

# A!

Aalto University  
School of Business

# Business Analytics

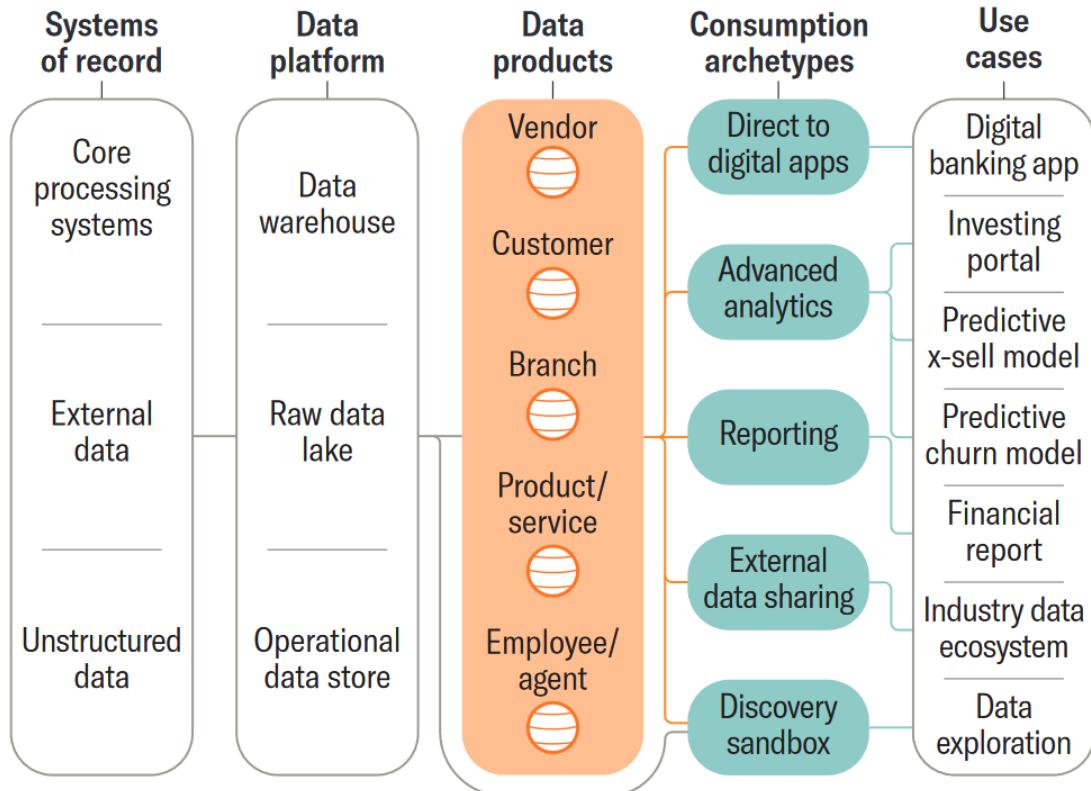
*Lecturer: Assoc. Prof. Yong Liu ([yong.liu@aalto.fi](mailto:yong.liu@aalto.fi));  
37E00500*

# Agenda

- 1. Understand different sorts of analytic techniques**
- 2. Machine learning and AI**
- 3. Analytics and Privacy**
- 4. Business Analytics in Real Life**
- 5. BI for New Businesses**
- 6. Future Competition**
- 7. BI Leadership Qualification**
- 8. Closure of the course**

# Data Product Approach

## Data product approach



In a data product approach, use case teams build solutions by leveraging standardized data products and wiring technologies together following consumption archetype patterns, which reduces work, simplifies the enterprise data architecture, and decreases the time it takes to realize the value.

Companies that treat data like a product can reduce the time it takes to implement it in new use cases by as much as **90%**, decrease their total ownership (technology, development, and maintenance) costs by up to **30%**, and reduce their risk and data governance burden.

# Section 1: Different Analytic Techniques

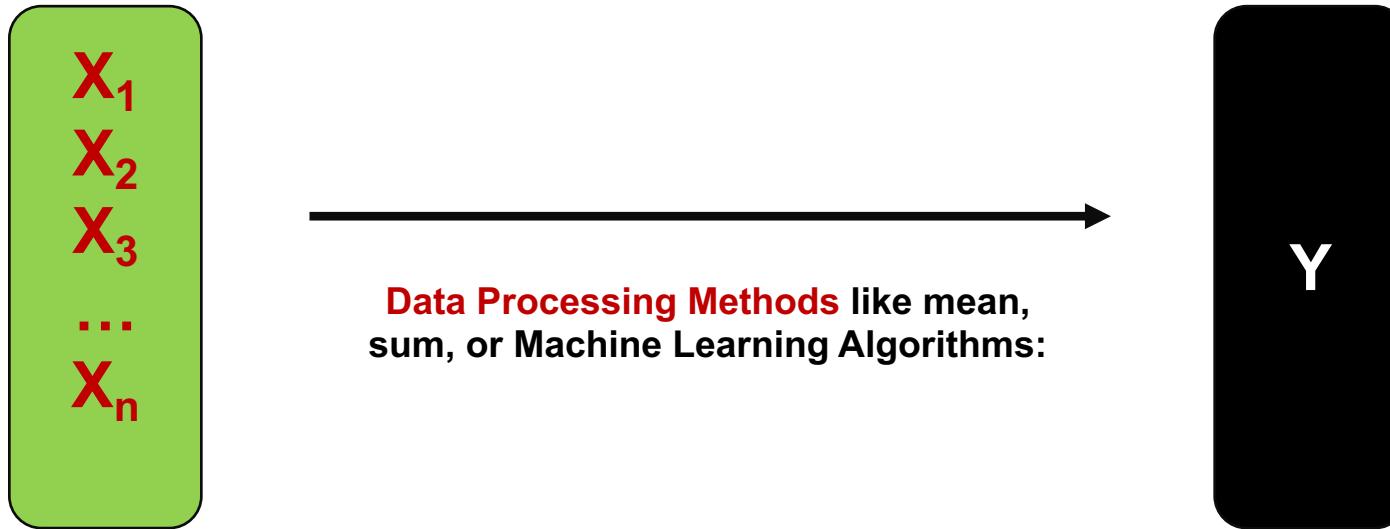


# What is the nature of data analytics?

**Data analytics** is the systematic process of collecting, transforming, and organizing raw data to draw meaningful insights, make predictions, identify patterns, and drive informed decision-making.

- Regression analysis
- Monte Carlo simulation
- Factor analysis
- Cluster analysis
- Time series analysis
- Sentiment analysis
- Machine learning & AI
- Etc.

# Key Components of Data Analytics



**Input:** Independent  
variables

**Variable of interest (VoF):**  
such as Dependent  
Variable

# Depending on Our Interests

New VoF

- **New DV:** A variable does not exist in the raw dataset
- Average, Sum, A/B test, Customer lifetime value analysis; **unsupervised** cluster analysis

Impact of an IV

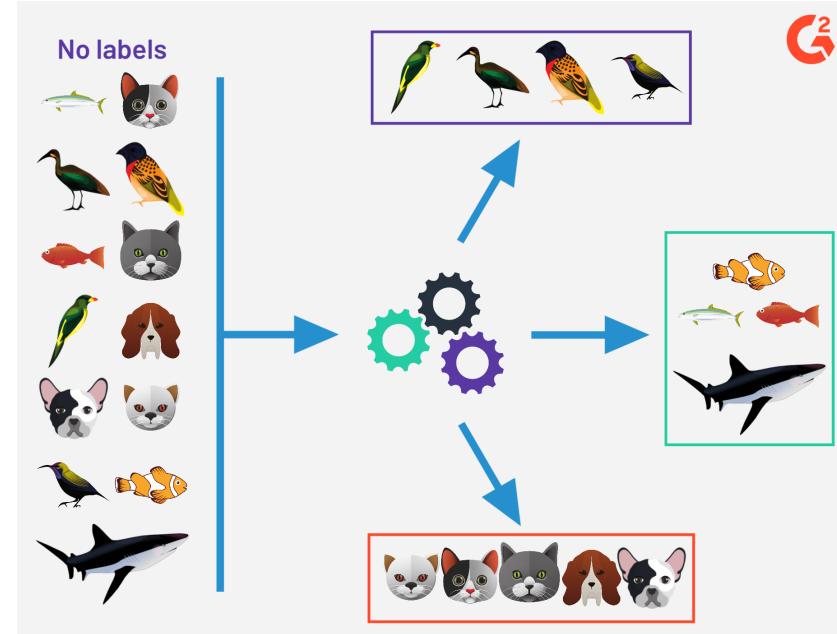
- **Impact of an IV:** normally measured as a path coefficient, computed via using regression methods, like Linear regression, Logistic regression, etc.

New values for  
an existing VoF

- **New values for an existing VoF:** The variable already exists in raw data, like supervised learning

# New VoF

- **New DV:** A variable does not exist in the raw dataset
- Judgement: You add a new column in the data table.
- Average, Sum, A/B test, Customer lifetime value analysis; customer segmentation (e.g., unsupervised cluster analysis)
- Typically associated with the **most useful business insights.**



# Impact of an IV

**Impact of an IV:** normally measured as a path coefficient, computed via using regression methods

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Diagram illustrating the components of a linear regression equation:

- Dependent Variable ( $Y_i$ )
- Population Y intercept
- Population Slope Coefficient ( $\beta_1$ )
- Independent Variable ( $X_i$ )
- Random Error term ( $\epsilon_i$ )

The equation is decomposed into:

- $\beta_0 + \beta_1 X_i$  is labeled as the Linear component.
- $\epsilon_i$  is labeled as the Random Error component.

For instance:

$y <- 5 + 2 * x1 + 3 * x2 - 4 * x3$

```
> linear_model <- lm(y ~ x1 + x2 + x3, data = data)
> summary(linear_model)
```

Call:

```
lm(formula = y ~ x1 + x2 + x3, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.49138	-0.65392	0.05664	0.67033	2.53210

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.01933	0.10734	-0.18	0.858
x1	1.94455	0.11688	16.64	<2e-16 ***
x2	3.04622	0.10946	27.83	<2e-16 ***
x3	-4.05739	0.11223	-36.15	<2e-16 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.052 on 96 degrees of freedom

Multiple R-squared: 0.9621, Adjusted R-squared: 0.9609

F-statistic: 812.9 on 3 and 96 DF, p-value: < 2.2e-16

# Example: Sales Prediction in Retail

Imagine that you work for a retail company and want to understand how advertising spend affects monthly sales for a particular product. In this scenario, multiple linear regression can be a valuable tool.

- **Dependent Variable (Response): Monthly Sales**
- **Independent Variables (Predictors):**
  - Advertising Spend: Amount spent on advertising for the product.
  - Store Size: The square footage of the store.
  - Average Customer Income: Average income of customers in the area.

# Impact of an IV

VoF	Regression Method	Example
Normal Distribution	Ordinary Least Squares (OLS) Regression	Monthly Sales
Binary Variable	Logistic Regression	Customer Churn
Multinomial distribution	Multinomial logistic regression	Customer Preference (Electronics, Clothing, Home Goods)
Count Variable	Poisson Regression	Number of Customer Support Tickets Resolved per Day
Count Variable with lots of zeros	Zero-Inflated Poisson Regression	Insurance Claims Frequency

# Difficulty in implementing different regression models?

VoF	Regression Method	R-code
Normal Distribution	Ordinary Least Squares (OLS) Regression	Model <- <b>lm</b> (outcome ~ Predictor1 + Predictor2, data = data) summary(Model)
Binary Variable	Logistic Regression	Model <- <b>glm</b> (outcome ~ Predictor1 + Predictor2, <b>family = binomial</b> , data = data) Summary(Model)
Multinomial distribution	Multinomial logistic regression	Model <- <b>multinom</b> (outcome ~ Predictor1 + Predictor2, data = data) Summary(Model)
Count Variable	Poisson Regression	Model <- <b>glm</b> (outcome ~ Predictor1 + Predictor2, <b>family = poisson</b> , data = data) Summary(Model)
Count Variable with lots of zeros	Zero-Inflated Poisson Regression	Model <- <b>zeroinfl</b> (outcome ~ Predictor1 + Predictor2, data = data) summary(Model)

# Difficulty in implementing different regression models?

Model <- `lm`(outcome ~ Predictor1 + Predictor2, data = data)

Model <- `glm`(outcome ~ Predictor1 + Predictor2, `family = binomial`, data = data)

Model <- `multinom`(outcome ~ Predictor1 + Predictor2, data = data)

Model <- `glm`(outcome ~ Predictor1 + Predictor2, `family = poisson`, data = data)

Model <- `zeroinfl`(outcome ~ Predictor1 + Predictor2, data = data)

# Section 2: Machine Learning and AI



# ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



- Vision
- Text to Speech
- Motion

## MACHINE LEARNING

Machine learning begins to flourish.



## DEEP LEARNING

Deep learning breakthroughs drive AI boom.



1950's

1960's

1970's

1980's

1990's

2000's

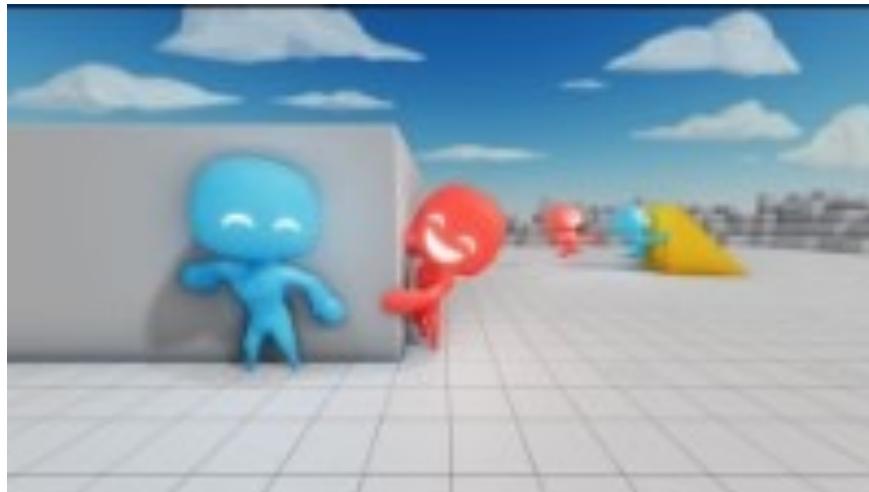
2010's

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

# Machine Learning in the Physical World



- When a Physical robot operates in the real world, there is often no preexisting input data.



<https://www.youtube.com/watch?v=kopoLzvh5jY>



<https://www.youtube.com/watch?v=xAXvfVTgqr0>

# Guess why this autopilot car crashed?

A truck rolled over to its side on the highway, leaving the roof of its box exposed to upcoming traffic. A car owner reportedly driving with some Autopilot driver-assist features didn't see the truck, and the safety feature didn't stop a collision.





You

why autopilot does not perform well in a road with a completely unfamiliar environment



ChatGPT

Autopilot systems, used in autonomous vehicles, rely on various sensors and pre-trained models to navigate and make decisions. When faced with a completely unfamiliar environment, several challenges can impact the performance of autopilot systems:

**1. Lack of Training Data:**

- Autopilot systems are typically trained on diverse datasets that include a variety of road scenarios, traffic conditions, and environmental factors. If the system encounters a situation for which it has not been adequately trained, it may struggle to make accurate predictions.

**2. Limited Generalization:**

- While autopilot systems aim to generalize well to different environments, there are limits to their ability to adapt to entirely new and unforeseen conditions. The lack of familiar patterns can make it challenging for the system to make accurate predictions.

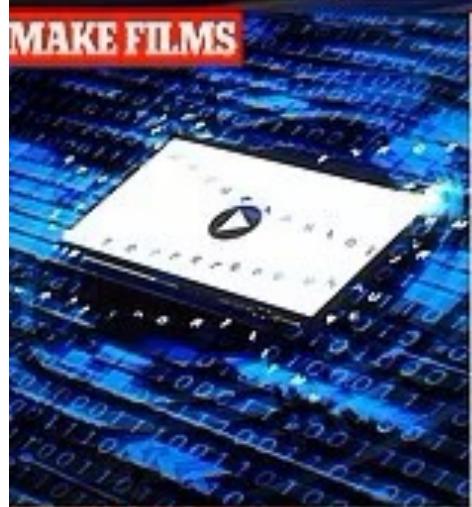
**3. Sensor Limitations:**

The reason why “vehicles on driver-assist features hit stationary objects on the road while at highway speed like that is due to them trying to reduce the number of false-positive braking events.

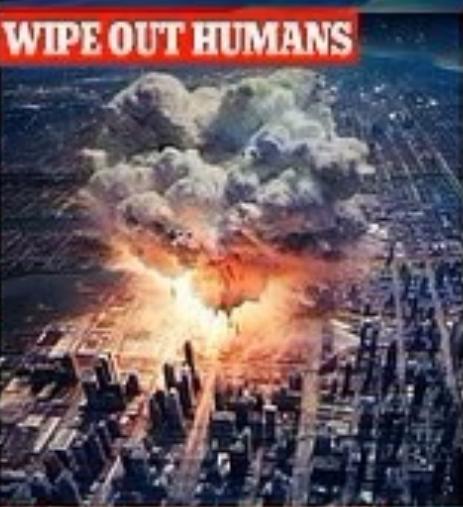
It will improve over time, but it's a good reminder that drivers have to stay attentive and ready to take control at all times”.

# HOW WILL AI CHANGE THE WORLD BY 2030?

MAKE FILMS



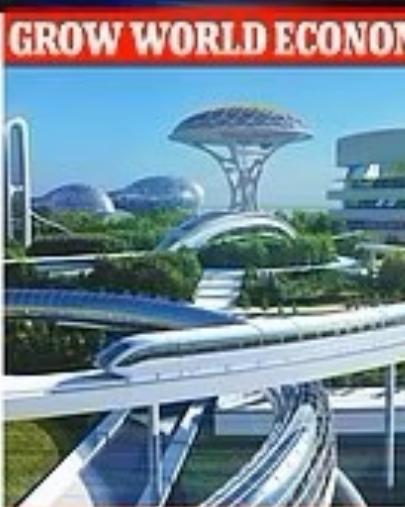
WIPE OUT HUMANS



CARE FOR ELDERLY



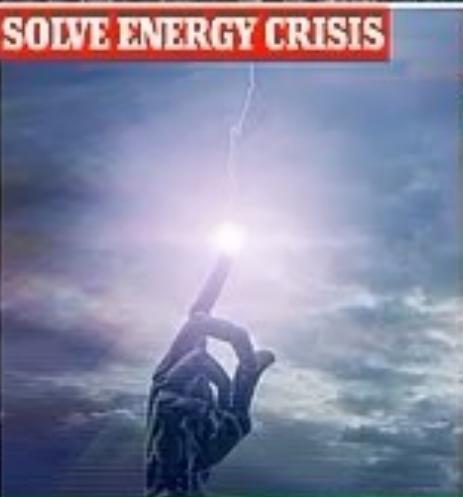
GROW WORLD ECONOMY



TEACH LESSONS



SOLVE ENERGY CRISIS



PREDICT ILLNESS



CLEVER AS HUMANS



# Section 3: Analytics and Privacy



# Facebook likes predict your personalities better than your family members

## PARTICIPANTS' PERSONALITY

Measured using 100-item IPIP Five-Factor Model questionnaire (for 70,520 participants)

User	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
User 1	2.1	4.2	1.9	5.0	3.2
User 2	1.9	2.1	3.2	4.2	4.3
User 3	...	...	...	...	...
(...)	...	...	...	...	...
User X	...	...	...	...	...

90% of participants

10%

- 1 Take personality scores and Likes of 90% of the participants and build linear regression models for the five personality traits using LASSO variable selection

## PARTICIPANTS' LIKES

Obtained from Facebook profiles

User	Running	Ford Explorer	Barack Obama	(...)	Dancing
User 1	1	1	0	-	0
User 2	0	1	1	-	1
User 3	1	0	1	-	1
(...)	...	...	...	...	...
User X	...	...	...	...	...

90% of participants

10%

## LINEAR REGRESSION MODELS

A regression formula with a coefficient for each Like is generated for each of the five personality traits  
e.g. Openness =  $a + \beta_1 * \text{running} + \beta_2 * \text{Obama} + \dots + e$

Regression Coefficients	Running	Ford Explorer	Barack Obama	(...)	Dancing
Openness	.3	.2	0	-	.2
Conscientiousness	.7	.1	.6	-	.7
Extraversion	.1	0	.1	-	.2
Agreeableness	...	...	...	...	...
Neuroticism	...	...	...	...	...

2

- 2 Take the Likes of the remaining 10% of the participants and use the linear regression models to predict scores for the five personality traits

## COMPUTERS' JUDGMENTS

Made using participants' Likes

User	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
User 1	2.1	4.2	1.9	5.0	3.2
User 2	1.9	2.1	3.2	4.2	4.3
User 3	...	...	...	...	...
(...)	...	...	...	...	...
User X	1.9	2.1	3.2	4.2	4.3

90% of participants

10%

Repeat 10 times to make judgments for all participants



Humans' Judgments

Humans' Accuracy



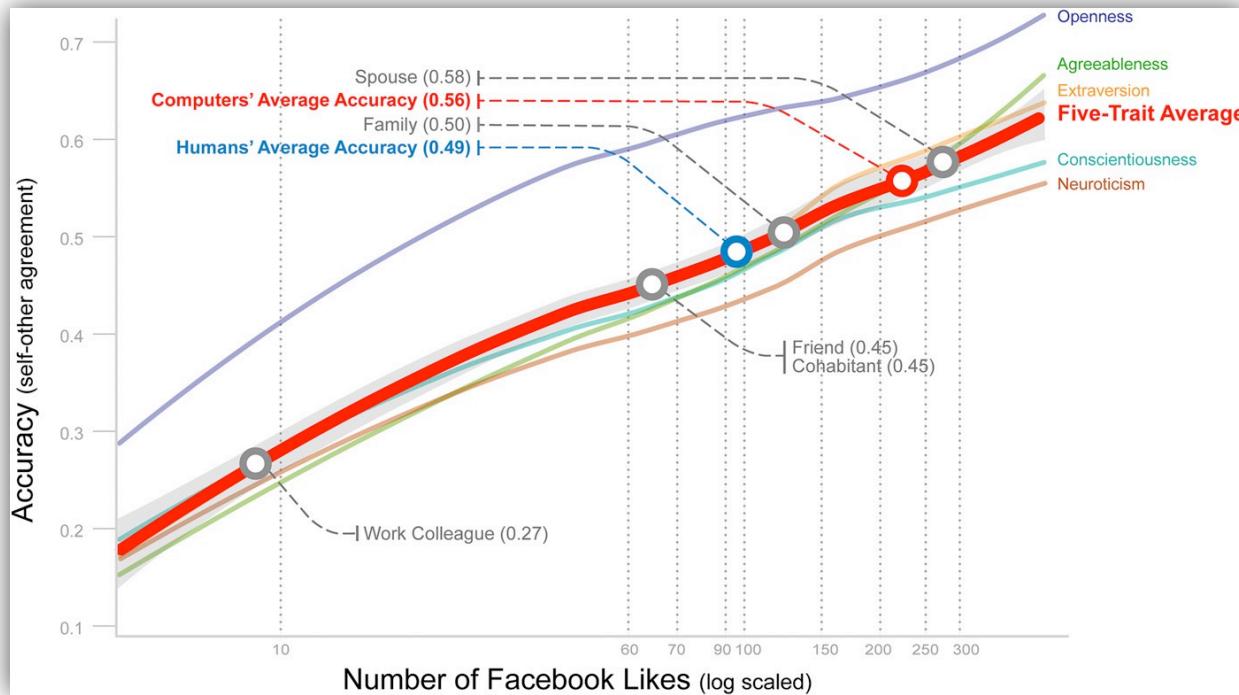
Self-ratings

Computers' Accuracy



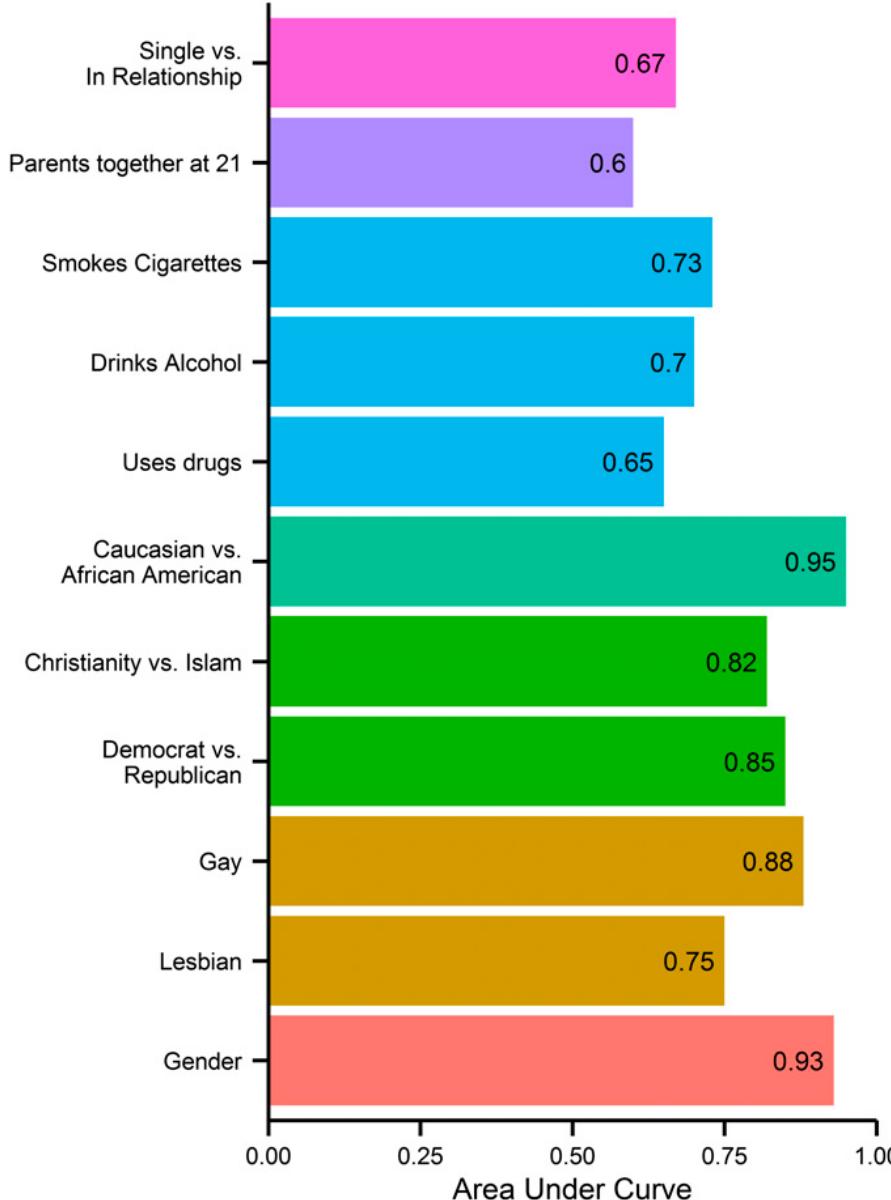
Computers' Judgments

Youyou et al. (2015) Computer-based personality judgments are more accurate than those made by humans, PNAS. 112(4), 1036-1040



Based on a sample of 86,220 Facebook users:

“Computers’ judgments of people’s personalities based on their digital footprints are more accurate and valid than judgments made by their close others or acquaintances” (friends, family, colleagues, etc.).



“Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.”

People are individually less  
unpredictable but  
collectively predictable

# Section 4: Business Analytics in Real Life



# Data Scientist Project in School

- Well-defined business problem
- Well-defined analytic problem
- You know where/how to get data
- All the variables are available
- Data is devoid of error
- Variable distribution does not matter or is in line with expectation.
- Code is prepared, devoid of bug.
- Evaluation/comparison is often unnecessary.
- Deployment is unnecessary.

# Data Scientist Project in Reality

- Business problem is poorly defined.
- Business problem cannot be converted to be a analytic problem.
- Difficulty to get data / data need to be collected.
- Not all the need variables are available – you have computed the variable you needed. Or you have too much data and don't know where to start.
- Data is with different sorts of problems or sometime errors:
  - Data converting problem [e.g., 3G to 3]
  - Outliers
  - Human typing errors
- Variable distribution varies a lot!
- Code / function used has problem / bug.
- Evaluation (model performance) is important but can be underestimated.
- Deployment is often necessary.

# Too much machine learning?

1. Omega-3 supplements
2. Fish oil capsules
3. Krill oil
4. Omega-3 softgels
5. Omega-3 fatty acids
6. DHA and EPA supplements
7. Omega-3 rich foods
8. Vegan omega-3
9. Algal oil supplements

# When analytics get complicated?

1. **Analytics with well-defined independent variables (IV) and dependent variables (DV) are simple.**
2. **Often we don't have well-defined DVs or IVs. This requires a capability or experience that turns a business problem into a data/analytic problem.**

# Case 1: COVID-19 apps

**Case: COVID-19 apps – an app can be developed in e.g., 3 weeks**

## How does it work?

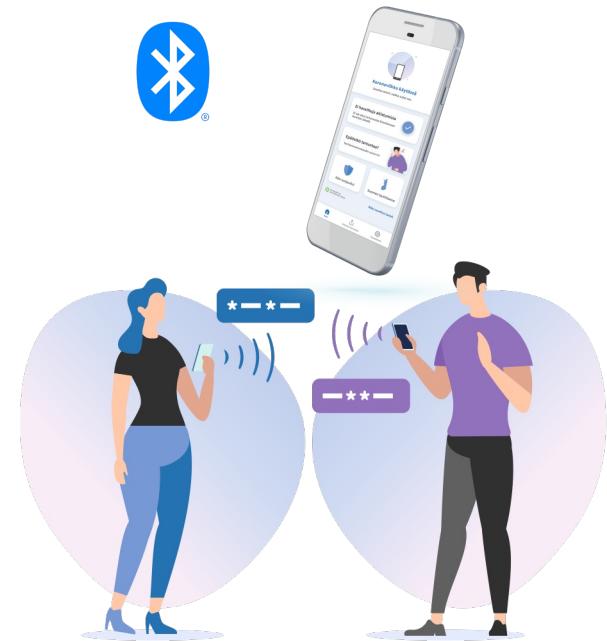
- You install the app and keep the Bluetooth function open. You will receive an alert if you have been exposed to a risky person for, e.g., 15 minutes.

## Data collection Problem

- What data would you need?
- How should the COVID-19 app collect data?

## Data Manipulation Problem:

- If you can collect the data you need, how can you identify the risky person?





# Case of COVID-19 app

## Data collection Problem:

### System input

1. A phone identifies nearby phones by Bluetooth (very mature technology).
2. Communicate with (identified) nearby phones every single (2 or 5) minutes.

Z A ↓	Person1	A Z ↓	Person2	ContactTime
A	B			2021-01-23 12:10:00
A	B			2021-01-23 12:15:00
A	B			2021-01-23 12:25:00
A	B			2021-01-23 12:05:00
A	B			2021-01-23 12:20:00
A	C			2021-01-23 12:25:00
A	C			2021-01-23 12:20:00
A	C			2021-01-23 12:35:00
A	D			2021-01-23 12:30:00
A	D			2021-01-23 12:35:00



# Case of COVID-19 app

Data collection Problem: how to identify risky person?

**Risky person:** those who have exposure to a person with COVID-19 for at least 15 consecutive minutes.

Z A	Person1	A Z	Person2	ContactTime
A		B		2021-01-23 12:10:00
A		B		2021-01-23 12:15:00
A		B		2021-01-23 12:25:00
A		B		2021-01-23 12:05:00
A		B		2021-01-23 12:20:00
A		C		2021-01-23 12:25:00
A		C		2021-01-23 12:20:00
A		C		2021-01-23 12:35:00
A		D		2021-01-23 12:30:00
A		D		2021-01-23 12:35:00

Raw data



As a DS, your task to identify the risky person who are likely affected by the person A who got COVID-19. SO HOW?



Aalto-yliopisto  
Aalto-universitetet  
Aalto University

Z A	Person1	A Z	Person2	ContactTime
A			B	2021-01-23 12:10:00
A			B	2021-01-23 12:15:00
A			B	2021-01-23 12:25:00
A			B	2021-01-23 12:05:00
A			B	2021-01-23 12:20:00
A			C	2021-01-23 12:25:00
A			C	2021-01-23 12:20:00
A			C	2021-01-23 12:35:00
A			D	2021-01-23 12:30:00
A			D	2021-01-23 12:35:00



Person1	A Z	Person2	A Z	ContactTime	FiveMinuteAgo	TenMinuteAgo
A		B		2021-01-23 12:05:00	(NULL)	(NULL)
A		B		2021-01-23 12:10:00	2021-01-23 12:05:00	(NULL)
A		B		2021-01-23 12:15:00	2021-01-23 12:10:00	2021-01-23 12:05:00
A		B		2021-01-23 12:20:00	2021-01-23 12:15:00	2021-01-23 12:10:00
A		B		2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A		C		2021-01-23 12:20:00	(NULL)	(NULL)
A		C		2021-01-23 12:25:00	2021-01-23 12:20:00	(NULL)
A		C		2021-01-23 12:35:00	(NULL)	2021-01-23 12:25:00
A		D		2021-01-23 12:30:00	(NULL)	(NULL)
A		D		2021-01-23 12:35:00	2021-01-23 12:30:00	(NULL)

Raw data

**Risky person:** those who have exposure to a person with COVID-19 for at least 15 consecutive minutes.



**Risky person:** Pairs of persons that has not null values in the variables of **ContactTime**, **FiveMinuteAgo** and **TenMinuteAgo**

DV (in theory)

DV (measurable)

Create table CoronatracingX AS

(Select Person1, Person2, ContactTime,

DATE\_ADD(ContactTime, INTERVAL -5 MINUTE) AS FiveMinuteAgo,

DATE\_ADD(ContactTime, INTERVAL -10 MINUTE) AS TenMinuteAgo

FROM Coronatracing)

Z A↓	Person1	A Z↓	Person2	ContactTime
	A		B	2021-01-23 12:10:00
	A		B	2021-01-23 12:15:00
	A		B	2021-01-23 12:25:00
	A		B	2021-01-23 12:05:00
	A		B	2021-01-23 12:20:00
	A		C	2021-01-23 12:25:00
	A		C	2021-01-23 12:20:00
	A		C	2021-01-23 12:35:00
	A		D	2021-01-23 12:30:00
	A		D	2021-01-23 12:35:00



Person1	A Z↓	Person2	ContactTime	FiveMinuteAgo	TenMinuteAgo
A		B	2021-01-23 12:10:00	2021-01-23 12:05:00	2021-01-23 12:00:00
A		B	2021-01-23 12:15:00	2021-01-23 12:10:00	2021-01-23 12:05:00
A		B	2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A		B	2021-01-23 12:05:00	2021-01-23 12:00:00	2021-01-23 11:55:00
A		B	2021-01-23 12:20:00	2021-01-23 12:15:00	2021-01-23 12:10:00
A		C	2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A		C	2021-01-23 12:20:00	2021-01-23 12:15:00	2021-01-23 12:10:00
B		C	2021-01-23 12:35:00	2021-01-23 12:30:00	2021-01-23 12:25:00
A		D	2021-01-23 12:30:00	2021-01-23 12:25:00	2021-01-23 12:20:00
A		D	2021-01-23 12:35:00	2021-01-23 12:30:00	2021-01-23 12:25:00

**UPDATE CoronatracingX SET FiveMinuteAgo = NULL  
WHERE (Person1,Person2,FiveMinuteAgo) Not IN (select Person1,Person2,ContactTime FROM Coronatracing)**

Person1	A ↓ Z ↓	Person2	A ↓ Z ↓ ContactTime	FiveMinuteAgo	TenMinuteAgo
A	B		2021-01-23 12:05:00	(NULL)	2021-01-23 11:55:00
A	B		2021-01-23 12:10:00	2021-01-23 12:05:00	2021-01-23 12:00:00
A	B		2021-01-23 12:15:00	2021-01-23 12:10:00	2021-01-23 12:05:00
A	B		2021-01-23 12:20:00	2021-01-23 12:15:00	2021-01-23 12:10:00
A	B		2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A	C		2021-01-23 12:20:00	(NULL)	2021-01-23 12:10:00
A	C		2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A	C		2021-01-23 12:35:00	(NULL)	2021-01-23 12:25:00
A	D		2021-01-23 12:30:00	(NULL)	2021-01-23 12:20:00
A	D		2021-01-23 12:35:00	2021-01-23 12:30:00	2021-01-23 12:25:00

**UPDATE CoronatracingX SET TenMinuteAgo = NULL  
WHERE (Person1,Person2,TenMinuteAgo) NOT IN (select Person1,Person2,ContactTime FROM Coronatracing)**

Person1	A ↓ Z ↓	Person2	A ↓ Z ↓ ContactTime	FiveMinuteAgo	TenMinuteAgo
A	B		2021-01-23 12:05:00	(NULL)	(NULL)
A	B		2021-01-23 12:10:00	2021-01-23 12:05:00	(NULL)
A	B		2021-01-23 12:15:00	2021-01-23 12:10:00	2021-01-23 12:05:00
A	B		2021-01-23 12:20:00	2021-01-23 12:15:00	2021-01-23 12:10:00
A	B		2021-01-23 12:25:00	2021-01-23 12:20:00	2021-01-23 12:15:00
A	C		2021-01-23 12:20:00	(NULL)	(NULL)
A	C		2021-01-23 12:25:00	2021-01-23 12:20:00	(NULL)
A	C		2021-01-23 12:35:00	(NULL)	2021-01-23 12:25:00
A	D		2021-01-23 12:30:00	(NULL)	(NULL)
A	D		2021-01-23 12:35:00	2021-01-23 12:30:00	(NULL)

# Case 2: Business Problem vs. Analytic Problem

- You are operating a chain hotel brand.
- You have data about who & when and which hotel a customer visited.
- Nearly all customers leave reviews of your hotel performance, such as service rating, room rating, location rating, etc.
- You want to know how different attribute rating affects their revisit behavior.
- Revisit behavior is coded as 0 (lead to no revisit), and 1 (lead to revise).



# We need to code a new variable of “revisit”, **BUT HOW?**

	Hotel_ID	User_ID	Service_Rating	Location_rating	Hotel_visit_date	Revisit (?)
1	1968430	Emily Johnson	5	4	16/01/2022	
2	1968430	Emily Johnson	4	5	25/02/2023	
3	582794	James Smith	3	5	21/02/2023	
4	582794	James Smith	4	5	01/09/2022	
5	603666	Olivia Davis	4	4	30/03/2022	
6	603666	Olivia Davis	3	5	20/06/2023	
7	603666	Olivia Davis	4	4	20/01/2024	
8	1094260	Ethan Williams	5	5	01/09/2023	
9	1094260	Daniel Harris	4	5	21/11/2023	

Provide specify the correct values for DV of each row, and submit your answers at [presemo.aalto.fi/bic](http://presemo.aalto.fi/bic)

**Revisit behavior is coded as 0 (lead to no revisit), and 1 (lead to revise).**

Example: 01010101010

# Again: Business Problem → Data Problem

	Hotel_ID	User_ID	Service_Rating	Location_rating	Hotel_visit_date	Revisit (?)
1	1968430	Emily Johnson	5	4	16/01/2022	1
2	1968430	Emily Johnson	4	5	25/02/2023	0
3	582794	James Smith	3	5	21/02/2023	0
4	582794	James Smith	4	5	01/09/2022	1
5	603666	Olivia Davis	4	4	30/03/2022	1
6	603666	Olivia Davis	3	5	20/06/2023	1
7	603666	Olivia Davis	4	4	20/08/2023	0
8	1094260	Ethan Williams	5	5	01/09/2023	0
9	1094260	Daniel Harris	4	5	21/11/2023	0

We cannot label all the data by hand, so how can we use a computer to label the data? For instance, how can you use computer language to label the data correctly and automatically?

# Case 3: Reima Sale Prediction

- Predicting product sales seems easy, but real-life situations can complicate things.
- As a BA of the company, we are tasked with presenting a Business Intelligence report that provides recommendations on the **ideal number** of productions of various products for the upcoming year.



Lasten kurahousut Lammikko  
9 väriä  
29,95 €



Lasten vedenpitävä välikausihalari  
Kapelli  
5 väriä  
99,95 €

**Sales volume:** The number of clothes products sold in one year.



Taaperoiden softshell-haalari Mjosa

8 väriä

59,95 €



Taaperoiden vedenpitävä  
kevyttopahaalari Marte Mid

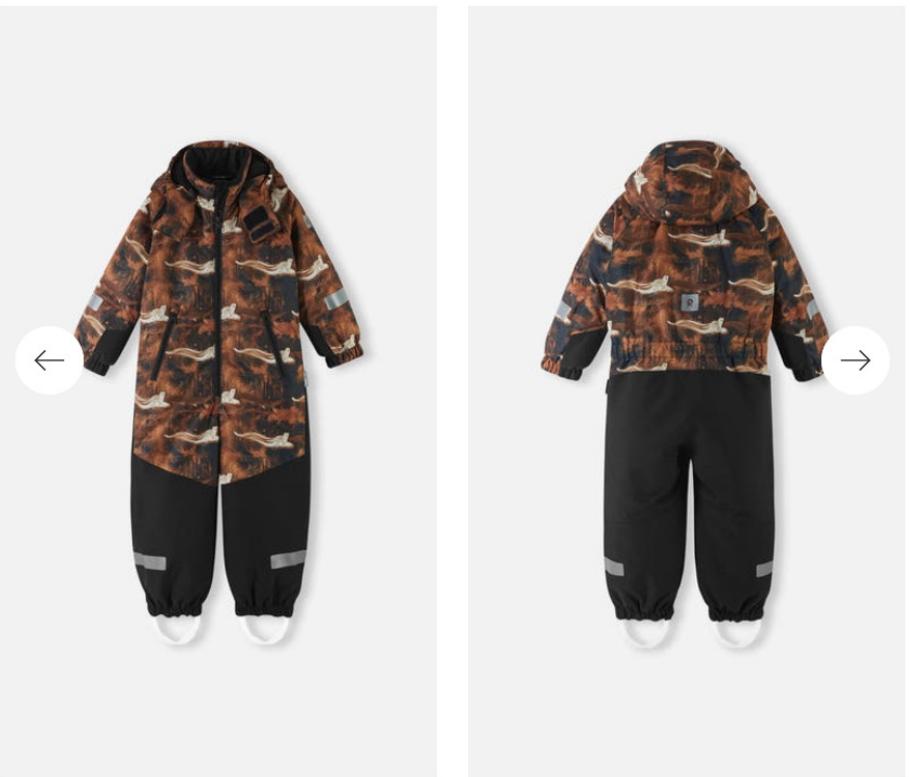
8 väriä

99,95 €

## Sales volume

How about sales of  
products on a global  
sale?

# Should we consider the difference in the sales by colors?



## Lasten vedenpitävä toppahaalari Kurikka

83,97 €

Alin hinta viimeisten 30 päivän aikana: 83,97 €

★★★★★ 4.6/5 (53)

SAATAVILLA 6 VÄRIÄ: CINNAMON BROWN



● Koko on saatavilla

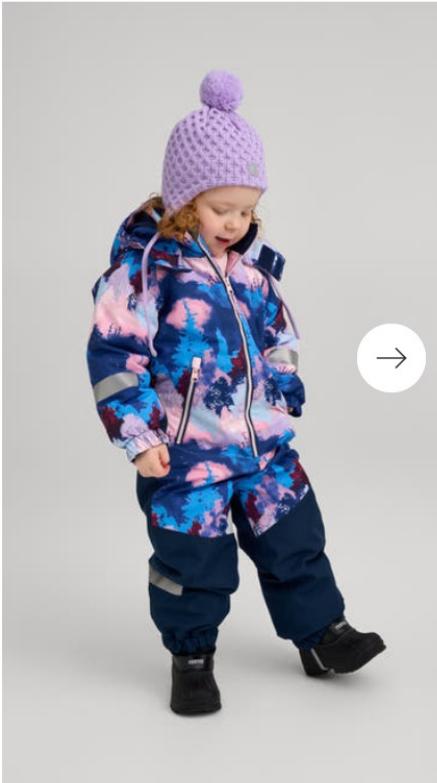
Istuvuuus: normaali

**Kasvuvara:** 6 cm [Lue lisää](#)

[Löydä sopivin koko](#) [Kokotaulukko](#)

[Lisää ostoskoriin](#)

# Should we use the number of products sold by color and size as an indicator of sales performance?



Lasten vedenpitävä toppahaalari  
Kurikka

83,97 €

Alin hinta viimeisten 30 päivän aikana: 83,97 €

★★★★★ 4.6/5 (53)

SAATAVILLA 6 VÄRIÄ: LILAC AMETHYST



VALITSE KOKO: 92



● Koko on saatavilla

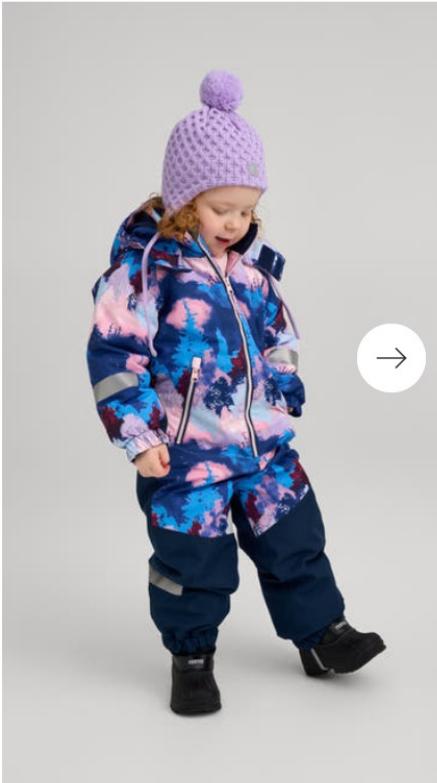
Istuvuus: normaali

**Kasvuvara:** 6 cm [Lue lisää](#)

[Löydä sopivin koko](#) [Kokotaulukko](#)

[Lisää ostoskoriin](#)

# How Would You Define Sales With Regard to Those Products Sold Out? Note that different products have different caps!



Lasten vedenpitävä toppahaalari  
Kurikka

83,97 €

Alin hinta viimeisten 30 päivän aikana: 83,97 €

★★★★★ 4.6/5 (53)

SAATAVILLA 6 VÄRÄÄ: LILAC AMETHYST



VALITSE KOKO: 92



● Koko on saatavilla

Istuvuuus: normaali

**Kasvuvara:** 6 cm [Lue lisää](#)

[Löydä sopivin koko](#) [Kokotaulukko](#)

[Lisää ostoskoriin](#)

**Would you think the products sold out in particular colors are the most popular products among all the colors?**

**In fact, the number of available products is different for different colors, which is affected by the availability of textiles. The procurement of textiles affects the number of available products.**

**Maybe we should differentiate products into two board categories: sold-out ones vs. not sold-out ones.**

**Should we recommend producing certain products more because of a high volume of sales?**

Some products get a good sale after having a heavy discount.  
It's not necessarily a yes or no question. However, avoiding producing products that do not make a profit is necessary.

## Section 5: Prepare for the Future

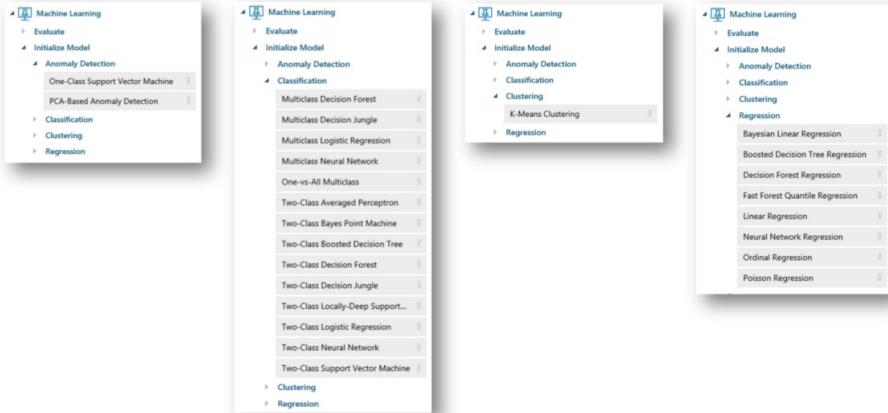


# Machine Learning is an Experimental Science!

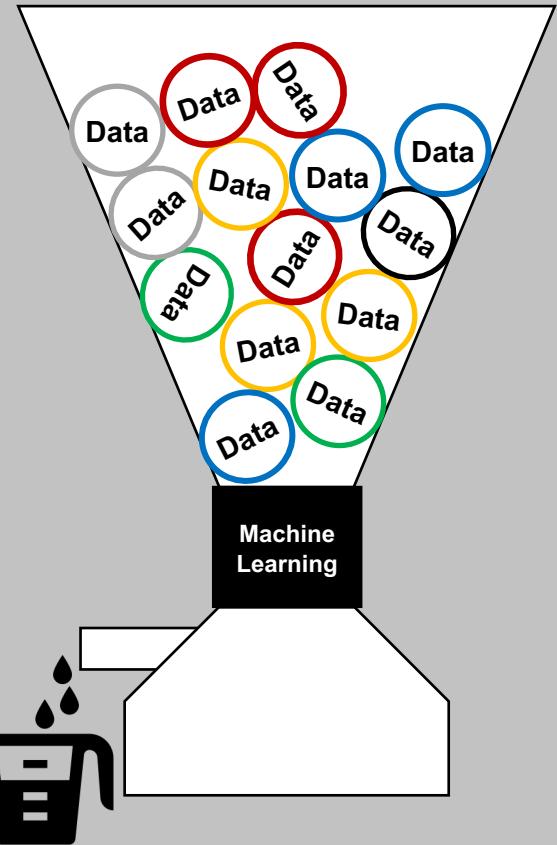
- Different machine learning algorithms have distinct strengths and weaknesses.
- No universal model performs optimally on all types of datasets and tasks.
- Experimentation helps identify the most suitable approach for a specific problem.
- Real-world datasets are often complex, containing diverse patterns and relationships. Trying various models allows you to explore different ways of capturing and understanding these complexities.

# How has machine learning been used in business contexts?

## Azure Machine Learning Algorithms



<https://towardsdatascience.com/microsoft-azure-machine-learning-336f83b0f27b>



- Output of ML
- Algorithm
  - Prediction

# For a data scientist using machine learning in business, whether s/he should also be a machine learning algorithm developer?

 Yo

## You

For a data scientist using machine learning in business, whether s/he should also be a machine learning algorithm developer? simple answer



## ChatGPT

Not necessarily. While having some algorithm development skills can be beneficial, it depends on the specific job requirements and the team structure. Some data scientists primarily focus on using existing algorithms and interpreting results rather than developing new ones.

**From a view of machine learning,  
explain why ChatGPT can  
succeed.**

There are approximately **86 billion** neurons in the human brain.



From the [OpenAI website](#), we know that ChatGPT is a fine-tuned version of GPT-3.5 ([text-davinci-002](#)).

1



On the [GPT-3.5 presentation page](#), they mention that the number of parameters is 175B (in the footnote of the table there, we can read that there may be slight differences with the actual model used in the API, though), which matches the size of GPT-3. Therefore I understand that ChatGPT is the same size as GPT-3.



The details of the architecture of GPT-3 have been made public in the [paper](#):

Model Name	$n_{\text{params}}$	$n_{\text{layers}}$	$d_{\text{model}}$	$n_{\text{heads}}$	$d_{\text{head}}$	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	$6.0 \times 10^{-4}$
GPT-3 Medium	350M	24	1024	16	64	0.5M	$3.0 \times 10^{-4}$
GPT-3 Large	760M	24	1536	16	96	0.5M	$2.5 \times 10^{-4}$
GPT-3 XL	1.3B	24	2048	24	128	1M	$2.0 \times 10^{-4}$
GPT-3 2.7B	2.7B	32	2560	32	80	1M	$1.6 \times 10^{-4}$
GPT-3 6.7B	6.7B	32	4096	32	128	2M	$1.2 \times 10^{-4}$
GPT-3 13B	13.0B	40	5140	40	128	2M	$1.0 \times 10^{-4}$
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	$0.6 \times 10^{-4}$

**Table 2.1:** Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Also, in the paper they mention that:

All models use a context window of  $n_{\text{ctx}} = 2048$  tokens

So:

- Number of layers: 96
- Number of attention heads: 96
- Dimensions of its hidden layers: 12288
- Sequence length: 2048
- Number of parameters: 175B

Note that, as ChatGPT comes from GPT-3 and GPT-3 is an evolution of GPT-2, we also know from their papers that the architecture differs somewhat from a standard Transformer decoder. See [this answer](#) for details on the specific architectural differences.

<https://datascience.stackexchange.com/questions/11827/3/specifc-about-chatgpts-architecture>

[https://www.nature.com/scitable/blog/brain-metrics/are\\_there\\_really\\_as\\_many/](https://www.nature.com/scitable/blog/brain-metrics/are_there_really_as_many/)

YO

**You**

r code to compare similarity the two strings of texts

**ChatGPT**

Certainly! To compare the similarity between two strings of text in R, you can use a variety of text similarity measures. Here's an example using the `'stringdist'` package and the Jaccard similarity:

R

Copy code

```
# Install and load necessary package
install.packages("stringdist")
library(stringdist)

# Function to calculate Jaccard similarity between two text
calculate_jaccard_similarity <- function(text1, text2) {
  # Tokenize the input strings
  tokens1 <- unlist(strsplit(tolower(text1), " "))
  tokens2 <- unlist(strsplit(tolower(text2), " "))

  # Calculate Jaccard similarity
  jaccard_similarity <- stringdist::stringdistmatrix(tokens1, tokens2)
  jaccard_similarity_score <- 1 - jaccard_similarity[1, 1]

  return(jaccard_similarity_score)
}

# Example texts
text1 <- "The stock market reached a new high today as investors"
text2 <- "Economic indicators contributed to a surge in the"
```

**As AI becomes increasingly proficient in data analytics, how should we (as business analysts, data scientists, or BI leaders) prepare for this evolving landscape?**

# Collective wisdom!

Do you agree?  
How should we prepare?



artificial intelligence.

[https://www.youtube.com/shorts/cmHrD\\_yVlGo](https://www.youtube.com/shorts/cmHrD_yVlGo)



The year is 1999. These students are working in the "new" computer lab. Who remembers when computers looked like that?



9:40 PM · Mar 30, 2017

---

1 Repost 1 Like 1 Bookmark

<https://twitter.com/SaintLucasMKE/status/847518895435124736>

# Section 6: BI for New Businesses



# BI with large machinery



# 'Moneyball For Law' Prediction Startup, Ex Parte, Bags \$7.5m

1st March 2022 · artificiallawyer · Litigation Prediction · 1



The logo for ExParte features a white stylized lightbulb icon containing a scale of justice, positioned above the company name "ExParte" in a large, white, sans-serif font.

We're using artificial intelligence and machine learning to predict the outcome of litigation and provide our customers with a winning advantage.

**Ex Parte**, a 'Moneyball for Law' startup, has bagged \$7.5m in Series A funding. It says that it can 'forecast the outcome of cases with approximately 85% accuracy' – which this site has to say seems doubtful in all but the simplest of cases, but is happy to be proved wrong.

The US company, which got going in 2017 and has four staff listed on LinkedIn, also claims to be 'the world's first company to leverage AI to predict the outcome of litigation'. But, we also have Premonition, which makes... (er... made?), very similar claims in the past and officially started back in 2014.

AI and Machine Learning as a product or service!

.... *Data engineering, strategies ... are lacking importance.*

VentureBeat

Security ▾ Data Infrastructure ▾ Automation ▾ Enterprise Anal

## This AI judge correctly predicts court case results 80% of time

A team of computer scientists and legal professionals has created artificial intelligence with the ability to accurately predict the outcome of 79 percent of cases decided upon by the European Court of Human Rights.

The performance of the artificial intelligence was published Monday by [PeerJ Computer Science](#).

11.4.2024

60

<https://venturebeat.com/business/this-ai-judge-correctly-predicts-court-case-results-80-of-time/>

<https://www.artificiallawyer.com/2022/03/01/moneyball-for-law-prediction-startup-ex-part-e-bags-7-5m/>

# Facebook likes predict your personalities better than your family members

## PARTICIPANTS' PERSONALITY

Measured using 100-item IPIP Five-Factor Model questionnaire (for 70,520 participants)

User	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
User 1	2.1	4.2	1.9	5.0	3.2
User 2	1.9	2.1	3.2	4.2	4.3
User 3	...	...	...	...	...
(...)	...	...	...	...	...
User X	...	...	...	...	...

90% of participants

10%

- 1 Take personality scores and Likes of 90% of the participants and build linear regression models for the five personality traits using LASSO variable selection

## PARTICIPANTS' LIKES

Obtained from Facebook profiles

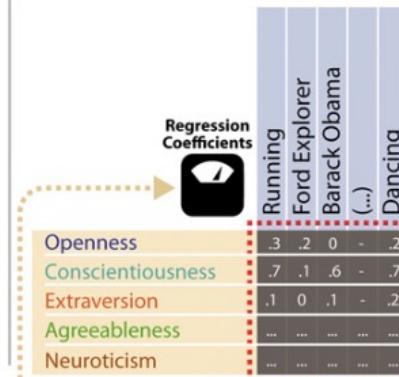
User	Running	Ford Explorer	Barack Obama	(...)	Dancing
User 1	1	1	0	-	0
User 2	0	1	1	-	1
User 3	1	0	1	-	1
(...)	...	...	...	...	...
User X	...	...	...	...	...

90% of participants

10%

## LINEAR REGRESSION MODELS

A regression formula with a coefficient for each Like is generated for each of the five personality traits  
e.g. Openness =  $a + \beta_1 * \text{running} + \beta_2 * \text{Obama} + \dots + e$



2

- 2 Take the Likes of the remaining 10% of the participants and use the linear regression models to predict scores for the five personality traits

## COMPUTERS' JUDGMENTS

Made using participants' Likes

User	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
User 1	2.1	4.2	1.9	5.0	3.2
User 2	1.9	2.1	3.2	4.2	4.3
User 3	...	...	...	...	...
(...)	...	...	...	...	...
User X	1.9	2.1	3.2	4.2	4.3

90% of participants

10%

Repeat 10 times to make judgments for all participants



Humans' Judgments

Humans' Accuracy



Self-ratings

Computers' Accuracy



Computers' Judgments

# The case of Webank

1. Webank offers personal load online within a few minutes.
2. "Webank" used thousands of dimensions of data to build a model, including credit data, social data, public security data, population registration data, academic data, transaction records, etc., comprehensively evaluate the credit rating of the applicant, decide whether to grant loans and grant the amount. All of this is completed within a few seconds after the customer opens the "Xiwei Loan" page."

# Section 7: Future Competition



# Computing Power and Electricity will decide the competition between industries and countries.

## Artificial Intelligence Could Fuel Robust Demand for Power. Here Are 2 Under-the-Radar Ways to Cash In on AI's Energy Needs.

February 18, 2024 — 06:17 am EST

Written by Matt DiLallo for The Motley Fool →



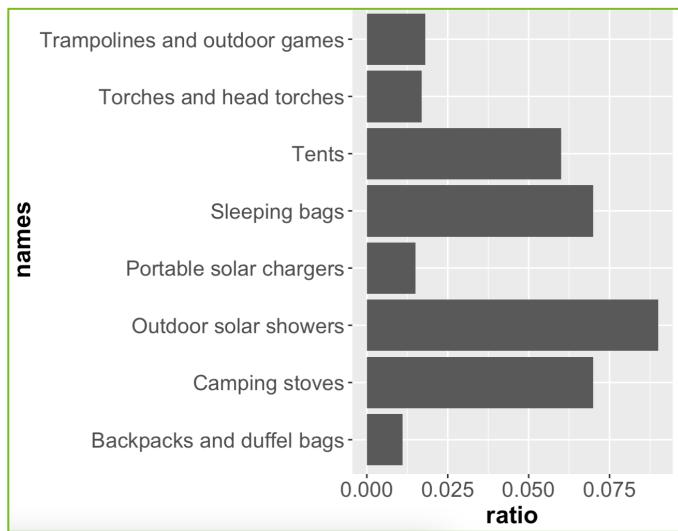
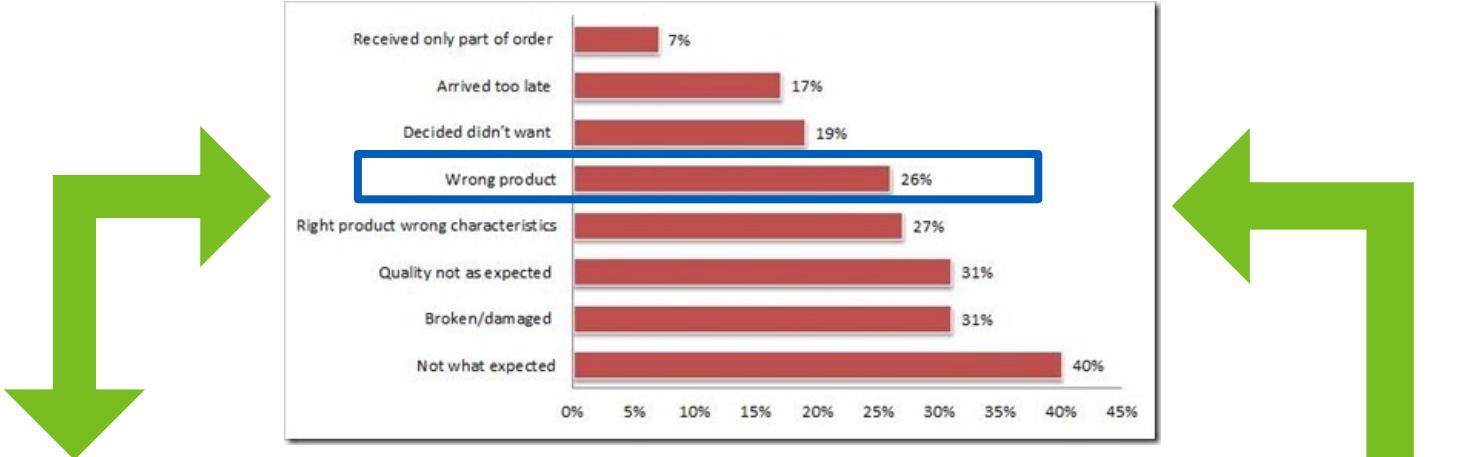
Artificial intelligence (AI) is a multi-trillion-dollar megatrend. According to an estimate by PwC, AI has the potential to provide a \$15.7 trillion boost to the global economy by the end of this decade. That's more than the current economic output of India and China combined.

To reach its full potential, [AI](#) requires two crucial components: computing power and electricity. While most investors are currently focusing on the first aspect (e.g., [semiconductors](#)), many don't yet realize the importance of power. According to one estimate, data centers will grow from consuming 2% of the world's electricity to 10% by 2030. That's the equivalent of adding the electric generation capacity of the current U.S. power grid. While [renewable energy will be crucial in powering AI](#), it can't carry that load alone. That drives the view that cleaner-burning [natural gas](#) will be vital in helping power AI's growth. Two under-the-radar ways to cash in on that trend are gas [pipeline](#) giants [Kinder Morgan \(NYSE: KMI\)](#) and [Williams \(NYSE: WMB\)](#).

11.4.2024

64

## Reasons for Product Returns



**KPI**

- Revenue
- Customer satisfaction

# Section 8: BI Leadership Qualification



# BI Leadership Qualification Survey

The survey includes seven questions. Please use strongly disagree, some disagree, neutral, somewhat agree, and strongly agree to evaluate the following statement.

1. The IT department should take a leading role in operating data quality projects.
2. The data used for creating dashboard reports must be accurately precise.
3. The success of the BI department should be measured by the number of report generation reports that they have generated to meet the requests of other departments.

# BI Leadership Qualification Survey

4. Business analysts in my company who generate BI reports should be responsible for the quality of data that they use.
5. Apart from doing e.g., machine learning, data scientists in my company are requested to understand the company's business processes.
6. Machine learning & AI are key to improving the management effectiveness of my company.

# BI Leadership Qualification Survey

The survey includes seven questions. Please use strongly disagree (1), some disagree (2), neutral (3), somewhat agree (4), and strongly agree (5) to evaluate the following statement.

1. The IT department should take a leading role in operating data quality projects. **Strongly Disagree (1) Strongly Agree (5)**
2. The data used for creating dashboard reports must be accurately precise. **Strongly Disagree (1) Strongly Agree (5)**
3. The success of the BI department should be measured by the number of report generation reports that they have generated to meet the requests of other departments. **Strongly Disagree (1) Strongly Agree (5)**

# BI Leadership Qualification Survey

4. Business analysts in my company who generate BI reports should be responsible for the quality of data that they use.

**Strongly Disagree (1) Strongly Agree (5)**

5. Apart from doing e.g., machine learning, data scientists in my company are requested to understand the company's business processes. **Strongly Agree (1) Strongly Disagree (5)**

6. Machine learning & AI are key to improving the management effectiveness of my company.

**Strongly Disagree (1) Strongly Agree (5)**

# Course Closure

**The objective of the course** is to give the students an understanding of the data informed decision-making. The focus is on business-oriented analytical skills, which enable the students to work with the data, to understand them and turn them into intelligence and actions.

**After the course, you will be able to:**

- Obtain an understanding of the data informed decision-making in enterprises.
- Obtain business-oriented analytical skills via the use of Power B and Tableau.
- Understand the challenges to work with the data, to understand them and turn them into intelligence and actions.

# Course Closure

- 1. Thank you for attending this course.**
- 2. Tableau Challenge Event**
  - Please indicate whether you are willing to participate in the challenge event to gain extra bonus points in the assignment.
- 3. Looking forward to seeing you again at the challenge event.**
- 4. Sharing your story with me if you have a good story about BI after your graduation!**

# Please fill in course feedback survey

- Log in Mycourse page
- Click “Course feedback” in the top menu of the page

