Title: A Matter of Tech Security
Predicting trends from podcasts

PRESENTER:
Isa Lykke Hansen

BACKGROUND:

The podcast **Security Now** has been running every week since mid-2005. Here it is used as a novel source of NLP data.

Using **LDA**, transcripts of the podcasts can be used to investigate how topics on tech security change over time. Could it be that Leo and Steve know about a new trend before Google does?

PIPELINE:

Scraping .txt files using **beautiful soup**

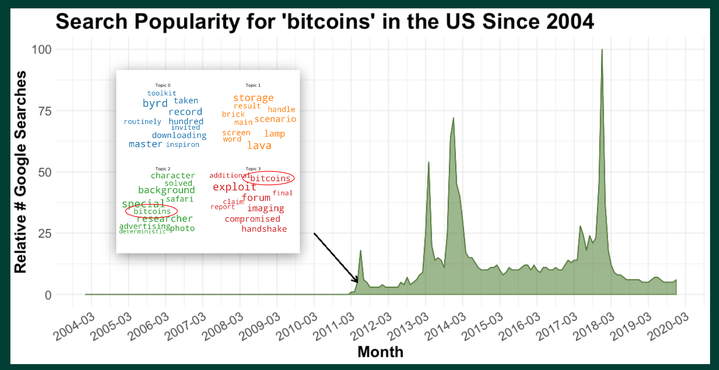
(Meta)data extraction in **R**.

Text cleanup using **nltk**. Stop word removal + lemmatization

LDA topic modelling using **gensim**

Visualization using **matplotlib** and **ggplot**

Can these two nerds predict the future?



Visit the github repository for this project

Get the full experience! Listen to the podcast

WORD2VEC

words most similar to 'woman':
[('pregnant', 0.8298099040985107), ('girl', 0.7716690897941589), ('leash', 0.7452483177185059), ('nephew', 0.7448058724403381), ('elderly', 0.7404824495315552), ('boyfriend', 0.7390583157539368), ('eating', 0.7377386093139648), ('perez', 0.7341799139976501), ('amusement', 0.7306317090988159), ('shatner', 0.729703426361084)]
words most similar to guy':
[('engineered', 0.712701678276062), ('prop', 0.6763680577278137), ('hacker', 0.6576520204544067), ('folk', 0.6571478247642517), ('incentivize', 0.6543788313865662), ('jerk', 0.6449341177940369), ('sabri', 0.6437070369720459), ('academician', 0.6413080096244812), ('goose', 0.6408295631408691), ('overselling', 0.6370724439620972)]

THINGS TO LOOK INTO

- Informative **cutoffs**
- Number of **topics**
- **Bigram** models (e.g "crypto currency")
- _____
- _____
- _____
- _____
- _____
- _____
- _____
- _____

DATA STATEMENT

*Data consists of transcripts of **730** "Security Now" podcast episodes, scraped from the public website www.grc.com/securitynow.htm. Data was gathered strictly for academic research purposes. Speaker demographic is predominantly **homogenic** (white, male, American, 30+) and consistent throughout the dataset. The texts are of semi-unstructured, conversational nature, and the topic range is narrow; each episode deals with issues of personal computer security, but due to the format are sometimes more private in nature. Data was scraped for the period of **August 2015** to **October 2019**.*

AARHUS UNIVERSITY