

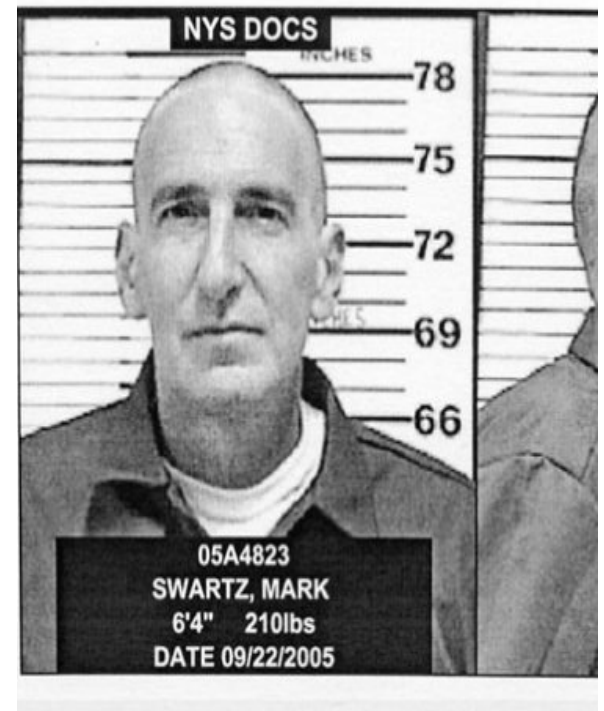
ALGORITHMIC ACCOUNTABILITY REPORTING:

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 acordo 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: Wikipedia 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)

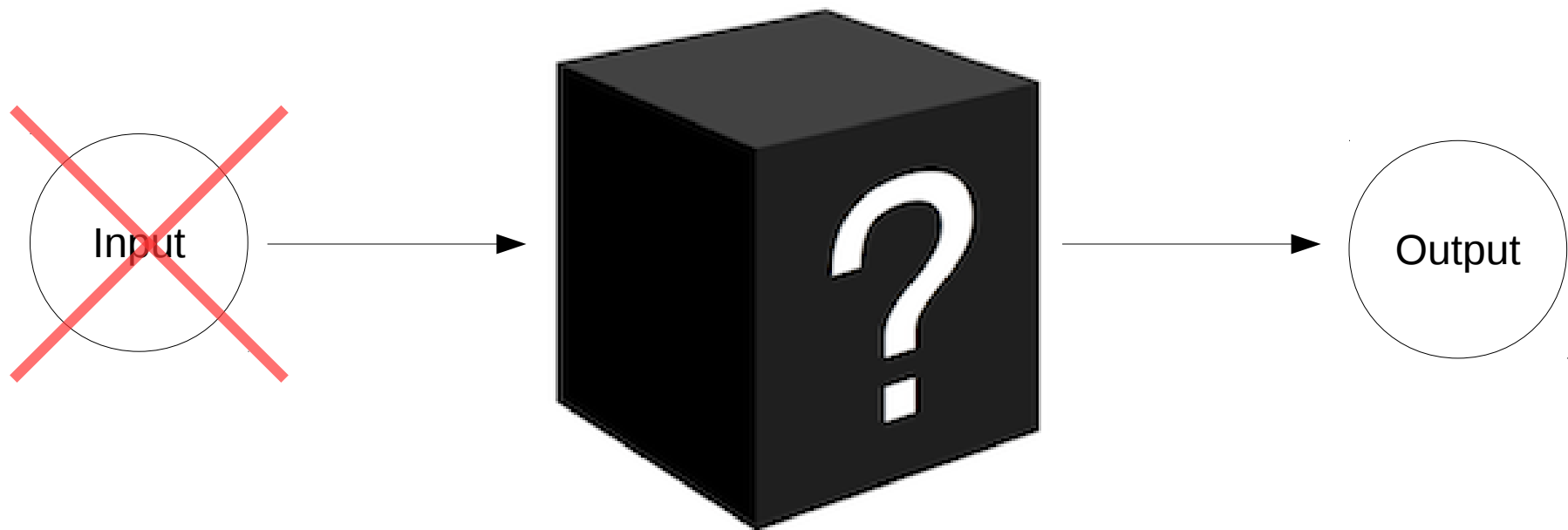
Engenharia Reversa (Teoria)

- 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

The Google logo, consisting of the word "Google" in its characteristic multi-colored font.

google is

google is **evil**

google is **god**

google is **your friend**

google is **skynet**

google is **acting weird**

google is **down**

google is **awesome**

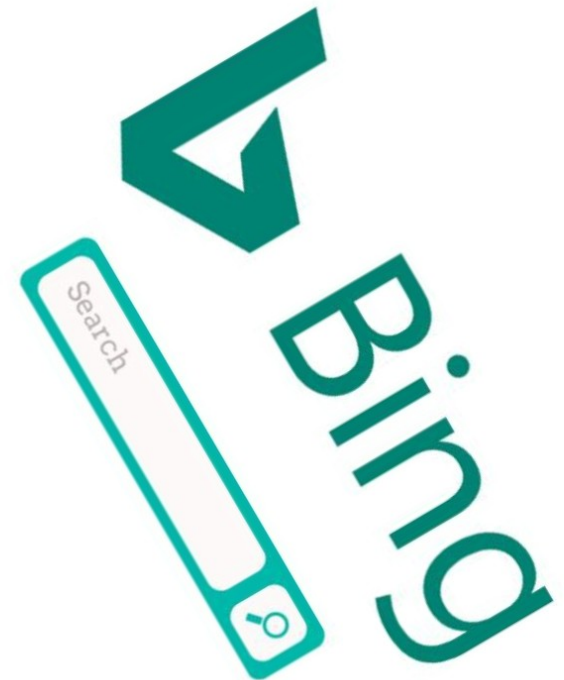
google is **taking over the world**

google is **watching you**

google is **cia**

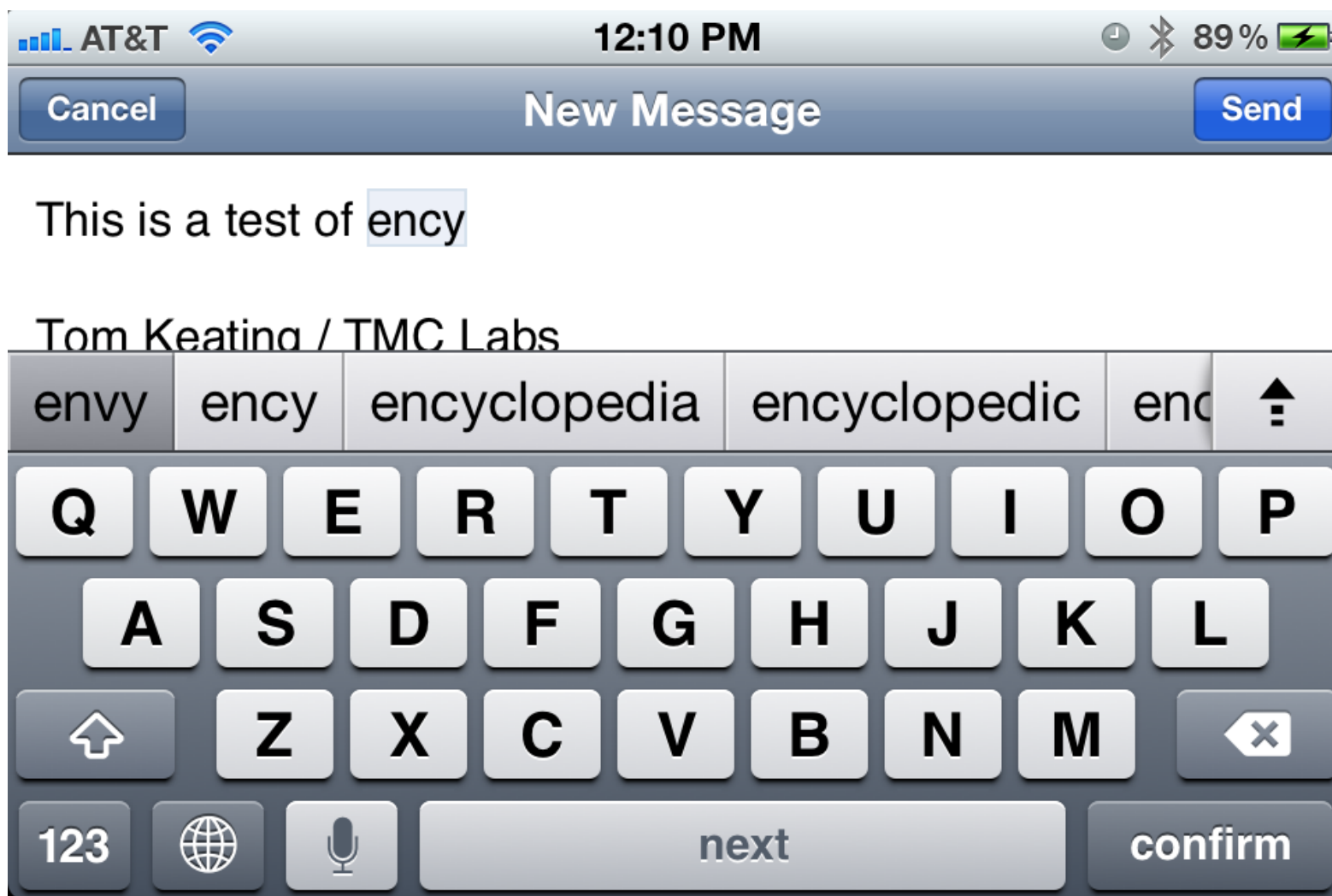
Google Search

I'm Feeling Lucky



Engenharia Reversa (Na pratica)

Autocorreção no iPhone



Engenharia Reversa (Na pratica)

E-mails políticos direcionados



<input type="checkbox"/>	5/10/11	★ SUBJECT:	Fixing what's broken
<input type="checkbox"/>	6/15/11	★ SUBJECT:	Dinner?
<input type="checkbox"/>	8/31/11	★ SUBJECT:	Frustrated
<input type="checkbox"/>	9/14/11	★ SUBJECT:	Let's meet
<input type="checkbox"/>	9/30/11	★ SUBJECT:	Stronger for it
<input type="checkbox"/>	12/31/11	★ SUBJECT:	Hey
<input type="checkbox"/>	3/27/12	★ SUBJECT:	If you're ready
<input type="checkbox"/>	3/31	★ SUBJECT:	Hey
<input type="checkbox"/>	4/30	★ SUBJECT:	Last call
<input type="checkbox"/>	5/11	★ SUBJECT:	My best friend
<input type="checkbox"/>	5/22	★ SUBJECT:	Wow
<input type="checkbox"/>	5/31	★ SUBJECT:	Not going to happen
<input type="checkbox"/>	5/31	★ SUBJECT:	Hey
<input type="checkbox"/>	5/31	★ SUBJECT:	Hey again
<input type="checkbox"/>	5/31	★ SUBJECT:	Aloha
<input type="checkbox"/>	6/11	★ SUBJECT:	I'm saving you a seat
<input type="checkbox"/>	6/11	★ SUBJECT:	Meet me for dinner
<input type="checkbox"/>	6/15	★ SUBJECT:	Rain check?
<input type="checkbox"/>	6/25	★ SUBJECT:	I will never stop fighting
<input type="checkbox"/>	6/28	★ SUBJECT:	Say you're with me
<input type="checkbox"/>	6/28	★ SUBJECT:	Today
<input type="checkbox"/>	6/28	★ SUBJECT:	Change is possible
<input type="checkbox"/>	6/30	★ SUBJECT:	To be frank
<input type="checkbox"/>	6/30	★ SUBJECT:	This is important
<input type="checkbox"/>	7/26	★ SUBJECT:	Hey
<input type="checkbox"/>	7/26	★ SUBJECT:	I don't get to tell you this enough
<input type="checkbox"/>	7/26	★ SUBJECT:	How grateful I am
<input type="checkbox"/>	7/31	★ SUBJECT:	So
<input type="checkbox"/>	7/31	★ SUBJECT:	This is critical
<input type="checkbox"/>	8/2	★ SUBJECT:	Are you in?
<input type="checkbox"/>	8/2	★ SUBJECT:	Say you're with me
<input type="checkbox"/>	8/9	★ SUBJECT:	This isn't going to stop

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?

