OPPORTUNITIES AND ETHICS IN DATA WAREHOUSING AND DATA MINING

IT 116: INFORMATION MANAGEMENT 2

JENNIFER L. LLOVIDO, DIT

1ST SEMESTER SY 2022-2023

ETHICAL CONCERNS IN DATA WAREHOUSING

- Data warehousing takes information from different databases as well as external sources and puts them inside a repository which can be accessed by end-users who need decision support.
- There are ethics to consider especially when some data ***may*** **be accessed only at the departmental or only at certain levels**.
- There is a **chance** that end-users may have access to information that they should not be examining. They may be breaking privacy laws **without knowing it**.

ETHICAL CONCERNS IN DATA WAREHOUSING

- Let’s take this as an example. A customer fills out his personal details as well as his medical history at a certain clinic. He authorizes the doctor and the clinic to know where he lives, who co-pays his insurance. It is fine if these details stay in the clinic’s database. Eventually, the clinic is bought by a larger hospital. The hospital uses a data warehouse but is the **hospital allowed to access those records that the mentioned customer has**? If not, the data warehouse must make sure that those information remain private.

ETHICAL CONCERNS IN DATA WAREHOUSING

- Situations like what I have mentioned **are happening today**.
- Often, implementers of the technology are simply told to integrate the data. The project manager simply builds it to make it happen. There's nothing wrong with that, they are simply doing their jobs.
- However, issues in privacy must be addressed **throughout every stage** of the Kimball Cycle.

ETHICAL CONCERNS IN DATA WAREHOUSING

- *Question #1:* While testing the data warehouse, is it alright to move small data sets from source systems to target systems **for testing purposes**?
- *The answer is this:* **It is not actually ethical to do so**. While testing, sometimes users are learning things they shouldn't know or things they aren't allowed to know.

ETHICAL CONCERNS IN DATA WAREHOUSING

- *Question #2:* In creating a data warehouse, we are allowed to get external data and pull it into the repository. These are **publically available information**. Is it ethical to integrate everything into the data warehouse?
- *The answer:* The project manager must decide which of the information is acceptable to integrate. Although the information is publically available, using some of them might raise **ethical considerations**. The ethics would focus on **how** the information is used, and by **whom**.

ETHICAL CONCERNS IN DATA WAREHOUSING

- *Question #2:* In creating a data warehouse, we are allowed to get external data and pull it into the repository. These are **publically available information**. Is it ethical to integrate everything into the data warehouse?
- *The answer:* The project manager must decide which of the information is acceptable to integrate. Although the information is publically available, using some of them might raise **ethical considerations**. The ethics would focus on **how** the information is used, and by **whom**.

ETHICAL CONCERNS IN DATA WAREHOUSING: CHECKLIST

Here is a checklist of items project managers and technology implementers can use to manage ethical concerns:

- Develop **service level agreements** with end users that define who has access to what levels of information
- Have **end-users involved** in defining the ethical standards of use for the data that will be delivered.
- Define the **bounds** around the integration efforts of public data, where it will be integrated and where it will not – so as to avoid conflicts of interest.
- Do not use “live” or real data for testing purposes – or **lock down the test environment**; too often test environments are left wide-open and accessible to too many individuals.

ETHICAL CONCERNS IN DATA WAREHOUSING: CHECKLIST

- Define where, how, and who will be using Data Mining – **restrict** the mining efforts to specific sets of information. Build a **notification** system to monitor data mining usage.
- Allow customers to **"block"** the integration of their own information (this one is questionable) depending on if the customer information after integration will be made available on the web.
- Remember that any efforts made are still subject to **governmental laws**. What laws do we have right now concerned with data privacy? Note that future laws could also be developed and we must be aware of those.

CONCLUSION

- While we can always mean well in our business intelligence efforts, we must always make sure that we consider ethical norms in our processes. Data warehouses are not exempt from these.

REFERENCES:

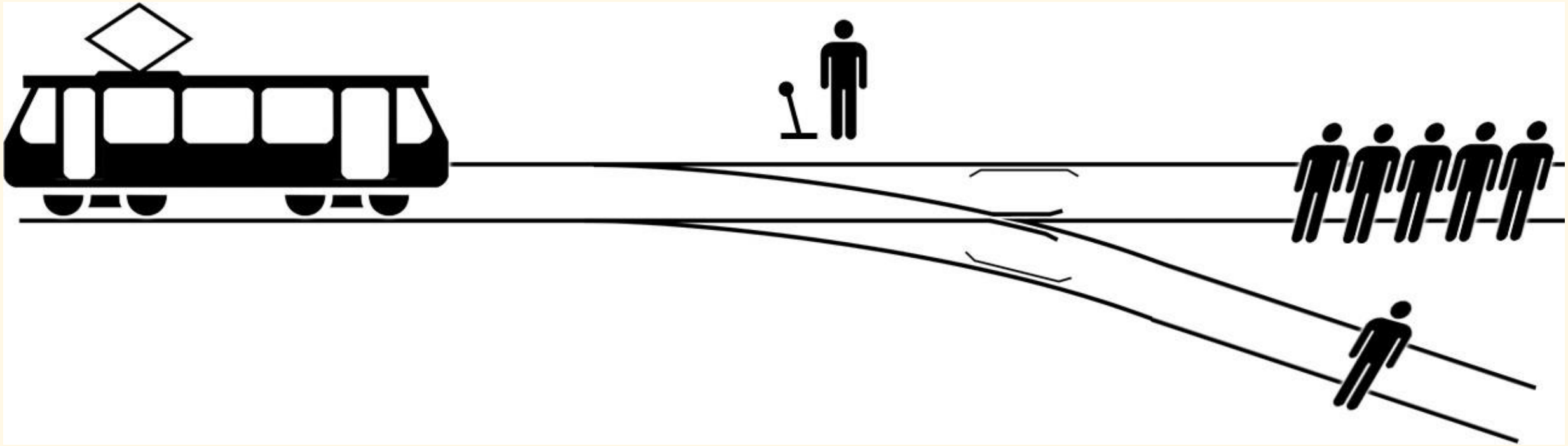
- http://tdan.com/data-warehousing-ethical-concerns-security-access-and-control/5186
- https://phlconnect.ched.gov.ph/admin/uploads/f197002b9a0853eca5e046d9ca4663d5/Fundamentals-of-Data-Warehousing-Opportunities-and-Ethics.pdf

ETHICAL ISSUES OF DATA MINING

INTRODUCTION

What is Ethics?

INTRODUCTION: WHAT IS ETHICS?



INTRODUCTION: WHAT IS ETHICS?

Code of Ethics

INTRODUCTION: WHAT IS ETHICS?

Legal Obligations

Why Ethics?

CONCERNS

Discrimination & Bias

Integrity

Lack of Transparency

Data Privacy Law

CONCERNS

Discrimination & Bias

Discrimination –
acts, practices or policies that impose a relative disadvantage on persons because of their membership of a salient social or recognized vulnerable group based on gender, race, skin color, language, religion, political opinion, ethnic minority etc.

direct or indirect relation between a protected attribute and the resulting prediction/classification/suggested decision

CONCERNS

Discrimination & Bias

direct discrimination
procedures that discriminate against minorities or disadvantaged groups on the basis of sensitive discriminatory attributes related to group membership such as race, gender or sexual orientation

indirect discrimination
procedures that might intentionally or accidentally discriminate against a minority, while not explicitly mentioning discriminatory attributes

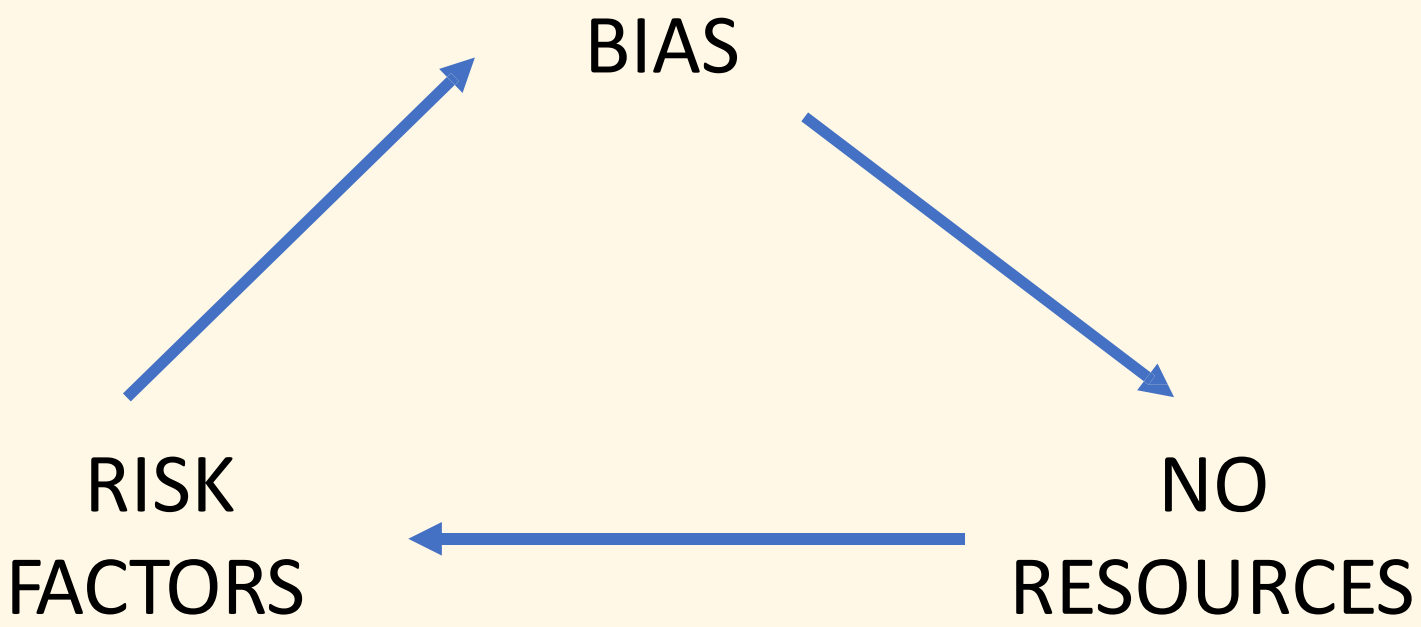
Table 4 Causes of discrimination in data analytics

From: Big Data and discrimination: perils, promises and solutions. A systematic review

Causes of discrimination	Related articles
1. Algorithmic causes	
1.1. Definition of the target variable	Barocas and Selbst 2016 [8], d'Alessandro et al. 2017 [25]
1.2. *Data issues* Training data (Historically biased data sets)	Kamiran and Calders 2012 [42], Barocas and Selbst 2016 [8], Brayne 2017 [14], d'Alessandro et al. 2017 [25]
1.3. *Data issues* Training data (manual assignment of class labels)	Barocas and Selbst 2016 [8], d'Alessandro et al. 2017 [25]
1.4. *Data issues* Data collection (Overrepresentation and underrepresentation)	Barocas and Selbst 2016 [8], d'Alessandro et al. 2017 [25]
1.5. Proxies	Schermer 2011 [73], Kamiran and Calders 2012 [42], Barocas and Selbst 2016 [8], Zliobaite and Custers 2016 [95], d'Alessandro et al. 2017 [25]
1.6. Feedback loop	Mantelero 2016 [54], Brayne 2017 [14], d'Alessandro et al. 2017 [25]
1.7. Overfitting	Kamiran and Calders 2012 [42], Mantelero 2016 [54]
1.8. Feature selection	Barocas and Selbst 2016 [8]
1.9. *Cost function* Error by omission	d'Alessandro et al. 2017 [25]
1.10 *Masking* Proxies	Peppet 2014 [61], Zarsky 2014 [93], Barocas and Selbst 2016 [8], Zliobaite and Custers 2016 [95], Kroll et al. 2017 [45]
2. Digital divide	
2.1. Skills	Boyd and Crawford 2012 [12], Casanas i Comabella and Wanat 2015[18]
2.2. Resources	Barocas and Selbst 2016 [8], Pak et al. 2017 [60]
2.3. Geographical location	Casanas i Comabella and Wanat 2015 [18], Barocas and Selbst 2016 [8], Pak et al. 2017 [60]
2.4. Age	Casanas i Comabella and Wanat 2015 [18]
2.5. Income	Barocas and Selbst 2016 [8], Pak et al. 2017 [60]
2.6 Gender	Boyd and Crawford 2012 [12]
2.7. Education	Boyd and Crawford 2012 [12]
2.8 Race	Bakken and Reame 2016 [6], Sharon 2016 [74]
3. Data linkage	Susewind 2015 [76], Cato et al. 2016 [19], Zarate et al. 2016 [91], Ploug and Holm 2017 [64]

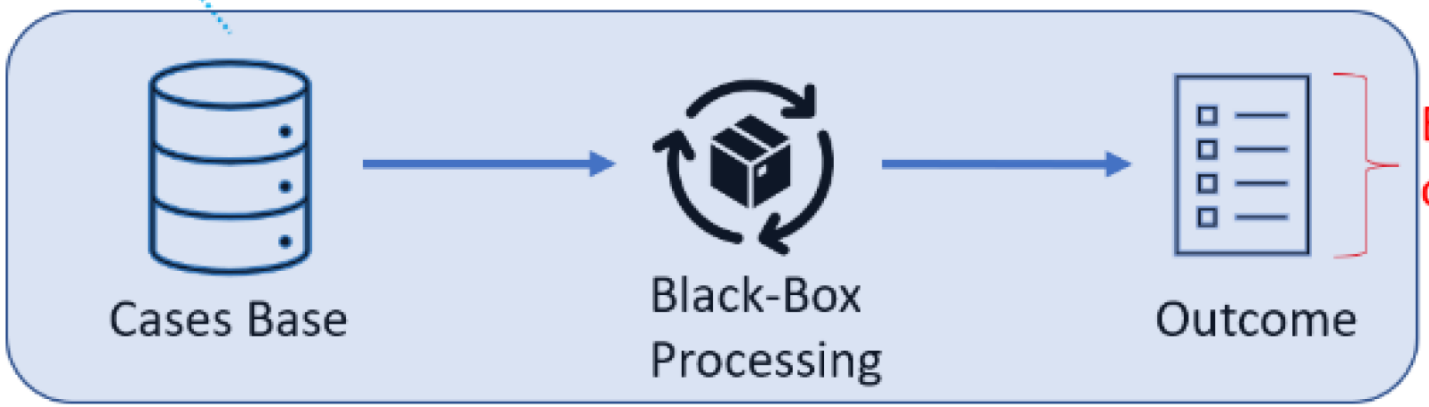
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4

CONCERNS: DISCRIMINATION



Expression of bias

Tuple	Personal Attributes								Features				Class
	Sensitive				Not-sensitive								
	Att. 1	Att.2	...	Att.n	Att. 1	Att.2	...	Att.n	F1	F2	...	Fn	CLi
T1													
T2													
...													
Tn													



CONCERNS

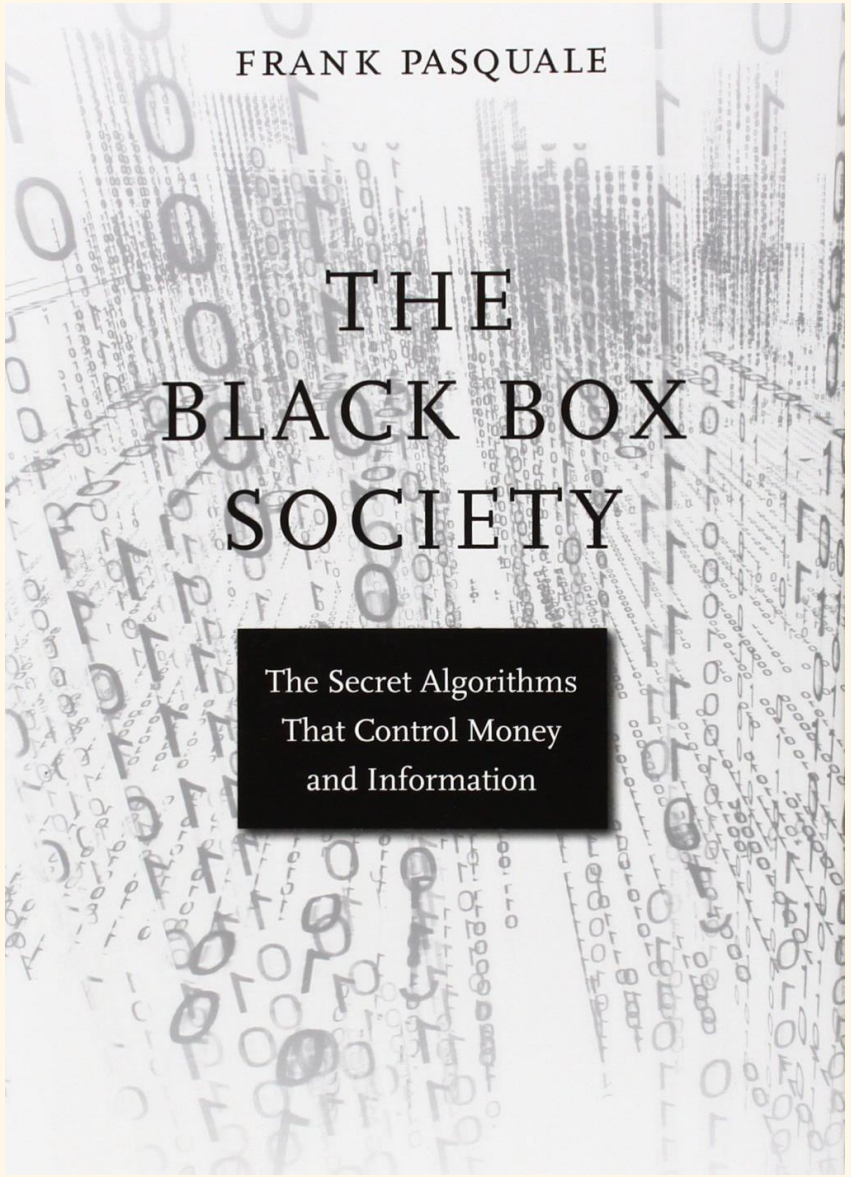
Integrity

CONCERNS

Transparency

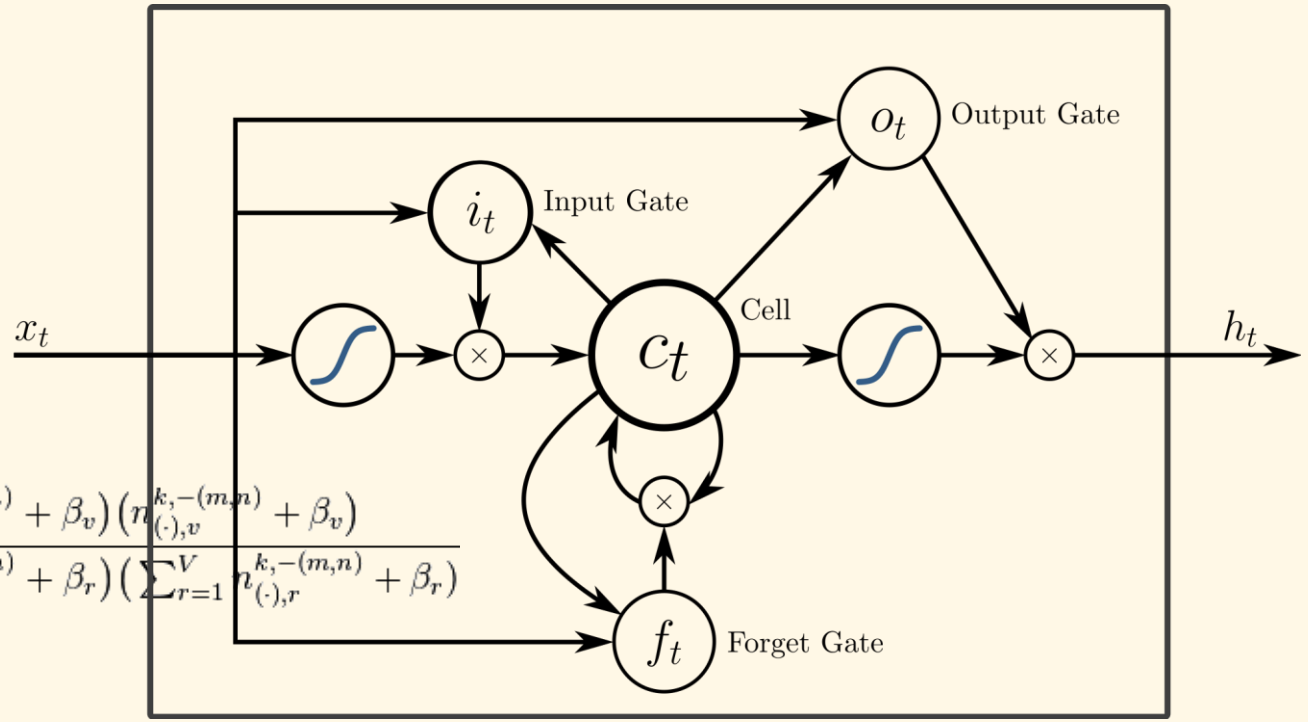
CONCERNS: TRANSPARENCY

- What is black box?
- Can be convenient
- But opaque



CONCERNS: TRANSPARENCY

$$
\begin{aligned}
&\propto \frac{\prod_{i\neq k}\Gamma(n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)}{\Gamma\big((\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)+1\big)}\prod_{i\neq k}\frac{\Gamma(n_{(\cdot),v}^{i,-(m,n)}+\beta_v)}{\Gamma\big(\sum_{r=1}^{V} n_{(\cdot),r}^{i,-(m,n)}+\beta_r\big)}\\
&\times\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1)\frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\big((\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r)+1\big)}\\
&\propto \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k+1)}{\Gamma\big((\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i)+1\big)}\frac{\Gamma(n_{(\cdot),v}^{k,-(m,n)}+\beta_v+1)}{\Gamma\big((\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r)+1\big)}\\
&= \frac{\Gamma(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k)\big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big)}{\Gamma\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)}\frac{\Gamma\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\Gamma\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)}\\
&\propto \frac{\big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big)}{\big(\sum_{i=1}^{K} n_{m,(\cdot)}^{i,-(m,n)}+\alpha_i\big)}\frac{\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)}\\
&\propto \big(n_{m,(\cdot)}^{k,-(m,n)}+\alpha_k\big)\frac{\big(n_{(\cdot),v}^{k,-(m,n)}+\beta_v\big)}{\big(\sum_{r=1}^{V} n_{(\cdot),r}^{k,-(m,n)}+\beta_r\big)}.
\end{aligned}
$$



CONCERNS

Privacy

Privacy
is when the system protects a person's identity and the integrity of data, indicates access permission and methods, data retention periods, and how data will be destroyed at the end of such period, which ensures a person's right to be forgotten

CONCERNS

Data Privacy Act

The Data Privacy Act of 2012
RA 10173

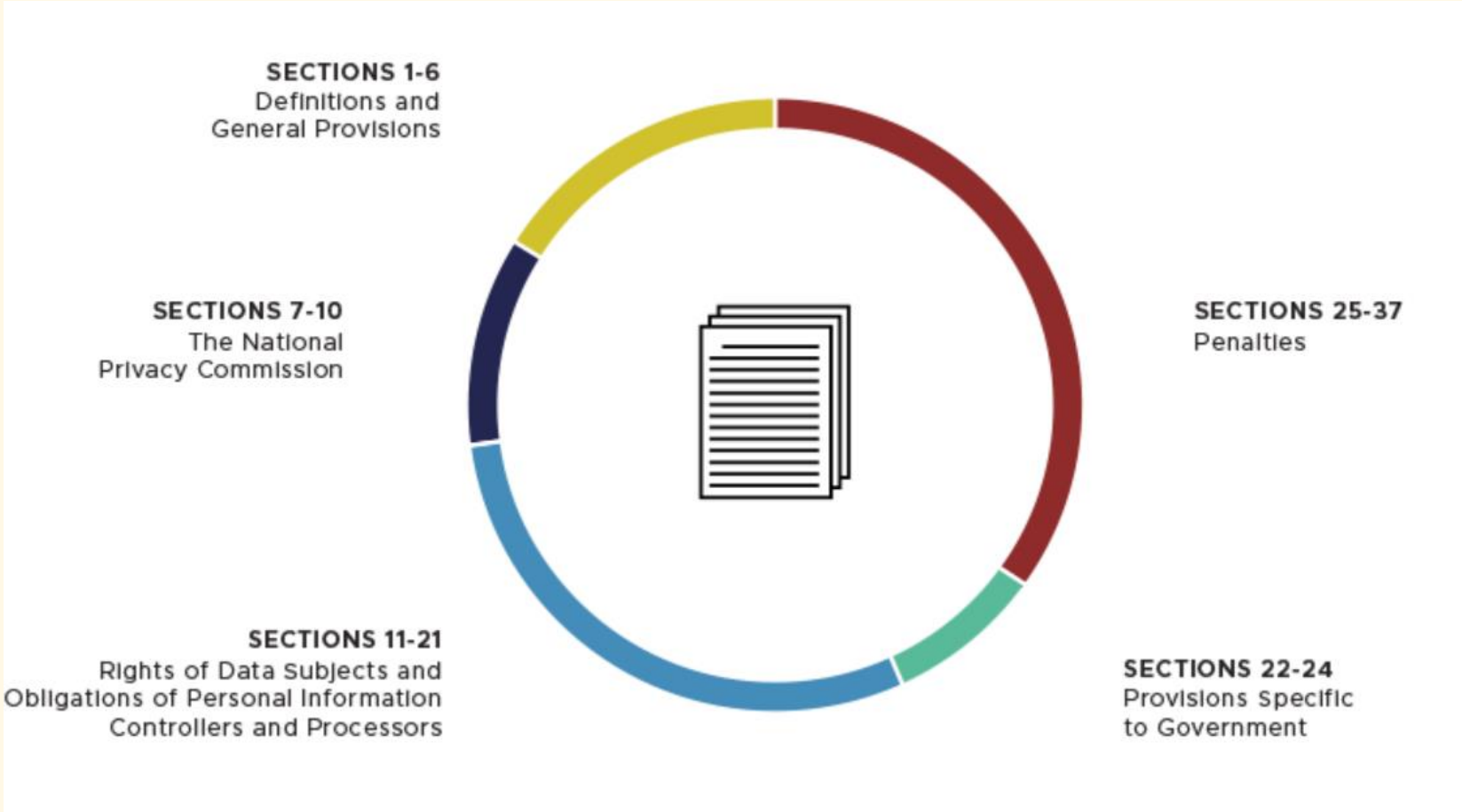
CONCERNS

What is **Data Privacy**?

The right of an individual not to have private information about himself disclosed, and to live freely from surveillance and intrusion.

Version#d

THE STRUCTURE OF THE DATA PRIVACY ACT



https://www.privacy.gov.ph/data-privacy-act-primer/

ADDRESSING CONCERNS

Consent Required

ADDRESSING CONCERNS: DATA PRIVACY ACT

The Right to be Informed

ADDRESSING CONCERNS: DATA PRIVACY ACT

The Right to Access

ADDRESSING CONCERNS: DATA PRIVACY ACT

The Right to Rectify

ADDRESSING CONCERNS: DATA PRIVACY ACT

The Right to Complain

ADDRESSING CONCERNS: DATA PRIVACY ACT

Data Protection

SUMMARY

- Ethical dilemmas
- Discrimination/Bias, Integrity, Transparency, Privacy
- Professional Codes/Voluntary Codes
- Legal Duties under the Data Privacy Act

REFERENCES

- https://networks.upou.edu.ph/11909/ethical-implications-of-business-analytics-dominic-ligot/
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0177-4
- https://encyclopedia.pub/entry/24727
- https://www.privacy.gov.ph/data-privacy-act-primer/

FINAL PROJECT REQUIREMENTS: CASE STUDY #2

RESEARCH PAPER ON DATA MINING/ DATA WAREHOUSING

- **The paper should include:**
  - **Title: The title should be specific and indicate the problem the research project addresses using keywords that will be helpful in literature reviews in the future**
  - **Abstract: It describes the main synopsis of your paper. The abstract is used by readers to quickly review the overall content of the paper. Journals typically place strict word limits on abstracts, such as 200 words, making them a challenge to write. The abstract should provide a complete synopsis of the research paper and should introduce the topic and the specific research question, provide a statement regarding methodology and should provide a general statement about the results and the findings. Because it is really a summary of the entire research paper, it is often written last.**

FINAL PROJECT REQUIREMENTS: CASE STUDY #2

RESEARCH PAPER ON DATA MINING/ DATA WAREHOUSING

- **The paper should include:**
  - **Introduction: It provides background information necessary to understand the research and getting readers interested in your subject. The introduction is where you put your problem in context. It begins by introducing the broad overall topic and providing basic background information. It then narrows down to the specific research question relating to this topic. It provides the purpose and focus for the rest of the paper and sets up the justification for the research.**
  - **Related Works**
  - **Methods : Describe your methods here. Provide as much detail as possible regarding conduct of your experiments etc. Summarize the algorithms generally, identify and discuss procedures, tools, highlight features relevant to your project, and refer readers to your references for further details.**

FINAL PROJECT REQUIREMENTS: CASE STUDY #2

RESEARCH PAPER ON DATA MINING/ DATA WAREHOUSING

- **The paper should include:**
  - **Results and Discussion : This section is the most important part of your paper. It is here that you demonstrate the work you have accomplished on this project and explain its significance. The quality of your analysis will impact your final grade more than any other component on the paper. You should therefore plan to spend the bulk of your project time not just gathering data, but determining what it ultimately means and deciding how best to showcase these findings.**

FINAL PROJECT REQUIREMENTS: CASE STUDY #2

RESEARCH PAPER ON DATA MINING/ DATA WAREHOUSING

- **The paper should include:**
  - **Conclusion: The conclusion should give your reader the points to “take home” from your paper. It should state clearly what your results demonstrate about the problem you were tackling in the paper. It should also generalize your findings, putting them into a useful context that can be built upon. All generalizations should be supported by your data, however; the discussion should prove these points, so that when the reader gets to the conclusion, the statements are logical and seem self-evident.**

FINAL PROJECT REQUIREMENTS: CASE STUDY #2

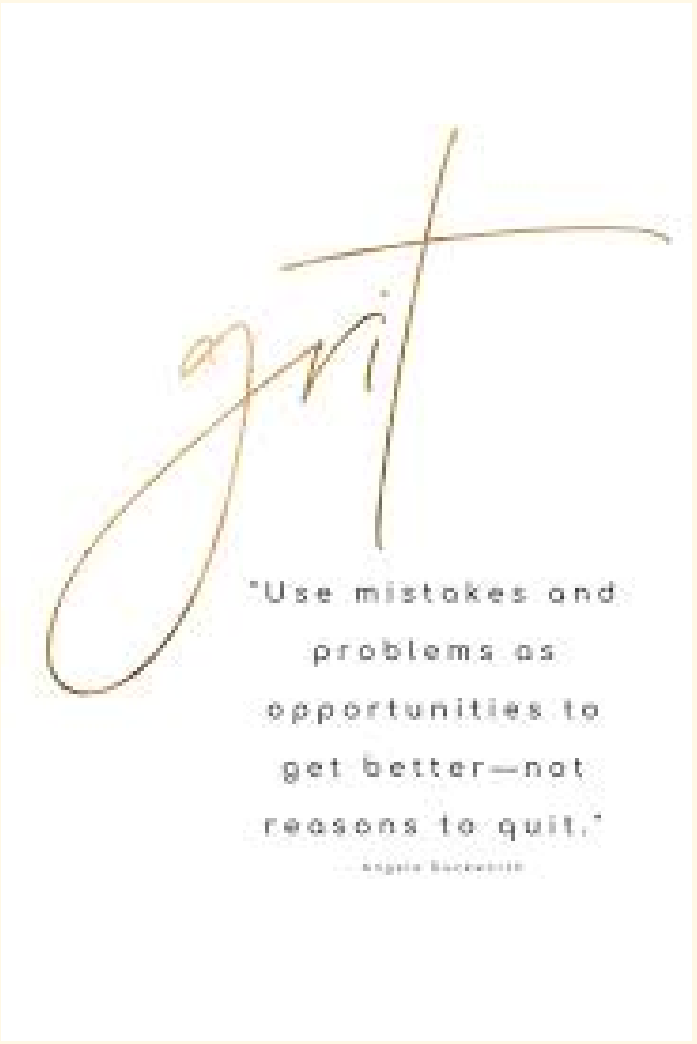
RESEARCH PAPER ON DATA MINING/ DATA WAREHOUSING

- **The paper should include:**
  - **Future Works/Recommendations: This section is a place for you to explain to your readers where you think the results can lead you. What do you think are the next steps to take? Recommendations are the added suggestions that you want people to follow when performing future studies.**
  - **Bibliography: Refer to any reference that you used in your assignment. Citations in the body of the paper should refer to a bibliography at the end of the paper.**

LOOK FOR VALID SOURCES

- ieeexplore.ieee.org
- http://scholar.google.com/
- http://citeseerx.ist.psu.edu
- https://dl.acm.org/
- http://www.sciencedirect.com/


Great works are performed, not by strength, but by perseverance.
- Samuel Johnson
HAPPIERHUMAN.COM


grit
"Use mistakes and problems as opportunities to get better—not reasons to quit."

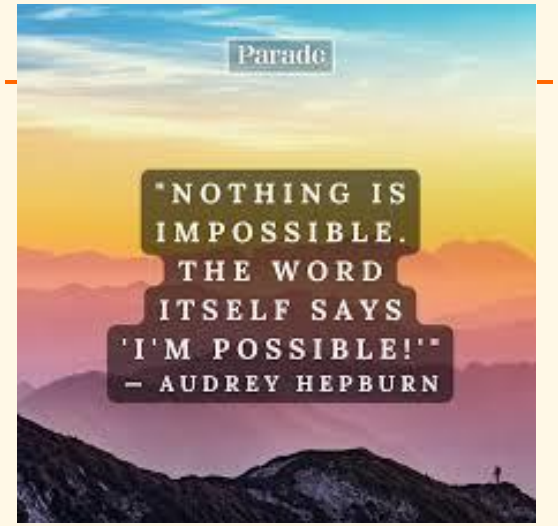

Grit is living life like it's a marathon, not a sprint.
Angela Duckworth
BrainyQuote


INSIGHT
Grit is the "extra something" that separates the most successful people from the rest. It's the passion, perseverance, and stamina that we must channel in order to stick with our dreams until they become a reality.
Travis Bradberry - Biographer

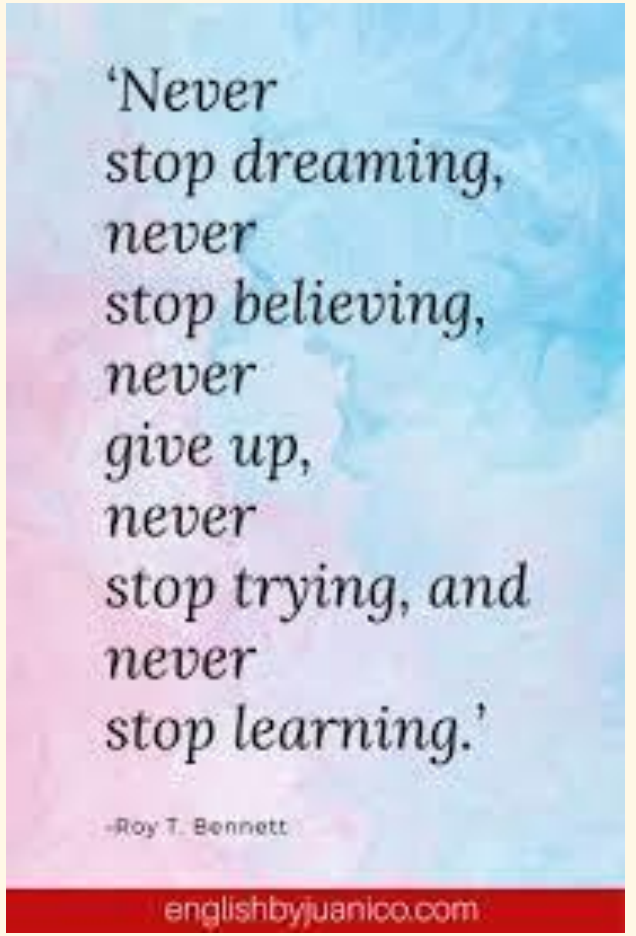

DEVELOP GRIT!
Excellence does not mean perfection; rather, it means never giving in, never giving up, and ultimately achieving the goal


G.R.I.T.
GUTS, RESILIENCE, INITIATIVE, TENACITY


"The best way to predict your future is to create it."
—Abraham Lincoln
peppyzing.com


Parade
"NOTHING IS IMPOSSIBLE. THE WORD ITSELF SAYS 'I'M POSSIBLE!'"
— AUDREY HEPBURN


TRUST YOURSELF
YOU KNOW MORE
THAN YOU THINK
YOU DO!



THANK YOU AND ALL THE BEST! ☺

JENNIFER L. LLOVIDO