# Algorithmic Justice for Development:

## Using Machine Learning to Identify and Mitigate Bias in Indian Courts

Elliott Ash (ETH Zurich), Sam Asher (John Hopkins),
Sandeep Bhupatiraju (World Bank), **Daniel L. Chen** (Toulouse),
Tanaya Devi (Harvard), Paul Novosad (Dartmouth),
Arianna Ornaghi (Warwick), Bilal Siddiqi (Berkeley)

MIT GOV/LAB
November 2019

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)
- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers
- Cycle of uneven/uncertain justice → distrust → lack of reliance → lower investment → economic inefficiency
- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)
- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers
- Cycle of uneven/uncertain justice → distrust → lack of reliance → lower investment → economic inefficiency
- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)
- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers
- Cycle of uneven/uncertain justice $\rightarrow$ distrust $\rightarrow$ lack of reliance $\rightarrow$ lower investment $\rightarrow$ economic inefficiency
- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)
- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers
- Cycle of uneven/uncertain justice $\rightarrow$ distrust $\rightarrow$ lack of reliance $\rightarrow$ lower investment $\rightarrow$ economic inefficiency
- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)
- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers
- Cycle of uneven/uncertain justice $\rightarrow$ distrust $\rightarrow$ lack of reliance $\rightarrow$ lower investment $\rightarrow$ economic inefficiency
- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# Motivation

- Strong institutions encourage investment and growth (e.g., Rodrik 2000; Pande and Udry 2006)

- Courts in developing countries face numerous challenges to providing efficient and fair justice to citizens and firms (e.g. Djankov et al., 2003; La Porta et al., 2008).
  - transplanted legal codes
    - preferences for informal mechanisms
  - low infrastructure in court system
    - low-quality representation
    - corruption
    - **implicit or explicit bias** among judicial officers

- Cycle of uneven/uncertain justice → distrust → lack of reliance → lower investment → economic inefficiency

- Scarce empirical research on courts in developing countries (e.g. Ponticelli and Alencar, 2016).

# New Opportunities

1. Court rulings and judge biographies are increasingly digitized, allowing the construction of large-scale datasets.
2. Natural language processing (NLP) tools can produce interpretable data from unstructured text – including written judicial opinions.
3. ML can predict judge decision-making and uncover bias.

# Data

- A new database on the universe of judicial proceedings (70 million hearings, 14 million cases, and 10 million written judicial decisions)
- Supreme Court of India, 24 High Courts, 3,000+ subordinate courts.
  - World's largest democracy and largest common law legal system

# Data

- A new database on the universe of judicial proceedings (70 million hearings, 14 million cases, and 10 million written judicial decisions)
- Supreme Court of India, 24 High Courts, 3,000+ subordinate courts.
  - World's largest democracy and largest common law legal system

# Data

- A new database on the universe of judicial proceedings (70 million hearings, 14 million cases, and 10 million written judicial decisions)
- Supreme Court of India, 24 High Courts, 3,000+ subordinate courts.
  - World's largest democracy and largest common law legal system

# Project #1

- An empirical analysis of biased justice due to social (dis)advantage.
- Disparities in:
  - judicial representation
  - judicial treatment
  - judicial outcomes
- By group membership:
  - male vs female
  - hindu vs muslim
  - upper-caste vs lower-caste
- Data explorer
- Three policy issues
  - Court congestion
  - Environment
  - Network analysis of lawyers and judges

# Project #1

- An empirical analysis of biased justice due to social (dis)advantage.
- Disparities in:
  - judicial representation
  - judicial treatment
  - judicial outcomes
- By group membership:
  - male vs female
  - hindu vs muslim
  - upper-caste vs lower-caste
- Data explorer
- Three policy issues
  - Court congestion
  - Environment
  - Network analysis of lawyers and judges

# Project #1

- An empirical analysis of biased justice due to social (dis)advantage.
- Disparities in:
  - judicial representation
  - judicial treatment
  - judicial outcomes
- By group membership:
  - male vs female
  - hindu vs muslim
  - upper-caste vs lower-caste
- Data explorer
- Three policy issues
  - Court congestion
  - Environment
  - Network analysis of lawyers and judges

# Measuring Stereotypes in Judicial Language

# Lexical slant

- Google translate
  - "he/she is a doctor" (turkish) -> "he is a doctor" (english)
  - "he/she is a nurse" (turkish) -> "she is a nurse" (english)
- A truck driver should plan his route carefully.
- A truck driver should plan the travel route carefully.

# Lexical slant

- Google translate
  - “he/she is a doctor” (turkish) -> “he is a doctor” (english)
  - “he/she is a nurse” (turkish) -> “she is a nurse” (english)
- A truck driver should plan his route carefully.
- A truck driver should plan the travel route carefully.

# Lexical slant

- Google translate
  - “he/she is a doctor” (turkish) -> “he is a doctor” (english)
  - “he/she is a nurse” (turkish) -> “she is a nurse” (english)
- A truck driver should plan his route carefully.
- A truck driver should plan the travel route carefully.

# Lexical slant

- Google translate
  - “he/she is a doctor” (turkish) -> “he is a doctor” (english)
  - “he/she is a nurse” (turkish) -> “she is a nurse” (english)
- A truck driver should plan his route carefully.
- A truck driver should plan the travel route carefully.

# "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect real-world decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to disparate treatment?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does training affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

# "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect real-world decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to disparate treatment?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does training affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

## "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect real-world decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to disparate treatment?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does training affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

## "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect real-world decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to disparate treatment?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does training affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

# "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect real-world decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to disparate treatment?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does training affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

## "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist in the **wild**?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect **judicial** decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to **disparate treatment of other judges**?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does **diversity** affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

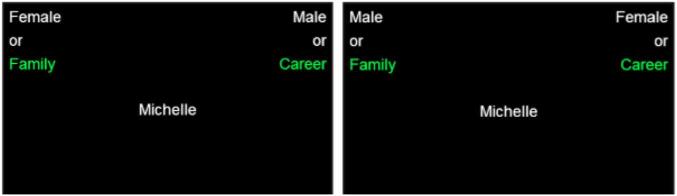
- Does implicit bias exist in the **wild**?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect **judicial** decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to **disparate treatment of other judges**?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does **diversity** affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

## "Attitudes that affect our understanding, actions, and decisions in an unconscious manner" Implicit bias (Kirnan institute OSU)

- Does implicit bias exist in the **wild**?
  - Ottaway et al. 2001, Rothermund et al. 2004, Arkes et al. 2004, Blanton et al. 2006
- Does it affect **judicial** decisions?
  - police (Correll et al. 2002); physicians (Green et al. 2007); resume screening (Bertrand et al. 2005)
- Does it lead to **disparate treatment of other judges**?
  - patients' feelings (Penner et al. 2010); grocery cashiers (Glover et al. 2017); students (Carlana 2018)
- Does **diversity** affect implicit attitudes?
  - exposure to female leaders (Beaman et al. 2009)

# Implicit attitudes

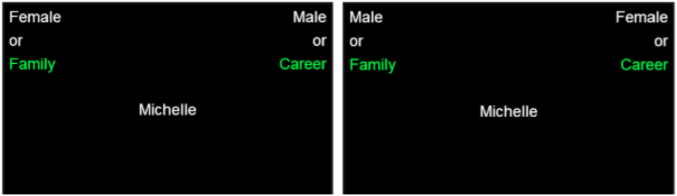
- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)




- Comparing reaction times across trials with different pairings
  - subjects are faster and make fewer errors on stereotype-consistent trials
  - difference yields "IAT score"

# Implicit attitudes

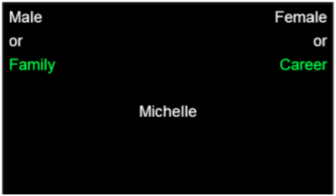
- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)




- Comparing reaction times across trials with different pairings
  - subjects are faster and make fewer errors on stereotype-consistent trials
  - difference yields "IAT score"

# Implicit attitudes

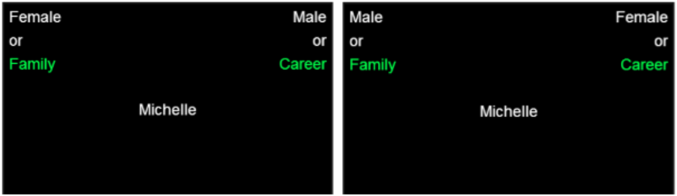
- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)


- Comparing reaction times across trials with different pairings
  - subjects are faster and make fewer errors on stereotype-consistent trials
  - difference yields "IAT score"

# Implicit attitudes

- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)




- Comparing reaction times across trials with different pairings
  - subjects are faster and make fewer errors on stereotype-consistent trials
  - difference yields "IAT score"

# Challenges of studying implicit attitudes

- Challenge: how can we measure implicit attitudes for the judiciary?
  - But we cannot elicit IAT scores from sitting judges (yet :-) )
- Proposed solution: proxy for IAT using large amounts of written text
  - Represent judicial language in vector space
  - Are words representing different groups associated to certain attributes?

# Challenges of studying implicit attitudes

- Challenge: how can we measure implicit attitudes for the judiciary?
  - But we cannot elicit IAT scores from sitting judges (yet :-) )
- Proposed solution: proxy for IAT using large amounts of written text
  - Represent judicial language in vector space
  - Are words representing different groups associated to certain attributes?

# Challenges of studying implicit attitudes

- Challenge: how can we measure implicit attitudes for the judiciary?
  - But we cannot elicit IAT scores from sitting judges (yet :-) )
- Proposed solution: proxy for IAT using large amounts of written text
  - Represent judicial language in vector space
  - Are words representing different groups associated to certain attributes?

# Challenges of studying implicit attitudes

- Challenge: how can we measure implicit attitudes for the judiciary?
  - But we cannot elicit IAT scores from sitting judges (yet :-) )
- Proposed solution: proxy for IAT using large amounts of written text
  - Represent judicial language in vector space
  - Are words representing different groups associated to certain attributes?

# Challenges of studying implicit attitudes

- Challenge: how can we measure implicit attitudes for the judiciary?
  - But we cannot elicit IAT scores from sitting judges (yet :-) )
- Proposed solution: proxy for IAT using large amounts of written text
  - Represent judicial language in vector space
  - Are words representing different groups associated to certain attributes?

# Words closest to female and male dimension


- Kerosene, petroleum, poured, modesty, cooperative, torture, harassed

Word-Embedding Association Test: $WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$ (Caliskan et al. 2017)

distance between IAT vectors correlate with behavioral delays

- $X$, $Y$ are **male** (his, he, him, mr, himself) vs. **female** words (her, she, ms, women, woman)
- $A$, $B$ are **career** (company, work, business, service, pay) vs. **family** (family, wife, husband, mother, father)

# Words closest to female and male dimension


- Kerosene, petroleum, poured, modesty, cooperative, torture, harassed

Word-Embedding Association Test: $WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$ (Caliskan et al. 2017)

distance between IAT vectors correlate with behavioral delays

- $X$, $Y$ are **male** (his, he, him, mr, himself) vs. **female** words (her, she, ms, women, woman)
- $A$, $B$ are **career** (company, work, business, service, pay) vs. **family** (family, wife, husband, mother, father)

# Words closest to female and male dimension


- Kerosene, petroleum, poured, modesty, cooperative, torture, harassed

Word-Embedding Association Test: $WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$ (Caliskan et al. 2017)

distance between IAT vectors correlate with behavioral delays

- $X$, $Y$ are **male** (his, he, him, mr, himself) vs. **female** words (her, she, ms, women, woman)
- $A$, $B$ are **career** (company, work, business, service, pay) vs. **family** (family, wife, husband, mother, father)

# Words closest to female and male dimension


- Kerosene, petroleum, poured, modesty, cooperative, torture, harassed

Word-Embedding Association Test: $WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$ (Caliskan et al. 2017)

distance between IAT vectors correlate with behavioral delays

- $X$, $Y$ are **male** (his, he, him, mr, himself) vs. **female** words (her, she, ms, women, woman)
- $A$, $B$ are **career** (company, work, business, service, pay) vs. **family** (family, wife, husband, mother, father)

# Words closest to female and male dimension


- Kerosene, petroleum, poured, modesty, cooperative, torture, harassed

Word-Embedding Association Test: $WEAT = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$ (Caliskan et al. 2017)

distance between IAT vectors correlate with behavioral delays

- $X$, $Y$ are **male** (his, he, him, mr, himself) vs. **female** words (her, she, ms, women, woman)
- $A$, $B$ are **career** (company, work, business, service, pay) vs. **family** (family, wife, husband, mother, father)

# How to represent text as data?

- (obama speaks media illinois) is orthogonal to (president greets press chicago) according to **cosine similarity**

- But **word embeddings** capture contextual similarities between words

1. Finding the degree of similarity between two words.
```
model.similarity('woman','man')
0.73723527
```
2. Finding odd one out.
```
model.doesnt_match('breakfast cereal dinner
lunch';.split())
'cereal'
```
3. Amazing things like woman+king-man =queen
```
model.most_similar(positive=
['woman','king'],negative=['man'],topn=1)
queen: 0.508
```
4. Probability of a text under the model
```
model.score(['The fox jumped over the lazy
dog'.split()])
0.21
```

- If we know the words having similar meanings in different languages, word embeddings can be used to (Google) translate!

"The Great AI Awakening", New York Times, Dec 14, 2016

# How to represent text as data?

- (obama speaks media illinois) is orthogonal to (president greets press chicago) according to **cosine similarity**

- But **word embeddings** capture contextual similarities between words

1. Finding the degree of similarity between two words.
```
model.similarity('woman','man')
0.73723527
```
2. Finding odd one out.
```
model.doesnt_match('breakfast cereal dinner
lunch';.split())
'cereal'
```
3. Amazing things like woman+king-man =queen
```
model.most_similar(positive=
['woman','king'],negative=['man'],topn=1)
queen: 0.508
```
4. Probability of a text under the model
```
model.score(['The fox jumped over the lazy
dog'.split()])
0.21
```

- If we know the words having similar meanings in different languages, word embeddings can be used to (Google) translate!

"The Great AI Awakening", New York Times, Dec 14, 2016

# How to represent text as data?

- (obama speaks media illinois) is orthogonal to (president greets press chicago) according to **cosine similarity**

- But **word embeddings** capture contextual similarities between words

1. Finding the degree of similarity between two words.
   ```
   model.similarity('woman','man')
   0.73723527
   ```
2. Finding odd one out.
   ```
   model.doesnt_match('breakfast cereal dinner
   lunch';.split())
   'cereal'
   ```
3. Amazing things like woman+king-man =queen
   ```
   model.most_similar(positive=
   ['woman','king'],negative=['man'],topn=1)
   queen: 0.508
   ```
4. Probability of a text under the model
   ```
   model.score(['The fox jumped over the lazy
   dog'.split()])
   0.21
   ```

- If we know the words having similar meanings in different languages, word embeddings can be used to (Google) translate!

"The Great AI Awakening", New York Times, Dec 14, 2016

# How to represent text as data?

- (obama speaks media illinois) is orthogonal to (president greets press chicago) according to **cosine similarity**
- But **word embeddings** capture contextual similarities between words

1. Finding the degree of similarity between two words.
   ```
   model.similarity('woman','man')
   0.73723527
   ```
2. Finding odd one out.
   ```
   model.doesnt_match('breakfast cereal dinner
   lunch';.split())
   'cereal'
   ```
3. Amazing things like woman+king-man =queen
   ```
   model.most_similar(positive=
   ['woman','king'],negative=['man'],topn=1)
   queen: 0.508
   ```
4. Probability of a text under the model
   ```
   model.score(['The fox jumped over the lazy
   dog'.split()])
   0.21
   ```

- If we know the words having similar meanings in different languages, word embeddings can be used to (Google) translate!

"The Great AI Awakening", New York Times, Dec 14, 2016

# Distance encodes semantic similarity between words

- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right) \left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Distance encodes semantic similarity between words

- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right)\left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - ★ 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Distance encodes semantic similarity between words

- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right)\left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Distance encodes semantic similarity between words

- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right)\left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Distance encodes semantic similarity between words

- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right) \left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - ★ 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Distance encodes semantic similarity between words

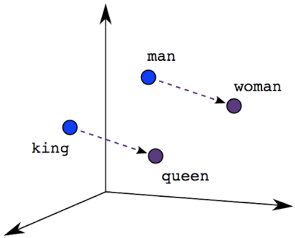
- GloVe (Global Vectors)
  - Based on intuition that co-occurrence probabilities convey meaning
  - Begins by contructing a co-occurence matrix using a fixed window
  - Obtains word vectors $w_i \in (-1, 1)^{300}$ that minimize

$$J(\boldsymbol{w}) = \sum_{i,j} f\left(X_{ij}\right) \left(w_i^T w_j - \log\left(X_{ij}\right)\right)^2$$

  - $X_{ij}$ is the co-occurrence count between words $i$ and $j$
  - $f(\cdot)$ is a weighting function that down-weights frequent words
  - Objective function $J(\cdot)$ trains word vectors to minimize squared difference between dot product of vectors representing two words and their empirical co-occurrence
  - Minimize $J(\cdot)$ by stochastic gradient descent (Pennington et al. 2014)
    - ★ 300-dimensional vectors, 50K vocabulary, window of 10 words, 0.05 learning rate, 20 epochs

# Word embeddings identify cultural dimensions

- Identify cultural dimension by taking difference between pairs of words



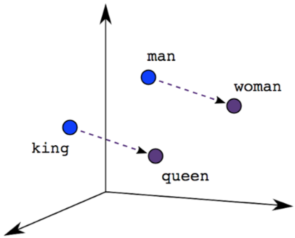

- $\overrightarrow{man} - \overrightarrow{woman}$ identifies a step in masculine direction

$$\overrightarrow{male} - \overrightarrow{female} = \frac{\sum_n \overrightarrow{male\ word_n}}{|N_{male}|} - \frac{\sum_n \overrightarrow{female\ word_n}}{|N_{female}|}$$

where $|N_{male}|$ is number of words used to identify the male dimension, e.g. $\overrightarrow{boy} - \overrightarrow{girl}$, $\overrightarrow{he} - \overrightarrow{she}$, etc.

# Word embeddings identify cultural dimensions

- Identify cultural dimension by taking difference between pairs of words

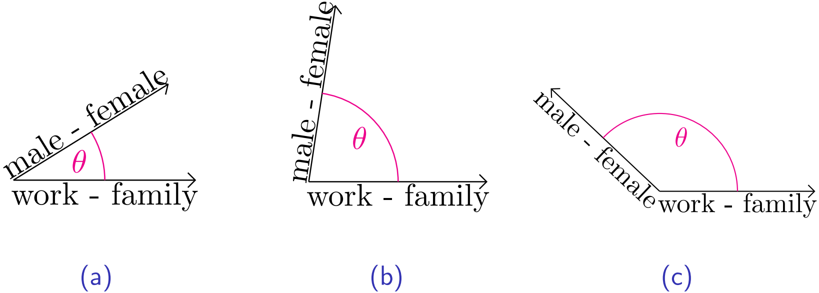



- $\overrightarrow{man} - \overrightarrow{woman}$ identifies a step in masculine direction

$$\overrightarrow{male} - \overrightarrow{female} = \frac{\sum_n \overrightarrow{male\ word_n}}{|N_{male}|} - \frac{\sum_n \overrightarrow{female\ word_n}}{|N_{female}|}$$

where $|N_{male}|$ is number of words used to identify the male dimension, e.g. $\overrightarrow{boy} - \overrightarrow{girl}$, $\overrightarrow{he} - \overrightarrow{she}$ , etc.

Measuring Gender Stereotypes using Cosine Similarity

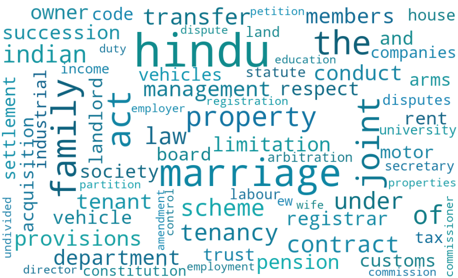


# Religion Dimension

| | |
|---|---|
| Hindu | hindu, hindus, hinduism |
| Muslim | muslim, muslims, islam, islamic |

- Highest positive and negative correlation to hindu-muslim dimension:

Words most correlated to $\overrightarrow{hindu} - \overrightarrow{muslim}$

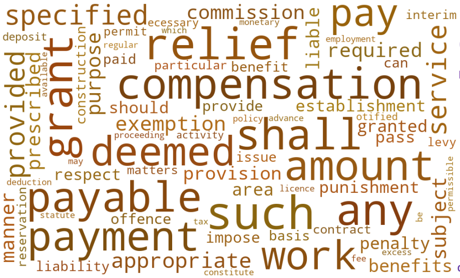



Words most correlated to $\overrightarrow{muslim} - \overrightarrow{hindu}$

# Stereotypes: The Career-Family Dimension

| | |
|---|---|
| **Career** | company, inc, work, business, service, pay, corp, employee, employment, benefits |
| **Family** | family, wife, husband, mother, father, parents, son, brother, parent, brothers |

Words most correlated to $\overrightarrow{career} - \overrightarrow{family}$




Words most correlated to $\overrightarrow{family} - \overrightarrow{career}$

# Prejudice: The Pleasant-Unpleasant Dimension

| | |
|---|---|
| **Pleasant** | good,better,best,pleasant,desirable,joy, love, peace, wonderful, |
| **Unpleasant** | bad,worse,worst,unpleasant,undesirable, terrible, horrible, nasty, war, failure |

Words most correlated to $\overrightarrow{pleasant} - \overrightarrow{unpleasant}$ Words most correlated to $\overrightarrow{unpleasant} - \overrightarrow{pleasant}$


What we have done in U.S. Courts

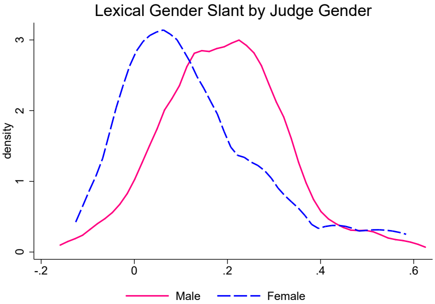
# Words closest to female and male dimension


- Migraine, hysterical, morbid, obese, terrified, unemancipated, battered
- Reserve, industrial, honorable, commanding, armed, conscientious, duty

# Words closest to female and male dimension


- Migraine, hysterical, morbid, obese, terrified, unemancipated, battered
- Reserve, industrial, honorable, commanding, armed, conscientious, duty

# Figure: Gender Slant, by Gender




Notes: The graphs show the distribution of the slant measure (cosine similarity between the gender and career-family dimensions), by judge gender. (p=0.012)

# Judges with more lexical slant are less likely to vote in favor of women's interests

| Dataset | Epstein et al. (2013) Data | | | Glynn and Sen (2015) Data | | |
|---|---|---|---|---|---|---|
| Gender Slant | -0.041*** | -0.041*** | -0.066*** | -0.053*** | -0.054*** | -0.058** |
| | (0.013) | (0.013) | (0.018) | (0.019) | (0.019) | (0.023) |
| Democrat | 0.150*** | 0.142*** | 0.185*** | 0.257*** | 0.259*** | 0.263*** |
| | (0.031) | (0.031) | (0.035) | (0.044) | (0.046) | (0.056) |
| Female | 0.122*** | 0.143*** | 0.089*** | 0.079** | 0.105*** | 0.096** |
| | (0.026) | (0.036) | (0.022) | (0.035) | (0.037) | (0.041) |
| Observations | 2335 | 2335 | 2335 | 1719 | 1719 | 1719 |
| Clusters | 112 | 112 | 112 | 109 | 109 | 109 |
| Outcome Mean | 0.4167 | 0.417 | 0.417 | 0.383 | 0.383 | 0.383 |
| Circuit-Year FE | X | X | X | X | X | X |
| Topic FE | X | X | X | X | X | X |
| Demographic Controls | X | X | X | X | X | X |
| + Interactions | | X | | | X | |
| Career FE (judge bio) | | | X | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow$20% pro-women's rights vote

## Judges with more lexical slant are less likely to vote in favor of women's interests

| Dataset | Epstein et al. (2013) Data | | | Glynn and Sen (2015) Data | | |
|---|---|---|---|---|---|---|
| Gender Slant | -0.041*** | -0.041*** | -0.066*** | -0.053*** | -0.054*** | -0.058** |
| | (0.013) | (0.013) | (0.018) | (0.019) | (0.019) | (0.023) |
| Democrat | 0.150*** | 0.142*** | 0.185*** | 0.257*** | 0.259*** | 0.263*** |
| | (0.031) | (0.031) | (0.035) | (0.044) | (0.046) | (0.056) |
| Female | 0.122*** | 0.143*** | 0.089*** | 0.079** | 0.105*** | 0.096** |
| | (0.026) | (0.036) | (0.022) | (0.035) | (0.037) | (0.041) |
| Observations | 2335 | 2335 | 2335 | 1719 | 1719 | 1719 |
| Clusters | 112 | 112 | 112 | 109 | 109 | 109 |
| Outcome Mean | 0.4167 | 0.417 | 0.417 | 0.383 | 0.383 | 0.383 |
| Circuit-Year FE | X | X | X | X | X | X |
| Topic FE | X | X | X | X | X | X |
| Demographic Controls | X | X | X | X | X | X |
| + Interactions | | X | | | X | |
| Career FE (judge bio) | | | X | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow 20\%$ pro-women's rights vote

# Panels with more slanted senior judges are less likely to assign opinions to women

| | | | | | | |
|---|---|---|---|---|---|---|
| Gender Slant | -0.020** | -0.020** | -0.015* | -0.023*** | -0.023*** | -0.026** |
| | (0.008) | (0.008) | (0.008) | (0.008) | (0.007) | (0.010) |
| Democrat | -0.065** | -0.033 | -0.080** | -0.067** | -0.059** | -0.049 |
| | (0.029) | (0.034) | (0.033) | (0.030) | (0.026) | (0.036) |
| Female | 0.137*** | 0.146*** | 0.160*** | 0.137*** | 0.135*** | |
| | (0.015) | (0.018) | (0.016) | (0.016) | (0.016) | |
| Observations | 32052 | 32052 | 32052 | 31858 | 36939 | 19940 |
| Clusters | 125 | 125 | 125 | 123 | 125 | 125 |
| Outcome Mean | 0.383 | 0.383 | 0.383 | 0.383 | 0.383 | 0.4325 |
| Circuit-Year FE | X | X | X | X | X | X |
| Demographic Controls | X | X | X | X | X | X |
| + Interactions | | X | | | | |
| Career FE | | | X | | | |
| Liberal % (Songer-Auburn) | | | | X | | |
| Includes 2-1 | | | | | X | |
| Excludes Female Senior Judge | | | | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow 10\%$ female assigned authorship

## Panels with more slanted senior judges are less likely to assign opinions to women

| | | | | | | |
|---|---|---|---|---|---|---|
| Gender Slant | -0.020** | -0.020** | -0.015* | -0.023*** | -0.023*** | -0.026** |
| | (0.008) | (0.008) | (0.008) | (0.008) | (0.007) | (0.010) |
| Democrat | -0.065** | -0.033 | -0.080** | -0.067** | -0.059** | -0.049 |
| | (0.029) | (0.034) | (0.033) | (0.030) | (0.026) | (0.036) |
| Female | 0.137*** | 0.146*** | 0.160*** | 0.137*** | 0.135*** | |
| | (0.015) | (0.018) | (0.016) | (0.016) | (0.016) | |
| Observations | 32052 | 32052 | 32052 | 31858 | 36939 | 19940 |
| Clusters | 125 | 125 | 125 | 123 | 125 | 125 |
| Outcome Mean | 0.383 | 0.383 | 0.383 | 0.383 | 0.383 | 0.4325 |
| Circuit-Year FE | X | X | X | X | X | X |
| Demographic Controls | X | X | X | X | X | X |
| + Interactions | | X | | | | |
| Career FE | | | X | | | |
| Liberal % (Songer-Auburn) | | | | X | | |
| Includes 2-1 | | | | | X | |
| Excludes Female Senior Judge | | | | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow 10\%$ female assigned authorship

# Judges with more lexical slant cite female judges less

| Dependent Variable | Cites at Least One Female Judge | | | |
|---|---|---|---|---|
| Gender Slant | -0.009* | -0.008* | -0.010* | -0.010* |
| | (0.005) | (0.005) | (0.006) | (0.005) |
| Democrat | -0.021 | -0.030* | -0.046*** | -0.026* |
| | (0.015) | (0.015) | (0.015) | (0.015) |
| Female | 0.123*** | 0.107*** | 0.134*** | 0.122*** |
| | (0.015) | (0.017) | (0.013) | (0.015) |
| Observations | 107923 | 107923 | 107923 | 106557 |
| Clusters | 139 | 139 | 139 | 136 |
| Outcome Mean | 0.383 | 0.383 | 0.383 | 0.381 |
| Circuit-Year FE | X | X | X | X |
| Demographic Controls | X | X | X | X |
| Interacted Demographic Controls | | X | | |
| Career FE | | | X | X |
| Liberal % (Songer-Auburn) | | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow 6\%$ citing a female

# Judges with more lexical slant cite female judges less

| Dependent Variable | Cites at Least One Female Judge | | | |
|---|---|---|---|---|
| Gender Slant | -0.009* | -0.008* | -0.010* | -0.010* |
| | (0.005) | (0.005) | (0.006) | (0.005) |
| Democrat | -0.021 | -0.030* | -0.046*** | -0.026* |
| | (0.015) | (0.015) | (0.015) | (0.015) |
| Female | 0.123*** | 0.107*** | 0.134*** | 0.122*** |
| | (0.015) | (0.017) | (0.013) | (0.015) |
| Observations | 107923 | 107923 | 107923 | 106557 |
| Clusters | 139 | 139 | 139 | 136 |
| Outcome Mean | 0.383 | 0.383 | 0.383 | 0.381 |
| Circuit-Year FE | X | X | X | X |
| Demographic Controls | X | X | X | X |
| Interacted Demographic Controls | | X | | |
| Career FE | | | X | X |
| Liberal % (Songer-Auburn) | | | | X |

$2\sigma$ of gender slant $\Rightarrow \downarrow 6\%$ citing a female

# Judges with more lexical slant reverse female district judges more

| | | | | |
|---|---|---|---|---|
| Gender Slant * Female District Judge | 0.010*** | 0.010*** | 0.012*** | 0.012*** |
| | (0.004) | (0.004) | (0.004) | (0.004) |
| Democrat * Female District Judge | -0.009 | -0.024** | -0.006 | -0.007 |
| | (0.014) | (0.009) | (0.014) | (0.013) |
| Female * Female District Judge | -0.009 | -0.022*** | -0.007 | -0.011 |
| | (0.009) | (0.008) | (0.009) | (0.010) |
| Democrat * Female * Female District Judge | | 0.152*** | | |
| | | (0.015) | | |
| Observations | 145862 | 145862 | 144965 | 145563 |
| Clusters | 133 | 133 | 130 | 133 |
| Outcome Mean for Male Judges | 0.180 | 0.180 | 0.180 | 0.180 |
| Outcome Mean for Female Judges | 0.157 | 0.157 | 0.157 | 0.157 |
| Circuit-Year FE | X | X | X | X |
| Judge FE | X | X | X | X |
| District Judge FE | X | X | X | X |
| Demographic Controls | X | X | X | X |
| + Interactions | | X | | |
| Liberal Score Interaction | | | X | |
| District-Year FE | | | | X |

# But female judges are 3.6% less likely to be reversed

| | | | | |
|---|---|---|---|---|
| Gender Slant * Female District Judge | 0.010*** | 0.010*** | 0.012*** | 0.012*** |
| | (0.004) | (0.004) | (0.004) | (0.004) |
| Democrat * Female District Judge | -0.009 | -0.024** | -0.006 | -0.007 |
| | (0.014) | (0.009) | (0.014) | (0.013) |
| Female * Female District Judge | -0.009 | -0.022*** | -0.007 | -0.011 |
| | (0.009) | (0.008) | (0.009) | (0.010) |
| Democrat * Female * Female District Judge | | 0.152*** | | |
| | | (0.015) | | |
| Observations | 145862 | 145862 | 144965 | 145563 |
| Clusters | 133 | 133 | 130 | 133 |
| Outcome Mean for Male Judges | 0.180 | 0.180 | 0.180 | 0.180 |
| Outcome Mean for Female Judges | 0.157 | 0.157 | 0.157 | 0.157 |
| Circuit-Year FE | X | X | X | X |
| Judge FE | X | X | X | X |
| District Judge FE | X | X | X | X |
| Demographic Controls | X | X | X | X |
| + Interactions | | X | | |
| Liberal Score Interaction | | | X | |
| District-Year FE | | | | X |

# Daughters Reduce Gender Slant

| | | |
|---|---|---|
| Daughter | -0.477* | -0.468* |
| | (0.274) | (0.278) |
| Democrat | -0.016 | -0.069 |
| | (0.535) | (0.613) |
| Female | -0.659*** | -0.683*** |
| | (0.232) | (0.239) |
| Democrat * Female | | 0.321 |
| | | (0.631) |
| Observations | 98 | 98 |
| Outcome Mean | -0.085 | -0.085 |
| Adjusted R2 | 0.528 | 0.520 |
| Circuit FE | X | X |
| Number of Children FE | X | X |
| Demographic Controls | X | X |
| Interacted Demographic Controls | | X |

Conditional on number of children, having a daughter as good as random.

# Daughters Reduce Gender Slant

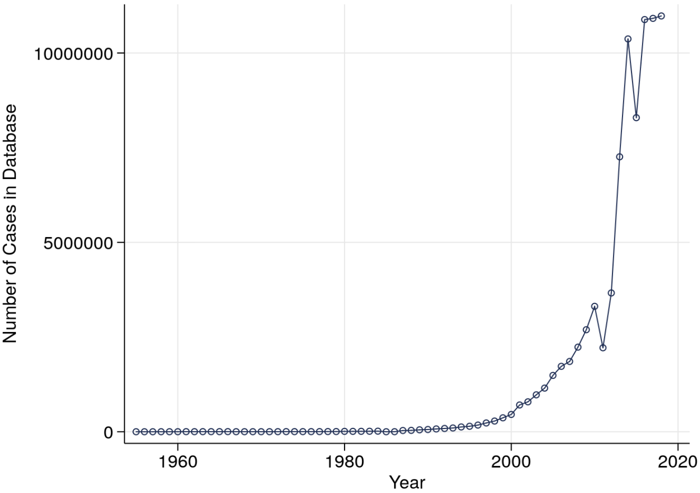
| | | |
|---|---|---|
| Daughter | -0.477* | -0.468* |
| | (0.274) | (0.278) |
| Democrat | -0.016 | -0.069 |
| | (0.535) | (0.613) |
| Female | -0.659*** | -0.683*** |
| | (0.232) | (0.239) |
| Democrat * Female | | 0.321 |
| | | (0.631) |
| Observations | 98 | 98 |
| Outcome Mean | -0.085 | -0.085 |
| Adjusted R2 | 0.528 | 0.520 |
| Circuit FE | X | X |
| Number of Children FE | X | X |
| Demographic Controls | X | X |
| Interacted Demographic Controls | | X |

Conditional on number of children, having a daughter as good as random.

What we are doing in Indian Courts

# India E-Courts

Figure: Number of Cases per Year, India E-Courts



# Meta-Data

- We have parsed the cases and hearings to pull out relevant metadata (331 and 256 fields respectively).
  - dates, court, parties, case type, and judge identity
- Simple measures of court efficiency that can be generated from the administrative data
  - total caseload, trial duration, case disposal rates, backlogs, appeal rates, and proportion of appeals successfully upheld
- Measures of judicial outcomes include
  - resolution, ruling, and sentence

# Meta-Data

- We have parsed the cases and hearings to pull out relevant metadata (331 and 256 fields respectively).
  - dates, court, parties, case type, and judge identity
- Simple measures of court efficiency that can be generated from the administrative data
  - total caseload, trial duration, case disposal rates, backlogs, appeal rates, and proportion of appeals successfully upheld
- Measures of judicial outcomes include
  - resolution, ruling, and sentence

# Meta-Data

- We have parsed the cases and hearings to pull out relevant metadata (331 and 256 fields respectively).
  - dates, court, parties, case type, and judge identity
- Simple measures of court efficiency that can be generated from the administrative data
  - total caseload, trial duration, case disposal rates, backlogs, appeal rates, and proportion of appeals successfully upheld
- Measures of judicial outcomes include
  - resolution, ruling, and sentence

# Written Opinions 1930-2018

- case title/citation, dates, judges on the panel, author
- split into sections and paragraphs
- annotated citations to legal authorities, i.e. statutes and previous cases

Separate dataset - lists of first and last names digitized from the Anthropological Survey of India

- gender, social identity (subcaste or jati) and religion (i.e., for Muslims)
- data on judges collected from the National Judicial Data Grid

The Socioeconomic and Caste Census (2012) is a *census* that describes household income and assets for all Indians.

The Economic Census (2005 and 2012) describes the universe of firms with ten or more employees

# Written Opinions 1930-2018

- case title/citation, dates, judges on the panel, author
- split into sections and paragraphs
- annotated citations to legal authorities, i.e. statutes and previous cases

Separate dataset - lists of first and last names digitized from the Anthropological Survey of India

- gender, social identity (subcaste or jati) and religion (i.e., for Muslims)

- data on judges collected from the National Judicial Data Grid

The Socioeconomic and Caste Census (2012) is a *census* that describes household income and assets for all Indians.
The Economic Census (2005 and 2012) describes the universe of firms with ten or more employees

# Written Opinions 1930-2018

- case title/citation, dates, judges on the panel, author
- split into sections and paragraphs
- annotated citations to legal authorities, i.e. statutes and previous cases

Separate dataset - lists of first and last names digitized from the Anthropological Survey of India

- gender, social identity (subcaste or jati) and religion (i.e., for Muslims)

- data on judges collected from the National Judicial Data Grid

The Socioeconomic and Caste Census (2012) is a *census* that describes household income and assets for all Indians.
The Economic Census (2005 and 2012) describes the universe of firms with ten or more employees

# Written Opinions 1930-2018

- case title/citation, dates, judges on the panel, author
- split into sections and paragraphs
- annotated citations to legal authorities, i.e. statutes and previous cases

Separate dataset - lists of first and last names digitized from the Anthropological Survey of India

- gender, social identity (subcaste or jati) and religion (i.e., for Muslims)
- data on judges collected from the National Judicial Data Grid

The Socioeconomic and Caste Census (2012) is a *census* that describes household income and assets for all Indians.

The Economic Census (2005 and 2012) describes the universe of firms with ten or more employees

# Written Opinions 1930-2018

- case title/citation, dates, judges on the panel, author
- split into sections and paragraphs
- annotated citations to legal authorities, i.e. statutes and previous cases

Separate dataset - lists of first and last names digitized from the Anthropological Survey of India

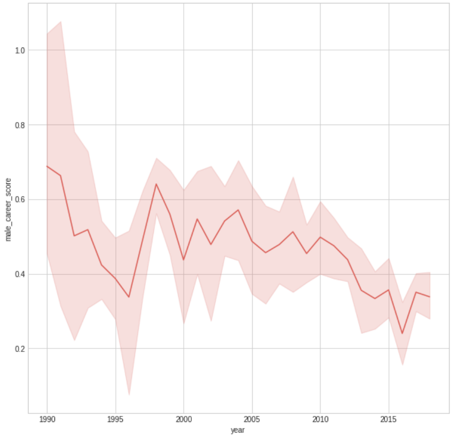
- gender, social identity (subcaste or jati) and religion (i.e., for Muslims)

- data on judges collected from the National Judicial Data Grid

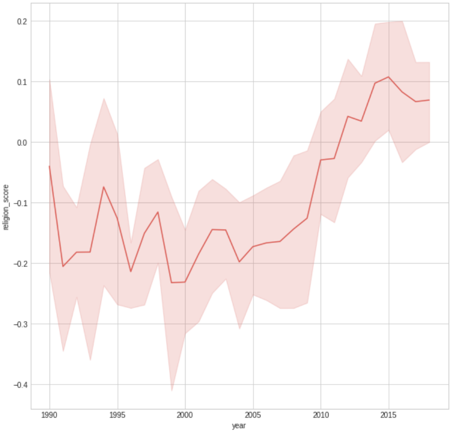
The Socioeconomic and Caste Census (2012) is a *census* that describes household income and assets for all Indians.
The Economic Census (2005 and 2012) describes the universe of firms with ten or more employees

# Gender Stereotypes and Religious Prejudice, 1990-2018

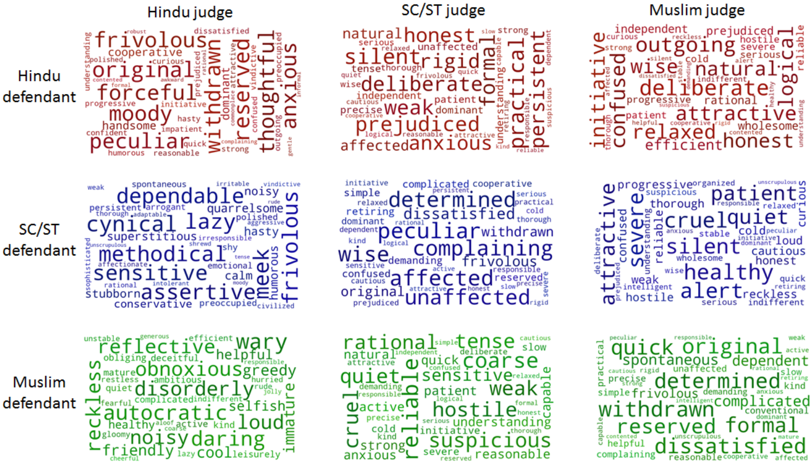
Male Association with Career
(Female Association with Family)



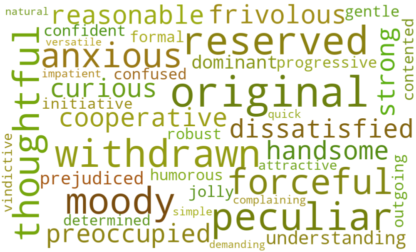

Hindu Association with Pleasant
(Muslim Association with Unpleasant)



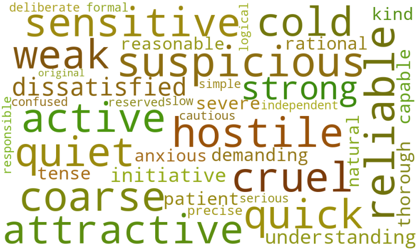
# Hindu, Muslim, and caste in India



# Sentiment analysis




Hindu judges describe Hindu litigants more positively




SC/ST judges describe Muslims more negatively

# Access to Justice

- Disparities in:
  - judicial representation
  - judicial treatment
  - judicial outcomes

# Access to Justice

Table: Distribution of Court Actors, By Social Group

| | *Counts and Percentages by Group* | | |
|---|---|---|---|
| | Hindus | | Muslims |
| | Non-Scheduled | Scheduled | |
| Civil Litigants | 3,837,066<br>87.5% | 179,613<br>4.1% | 366,278<br>8.3% |
| Criminal Defendants | 568,017<br>86.4% | 35,104<br>5.3% | 54,146<br>8.2% |
| Judges | 1,318,440<br>88.4% | 11,211<br>0.8% | 162,552<br>10.8% |

- None reflect the distribution in the population.
  - Scheduled - 16%; Muslims - 14%
  - Scheduled - 30% more likely to appear as defendants than litigants

# Access to Justice

Table: Distribution of Court Actors, By Social Group

| | *Counts and Percentages by Group* | | |
|---|---|---|---|
| | Hindus | | Muslims |
| | Non-Scheduled | Scheduled | |
| Civil Litigants | 3,837,066<br>87.5% | 179,613<br>4.1% | 366,278<br>8.3% |
| Criminal Defendants | 568,017<br>86.4% | 35,104<br>5.3% | 54,146<br>8.2% |
| Judges | 1,318,440<br>88.4% | 11,211<br>0.8% | 162,552<br>10.8% |

- None reflect the distribution in the population.
  - Scheduled - 16%; Muslims - 14%
  - Scheduled - 30% more likely to appear as defendants than litigants

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Random Assignment

- Random assignment of judges is not a universal feature of the Indian court system, but it appears in many of the subordinate courts
- Focus on Delhi courts [pop. $\sim$ Netherlands, 2x Sweden, 4x Norway]
  - all cases filed under the Indian Penal Code Act of 1860
  - all brought by the state (so defendant = respondent)
- District and Sessions Judge (the highest court in the district)
  - principal court of civil jurisdiction (most serious cases)
- Chief Metropolitan Magistrate
  - cases punishable with imprisonment for a term up to seven years.
- Court - Dwarka, Karkardooma, Patiala, Rohini, Saket, Tis Hazari.
  - Condition on this, judges appear exogenously assigned

# Outcomes + Treatment

- **Outcome:** Negative Disposition
  - For bail hearings, bail is not allowed
  - For non-bail hearings, convicted or guilty
- **Treatment:** Duration between hearings
- **Treatment:** Number of hearings per case

Table 1(a): Summary Statistics by Gender

| | Full Sample | Female Respondent | Male Respondent | *p-value* | Female Judge | Male Judge | *p-value* |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| *Case Characteristics:* | | | | | | | |
| Female | 0.087 | 1.000 | 0.000 | . | 0.059 | 0.077 | 0.074 |
| Hindu | 0.843 | 0.846 | 0.843 | 0.732 | 0.862 | 0.847 | 0.238 |
| Person Crime | 0.429 | 0.318 | 0.440 | 0.000 | 0.445 | 0.458 | 0.549 |
| Property Crime | 0.395 | 0.445 | 0.390 | 0.000 | 0.366 | 0.381 | 0.480 |
| Other Crime | 0.455 | 0.560 | 0.445 | 0.000 | 0.552 | 0.466 | 0.001 |
| Bail Hearing | 0.192 | 0.256 | 0.186 | 0.000 | 0.136 | 0.164 | 0.407 |
| Joint F-stat | | | | | | | 0.110 |
| *Court/Judge Characteristics:* | | | | | | | |
| Female Judge | 0.247 | 0.206 | 0.251 | 0.006 | 1.000 | 0.000 | . |
| Hindu Judge | 0.982 | 0.981 | 0.982 | 0.709 | 0.958 | 0.970 | 0.577 |
| Chief Metropolitan Magistrate | 0.391 | 0.194 | 0.409 | 0.000 | 0.679 | 0.419 | 0.000 |
| District and Sessions Judge | 0.609 | 0.806 | 0.590 | 0.000 | 0.321 | 0.562 | 0.000 |
| Year: 2015 | 0.042 | 0.022 | 0.044 | 0.000 | 0.075 | 0.062 | 0.423 |
| Year: 2016 | 0.335 | 0.328 | 0.336 | 0.644 | 0.360 | 0.382 | 0.549 |
| Year: 2017 | 0.204 | 0.220 | 0.202 | 0.071 | 0.142 | 0.180 | 0.133 |
| Year: 2018 | 0.419 | 0.429 | 0.418 | 0.486 | 0.422 | 0.376 | 0.242 |
| Court: Dwarka | 0.111 | 0.122 | 0.110 | 0.388 | 0.074 | 0.141 | 0.026 |
| Court: Karkrdooma | 0.352 | 0.319 | 0.355 | 0.076 | 0.226 | 0.224 | 0.970 |
| Court: Patiala House | 0.026 | 0.024 | 0.027 | 0.388 | 0.082 | 0.097 | 0.583 |
| Court: Rohini | 0.190 | 0.234 | 0.186 | 0.011 | 0.156 | 0.199 | 0.272 |
| Court: Saket | 0.133 | 0.133 | 0.133 | 0.972 | 0.211 | 0.162 | 0.227 |
| Court: Tis Hazari/Rouse | 0.187 | 0.168 | 0.189 | 0.177 | 0.251 | 0.177 | 0.078 |
| Avg. No. of Cases | | | | | 112.769 | 169.577 | 0.002 |
| *Outcomes:* | | | | | | | |
| % Negative Disposition | 0.182 | 0.192 | 0.181 | 0.246 | 0.166 | 0.166 | 0.988 |
| Duration Bt. Hearings | 28.145 | 22.049 | 28.607 | 0.000 | 39.509 | 25.827 | 0.000 |
| No. of Hearings | 4.265 | 3.334 | 4.354 | 0.000 | 5.936 | 5.694 | 0.697 |
| Observations | 61,236 | 5,315 | 55,921 | | 134 | 272 | |

Notes: This table presents summary statistics for different samples considered for the main analysis. Column (1) includes the full sample of all court cases filed under the Indian Penal Code Act between 2015 and 2018 in any district court in Delhi. The full sample only considers those cases that have been resolved and have no missing data. Column (2) includes only those cases that have a female respondent. Column (3) includes those that have a male respondent. Column (4) reports p-values on the difference of the mean of any characteristic as reported in columns (2) and (3). Column (5) includes those cases that have been decided by a female judge. Column (6)

# Comments

- Other crime = "cruelty by husband/relatives" - which is related to dowry
  - respondent is mother-in-law or husband's sister, thus more often female
  - and handled by district-session judges who are male

Table 2: Judicial Outcomes by Gender

| | | Duration Between Hearings | | | Number of Hearings | | | Negative Outcome | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| (1) | Male Judge | 11.624*** | 14.577*** | 16.338*** | 2.261*** | 2.220*** | 2.187*** | 0.063*** | 0.058 | 0.036 |
| | | (1.580) | (2.544) | (2.605) | (0.250) | (0.340) | (0.347) | (0.023) | (0.040) | (0.045) |
| (2) | Female Judge | 39.128*** | 34.930*** | 38.000*** | -2.403*** | -3.480*** | -4.424*** | 0.162*** | 0.142** | 0.141* |
| | | (6.449) | (6.166) | (6.381) | (0.889) | (0.835) | (0.925) | (0.056) | (0.068) | (0.074) |
| (3) | Female*Male Judge | -0.477 | -0.381 | -0.471 | -0.257*** | -0.222*** | -0.235*** | -0.013* | -0.012* | -0.011 |
| | | (0.733) | (0.711) | (0.726) | (0.072) | (0.070) | (0.070) | (0.007) | (0.007) | (0.007) |
| (4) | Female*Female Judge | -4.387*** | -4.211*** | -4.038*** | -0.422*** | -0.378** | -0.363** | -0.012 | -0.012 | -0.011 |
| | | (1.561) | (1.538) | (1.497) | (0.157) | (0.156) | (0.153) | (0.013) | (0.013) | (0.012) |
| | *p-value:* (3) = (4) | 0.024 | 0.024 | 0.033 | 0.340 | 0.362 | 0.445 | 0.935 | 0.996 | 0.988 |
| | *Fixed Effects* | | | | | | | | | |
| | Year | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Court | N | Y | Y | N | Y | Y | N | Y | Y |
| | Year*Court | N | N | Y | N | N | Y | N | N | Y |
| | Observations | 36,871 | 36,871 | 36,871 | 61,236 | 61,236 | 61,236 | 51,560 | 51,560 | 51,560 |

Notes: This table reports differences in judicial outcomes by gender of respondent and judge presiding over the case. For columns (1)-(3), the outcome is the average duration between any two consecutive hearings for a case. For columns (4) - (6), the outcome is the number of hearings per case. Finally, for columns (7)- (9), the outcome is whether the respondent pled guilty/was convicted versus other judgements if the purpose of the hearing was not bail while the outcome is whether the bail was dismissed versus other judgements if the purpose of hearing was bail. For each outcome, three seperate OLS regressions were run. The first set of regressions regresses the outcome on a dummy variable indicating male judge, a dummy indicating female judge, an interaction varible between female respondent and male judge, an interaction variable between female respondent and female judge, offense type interacted with male judge, offense type interacted with female judge, filing year interacted with male and female judge respectively. The second set of regressions conducts the same regression but add controls for court interacted with male and female judge, respectively. The third and final set of regressions conducts the same regression but adds controls for court interacted with year interacted with male and female judge, respectively. The fifth row reports p-values for the difference in coefficients reported in row (3) and row (4). All regressions cluster standard errors at judge level. The sample considered is all cases filed under Indian Penal Code Act between 2015 and 2018 in any district court in Delhi.

# Gender Comments

- Female judges take longer between hearings and hold fewer hearings
  - Female defendants get fewer hearings and shorter delay between hearings
  - Female judges are especially faster for female defendants
  - No difference-in-difference regarding outcomes (on average)
  - Female judges are harsher
- Randomization controls are offense type interacted with male judge, offense type interacted with female judge, filing year interacted with male and female judge. (1)
  - + court interacted with male and female judge. (2)
  - + court interacted with year interacted with male and female judge. (3)

# Crime Categories

- **Person Crime** - Any offense affecting the human body, namely,
  - (a) Murder and Culpable homicide
  - (b) Causing miscarriage, injuries to unborn children, exposure of infants and the concealment of births
  - (c) Hurt
  - (d) Wrongful restraint and confinement
  - (e) Criminal force and assault
  - (f) Kidnapping, abduction, slavery and forced labor
  - (g) Sexual offenses including rape and sodomy
- **Property Crime** - Any offense against property, namely,
  - (a) Theft
  - (b) Extortion
  - (c) Robbery and dacoity
  - (d) Criminal misappropriation of property
  - (e) Criminal breach of trust
  - (f) Receiving of stolen property
  - (g) Cheating
  - (h) Fraudulent deeds and disposition of property
  - (i) Mischief
  - (j) Criminal trespass

# Crime Categories

- **Person Crime** - Any offense affecting the human body, namely,
  - (a) Murder and Culpable homicide
  - (b) Causing miscarriage, injuries to unborn children, exposure of infants and the concealment of births
  - (c) Hurt
  - (d) Wrongful restraint and confinement
  - (e) Criminal force and assault
  - (f) Kidnapping, abduction, slavery and forced labor
  - (g) Sexual offenses including rape and sodomy
- **Property Crime** - Any offense against property, namely,
  - (a) Theft
  - (b) Extortion
  - (c) Robbery and dacoity
  - (d) Criminal misappropriation of property
  - (e) Criminal breach of trust
  - (f) Receiving of stolen property
  - (g) Cheating
  - (h) Fraudulent deeds and disposition of property
  - (i) Mischief
  - (j) Criminal trespass

Table 3: Judicial Outcomes by Subsamples

| | Duration Between Hearings | Number of Hearings | Negative Outcome |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Theft/Robbery* | | | |
| Female*Male Judge | -0.449 | -0.010 | 0.019 |
| | (1.776) | (0.308) | (0.025) |
| Female*Female Judge | -1.203 | -0.573 | 0.009 |
| | (2.760) | (0.551) | (0.038) |
| *p-value:* | 0.818 | 0.373 | 0.819 |
| Observations | 6,701 | 10,233 | 9,127 |
| *Murder and Culpable Homicide* | | | |
| Female*Male Judge | 1.121 | -0.341* | -0.022 |
| | (1.749) | (0.181) | (0.014) |
| Female*Female Judge | 2.682 | -0.099 | 0.065*** |
| | (2.291) | (0.394) | (0.025) |
| *p-value:* | 0.589 | 0.576 | 0.002 |
| Observations | 5,168 | 8,397 | 7,394 |
| *Public Health, Safety, Convenience, Decency, and Morals* | | | |
| Female*Male Judge | 1.534 | 0.163 | 0.007 |
| | (3.331) | (0.312) | (0.034) |
| Female*Female Judge | 2.122 | -0.516 | 0.158* |
| | (6.480) | (0.672) | (0.084) |
| *p-value:* | 0.936 | 0.360 | 0.099 |

# Gender Sub-Group Analysis

- Evidence of anti- in-group bias (threatened egoism?)
  - Females deciding harsher on females for murder, morals (drugs, negligence, obscenity), cruelty by relatives of husband
- Evidence of in-group bias
  - Females being lenient to females on cheating, sexual offenses

Table 3: Judicial Outcomes by Subsamples

| | Duration Between Hearings | Number of Hearings | Negative Outcome |
|---|---|---|---|
| | (1) | (2) | (3) |
| *Theft/Robbery* | | | |
| Hindu*Hindu Judge | 1.648 | -0.275$^{*}$ | -0.003 |
| | (1.441) | (0.161) | (0.011) |
| Hindu*Non-Hindu Judge | -22.224$^{***}$ | -1.085 | 0.080 |
| | (3.786) | (0.673) | (0.081) |
| *p-value:* | 0.000 | 0.243 | 0.306 |
| Observations | 3,347 | 5,364 | 4,726 |
| *Murder and Culpable Homicide* | | | |
| Hindu*Hindu Judge | 1.752$^{*}$ | 0.289 | -0.000 |
| | (0.979) | (0.176) | (0.010) |
| Hindu*Non-Hindu Judge | 4.439$^{*}$ | -0.482 | -0.184 |
| | (2.586) | (1.113) | (0.194) |
| *p-value:* | 0.332 | 0.495 | 0.345 |
| Observations | 2,912 | 4,800 | 4,199 |
| *Public Health, Safety, Convenience, Decency, and Morals* | | | |
| Hindu*Hindu Judge | -0.661 | -0.069 | 0.021 |
| | (2.521) | (0.205) | (0.023) |
| Hindu*Non-Hindu Judge | 15.091 | 0.483 | 0.098 |
| | (9.734) | (0.441) | (0.091) |
| *p-value:* | 0.119 | 0.257 | 0.413 |

# Religion Comments

- Muslim judges take longer between hearings and hold fewer hearings
  - No difference-in-difference regarding outcomes (on average)
  - Hindu judges are harsher

Table 1(b): Summary Statistics by Religion

| | Full Sample | Hindu Respondent | Non-Hindu Respondent | *p-value* | Hindu Judge | Non-Hindu Judge | *p-value* |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| *Case Characteristics:* | | | | | | | |
| Female | 0.098 | 0.098 | 0.096 | 0.751 | 0.076 | 0.057 | 0.283 |
| Hindu | 0.843 | 1.000 | 0.000 | . | 0.849 | 0.888 | 0.286 |
| Person Crime | 0.411 | 0.407 | 0.428 | 0.072 | 0.436 | 0.531 | 0.193 |
| Property Crime | 0.384 | 0.384 | 0.385 | 0.947 | 0.360 | 0.385 | 0.762 |
| Other Crime | 0.501 | 0.517 | 0.415 | 0.000 | 0.535 | 0.552 | 0.807 |
| Bail Hearing | 0.205 | 0.206 | 0.203 | 0.903 | 0.163 | 0.075 | 0.241 |
| Joint F-stat | | | | | | | 0.297 |
| *Court/Judge Characteristics:* | | | | | | | |
| Female Judge | 0.226 | 0.225 | 0.229 | 0.843 | 0.309 | 0.417 | 0.456 |
| Hindu Judge | 0.982 | 0.981 | 0.988 | 0.095 | 1.000 | 0.000 | . |
| Chief Metropolitan Magistrate | 0.376 | 0.374 | 0.391 | 0.471 | 0.514 | 0.417 | 0.504 |
| District and Sessions Judge | 0.623 | 0.626 | 0.608 | 0.474 | 0.475 | 0.583 | 0.457 |
| Year: 2015 | 0.039 | 0.038 | 0.041 | 0.425 | 0.060 | 0.070 | 0.816 |
| Year: 2016 | 0.332 | 0.334 | 0.322 | 0.532 | 0.371 | 0.368 | 0.973 |
| Year: 2017 | 0.207 | 0.203 | 0.224 | 0.099 | 0.170 | 0.096 | 0.073 |
| Year: 2018 | 0.423 | 0.425 | 0.413 | 0.527 | 0.398 | 0.466 | 0.552 |
| Court: Dwarka | 0.119 | 0.131 | 0.054 | 0.000 | 0.116 | 0.083 | 0.686 |
| Court: Karkrdooma | 0.343 | 0.319 | 0.472 | 0.000 | 0.224 | 0.064 | 0.015 |
| Court: Patiala House | 0.022 | 0.023 | 0.018 | 0.071 | 0.086 | 0.173 | 0.382 |
| Court: Rohini | 0.188 | 0.203 | 0.107 | 0.000 | 0.181 | 0.416 | 0.103 |
| Court: Saket | 0.143 | 0.137 | 0.173 | 0.022 | 0.193 | 0.014 | 0.000 |
| Court: Tis Hazari/Rouse | 0.185 | 0.186 | 0.177 | 0.607 | 0.200 | 0.250 | 0.694 |
| Avg. No. of Cases | | | | | 113.678 | 63.750 | 0.023 |
| *Outcomes:* | | | | | | | |
| % Negative Disposition | 0.191 | 0.191 | 0.192 | 0.945 | 0.171 | 0.153 | 0.768 |
| Duration Bt. Hearings | 28.025 | 28.165 | 27.289 | 0.472 | 30.193 | 29.533 | 0.895 |
| No. of Hearings | 4.083 | 4.081 | 4.091 | 0.932 | 5.430 | 6.032 | 0.719 |
| Observations | 42,371 | 35,721 | 6,650 | | 366 | 12 | |

Notes: This table presents summary statistics for different samples considered for the main analysis. Column (1) includes the full sample of all court cases filed under the Indian Penal Code Act between 2015 and 2018 in any district court in Delhi. The full sample only considers those cases that have been resolved and have no missing data. Column (2) includes only those cases that have a Hindu respondent. Column (3) includes those that have a non-Hindu respondent. Column (4) reports p-values on the difference of the mean of any characteristic as reported in columns (2) and (3). Column (5) includes those cases that have been decided by a Hindu judge. Column (6) includes those that have been decided by a non-Hindu judge. Column (7) reports p-values on the difference of the mean of any characteristic as reported in columns (5) and (6).

Table 2: Judicial Outcomes by Religion

| | | Duration Between Hearings | | | Number of Hearings | | | Negative Outcome | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| (1) | Hindu Judge | 30.129$^{***}$ | 28.397$^{***}$ | 28.552$^{***}$ | 10.656$^{***}$ | 11.103$^{***}$ | 10.924$^{***}$ | 0.163$^{***}$ | 0.156$^{***}$ | 0.156$^{***}$ |
| | | (3.268) | (3.658) | (3.707) | (1.061) | (1.048) | (1.080) | (0.037) | (0.046) | (0.049) |
| (2) | Non-Hindu Judge | 21.320$^{**}$ | 10.144 | 62.809$^{***}$ | 8.014$^{***}$ | 2.936$^{*}$ | 7.952$^{***}$ | 0.150 | 0.250$^{*}$ | 0.034 |
| | | (9.448) | (15.911) | (5.601) | (1.419) | (1.747) | (1.383) | (0.165) | (0.132) | (0.136) |
| (3) | Hindu*Hindu Judge | 0.531 | 1.518$^{*}$ | 1.616$^{*}$ | 0.070 | 0.114$^{*}$ | 0.134$^{**}$ | -0.001 | 0.004 | 0.004 |
| | | (1.005) | (0.900) | (0.897) | (0.080) | (0.067) | (0.065) | (0.006) | (0.006) | (0.006) |
| (4) | Hindu*Non-Hindu Judge | 0.997 | 2.070 | 1.603 | -0.628 | -0.101 | -0.430 | -0.074$^{***}$ | -0.033 | -0.029 |
| | | (3.745) | (3.552) | (3.587) | (0.746) | (0.485) | (0.561) | (0.022) | (0.020) | (0.022) |
| | *p-value:* (3) = (4) | 0.904 | 0.880 | 0.997 | 0.353 | 0.660 | 0.319 | 0.002 | 0.082 | 0.158 |
| | *Fixed Effects* | | | | | | | | | |
| | Year | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| | Court | N | Y | Y | N | Y | Y | N | Y | Y |
| | Year*Court | N | N | Y | N | N | Y | N | N | Y |
| | Observations | 24,993 | 24,993 | 24,993 | 42,371 | 42,371 | 42,371 | 35,226 | 35,226 | 35,226 |

Notes: This table reports differences in judicial outcomes by religion of respondent and judge presiding over the case. For columns (1)-(3), the outcome is the average duration between any two consecutive hearings for a case. For columns (4) - (6), the outcome is the number of hearings per case. Finally, for columns (7)- (9), the outcome is whether the respondent pled guilty/was convicted versus other judgements if the purpose of the hearing was not bail while the outcome is whether the bail was dismissed versus other judgements if the purpose of hearing was bail. For each outcome, three seperate OLS regressions were run. The first set of regressions regresses the outcome on a dummy variable indicating Hindu judge, a dummy indicating non-Hindu judge, an interaction varible between Hindu respondent and Hindu judge, an interaction variable between Hindu respondent and non-Hindu judge, offense type interacted with Hindu judge, offense type interacted with non-Hindu judge, filing year interacted with Hindu and non-Hindu judge respectively. The second set of regressions conducts the same regression but add controls for court interacted with Hindu and non-Hindu judge, respectively. The third and final set of regressions conducts the same regression but adds controls for court interacted with year interacted with Hindu and non-Hindu judge, respectively. The fifth row reports p-values for the difference in coefficients reported in row (3) and row (4). All regressions cluster standard errors at judge level. The sample considered is all cases filed under Indian Penal Code Act between 2015 and 2018 in any district court in Delhi.

# Data Explorer

- https://explore-ecourts.herokuapp.com/
- Court backlog
- Environment
- Network analysis of lawyers and judges

# Data Explorer

- https://explore-ecourts.herokuapp.com/
- Court backlog
- Environment
- Network analysis of lawyers and judges

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'

# Using Machine Learning to Identify and Mitigate Bias

- Backlash to AI vs. Incremental AI
- In Stage 0, assess judges vs. a bootstrapped judge (predicted decision-maker)
- In Stage 1, people use AI as a support tool, speeding up existing processes (for example, by prefilling forms)
- Once they're used to this, they can more easily accept an added functionality (Stage 2) in which AI becomes a choice monitor, pointing out choice inconsistencies (pay more attention / be less indifferent)
- Stage 3 elevates the AI to the role of a more general coach, providing outcome feedback on choices and highlighting decision patterns.
  - Transparent + explanable | explain why deviate
- Then, in Stage 4, the AI brings in other people's decision histories and patterns, serving as a platform for a community of experts.
  - (1) self-image, (2) self-improvement, (3) self-understanding, (4) ego
  - (0) ↓bias, (1) ↑autonomy, (2) ↑learning, (3) ↑transparency, ↓status quo, ↓adversarial attack
- Only in Stage 5, recommend the 'optimal decision'