BIG DATA:
A Tool for Fighting Discrimination and Empowering Groups

PREFACE

In May 2014, the Executive Office of the President concluded its 90-day study of Big Data and privacy and released a report entitled *Big Data: Seizing Opportunities, Preserving Values.* The report highlighted certain positive uses of Big Data, such as identifying health risks at an early stage, creating efficiencies in energy distribution, and uncovering fraud through predictive analysis. However, it also concluded that Big Data analytics could facilitate discrimination in housing, credit, employment, health, education, and a range of other markets. These potential benefits and drawbacks underscore the need to better understand how Big Data will shape our lives in years to come.

Recognizing the 50th anniversary of the Civil Rights Act and the challenges to fighting discrimination in the 21st century, the case studies included in this report show how businesses, governments, and civil society organizations are leveraging Big Data (and other data sets[1]) to protect and empower vulnerable groups, including by providing access to job markets, uncovering discriminatory practices, and creating new tools to improve education and assist those in need. While by no means an exhaustive list of Big Data's potential to uncover and fight discrimination, we offer these examples to show how Big Data already is redefining efforts to ensure equal opportunity for all.

We would like to thank the Anti-Defamation League for its partnership in preparing this report, Jared Bomberg and Julian Flamant at Hogan Lovells US LLP for providing essential research and drafting support, and members of the FPF Advisory Board for reviewing drafts and providing guidance.

We hope these examples will contribute to discussions about Big Data's impact on discrimination.

Jules Polonetsky
Executive Director and Co-Chair
Future of Privacy Forum

Christopher Wolf
Founder and Co-Chair
Future of Privacy Forum

[1] Some of the datasets being used by businesses, governments, and civil society organization will be considered by some to be more appropriately classified as "small data" as they are built, in some cases, from fixed and pre-existing datasets or rely on limited data inputs. We ask that our readers recognize the value of evolving uses and usefulness of data as exposed by these cases and imagine how that value will be compounded as applications of Big Data catch up to technological capabilities.

TABLE OF CONTENTS

I. Seeing Beyond Bias to Provide New Opportunities 1

Case Study 1: Workplace Diversity (Entelo) 1

Case Study 2: Opportunity for Advancement (Google) 2

Case Study 3: Allocation of Public Works (Cedar Grove Institute) 3

Case Study 4: Demographics of Health (State of New York) 4

II. Transparency is a Necessary Disinfectant 5

Case Study 5: Hate Crime Report (Federal Bureau of Investigation) 5

Case Study 6: *McClesky v. Kemp* 6

Case Study 7: Discrimination Complaint Data (EEOC) 7

Case Study 8: *United States v. Sterling* 8

Case Study 9: Mapping Public School Segregation (Urban Institute) 9

III. Long-Term Problems Require Innovative Solutions 10

Case Study 10: Education for All (NSBA) 10

Case Study 11: Tracking Migratory Patterns (United Nations) 11

Case Study 12: Economic Development and Equality (OECD) 12

Case Study 13: Finding Missing and Exploited Children (Palantir/NCMEC) 13

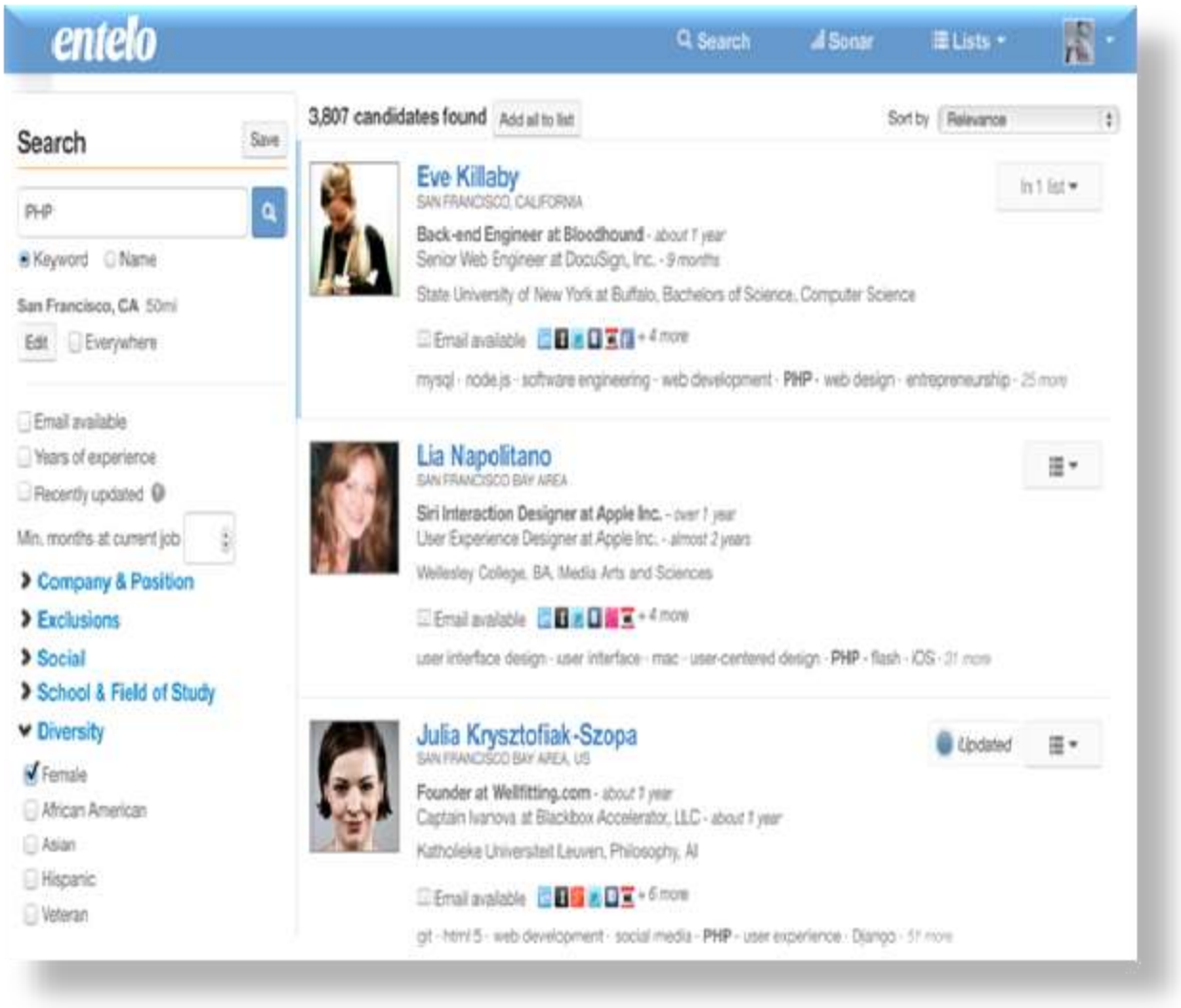
Case Study 14: Human Trafficking (Palantir/Polaris Project) 14

I. SEEING BEYOND BIAS TO PROVIDE NEW OPPORTUNITIES

Case Study 1: Workplace Diversity (Entelo)

Entelo Diversity, a candidate recruiting platform launched in April 2014, is improving workplace diversity by empowering recruiters to search for job candidates from within underrepresented segments of the population. Using a proprietary algorithm, this workplace diversity tool sifts through publicly available data—pulled from social media platforms—to match recruiters with candidates who hold necessary qualifications, but also meet particular diversity requirements. The tool can filter candidates based on gender, race, and military history in five categories: Female, African American, Asian, Hispanic, and Veteran.

Example of an Entelo Search



Source: http://blog.entelo.com/company-news/announcing-entelo-diversity

Case Study 2: Opportunity for Advancement (Google)

A challenge for the technology industry is ensuring diversity in the workplace. Twitter has recently reported that 90% of its global "tech" employees are male and Google admits "[it's] not where [it] wants to be when it comes to diversity," with only 17% women among its tech workforce.

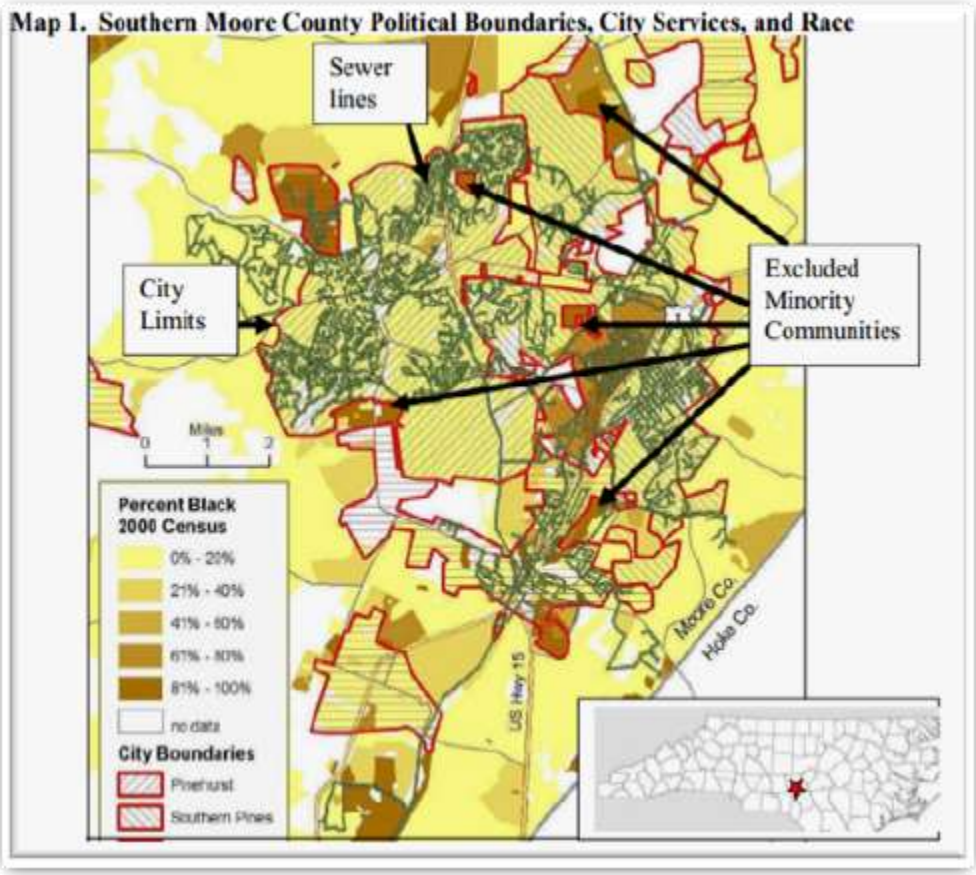
The challenge for Google is apparent within its management and leadership ranks where the workforce is dominated—to an extent—by men. Recognizing the value of a diverse workforce, Google is leveraging its data analytics capabilities to help change those numbers. Through analytics and research, the company identified that its employee advancement conventions, which in part call on employees to nominate themselves for promotions favor men, who are more likely to 'raise their hands' than women. Using the lessons gleaned from workplace analytics, Google has implemented programs to encourage women to apply for promotions and has reformed its hiring practices to ensure that female candidates meet female employees, with whom they are more likely to highlight their career achievements and credentials.

Source: http://www.google.com/diversity/at-google.html#tab=tech

Case Study 3: Allocation of Public Works (Cedar Grove Institute)

The Cedar Grove Institute for Sustainable Communities is a non-profit organization that leverages open data to explore disparities in allocation of geographic boundaries and public works of various communities. Cedar Grove uses a combination of demographic analysis, contextual investigation, housing and economic analysis, and geographic information systems to explore potentially discriminatory implications of public policy decisions. In a 2005 study, *Segregation in the Modern South: A Case Study of Southern Moore County*, Cedar Grove combined census data and other publicly available surveys and demographic information to explore the impact on community development of land annexation policies of Moore County North Carolina.

Map Showing Excluded Minority Communities



Source: http://www.cedargroveinst.org/

Case Study 4: Demographics of Health (State of New York)

In 2011, the Institute of Medicine, the health arm of the National Academy of Sciences, reported that lesbian, gay, bisexual, and transgender (LGBT) individuals have unique health experiences and needs, but as a nation, we do not know exactly what these needs are. The IOM also reported that clinicians and researchers are faced with incomplete information regarding the health status of LGBT individuals and that current research has not adequately examined subpopulations, particularly racial and ethnic groups and peoples' health needs based on age.

In response, the State of New York launched a coordinated, multi-agency effort to strengthen data collection regarding LGBT individuals in New York. The campaign will rely on data collected on a self-reporting basis by the New York's Department of Health, Department of Corrections and Community Supervision, Office for the Aging, Office of Mental Health, Office of Alcohol and Substance Abuse Services, Office of Temporary and Disability Assistance, Office of Children and Family Services, and Office for People with Developmental Disabilities. The data collected will be shared among the eight agencies to create a comprehensive method for identifying the needs of LGBT individuals. It is hoped that stronger data sets will empower the State and others to create more tailored approaches to reduce health disparities impacting LGBT individuals.

News Article
Wednesday, July 23, 2014

GOVERNOR CUOMO ANNOUNCES MULTI-AGENCY STATE EFFORT TO ADDRESS LGBT DISPARITIES

New York becomes first state in the nation with coordinated statewide strategy to improve LGBT data collection

Governor Andrew M. Cuomo announced that New York State is undertaking a coordinated, multi-agency effort to strengthen data collection for lesbian, gay, bi-sexual and transgender (LGBT) New Yorkers. Outlined in the first report by the State's Interagency LGBT Task Force, this statewide effort to include sexual orientation and gender identity information in data collections will allow the state to better tailor services to meet LGBT needs, ultimately improving the health and lives of thousands of New Yorkers. This effort makes New York the first state in the nation to employ a coordinated strategy to develop its data collection procedures for the LGBT community.

"New York State has a long history of advancing progressive ideals, and today we are continuing to lead the nation by identifying new ways to improve services and better meet the needs of the LGBT community," Governor Cuomo said. "By being more inclusive with how state agencies monitor the demographics of those they serve, we can address health and financial disparities, safety concerns, and a myriad of other issues that impact LGBT New Yorkers. This is another step forward for an important community in New York, and our administration will continue standing up for all New Yorkers, regardless of their sexual orientation or gender identity."

The Institute of Medicine in its March 2011 report, The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding, emphasized the need for collection of gender identity

Source: http://www.ocfs.state.ny.us/main/view_article.asp?ID=833

II. TRANSPARENCY IS A NECESSARY DISINFECTANT

Case Study 5: Hate Crime Report (Federal Bureau of Investigation)

The FBI's Uniform Crime Reporting program for hate crimes is a nationwide effort of more than 13,000 city, university and college, county, state, tribal and federal law enforcement agencies voluntarily reporting data on crimes brought to their attention. The data has become one of the country's primary methods of tracking, analyzing, and responding to hate crime violence. Hate crime incidents are broken down into various categories such as offense type, location, bias motivation, victim type, number of individual victims, number of offenders, and the race of the offenders. The streamlined and searchable nature of the data provides law enforcement and civil society groups an ability to monitor and analyze hate crimes and better direct training, advocacy, and legal efforts to reduce the number of hate crimes and improve the response to hate crime incidents.

Sources: http://www.fbi.gov/about-us/cjis/ucr/hate-crime/2012/hate-crime http://www.adl.org/press-center/press-releases/hate-crimes/adl-welcomes-fbi-hate-crime.html

Case Study 6: *McClesky v. Kemp*

McCleskey, an African American man, was sentenced to death after being convicted of armed robbery and the murder of a white police officer. In a writ of habeas corpus, McClesky argued that the Georgia capital sentencing process was administered in a racially discriminatory manner in violation of the Eighth and Fourteenth Amendments. In support of the claim, McClesky offered a statistical study (the Baldus study) to show disparities in the imposition of the death sentence in Georgia based on the murder victim's race and the defendant's race. The study was based on over 2,000 murder cases that occurred in Georgia during the 1970's, and involved data relating to the victim's race, the defendant's race, and the various combinations of such persons' races. The study found a consistent pattern of discrimination in the use of the death penalty against defendants who were charged with killing white victims compared to defendants who were charged with killing African American victims.

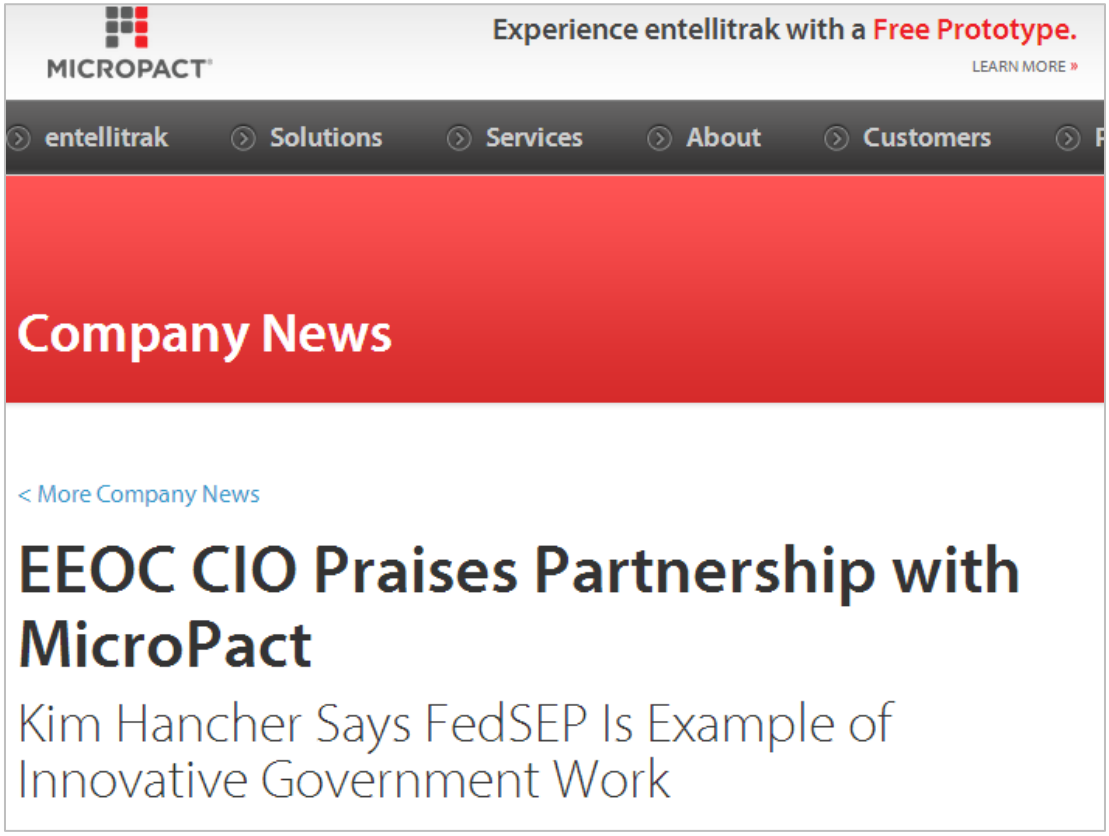
While the court ultimately found against McClesky, the case has been described as a turning point in the debate over the death penalty in the United States. The Baldus study has been replicated in numerous jurisdictions with similar findings. Race is now a powerful issue in debates over the death penalty because of studies like the Baldus study, which show that race can affect death penalty decisions.

Source: McClesky v. Kemp, 481 U.S. 279 (1987).

Case Study 7: Discrimination Complaint Data (EEOC)

The Equal Employment Opportunity Commission (EEOC) is responsible for enforcing federal laws that make it illegal to discriminate against a job applicant or an employee because of the person's race, color, religion, sex (including pregnancy), national origin, age (40 or older), disability, or genetic information.

In March 2013 the EEOC unveiled FedSEP, an electronic portal through which more than 325 federal agencies interact with the EEOC relative to their workforce and complaint data. This gateway provides each agency's staff with a single point of access to EEOC data collection systems and provides a new tool to collect and analyze government-agency data on workplace discrimination charges. By streamlining the submission process for so many documents and aggregating data from many sources, FedSEP allows EEO professionals greater ability to spot trends and uncover discrimination across the federal government.



Source: http://www.entellitrak.com/blog/detail/eeoc-cio-praises-partnership-with-micropact/

Case Study 8: *United States v. Sterling*

In 2006, Donald Sterling, his wife, and their family-trust real-estate company, *Beverly Hills Properties*, were accused of engaging in discriminatory practices in violation of the Fair Housing Act and Title VIII of the Civil Rights Act of 1968. The United States alleged, *inter alia*, that the defendants refused to rent portions of their 27-building development in the "Koreatown" neighborhood of Los Angeles to non-Koreans (*e.g., Hispanics and blacks)*. A study by Dr. Shelley Lapkoff used a database of tenant information released by the defendants to show that the number of Korean tenants across the 27 buildings had increased "significantly," from 64 percent to 83 percent, within the year following acquisition of those buildings by the defendants. Dr. Lapkoff's report also included an analysis of census data to determine whether the defendant's claim that the changing demographic distribution of Koreatown could explain the decreasing diversity of its tenants.

Dr. Lapkoff's analysis found that the overall demographics of Koreatown remained relatively stable during the period in question, and that Hispanics remained the dominant race in the area. The report concluded that, absent external changes, the increase in Korean tenants was consistent with the United States' allegation of housing discrimination. A later study by Dr. Lapkoff also used census data to show that there were no major shifts in household income or Korean households in the area that could explain the increase of Korean renters.

The studies helped lead to a settlement agreement that included a number of measures aimed at ending discriminatory renting practices in the 27 buildings owned by Beverly Hills Properties and required the defendants to pay $2,625,000 to be disbursed among aggrieved persons and a $100,000 civil penalty.

Table 3

Percentage of New Tenants Who Were Korean, Before and After Acquisition

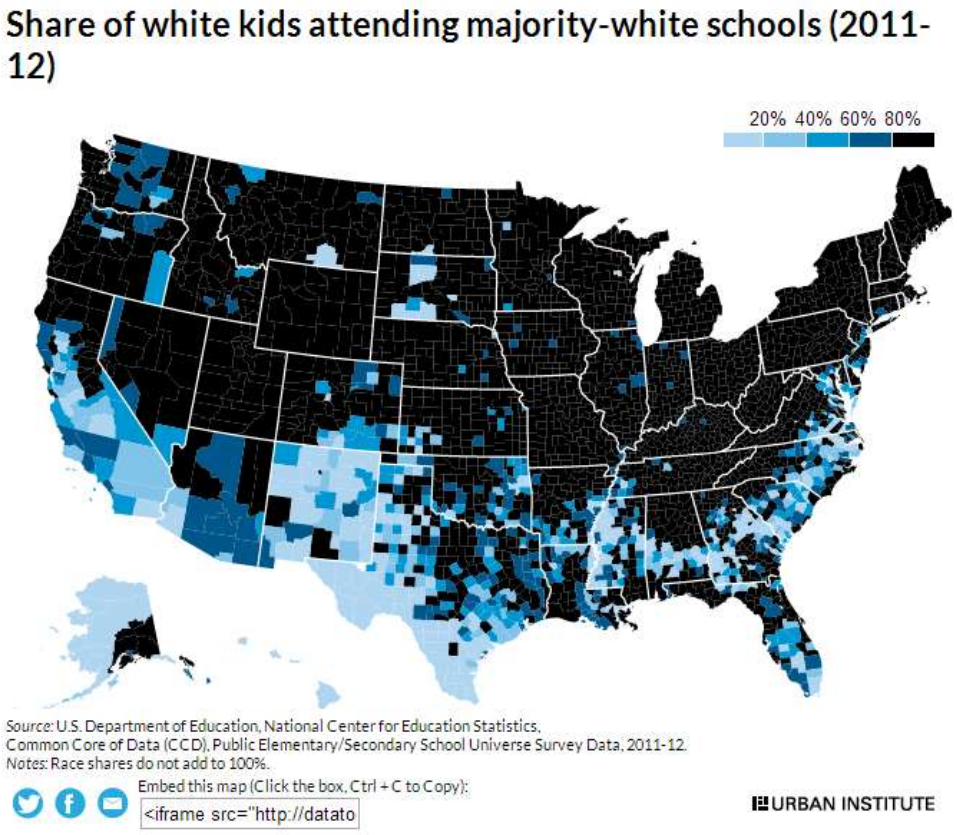
Building #	Year Before Acquisition Total	Korean	% Korean	Year After Acquisition Total	Korean	% Korean
Buildings Where Less than 80 percent of New Tenants are Korean at Time of Acquisition						
101	12	0	0%	16	4	25%
112	4	0	0%	5	5	100%
93	14	1	7%	18	13	72%
96	9	1	11%	11	9	82%
102	25	3	12%	22	8	36%
105	8	2	25%	7	5	71%
108	20	5	25%	30	28	93%
85	12	4	33%	14	13	93%
106	20	7	35%	18	11	61%
98	7	4	57%	11	11	100%
88	43	26	60%	41	34	83%
97	8	5	63%	10	8	80%
82	29	20	69%	27	25	93%
95	33	24	73%	44	38	86%
104	19	15	79%	28	25	89%
Subtotal	263	117	44%	303	238	79%
Buildings Where More than 90 percent of New Tenants are Korean at Time of Acquisition						
87	40	36	90%	45	43	96%
89	12	11	92%	13	13	100%
83	11	11	100%	9	9	100%
84	18	18	100%	17	14	82%
90	8	8	100%	19	17	89%
91	9	9	100%	4	4	100%
94	5	5	100%	10	10	100%
103	14	14	100%	18	16	89%
107	14	14	100%	10	9	90%
109	14	14	100%	20	18	90%
110	12	12	100%	7	6	86%
Total	157	152	97%	172	159	92%

Building # 86 Had No Rentals the Year Before Acquisition for Comparison

Source: United States v. Sterling, No. 2:06CV04885, (C.D. Cal. Nov. 12, 2009).

Case Study 9: Mapping Public School Segregation (Urban Institute)

Even as the country becomes more diverse – this year nonwhite students will account for the majority of public school students – black and Hispanic students often remain segregated from white students at historic levels. Drawing from the Department of Education's National Center for Education Statistics, the Urban Institute provides interactive county-level maps that track and visualize public-school segregation. The maps aggregate primary and secondary public-school enrollment by county and identify where white children predominantly attend majority-white schools and where minorities attend schools with predominantly minority classmates. The data is compiled using demographic information and a combination of five school surveys, covering the universe of all free public schools and school districts in the United States. It shows that despite the country's growing diversity, even extremely diverse regions of the country still have segregated school systems.



Sources: http://blog.metrotrends.org/2014/08/americas-public-schools-remain-highly-segregated/
http://www.vox.com/2014/8/19/6031279/majority-minority-public-schools
http://nces.ed.gov/ccd/pdf/psu12pgen.pdf

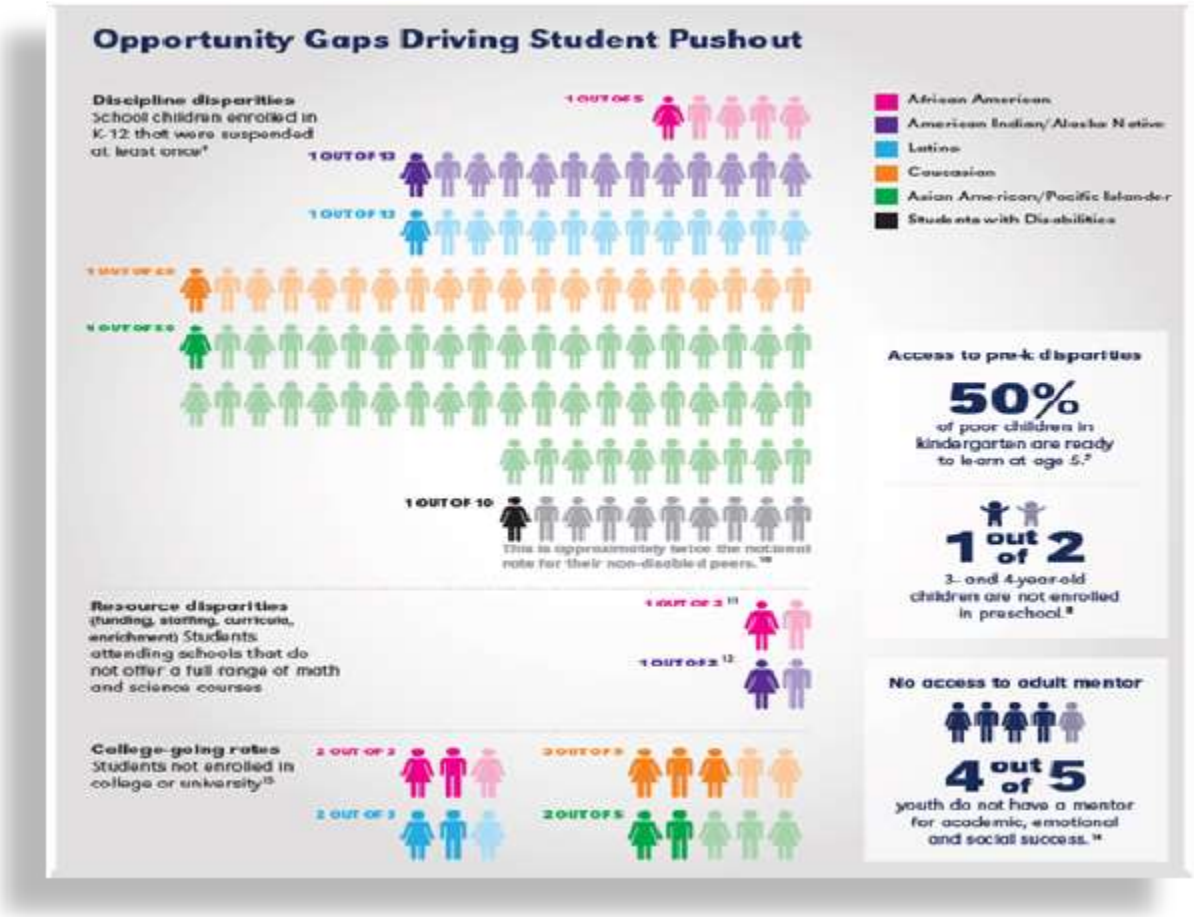
III. LONG-TERM PROBLEMS REQUIRE INNOVATIVE SOLUTIONS

Case Study 10: Education for All (NSBA)

A recent report by National School Boards Association (NSBA) offers novel policy solutions for increasing education rates in America. The report, *Partnerships, not Pushouts,* combines census data with data collected by various organizations to identify factors—known as "pushouts"—that may be responsible for driving young people away from education. Pushout factors can be more common among different segments of the population. For example, school suspensions—considered a major "pushout" factor, affect one out of five African American students and only one out twenty Caucasian students, which may partly explain the large discrepancy between graduation rates of those two groups.

To increase education levels among American youth, the NSBA proposes a variety student-centered "Personal Opportunity Plans" (POPs). To be effective, POPs are tailored to meet the needs of students on an individualized basis, addressing the pushout factor(s) most threatening to a particular student's academic success.

Discipline Disparities



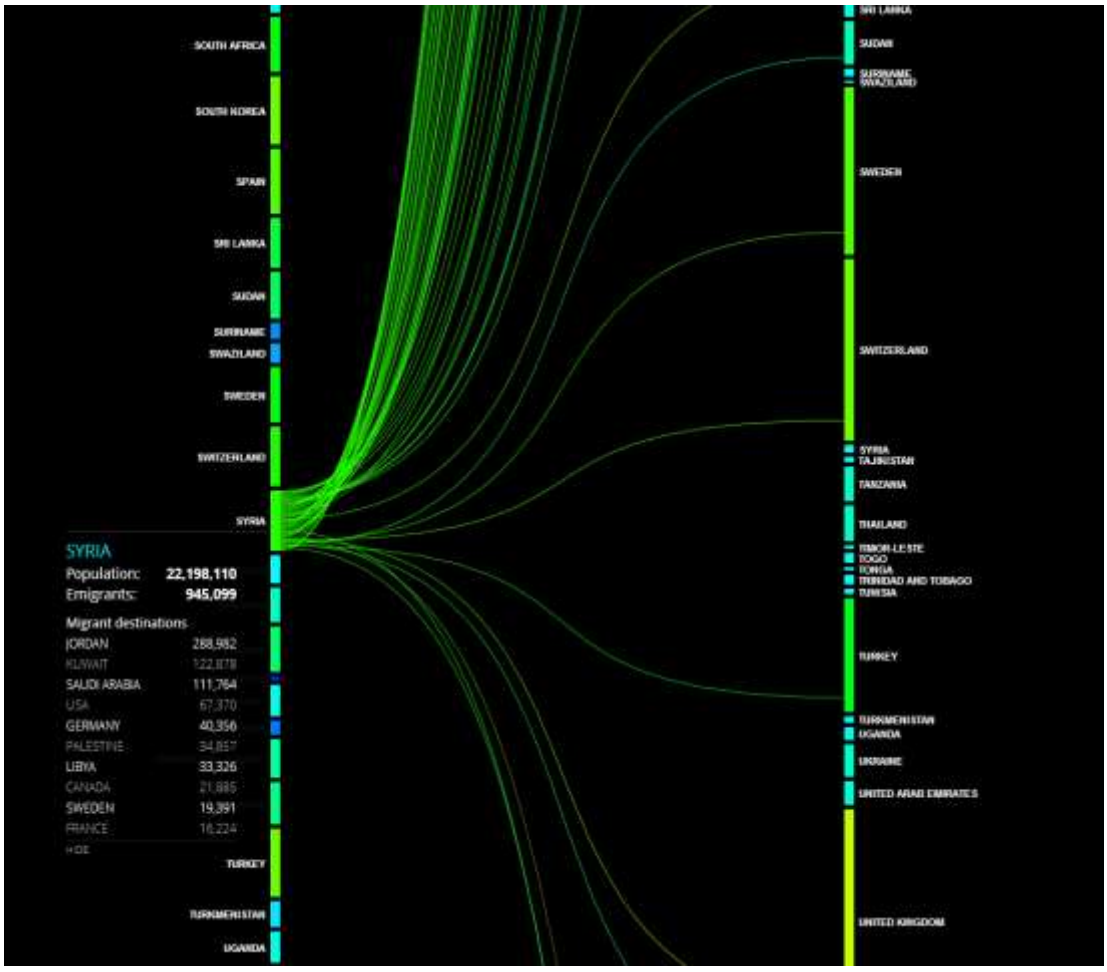
Source: http://www.nsba.org/sites/default/files/reports/Partnerships_Not_Pushouts_Guide.pdf

Case Study 11: Tracking Migratory Patterns (United Nations)

The United Nations has highlighted the social benefits of tracking migratory patterns of diverse peoples. For example, tracking the movement of displaced populations can empower humanitarian groups to provide better aid to those populations. As a new project under the *UN Global Pulse* banner, the organization is exploring new ways to track displaced populations using Big Data. The organization cites "significant shortcomings" with traditional methods of migratory benchmarking such as censuses, demographic and thematic surveys and administrative registers, which quickly become outdated.

In its review, Global Pulse highlights a number of studies that rely on Big Data collected from social media sites or open data initiatives to draw important conclusions about population movements. In the example below, PeopleMovin, repurposes "open" migration, refugee and asylum, and world population data to create an interactive tool allowing users to quickly identify international movement patterns and identify where relief efforts are most valuable.

Migration Patterns from Syria:



Source: http://peoplemov.in/#f_SY

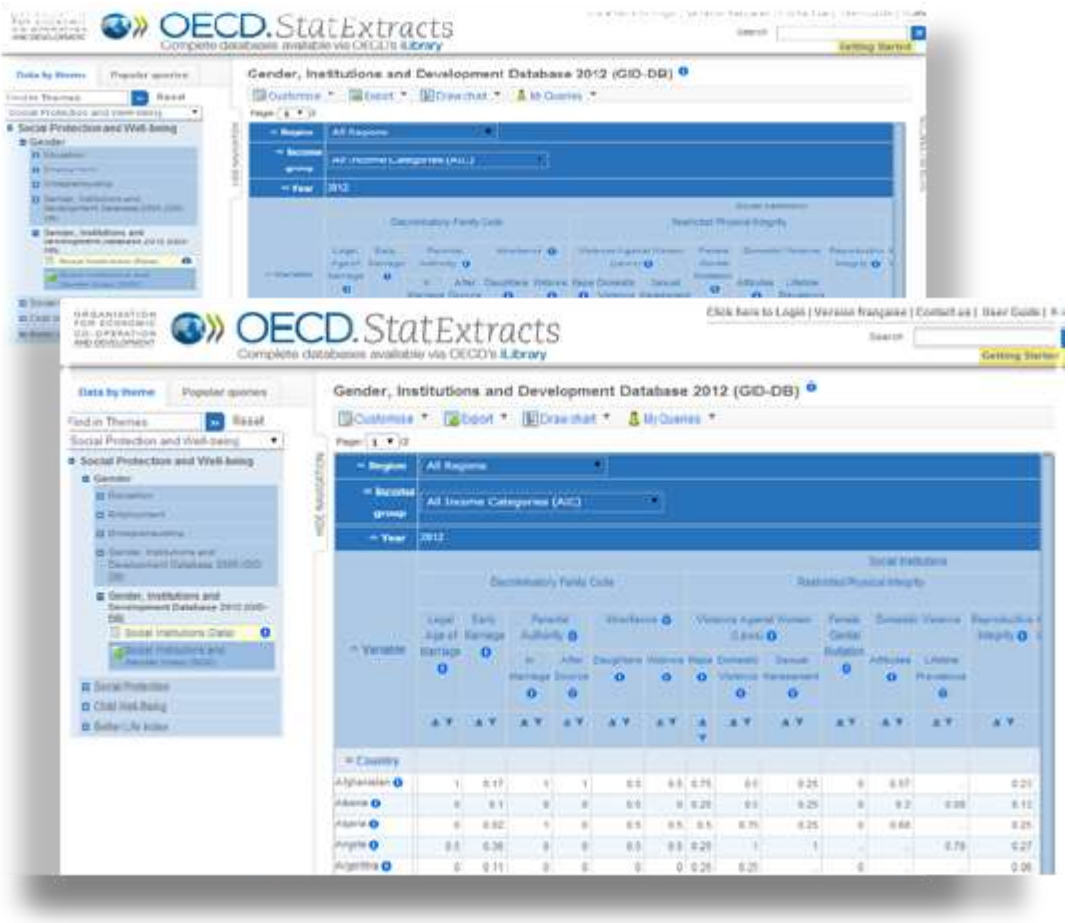
Case Study 12: Economic Development and Equality (OECD)

Since 2009, the Organization for Economic Co-Operation and Development has offered the publicly available *Gender, Institutions and Development Database.* The database compiles gender-discrimination data from 160 countries to provide researchers and policymakers with an analysis of 60 detailed variables, ranging from factors like "Discriminatory Family Code" to "Restricted Civil Liberties," that are likely to impact women's engagement in society and the economy.

A defining feature of the GID-DB is that, in addition to traditional quantitative analyses, the database uses an innovative scoring system to evaluate discriminatory institutional features. For example, while traditional studies of "early marriage" analyze rates of marriage among various age groups, the GID-DB has created a scaled system that combines rates of early marriage with an analysis of legal, traditional and religious customs to provide a much deeper look at gender discrimination. An example of the scaled system is provided below:

- 0: The law on the minimum age of marriage does not discriminate against women.
- 0.5: The law on the minimum age of marriage discriminates against some women, for example through customary, traditional and religious law.
- 1: The law on the minimum age of marriage discriminates against all women or there is no law on the minimum age of marriage.

Gender, Institutions and Development Database (GID-DB)



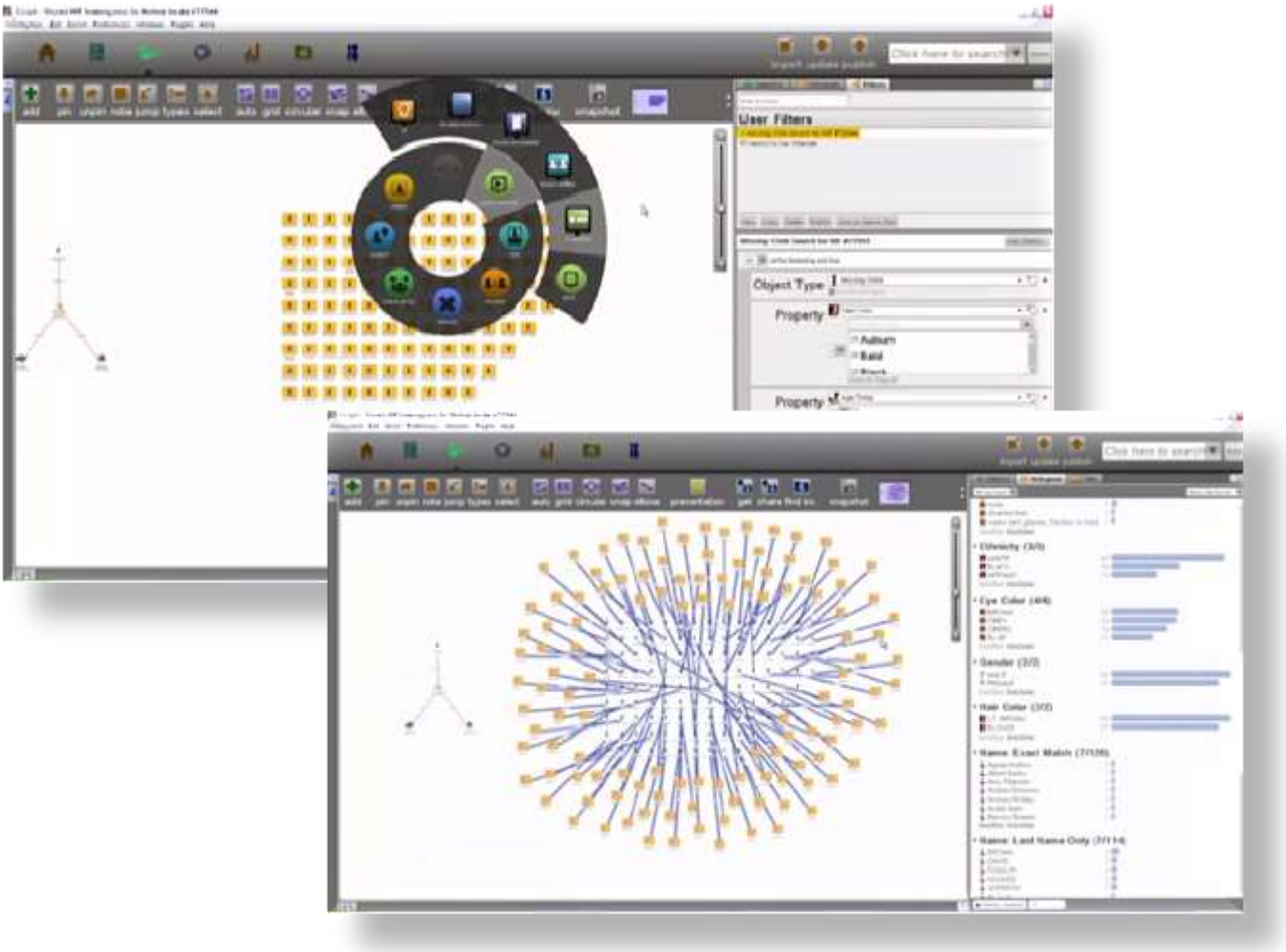
Source: http://stats.oecd.org/Index.aspx?DataSetCode=GIDDB2012

Case Study 13: Finding Missing and Exploited Children (Palantir/NCMEC)

The National Center for Missing & Exploited Children (NCMEC) compiles a wide variety of information from law enforcement, social media, and proprietary databases. Much of this information has traditionally been stored in siloed databases, requiring analysts to manually query each database when investigating a case. The Big Data analytics tool, developed in 2010 by Palantir, empowers NCMEC analysts to query a range of databases simultaneously.

The below case illustrates how the NCMEC uses Big Data to save children:

> A 17-year-old girl was reported missing and suspected of being a victim of sex trafficking. Through various searches, a NCMEC analyst was able to find multiple posts online that advertised this missing child for sex. Through information in the ads, the analyst was able to tie them to other posts from the same pimp. The analysis included over 50 advertisements, 9 different females, and a trail covering 5 states. A Link Analysis graph was created using Palantir that helped law enforcement to easily see the large scope of the ring. This insight helped law enforcement link the pimp to a multitude of other crimes and other girls that he victimized.



Sources: https://www.palantir.com/wp-assets/wp-content/uploads/2014/01/NCMEC-Impact-Study.pdf https://www.youtube.com/watch?v=TKpam_iy3Fo

Case Study 14: Human Trafficking (Palantir/Polaris Project)

Human Trafficking is a problem facing tens of millions of people and their families around the world. According to the *Polaris Project*, each year 21 million people are enslaved worldwide to generate a profit of $32 billion for their captors. To combat this global problem, organizations like Polaris Project maintain extensive databases of information collected from various public and private sources. The organization reports that it may collect up to 170 different quantitative and qualitative variables per case record, including first-hand data obtained through its National Human Trafficking Resource Center Hotline. In 2013, the NHTRC received 31,945 phone calls, 1,488 e-mails, 1,669 tips from online form submissions, and 787 SMS threads. In order to leverage this vast amount of data, the organization uses the Palantir Gotham analytics platform to track trafficking rings, quickly identify discrete human-trafficking events, and mobilize appropriate response units.

Sources: http://www.polarisproject.org/resources/hotline-statistics/human-trafficking-trends-in-the-united-states https://www.youtube.com/watch?v=kdQrLMEF-Eg#t=82


FPF
FUTURE OF
PRIVACY FORUM


ADL
Anti-Defamation League