



INDONESIA JAYA

Data Analytics

Program Zenius Studi Independen Bersertifikat Bersama Kampus Merdeka



Quick Intro

Theo Jeremiah

Roles:

- CURRENTLY | Data Scientist at AirAsia
- 20 - 23 | Data Scientist at Allianz Indonesia
- 19 - 20 | Business Development at Mineski Indonesia
- 18 - 19 | Data Analyst at Excite Indonesia

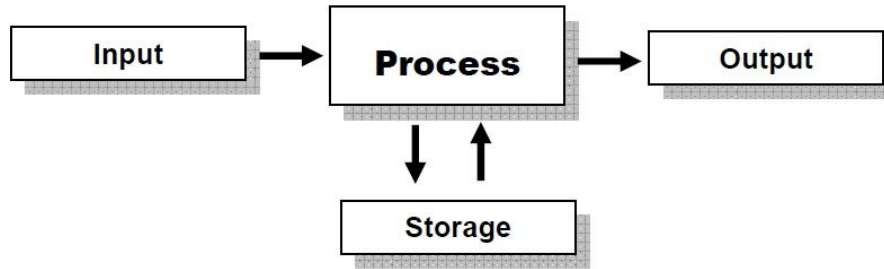


www.linkedin.com/in/theojeremiah/

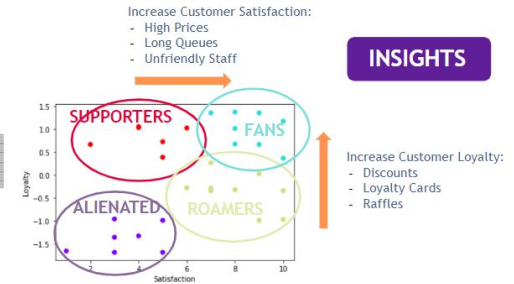
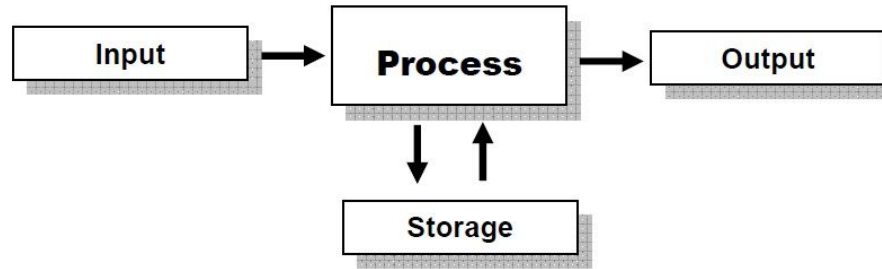
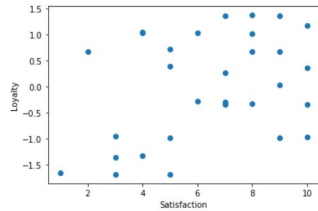
- 1. Important Concepts**
- 2. Roles in Data Science**
- 3. How to Build Your Portfolio**
- 4. Intro to Kaggle and Github**

Important Concepts

Data Analytics



Data Analytics



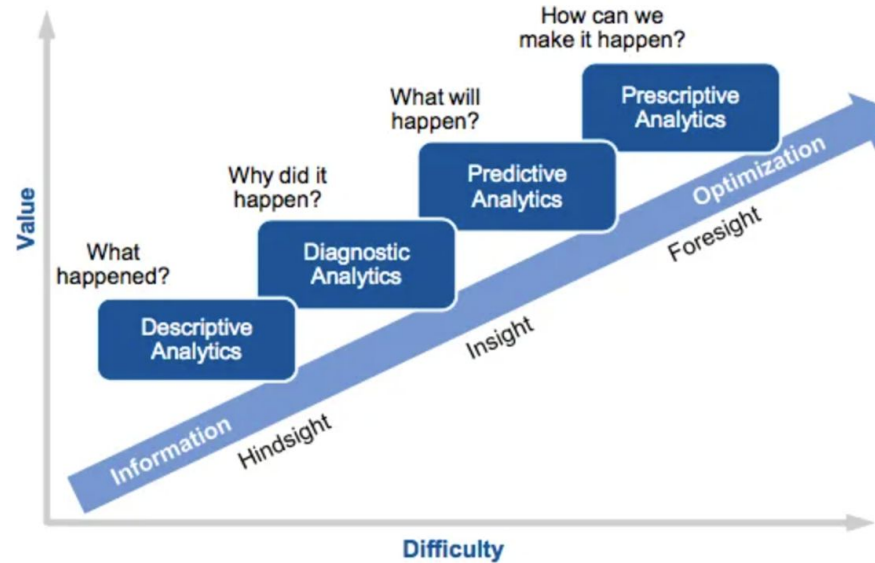
Data Analytics

“Garbage in, garbage out”



Your analysis is as good as your data.

Data Analytics



Source : Gartner Analytics Ascendancy Model

<https://www.clickz.com/how-can-ai-allow-marketers-to-predict-the-future/112268/gartner-analytic-ascendancy-model/>

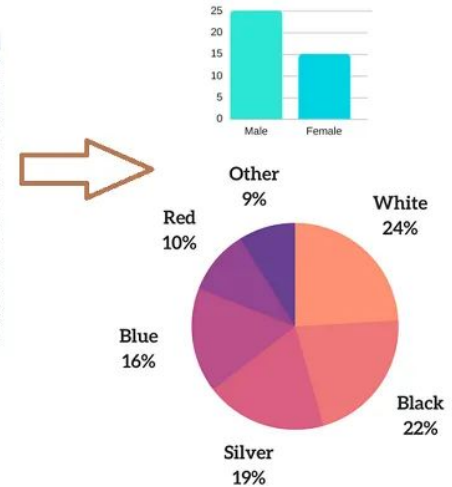
<https://www.gartner.com/en/topics/data-and-analytics>

Descriptive Analytics

- Describing the data
- Common Calculation :
 - Sums
 - Counts
 - Averages
- Typical Reports :
 - Tables
 - Bar Charts
 - Pie Charts
 - Narratives

	A	B	C	D
1	Respondent Number	Age	Gender	Favorite Car Color
2	1	22	M	White
3	2	37	F	Silver
4	3	45	F	Black
5	4	62	F	Gray
6	5	28	M	Red
7	6	45	M	Green
8	7	88	F	Brown
9	8	61	M	White
10	9	95	M	Black
11	10	27	M	White
12	11	39	F	Green
13	12	43	M	Brown
14	13	55	F	Black
15	14	59	F	White

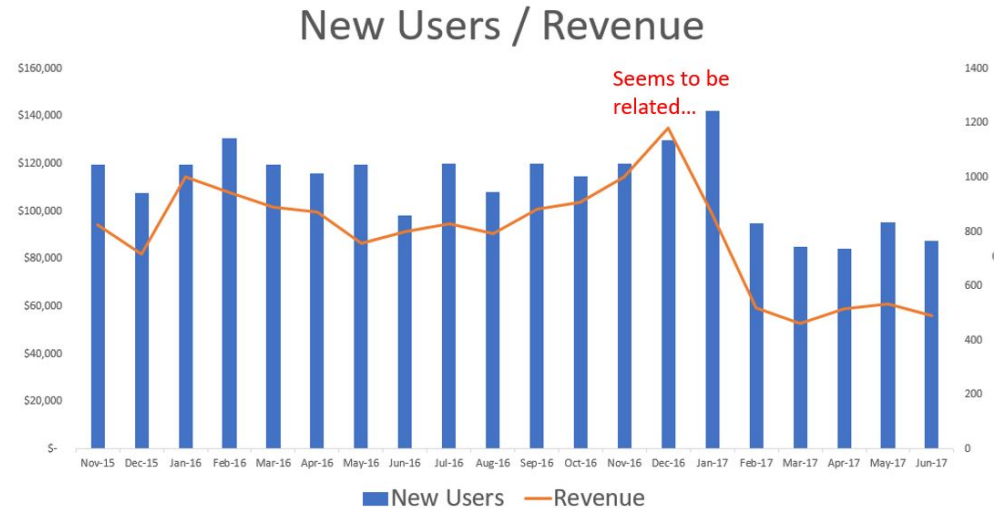
RAW DATA



Descriptive Statistics

Diagnostic Analytics

- Answers “Why did it happen?”
- Drill Down Techniques
- Data Discovery
- Correlations
- Combining Charts
- Discover Related Metrics



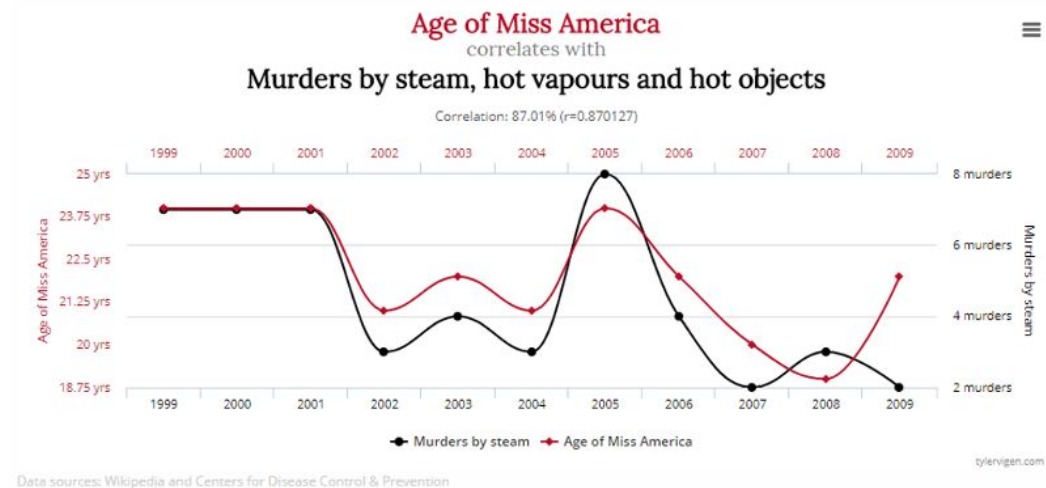
Diagnostic Analytics

Correlation **doesn't prove** Causation

Correlation will tell you when two variables (say clicks and conversions) move **in sync** with one another

While it's tempting to draw conclusions from that fact, the **correlation must also make sense** before it can be considered as **causal evidence**

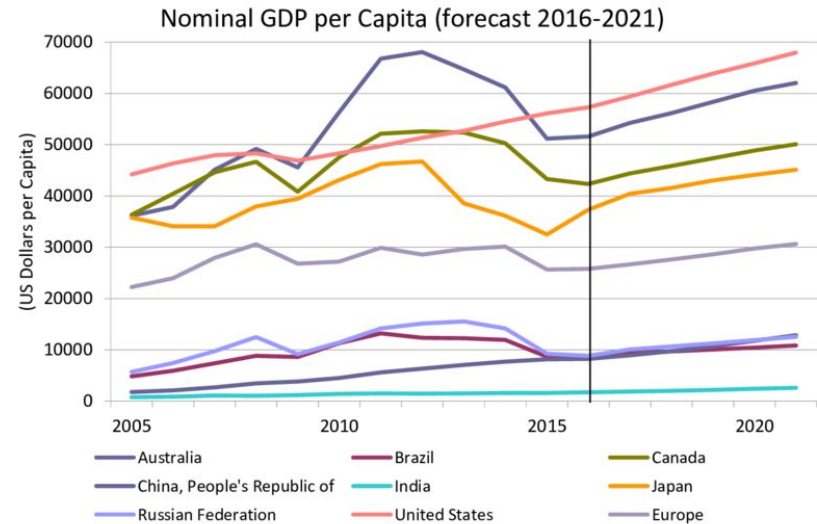
That's why we need **Business Acumen**.



Predictive Analytics

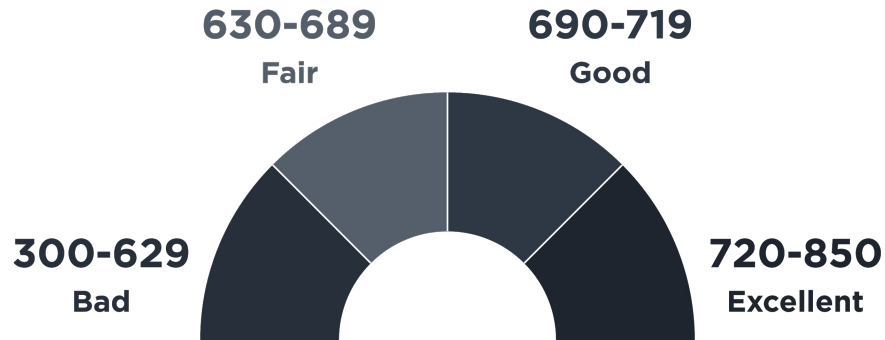
a variety of **statistical techniques** from **data mining**, **predictive modelling**, and **machine learning**, that analyze **current** and **historical facts** to make **predictions about future** or otherwise unknown events.

- **exploiting patterns** found in historical and transactional data
- **identifying** risks and opportunities
- **capturing relationships** among many factors to the target
- **guiding** decision-making



Predictive Analytics

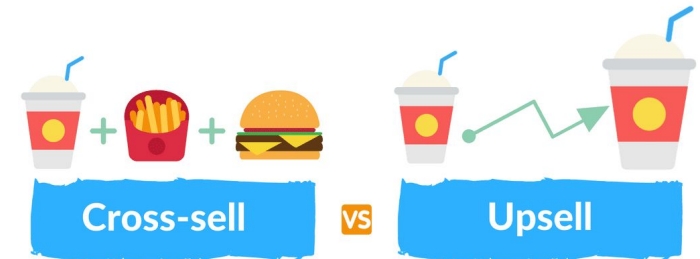
Credit Risk Scoring



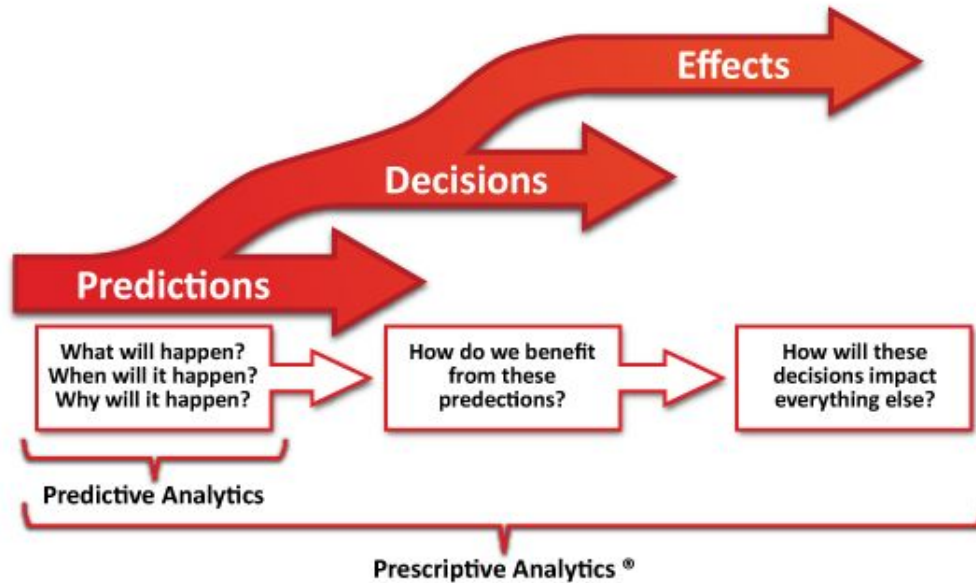
Sentiment Analysis

Word	Sentiment
good	0.5
great	0.8
terrible	-0.8
alright	0.1

Cross-Selling/Upselling



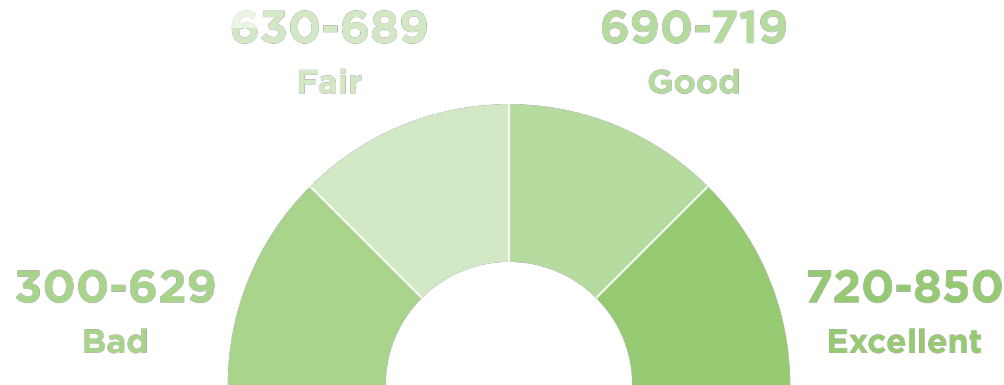
Prescriptive Analytics



also include **Optimization**.

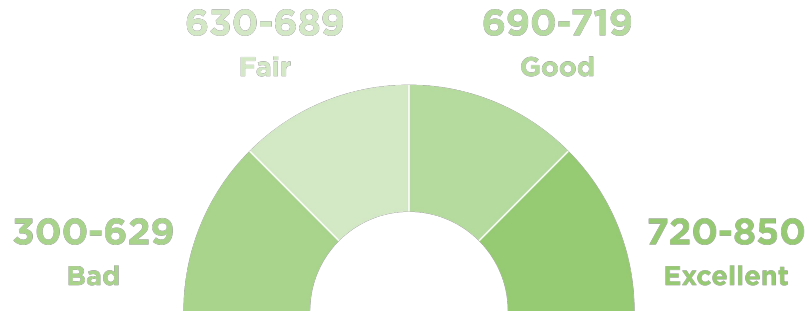
Prescriptive Analytics

Credit Risk Scoring



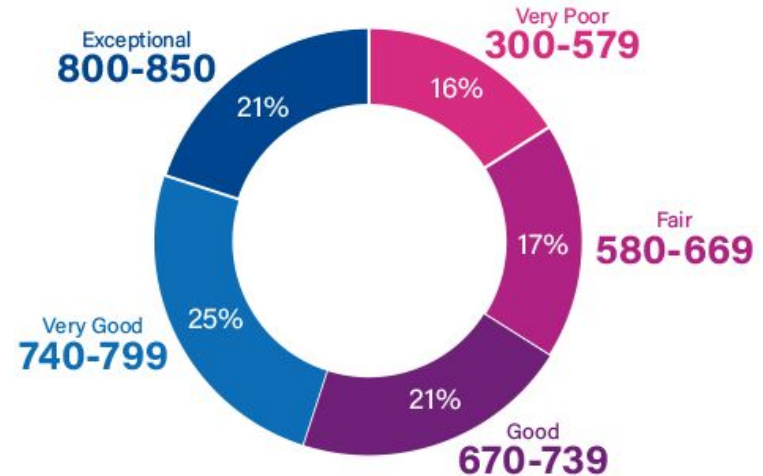
- How much is the **Expected Credit Loss (ECL)** ?
- How about the **Probability of Default (PD)** ?
- Where is **the best cut-off** for Bad and Good given X risk appetite ?
- When someone is **accepted** for a loan, will someone with **840** credit score has the same **LTV (Lifetime Value)** as other one who has **700** ?

Prescriptive Analytics



Customer with no credit history

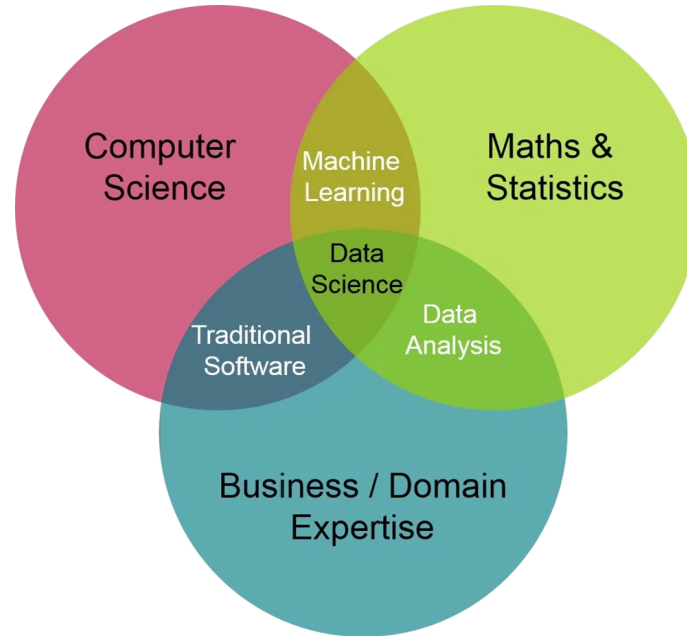
Customer with credit history



Different Product, Different Credit Scorecard
Different Region, Different Credit Scorecard
Unbankable vs Bankable Customer Credit Scorecard

Roles in Data Science

Roles in Data Science

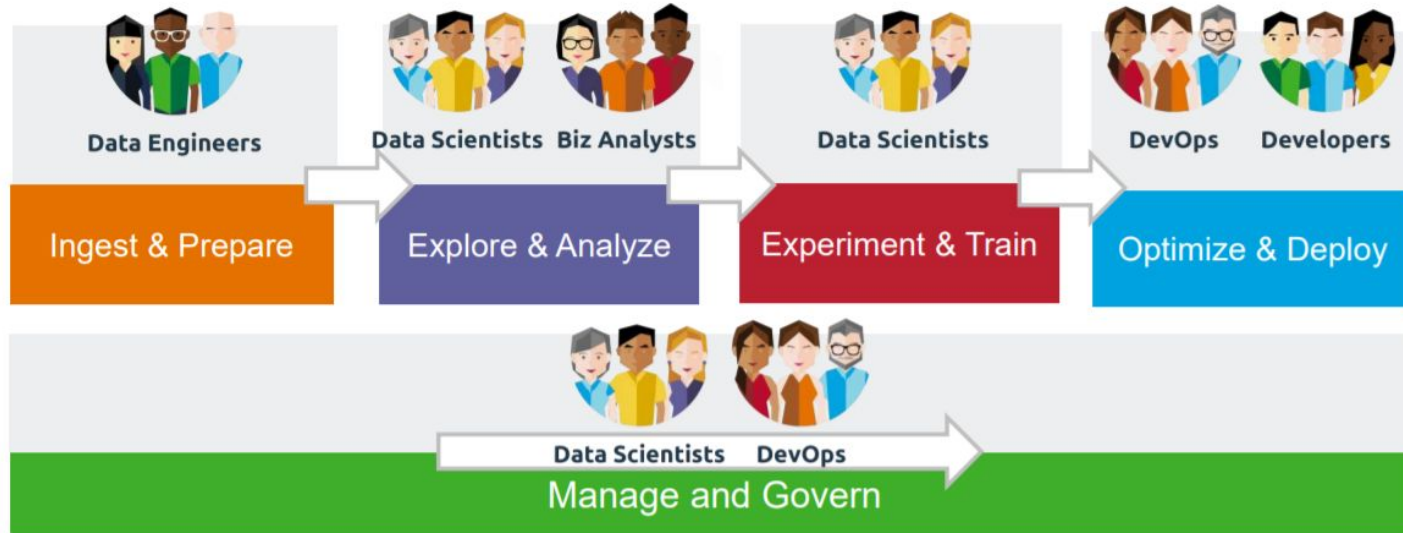


Roles in Data Science

"A data scientist is someone who is better at statistics than any programmer and better at programming than any statistician"

- a random stranger on twitter

Data Science : Team Sport



Data Scientist

<h2>Data Scientist</h2> <p>also known as Data Managers, statisticians.</p>  <p>A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.</p> <p>Skills: Mathematics, Programming, Communication</p> <div></div> <p>Will use programmes such as: SQL, Python, R</p>	<h2>Data Engineers</h2> <p>also known as database administrators and data architects.</p>  <p>They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.</p> <p>Skills: Programming, Mathematics, Big data</p> <div></div> <p>Will use programmes such as: Hadoop, NoSQL, and Python</p>	<h2>Data Analysts</h2> <p>also known as business Analysts.</p>  <p>They typically help people from across the company understand specific queries with charts.</p> <p>Skills: Statistics, Communication, Business knowledge</p> <div></div> <p>Will use programmes such as: Excel, Tableau, SQL</p>
---	---	--

Data Analyst



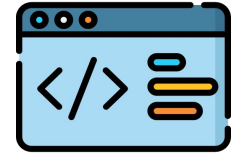
Analytics

Problem Solving, Data Exploration



Visualization

Right Plot for The Right Purpose



Programming & Tools

SQL, Python, Excel



Statistics

Uni/Bi/Multi-variate, Hypothesis Testing



Business Acumen

Understanding The Subject Matter Deeply

How to Build Your Portfolio

How To Build Your Portfolio

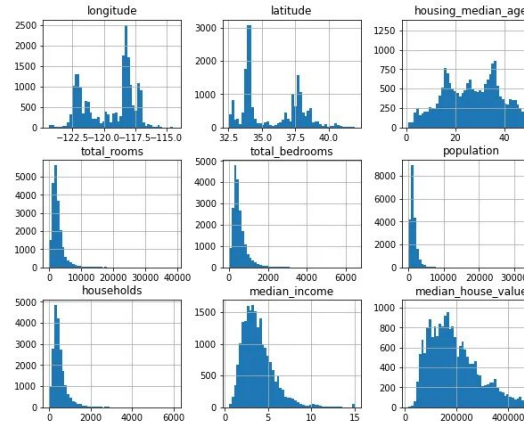
Building a portfolio is an essential part to conquer the career struggle.



- **Choose a topic you're interested in**
Don't get complicated, make sure the topic to analyze is something you know well, to ease the way.
- **End-to-end**
Make a complete portfolio from start to finish
- **Explainable**
Make sure your audience/interviewer able to understand what you're making and how you make it.
- **Make a story and publish it!**
Use platform like kaggle, github, medium and linkedin to spread the awesome stuff and the journey!

Example of a Good Portfolio and Projects

Distribution of Amazon Product Ratings



<https://amankharwal.medium.com/130-machine-learning-projects-solved-and-explained-605d188fb392>

***credit to the owner.**

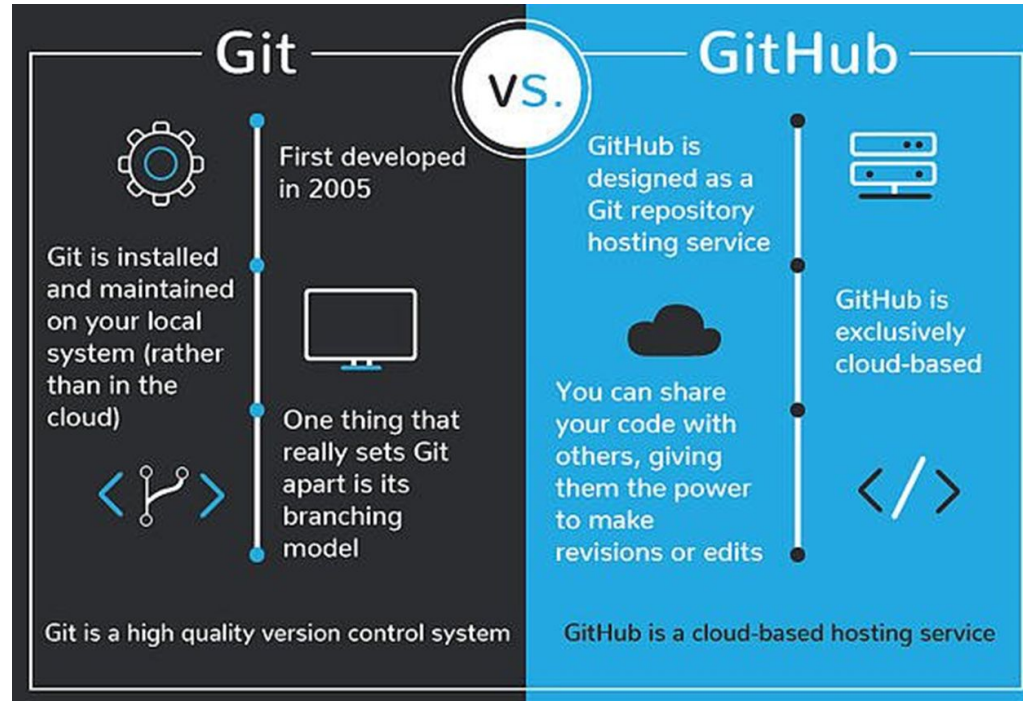
Intro to Kaggle and Github

Github

At a high level, GitHub is a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code.



Github

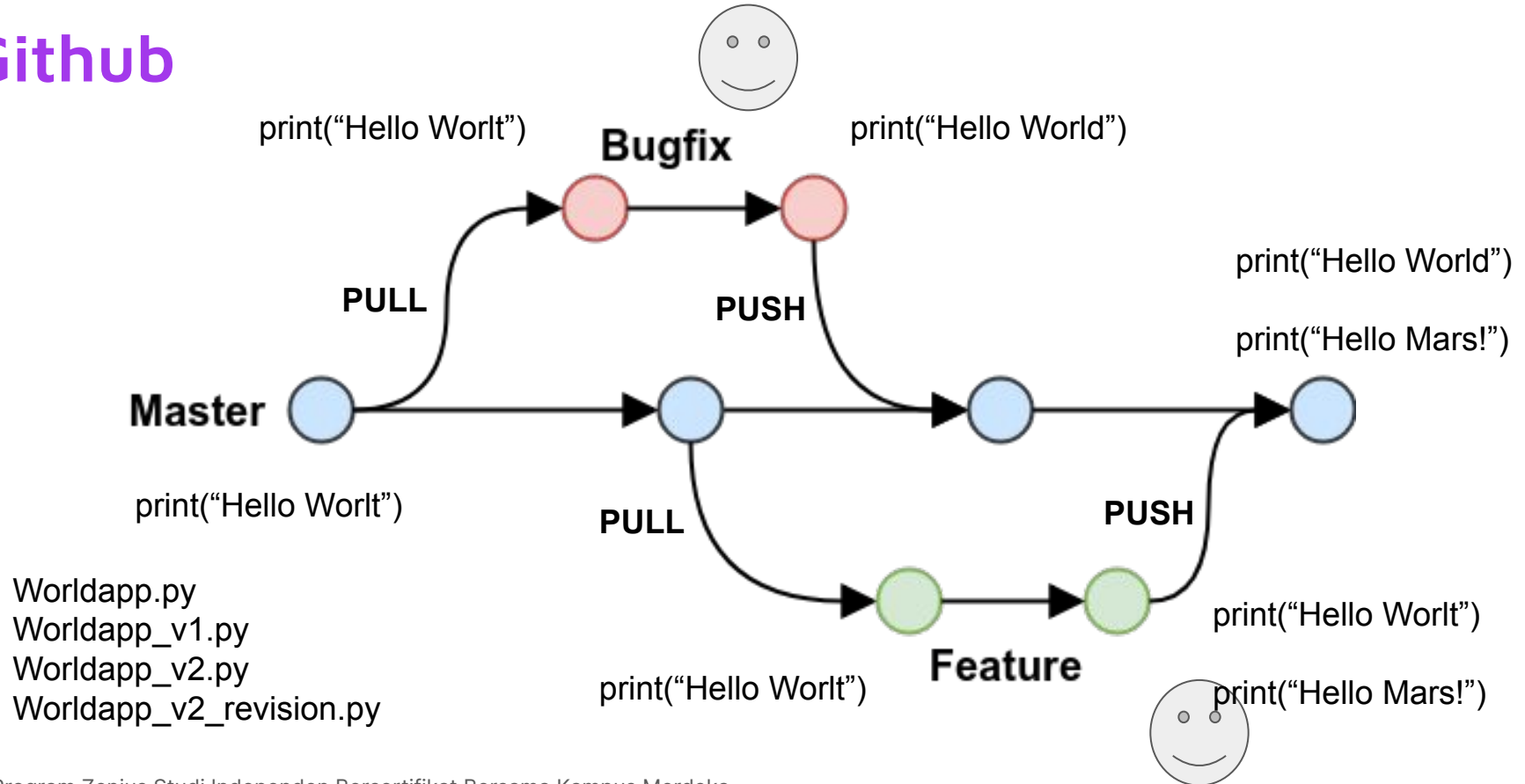


Github

command	description
<code>git clone <i>url</i> [<i>dir</i>]</code>	copy a git repository so you can add to it
<code>git add <i>files</i></code>	adds file contents to the staging area
<code>git commit</code>	records a snapshot of the staging area
<code>git status</code>	view the status of your files in the working directory and staging area
<code><u>git</u> diff</code>	shows diff of what is staged and what is modified but <u>unstaged</u>
<code>git help [<i>command</i>]</code>	get help info about a particular command
<code>git pull</code>	fetch from a remote repo and try to merge into the current branch
<code>git push</code>	push your new branches and data to a remote repository
others: <u>init</u> , reset, branch, checkout, merge, log, tag	

Some of git commands.

Github



Worldapp.py
Worldapp_v1.py
Worldapp_v2.py
Worldapp_v2_revision.py

Version Control

Version Control is the general term.

Version control lets developers safely work through branching and merging.

With **branching**, a developer duplicates part of the source code (called the repository). The developer can then safely make changes to that part of the code without affecting the rest of the project.

Then, once the developer gets his or her part of the code working properly, he or she can **merge (merging)** that code back into the main source code to make it official.

All of these changes are then tracked and can be reverted if need be.

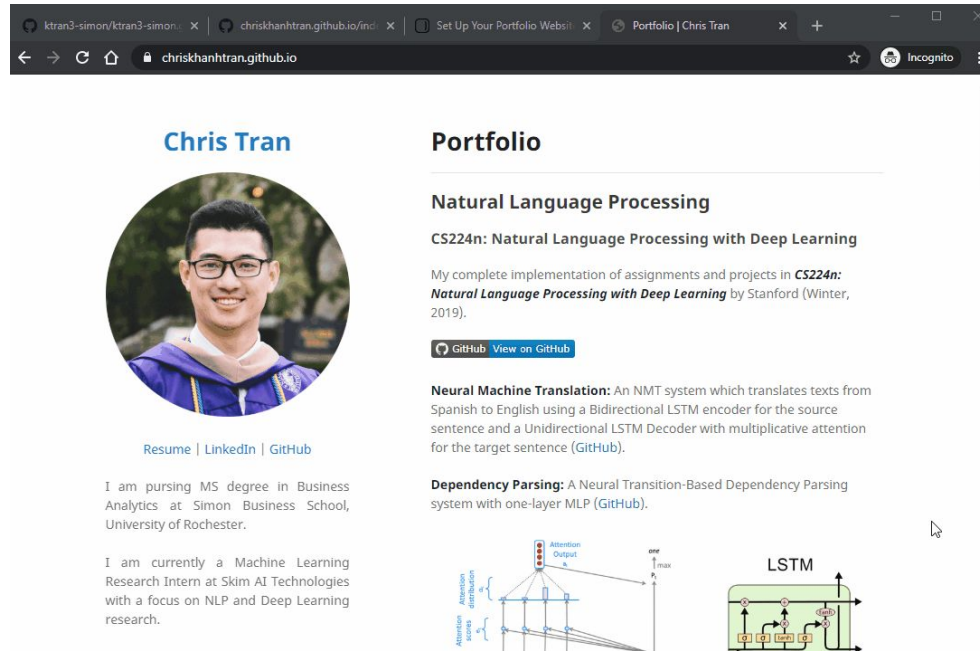
Git

Git is a specific open-source version control system created by Linus Torvalds in 2005.

Specifically, **Git** is a distributed version control system, which means that the entire codebase and history is available on every developer's computer, which allows for easy branching and merging.

Let's try to create a github account and do some demos there ! Link : <https://github.com/>

Github Portfolio



The screenshot shows a web browser window with the address bar displaying 'chriskhanhtran.github.io'. The page features a profile section for Chris Tran, including a circular profile picture, a bio, and links to his Resume, LinkedIn, and GitHub. The main content area is titled 'Portfolio' and lists two projects: 'Natural Language Processing' and 'Neural Machine Translation'. The 'Natural Language Processing' project is described as a complete implementation of assignments and projects in 'CS224n: Natural Language Processing with Deep Learning' by Stanford (Winter, 2019). Below the project descriptions are two diagrams: one showing an attention mechanism with 'Attention weights' and 'Attention distribution', and another showing an 'LSTM' cell structure.

Chris Tran

[Resume](#) | [LinkedIn](#) | [GitHub](#)

I am pursuing MS degree in Business Analytics at Simon Business School, University of Rochester.

I am currently a Machine Learning Research Intern at Skim AI Technologies with a focus on NLP and Deep Learning research.

Portfolio

Natural Language Processing

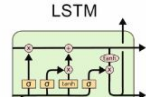
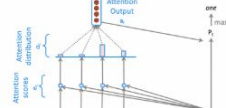
CS224n: Natural Language Processing with Deep Learning

My complete implementation of assignments and projects in **CS224n: Natural Language Processing with Deep Learning** by Stanford (Winter, 2019).

[GitHub](#) [View on GitHub](#)

Neural Machine Translation: An NMT system which translates texts from Spanish to English using a Bidirectional LSTM encoder for the source sentence and a Unidirectional LSTM Decoder with multiplicative attention for the target sentence ([GitHub](#)).

Dependency Parsing: A Neural Transition-Based Dependency Parsing system with one-layer MLP ([GitHub](#)).



Kaggle

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

<https://www.kaggle.com/>

The Kaggle logo, featuring the word "kaggle" in a lowercase, blue, sans-serif font.

Kaggle

The screenshot displays the Kaggle notebook interface for a competition titled "Predict Malicious Websites: XGBoost". The interface includes a top navigation bar with a file menu, a code editor, and a right-hand sidebar with session and workspace information.

Code Editor:

```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()

# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(tree_method = "gpu_exact")

# fit our model
model.fit(y=Y, X=X)

In[]: # split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]

# predictions & actual values, from test set
predictions = model.predict(test_X) > 0
actual = test_Y
```

Sessions:

Session	Time	GPU
Draft Session	1m/9h	Off

Workspace:

- input (read-only data)

Settings:

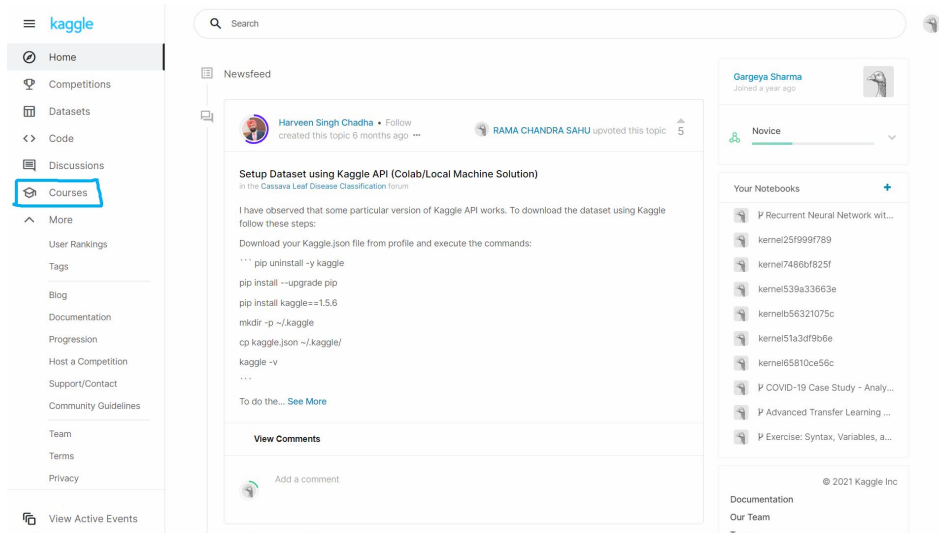
- Sharing: Private
- Language: Python
- Docker: Latest Available
- GPU: On
- Internet: Off
- Packages: Install...
- BigQuery: Enable...

Console:

Draft Session (0m10s) | CPU: 45% | GPU: Off | RAM: 4.5/8GB | Disk: 32MB/128GB

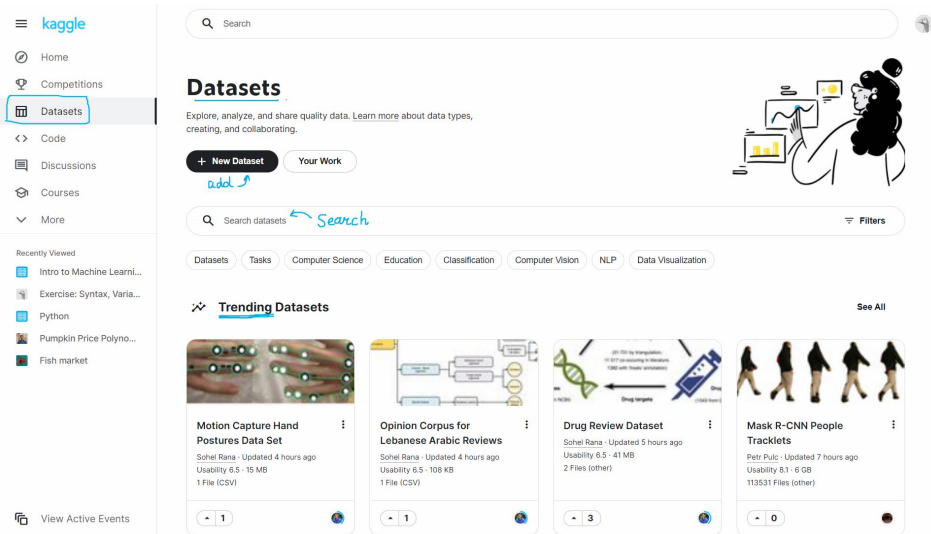
Advantages of Using Kaggle

1. Free Courses and free certificates available



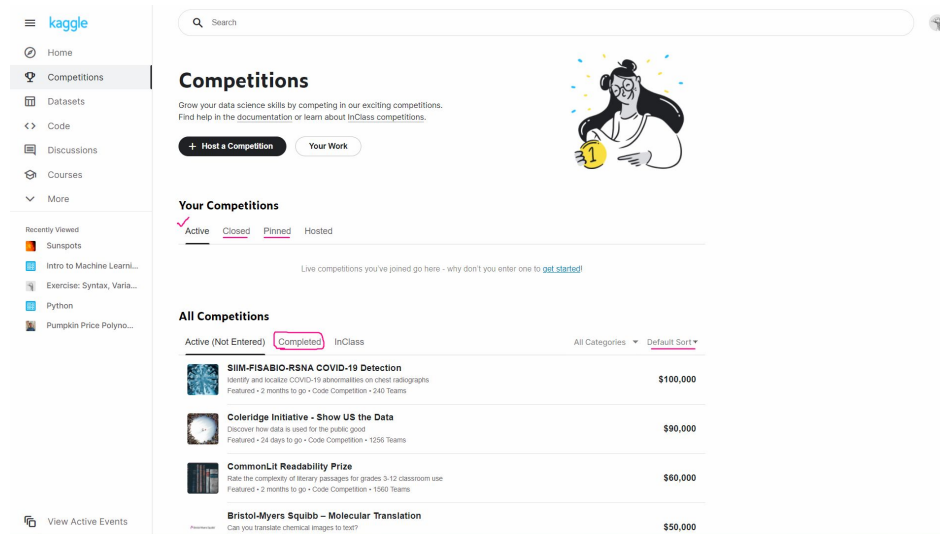
Advantages of Using Kaggle

2. A Huge collection of publicly available/ contributed datasets to practice/ work on



Advantages of Using Kaggle

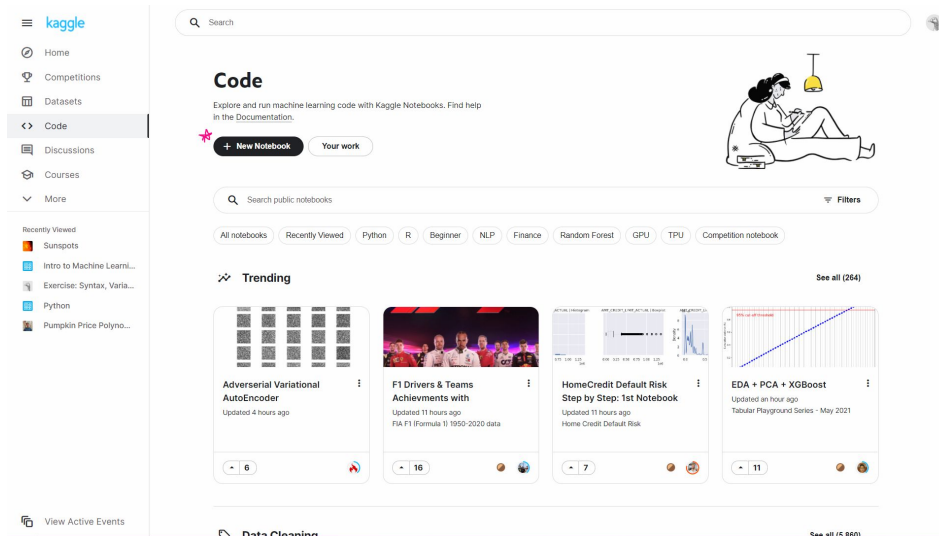
3. Data Science/ Machine Learning / Deep learning Competitions



The screenshot shows the Kaggle website's 'Competitions' page. The left sidebar contains navigation links: Home, Competitions (selected), Datasets, Code, Discussions, Courses, and More. Below these are 'Recently Viewed' items: Sunspots, Intro to Machine Learning..., Exercise: Syntax, Variables..., Python, and Pumpkin Price Polynomial Regression. The main content area has a search bar and a 'Competitions' header with a sub-header: 'Grow your data science skills by competing in our exciting competitions. Find help in the documentation or learn about InClass competitions.' There are two buttons: 'Host a Competition' and 'Your Work'. Below this is a 'Your Competitions' section with tabs for 'Active', 'Closed', 'Pinned', and 'Hosted'. A message states: 'Live competitions you've joined go here - why don't you enter one to [get started!](#)'. The 'All Competitions' section has tabs for 'Active (Not Entered)', 'Completed' (highlighted with a red box), and 'InClass'. It lists three competitions: 1. 'SIIM-FISABIO-RSNA COVID-19 Detection' with a prize of \$100,000, 2. 'Coleridge Initiative - Show US the Data' with a prize of \$80,000, and 3. 'CommonLit Readability Prize' with a prize of \$60,000. A fourth competition, 'Bristol-Myers Squibb - Molecular Translation', is partially visible with a prize of \$50,000. The bottom of the page has a 'View Active Events' link.

Advantages of Using Kaggle

4. Kaggle Notebooks / Code



Assignment

Assignment 1

Instruksi Assignment 1A

Buatlah sebuah tulisan mengenai use case data science di industri tertentu yang dapat kalian pilih di bawah ini dan bahas dengan detail bagaimana data science diaplikasikan, metode yang digunakan, contoh data yang dipakai, dan impact terhadap bisnis di industri tersebut.

Pilihan industri:

- Banking
- Retail
- Healthcare
- Supply Chain

Minimal 2 halaman, Times New Roman 12, gunakan pula ilustrasi/ diagram untuk melengkapi penjelasan yang kalian berikan.

Instruksi Assignment 1B

Buatlah sebuah github page

Portfolio boleh diisi oleh file **.ipynb** kosong / dummy terlebih dahulu.

Baca juga "**Important Links and Software**" pada Module sebagai referensi pembuatan github page ini.

Buatlah juga akun LinkedIn untuk menjadi tempat showcase portfolio dan experience kalian

Submission: kumpulkan tugas dalam file pdf, lampirkan url github dan LinkedIn kalian di dalam pdf tersebut.

Available from	Until
Mar 15 at 09.00 PM	Mar 20 at 11.59PM

Thanks!
Any Questions?

zenius



**Kampus
Merdeka**
INDONESIA JAYA

