UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Dark Text Arts

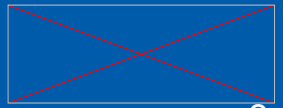
Inputs:
- A Modupe [PhD Candidate]
- A Moodley [MIT Big Data

Dr. Vukosi Marivate

Make today matter

Data Science for Social Impact

Dark Text Arts?



Your thoughts

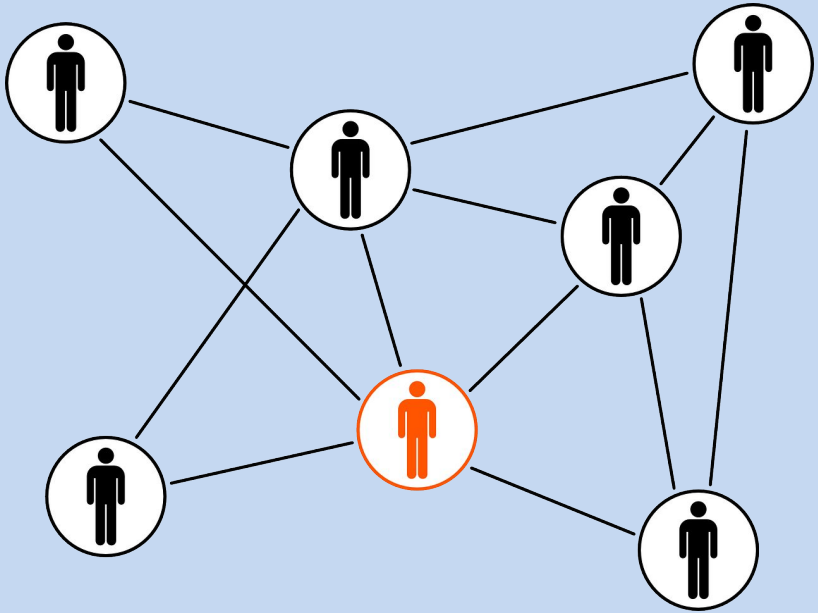
Can you think of a few examples of **Dark Text Arts**?

The Internet

Has brought unprecedented connection of people.

Sharing of information.

Shaping society's evolution.



The Internet + People

The internet can be a dark place

- Spam
- Hate Speech
- Sexual Predators
- Misinformation

Abuse

SPAM

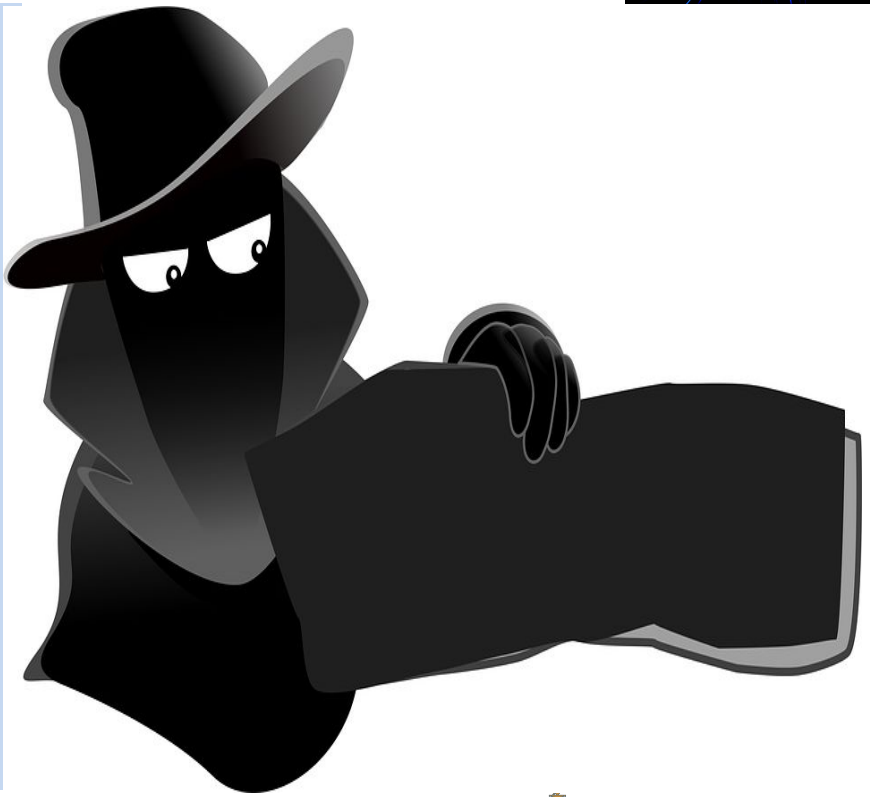
- Email

Gaming Social Systems

- User generated content
- Fake Reviews

Abuse - Chat

Sexual Predators



Abuse - Speech/Comments

Hate Speech

Stop outsourcing the regulation of hate speech to social media

March 27, 2019 10.51pm GMT

Regulating hate speech on the internet should not be left to private corporations. Shutterstock

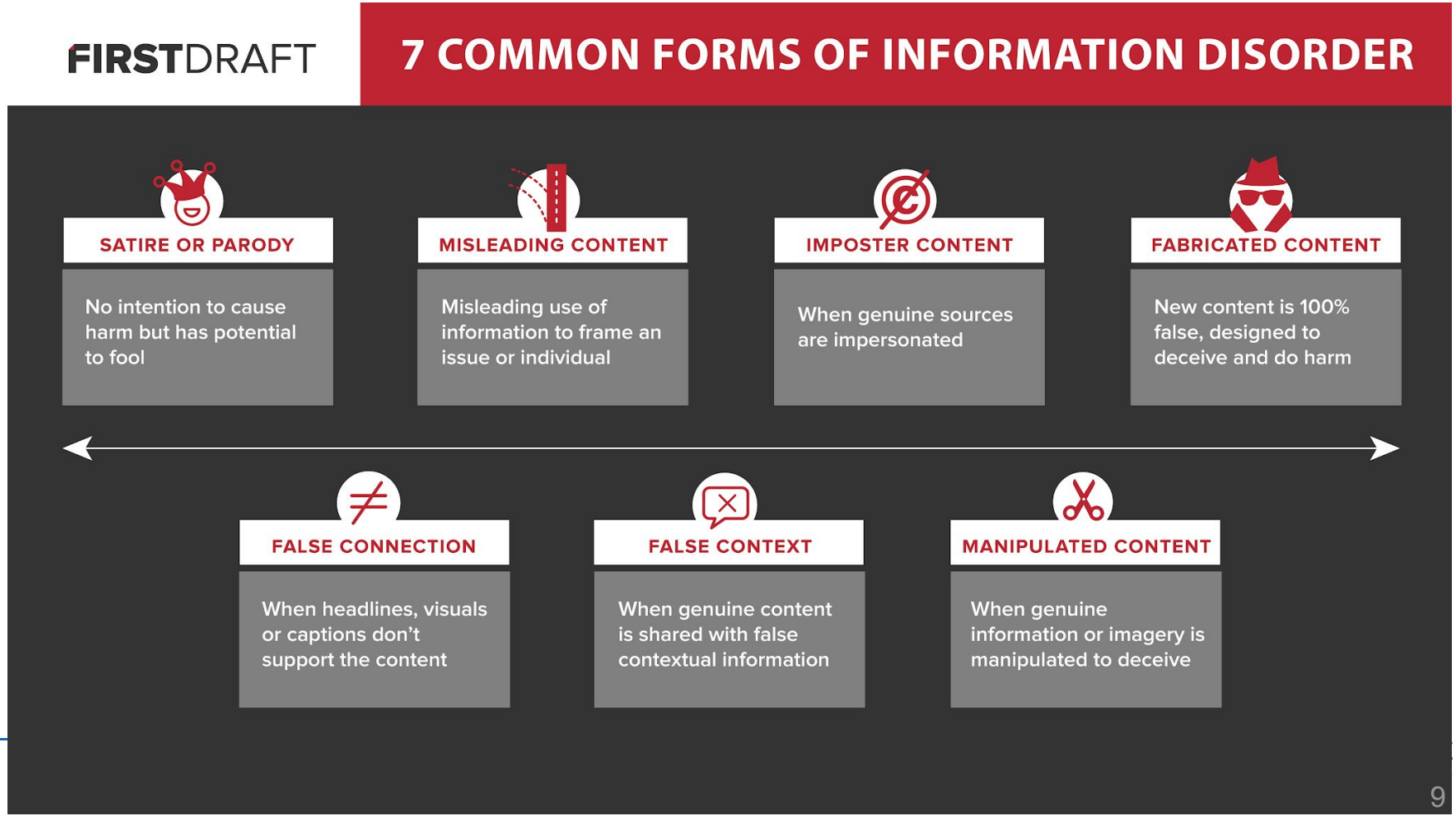
Email
Twitter 9
Facebook 228
LinkedIn
Print

When it comes to dealing with online hate speech, we've ended up in the worst of all possible worlds.

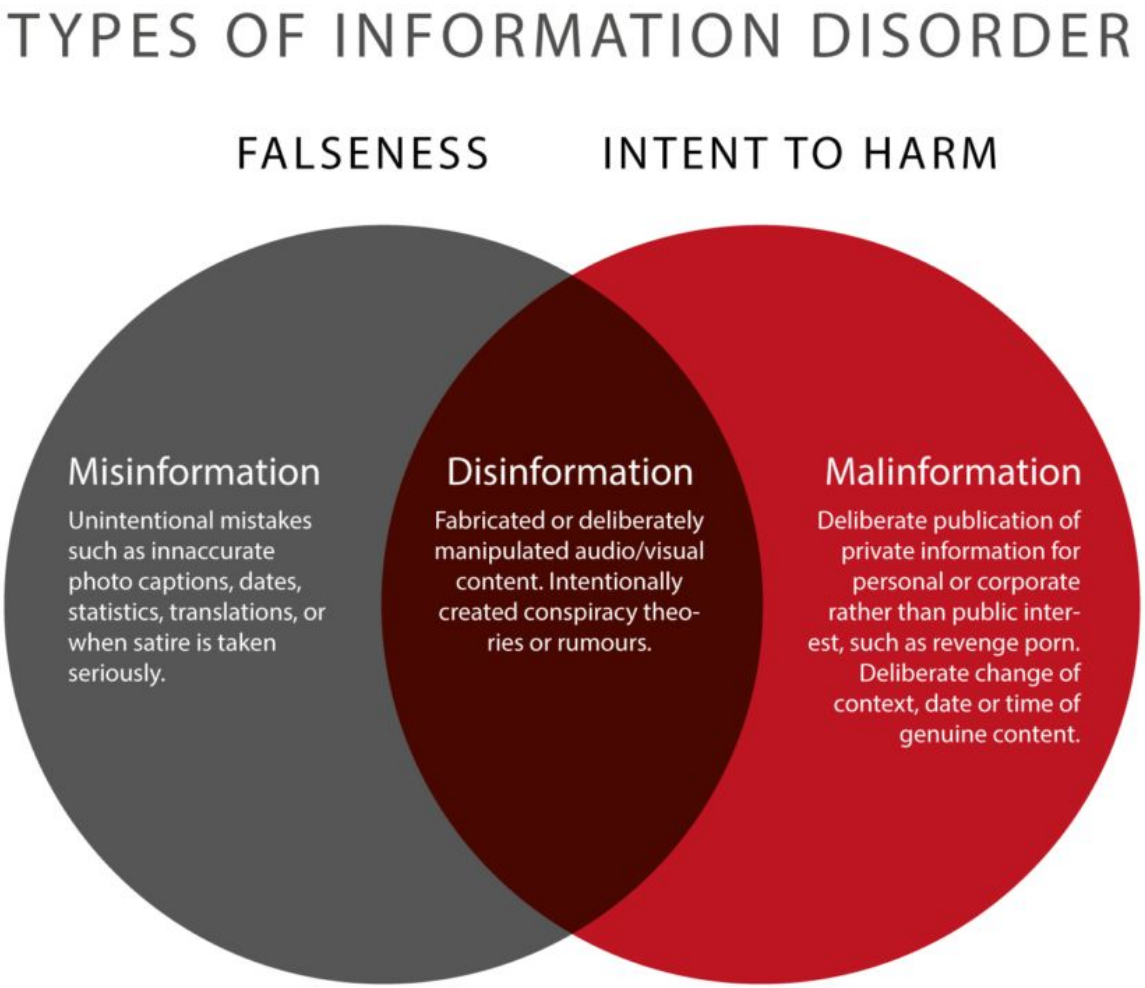
On the one hand, you have social media platforms like Facebook and Twitter that seem extremely reluctant to ban white supremacists and actual neo-Nazis, but enthusiastically enforce their own capricious terms of service to keep adults safe from such harmful things as the female nipple. That is, until something horrific happens, such as the

https://theconversation.com/stop-outsourcing-the-regu

Disinformation



Fake News


TYPES OF INFORMATION DISORDER
FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

Others?

Do you have any other examples?

Course Overview

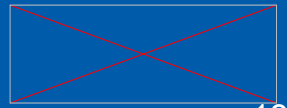
Understanding

- Disinformation
- Abuse
- Fake News etc.

Through

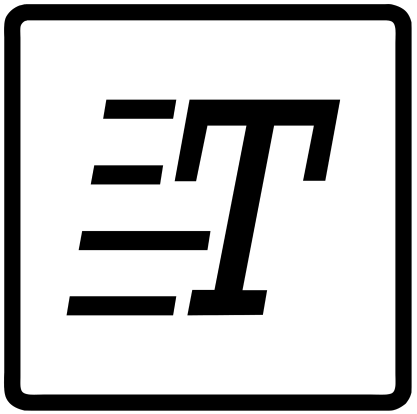
- Natural Language Processing
- Machine learning
  - *NB: This class **assumes** you are comfortable with **Machine Learning.***

Datasets and Problems



Interesting Dark Text Problems?

- Online Abuse
  - Hate Speech
  - Sexual Predators
- Deception
  - Multi-Author Text
- Disinformation
- Fake News



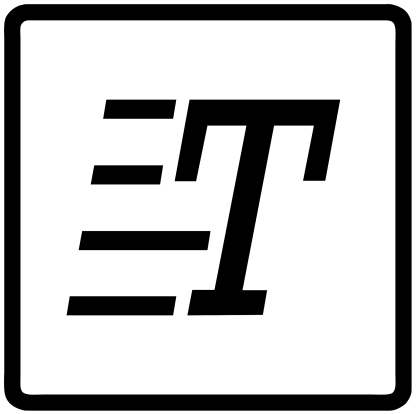
Where is the Data? Some Examples

General

- NLP Progress https://nlpprogress.com/

Hate Speech

- Hate speech dataset from a white supremacist forum https://github.com/aitor-garcia-p/hate-speech-dataset



Fake News

- Fake News https://github.com/KaiDMML/FakeNewsNet
- Real 411 https://www.real411.org/complaints [South Africa]

Author Identification

- Style Change https://pan.webis.de/clef19/pan19-web/style-change-detection.html

Data Science Approach

Exploratory Data Analysis

- Word Clouds
- Topic Modelling
- Top Phrases, Words
- Etc.
-

Problem Formulation and Iteration

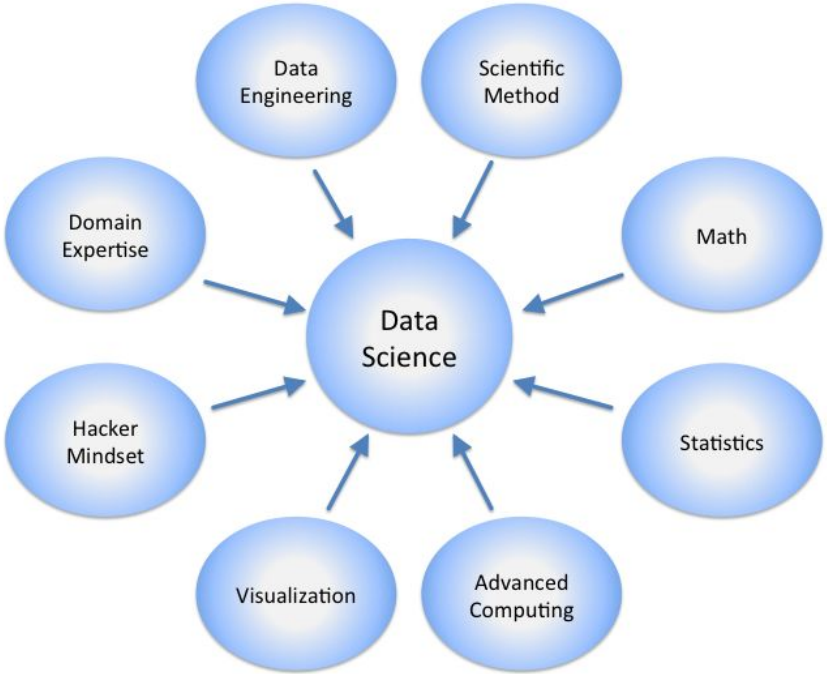
- What are you trying to answer/find?

Modelling

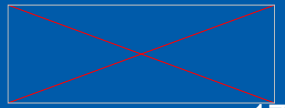
- Classification
- Categorisation
- Behavioural Modelling etc.

Communicate/Visualise/Deploy

Generate More Questions



Where to from here?



Projects

Early Details [More details will be available later]

- Will be an individual effort
- Find an interesting problem
  - Good Problem
  - Good Data
  - Good Approach
- You will submit ideas early in the project process
- Work on the project
- Submit Report
- Present findings

Resources

NLP General

- https://github.com/fastai/course-nl
- Stanford Coursera NLP Slides
  https://web.stanford.edu/~jurafsky/NLPCourseraSlides.html

Python Libraries

- SKLearn NLP (Working With Text Data) - URL (Nice tutorial)
- spaCY: Industrial-Strength Natural Language Processing - URL
- NLTK

Thank You

Dr. Vukosi Marivate
vukosi.marivate@cs.up.ac.za
https//dsfsi.github.io
@vukosi