An Improved Method of Automated Nonparametric Content Analysis for Social Science[1]

Gary King[2]

Institute for Quantitative Social Science
Harvard University

Texas A&M Inaugural STATA Lecture, 1/19/2017

[1]Based on joint work with Connor Jerzak and Anton Strezhnev
[2]GaryKing.org

Mortality Data, Developed Countries:

Mortality Data, Developed Countries: Death Certificates

Mortality Data, Developed Countries: Death Certificates

LOCAL CERTIFIED RECORD OF DEATH

CITY-COUNTY DEPARTMENT OF HEALTH
City of Evansville — Vanderburgh County
Evansville, Indiana

№ 14397

This is to Certify, that our records show Emily Elizabeth Hale died July 28 1952 (month, day, year) at 6 pm (hour of death) 31 E. Illinois St. (street, hospital or rural)

Age at death 92 (years) Sex female Race white widow (marital status)

Primary cause of death given was cardio vascular renal disease 5 years

Signed by L. B Miller (physician or coroner) city (address)

Place of burial or removal Memorial Park (name of cemetery) city (address)

Date of Burial 7-30-52 Johann (Funeral Director) city (address)

Signed Jane M. Hoopes MD Registrar.

Evansville, Indiana 8-20-80 (date) [SEAL]

NOTE: Recorded locally in Book No. 19 Page No. 33 Reg. No. 1041

FEE $2.00

Mortality Data, Most of the World:

Mortality Data, Most of the World: **?**

Mortality Data, Most of the World: **?**

Mortality Data, Most of the World: Verbal Autopsy

Verbal Autopsy Methods

Verbal Autopsy Methods

- The Problem

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods:

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts;

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight:

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!
- They care about:

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!
- They care about: % in categories,

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!
- They care about: <u>% in categories,</u> not <u>individual classification</u>

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!
- They care about: <u>% in categories,</u> not <u>individual classification</u>
- Statistical problem:

Verbal Autopsy Methods

- The Problem
  - Policymakers need the cause-specific mortality %'s to set research goals, donor priorities, and ameliorative policies
  - High quality death registration data: 23/192 countries
- Existing Approaches
  - Verbal Autopsy: Ask relatives or caregivers 50+ symptom questions
  - Ask physicians (low intercoder reliability)
  - Classification algorithms
    - Find deaths with medically certified causes at a local hospital,
    - Trace caregivers to their homes, ask the same symptom questions
    - Statistically classify deaths in community
    - (model-dependent, low accuracy)
  - Summary of existing methods: huge efforts; few reliable results
- Our Key insight: in public health, no one cares about you!
- They care about: % in categories, not individual classification
- Statistical problem: use labeled set to estimate %s in unlabeled set

Social Media Analytics

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy?

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%?

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%?
  ($\approx$ the accuracy of Google or Bing)

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX $\rightsquigarrow$ biased!

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX $\rightsquigarrow$ biased!
  - Spam detectors are tuned for you: You're more annoyed missing an important email then seeing spam in your INBOX

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX $\rightsquigarrow$ biased!
  - Spam detectors are tuned for you: You're more annoyed missing an important email then seeing spam in your INBOX
  - Removing bias requires adjustment

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
    - 5% don't like Trump because of foreign policy
    - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
    - Quantity of interest: % spam received
    - Estimator: % spam in INBOX $\rightsquigarrow$ biased!
    - Spam detectors are tuned for you: You're more annoyed missing an important email then seeing spam in your INBOX
    - Removing bias requires adjustment
- Key insight:

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX $\rightsquigarrow$ biased!
  - Spam detectors are tuned for you: You're more annoyed missing an important email then seeing spam in your INBOX
  - Removing bias requires adjustment
- Key insight: no one cares what @StatPumpkin213 says on Twitter

Social Media Analytics

- Categories: List of 10 reasons to like or not like Donald Trump
- With a hand-coded training set: best classifier accuracy? 60%? ($\approx$ the accuracy of Google or Bing)
- Classify&Count estimate (for example):
  - 5% don't like Trump because of foreign policy
  - Truth is $5\% + 40\% = 45\%$
- Classification: great for Google, useless for some social science
- E.g., Spam in your INBOX
  - Quantity of interest: % spam received
  - Estimator: % spam in INBOX $\rightsquigarrow$ biased!
  - Spam detectors are tuned for you: You're more annoyed missing an important email then seeing spam in your INBOX
  - Removing bias requires adjustment
- Key insight: no one cares what @StatPumpkin213 says on Twitter
- Statistical problem: use labeled set to estimate %s in unlabeled set

The Statistical Problem

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining:

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification,"

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation,"

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting,"

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation,"

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set
  - Classification: category label for every document

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set
  - Classification: category label for every document
  - Our goal: % in each category

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set
  - Classification: category label for every document
  - Our goal: % in each category
- Challenges

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set
  - Classification: category label for every document
  - Our goal: % in each category
- Challenges
  - Unlabeled & labeled sets: not random samples from same population

The Statistical Problem

Estimating Aggregate Percentages: Distinguishing Characteristic of Social Science

- Names
  - In epidemiology: "prevalence estimation"
  - In computer science, machine learning, computational linguistics, and data mining: "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance."
- Given:
  - C mutually exclusive and exhaustive categories
  - Labeled set: documents with category labels (usually by hand, from one time or place)
  - One or more unlabeled sets from subsequent days or places
- Quantities of interest for each unlabeled set
  - Classification: category label for every document
  - Our goal: % in each category
- Challenges
  - Unlabeled & labeled sets: not random samples from same population
  - Unlabeled sets can change over time in unanticipated ways

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)
- crimson hexagon ⇝ 

⇝ **Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media**

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)
- crimson hexagon ⇝ 

⇝

Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named

- Worldwide cause-of-death estimates for

World Health Organization

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)
- crimson hexagon $\rightsquigarrow$ 

$\rightsquigarrow$ 
Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named
- Worldwide cause-of-death estimates for World Health Organization
- Open source software:

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)
- crimson hexagon $\rightsquigarrow$ 

$\rightsquigarrow$ 
Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named
- Worldwide cause-of-death estimates for World Health Organization
- Open source software: *VA: Verbal Autopsy Software*

Prior work: The Only Multicategory Method w/o Classification

- "Verbal Autopsy Methods with Multiple Causes of Death." (King & Lu, *Statistical Science*, 2008)
- "Designing Verbal Autopsy Studies" (King, Lu, & Shibuya, *Population Health Metrics*, 2010)
- "A Method of Automated Nonparametric Content Analysis for Social Science" (Hopkins & King, *AJPS*, 2010)
- U.S. Patent 8180717 (Hopkins, King, Lu, 2012)
- crimson hexagon ⇝ 

⇝ 
Fast Company Names Crimson Hexagon Number Seven on "The 10 Most Innovative Companies in Web" List Leading Social Intelligence Firm Recognized For Revolutionary Measurement of Consumer Opinions in Social Media

Published: Wednesday, 16 Mar 2011 | 9:20 AM ET Text Size

CAMBRIDGE, Mass., Mar 16, 2011 (BUSINESS WIRE) -- Fast Company named
- Worldwide cause-of-death estimates for World Health Organization
- Open source software: *VA: Verbal Autopsy Software* and *Readme: Software for Automated Content Analysis*

%'s Can be Estimated Better than "Classify & Count"

%'s Can be Estimated Better than "Classify & Count"

- Start with bad estimator $P(\widehat{D = 1})$, from classification or otherwise

%'s Can be Estimated Better than "Classify & Count"

- Start with bad estimator $P(\widehat{D=1})$, from classification or otherwise
- Decompose bad estimator with accounting identity

$$P(\widehat{D=1}) = \text{(sens)}\ \color{red}{P(D=1)} + (1-\text{spec})\ P(D=2)$$

%'s Can be Estimated Better than "Classify & Count"

- Start with bad estimator $\widehat{P(D=1)}$, from classification or otherwise
- Decompose bad estimator with accounting identity

$$\widehat{P(D=1)} = \text{(sens)}\ {\color{red}P(D=1)} + (1-\text{spec})\ P(D=2)$$

- Sensitivity: Prob(true positive)

%'s Can be Estimated Better than "Classify & Count"

- Start with bad estimator $\widehat{P(D=1)}$, from classification or otherwise
- Decompose bad estimator with accounting identity

$$\widehat{P(D=1)} = \text{(sens)}\; P(D=1) + (1-\text{spec})\; P(D=2)$$

- Sensitivity: Prob(true positive)
- 1 − Specificity: Prob(false negative)

%'s Can be Estimated Better than "Classify & Count"

- Start with bad estimator $\widehat{P(D=1)}$, from classification or otherwise
- Decompose bad estimator with accounting identity

$$\widehat{P(D=1)} = \text{(sens)}\ P(D=1) + \ (1-\text{spec})\ P(D=2)$$

  - Sensitivity: Prob(true positive)
  - $1-$ Specificity: Prob(false negative)

- Solve for "truth" to correct estimate:

$$P(D=1) = \frac{\widehat{P(D=1)} - (1-\text{spec})}{\text{sens} - (1-\text{spec})}$$

Generalizations: C Categories, No Individual Classification

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c') \, P(D = c')$$

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c') \; P(D = c')$$

Misclassification Probs

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c')\ P(D = c')$$

Misclassification Probs

- New accounting identity:

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c')\ P(D = c')$$

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D}=c) = \sum_{c'=1}^{C} P(\hat{D}=c|D=c')\; P(D=c')$$

Misclassification Probs

- New accounting identity:

$$P(S=s) \; = \sum_{c'=1}^{C} P(S=s|D=c')\; P(D=c')$$

Word stem profiles

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c')\ P(D = c')$$

Misclassification Probs

- New accounting identity:

$$P(S = s) \ = \sum_{c'=1}^{C} P(S = s | D = c')\ P(D = c')$$

Word stem profiles

- $S = s$: Deterministic profile of document

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c|D = c')\ P(D = c')$$

Misclassification Probs

- New accounting identity:

$$P(S = s)\ = \sum_{c'=1}^{C} P(S = s|D = c')\ P(D = c')$$

Word stem profiles

- $S = s$: Deterministic profile of document (Many options!)

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c')\ P(D = c')$$

Misclassification Probs

- New accounting identity:

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c')\ P(D = c')$$

Word stem profiles

- $S = s$: Deterministic profile of document (Many options!)
- $P(S = s)$: proportion of documents in profile s

Generalizations: C Categories, No Individual Classification

- Accounting identity for C categories:

$$P(\hat{D} = c) = \sum_{c'=1}^{C} P(\hat{D} = c | D = c')\ P(D = c')$$

Misclassification Probs

- New accounting identity:

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c')\ P(D = c')$$

Word stem profiles

Word stem profiles by category

- $S = s$: Deterministic profile of document (Many options!)
- $P(S = s)$: proportion of documents in profile s

Matrix Simplifications

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c')\, P(D = c')$$

Matrix Simplifications

$$P(S = s) = \sum_{c'=1}^{C} P(S = s|D = c')\, P(D = c')$$

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

Matrix Simplifications

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c') \, P(D = c')$$

$$\underset{W \times 1}{P(S)} = \underset{W \times C}{P(S|D)} \; \underset{C \times 1}{P(D)}$$

- Word stem profile

Matrix Simplifications

$$P(S = s) = \sum_{c'=1}^{C} P(S = s|D = c')\ P(D = c')$$

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category

Matrix Simplifications

$$P(S = s) = \sum_{c'=1}^{C} P(S = s | D = c')\ P(D = c')$$

$$\underset{W \times 1}{P(S)} = \underset{W \times C}{P(S|D)}\ \underset{C \times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved;

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved; make assumption: $P(S|D)^{\mathrm{L}} = P(S|D)^{\mathrm{U}}$

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved; make assumption: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}}$
  - W is too large;

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved; make assumption: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}}$
  - W is too large; take random subsamples and average

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)}\ \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved; make assumption: $P(S|D)^{\mathrm{L}} = P(S|D)^{\mathrm{U}}$
  - W is too large; take random subsamples and average
  - $P(D) \equiv \beta$ is on the simplex;

Matrix Simplifications

$$\underset{W\times 1}{P(S)} = \underset{W\times C}{P(S|D)} \; \underset{C\times 1}{P(D)}$$

- Word stem profile
- Word stem profile by category
- Quantity of interest, category proportions
- Alternative notation: $Y = X\beta$
- Solve for quantity of interest: $\beta = (X'X)^{-1}X'Y$
- The readme estimator:
  - $P(S|D) \equiv X$ is unobserved; make assumption: $P(S|D)^{\mathrm{L}} = P(S|D)^{\mathrm{U}}$
  - W is too large; take random subsamples and average
  - $P(D) \equiv \beta$ is on the simplex; use constrained LS

Statistical Properties

Statistical Properties

Assumptions

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}}$

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Statistical Properties

Assumptions

- Classification: $P(S,D)^{\mathsf{L}} = P(S,D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

- Like regression with random measurement error in X ($\rightsquigarrow$ attenuation)

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

- Like regression with random measurement error in X ($\rightsquigarrow$ attenuation)
- Unlike regression, it's Consistent:
  $\operatorname*{plim}_{n\to\infty} P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}}$

Statistical Properties

Assumptions

- Classification: $P(S,D)^{\mathsf{L}} = P(S,D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

- Like regression with random measurement error in X ($\rightsquigarrow$ attenuation)
- Unlike regression, it's Consistent:

$$\operatorname*{plim}_{n\to\infty} P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \quad \rightsquigarrow \quad \operatorname*{plim}_{n\to\infty} \widehat{P(D)} = P(D)$$

Statistical Properties

Assumptions

- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

- Like regression with random measurement error in X ($\rightsquigarrow$ attenuation)
- Unlike regression, it's Consistent:
  $$\operatorname*{plim}_{n\to\infty} P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \quad \rightsquigarrow \quad \operatorname*{plim}_{n\to\infty} \widehat{P(D)} = P(D)$$
- But it's biased: $E[\widehat{P(D)}] \neq P(D)$

Statistical Properties

Assumptions

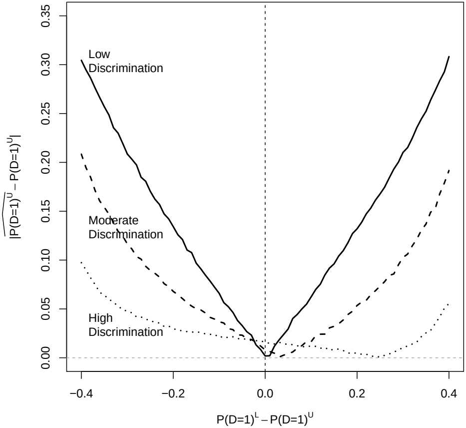
- Classification: $P(S, D)^{\mathsf{L}} = P(S, D)^{\mathsf{U}}$ (& measure all predictors!)
- Readme: $P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \rightsquigarrow \widehat{P(D)} = P(D)$
- Alternative DGP: labeled is random, unlabeled is fixed population
- Readme2: $E[P(S|D)^{\mathsf{L}}] = P(S|D)^{\mathsf{U}}$

Properties

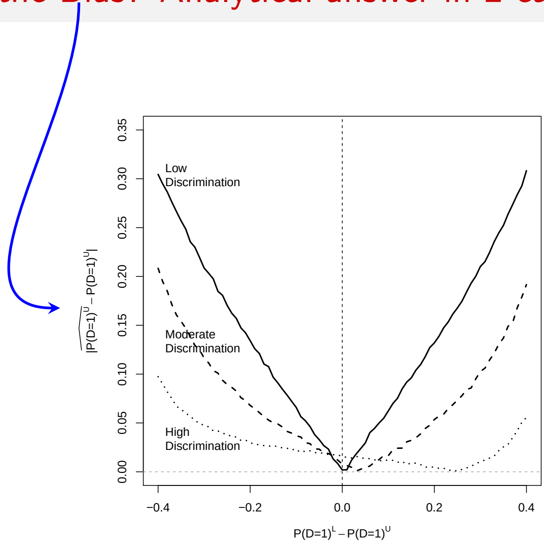
- Like regression with random measurement error in X ($\rightsquigarrow$ attenuation)
- Unlike regression, it's Consistent:
  $$\operatorname*{plim}_{n\to\infty} P(S|D)^{\mathsf{L}} = P(S|D)^{\mathsf{U}} \quad \rightsquigarrow \quad \operatorname*{plim}_{n\to\infty} \widehat{P(D)} = P(D)$$
- But it's biased: $E[\widehat{P(D)}] \neq P(D) \rightsquigarrow$ attenuation toward $P(D)^{\mathsf{L}}$

Where's the Bias? Analytical answer in 2 categories

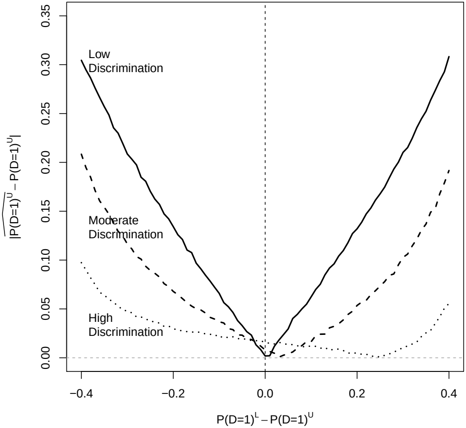
Where's the Bias? Analytical answer in 2 categories



Where's the Bias? Analytical answer in 2 categories

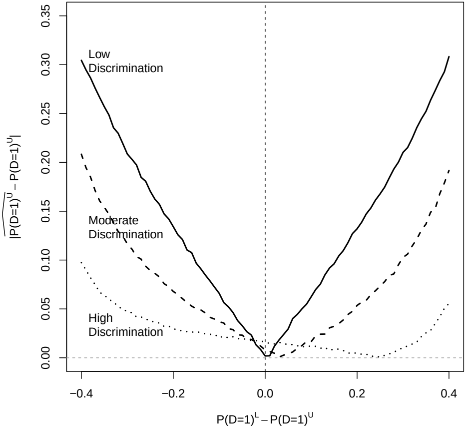


Where's the Bias? Analytical answer in 2 categories



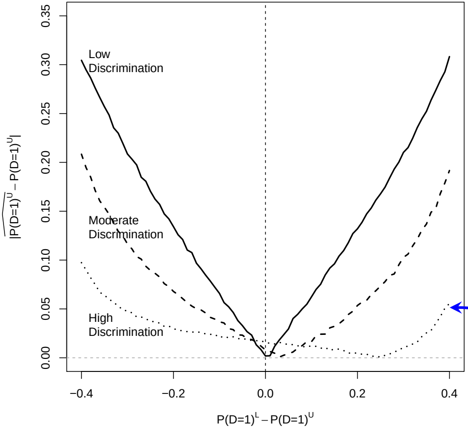
Try to:

Where's the Bias? Analytical answer in 2 categories



Try to: Reduce $P(D)$ divergence;

Where's the Bias? Analytical answer in 2 categories



Try to: Reduce $P(D)$ divergence; Increase $P(S|D)$ discrimination

A Proposed Readme2: Part 1 (of 2)

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"):

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}}$

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} \rightsquigarrow$ but it's sparse

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} \rightsquigarrow$ but it's sparse $\rightsquigarrow$ weights are too variable

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} \rightsquigarrow$ but it's sparse $\rightsquigarrow$ weights are too variable
  - We prove $\frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} = f(\text{Propensity score})$ (of labeled v. unlabeled set)

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \dfrac{P(D_\ell)^{\mathrm{U}}}{P(D_\ell)^{\mathrm{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \dfrac{P(S_\ell)^{\mathrm{U}}}{P(S_\ell)^{\mathrm{L}}} \leadsto$ but it's sparse $\leadsto$ weights are too variable
  - We prove $\dfrac{P(S_\ell)^{\mathrm{U}}}{P(S_\ell)^{\mathrm{L}}} = f(\text{Propensity score})$ (of labeled v. unlabeled set) $\leadsto$ Use PScore to smooth

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathsf{U}}}{P(D_\ell)^{\mathsf{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} \rightsquigarrow$ but it's sparse $\rightsquigarrow$ weights are too variable
  - We prove $\frac{P(S_\ell)^{\mathsf{U}}}{P(S_\ell)^{\mathsf{L}}} = f$(Propensity score) (of labeled v. unlabeled set) $\rightsquigarrow$ Use PScore to smooth
  - To increase discrimination, form propensity score using "important" words

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\text{U}}}{P(D_\ell)^{\text{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\text{U}}}{P(S_\ell)^{\text{L}}} \rightsquigarrow$ but it's sparse $\rightsquigarrow$ weights are too variable
  - We prove $\frac{P(S_\ell)^{\text{U}}}{P(S_\ell)^{\text{L}}} = f(\text{Propensity score})$ (of labeled v. unlabeled set)
    $\rightsquigarrow$ Use PScore to smooth
  - To increase discrimination, form propensity score using "important" words (with two lasso-regularized multivariate logistic models);

A Proposed Readme2: Part 1 (of 2)

- Goal: reduce $P(D)$ divergence, increase $P(S|D)$ discrimination
- Start with readme
- Add bootstrap aggregating ("bagging"): Improve stability, accuracy
- Add *Weighted* bagging:
  - Ideal (unavailable) weights to reduce divergence: $p_\ell \propto \frac{P(D_\ell)^{\mathrm{U}}}{P(D_\ell)^{\mathrm{L}}}$
  - We're estimating β in $Y = X\beta$ & know the true Y in the test set!
  - Weight on $p_\ell \propto \frac{P(S_\ell)^{\mathrm{U}}}{P(S_\ell)^{\mathrm{L}}} \rightsquigarrow$ but it's sparse $\rightsquigarrow$ weights are too variable
  - We prove $\frac{P(S_\ell)^{\mathrm{U}}}{P(S_\ell)^{\mathrm{L}}} = f(\text{Propensity score})$ (of labeled v. unlabeled set) $\rightsquigarrow$ Use PScore to smooth
  - To increase discrimination, form propensity score using "important" words (with two lasso-regularized multivariate logistic models); same logic as balancing for causal inference

A Proposed Readme2: Part 2 (of 2)

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
    - Most words have little effect

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
  - Most words have little effect
  - If no effect, $P(S|D) = P(S)$

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
  - Most words have little effect
  - If no effect, $P(S|D) = P(S)$
  - $\rightsquigarrow$ Shrink $P(S|D)$ toward prior of $P(S)^{\mathrm{U}}$

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
  - Most words have little effect
  - If no effect, $P(S|D) = P(S)$
  - $\rightsquigarrow$ Shrink $P(S|D)$ toward prior of $P(S)^{\mathsf{U}}$
  - (Details: Beta-binomial Bayesian model for cell counts)

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
  - Most words have little effect
  - If no effect, $P(S|D) = P(S)$
  - $\rightsquigarrow$ Shrink $P(S|D)$ toward prior of $P(S)^{\mathrm{U}}$
  - (Details: Beta-binomial Bayesian model for cell counts)
- Overall method: weighted bagging + PScore + Bayesian shrinkage

A Proposed Readme2: Part 2 (of 2)

- Form $P(S|D)$ by tabulating weighted bootstrap labeled set
- But $P(S|D)$ is sparse (for each bootstrapped sample)
- Use Bayesian model: mitigate sparseness, increase efficiency
  - Most words have little effect
  - If no effect, $P(S|D) = P(S)$
  - $\rightsquigarrow$ Shrink $P(S|D)$ toward prior of $P(S)^{\mathsf{U}}$
  - (Details: Beta-binomial Bayesian model for cell counts)
- Overall method: weighted bagging + PScore + Bayesian shrinkage
- Refinements: alternative numeric representations of text

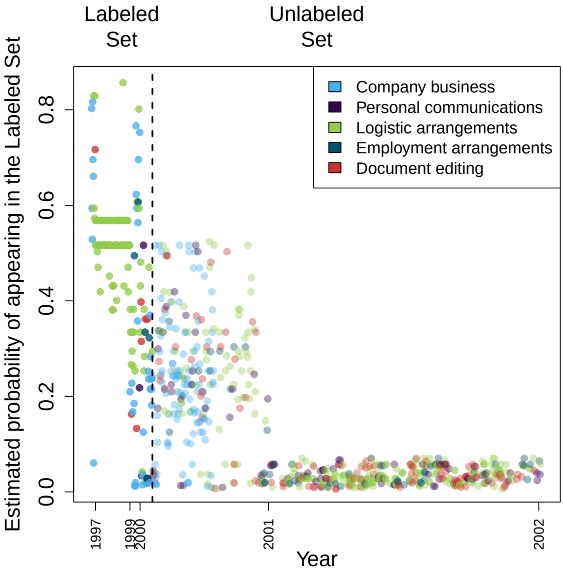
Example with Large $P(D)$ Divergence: Enron Emails

Example with Large $P(D)$ Divergence: Enron Emails

California energy crisis dramatically changes content

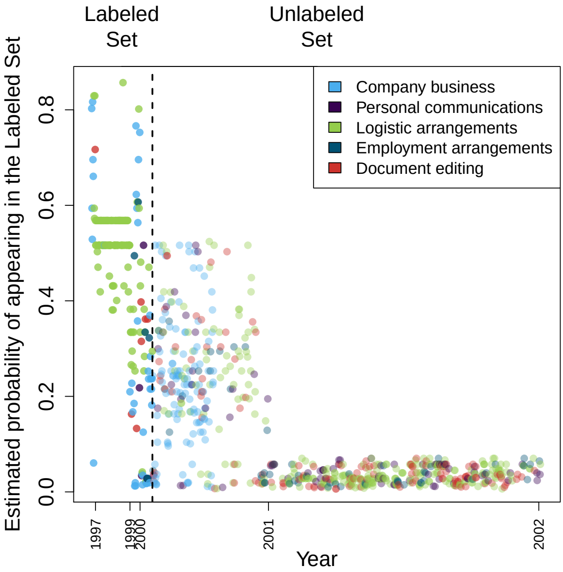
Example with Large $P(D)$ Divergence: Enron Emails

California energy crisis dramatically changes content



Example with Large $P(D)$ Divergence: Enron Emails

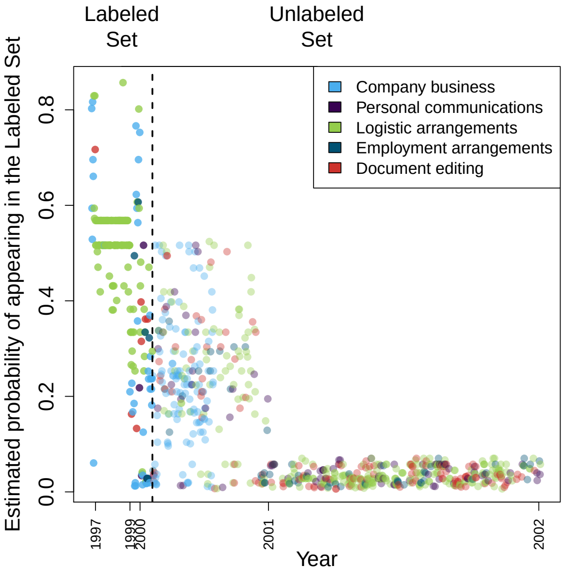
California energy crisis dramatically changes content



- Pscores vary considerably over time by category

Example with Large $P(D)$ Divergence: Enron Emails

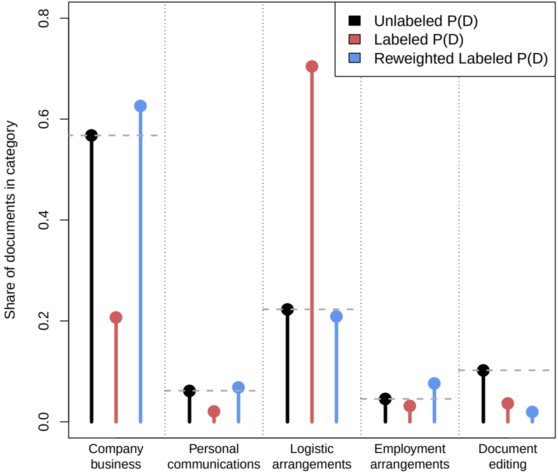
California energy crisis dramatically changes content



- Pscores vary considerably over time by category
- High $P(D)$ divergence

Example with Large $P(D)$ Divergence: Enron Emails

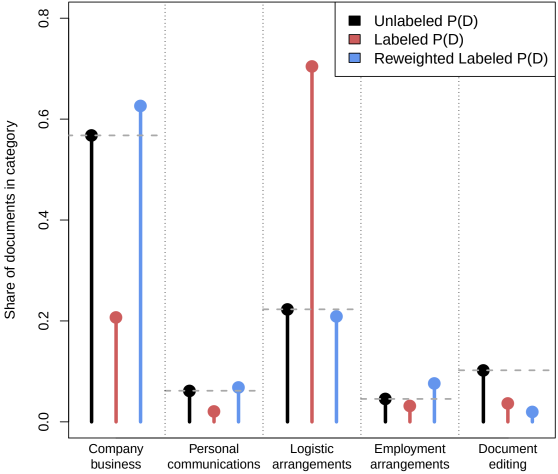
California energy crisis dramatically changes content



- Pscores vary considerably over time by category
- High $P(D)$ divergence

Example with Large $P(D)$ Divergence: Enron Emails

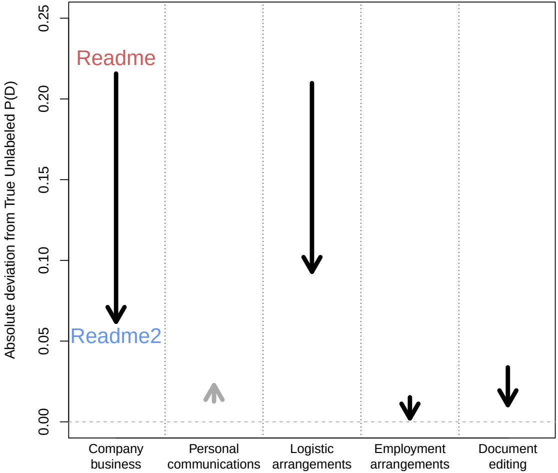
California energy crisis dramatically changes content



- Pscores vary considerably over time by category
- High $P(D)$ divergence
- Weighted bootstrapping eliminates most $P(D)$ divergence

Example with Large $P(D)$ Divergence: Enron Emails

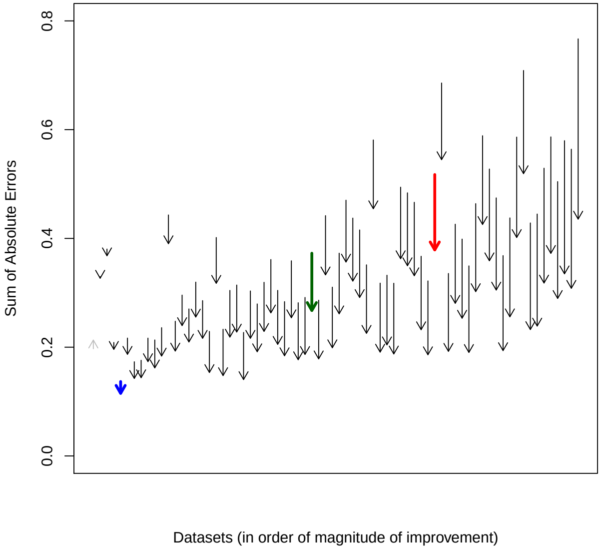
California energy crisis dramatically changes content



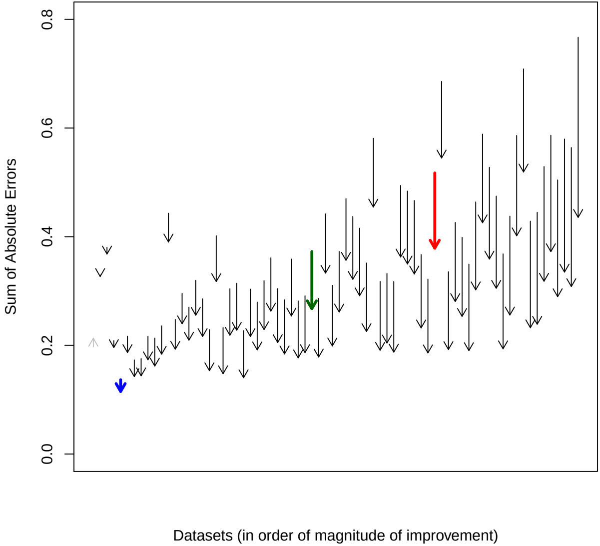
- Pscores vary considerably over time by category
- High $P(D)$ divergence
- Weighted bootstrapping eliminates most $P(D)$ divergence
- Large reduction in estimation error

Validation in 72 Data Sets

Validation in 72 Data Sets

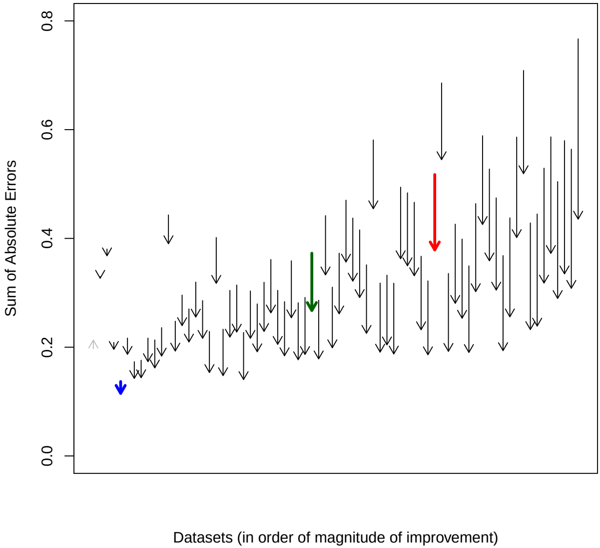


Validation in 72 Data Sets



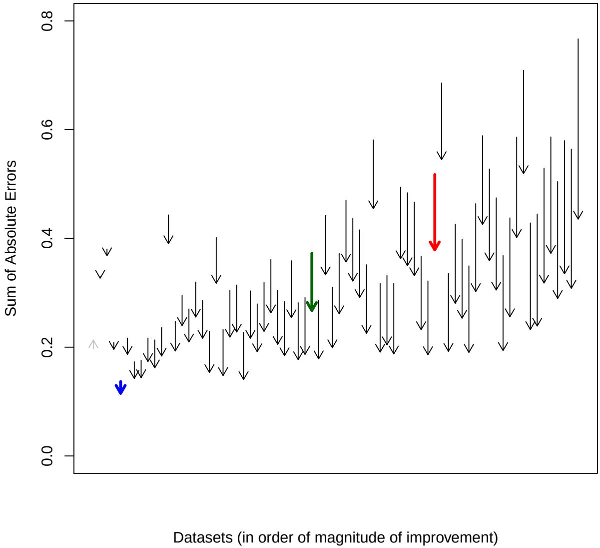
- Enron

Validation in 72 Data Sets



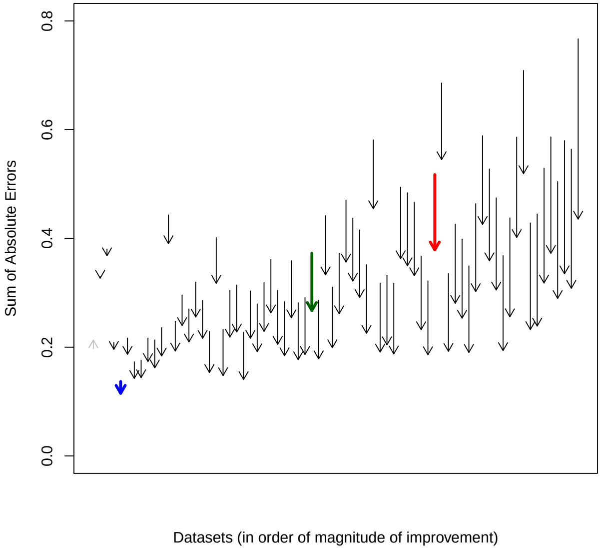
- Enron
- Hillary Clinton (2008)

Validation in 72 Data Sets



- Enron
- Hillary Clinton (2008)
- Immigration blogs

Validation in 72 Data Sets



- Enron
- Hillary Clinton (2008)
- Immigration blogs
- 69 Twitter data sets created by firms, governments, candidates, nonprofits, etc.

Conclusions

Conclusions

- Social science: about population aggregates,

Conclusions

- Social science: about population aggregates,
  not individual classification

Conclusions

- Social science: about population aggregates,
  not individual classification
- Estimate the quantity of interest;

Conclusions

- Social science: about population aggregates,
  not individual classification
- Estimate the quantity of interest;
  beware of adapting tools from strangers (with other goals)

Conclusions

- Social science: about population aggregates,
  not individual classification
- Estimate the quantity of interest;
  beware of adapting tools from strangers (with other goals)
- Readme $\rightsquigarrow$ Readme2

Conclusions

- Social science: about population aggregates,
  not individual classification
- Estimate the quantity of interest;
  beware of adapting tools from strangers (with other goals)
- Readme $\rightsquigarrow$ Readme2
  (software coming)

Conclusions

- Social science: about population aggregates,
  not individual classification
- Estimate the quantity of interest;
  beware of adapting tools from strangers (with other goals)
- Readme $\rightsquigarrow$ Readme2
  (software coming)

For more information:
GaryKing.org