Big Data, Machine Learning & the Social Sciences

Hanna Wallach

Microsoft Research & University of Massachusetts Amherst

hanna@dirichlet.net

DATA

QUESTIONS

MODELS

FINDINGS

DATA

QUESTIONS

MODELS

FINDINGS

What is Big Data?

> [B]ig data [...] refers to large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available.
>
> — NSF & NIH

What is Big Data?

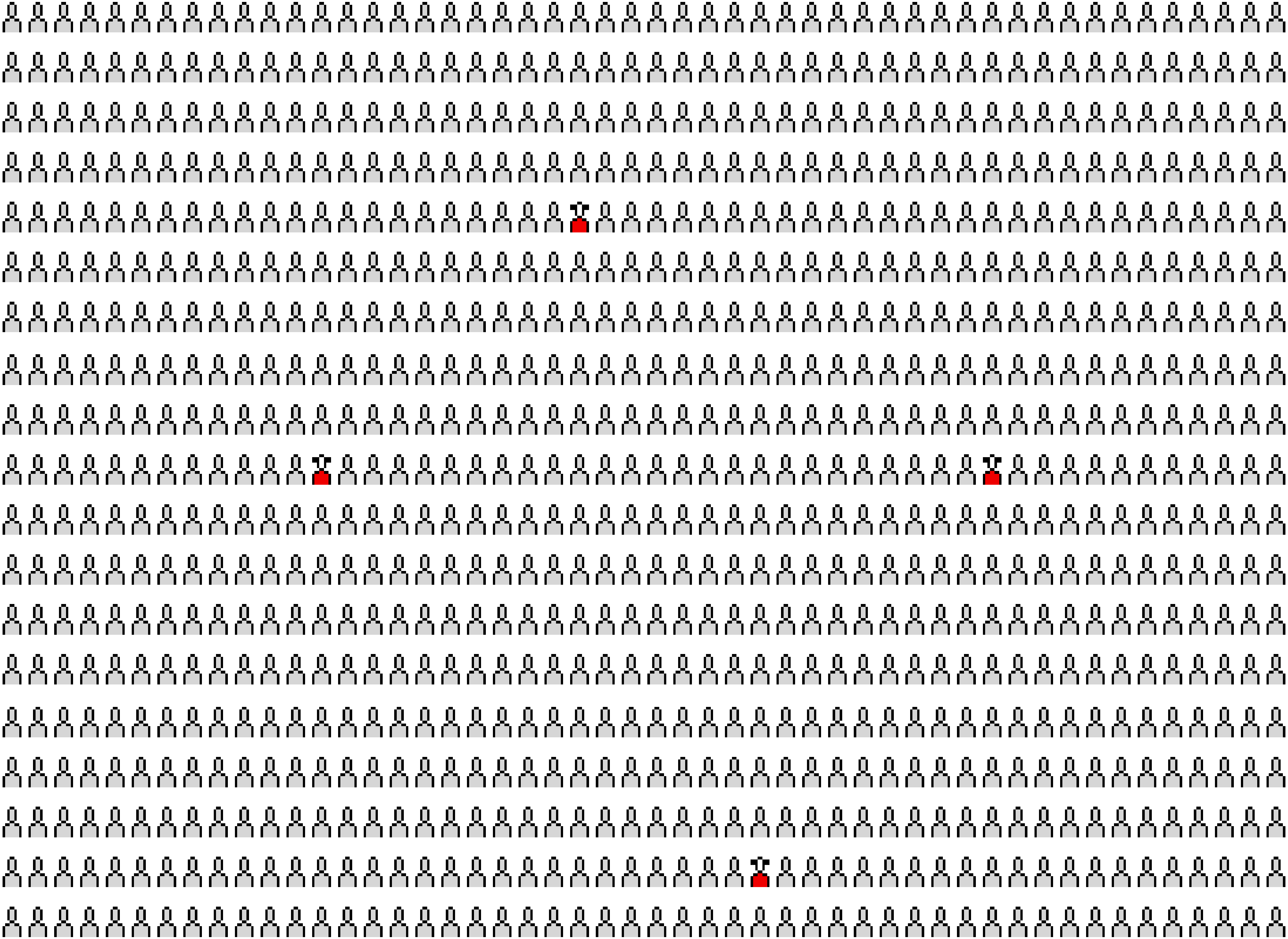
> “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making [and] insight discovery.
>
> — Gartner, Inc., 2012

Social Data

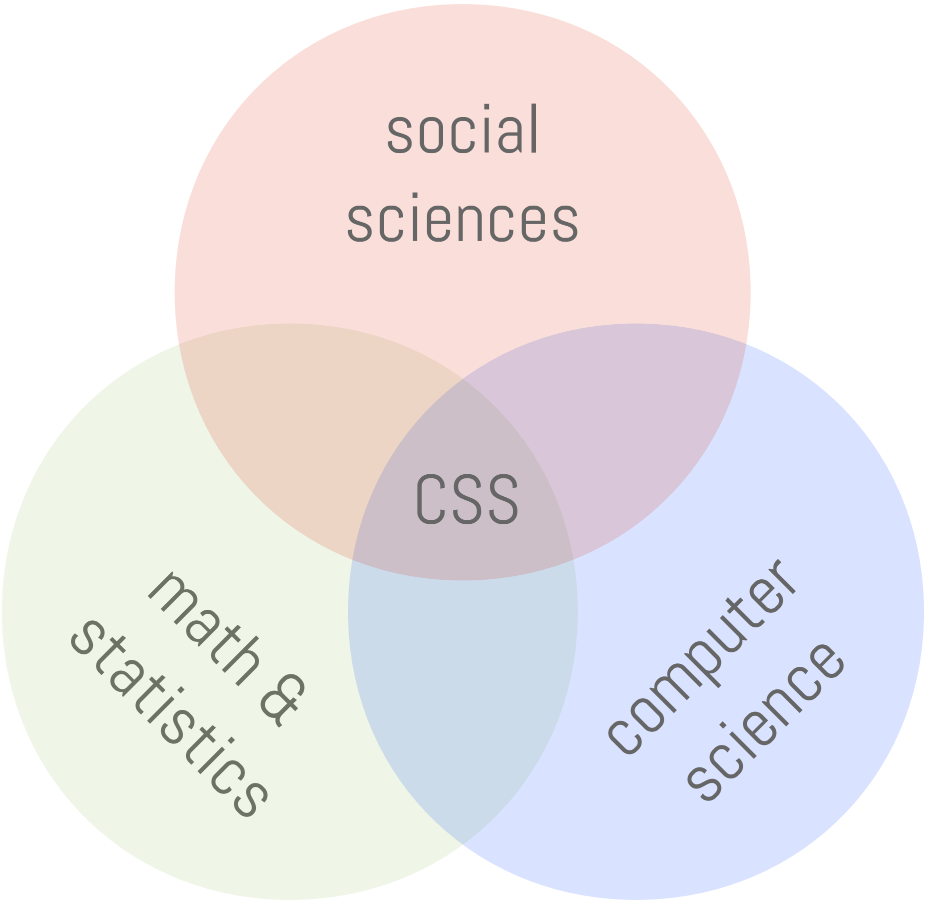
> Big data is the amassing of huge amounts of statistical information on **social and economic trends and human behavior**.
>
> — Michelle Chen, 2014

Granular Data

> “The issue is not just size – we’ve always had big data sets; the issue is **granularity**.
>
> — Michael Jordan, 2014



Computational Social Science



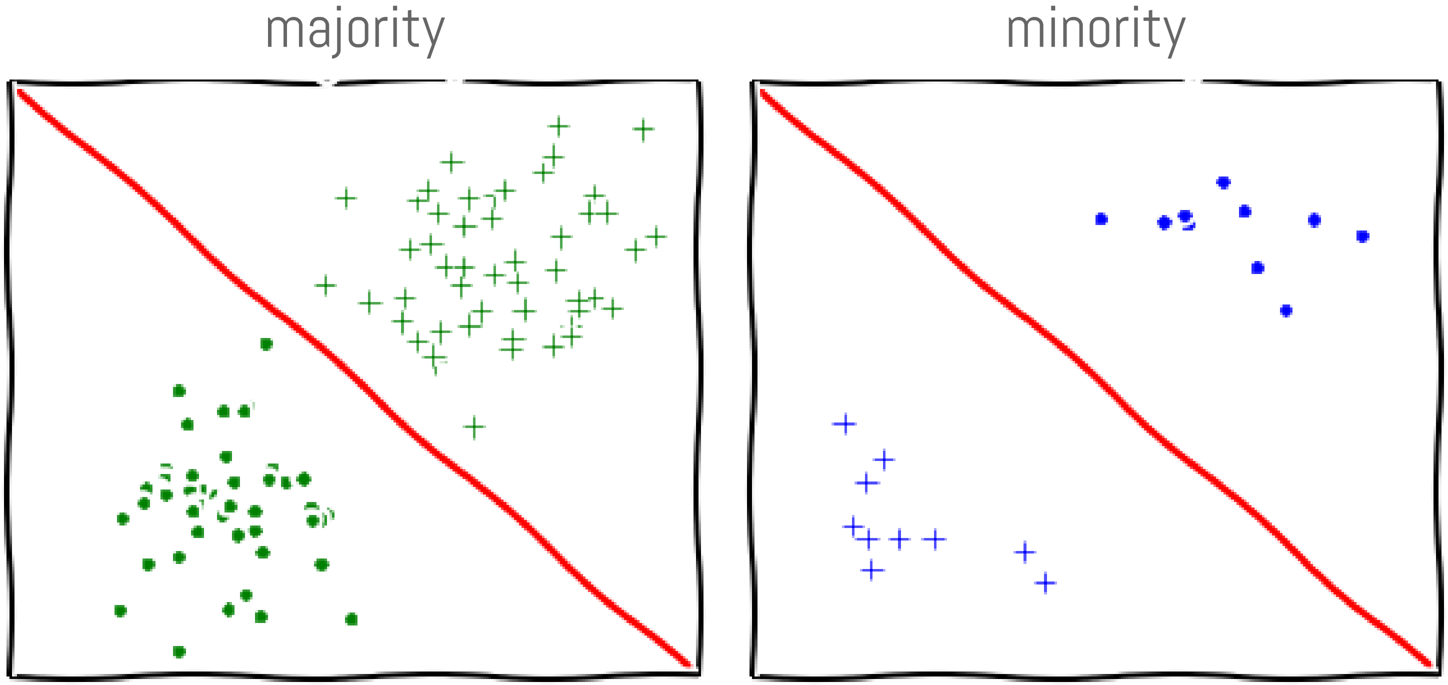
social sciences

CSS

math & statistics

computer science

On Average... Good?



DATA

DATA

DATA information data
DaTa DatA DAta daTA
INFORMATION dAtA

data → data → data →
data → data → data → ...

Data

DATA

QUESTIONS

MODELS

FINDINGS

DATA

QUESTIONS

MODELS

FINDINGS

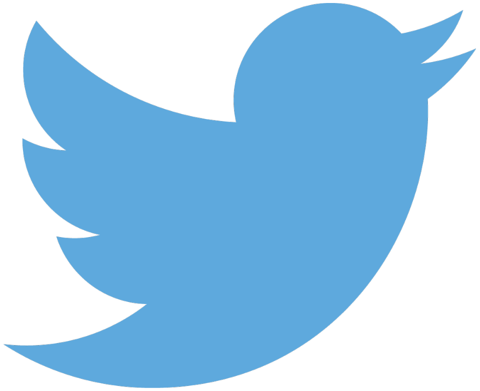
Small Patterns

> “The real, epistemological problem with big data is small patterns. [... But] small patterns may be significant only if properly aggregated. [...]
> So what we need is a better understanding of which data are worth preserving.
>
> — Luciano Floridi, 2013

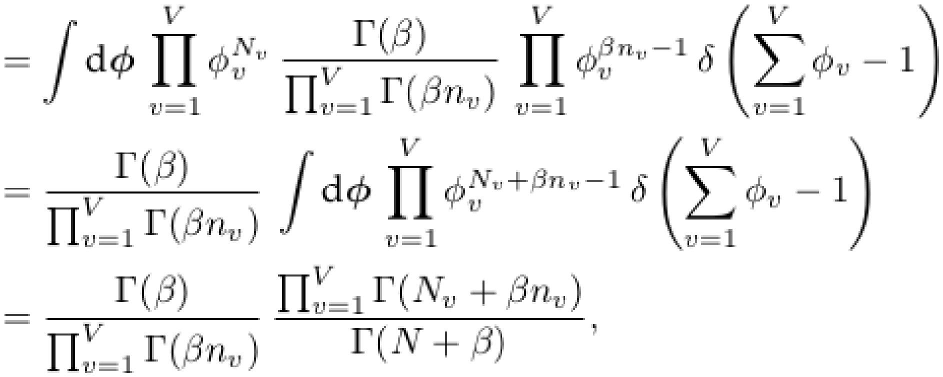
Big Questions

> “And this is a matter of grasping which questions are or will be interesting. [... U]ltimately, the game will be won by those who 'know how to ask and answer relevant questions.'
>
> — Luciano Floridi, 2013

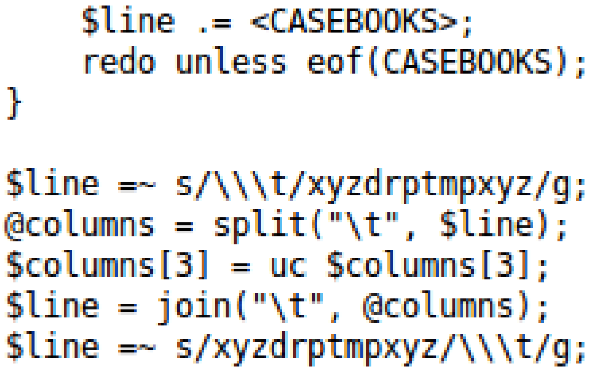
Data-first?



Method-first?



$$
\begin{aligned}
&= \int \mathrm{d}\boldsymbol{\phi} \prod_{v=1}^{V} \phi_v^{N_v} \frac{\Gamma(\beta)}{\prod_{v=1}^{V} \Gamma(\beta n_v)} \prod_{v=1}^{V} \phi_v^{\beta n_v - 1} \, \delta\left(\sum_{v=1}^{V} \phi_v - 1\right) \\
&= \frac{\Gamma(\beta)}{\prod_{v=1}^{V} \Gamma(\beta n_v)} \int \mathrm{d}\boldsymbol{\phi} \prod_{v=1}^{V} \phi_v^{N_v + \beta n_v - 1} \, \delta\left(\sum_{v=1}^{V} \phi_v - 1\right) \\
&= \frac{\Gamma(\beta)}{\prod_{v=1}^{V} \Gamma(\beta n_v)} \frac{\prod_{v=1}^{V} \Gamma(N_v + \beta n_v)}{\Gamma(N + \beta)},
\end{aligned}
$$



```
        $line .= <CASEBOOKS>;
        redo unless eof(CASEBOOKS);
    }

    $line =~ s/\\\t/xyzdrptmpxyz/g;
    @columns = split("\t", $line);
    $columns[3] = uc $columns[3];
    $line = join("\t", @columns);
    $line =~ s/xyzdrptmpxyz/\\\t/g;
```

~~Data-first~~

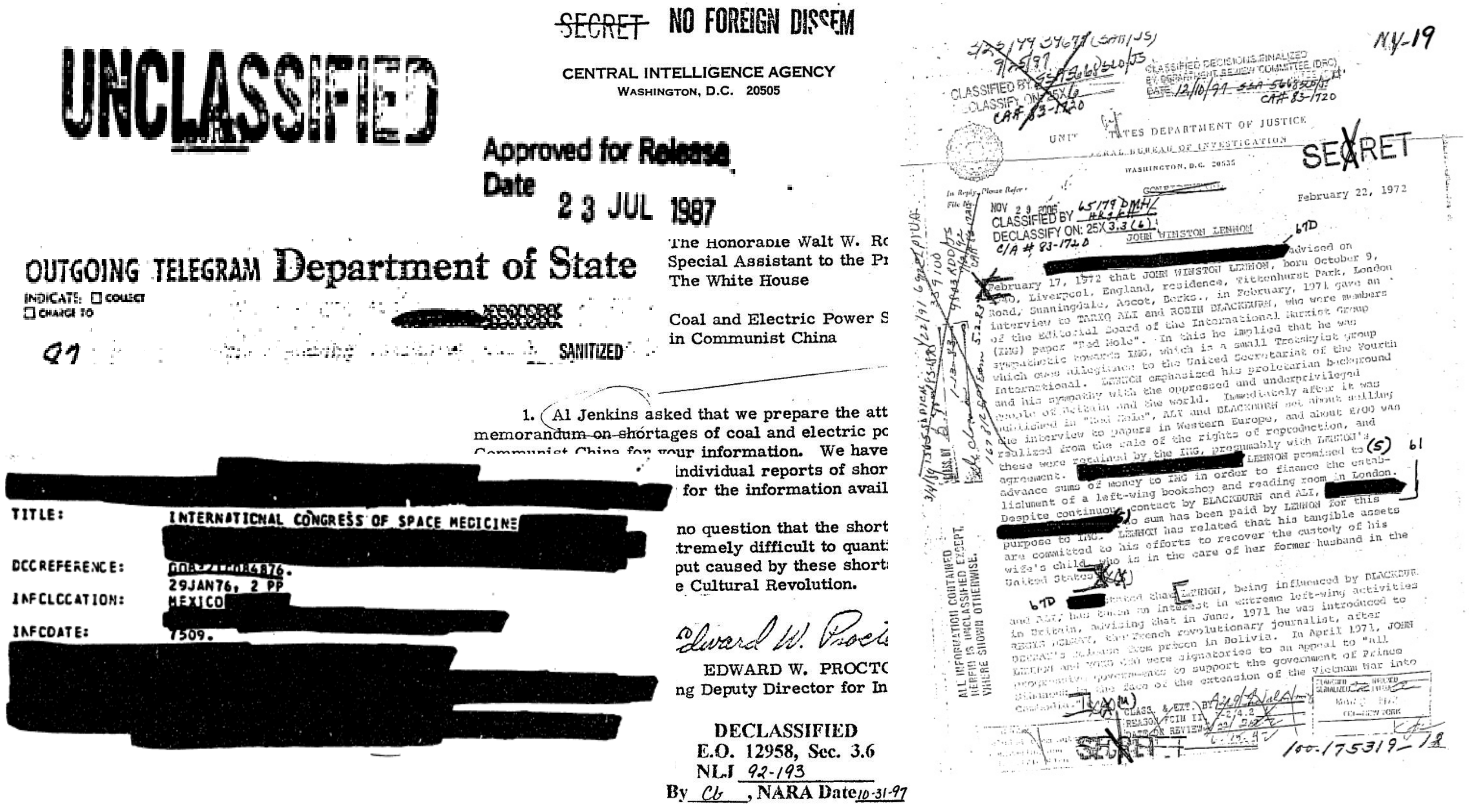
~~Method-first~~

QUESTION-FIRST

Gender in Local Government

Is there gender homophily? Do women occupy disadvantaged positions in the network? Are women less central in the dominant coalition?

“Push” Transparency



UNCLASSIFIED

OUTGOING TELEGRAM Department of State

INDICATE: □ COLLECT
□ CHARGE TO

[redacted]

91 SANITIZED

[redacted]

TITLE: INTERNATICNAL CONGRESS OF SPACE MEDICINE [redacted]

DCCREFERENCE: [illegible]4876.
29JAN76, 2 PP

INFCLCCATION: MEXICO [redacted]

INFCDATE: 7509.

[redacted]

~~SECRET~~ NO FOREIGN DISSEM

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

Approved for Release
Date 23 JUL 1987

The Honorable Walt W. Ro
Special Assistant to the Pr
The White House

Coal and Electric Power S
in Communist China

1. Al Jenkins asked that we prepare the att memorandum ~~on~~ shortages of coal and electric po Communist China for your information. We have individual reports of shor for the information avail

no question that the short tremely difficult to quant put caused by these short e Cultural Revolution.

Edward W. Proct
EDWARD W. PROCT
ng Deputy Director for In

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97

~~4/25/99 54678 (SAM/JS)~~
~~9/25/97~~
~~CLASSIFIED BY 55156 8510/JS~~
~~DECLASSIFY ON: 25X6~~
~~CA# 83-1720~~

CLASSIFIED DECISIONS FINALIZED BY DEPARTMENT REVIEW COMMITTEE (DRC)
DATE 12/10/97 ~~SSA 56685 JS~~
CA# 83-1720

NY-19

UNITED STATES DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION
WASHINGTON, D.C. 20535

~~SECRET~~

In Reply, Please Refer to
File No.

~~CONFIDENTIAL~~

February 22, 1972

NOV 29 2006
CLASSIFIED BY 65179 DMH/
DECLASSIFY ON: 25X 3.3 (6)
C/A # 83-1720

JOHN WINSTON LENNON

b7D

[redacted] advised on February 17, 1972 that JOHN WINSTON LENNON, born October 9, [illegible], Liverpool, England, residence, Tittenhurst Park, London Road, Sunningdale, Ascot, Berks., in February, 1971 gave an interview to TARIQ ALI and ROBIN BLACKBURN, who were members of the Editorial Board of the International Marxist Group (IMG) paper "Red Mole". In this he implied that he was sympathetic towards IMG, which is a small Trotskyist group which owes allegiance to the United Secretariat of the Fourth International. LENNON emphasized his proletarian background and his sympathy with the oppressed and underprivileged people of Britain and the world. Immediately after it was published in "Red Mole", ALI and BLACKBURN set about selling the interview to papers in Western Europe, and about £700 was realized from the sale of the rights of reproduction, and these were retained by the IMG, presumably with LENNON's agreement. [redacted] LENNON promised to advance sums of money to IMG in order to finance the establishment of a left-wing bookshop and reading room in London. Despite continuous contact by BLACKBURN and ALI, [redacted] (S) b1 no sum has been paid by LENNON for this purpose to IMG. LENNON has related that his tangible assets are committed to his efforts to recover the custody of his wife's child who is in the care of her former husband in the United States. (U)

b7D [redacted] stated that LENNON, being influenced by BLACKBURN and ALI, has shown an interest in extreme left-wing activities in Britain, advising that in June, 1971 he was introduced to REGIS DEBRAY, the French revolutionary journalist, after DEBRAY's release from prison in Bolivia. In April 1971, JOHN LENNON and YOKO ONO were signatories to an appeal to "all progressive governments to support the government of Prince Sihanouk in the face of the extension of the Vietnam War into Cambodia." (U)

CLASS. & EXT. BY
REASON FCIM II 1-2.4.2
DATE OF REVIEW 6-25-9?

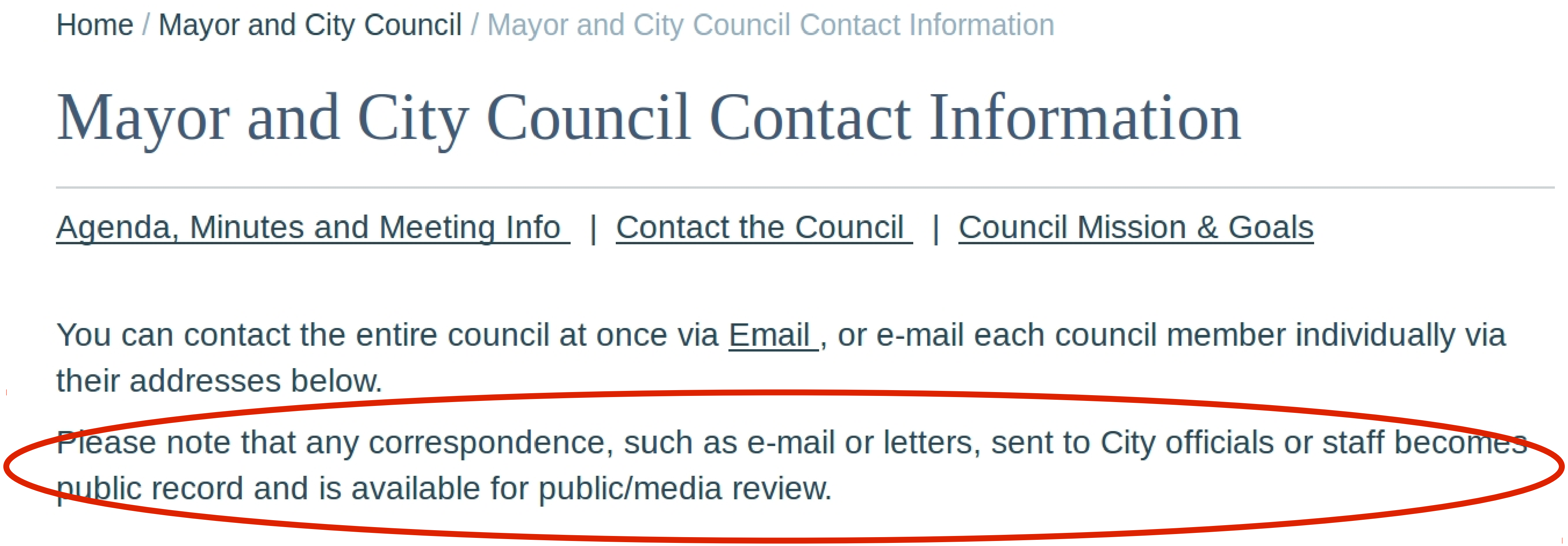
SEARCHED SERIALIZED INDEXED FILED
FBI—NEW YORK

~~SECRET~~

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED EXCEPT WHERE SHOWN OTHERWISE.

100-175319-18

“Pull” Transparency



Home / Mayor and City Council / Mayor and City Council Contact Information

Mayor and City Council Contact Information

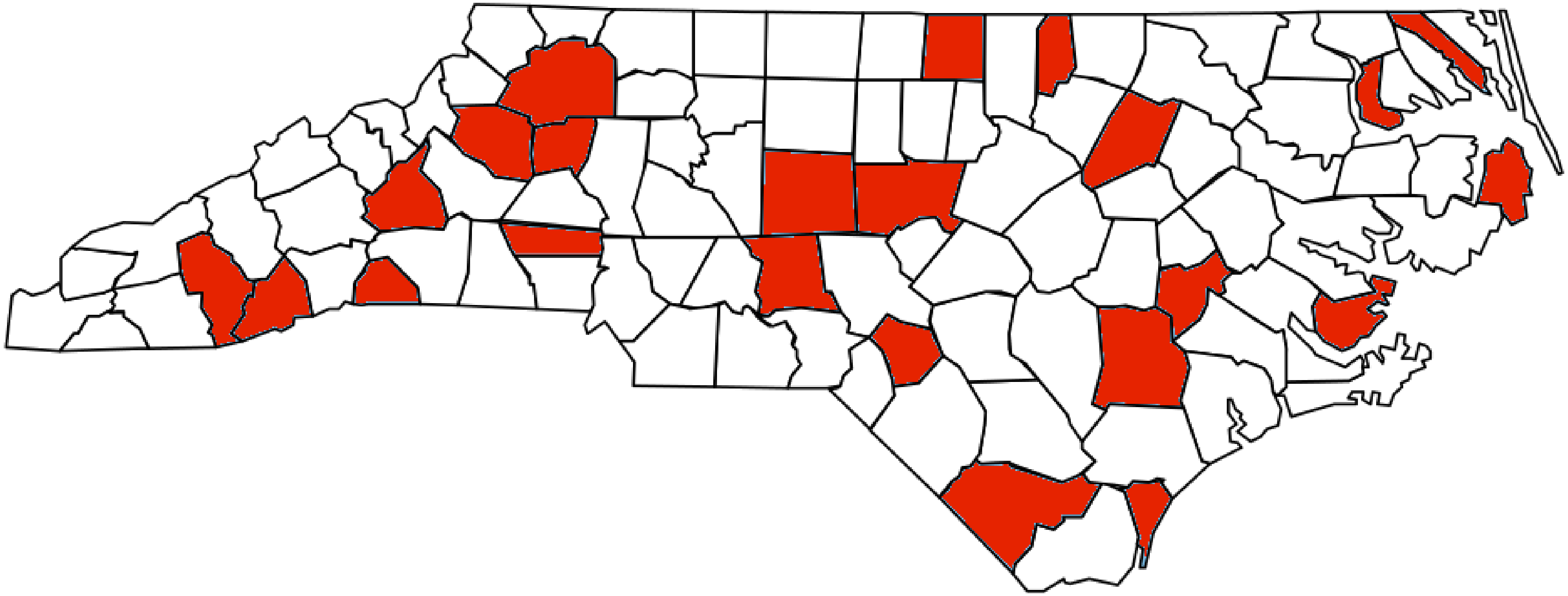
Agenda, Minutes and Meeting Info | Contact the Council | Council Mission & Goals

You can contact the entire council at once via Email , or e-mail each council member individually via their addresses below.

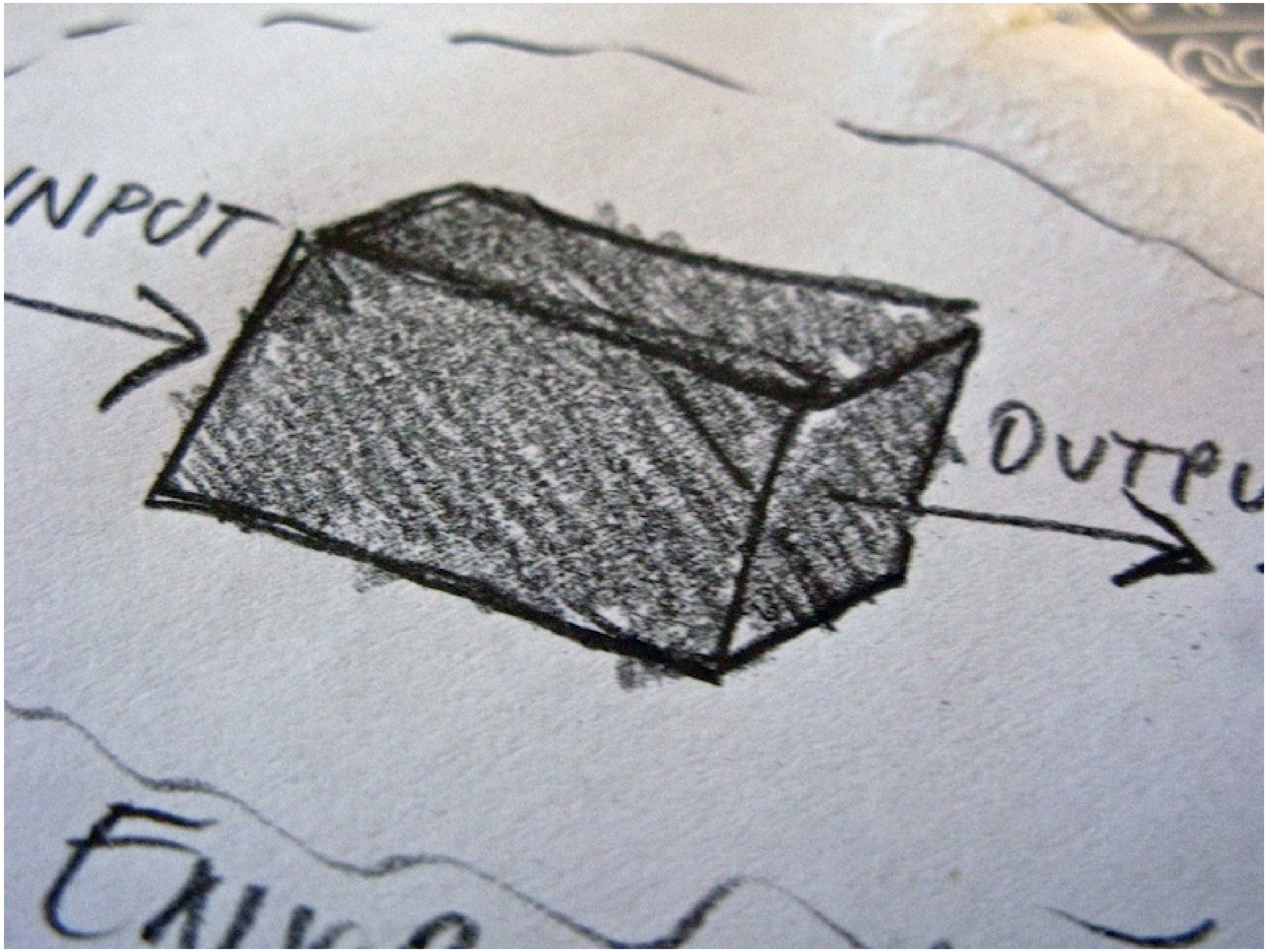
Please note that any correspondence, such as e-mail or letters, sent to City officials or staff becomes public record and is available for public/media review.

public records:
new data availability

23 Counties, 500,000+ Emails



Algorithmic Accountability



DATA

QUESTIONS

MODELS

FINDINGS

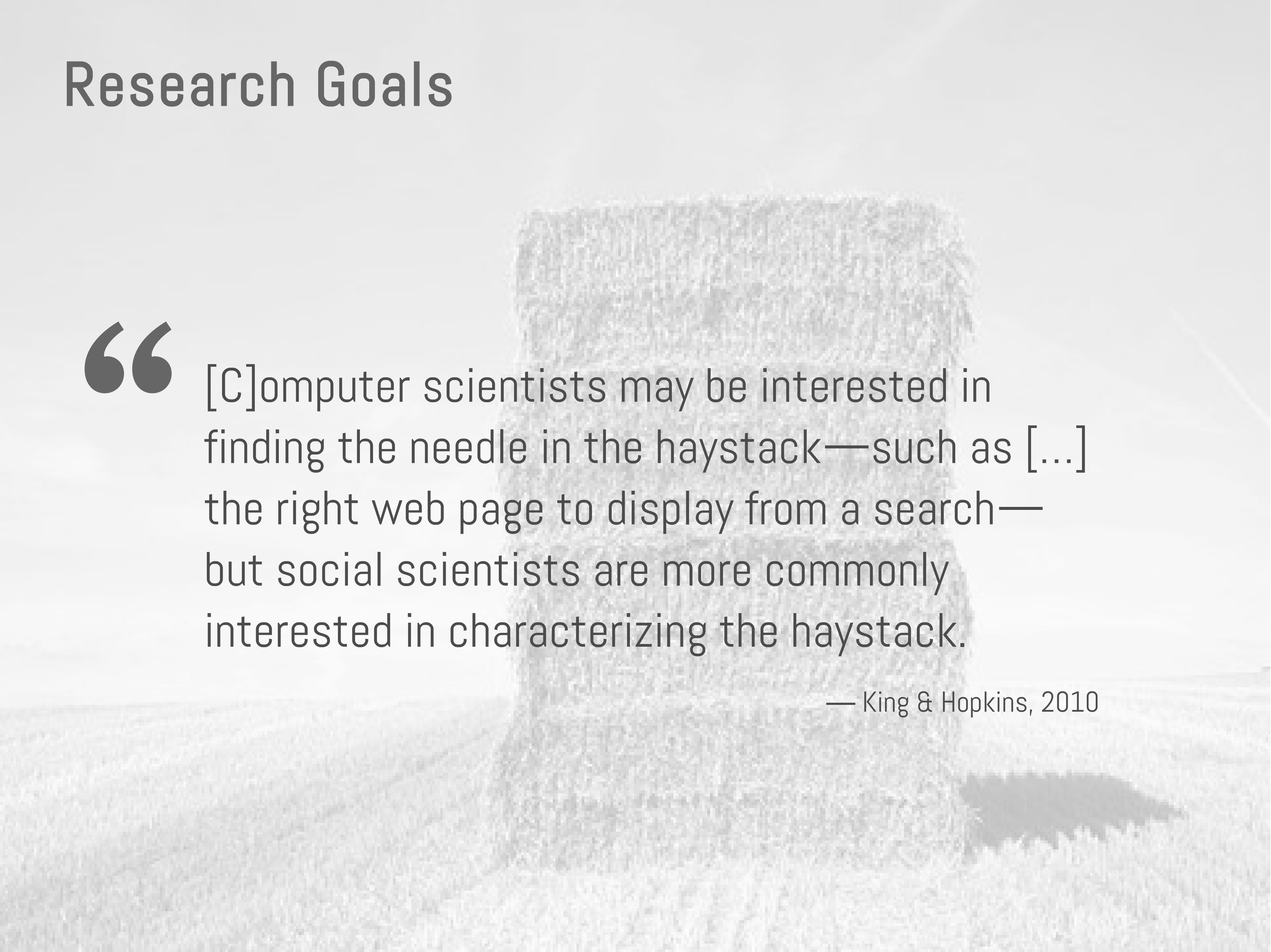
DATA

QUESTIONS

MODELS

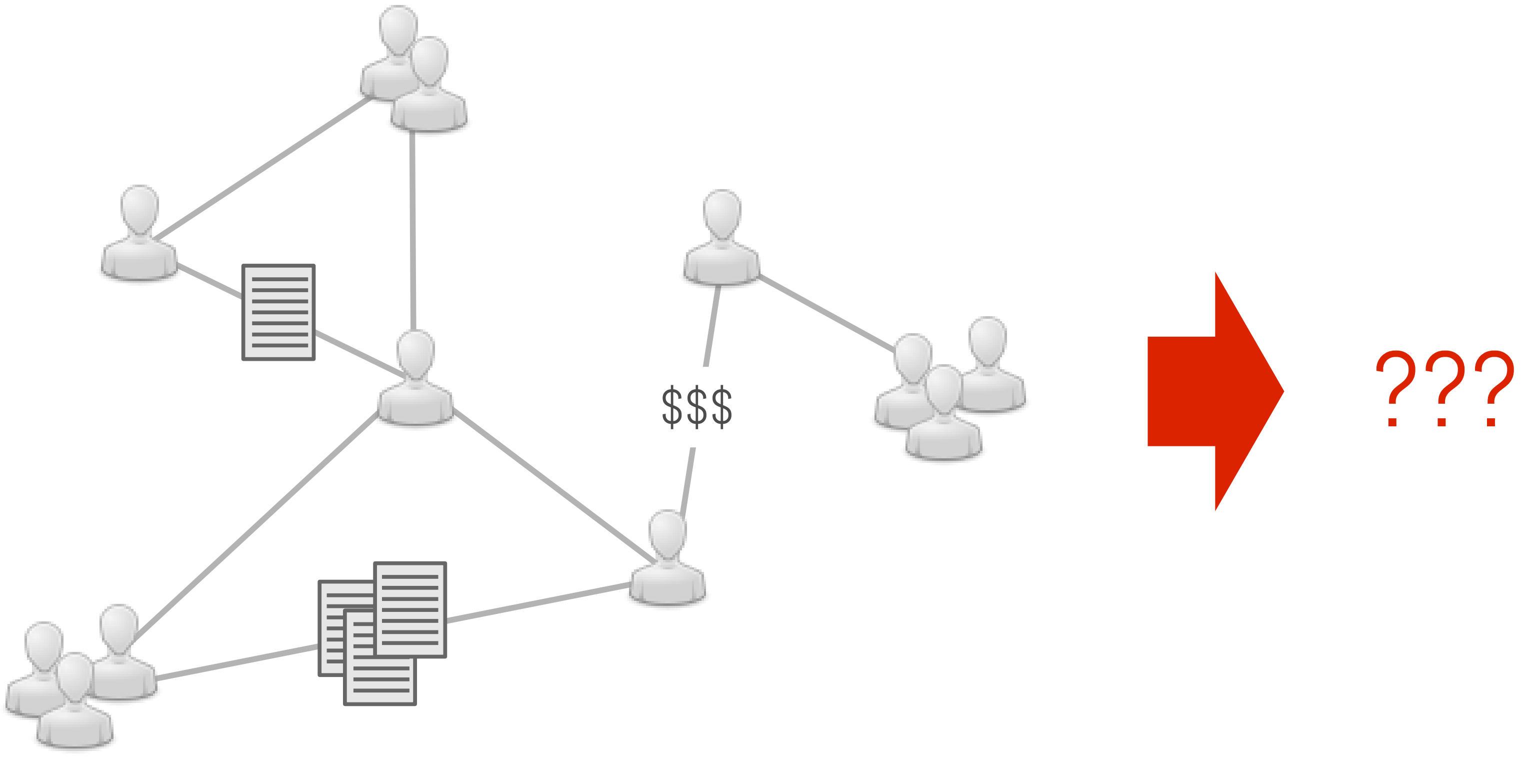
FINDINGS

Research Goals



> “ [C]omputer scientists may be interested in finding the needle in the haystack—such as [...] the right web page to display from a search—but social scientists are more commonly interested in characterizing the haystack.
>
> — King & Hopkins, 2010

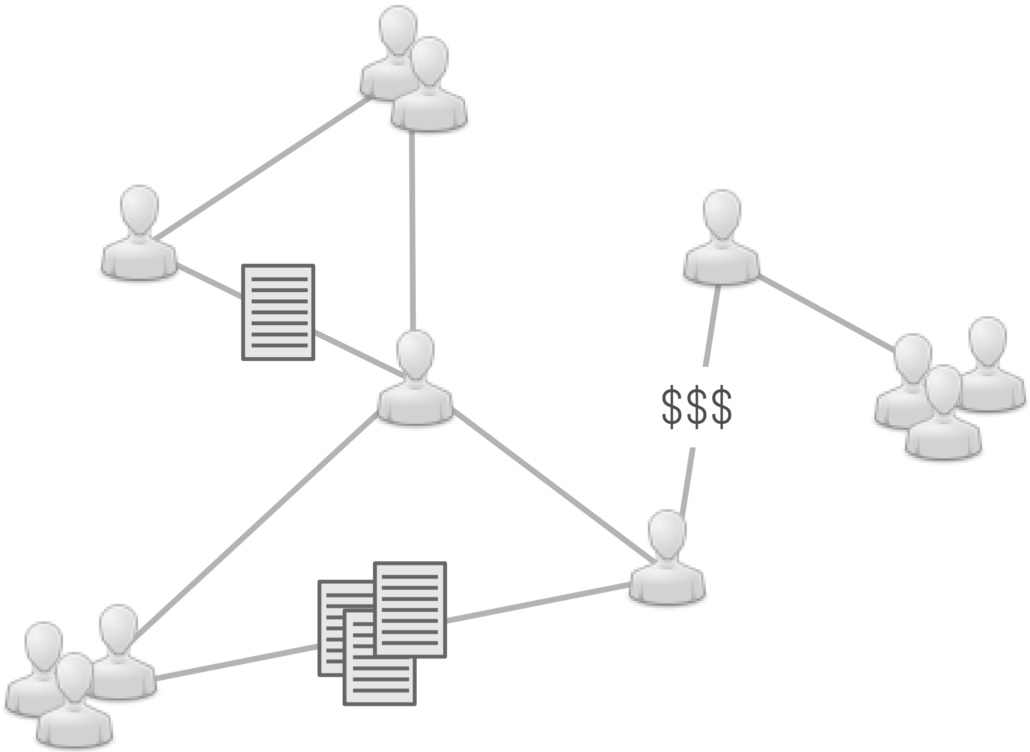
Prediction



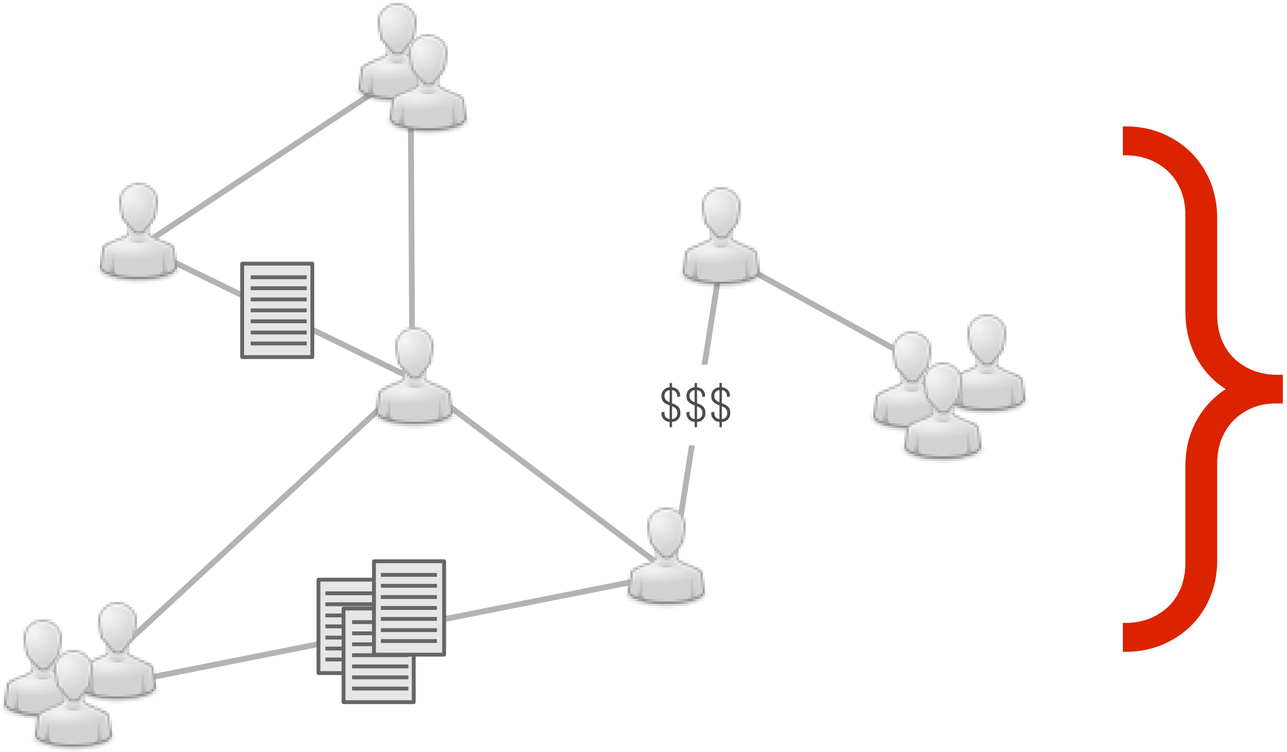
???

Explanation

???



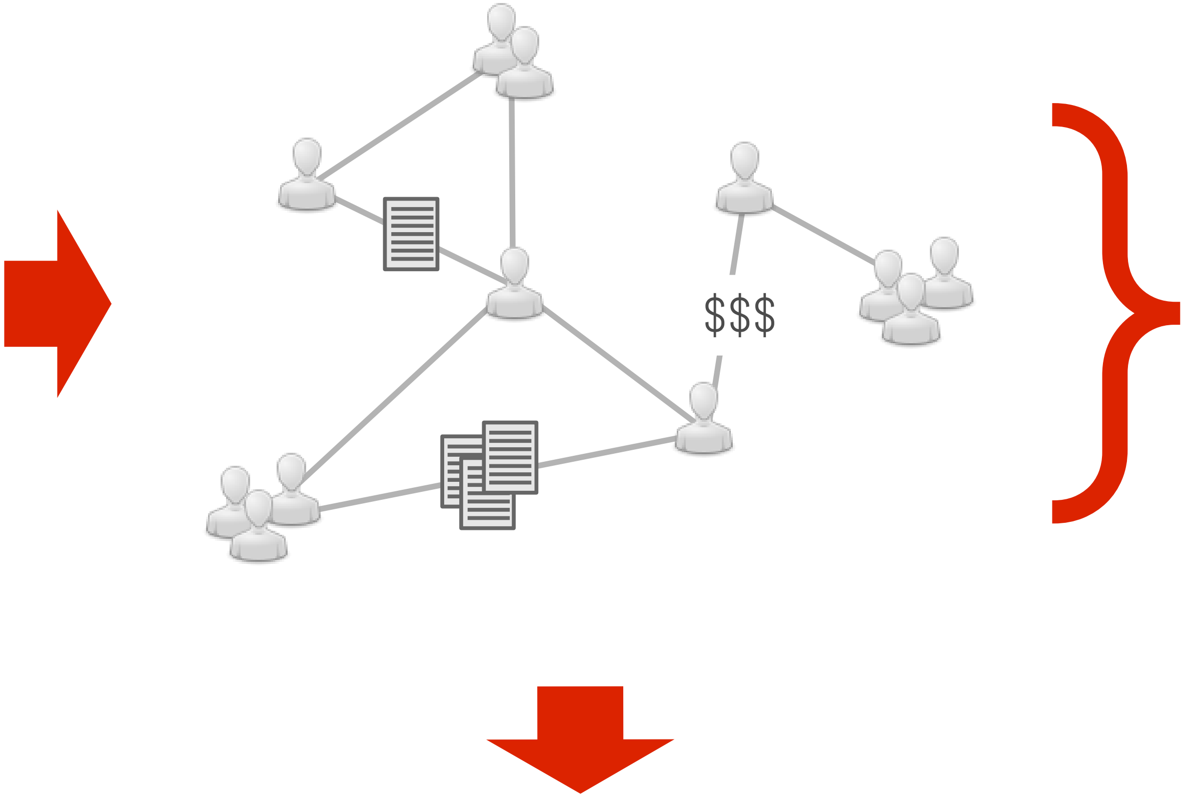
Exploration



???

Bias & Unfairness

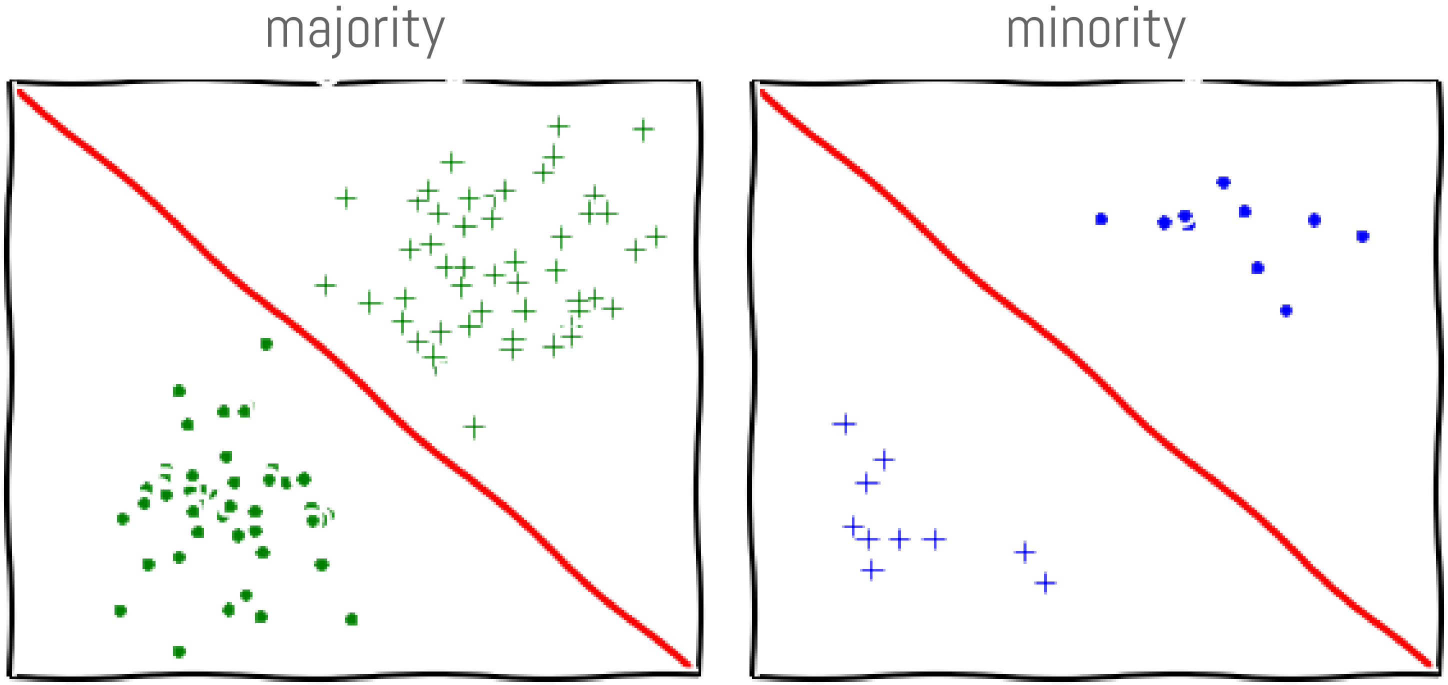
???



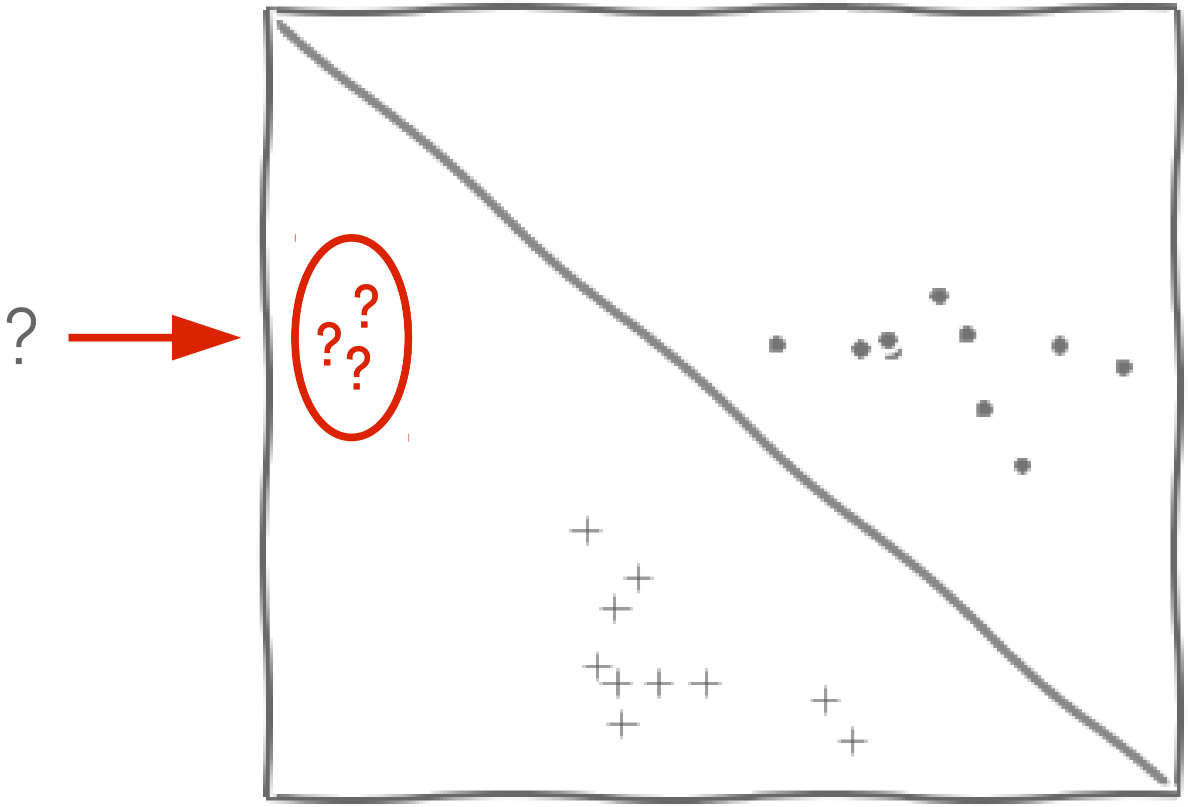
???

???

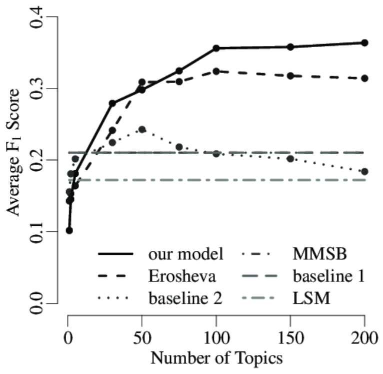
Error Analysis

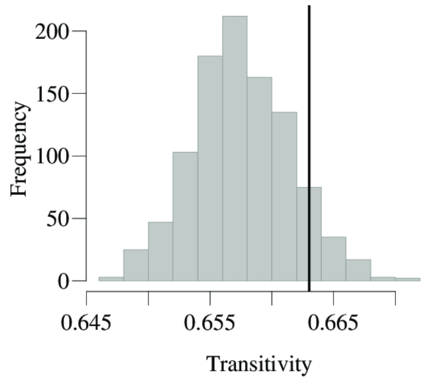


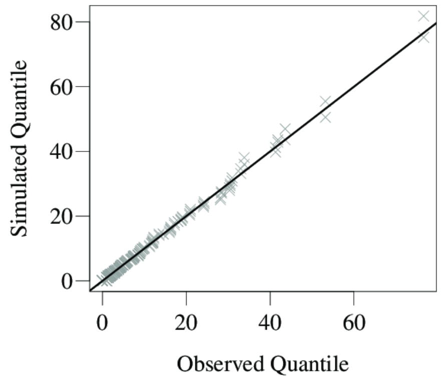
Uncertainty

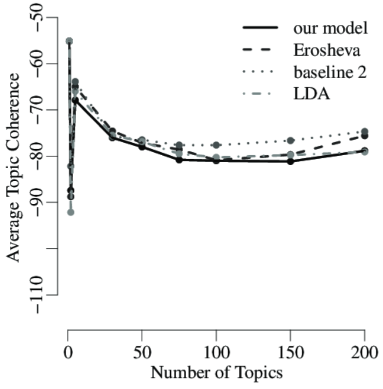


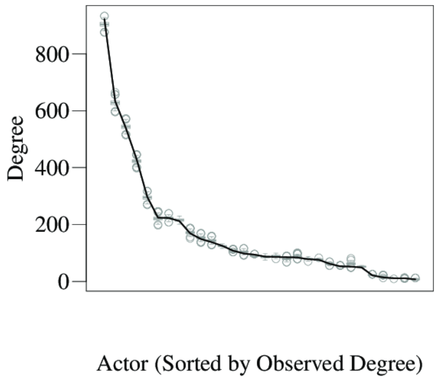
Exploratory Models

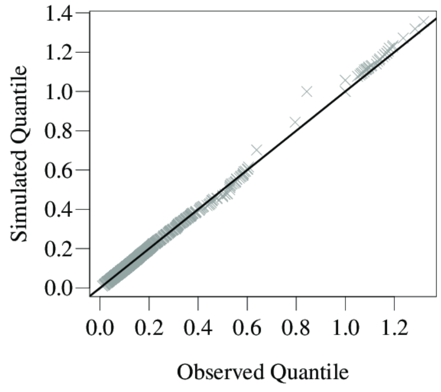












DATA

QUESTIONS

MODELS

FINDINGS

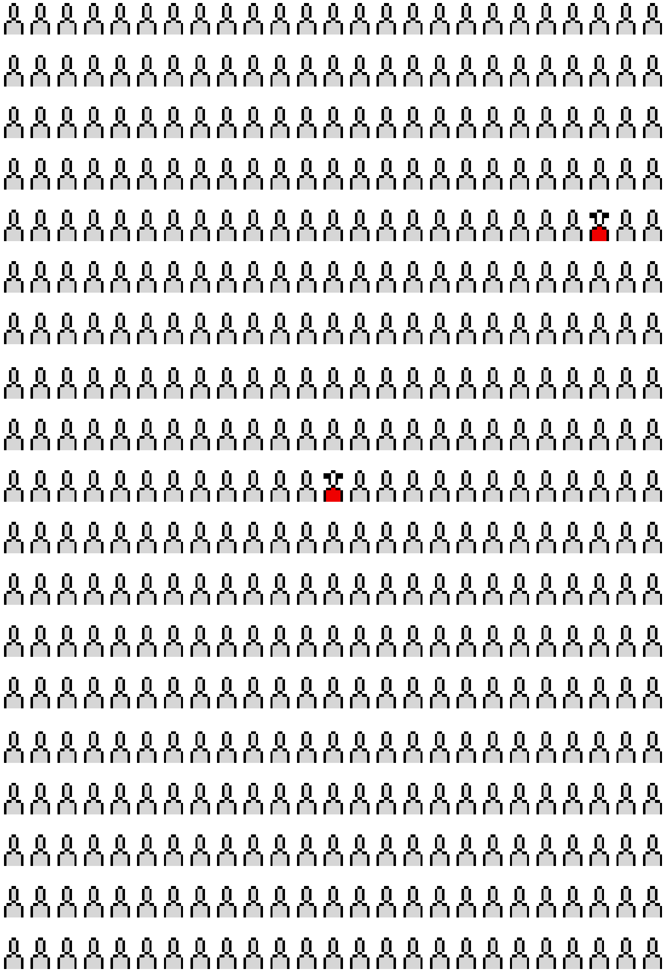
DATA

QUESTIONS

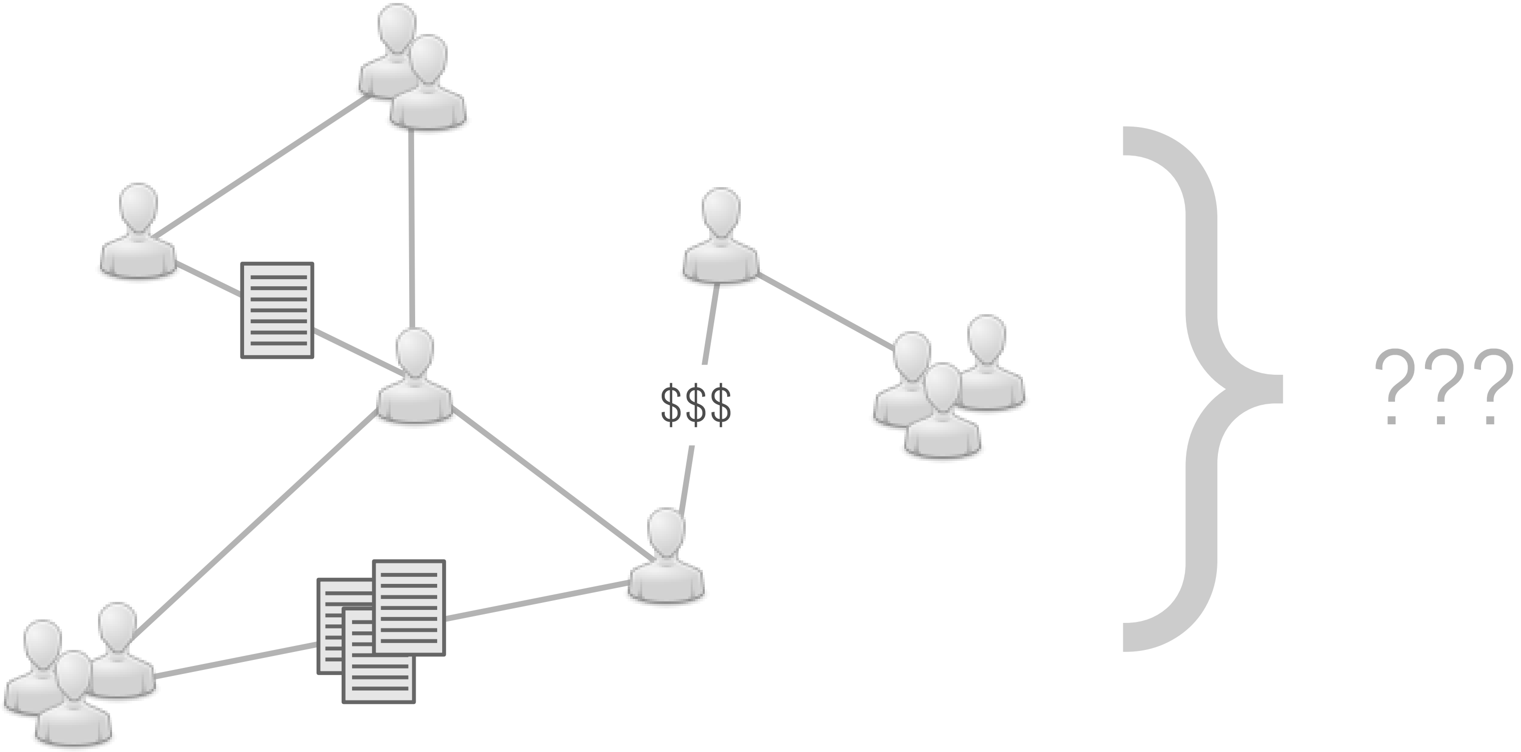
MODELS

FINDINGS

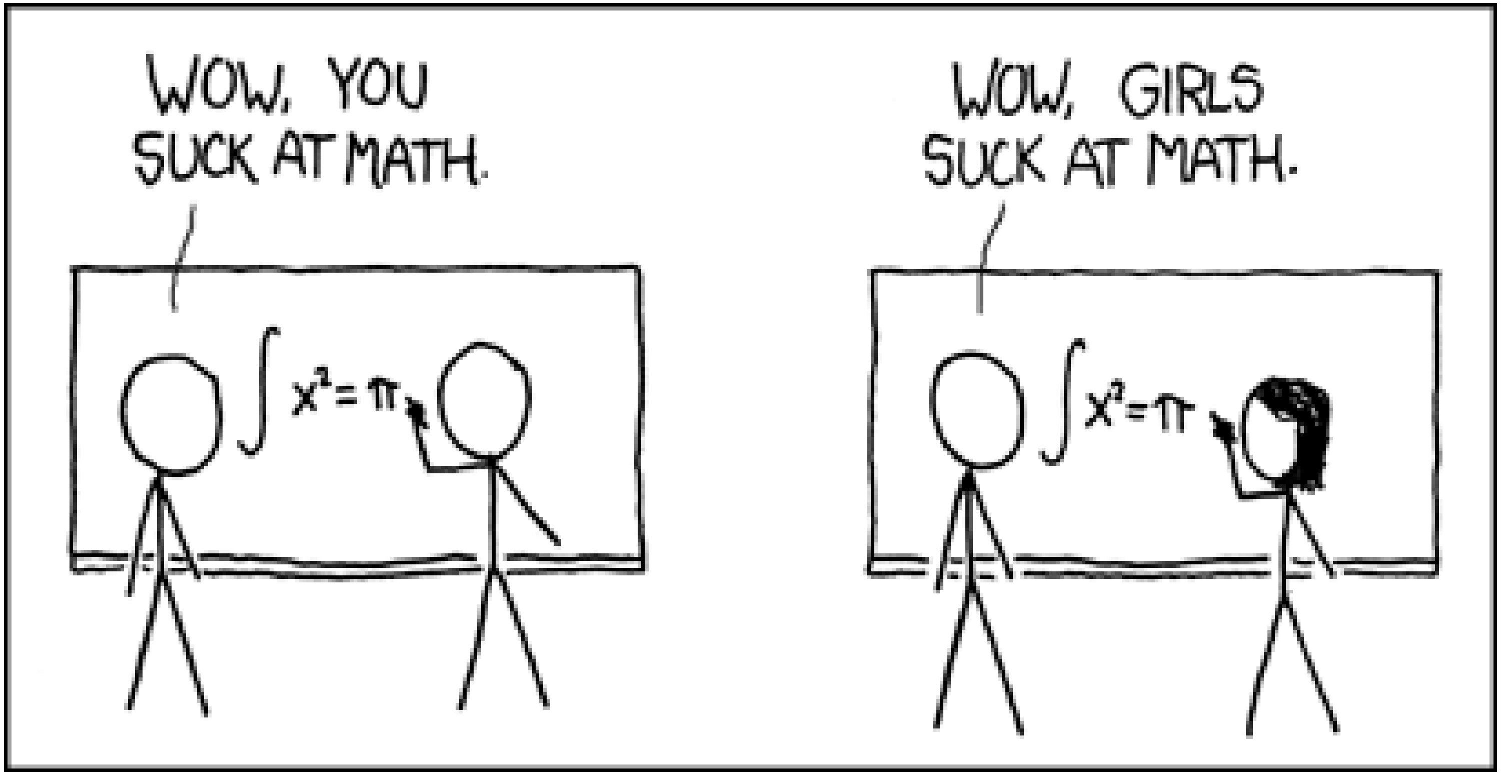
Intuition



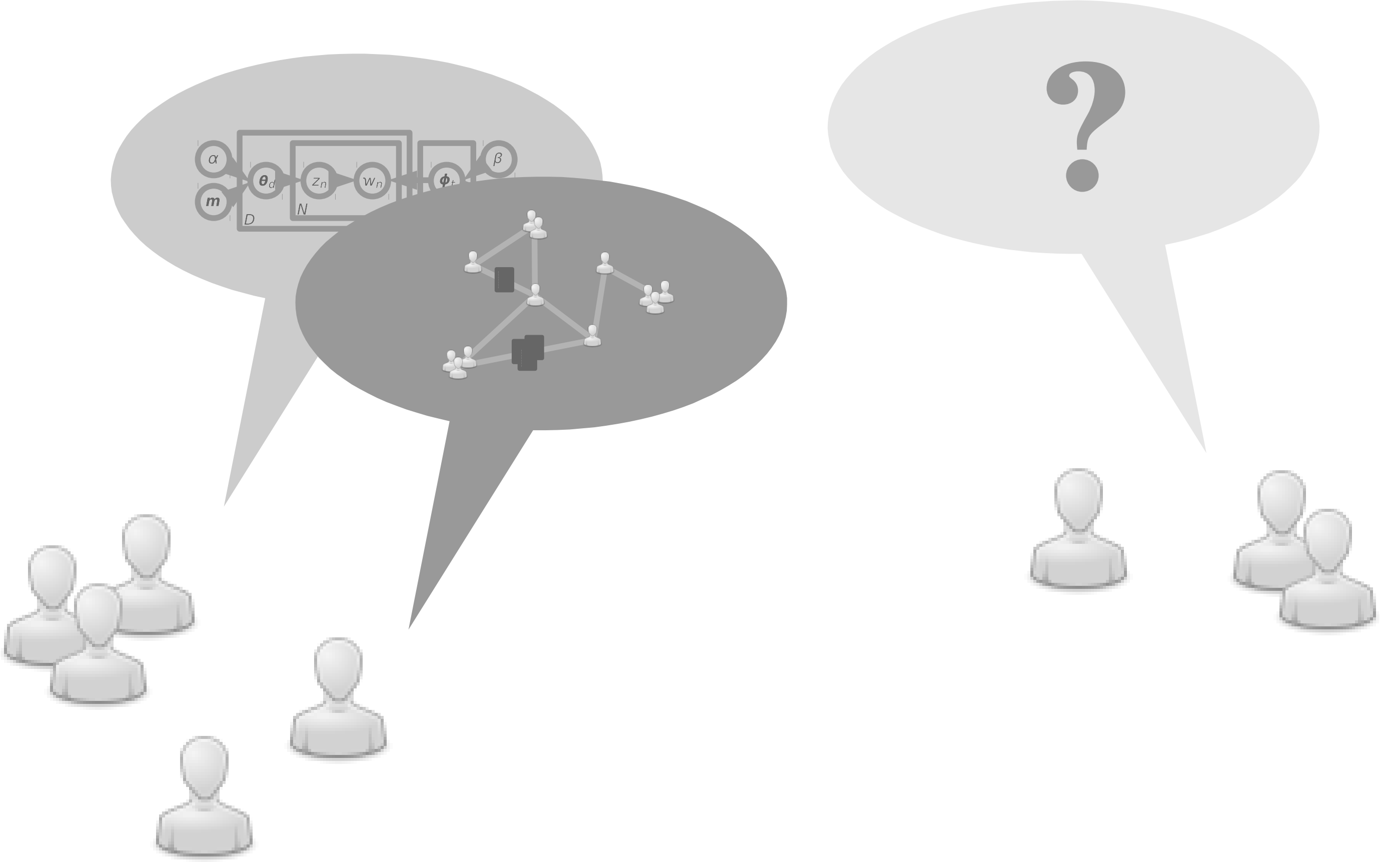
Confirmation Bias



e.g., Stereotype Threat



Scientific Communication



DATA

QUESTIONS

MODELS

FINDINGS

Thanks!

@hannawallach | http://dirichlet.net