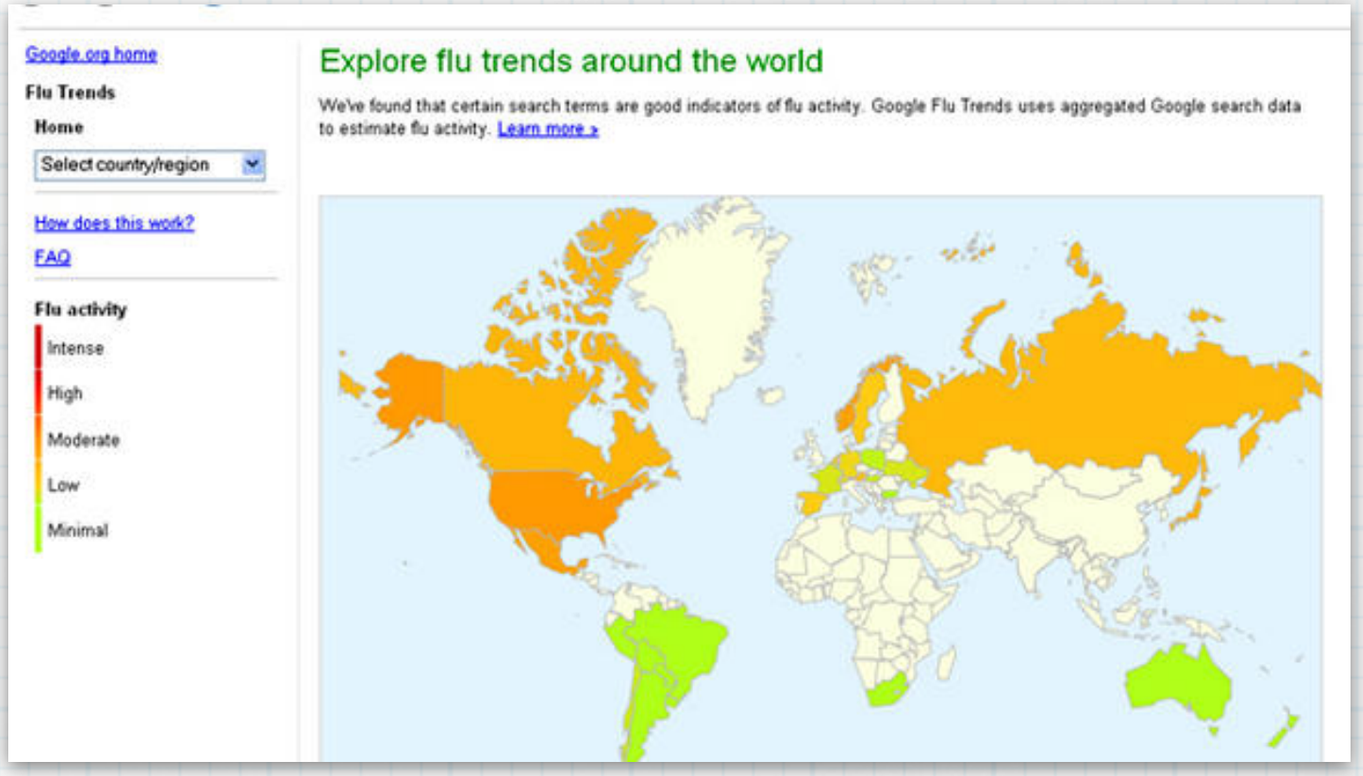






Failure Case: Google Flu Trends

* Use keywords and IP addresses from google search to predict flu outbreak
* Compared with CDC data
* Created log regression model to predict flu outbreak from 45 keywords
* Can predict flu outbreak in US 2 weeks before CDC in 2010



https://www.cbsnews.com/news/google-helps-doctors-track-flu-season-how/

Why Google Flu Trends fails?

* Sometimes very inaccurate
  * Consistently overestimated flu cases during 2011-2013
  * Predicted twice the number of flu cases in 2012-2013
* Reasons
  * Search terms related to flu but not flu, e.g., "fever", "cough"
  * Search terms influenced by other source, e.g., flu news
  * The model was too simple

Failure Case: Target

* "Pregnancy prediction" led to PR disaster
* "We are very conservative about compliance with all privacy laws. But even if you're following the law, you can do things where people get queasy."

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, FORBES STAFF
Welcome to The Not-So Private Parts where technology & privacy collide FULL BIO

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT +4.88%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on

TARGET

Target has got you in its aim

Data Science v.s. Statistics

* Many data science activities are based on Statistics
* Statistics concerns with theory, such as biasness or robustness
* Applied Statistics usually deal with smaller data sets and few data sources

Data Science v.s. AI

* Machine Learning techniques are usually employed to create model from data

* Machine Learning is a subset of AI (Artificial Intelligence)

* New Machine Learning Techniques greatly improve the accuracy of prediction model

Data Science v.s. Big Data

* Big Data is data science with 5Vs
* 5V
  * Volume
  * Velocity
  * Variety
  * Veracity
  * Value
* Needs specialized tools to deal with volume and velocity
* E.g. Hadoop, Spark, Cloud Computing

References

* Wikipedia, John Snow, retrieved from https://en.wikipedia.org/wiki/John_Snow, retrieved on 14 July 2020.
* Jason McNellis, You're likely investing a lot in marketing analytics, but are you getting the right insights?, 5 Nov 2019, retrieved from https://blogs.gartner.com/jason-mcnellis/2019/11/05/youre-likely-investing-lot-marketing-analytics-getting-right-insights/
* Timo Elliot, Predictive Is The Next Step In Analytics Maturity? It's More Complicated Than That!, 20 April 2018, retrieved from https://timoelliott.com/blog/2018/04/predictive-is-the-next-step-in-analytics-maturity-its-more-complicated-than-that.html.
* Kashmir Hill, How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, Forbes.com, 16 FEB 2012.
* John Dix, How UPS uses analytics to drive down costs (and no, it doesn't call it big data), NetworkWorld.com, 1 DEC 2014.
* Billy Beane, Wikipedia.com, retrieved from https://en.wikipedia.org/wiki/Billy_Beane
* Google Flu Trends, Wikipedia.com, retrieved from https://en.wikipedia.org/wiki/Google_Flu_Trends