Analysis Types in Data Science

Feb. 2020

~ Agenda ~

1. Introduction
2. Analytics Types
3. Toward Augmented Analytics

~ Agenda ~

1. Introduction
2. Analytics Types
3. Toward Augmented Analytics

Has anyone heard this word?

Data Scientist will be
the most sexy occupation
in the next 10 years

Hal Varian (2009),
Chief Economist at Google

Is that wrong?

~~Data Scientist~~ will be the most sexy occupation in the next 10 years

Hal Varian (2009),
Chief Economist at Google

Not Data Scientists, but statistians

Statistians will be
the most sexy occupation
in the next 10 years

Hal Varian (2009),
Chief Economist at Google

The New York Times :
「For Today's Graduate, Just One Word: Statistics」

Firms that make data-driven decisions are 5~6% more productive(*)

- 5-6% growth a year,
  you might think small
- But what if the growth will be
  5-6% every year?
- **200% growth in 15 years !!**

  (compound interest $(1.0+0.05)^{15} \sim \times 2.0$)

(*)Strength in Numbers:
How does data-driven decision-making affect firm performance?
(April, 2011)

Investment ROI for data analysis is 13 times(*)

- GAFA invests huge amount of R&D expenses every year

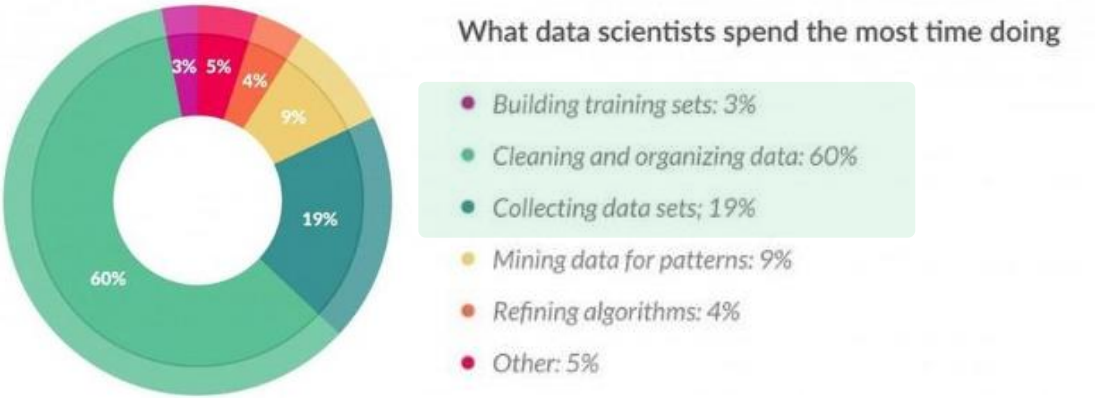
Firms		2018FY	
		R&D expenses [billion$]	R&D expenses /Sales
GAFA	Amazon	29·0	12%
	Google (Alphabet)	21·7	16%
	Apple	14·5	5%
	Facebook	10·0	18%
Japan	Toyota motors	9·0	3%
	SONY	4·5	6%
	HITACHI	2·7	3%

* Securities report of each firm (2018FY), 1dollar=110·5yen

(*) 「Analytics pays back $13·01 for every dollar spent」 September 17, 2014 - Nucleus Research

More than 80% spend time collecting and processing data

- CloudFlower's survey (2016)
- Daily work of Data Scientist:
  - 1. Data collection (19%)
  - 2. Data processing (63%)
  - 3. Model building (13%)
  - Other (5%) (What's?)



~ Agenda ~

1. Introduction
2. Analytics Types
3. Toward Augmented Analytics

Analysis are classified as following three types

- **① Descriptive Analysis (BI Tools):**
  - Use BI tools to understand current situations
- **② Diagnostic Analysis:**
  - Experts use statistical tools to find causes/factors
- **③ Predictive Analysis (AI):**
  - Predict the probability of what will happen in the future, by using ML and DL

BI tools can only give trivial results
① Descriptive Analysis:

- BI tools are used only to monitor and visualize various KPIs
  - Note that unexpected or surprising hypothetical factors cannot be made by BI tools !!

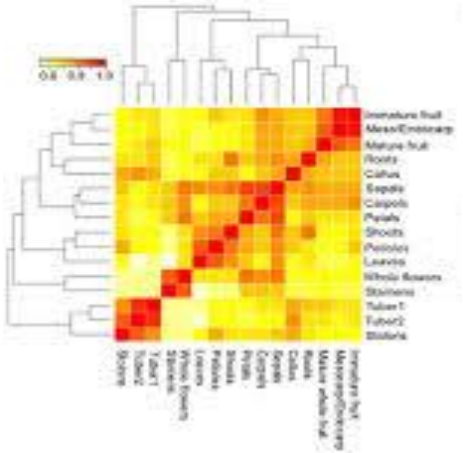
- Decision-makers can understand only "summary" and "visualization", not detail process

DA is to examine data and content to answer the question, "Why did it happen?"

② Diagnostic Analysis (DA):

- DA is characterized by the following techniques
  - Drill-down
  - Data-mining
  - Correlation Analysis
  - …

Correlation Heat Map



- In DA, It is important to come up with ideas for developing hypotheses!
  - Domain knowledge is also required

In DA, the concept of "lift" for averages is basic, but powerful !!

② Diagnostic Analysis (DA):

- It is essentially important to find some segments with high/low **lift** relative to the average
  - Also desirable that #data samples in the segment are sufficient

No	segment	KPI①			KPI②		
		value	average	lift	value	average	lift
1	A	70%	50%	1·40	100	55	1·82
2	B	50%	50%	1·00	55	55	1·00
3	C	30%	50%	0·60	10	55	0·18
⋮	⋮						

ML/DL models are easy to black-box
③ Predictive Analysis (AI):

- Using ML/DL model, we can predict the probability of what will happen
  - Automatically predict if have only dataset
  - AI-Automation can reduce man-hours
- But, we don't know how to improve current situations
  - ML/DL models are black-box
  - Only ML/DL models cannot lead to action

(Summary) Analysis Types

Types	①Descriptive Analysis	②Diagnostic Analysis	③Predictive Analysis
Usage	to understand current situation	to find causes/factors	to predict the probability of what will happen in the future
Tools	BI tools	Statistical tools	Machine Learning or Deep Learning model
Pros	✓ can understand situation without spending time	✓ can find causes/factors	✓ can predict if have only dataset
Cons	✓ BI tools cannot find causes/factors	✓ only experts can execute	✓ easy to black-box

~ Agenda ~

1. Introduction
2. Analytics Types
3. Toward Augmented Analytics

As AI-technology advances, need for citizen data scientists is more increasing

- **Citizen data scientists are "power users"**
  - who can perform both simple and moderately sophisticated analytical tasks that would previously have required more expertise
  - Complementary role to expert data scientists

- **Recently technology has gotten easier for non-specialists to use**
  - BI tools are extending their reach to incorporate easier accessibility to both data and analytics

Augmented Analytics (Gartner)

- In Augmented Analytics,
  the three points are automated
  - 1) Data preparation
  - 2) Insight generation
  - 3) Insight visualization

- So that eliminating the need for
  expert data scientists in many situations

"Explainable AI (XAi)" is also ranked in the top 10 in Gartner

- **Most advanced AI has turned into a complex black box**
  - cannot explain why a particular recommendation or decision was reached
- **XAi is required to ensure accountability and accuracy, fairness, stability, and transparency of decision making**
  - In practice, there are several methods for Xai
    - Explain later…

(Gartner) Top 10 Data and Analytics Technology Trends for 2019

1 Augmented Analytics
2 Augmented Data Management
3 Continuous Intelligence
4 Explainable AI (XAi)
5 Graph Analytics
6 Data Fabric
7 NLP/Conversational Analytics
8 Commercial AI and ML
9 Blockchain
10 Persistent Memory Servers

【Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019 (Feb. 2019)】

Explainable AI in practice

- Understanding the reasons behind predictions is quite important:
  - Case1: Plan to take action based on a prediction
  - Case2: Choose whether to adopt a new model

- Two techniques are typical:
  - (1)LIME (local interpretable model-agnostic explanations)
  - (2)SHAP (SHapley Additive exPlanations)

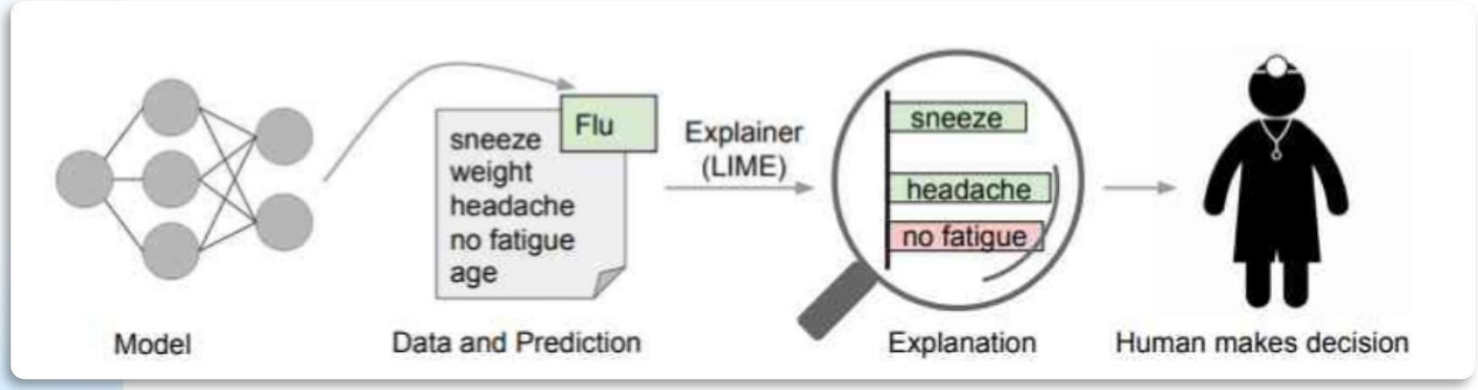
Explainable AI in practice

(1) LIME (Local Interpretable Model-agnostic Explainations):

- **LIME explains why the prediction was made**

  e.g.) Flu prediction:

  - LIME highlights the symptoms in the patient's history that led to the prediction



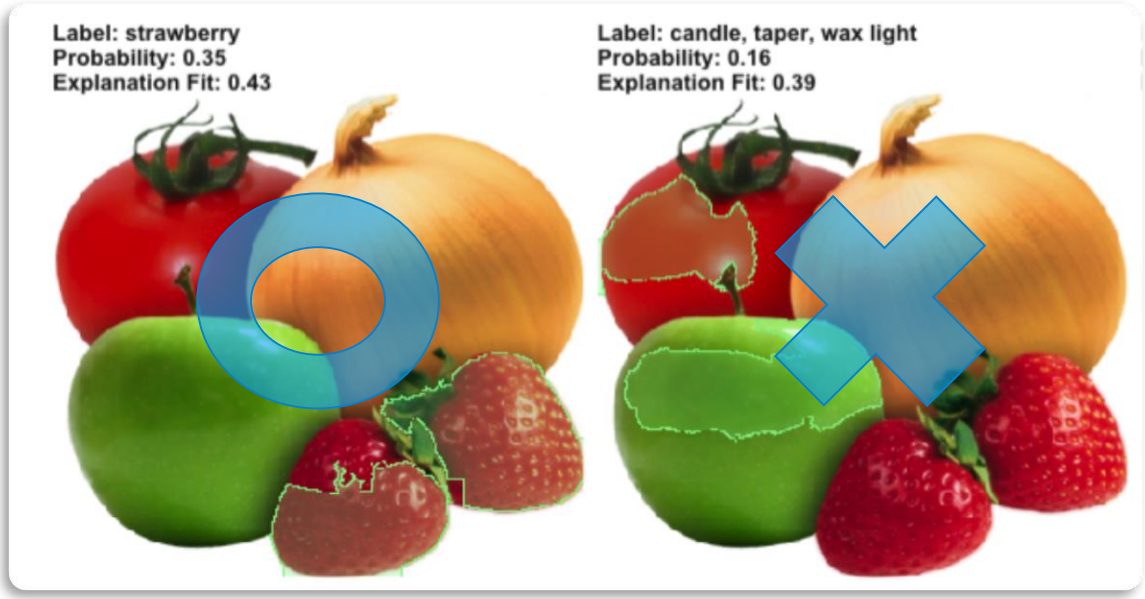
【"Why Should I Trust You?": Explaining the Predictions of Any Classifier】

- **Features of LIME:**
  - Applicable to any ML or DL model
  - Applicable to any data type (text, image, ...)

Explainable AI in practice

(1) LIME (Local Interpretable Model-agnostic Explainations):

- **LIME explains why the prediction was made**
  - Left) Basis for predicting a strawberry is surrounded by a yellow-green area
  - Right) Basis for predicting a candle, taper, wax…



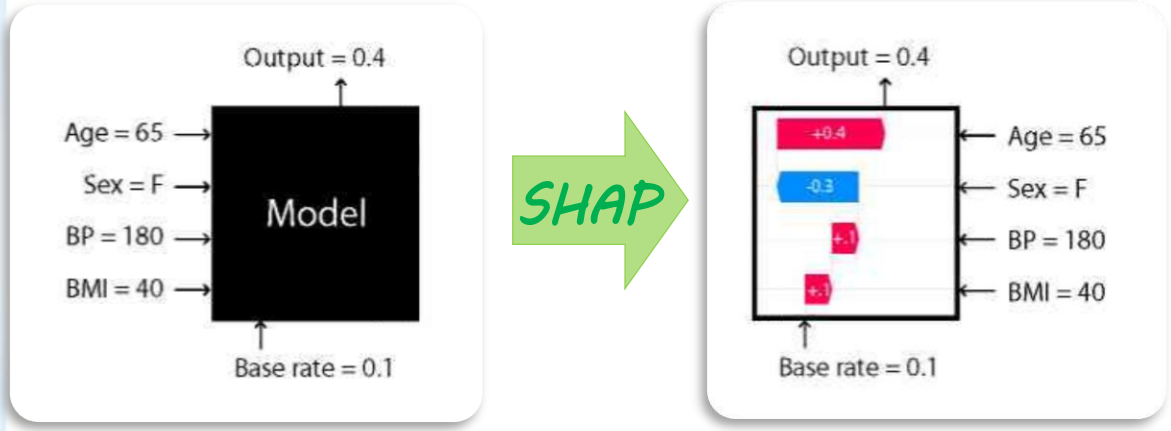
Explainable AI in practice

(2) SHAP (SHapley Additive exPlanations):

- **SHAP** is a game theoretic approach to explain the output of any ML model

SHAP

  - Use the classic "Shapley Values" from game theory



【github·com/slundberg/shap】

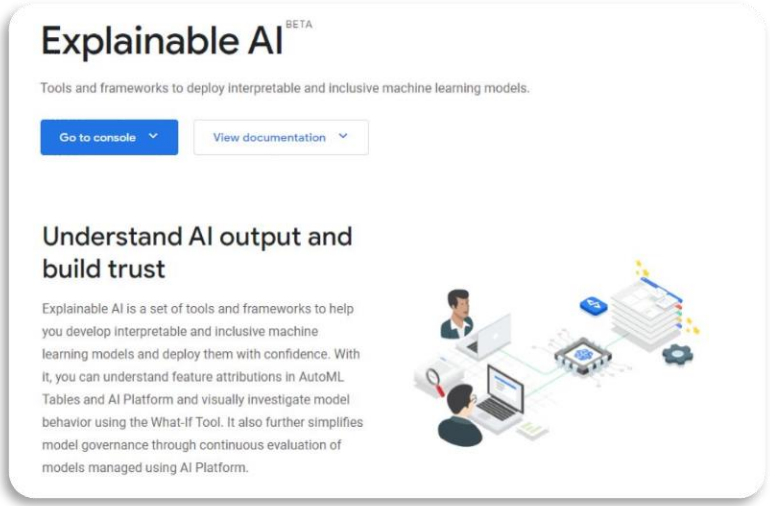
  - Each data coefficient is assigned a score indicating how much the ML model has been affected

Output(0.4) = 0.1(Base rate) + 0.4(Age=65) -0.3(Sex=F)
+ 0.1(BP=180) + 0.1(BMI=40)

Explainable AI in practice

Google Cloud Explainable AI (Beta-version):

- Explainable AI quantifies how each feature in the dataset affected the algorithm-derived results
  - Each data coefficient is assigned a score indicating how much the ML model has been affected



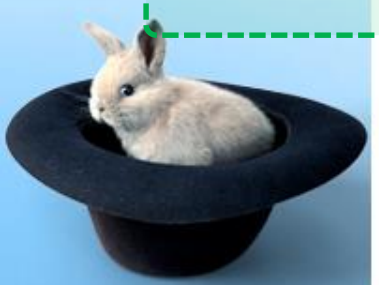
【Google Cloud Explainable AI】

In Augmented Analytics, citizen data scientist will be able to cover ②Diagnostic parts

①Descriptive Analysis	②Diagnostic Analysis	③Predictive Analysis

Expert Data Scientist

Citizen Data Scientist



(References): Global Companies whose mission is to democratize AI

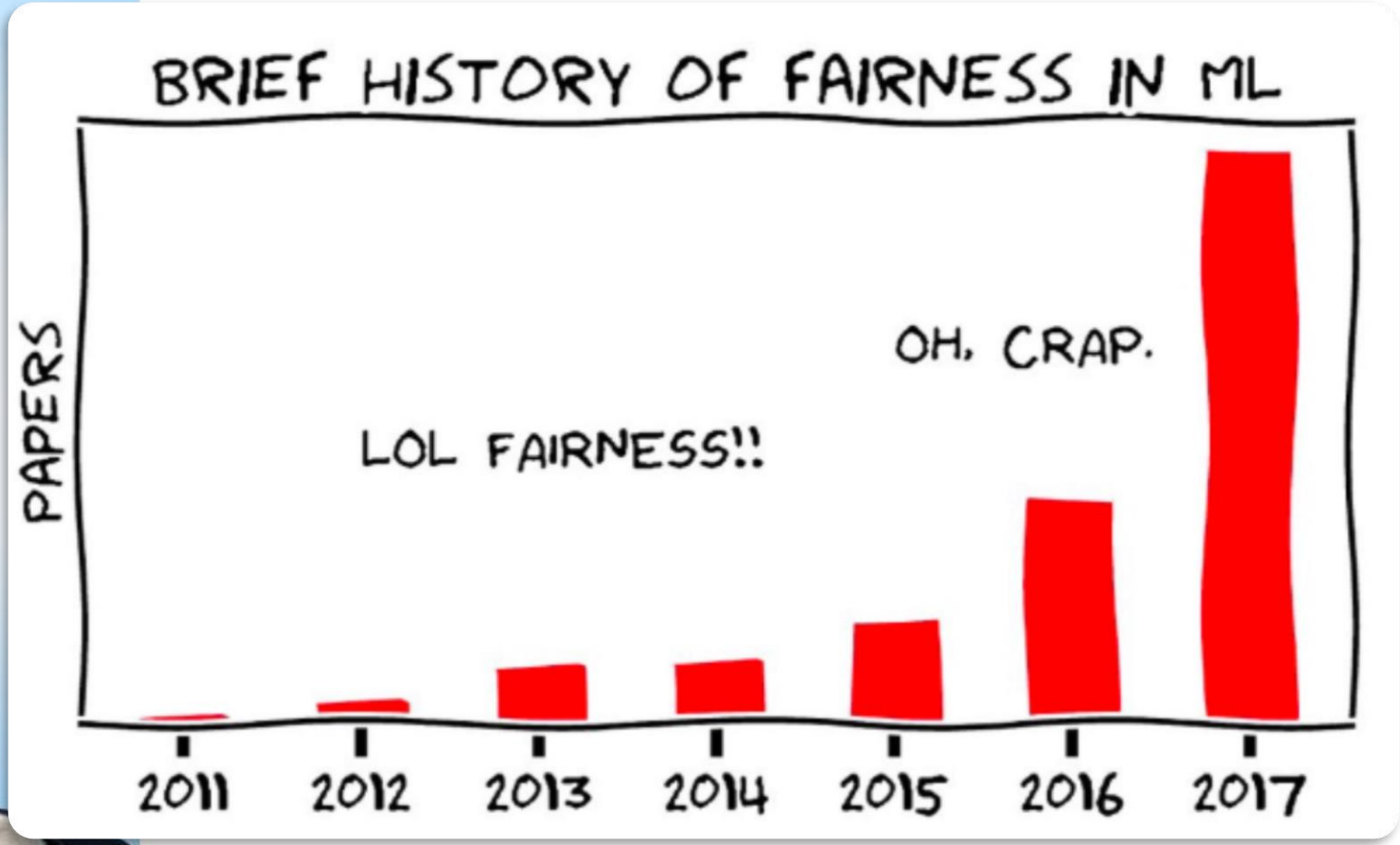
Company Name			Founded	Product Name	Mission	URL (except "https://")
USA	DataRobot, Inc.		2012y	[DataRobot]	Our mission is to change the way businesses all over the world make their most important decisions	www.datarobot.com
	dotData, Inc.	NEC	2018/02	[dotData]	Make An Impact Felt Around The World	dotdata.com
	Feature Labs	Alteryx	2018y	[Featuretools]	Feature Labs is on mission to put machine learning to work	www.featurelabs.com
	H2O.ai		2012y	[Driverless AI]	H2O.ai is Democratizing Artificial Intelligence	www.h2o.ai/company
Japan	DataVehicle Inc.		Nov-14	[dataDiver] [dataFerry]	Making data science familiar. It is our mission	www.dtvcl.com

A topic of "Fairness in Machine Learning" is recently increasing importance

- What's Fairness in ML?
  - "Fairness" refers to preventing disadvantages due to bias in the ML algorithm or learning data

- Cases where fairness should be considered
  1) Lending based on credit scoring
  2) Determining criminal sentencing
     - Bias by sensitive information such as race, gender, region, culture

Papers on Fairness in ML have started to increase about three years ago



【Machine Learning Model Fairness in Practice (Moritz Hardt)】

Ethics Guidelines for Trustworthy AI

European Commission issues ethical guidelines on AI in 2019y:

- The Guidelines list seven key requirements that AI systems should meet in order to be trustworthy:

1) Human agency and oversight
2) Technical robustness and safety
3) Privacy and Data governance
4) Transparency
5) Diversity, non-discrimination and fairness
6) Societal and environmental well-being
7) Accountability

【Ethics Guidelines for Trustworthy AI】

Fairness in Practice
"ML-fairness-gym" (Google 2020):

ex) Lending based on credit scoring

- Depending on goals, the effective threshold will vary

  case1) Aim to maximize profit

  case2) Seek fairness between different groups

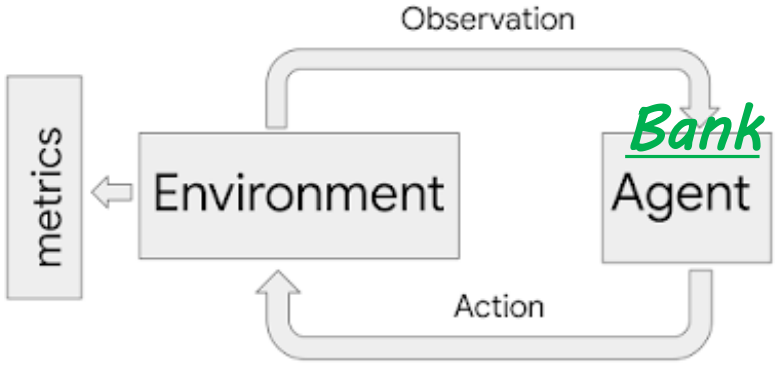
- Focused on short-term goals, it can have unintended and unfair consequences between groups

- So, need to check for unequal disparities in the criteria that the ML system outputs

【ML-fairness-gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems】

"ML-fairness-gym" is a tool for exploring long-term impacts of ML Systems

- "ML-fairness-gym" simulates the result in the following way
  - [Observation]:
    Get necessary data of the loan applicant,
    credit score, and group composition
  - [Action]: Allow or deny the loan application
  - [Metrics]:
    Model whether the applicant
    successfully pays off
    or goes bankrupt and
    adjusts the credit score



【Github:ML-fairness-gym】

Hal Varian's story has a continuation

He is likely to have predicted 2020y situation as of 2009y !!

a) This situation is expected to continue for decades to come

b) Data scientist is so important that elementary school students can learn

c) Statistians are just a few of these jobs

d) It is important that project managers have access to data

Bonus Slide

A "Business Data Scientist" is born by fusion of business and machine learning!

Two warriors fusion to become a stronger warrior!

Business Side ML Engineering

『DBZ (Devil Buu)』

End of Document