Fairness in Machine Learning
(or Fairness by Design)

July 14, 2016

Welcome!

A bit about myself…

THE LATEST | FORD LIVE EVENTS

Fairness by Design

How Can Big Data Advance Opportunity for All?

Today, it's easier than ever before to make predictions and decisions by analyzing an unprecedented wealth of data. While corporate use of these "algorithmic decision-making tools" is widespread, government agencies at all levels are also using these techniques to guide decisions, increase efficiency, inform policy, and shape service delivery. Yet with these advances come risks: Without careful design, application, and oversight, these tools could be used to harm vulnerable populations and reinforce existing inequities.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

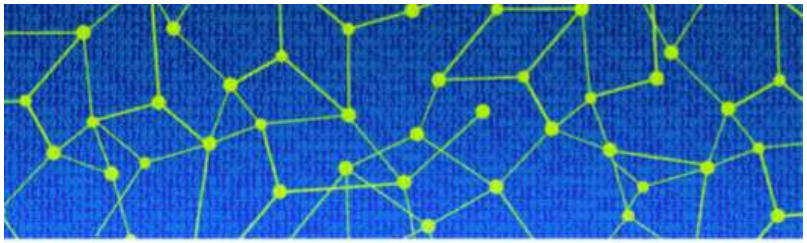
Executive Office of the President

May 2016

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016



EUROPEAN DATA PROTECTION SUPERVISOR

Opinion 7/2015

Meeting the challenges of big data

A call for transparency, user control, data protection by design and accountability

EDPS

19 November 2015

“It’s Discrimination, Stupid” (Gandy 1995)

Differential Treatment

- Opportunities
- Access
- Eligibility
  - Terms
- Price
  - Inducements and penalties
- Attention
  - Scrutiny
- Exposure

The Machine Learning *Process*

“Target Variables”

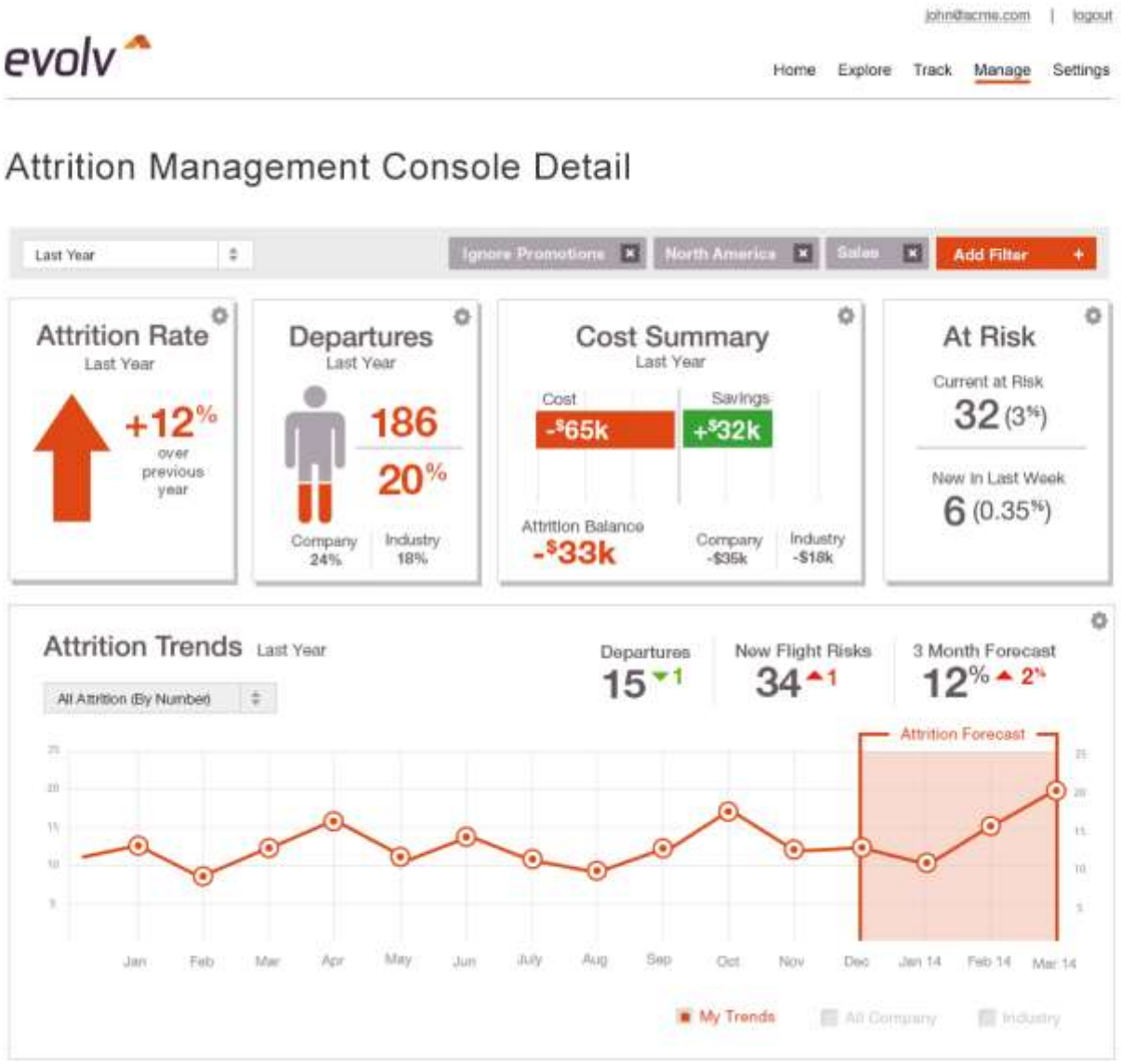
The “Art” of Data Mining

- Data miners must determine how to solve the problem at hand by translating it into a question about the value of some “target variable”
- The proper specification of the target variable is frequently not obvious, and it is the data miner’s task to define it
  - Construct validity
  - Type III error

Upfront Interpretative Work

- The definition of the target variable and its associated class labels will determine what data mining happens to find
  - Good employee
  - Ideal Customer
  - Relevance
- Can be far more consequential than the “accuracy” of these inferences or the details of the “formula”

Good Employee



The Ideal Customer

- The person who is most likely to click on an ad
- The person who is most likely to make a purchase
- The person who will establish a long-term relationship with the company
- The person who will generate the most profit for the company

Relevance

Google’s PageRank

Facebook’s EdgeRank

Forsake Formalization

- These moments of translation are opportunities to debate the very nature of the problem—and to be creative in parsing it
- The process of formalization *can* make explicit the beliefs, values, and goals that motivate a project

“Training Data”

Data mining is really a way to learn by example

The data that function as examples are known as “training data”—quite literally the data that train the model to behave in a certain way

2.A.

Skewed set of examples

(problems with data collection)

2.B.

Setting a bad example

(problems with the labeling of examples)

Uncounted, Unaccounted, Discounted

- The quality and representativeness of records might vary in ways that correlate with class membership
  - less involved in the formal economy and its data-generating activities
  - unequal access to and less fluency in the technology necessary to engage online
  - less profitable customers or less important constituents and therefore less interesting as targets of observation
- Convenience Sample
  - Data gathered for routine business purposes tend to lack the rigor of social scientific data collection

A Skewed Set of Examples: StreetBump

Crawford, K., 2013. The Hidden Biases in Big Data. *Harvard Business Review*.

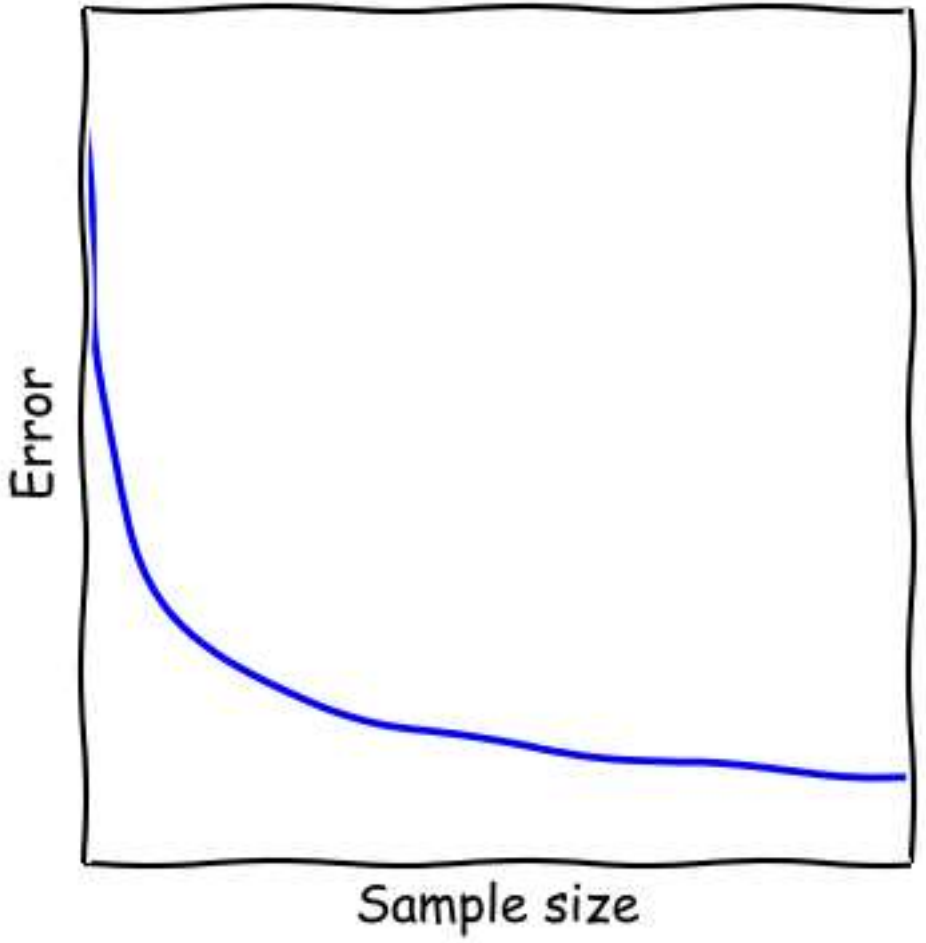
Skewing the Sample Frame and Limiting Future Contact

- These results may lead to decision procedures that limit the future contact an organization will have with specific groups, skewing still further the sample upon which subsequent analyses will be performed
  - Limiting contact with specific populations may deny members of those populations the opportunity to prove that they buck the apparent trend
- Over-representation can have similar effects

Correcting for Bias

- Unfortunately, the under- and over-representation of members of protected classes is not always evident
- The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions

Sample Size Disparity




NETFLIX

Labeling Examples

- Sometimes a rather straightforward affair
  - e.g., spam/not spam
- Sometimes a laborious process that is fraught with peril
  - e.g., default
  - e.g., good job candidate

Applying the Label

- Even where the class labels are uncontested or uncontroversial, they may present a problem because analysts will often face difficult choices in deciding which of the available labels best applies to a particular example.
  - Certain cases may present some, but not all, criteria for inclusion in a particular class.
  - At other times, the class labels may be insufficiently precise to capture meaningful differences between cases.

Bad Examples: Reproducing Past Prejudice

- So long as prior decisions affected by some form of prejudice or bias serve as examples of *correctly* rendered determinations, data mining will induce rules that exhibit the same prejudice

HOWARD
UNIVERSITY

WELLESLEY

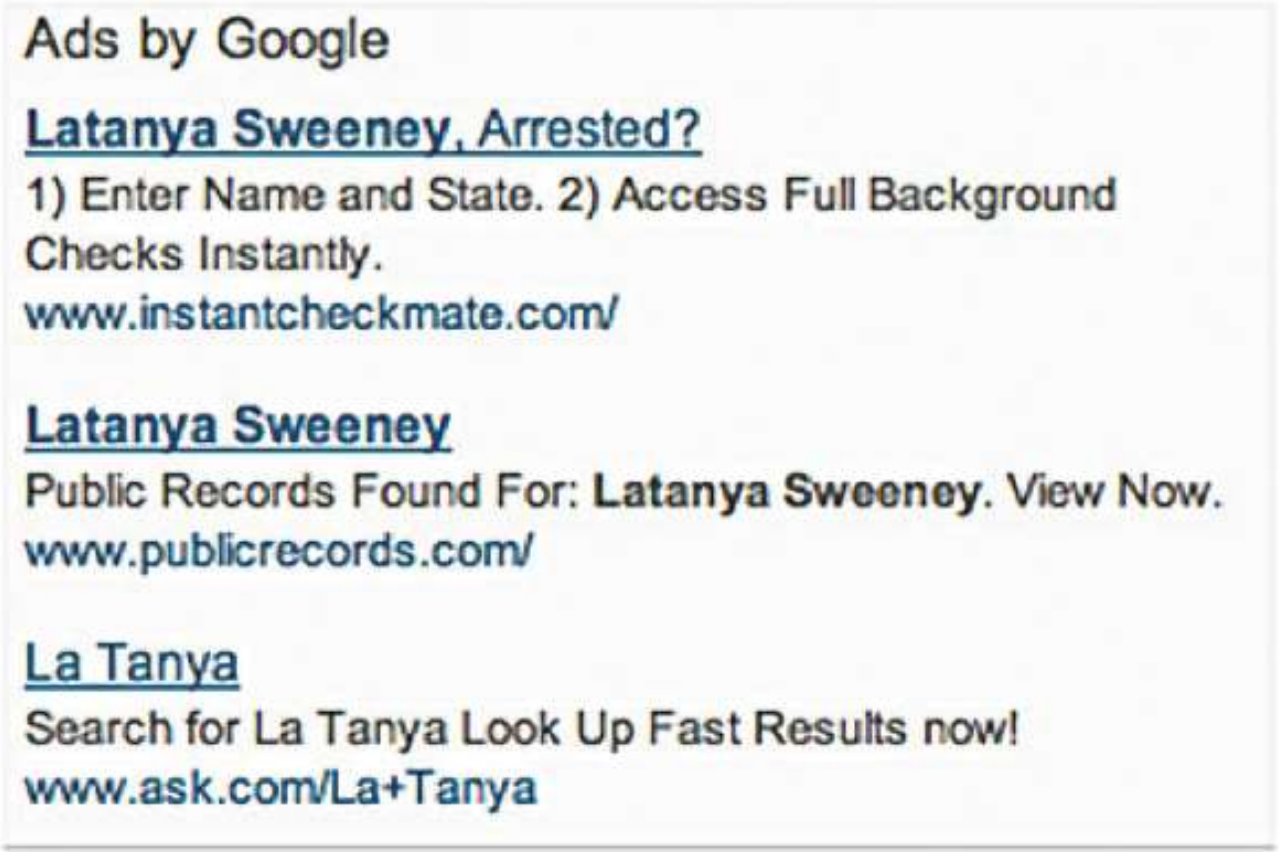
Identifying and Ridding the Taint

- Training data serve as ground truth
  - These would seem like well performing models according to standard evaluation methods
- What the objective assessment *should* have been
  - Accepted and rejected candidates may not differ only in terms of protected characteristics
- How someone *would* have performed under different, non-discriminatory circumstances
  - The difficulty in dealing with counterfactuals and correcting for past injustices

Reflect Current Prejudice

- Not only can data mining inherit *prior* prejudice, it can also reflect *current* prejudice
  - In catering to the demonstrated preferences of users, companies may unintentionally adopt the prejudices that guide users' behavior

Setting a Bad Example: Instant Checkmate



Sweeney, L., 2013. Discrimination in Online Ad Delivery. *Communications of the ACM*, 56(5), pp.44–54.

Correcting for Prejudice

- Platforms may only be able to purge the effects of prejudice from the data upon which their decisions depend if they commit to ensuring a minimum discrepancy in the impact those decisions have on different social groups
  - Ex-ante corrections (ridding taint) versus ex-post outcome-based test and active rebalancing

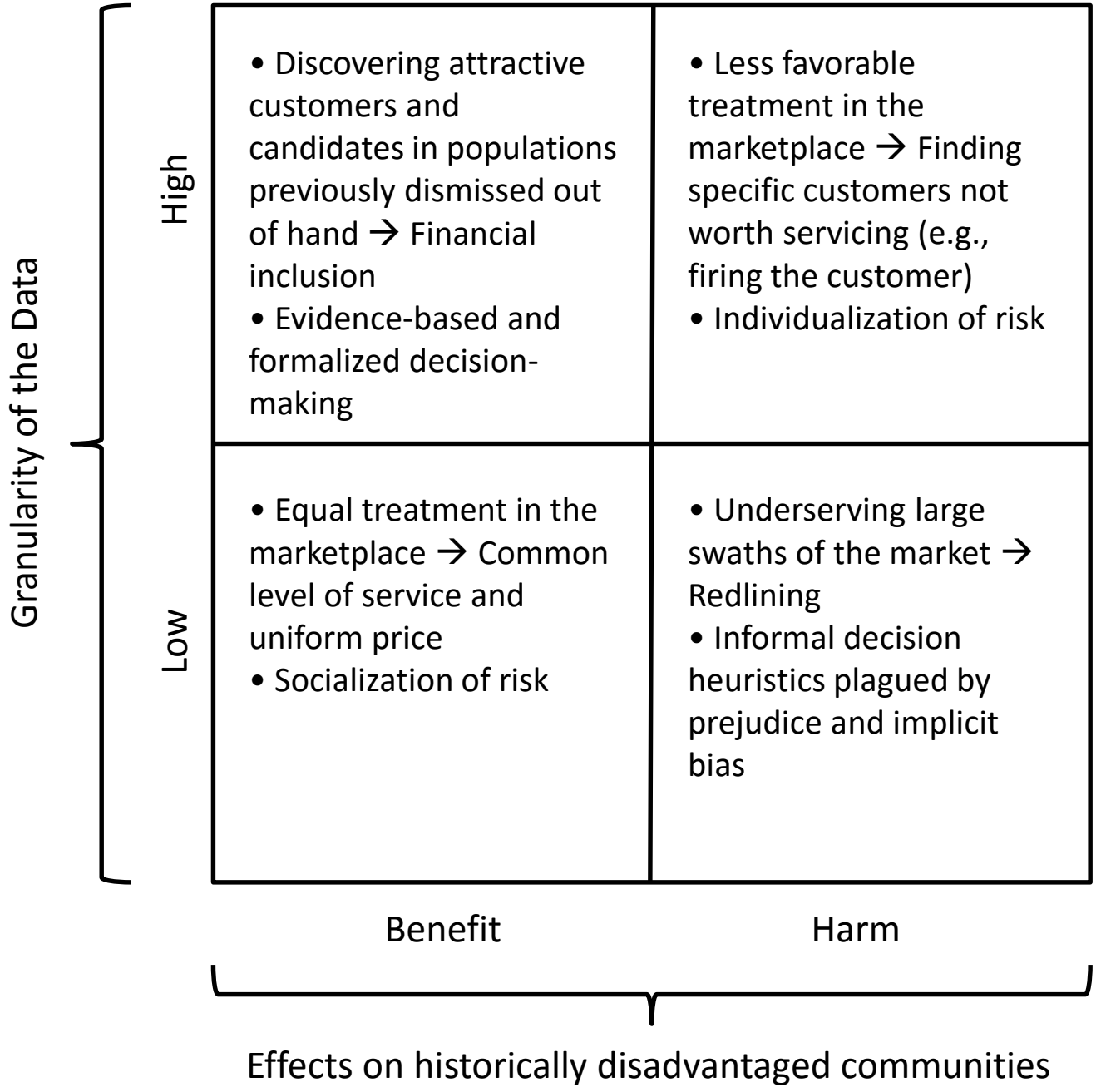
“Feature Selection”

Carving Up the Population

- Does the feature set provide sufficient detail to carve-up the population in a way that reveals relevant variations within each apparent sub-groups?
  - e.g., redlining
- In other words: how does the error rate vary across the population?
  - Discrimination can be an artifact of statistical reasoning rather than prejudice on the part of decision-makers or bias in the composition of the dataset
- …But there's no such thing as perfectly individualized decisions
  - e.g., insurance

At What Cost?

- Obtaining information that is sufficiently rich to draw precise distinctions can be difficult, expensive, or objectionable
  - Does this justify subjecting historically marginalized groups to erroneous decisions at higher rates?
- Any rigorous defense of data mining must justify ignoring the history that accounts for the higher costs involved in improving the accuracy of determinations for the least well-off

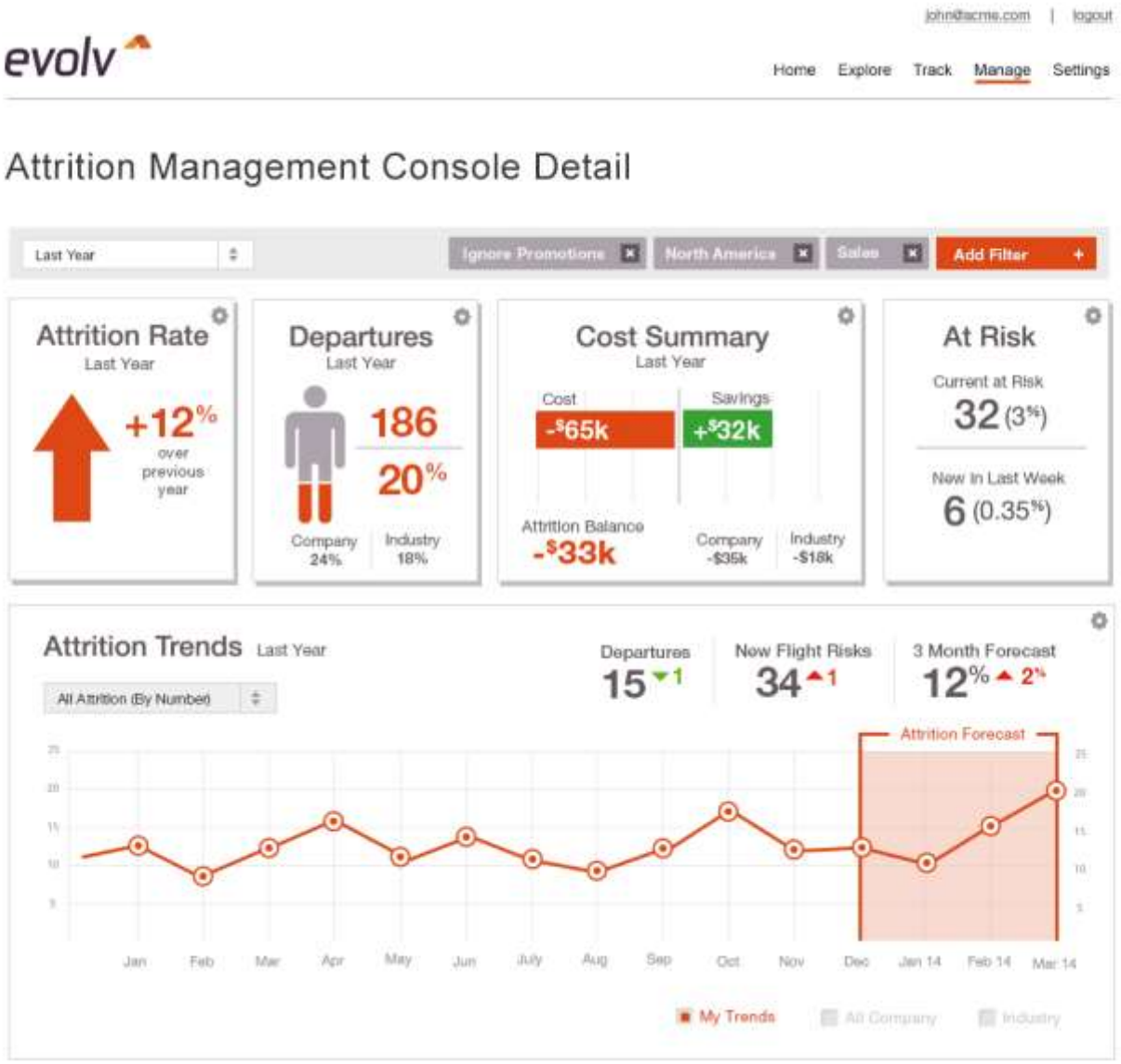

Granularity of the Data
High
Low
• Discovering attractive customers and candidates in populations previously dismissed out of hand → Financial inclusion
• Evidence-based and formalized decision-making
• Less favorable treatment in the marketplace → Finding specific customers not worth servicing (e.g., firing the customer)
• Individualization of risk
• Equal treatment in the marketplace → Common level of service and uniform price
• Socialization of risk
• Underserving large swaths of the market → Redlining
• Informal decision heuristics plagued by prejudice and implicit bias
Benefit
Harm
Effects on historically disadvantaged communities

“Proxies”

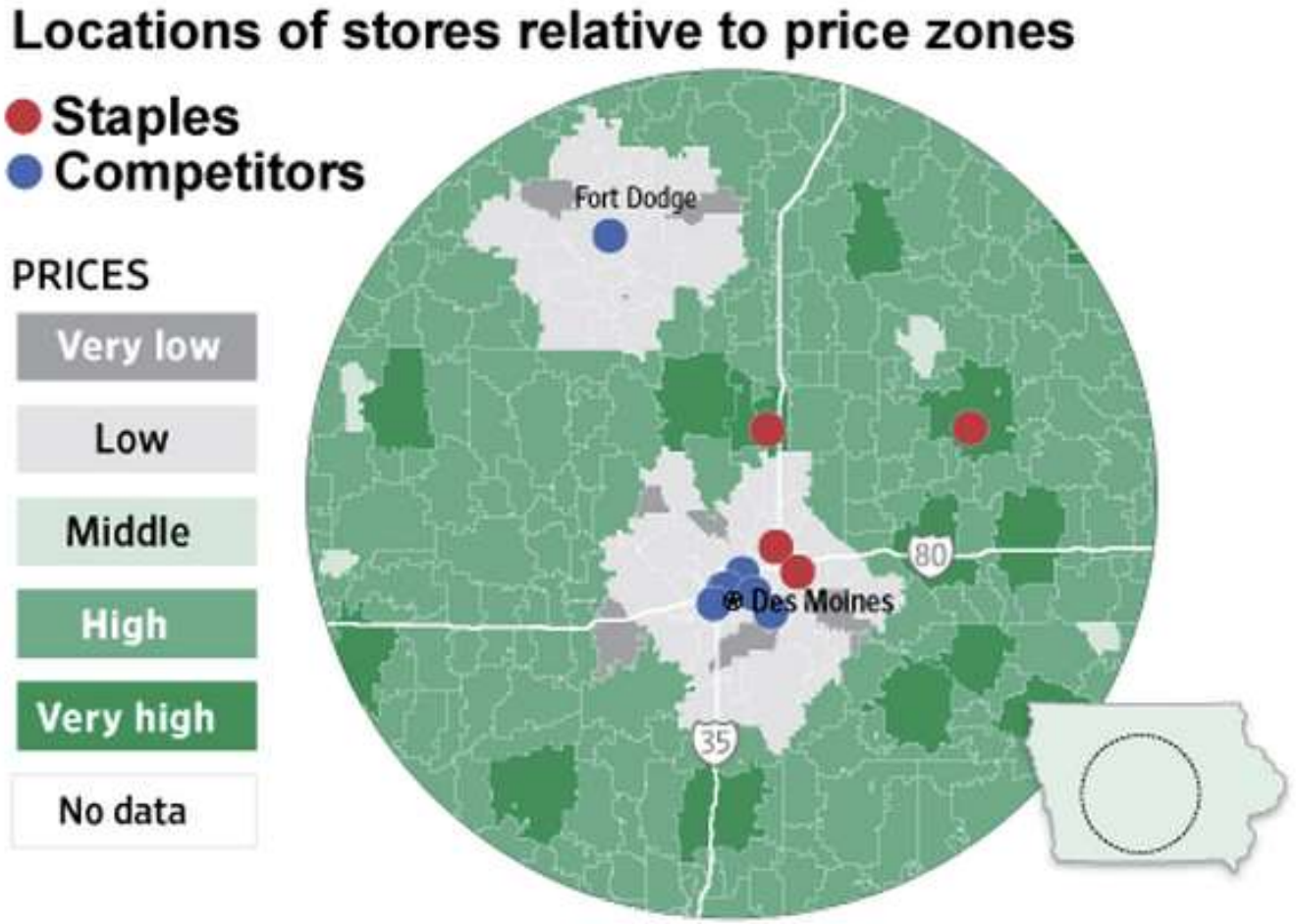
Dealing with “Redundant Encodings”

- In many instances, making accurate determinations will mean considering factors that are somehow correlated with legally proscribed features
  - There is no obvious way to determine how correlated a relevant attribute or set of attributes must be with proscribed features to be worrisome
  - Nor is there a self-evident way to determine when an attribute or set of attributes is sufficiently relevant to justify its consideration, despite the fact that it is highly correlated with these features

Good Employee



“Websites Vary Prices, Deals Based on Users' Information”



Recapitulating Inequality

- This reflects the simple fact of inequality, but an inequality that is not random
  - Members of protected classes are frequently the groups in the position of relative disadvantage
- Better data will more precisely map the contours of inequality
  - At some point, the data will be sufficiently rich that it will be irrelevant whether class membership is considered explicitly
- …and attempts to ensure procedural fairness will be in conflict with the imperative to ensure accurate determinations

Fairness/Accuracy Trade-Off

- Computer scientists have demonstrated—unwittingly—that the tidy distinction between equality of opportunity and equality of outcome may be impossible to uphold in practice
  - Dwork et al. “demonstrate a quantitative trade-off between fairness and utility”
- The pressing challenge does not lie with ensuring procedural fairness through a more thorough stamping out of prejudice and bias, but rather with developing ways of reasoning that can help adjudicate when and what amount of disparate impact is tolerable
  - Dwork et al.’s “Fairness through Awareness”

“Masking”

Masking

- Analytics could also breathe new life into traditional forms of intentional discrimination because decision-makers with prejudicial views can mask their intentions by exploiting each of the mechanisms enumerated above
  - Knowingly bias the collection of data to ensure that analytics suggests rules that are less favorable to members of protected classes
  - attempt to preserve the known effects of prejudice in prior decision-making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule
  - intentionally rely on features that only permit coarse-grain distinction-making—distinctions that result in avoidable and higher rates of erroneous determinations for members of a protected class
- Of course, also possible to simply infer class membership

SEP 17 2014

Using publicly available information to proxy for unidentified race and ethnicity

We ensure that lenders are complying with fair lending laws and are addressing discrimination across the consumer credit industry. Information on consumer race and ethnicity is required to conduct fair lending analysis of non-mortgage credit products, but auto lenders and other non-mortgage lenders are generally not allowed to collect consumers' demographic information. As a result, substitute, or "proxy" information is used to fill in information about consumers' demographic characteristics.

In this paper, we explain the construction of the proxy for race and ethnicity currently employed by our Office of Research and division of Supervision, Enforcement, and Fair Lending. This report also provides an assessment of the performance of the proxy method using a sample of mortgage applicants for whom race and ethnicity are reported.

Take a look at our methodology and assessment.

You can also take a look at the statistical software code and publicly available census data used to build the proxy.

How Data Mining Discriminates

- Target Variable
- Training data
  - Skewed samples
  - Tainted examples
- Feature selection
  - Limited and coarse features
- Proxies
- Masking

…but there’s more

Taking Unfair Advantage

- So-called “persuasion profiling”
- Objectionable because it allows firms to prey on the vulnerable or desperate
  - Intentionally and blatantly
  - Intentionally and undetectably
  - Unintentionally and blatantly
  - Unintentionally and undetectably

A Taxonomy of Concerns

1. Concerns with prejudicial decision-making and its masking
2. Concerns with bias and error—and the distribution of those errors across different social groups
3. Concerns with threats to solidarity and the perpetuation of inequality even in the absence of prejudice, bias, and error
4. Concerns with undue sway and bargaining power

Break

Old York University

- In an effort to curb crime, Old York has decided that it will take a more data-driven approach to policing. In particular, the police would like to predict the location of future criminal activity and have begun to train a model on 10 years of arrest records.
- Armed with the model, the police then plan to deploy a greater proportion of its officers to those specific areas that are predicted to experience relatively higher rates of crime.

Old York University

- First, the university will use information on its prior students' performance and background to determine what kind of high school students are likely to be qualified and interested candidates. It will then focus its recruitment efforts exclusively on these students.
- Second, the university will train a model on its prior students' transcripts to predict its currents students' interest and future success in different classes and majors. Advisers will then steer students toward the suggested course of study.