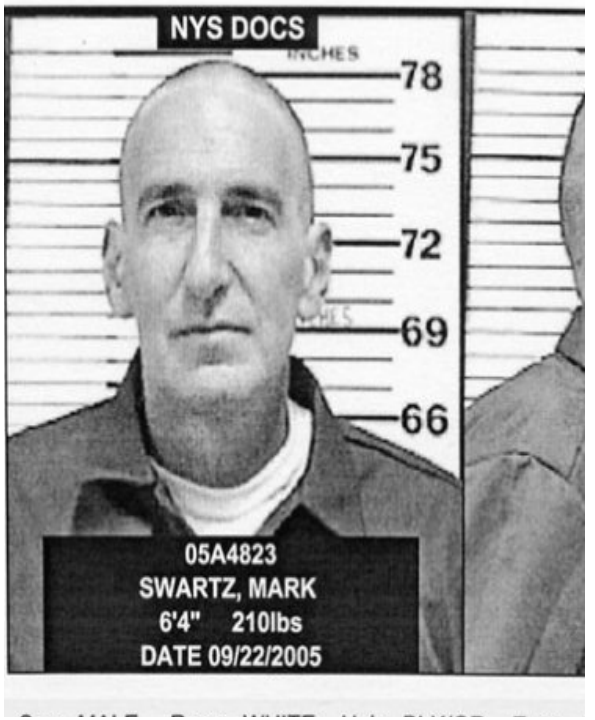
ALGORITHMIC ACCOUNTABILITY REPORTING:

On the investigation of black boxes

NICHOLAS DIAKOPOULOS, PH.D.

Introdução

- *“ Code is law ”*
- Consequências dos algoritmos em nossas vidas:
  - Negativas
  - Positivas
- Algoritmos e Transparência
- Engenharia Reversa



O poder dos algoritmos de decisão

Priorização

- Ordenar entidades de acordado com a prioridade
- Critérios de ordenação.
- Viés.
- Discriminação e preconceito.
- Critérios não são publicados. Intencional?

Classificação

- Classificar entidades em classes
- Machine learn ou Clustering
- Bias
- Dois tipos de erros em algoritmos de classificação
- Exemplo: Google's Content ID

Associação

- Relacionamento entre entidades.
- Exemplo: Wikepedia automatic Hyperlink
- Extração de informação semântica e conotativa
- Problemas com critérios e FP/FN

Filtragem

- Inclui ou exclui informação de acordo com algumas regras/critérios.
- Exemplos: apps de noticias personalizadas
- Usa-se para destacar ou censurar certas informações
- Caso: sistema de censura do Weibo

Responsabilidade algorítmica

Transparência

- Falsa transparência
- Sigilo do informação como negocio
- Grandes dificuldades e desafios (confiabilidade, etc)

Engenharia Reversa (Teoria)

- Alternativa de descobrir o poder de um algoritmo quando não se pode contar com a transparência
- É possível descobrir o efeito do poder de algoritmos intencionais e não intencionais
- Falar com o desenvolvedor do sistema trás informações uteis, alternativa a engenharia reversa

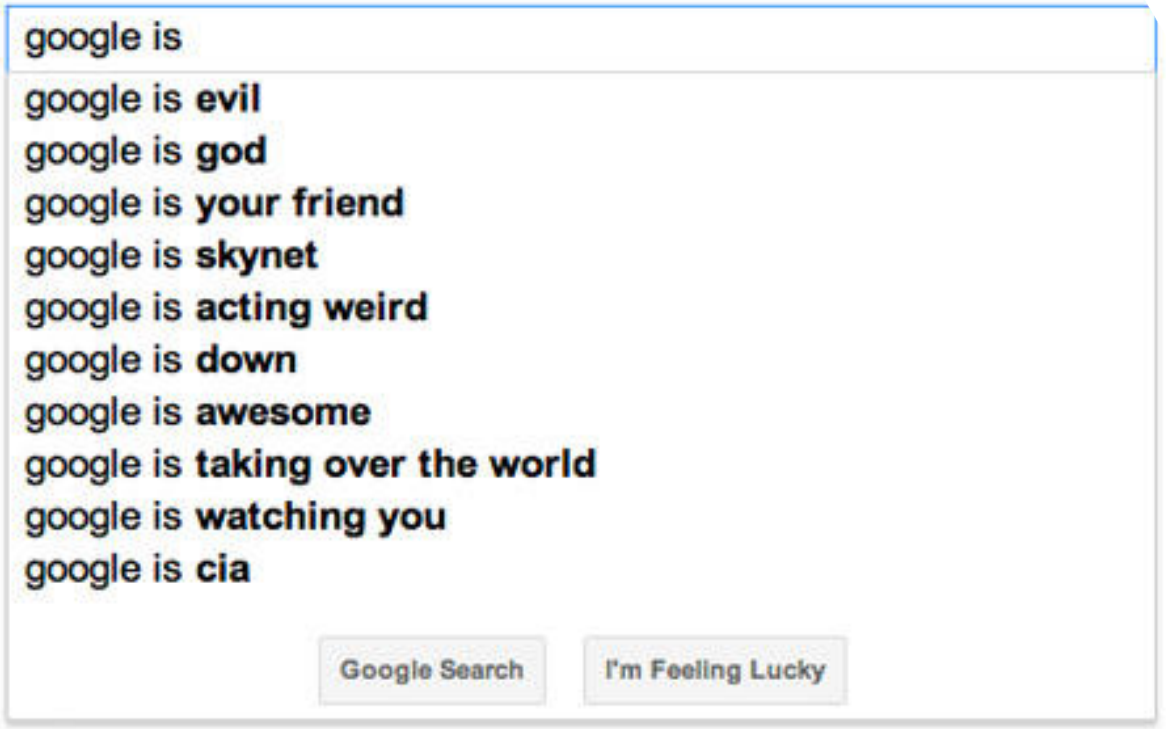
Engenharia Reversa (Teoria)

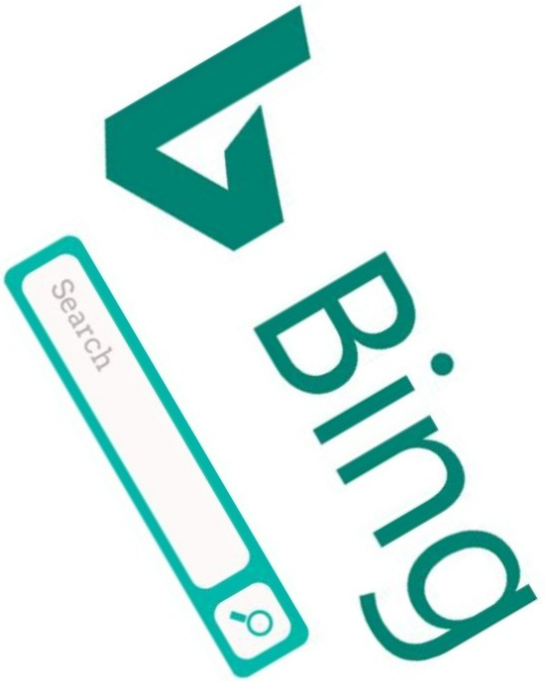
Engenharia Reversa (Teoria)

Engenharia Reversa (Na pratica)

Auto-conclusões no Google e no Bing

Google





Engenharia Reversa (Na pratica)

Autocorreção no iPhone

Engenharia Reversa (Na pratica)

E-mails políticos direcionados

Engenharia Reversa (Na pratica)

Discriminação de preços no comércio on-line

Como investigar a Responsabilidade de um Algoritmo?

Identificação

- Quais as consequências e impacto do algoritmo para o publico?
- Quantas pessoas são afetadas?
- Tem potencial para discriminação?
- FP? FN?
- Suposições

Amostragem

- Escolher amostar a relação de entrada saída de um algoritmo é o desafio chave;
- O que você *pode* amostrar vs o que *você gostaria*;
- As amostras tem que representar o público;
- Usar engenharia reversa;

Encontrando a historia

- Procurar e filtrar por insight interessantes;

Conclusão

Perguntas?