From Classroom to Clinic

Data Science for Social Good Fellowships and the Lessons Data Science Educators Can Learn from the Medical Profession

Matt Gee

National Academy of Sciences December 2018

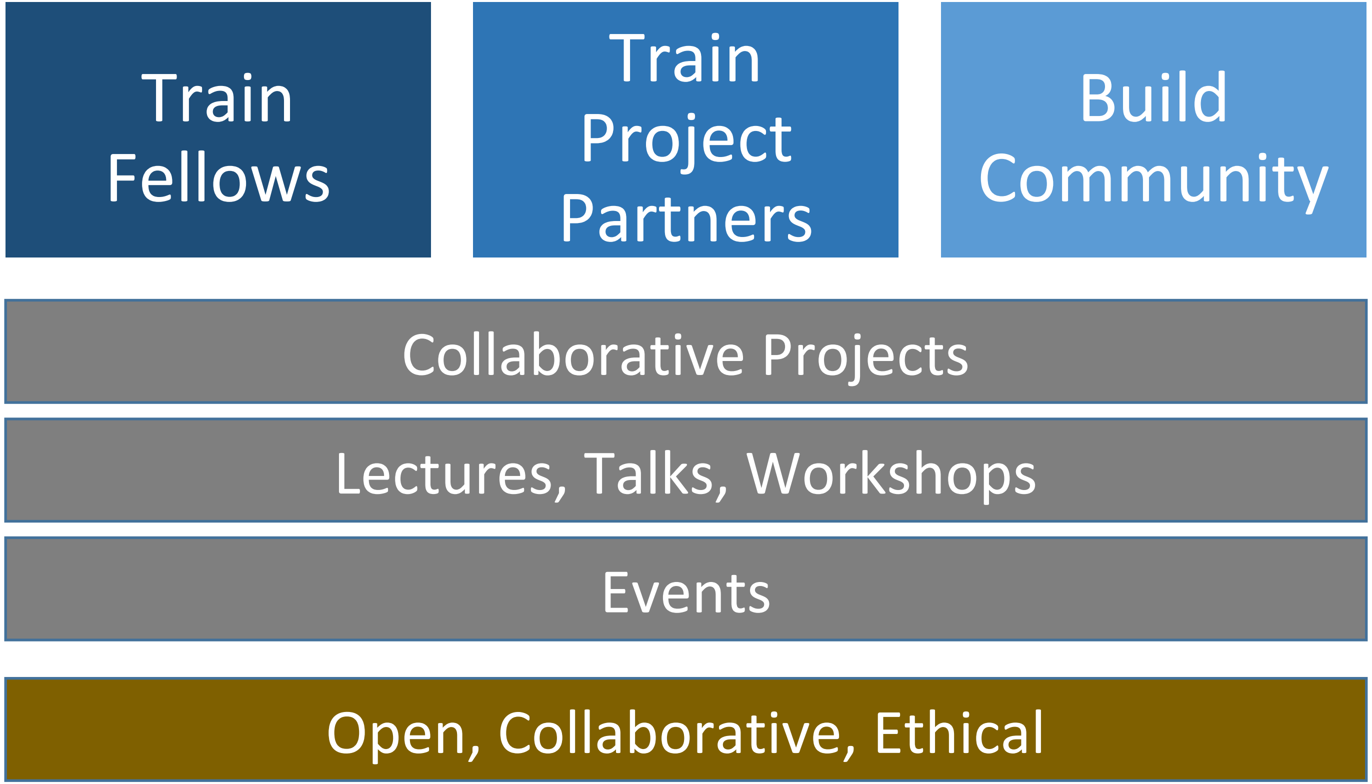
Data Science for….

Data Science For Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

Goals of DSSG



Train Fellows

Train Project Partners

Build Community

Collaborative Projects

Lectures, Talks, Workshops

Events

Open, Collaborative, Ethical

Step 1: Find Fellows

600+ applications in 3 weeks

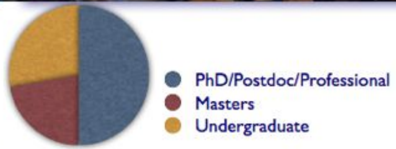
120 interviews

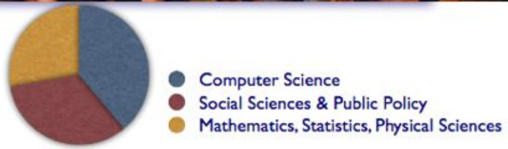
36 fellows selected

Stanford University
University of California-Irvine
Arizona State University
University of Texas-Austin
University of Wisconsin-Madison
University of Texas-Austin
Instituto Technológico Autónomo de México

University of Chicago
University of Illinois-Chicago
University of Michigan
University of Alabama
Georgia Institute of Technology
Carnegie Mellon University
Cornell University
Israel Institute of Technology

University of Maryland
City University of New York
Columbia University
Yale University
Harvard University
Massachusetts Institute of Technology
University of Cambridge



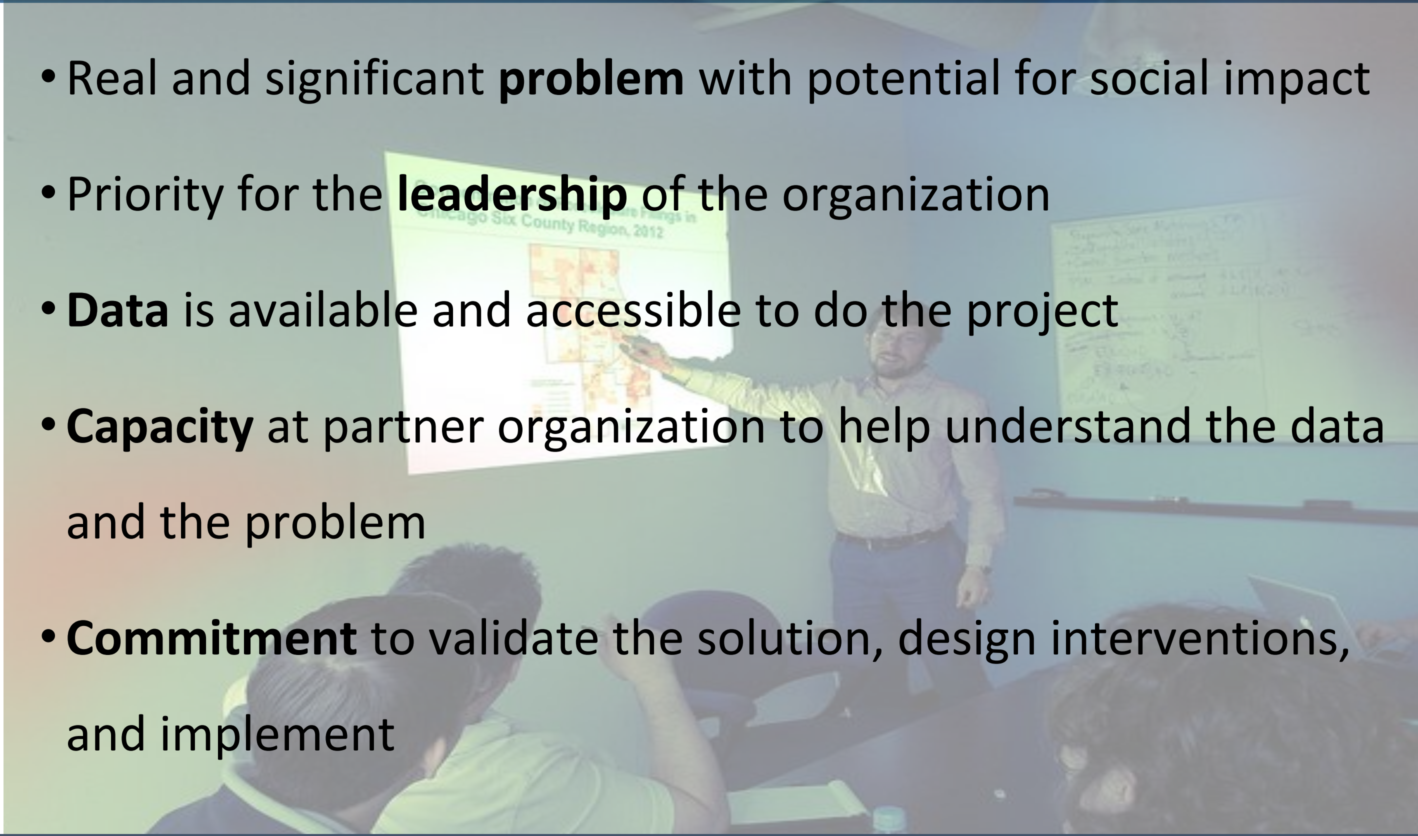


Step 2: Pick partners

30+ potential partners

60 discovery calls and partner interviews

12 partners



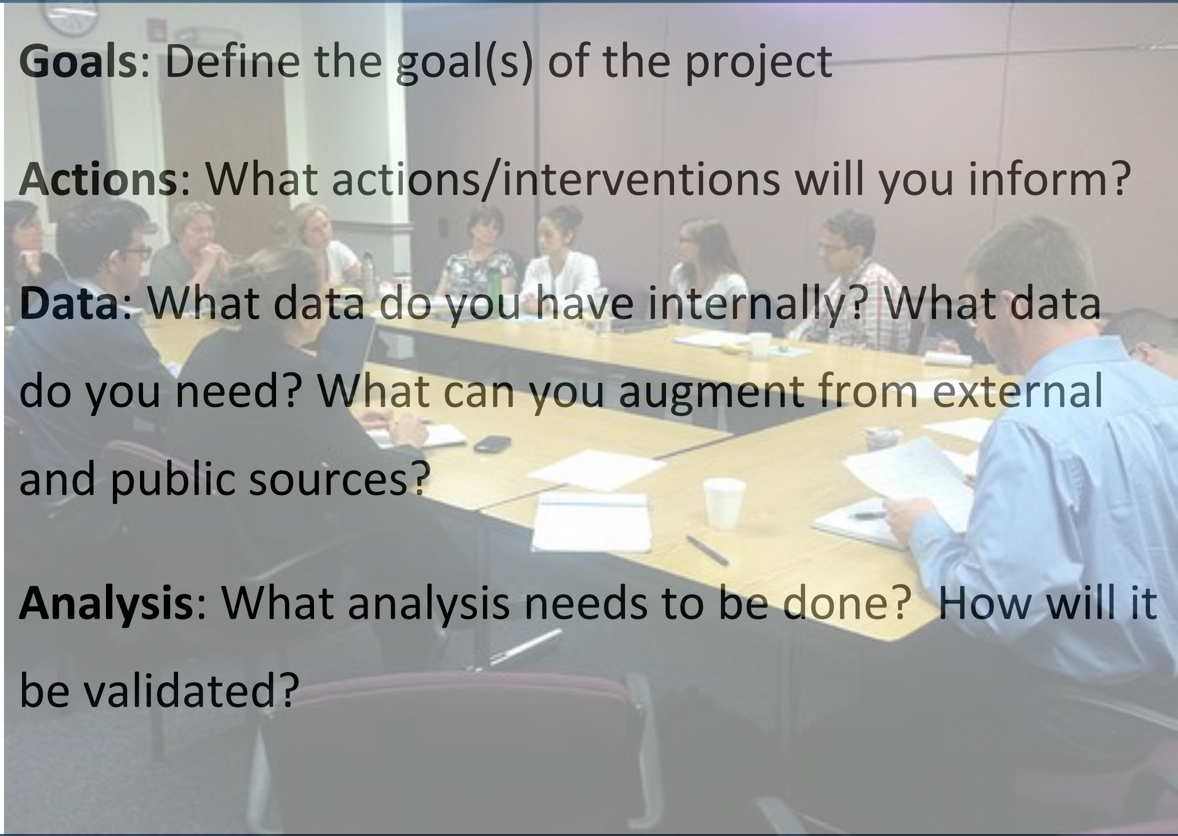
- Real and significant **problem** with potential for social impact
- Priority for the **leadership** of the organization
- **Data** is available and accessible to do the project
- **Capacity** at partner organization to help understand the data and the problem
- **Commitment** to validate the solution, design interventions, and implement

Step 3: Scope projects

250 hours of scoping calls, meetings, and negotiations

14 finalized projects

9 legal agreements signed by the first day of the fellowship



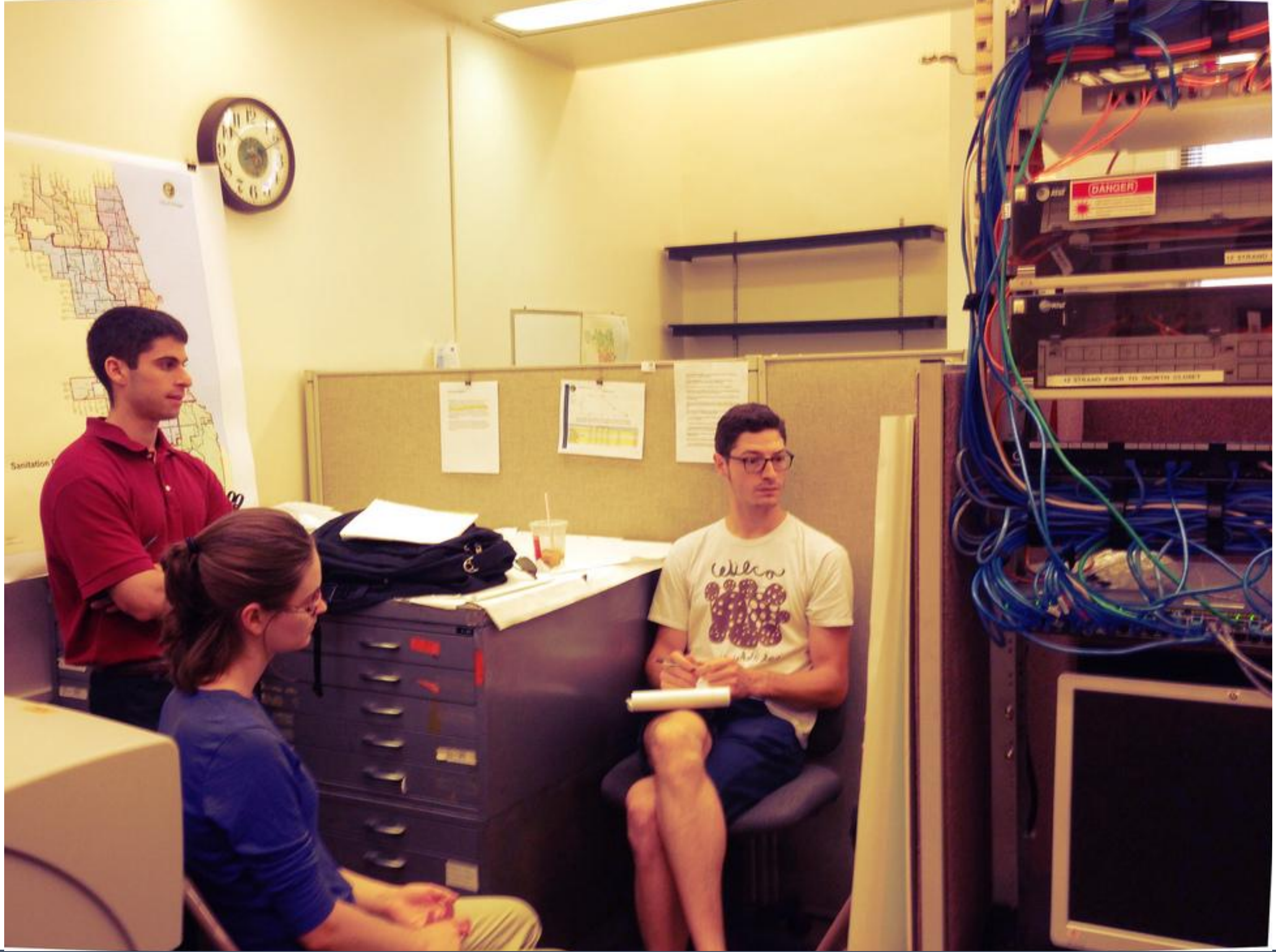
Goals: Define the goal(s) of the project

Actions: What actions/interventions will you inform?

Data: What data do you have internally? What data do you need? What can you augment from external and public sources?

Analysis: What analysis needs to be done? How will it be validated?

The Eric & Wendy Schmidt
Data Science for Social Good
Summer Fellowship 2013
Computation Institute
Urban Center for Computation and Data
The University of Chicago
Argonne National Laboratory
The Harris School
Public Policy | The University of Chicago
FIRE EXTINGUISHER


DANGER
Sanitation

ASUS
ASUS
Google
SUMMER OF
Social Good

322

Nurse-Family Partnership

Helping First-Time Parents Succeed®

“We are used to using data to justify funding decisions. Now we can use data to improve what we do.”

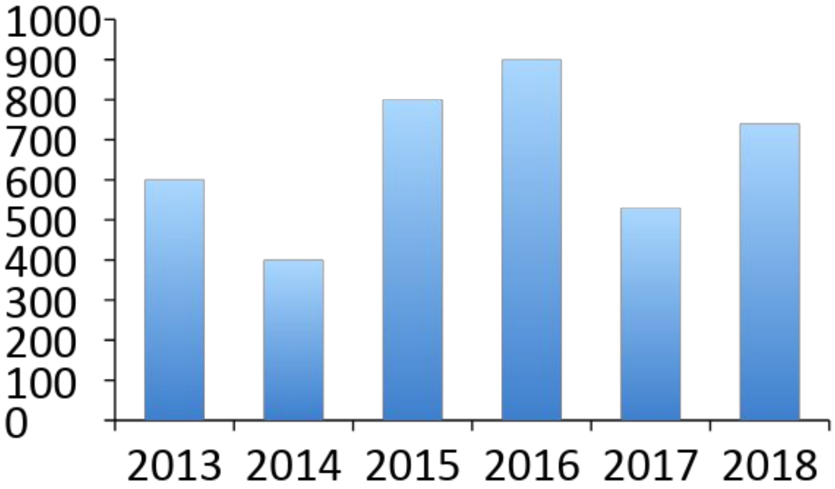
Bill Thorland
Nurse Family Partnership

DSSG Fellowship over the years

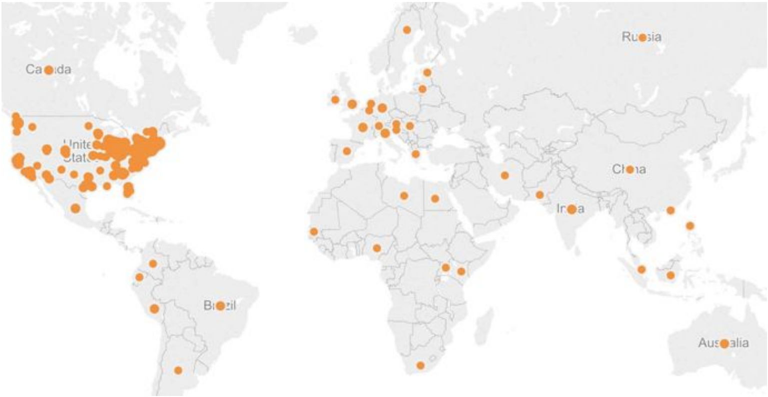
2013	2014	2015	2016	2017	2018
36 Fellows	48 Fellows	42 Fellows	42 Fellows	15 Fellows	38 Fellows
6 Mentors	8 Mentors	5 Mentors 3 PMs	6 Mentors 3 PMs	2 Mentors 2 PMs	5 Mentors 4 PMs
12 Projects	14 Projects	12 Projects	12 Projects	6 Projects	10 Projects
12 Weeks	12 Weeks	14 Weeks	13 Weeks	12 Weeks	12 Weeks
THE UNIVERSITY OF CHICAGO	THE UNIVERSITY OF CHICAGO	THE UNIVERSITY OF CHICAGO	THE UNIVERSITY OF CHICAGO	NOVA NOVA SCHOOL OF BUSINESS & ECONOMICS	NOVA NOVA SCHOOL OF BUSINESS & ECONOMICS THE UNIVERSITY OF CHICAGO

4000+ Applicants over the past 6 Summers

4000+ Applicants



Over 70 countries and 400 universities



Computer Science, Statistics, Math, Applied Math, Sociology, Economics, Public Policy, Political Science, Physics, Chemistry, Biology, Public Health, Psychology, Engineering, BioStatistics, Geography, Business, Neuroscience, …

224

Fellows

70 Projects

IDHS
Illinois Department of Human Services

FEEDING AMERICA

DEPARTMENT OF LABOR UNITED STATES OF AMERICA

VANCOUVER PUBLIC SCHOOLS
VPS

NOBLE

WAKE COUNTY PUBLIC SCHOOL SYSTEM

Perspectives Charter Schools

KIPP New Jersey

KIPP Chicago
COLLEGE PREP PUBLIC SCHOOLS

INFONAVIT
Hogares con valor

cclba

city of CINCINNATI
PERFORMANCE & DATA ANALYTICS

Montgomery County Public Schools
MCPS

CHICAGO PUBLIC SCHOOLS
CPS

POLICE
CHARLOTTE-MECKLENBURG

THE WHITE HOUSE
WASHINGTON

CHICAGO ALLIANCE TO END HOMELESSNESS

INSTITUTE FOR HOUSING STUDIES

ChapinHall
at the University of Chicago

JOSÉ DE MELLO · SAÚDE

City of Memphis
MEMPHIS TENNESSEE

WORLD BANK GROUP

CITY OF CHICAGO INCORPORATED 4TH MARCH 1837
URBS IN HORTO

cta

POLICE CITY OF CHICAGO INCORPORATED 4th MARCH 1837

SEAL OF COOK COUNTY ILLINOIS
JANUARY 1831

BERKELEY LAB
Lawrence Berkeley National Laboratory

METRO POLICE
NASHVILLE DAVIDSON COUNTY

Get Covered Illinois
The Official Health Marketplace

CHICAGO HARRIS
PUBLIC POLICY | THE UNIVERSITY OF CHICAGO

AGENTIS energy

AUSTRALIAN CONSERVATION FOUNDATION

THE CASE FOUNDATION

EPIC | ENERGY POLICY INSTITUTE AT THE UNIVERSITY OF CHICAGO

Pecan Street Inc.

Arlington Public Schools

SUNLIGHT FOUNDATION

Health Leads

MÉXICO
PRESIDENCIA DE LA REPÚBLICA
ESTADOS UNIDOS MEXICANOS

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

City of Milwaukee

Wiki Energy

mesa
PUBLIC SCHOOLS

HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH

Ushahidi

careerbuilder

NorthShore
University HealthSystem

ENROLL AMERICA

Nurse-Family Partnership
Helping First-Time Parents Succeed

SKILLS FOR CHICAGOLAND'S FUTURE

DIVVY

CONSERVATION INTERNATIONAL

CDOT
CHICAGO DEPARTMENT OF TRANSPORTATION

MuskingumValley
Educational Service Center

JOHNSON COUNTY
KANSAS

TULSA
PUBLIC SCHOOLS

NEW YORK STATE OF OPPORTUNITY
Department of Environmental Conservation

SEDESOL
SECRETARÍA DE DESARROLLO SOCIAL
ESTADOS UNIDOS MEXICANOS

spire

DigitalGlobe

WORLD ECONOMIC FORUM

planet.

New Vision for the Ocean

Syracuse
New York

City of Rotterdam

SANERGY

Rijkswaterstaat
Ministerie van Infrastructuur en Milieu

DATA SCIENCE FOR SOCIAL GOOD
ATLANTA 2016

Data Science For Social Good
Summer Fellowship
THE UNIVERSITY OF CHICAGO

KDD 2014
This year's special theme:
Data Science for Social Good

IBM Social Good Fellowship

UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD

SoGood 2016

BAYES IMPACT

Data For Good Exchange 2016

Volunteerism for the Data Generation: Using Data Science Superpowers for Social Good

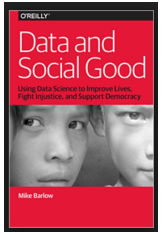
LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

Statistics for Social Good

We're a group of Stanford students, researchers, and faculty exploring the potential to promote social good through effective data analysis.

#Data4Good

NC DATA4GOOD
« data crunch for social good »



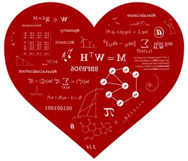
O'REILLY

Data and Social Good

Using Data Science to Improve Lives, Fight Injustice, and Support Democracy

Mike Barlow

DO GOOD DATA 2015



Partnership for Social Good

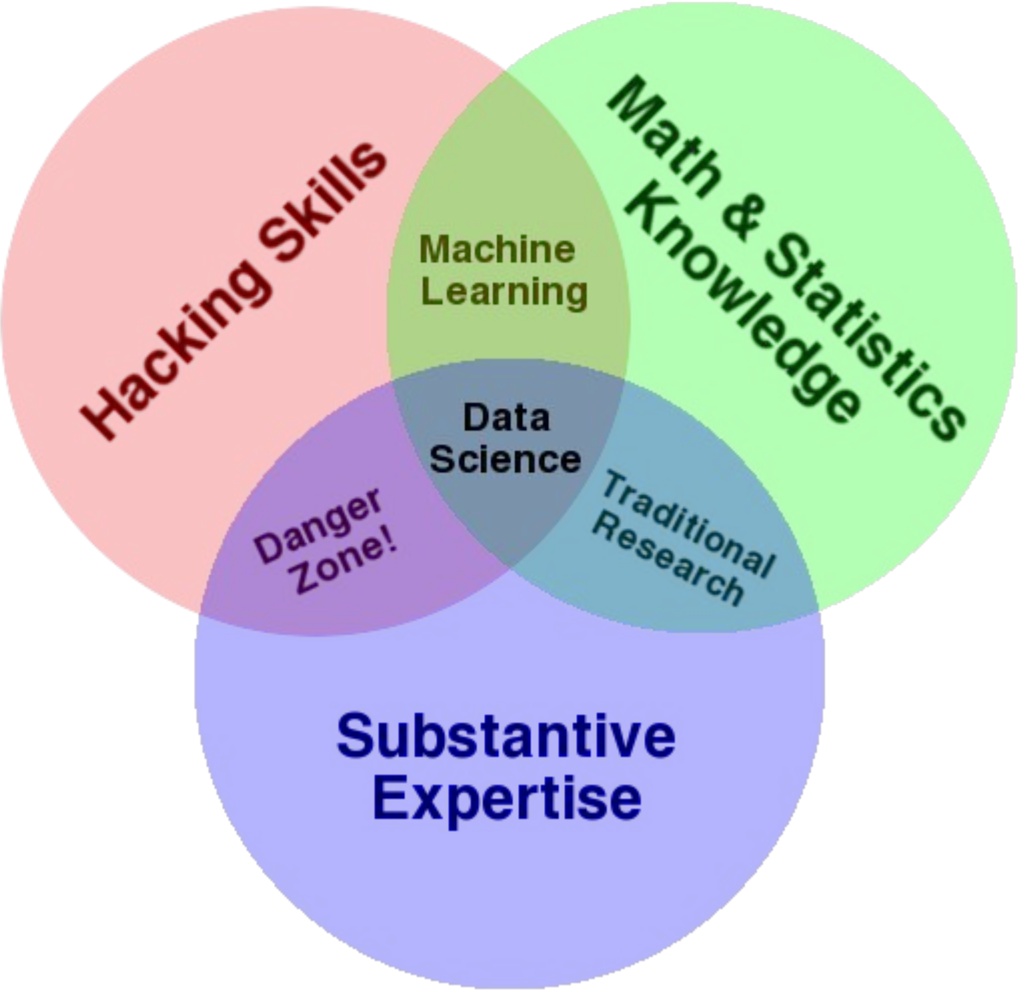
Data Science Initiative
UNC CHARLOTTE

Data Science For Social Good Europe
Summer Fellowship 2018
CASCAIS
NOVA NOVA SCHOOL OF BUSINESS & ECONOMICS
THE UNIVERSITY OF CHICAGO

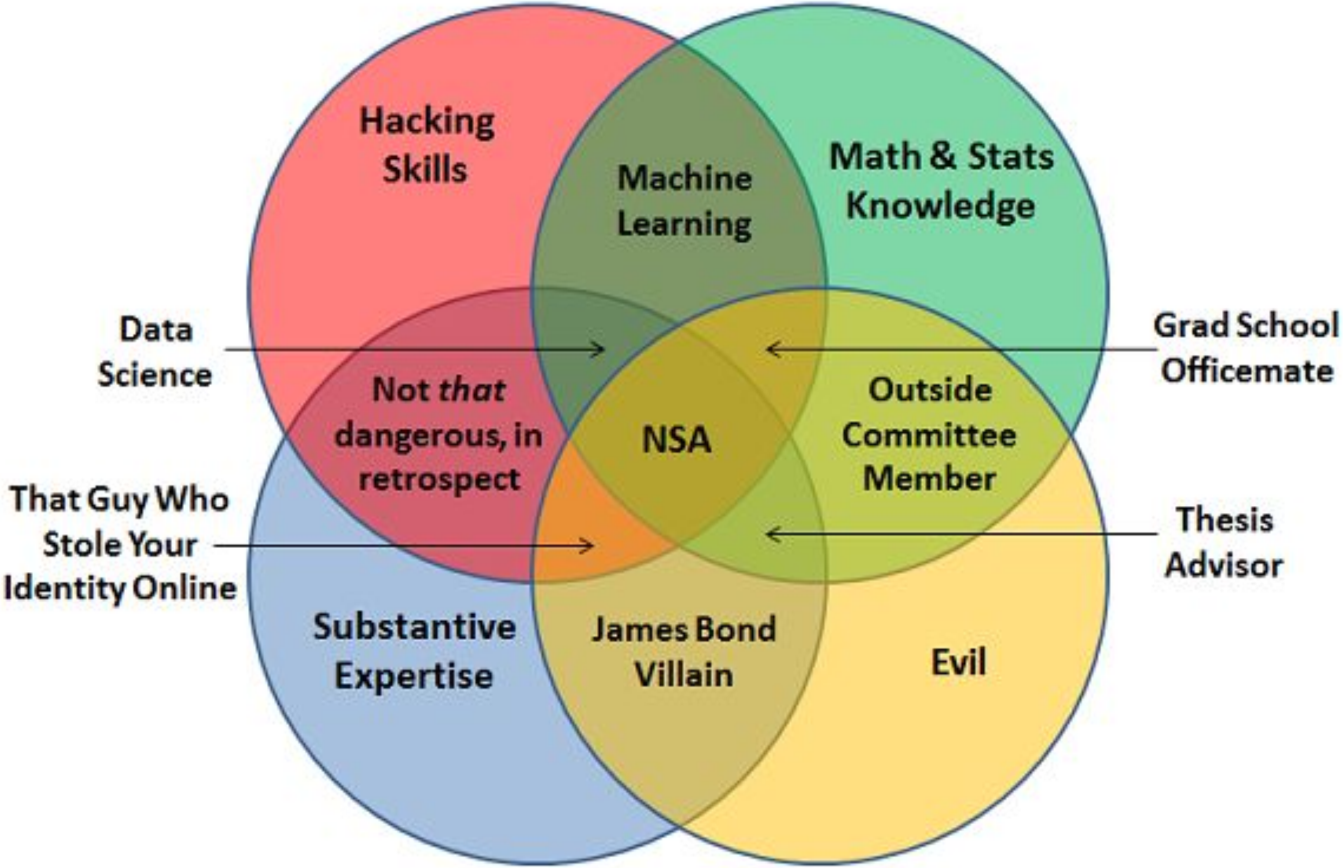
DataKind
USING DATA IN THE SERVICE OF HUMANITY

DRIVENDATA

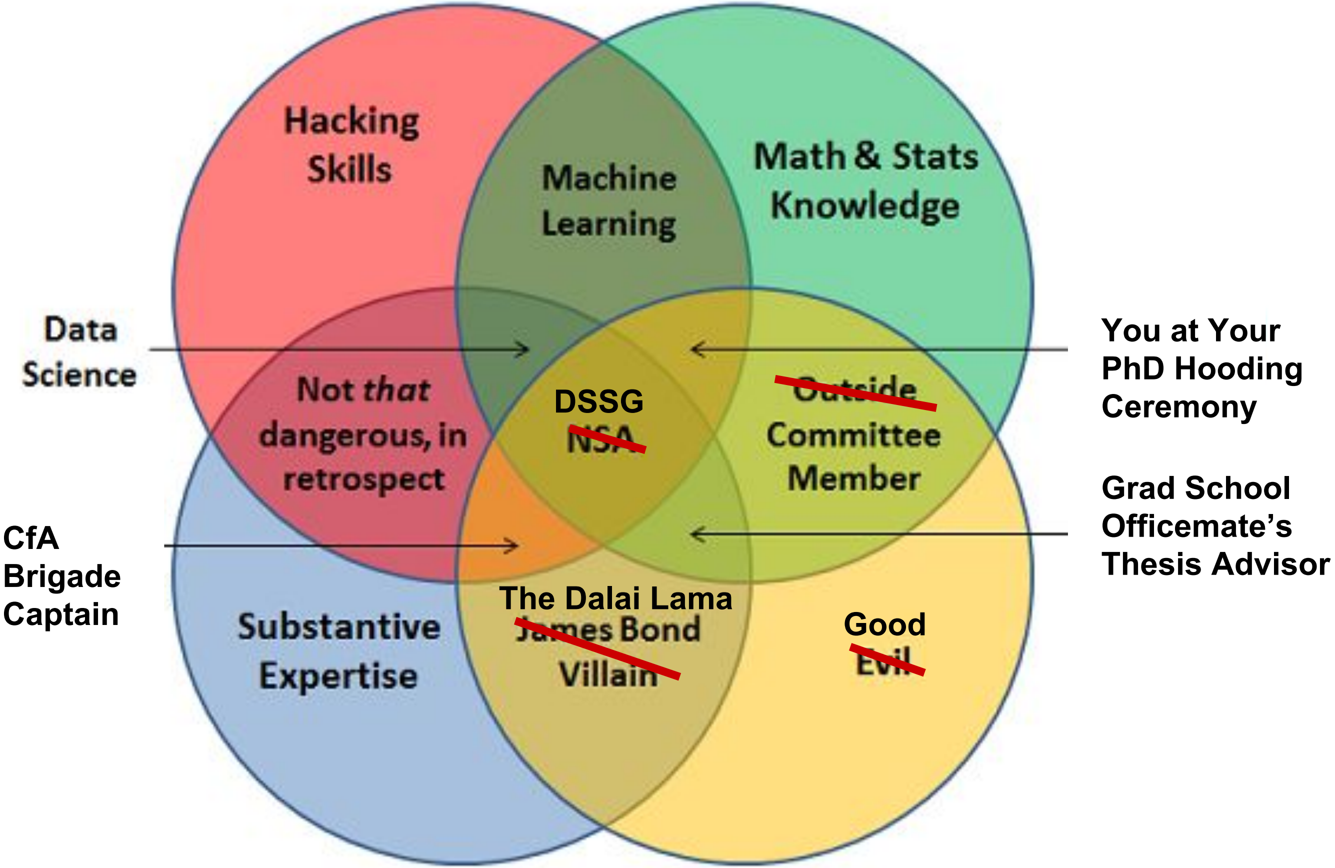
What we got wrong


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

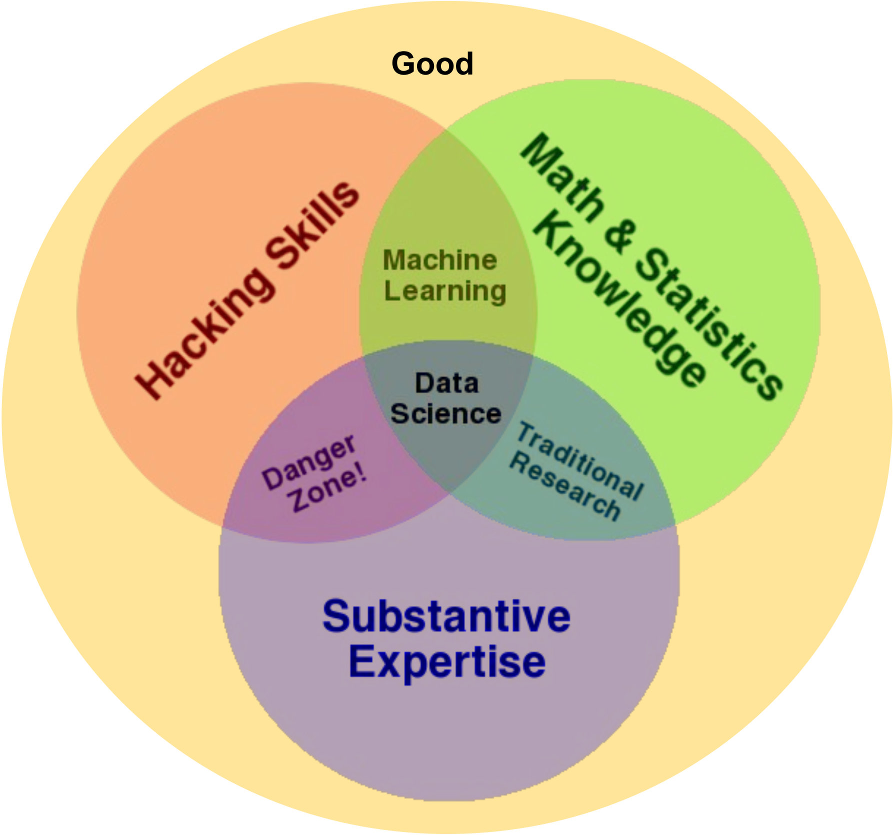
Drew Conway


Hacking Skills
Math & Stats Knowledge
Machine Learning
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil

Joel Grus


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
You at Your PhD Hooding Ceremony
Not *that* dangerous, in retrospect
DSSG
~~NSA~~
~~Outside~~
Committee Member
Grad School Officemate's Thesis Advisor
CfA Brigade Captain
Substantive Expertise
The Dalai Lama
~~James Bond Villain~~
Good
~~Evil~~

All data science must be grounded in a sense of the good



What will it take for socially beneficial data science to go from

doing data science for good

to

doing good data science?

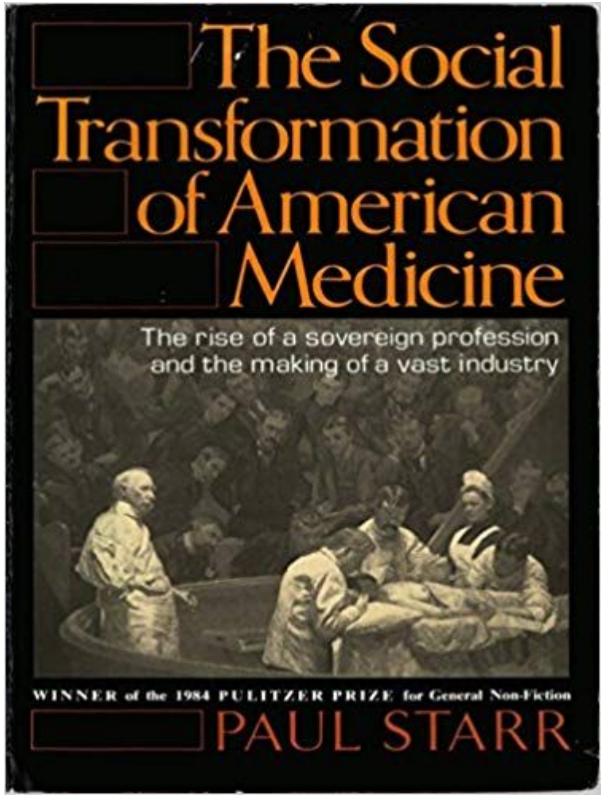
Learning from our early experience as well as that of older professions

Data Science For Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

+



The Social Transformation of American Medicine

The rise of a sovereign profession and the making of a vast industry

WINNER of the 1984 PULITZER PRIZE for General Non-Fiction

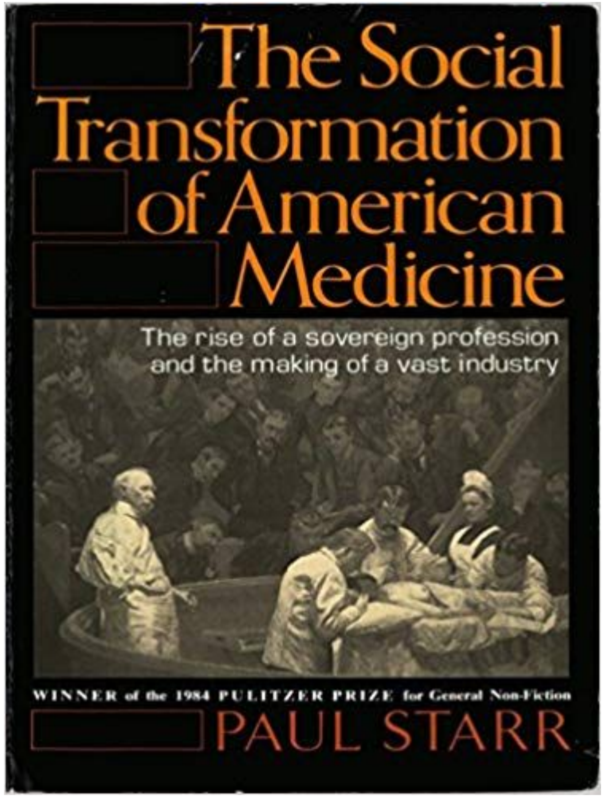
PAUL STARR

Three reasons why
all data science should be taught as data science for social good

Reason 1

As a check on the increasing power of the data scientist

“The dream of reason never took power into account.”

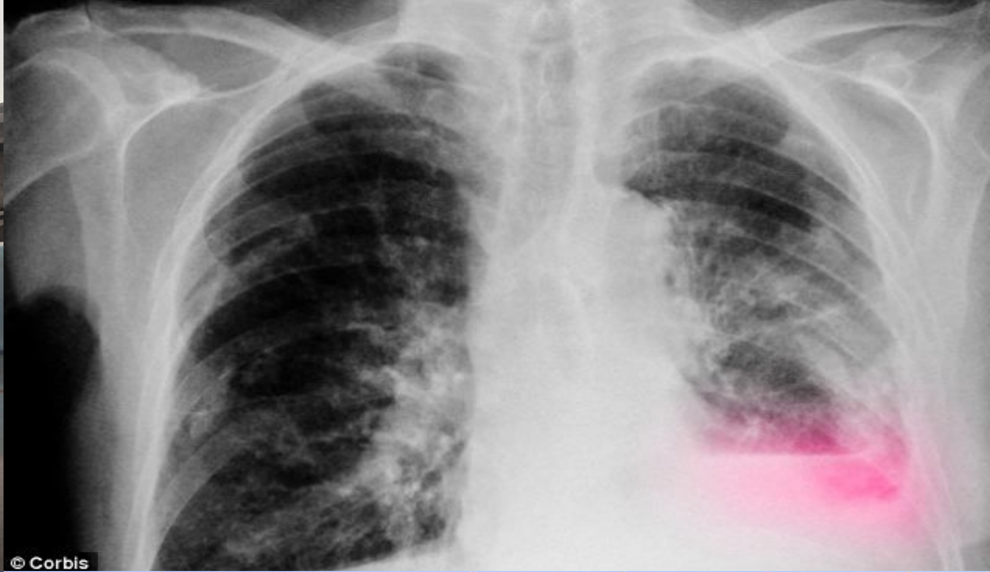

The Social Transformation of American Medicine
The rise of a sovereign profession and the making of a vast industry
WINNER of the 1984 PULITZER PRIZE for General Non-Fiction
PAUL STARR

```
public void Start(AModul Owner)
{
    m_Parent = Owner;
    jade.wrapper.AgentController [illegible] = null;
    try
    {
        //get emulator class
        //package.class -> package.Emu_class
        String parentclass = Owner.getClass().[illegible]me();
        int dot = parentclass.indexOf('.');
        String classname = [illegible]
        if(dot>0)
            classname = parentclass.substring([illegible]+1) + classname;
        classname += parentclass.substring(dot+1);

        //create arguments
        Object[] args = new Object[2];
        args[0] = Owner.getLocalName();
        args[1] = Owner.getAID();
        String name = "Emu " + Owner.getLocalName();

        //start emulator
        a = Owner.getContainerController().createNewAgent(name,classname,args);
```

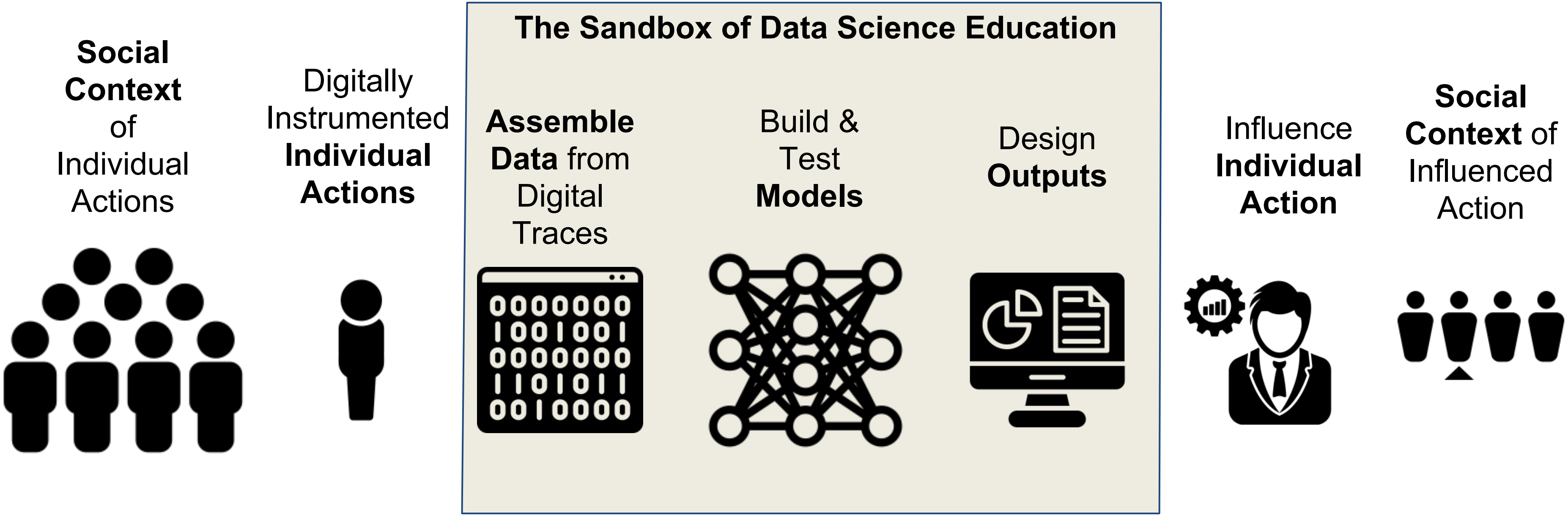
Google
self-driving car
HYBRID


© Corbis

MacBook Pro

THINK
ΣΚΕΨΟΥ
DENK
$2,000
$5,000
$5,000
Ken
WATSON
BRAD

We teach data science at a distance



Social Context of Individual Actions

Digitally Instrumented **Individual Actions**

The Sandbox of Data Science Education

Assemble Data from Digital Traces

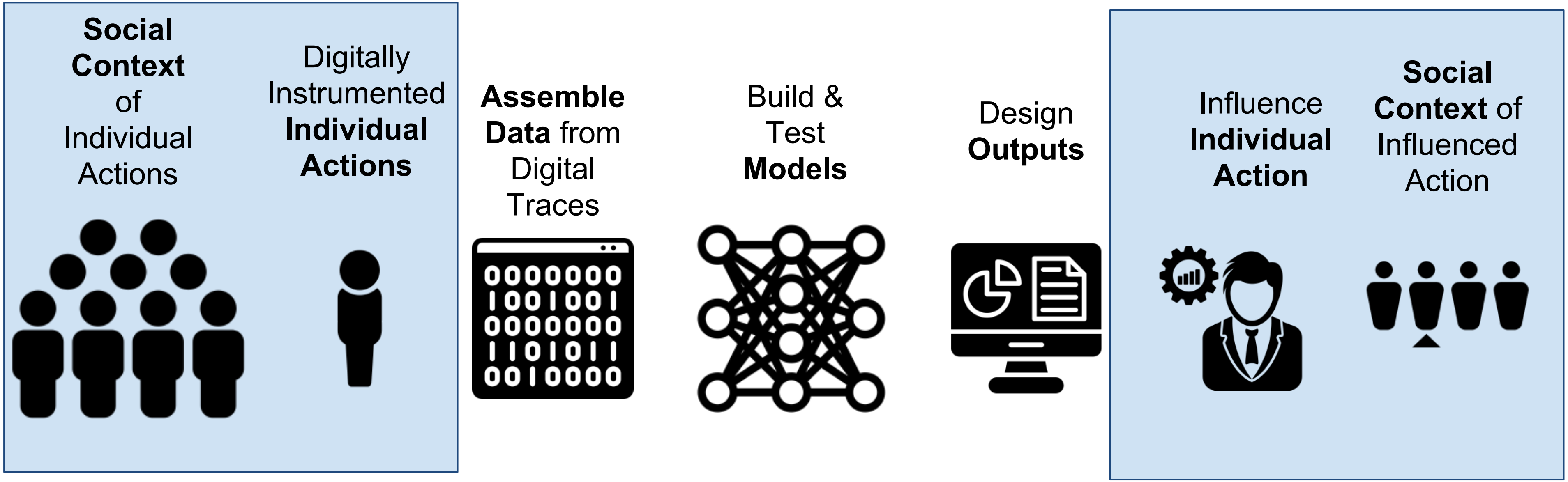
Build & Test **Models**

Design **Outputs**

Influence **Individual Action**

Social Context of Influenced Action

Data science happens in a social context



Social Context of Individual Actions

Digitally Instrumented **Individual Actions**

Assemble Data from Digital Traces

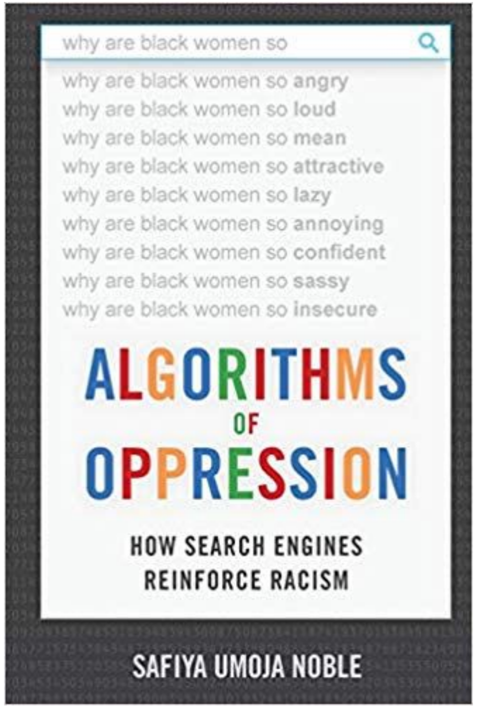
Build & Test **Models**

Design **Outputs**

Influence **Individual Action**

Social Context of Influenced Action

If you are a data scientist, your work will have social consequences



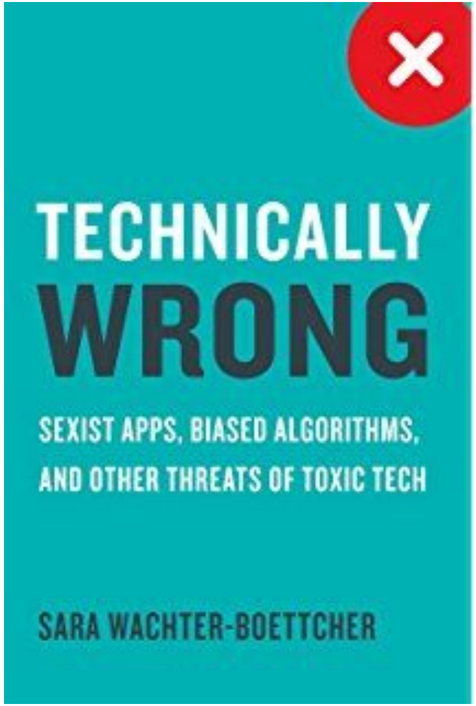
why are black women so

why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

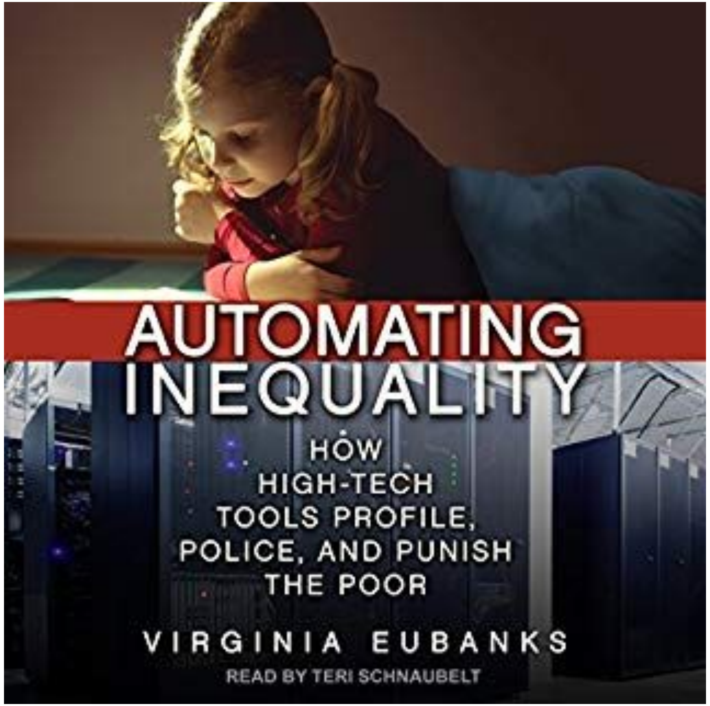
SAFIYA UMOJA NOBLE



TECHNICALLY WRONG

SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH

SARA WACHTER-BOETTCHER



AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR

VIRGINIA EUBANKS

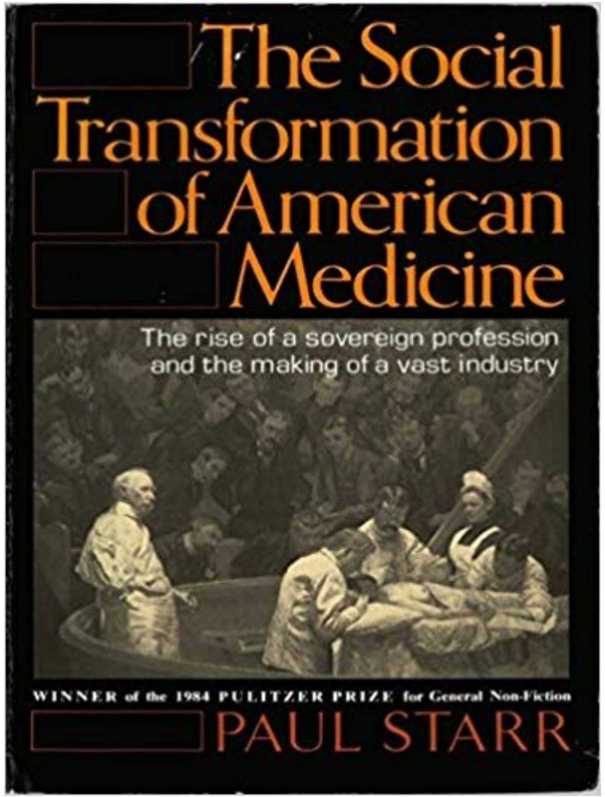
READ BY TERI SCHNAUBELT

EXIT

Reason 2

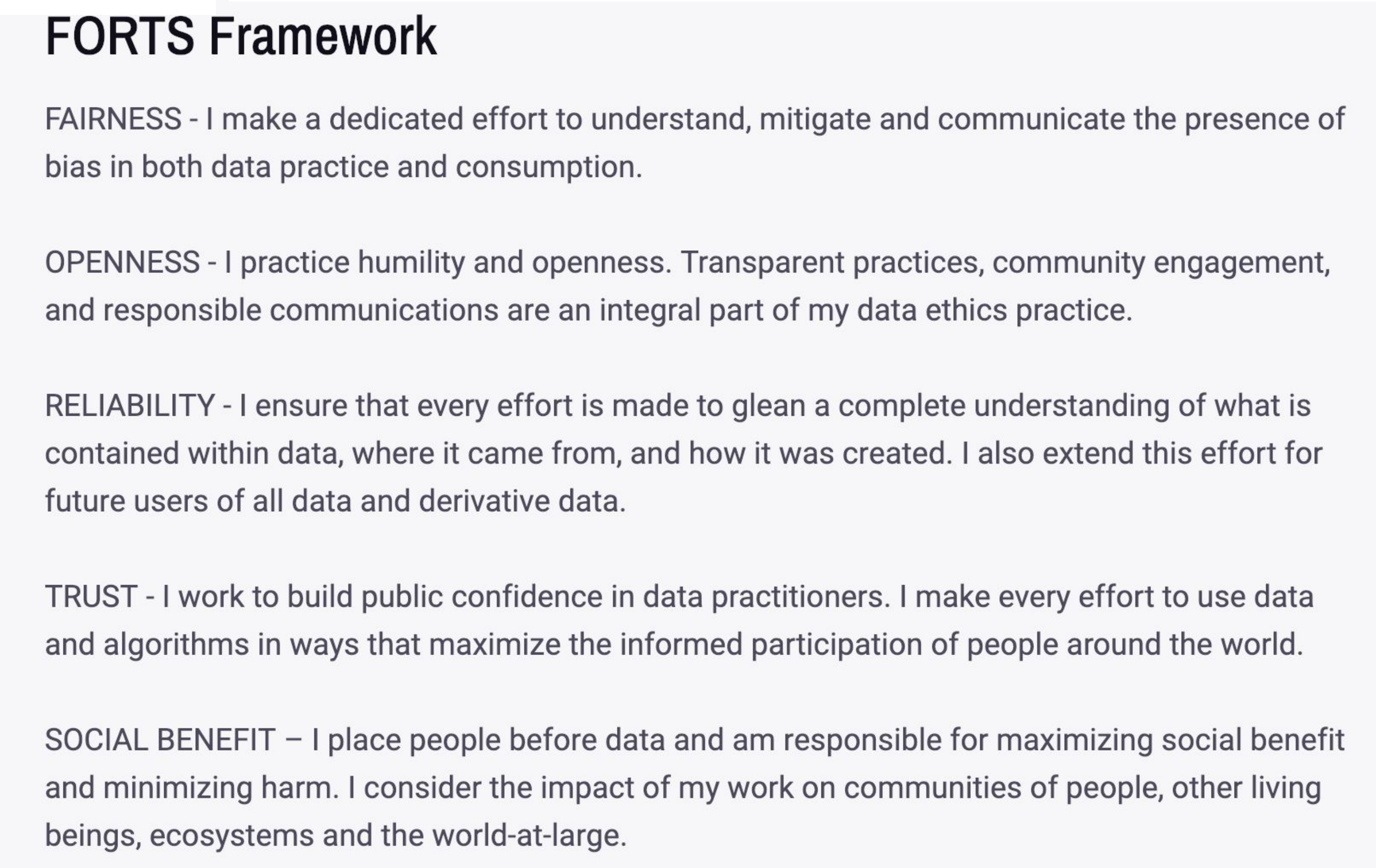
To establish professional norms

“A profession, sociologists have suggested, is an occupation that regulates itself through systematic, required training and collegial discipline;
that has a base in technical, specialized knowledge;
and that **has a service rather than profit orientation, enshrined in its code of ethics.**”


The Social Transformation of American Medicine
The rise of a sovereign profession and the making of a vast industry
WINNER of the 1984 PULITZER PRIZE for General Non-Fiction
PAUL STARR

Data for Democracy

The Global Data Ethics Project



FORTS Framework

FAIRNESS - I make a dedicated effort to understand, mitigate and communicate the presence of bias in both data practice and consumption.

OPENNESS - I practice humility and openness. Transparent practices, community engagement, and responsible communications are an integral part of my data ethics practice.

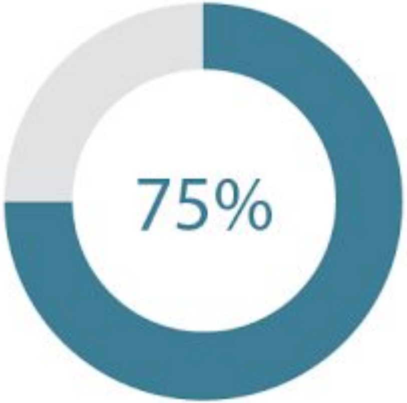
RELIABILITY - I ensure that every effort is made to glean a complete understanding of what is contained within data, where it came from, and how it was created. I also extend this effort for future users of all data and derivative data.

TRUST - I work to build public confidence in data practitioners. I make every effort to use data and algorithms in ways that maximize the informed participation of people around the world.

SOCIAL BENEFIT – I place people before data and am responsible for maximizing social benefit and minimizing harm. I consider the impact of my work on communities of people, other living beings, ecosystems and the world-at-large.

Reason 3

To attract and keep the best and brightest minds to our profession



% of adults under the age of 35 willing to take a pay cut to work at a job that has social purpose.

IMPACT

SOLVE A PROBLEM A REAL PERSON CARES ABOUT

Elissa

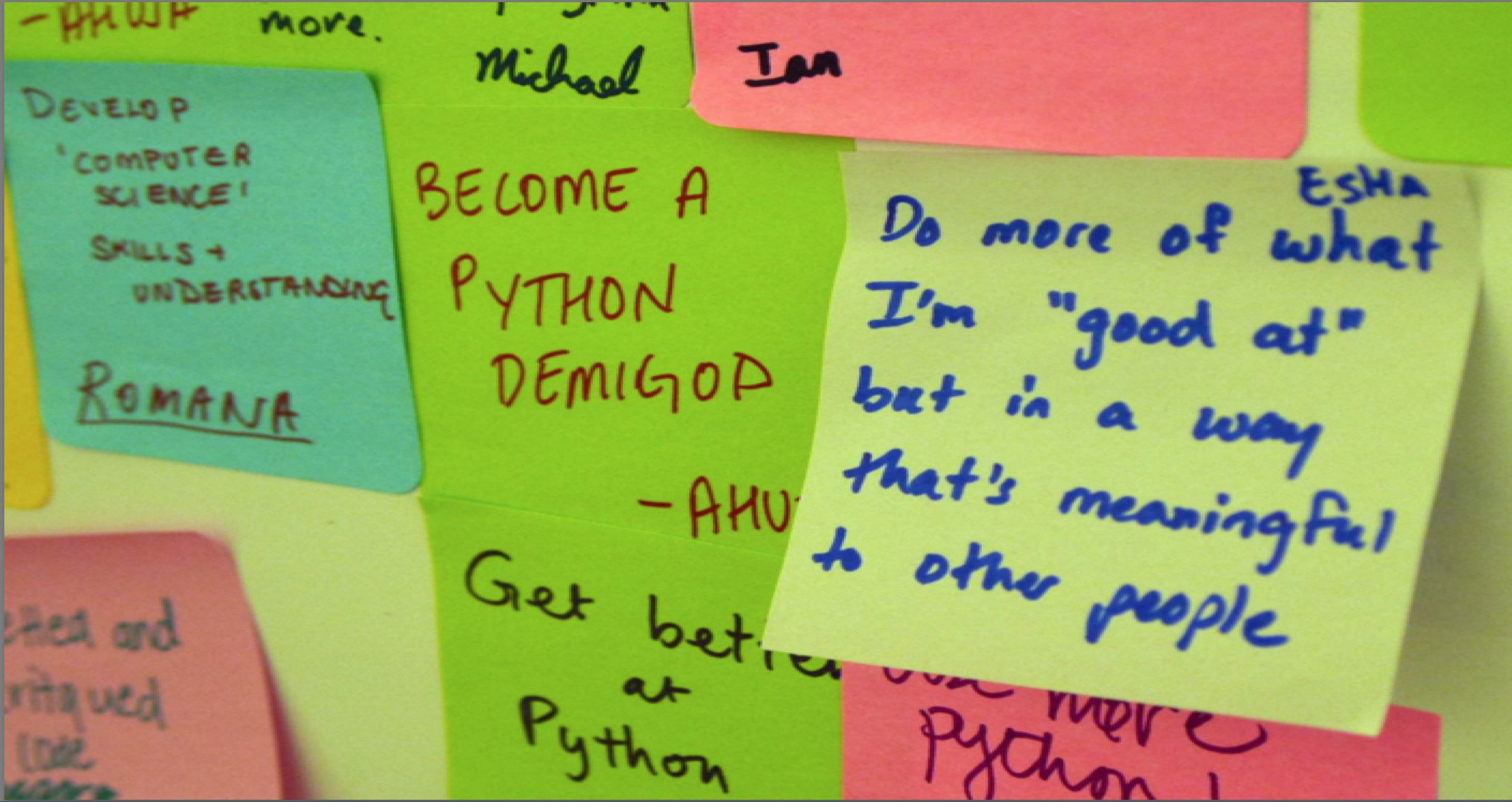
- Get out of sterile environment of academic research into real and current problem solving for social good!

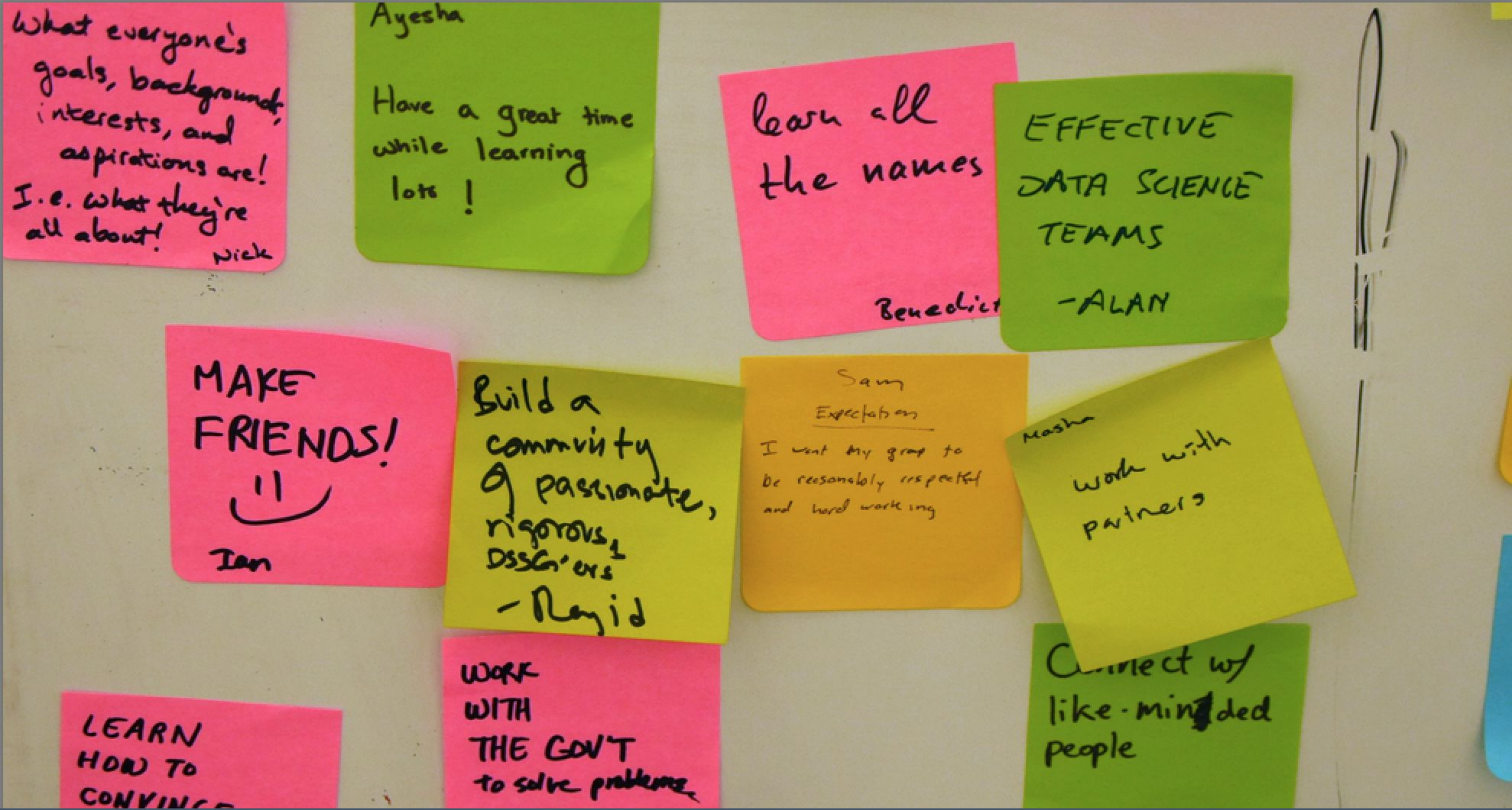
Apply my skills to a project with social impac

ASIF

Actionable Data Science

become better


-AHUJA more.
Michael
Ian
DEVELOP 'COMPUTER SCIENCE' SKILLS + UNDERSTANDING
ROMANA
BECOME A PYTHON DEMIGOD -AHU
Esha
Do more of what I'm "good at" but in a way that's meaningful to other people
Get better at Python
more Python!



What everyone's goals, background, interests, and aspirations are! I.e. what they're all about! Nick

Ayesha

Have a great time while learning lots !

learn all the names

Benedict

EFFECTIVE DATA SCIENCE TEAMS -ALAN

MAKE FRIENDS! :)

Ian

Build a community of passionate, rigorous, DSSG'ers -Rayid

Sam

Expectations

I want my group to be reasonably respectful and hard working

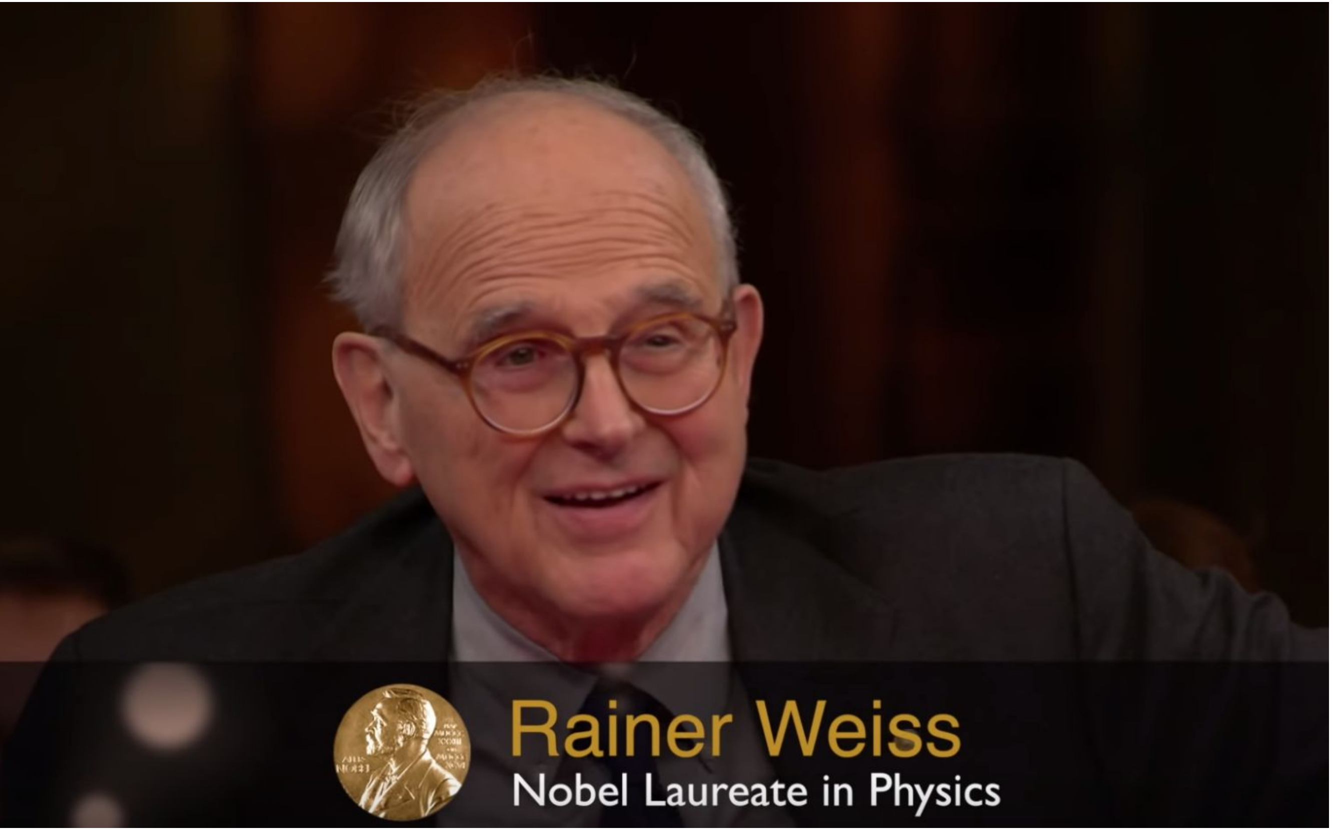
Masha

work with partners

LEARN HOW TO CONVINC

WORK WITH THE GOV'T to solve problems

Connect w/ like-minded people



"You do things that you get a kick out of, and luckily, sometimes they turn out to be important."

Nobel Minds, 2017

It’s time to update what it means to do good data science

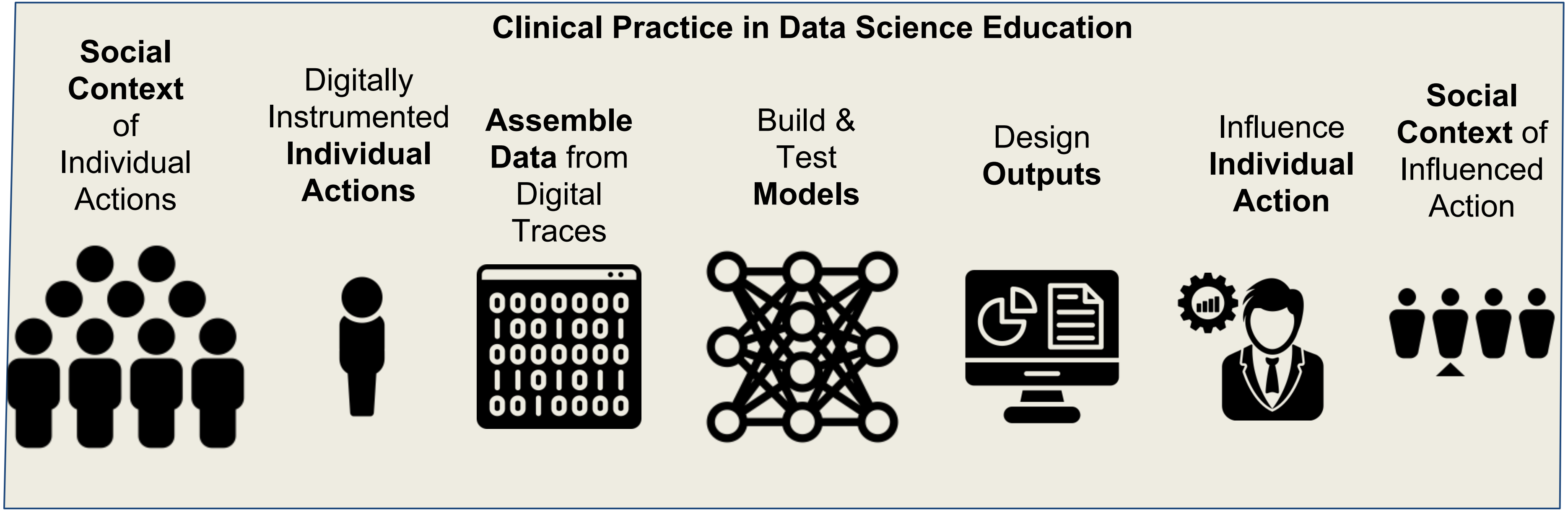
Three suggestions for how

all data science can be taught as
data science for social good

Suggestion 1

Add **clinical practice requirements** to all data science programs

From classroom to clinic
Putting data scientists into the social context of their algorithms



Clinical Practice in Data Science Education

Social Context of Individual Actions

Digitally Instrumented **Individual Actions**

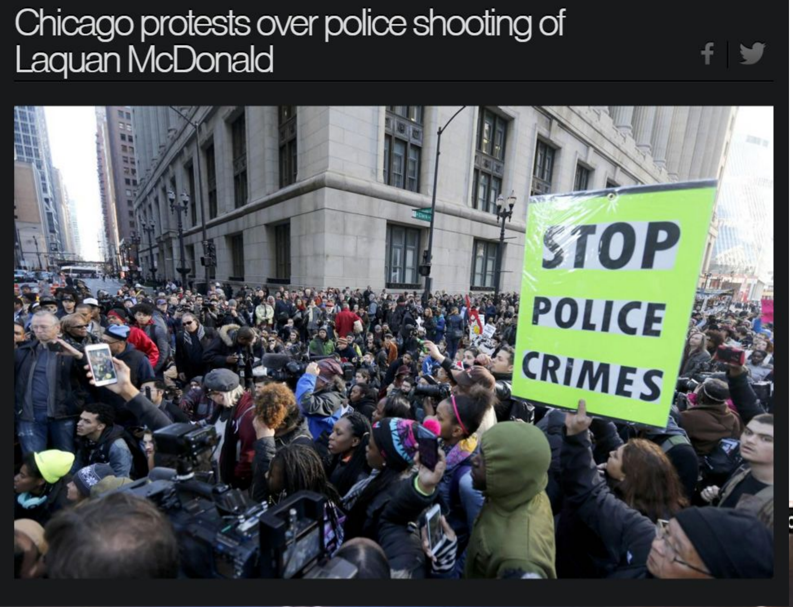
Assemble Data from Digital Traces

Build & Test **Models**

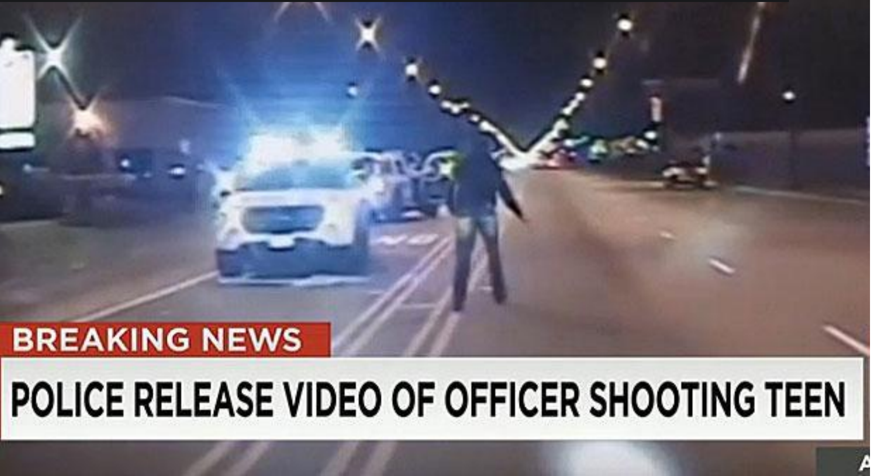
Design **Outputs**

Influence **Individual Action**

Social Context of Influenced Action



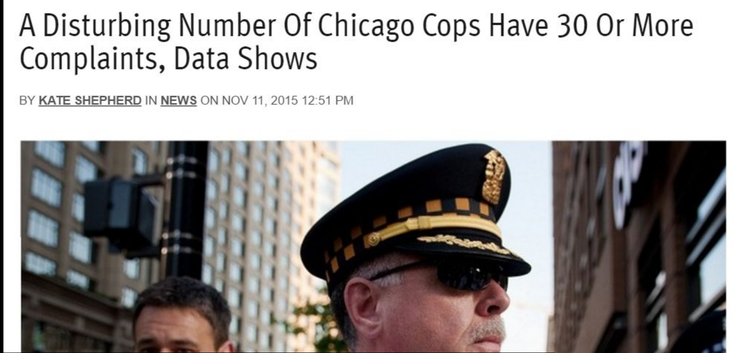
Chicago protests over police shooting of Laquan McDonald



cago Pol

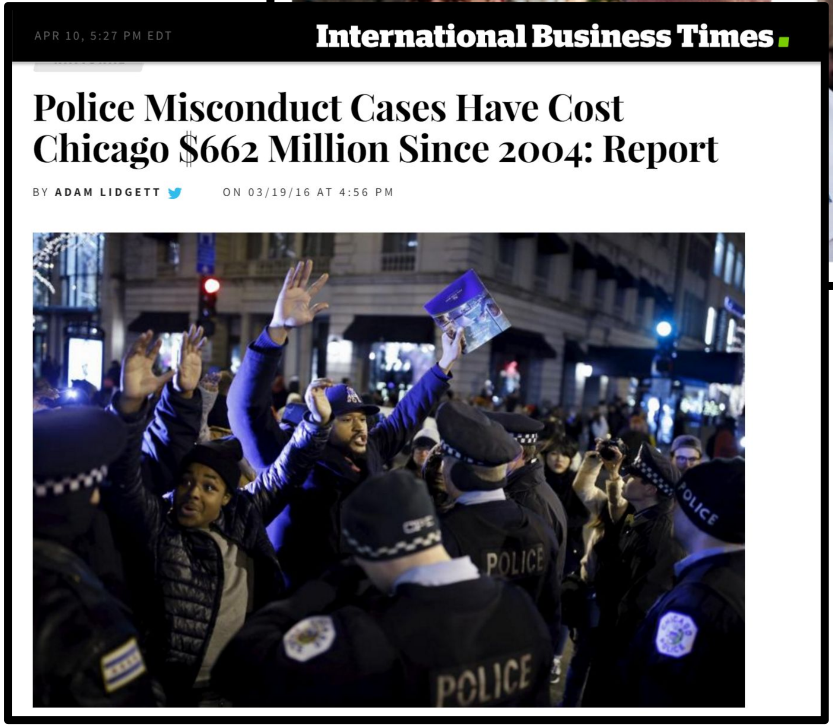
BREAKING NEWS

POLICE RELEASE VIDEO OF OFFICER SHOOTING TEEN



A Disturbing Number Of Chicago Cops Have 30 Or More Complaints, Data Shows

BY KATE SHEPHERD IN NEWS ON NOV 11, 2015 12:51 PM



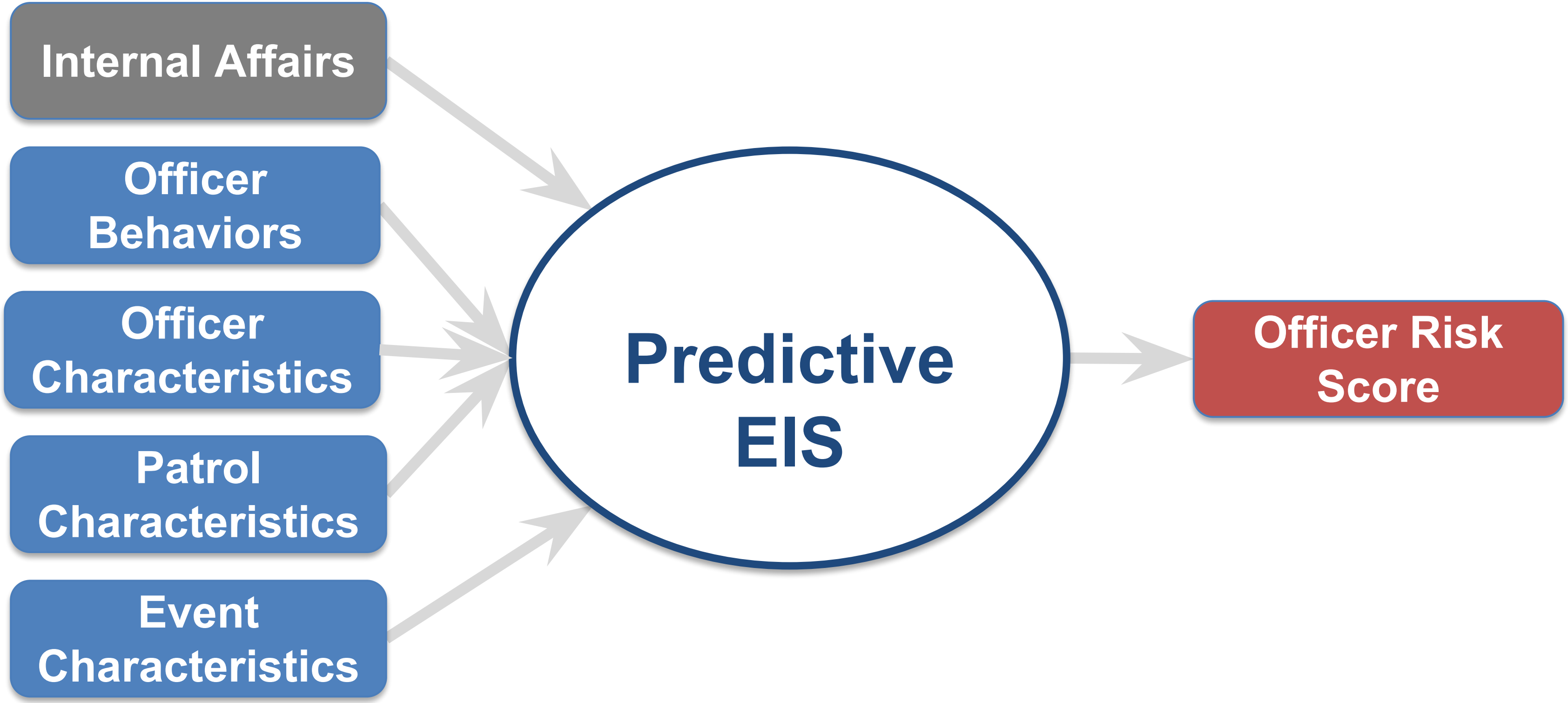
APR 10, 5:27 PM EDT

International Business Times

Police Misconduct Cases Have Cost Chicago $662 Million Since 2004: Report

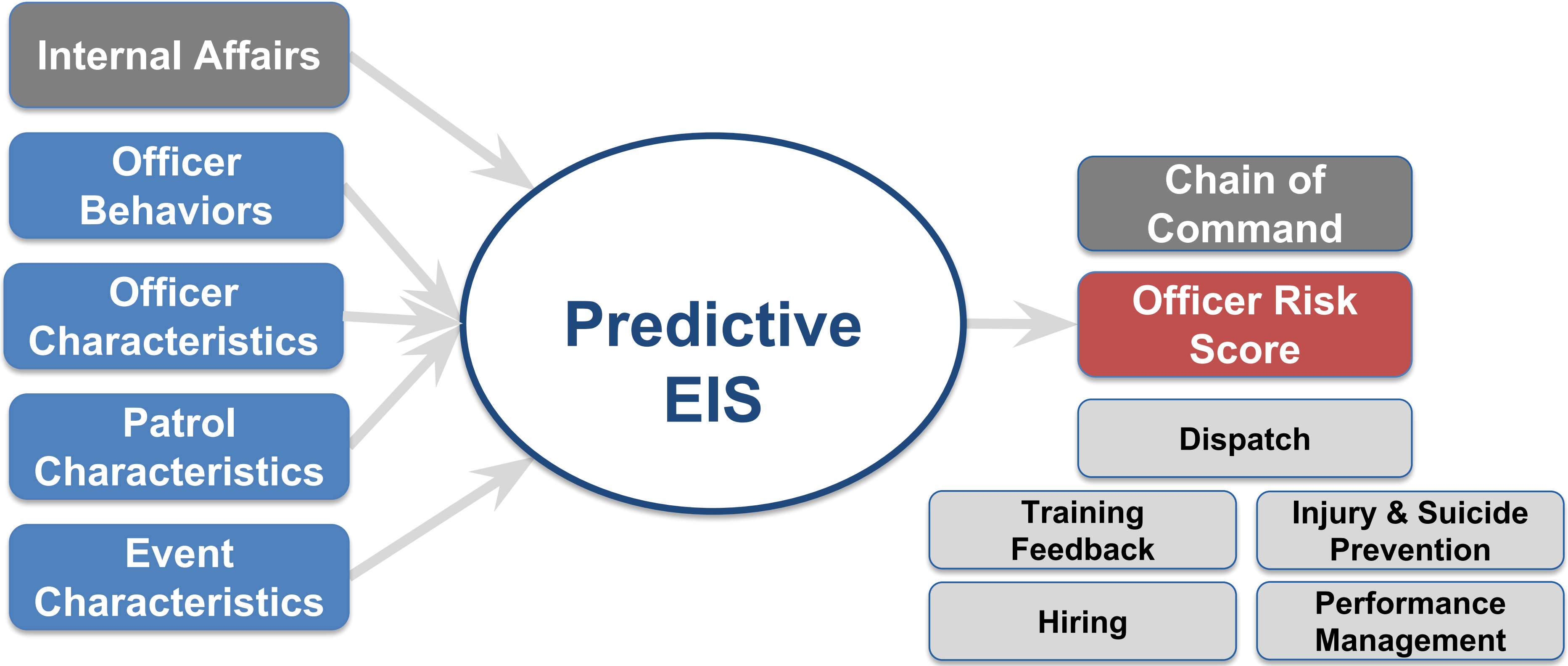
BY ADAM LIDGETT ON 03/19/16 AT 4:56 PM

Predicting officer's risk of adverse event



From classroom to clinic
Putting data scientists into the social context of their algorithms

Toward a more holistic solution



This will be difficult for many data science programs to do.

We’re here to help.

Suggestion 2

Add written or verbal **discussion of the social and ethical implications** of models into every problem set in data science coursework

Questions we’ve asked

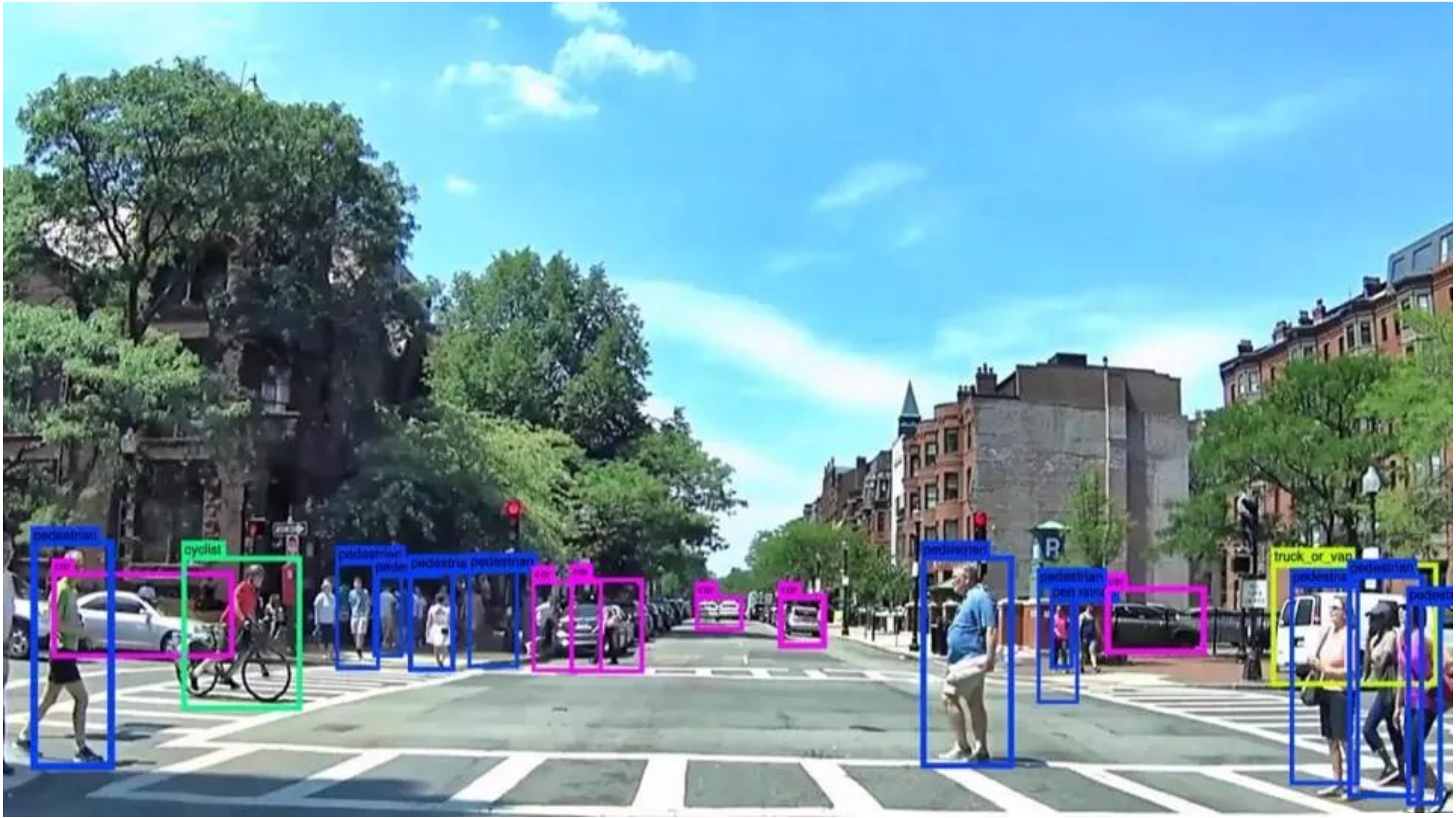
- What biases may exist in the data you’ve been given? How can you find out?
- How will your choices with tuning parameters affect different populations represented in the data?
- How do you know you aren’t getting the right answer to the wrong question?
- How would you justify what you’d built to someone whose welfare is made worse off by the implementation of your algorithm?

Predicting back-to-school enrollment for Chicago Public Schools
Predicting Social Se
Heyday

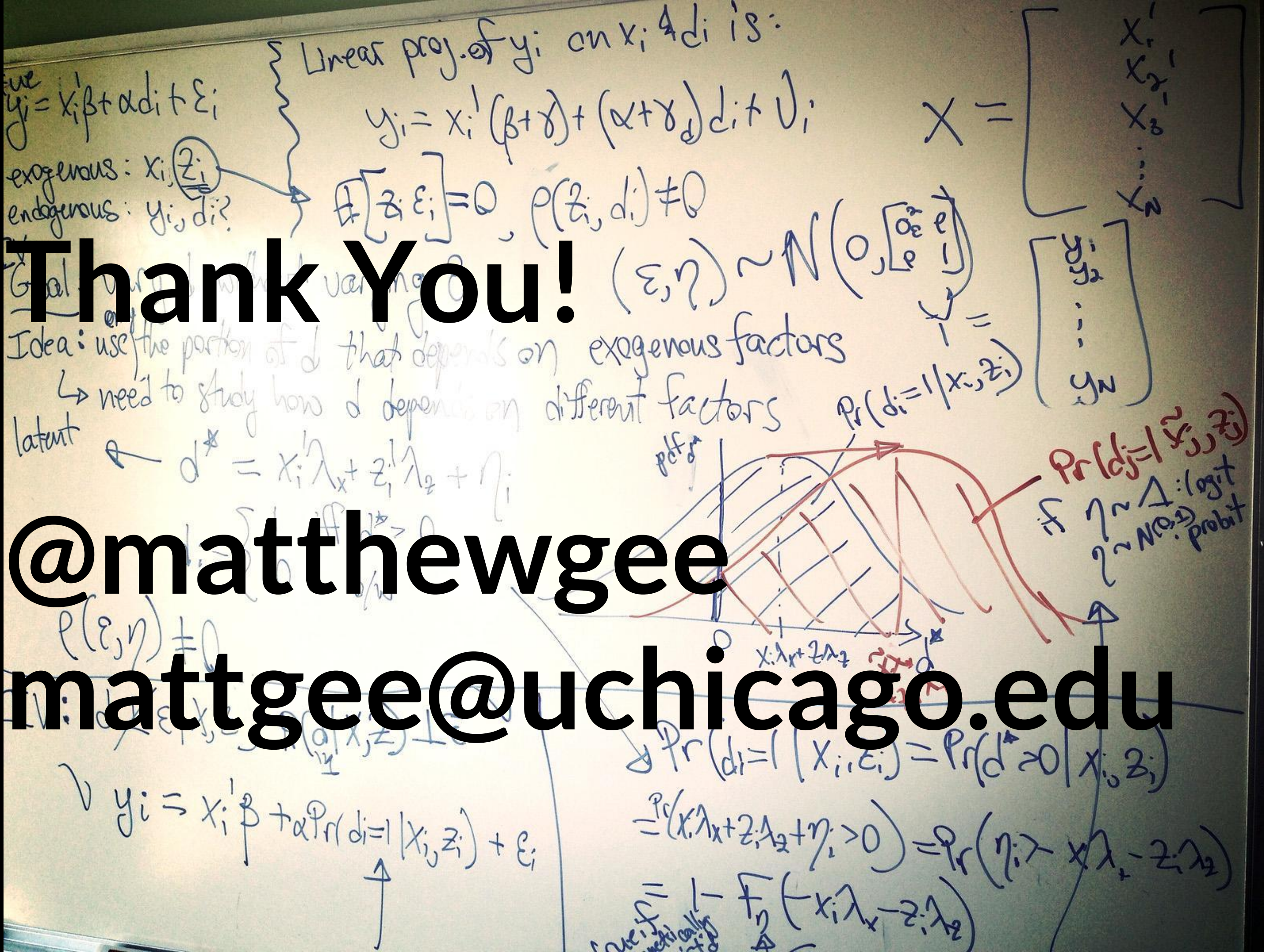
Suggestion 3

Provide guidance to employers for incorporating ethics case studies into hiring, apprenticeships, and mentorship opportunities.

What could happen if we teach data scientists to always be doing data science for social good



NUERALA, 2018


true $y_i = x_i'\beta + \alpha d_i + \varepsilon_i$
Linear proj. of y_i on x_i & d_i is:
$y_i = x_i'(\beta+\gamma) + (\alpha+\gamma_d)d_i + v_i$
$X = [x_1', x_2', x_3', \dots, x_N']$
exogenous: x_i, z_i
endogenous: y_i, d_i?
$E[z_i \varepsilon_i] = 0$, $\rho(z_i, d_i) \neq 0$
$(\varepsilon, \eta) \sim N(0, [\sigma_\varepsilon^2\ \rho;\ \rho\ 1])$
$Y = [y_1, y_2, \dots, y_N]$
Idea: use the portion of d that depends on exogenous factors
↳ need to study how d depends on different factors
latent $d^* = x_i'\lambda_x + z_i'\lambda_z + \eta_i$
pdf η
$Pr(d_i=1|x_i, z_i)$
$Pr(d_j=1|x_j, z_j)$
if $\eta \sim \Lambda$: logit
$\eta \sim N(0,1)$: probit
$\rho(\varepsilon, \eta) \neq 0$
$Pr(d_i=1|x_i, z_i) = Pr(d^*>0|x_i, z_i)$
$= Pr(x_i\lambda_x + z_i\lambda_z + \eta_i > 0) = Pr(\eta_i > -x_i\lambda_x - z_i\lambda_z)$
$= 1 - F_\eta(-x_i\lambda_x - z_i\lambda_z)$
$y_i = x_i'\beta + \alpha Pr(d_i=1|x_i, z_i) + \varepsilon_i$

Thank You!

@matthewgee

mattgee@uchicago.edu