# How Machines Learn to Discriminate

Solon Barocas

Microsoft Research

# Discrimination Law: Two Doctrines

| Disparate Treatment | Disparate Impact |
|---|---|
| Formal | Unjustified |
| Intentional | Avoidable |

"Protected Class"

1867
HOWARD
UNIVERSITY

WELLESLEY
W

# Dealing with Tainted Examples

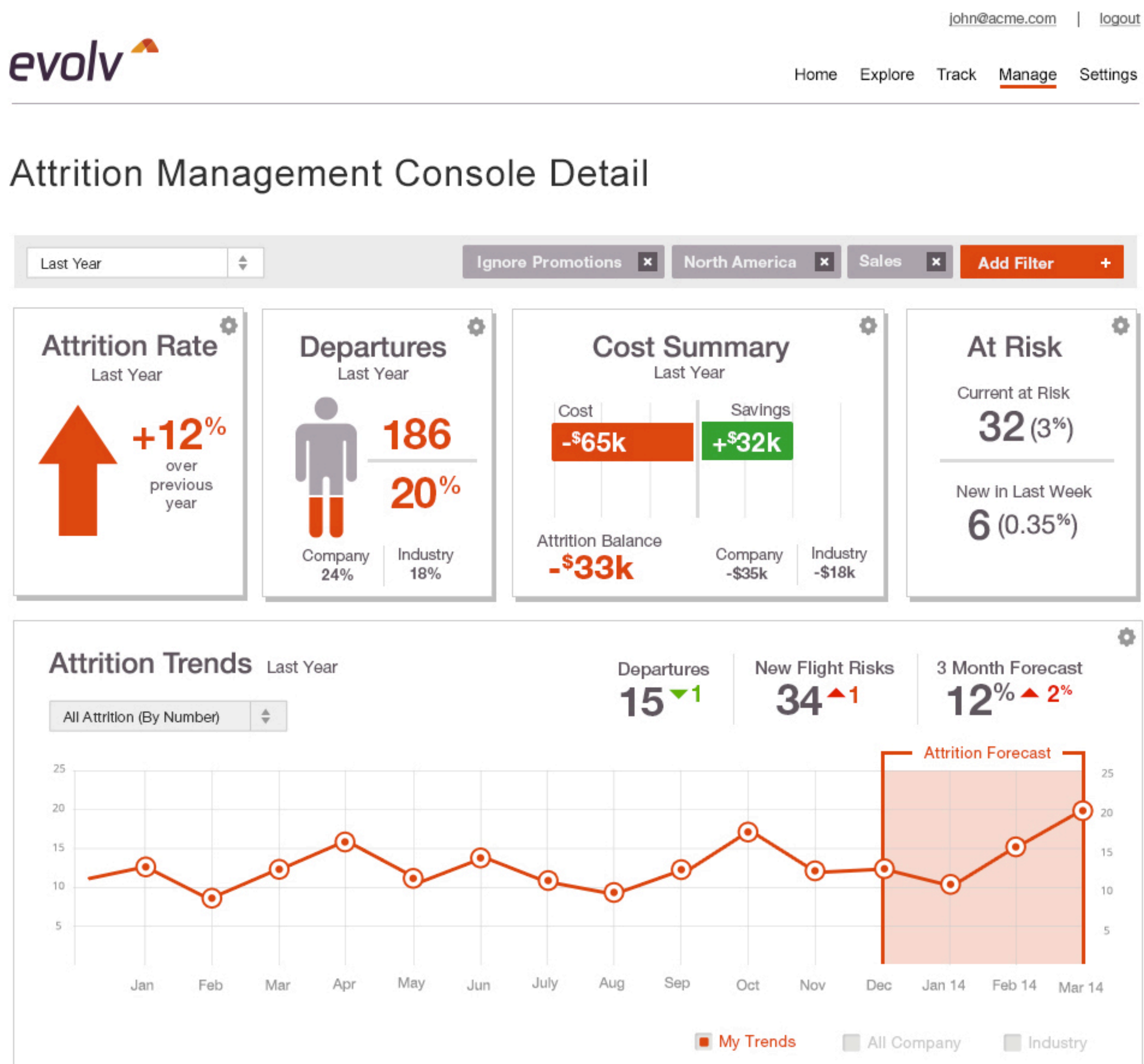
- Training data serve as ground truth
  - These would seem like well performing models according to standard evaluation methods
- What the objective assessment *should* have been
  - Accepted and rejected candidates may not differ only in terms of protected characteristics
- How someone *would* have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

# Settling on a Selection of Features

- Does the feature set provide sufficient information to carve-up the population in a way that reveals relevant variations within each apparent sub-group?
  - *Unintentional* 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
- Does the difficulty or cost involved in obtaining the information necessary to bring accuracy rates into closer parity justify subjecting certain populations to worse assessment?
  - Parity = Fair
  - Accurate = Fair

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

Effects on historically disadvantaged communities


john@acme.com | logout
evolv
Home
Explore
Track
Manage
Settings
Attrition Management Console Detail
Last Year
Ignore Promotions
North America
Sales
Add Filter +
Attrition Rate
Last Year
+12%
over previous year
Departures
Last Year
186
20%
Company 24%
Industry 18%
Cost Summary
Last Year
Cost
Savings
-$65k
+$32k
Attrition Balance
-$33k
Company -$35k
Industry -$18k
At Risk
Current at Risk
32 (3%)
New in Last Week
6 (0.35%)
Attrition Trends
Last Year
All Attrition (By Number)
Departures
15 1
New Flight Risks
34 1
3 Month Forecast
12% 2%
Attrition Forecast
25
20
15
10
5
Jan
Feb
Mar
Apr
May
Jun
July
Aug
Sep
Oct
Nov
Dec
Jan 14
Feb 14
Mar 14
My Trends
All Company
Industry

# 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

# Let’s not 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

Solon Barocas and Andrew Selbst,
"Big Data's Disparate Impact," *California Law Review*, Vol. 104, 2016

Solon Barocas
Microsoft Research
solon@microsoft.com