Big Data for Public Policy Analysis

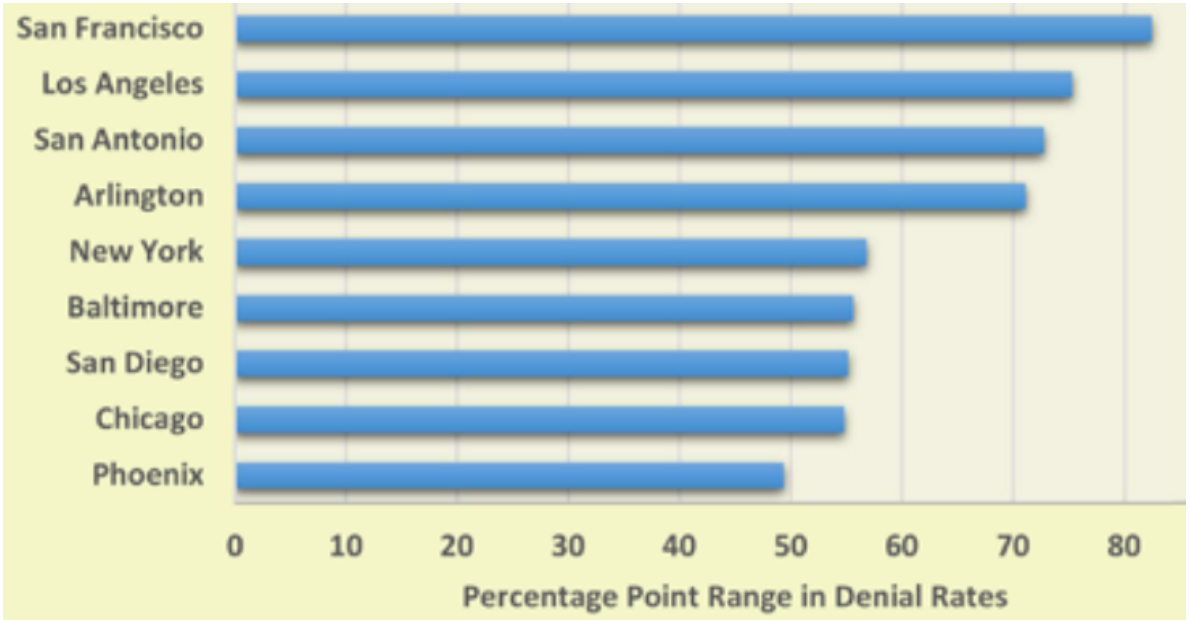
Instructors: Elliott Ash and Malka Guillot

1. Course Overview and Introduction

(Big) Data can diagnose policy problems.

(Big) Data can diagnose (and hopefully help solve) policy problems.

U.S. Asylum Courts: Disparities in Grant Rates



- In San Francisco, one judge grants 90.6% of asylum requests, while another judge grants just 2.9%!

Jailing Decisions Before/After Lunch Breaks

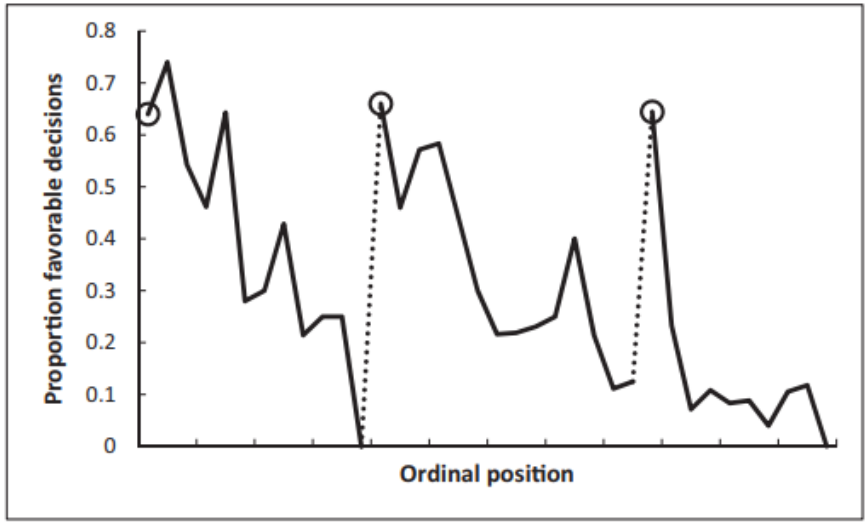
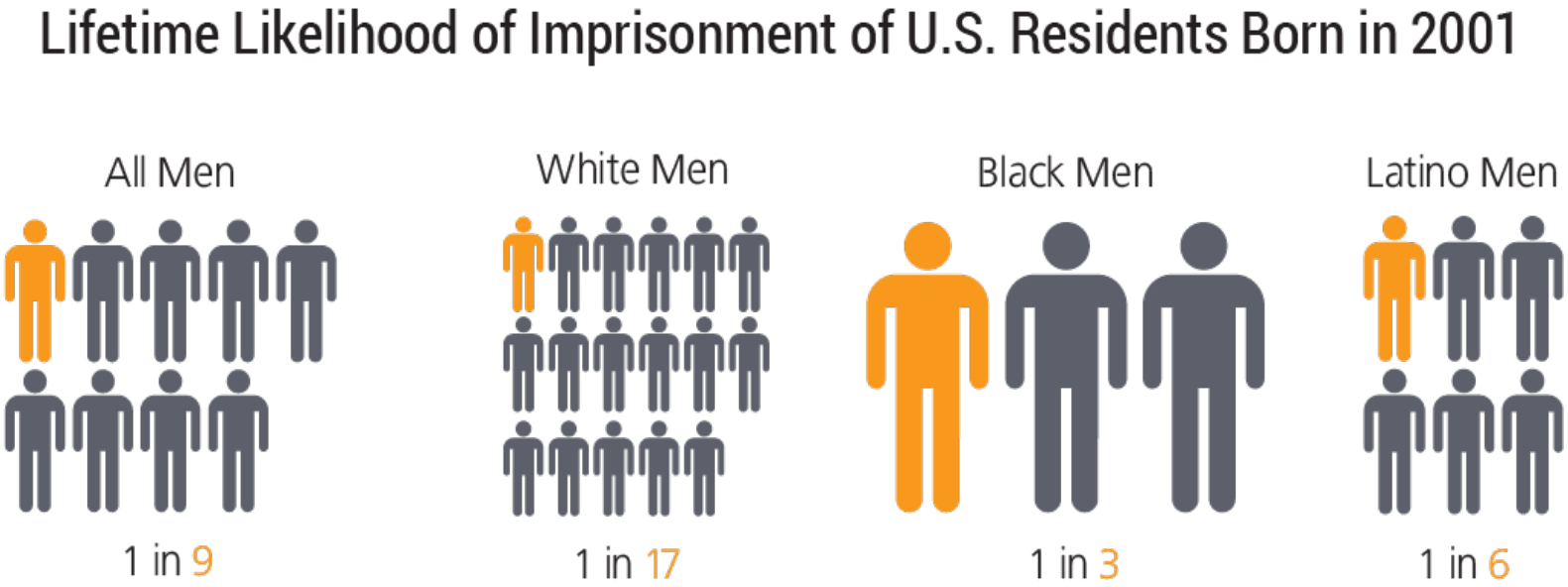


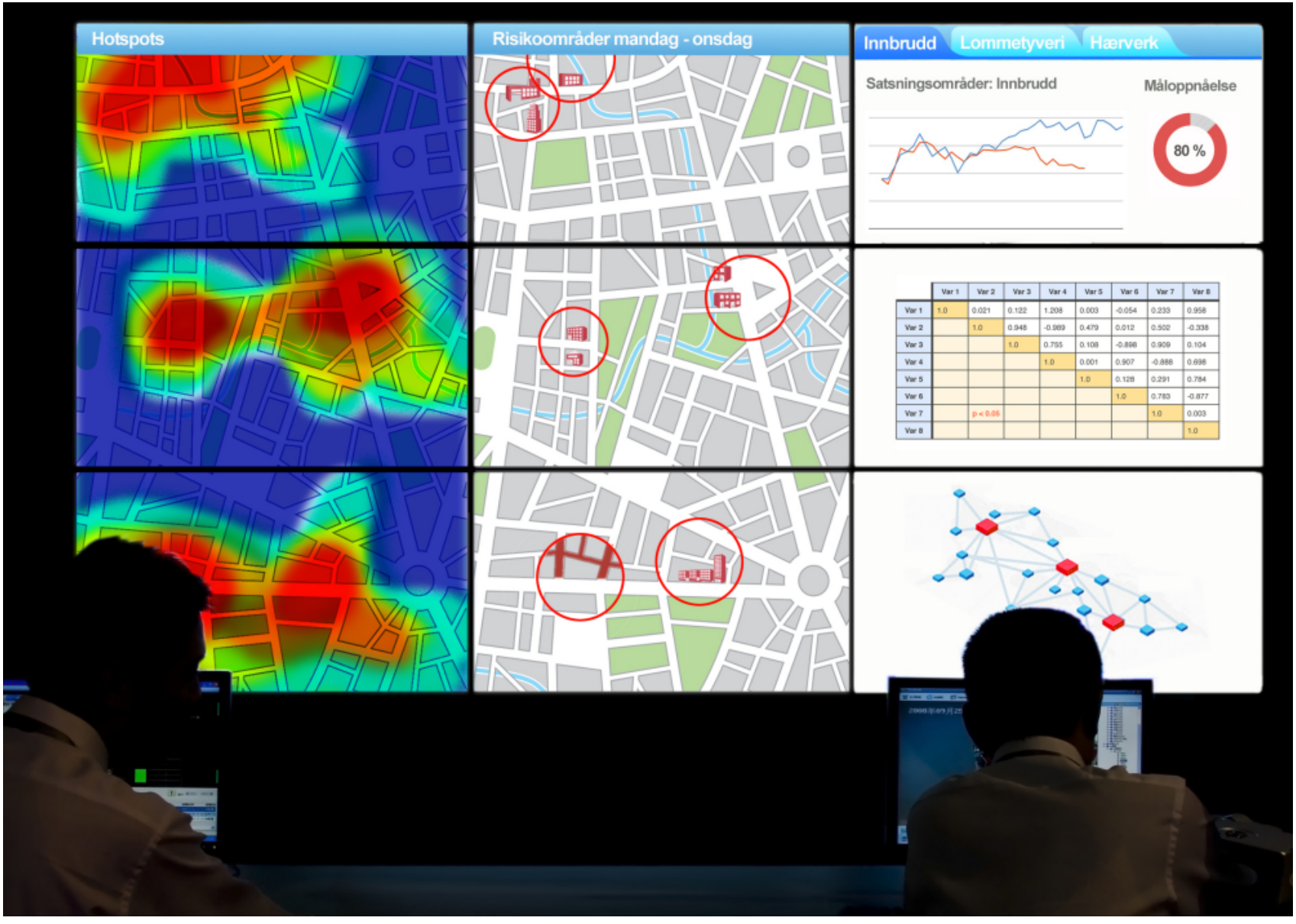
Fig. 1. Proportion of rulings in favor of the prisoners by ordinal position. Circled points indicate the first decision in each of the three decision sessions; tick marks on *x* axis denote every third case; dotted line denotes food break. Because unequal session lengths resulted in a low number of cases for some of the later ordinal positions, the graph is based on the first 95% of the data from each session.

Source: Danziger et al, PNAS 2011, Israel judges deciding on parole.



Source: The Sentencing Project.

Predictive Policing



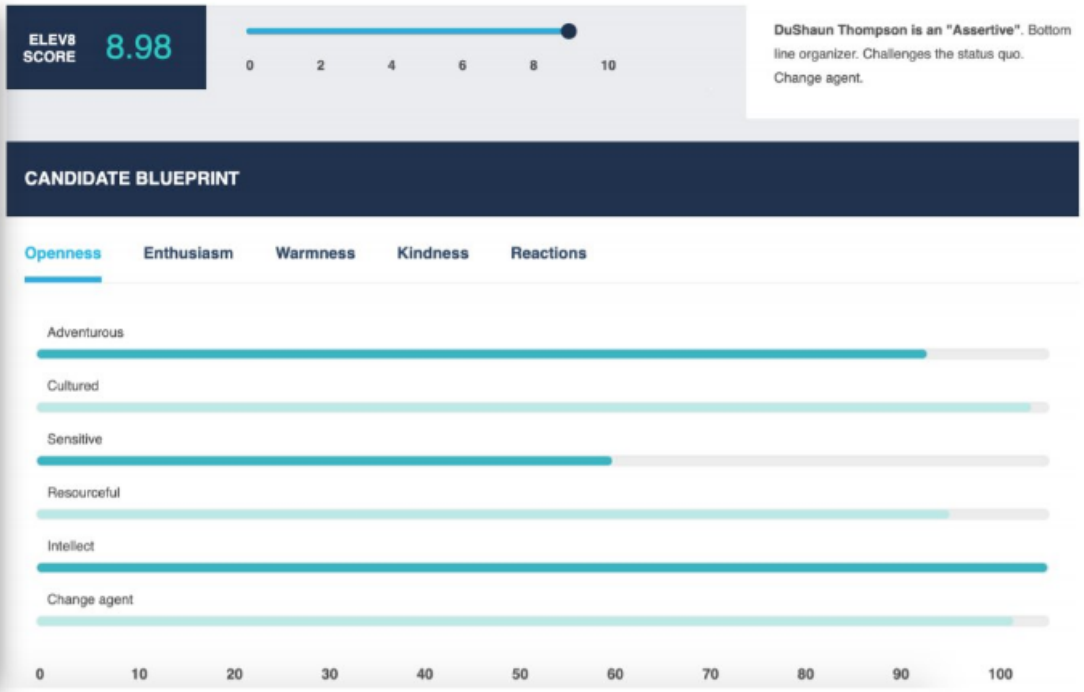
(Big) Data can cause (or magnify) problems.

Predictive policing poses discrimination risk, thinktank warns

Machine-learning algorithms could replicate or amplify bias on race, sexuality and age

▲ One officer said human biases including more stop and searches of black men were likely to be introduced into algorithm data sets. Photograph: Carl Court/Getty Images

Assessing personality & job suitability from 30-second video



Source: Narayanan slides.

OPENAI'S NEW MULTITALENTED AI WRITES, TRANSLATES, AND SLANDERS

A step forward in AI text-generation that also spells trouble

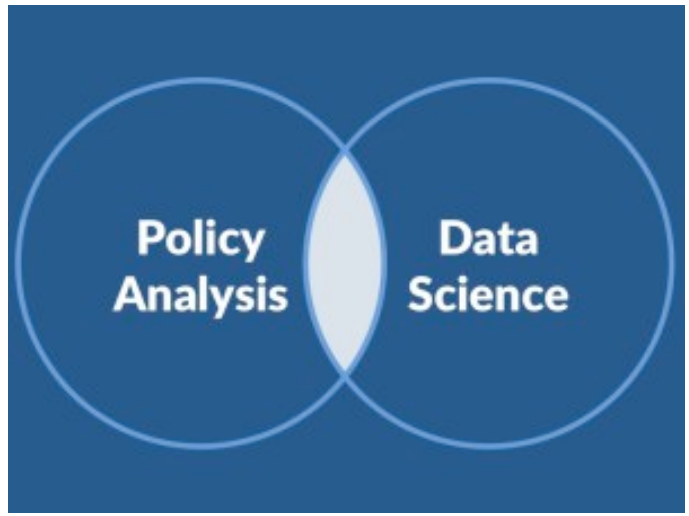
By James Vincent | Feb 14, 2019, 12:00pm EST

Howard, co-founder of Fast.AI agrees. "I've been trying to warn people about this for a while," he says. "We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter."

`https://transformer.huggingface.co/doc/distil-gpt2`

Welcome

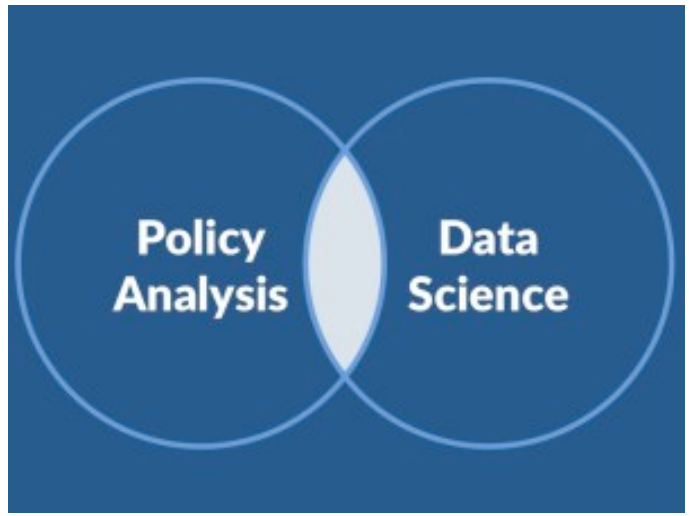
- This course focuses on applications of **big data tools** to **public policy analysis**



- Goals:
  - Equip the standard machine learning toolkit.

Welcome

- This course focuses on applications of **big data tools** to **public policy analysis**



- Goals:
  - Equip the standard machine learning toolkit.
  - Put it to work on a real-world policy project.

Objectives

- What is “Big Data”?

Objectives

- What is "Big Data"?
- What are new and nontraditional sources of data?
    - API's, web scraping, etc.

Objectives

- What is "Big Data"?
- What are new and nontraditional sources of data?
    - API's, web scraping, etc.
- What are the advantages and disadvantages of using large or new sources of data?
    - new problems
    - new solutions

Objectives

- What is "Big Data"?
- What are new and nontraditional sources of data?
  - API's, web scraping, etc.
- What are the advantages and disadvantages of using large or new sources of data?
  - new problems
  - new solutions
- What are the technical constraints of using large data sources?

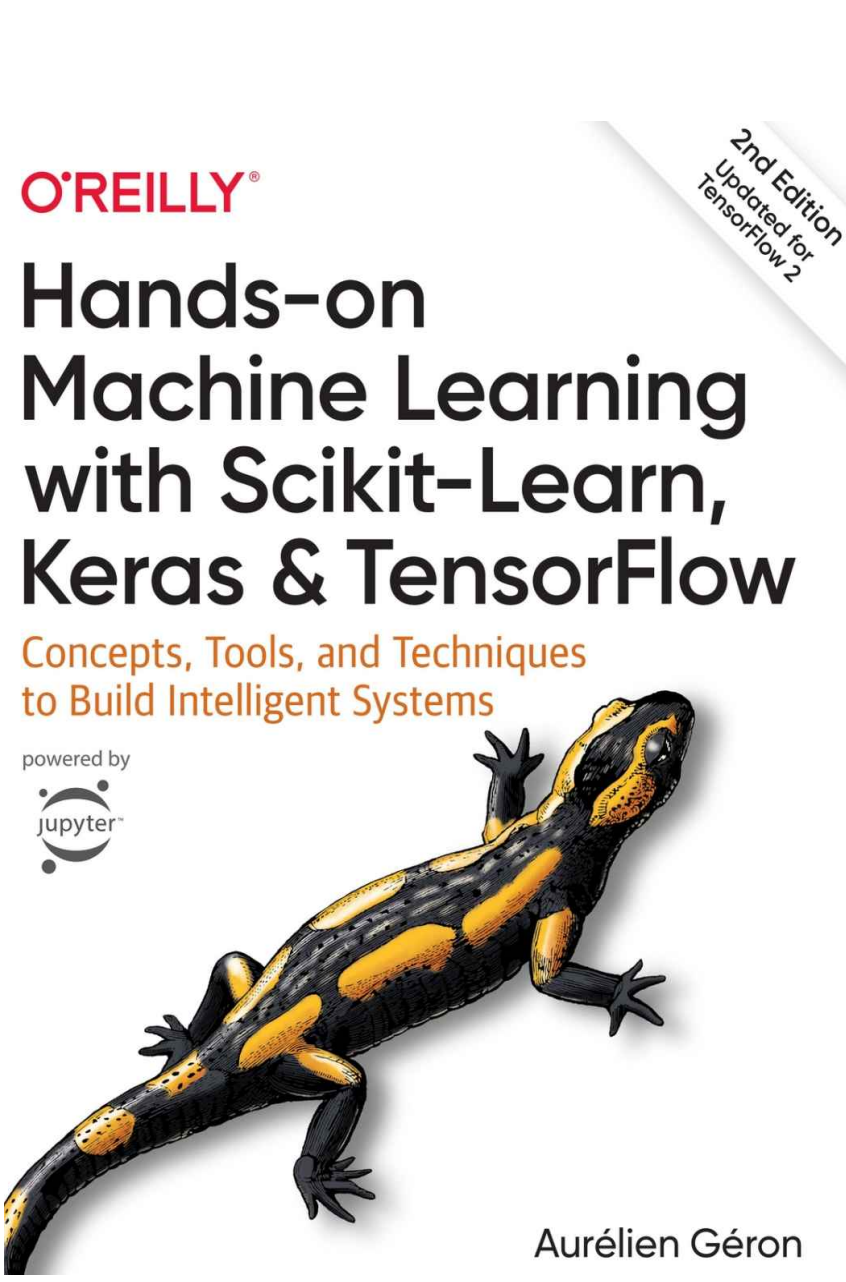
Outline

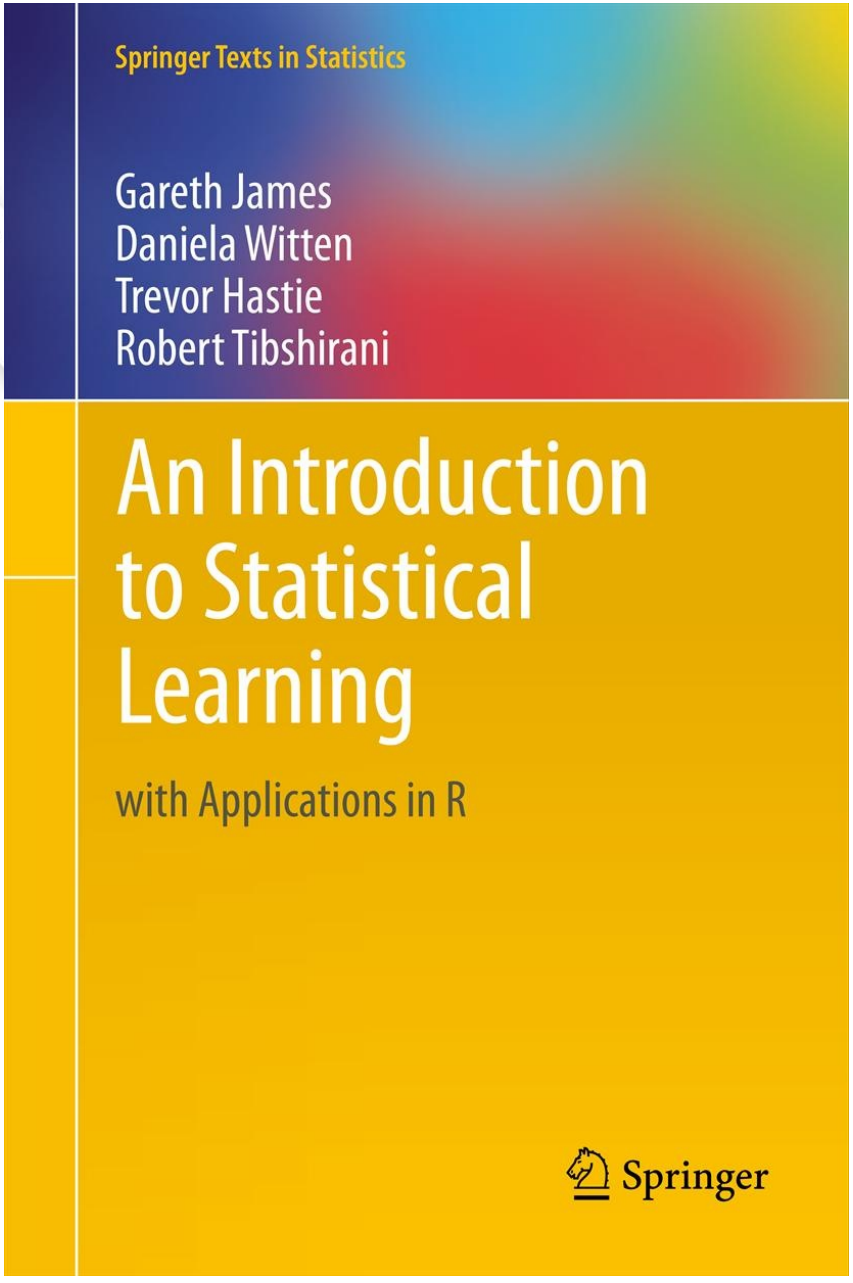
Lecture Times

- Tuesdays, 1:15pm-3pm
  - Location: UNO B11
- Office hours:
  - After lecture, or by appointment (`ashe@ethz.ch` & `guillotm@ethz.ch`).

Online Course Materials

- Course Syllabus:
  - `https://docs.google.com/document/d/1rjo0jNQtEyML0m_PU2kKlUdUzWE942V7eaXqryCyjQQ/edit?ts=5dd7cc31`
- GitHub Repo with Course Materials:
  - `https://github.com/elliottash/big_data_policy_2020`


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer

Programming Languages

- The instructors are most familiar with Python, which is ideal for machine learning and text data.
    - We recommend Anaconda Distribution: `continuum.io/downloads`
- You are welcome to use R instead.

Programming Languages

- The instructors are most familiar with Python, which is ideal for machine learning and text data.
  - We recommend Anaconda Distribution: `continuum.io/downloads`
- You are welcome to use R instead.
- Problem sets should be submitted to assignment dropbox as HTML exports of jupyter notebooks (for Python) or equivalent (e.g. R-Markdown) for R.

Programming Languages

- The instructors are most familiar with Python, which is ideal for machine learning and text data.
  - We recommend Anaconda Distribution: `continuum.io/downloads`
- You are welcome to use R instead.
- Problem sets should be submitted to assignment dropbox as HTML exports of jupyter notebooks (for Python) or equivalent (e.g. R-Markdown) for R.
- See `R-stata-cookbook-for-python.ipynb` for an intro to pandas, plotting and writing R code in python

Three Problem Sets

- Implement major methods in data collection, machine learning and text analysis from class.

Three Problem Sets

- Implement major methods in data collection, machine learning and text analysis from class.
- See due dates on syllabus.
  1. Predicting firm stock price using Twitter data
  2. Crime re-arrest problem
  3. ML / Econometrics Problem

Course Project

- The main course product is a public policy oriented data analysis application.
  - Can be done individually or in groups of 2.
  - In consultation with project advisors, form a research design using methods learned in the course.

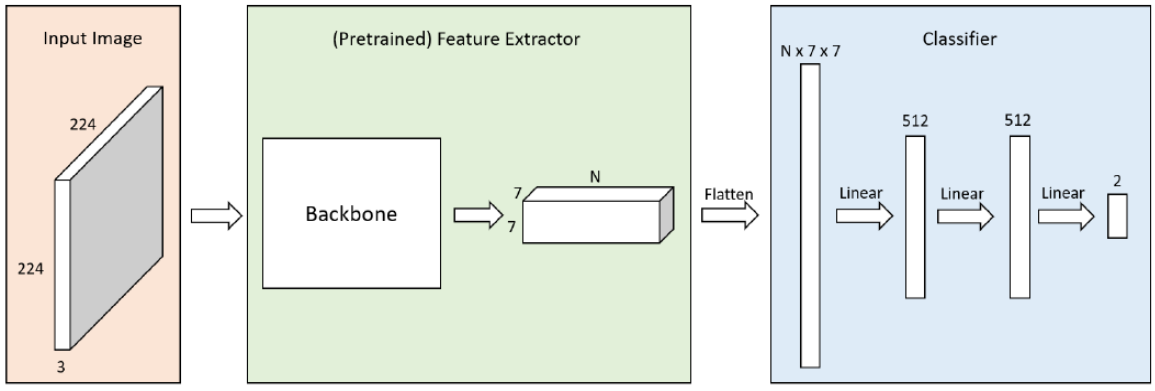
Course Project

- The main course product is a public policy oriented data analysis application.
  - Can be done individually or in groups of 2.
  - In consultation with project advisors, form a research design using methods learned in the course.
- Deliverables: See syllabus

Example Spring 2019 Project

Dominik Borer: Predicting Candidate Party from Political Television Ads

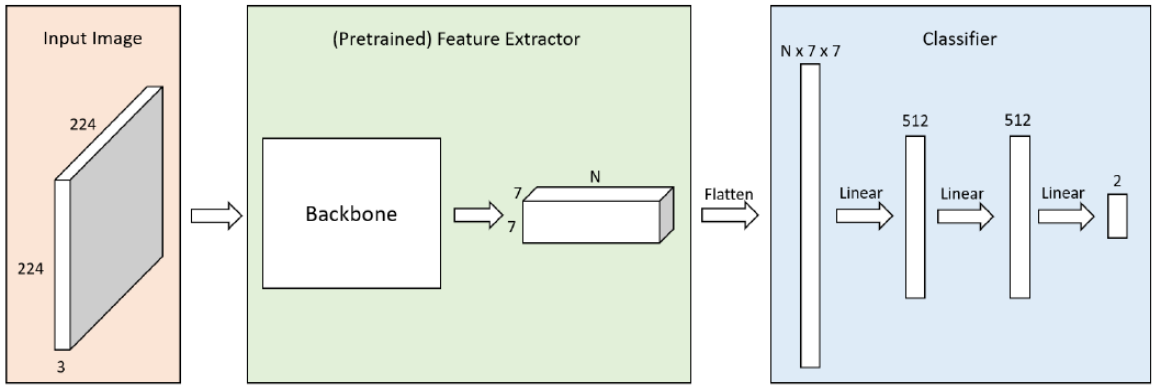
Figure 3: Overview of Model Architecture



Example Spring 2019 Project

Dominik Borer: Predicting Candidate Party from Political Television Ads

Figure 3: Overview of Model Architecture



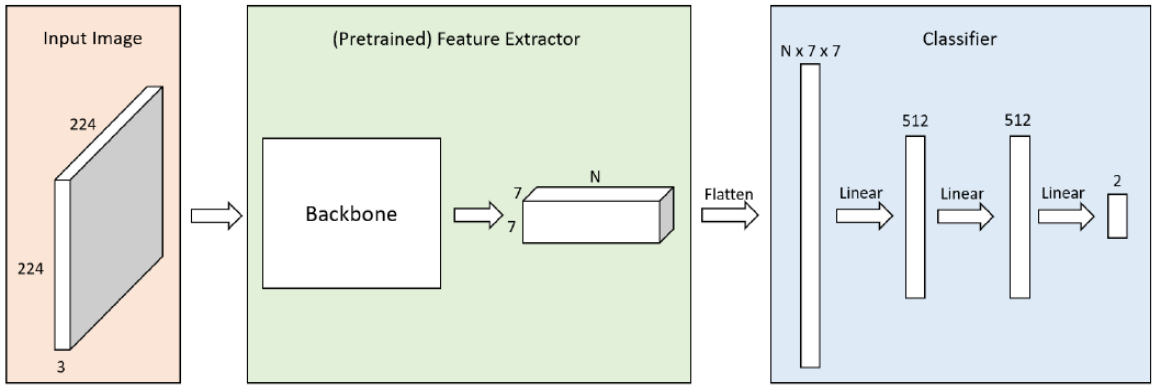
Panel C. Confusion matrix for test set

	Predicted Democratic	Predicted Republican
Actual Democratic	44.88% (793)	7.98% (141)
Actual Republican	15.73% (278)	31.41% (555)

Example Spring 2019 Project

Dominik Borer: Predicting Candidate Party from Political Television Ads

Figure 3: Overview of Model Architecture



Panel C. Confusion matrix for test set

	Predicted Democratic	Predicted Republican
Actual Democratic	44.88% (793)	7.98% (141)
Actual Republican	15.73% (278)	31.41% (555)

Panel A: News Show Images with Highest Democrat Slant

Panel B: News Show Images with Highest Republican Slant

Other Spring 2019 Projects

- One of the groups began building a legal research application for Swiss lawyers:
  - feature-rich legal search engine.
  - MSWord plugin that suggests cite references as you type.
  - Already partnered with Homberger law firm to test it out.

Other Spring 2019 Projects

- One of the groups began building a legal research application for Swiss lawyers:
  - feature-rich legal search engine.
  - MSWord plugin that suggests cite references as you type.
  - Already partnered with Homberger law firm to test it out.
- Another group partnered with a local company to build out their environmental-regulation analytics
  - won an Innosuisse grant.

New Project Ideas

- We have a list of potential project ideas, will share with interested students.

Spirit of the class

- Learning
  - by doing
  - how to learn (i.e. asking your favorite search engine)
- Collaboration (group project...)
- Don't expect us to answer all questions

Outline

Motivations

- We are seeing a **revolution** in policy analysis...
  - **new datasets**: administrative microdata, digitization of text archives, social media
  - **new methods**: causal inference, natural language processing, machine learning

Motivations

- We are seeing a **revolution** in policy analysis...
  - **new datasets**: administrative microdata, digitization of text archives, social media
  - **new methods**: causal inference, natural language processing, machine learning

.

.. which contribute to tackle **forecasting** and **public policy** evaluation with a new angle

Motivations

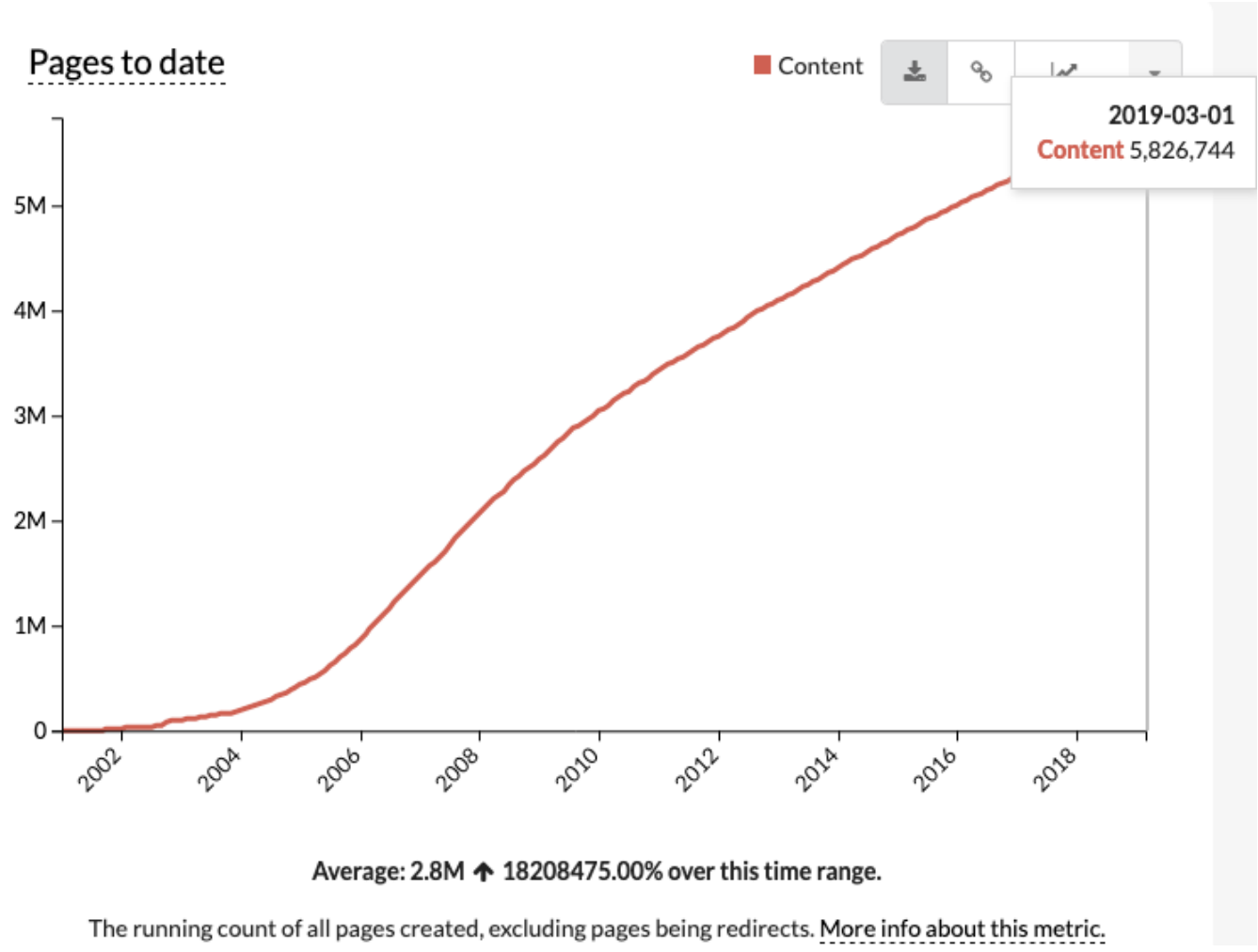
- We are seeing a **revolution** in policy analysis...
  - **new datasets**: administrative microdata, digitization of text archives, social media
  - **new methods**: causal inference, natural language processing, machine learning

.

.. which contribute to tackle **forecasting** and **public policy** evaluation with a new angle

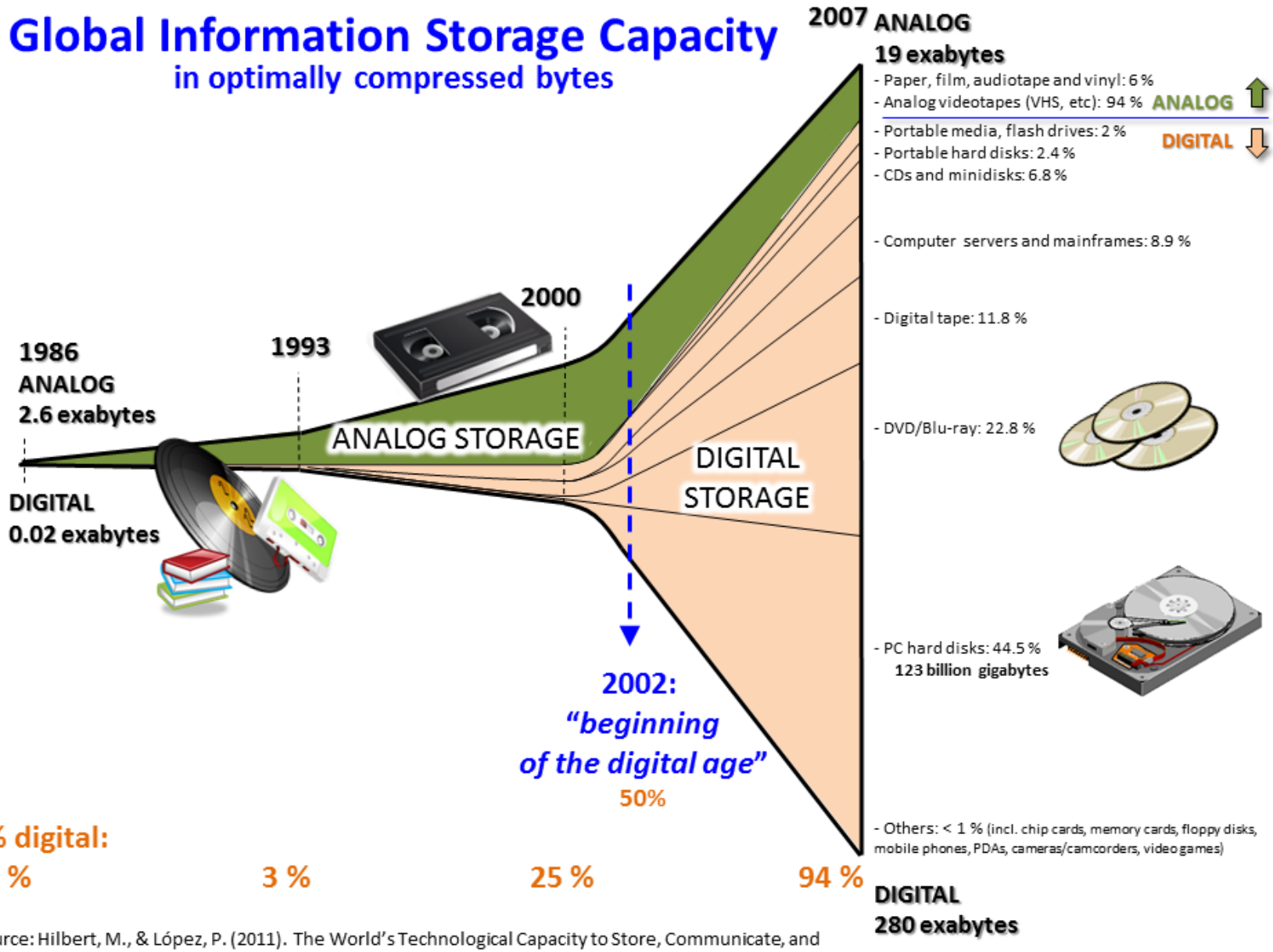
New possibilities: exciting!

of Wikipedia Pages, 2001-2019



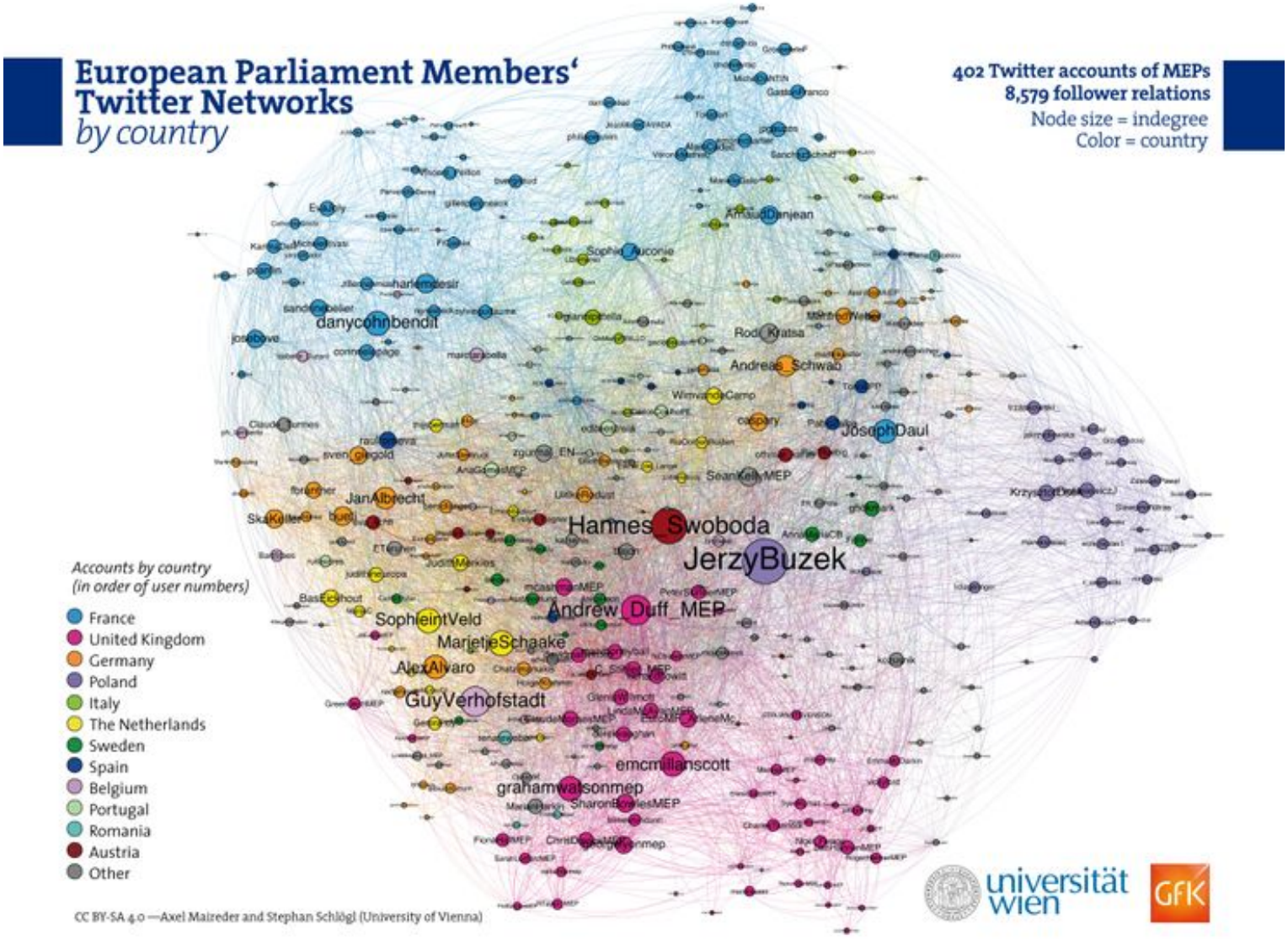
Global Information Storage Capacity

in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

New Data, New Possibilities



What is Big Data?

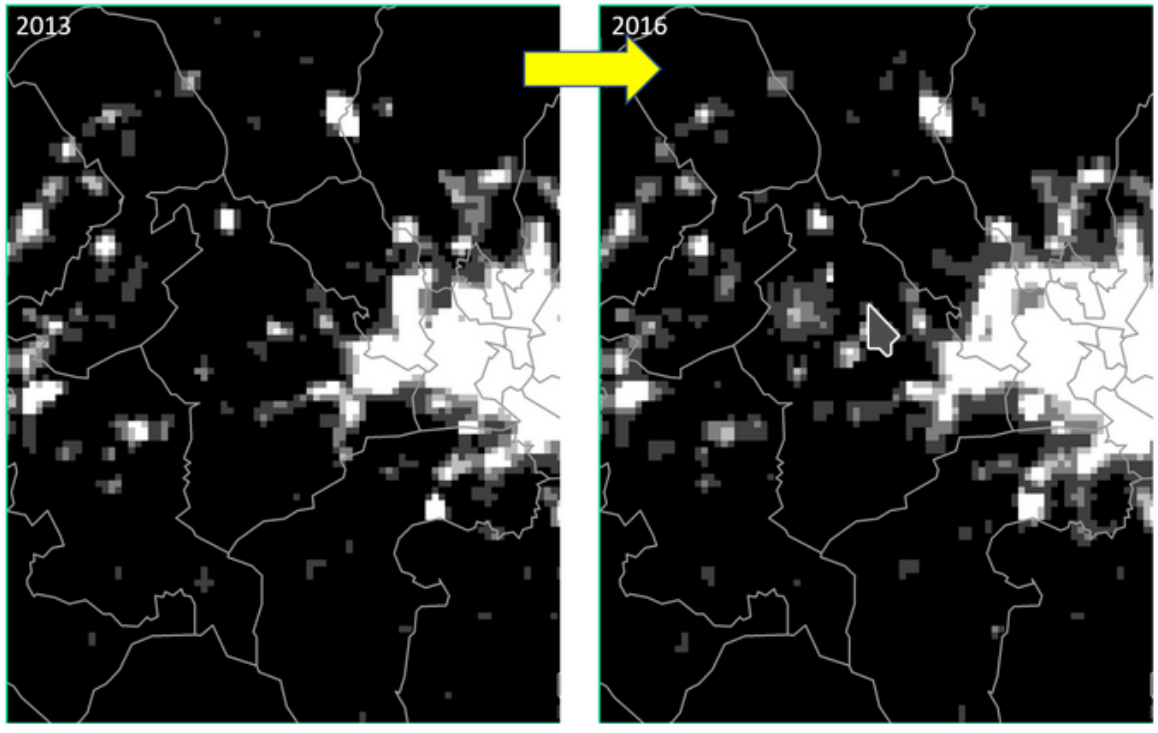
- Volume of data
- Format:
  - Structured
  - Unstructured
- Ready made vs. custommades

$\Rightarrow$ data will be even more important for policy making analyses

New Tools and Methods

- Data collection (eg. APIs, webscraping)
- Analysis (e.g. text analysis, machine learning)
- Visualization (e.g, maps, social networks, web apps)

New Measurements



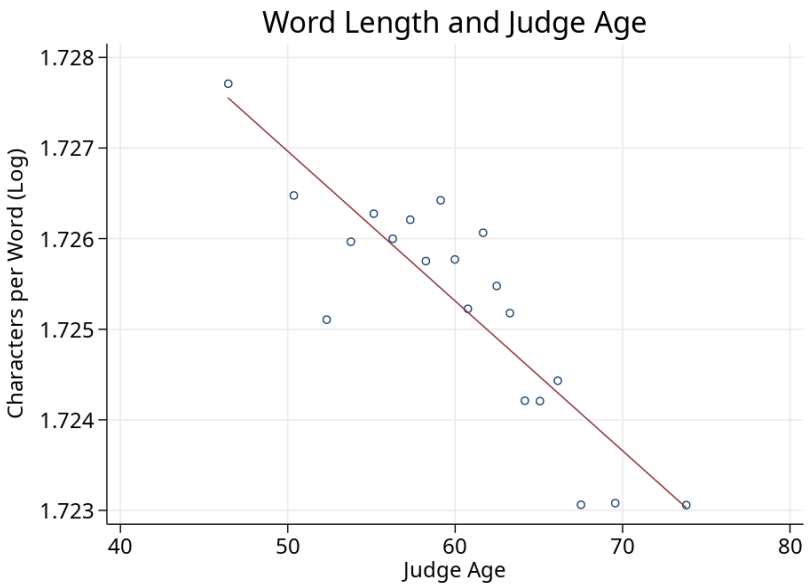
FYR Macedonia: changes in nighttime light intensity

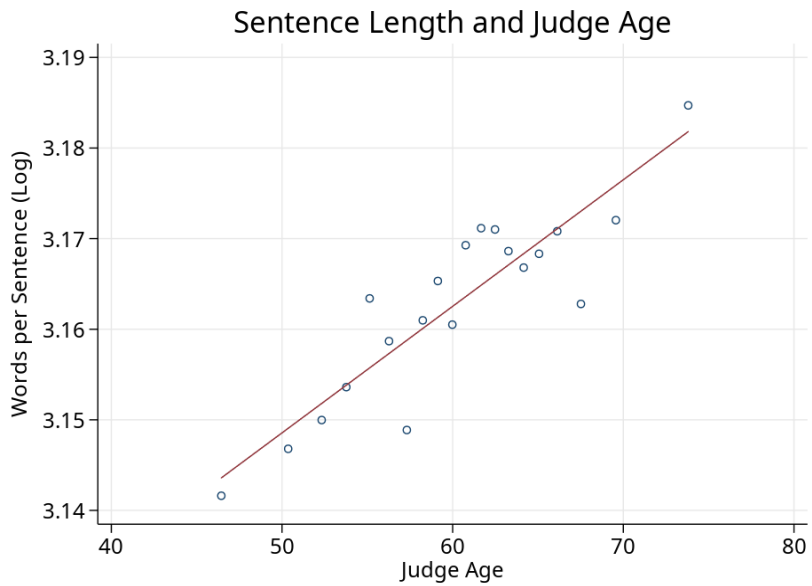
Source: World Bank.

New Measurements

Source: World Bank.

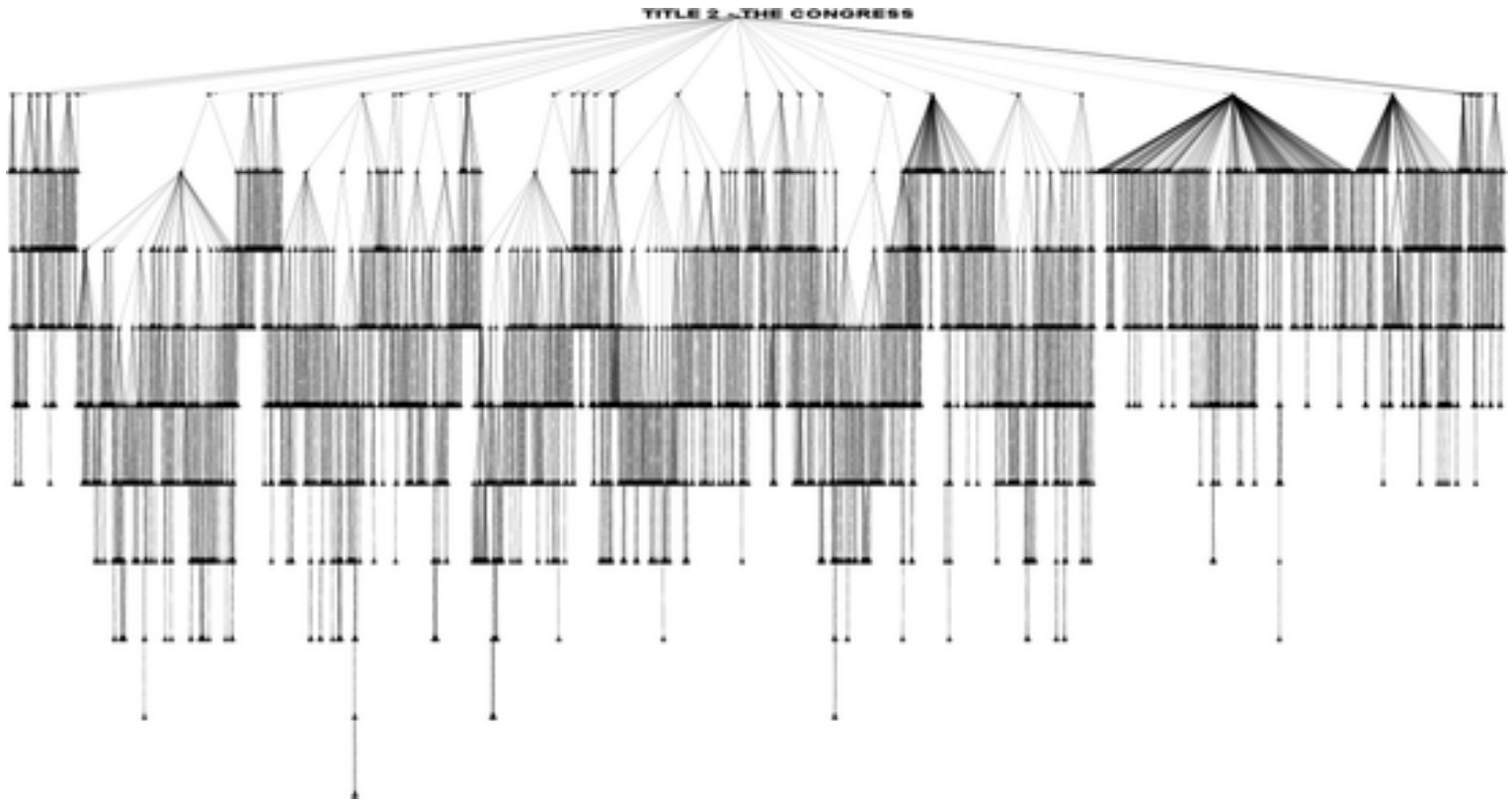
Judge Age and Writing Style





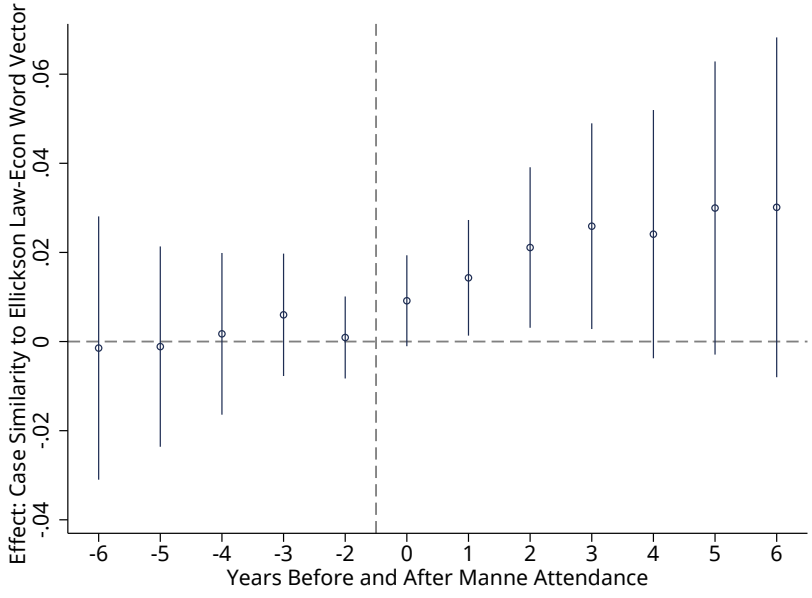
Ash and MacLeod (2020)

Complexity Analysis of Legislation



- The U.S. Code consists of 49 titles, which can be further subdivided into subtitle, chapter, subchapter, part, subpart, section, subsection, paragraph, subparagraph, clause, and subclause (Katz and Bommarito 2014)

Impact of Economics Training on Economics Language



After attendance, Economics Trained Judges increase use of a selection of terms related to law and economics (Ash, Chen, and Naidu 2019)

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (2016)

- Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (2016)

- Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.

- For each newspaper on each day since 1985, submit the following query:
  - 1. Article contains "uncertain" OR "uncertainty", AND
  - 2. Article contains "economic" OR "economy", AND
  - 3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

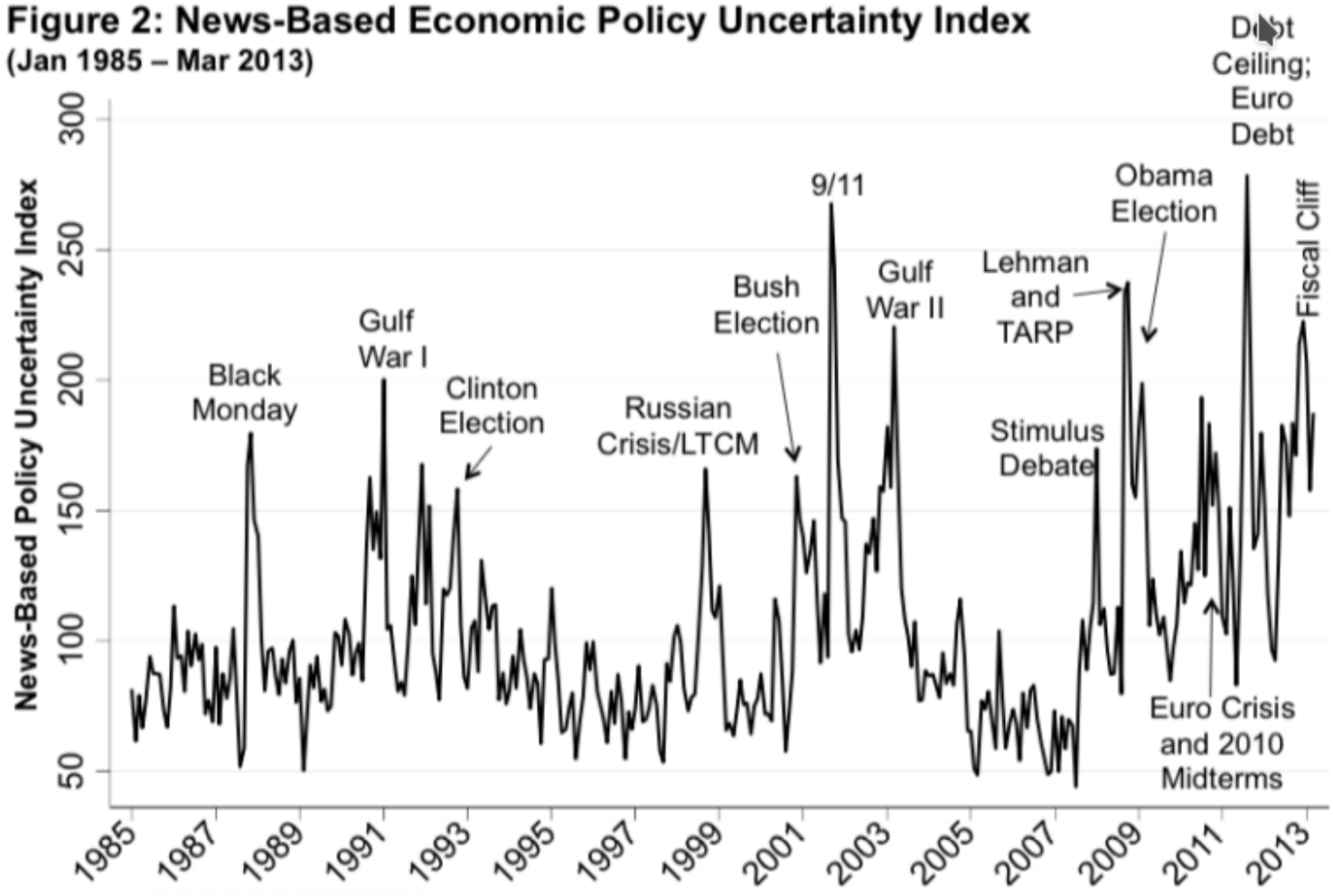
Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (2016)

- Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles.

- For each newspaper on each day since 1985, submit the following query:
  - 1. Article contains "uncertain" OR "uncertainty", AND
  - 2. Article contains "economic" OR "economy", AND
  - 3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"
- Normalize resulting article counts by total newspaper articles that month.

Measuring uncertainty in macroeconomy



Predicting U.S. Asylum Court Decisions

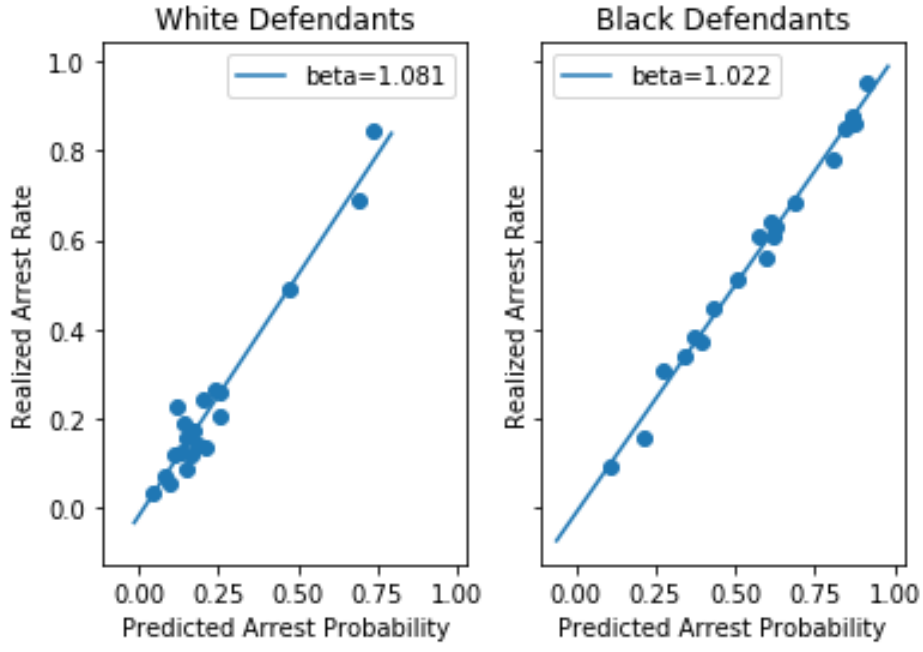
		Predicted	
		Denied	Granted
True	Denied	195,223	65,798
	Granted	73,269	104,406

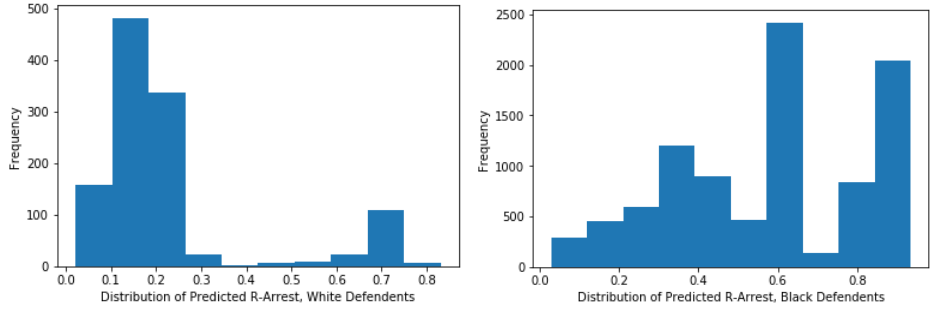
Accuracy = 68.3%, F1 = 0.60

- Prediction App (Beta):
  `https://floating-lake-11821.herokuapp.com/`

Predicting Re-Arrest, New Orleans Prosecutor Office

Amaranto, Ash, Chen, Ren, and Roper (2018)





- **Algorithms can correct for biases in human decision-making:**
    - for judges: all defendants get the same decision for the same evidence.

- **Algorithms can correct for biases in human decision-making:**
  - for judges: all defendants get the same decision for the same evidence.

- **What about *systematic* biases?**
  - e.g., those leading to racial disparities.

- **Algorithms can correct for biases in human decision-making:**
  - for judges: all defendants get the same decision for the same evidence.

- **What about *systematic* biases?**
  - e.g., those leading to racial disparities.
  - There is a risk these could be reproduced or even amplified by the algorithm.

- **Algorithms can correct for biases in human decision-making:**
  - for judges: all defendants get the same decision for the same evidence.

- **What about *systematic* biases?**
  - e.g., those leading to racial disparities.
  - There is a risk these could be reproduced or even amplified by the algorithm.
  - But algorithms can also be used to **detect** systematic bias, to **understand** it – and therefore to help **reduce** it.

Interpretable Machine Learning

- Key point:
  - Standard machine learning techniques cannot be interpreted easily.

Interpretable Machine Learning

- Key point:
  - Standard machine learning techniques cannot be interpreted easily.
- Users and decision subjects want to understand the model

Interpretable Machine Learning

- Key point:
  - Standard machine learning techniques cannot be interpreted easily.
- Users and decision subjects want to understand the model
- Other models/approaches improve interpretability:
  - Random Forests provide feature importance ranking.
  - LIME and related tools can help interpret any model (Ribeiro et al 2016).

- **Econometrics** (applied statistical causal inference):
    - y is one-dimensional, x is low-dimensional.

- **Econometrics** (applied statistical causal inference):
    - y is one-dimensional, x is low-dimensional.
    - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.

- **Econometrics** (applied statistical causal inference):
    - y is one-dimensional, x is low-dimensional.
    - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

      where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
    - ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
    - useful for policy evaluation.

- **Econometrics** (applied statistical causal inference):
  - y is one-dimensional, x is low-dimensional.
  - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

    where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
  - ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
  - useful for policy evaluation.
- **Machine learning**:
  - $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.

- **Econometrics** (applied statistical causal inference):
  - y is one-dimensional, x is low-dimensional.
  - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

    where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
  - ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
  - useful for policy evaluation.
- **Machine learning**:
  - $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
  - learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function

$$y_i = h(\boldsymbol{x}; \theta)$$

    that predicts y given covariates $\boldsymbol{x}$.

- **Econometrics** (applied statistical causal inference):
  - y is one-dimensional, x is low-dimensional.
  - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

    where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
  - ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
  - useful for policy evaluation.
- **Machine learning**:
  - $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
  - learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function

$$y_i = h(\boldsymbol{x}; \theta)$$

    that predicts y given covariates $\boldsymbol{x}$.
  - if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.

- **Econometrics** (applied statistical causal inference):
  - y is one-dimensional, x is low-dimensional.
  - estimate a low-dimensional **causal parameter** ρ using

$$y_i = \alpha_i + x_i \cdot \rho + \epsilon_i$$

    where i indexes over documents, α_i includes control variables (and fixed effects), $\cdot$ is dot product, and ϵ_i is the error residual.
  - ρ gives a prediction how outcome y would change if treatment variable x were **exogenously shifted**.
  - useful for policy evaluation.

- **Machine learning**:
  - $\boldsymbol{y}$ can be multi-dimensional, $\boldsymbol{x}$ can be high-dimensional.
  - learn a high-dimensional vector of parameters θ to approximate a (potentially non-linear) function

$$y_i = h(\boldsymbol{x}; \theta)$$

    that predicts y given covariates $\boldsymbol{x}$.
  - if we collected more data on $\boldsymbol{x}$, we could predict the associated $\hat{y}$.
  - but $h(\cdot)$ does not provide a *counterfactual prediction* – that is, how the outcome would change if $\boldsymbol{x}$'s were exogenously shifted.

Objectives

1. **What is the policy problem or research question?**

Objectives

1. **What is the policy problem or research question?**
2. Corpus and Data:
    - obtain, clean, preprocess, and link.
    - Produce descriptive visuals and statistics on the text and metadata

Objectives

1. **What is the policy problem or research question?**
2. Corpus and Data:
    - obtain, clean, preprocess, and link.
    - Produce descriptive visuals and statistics on the text and metadata
3. Machine learning:
    - Select a model and train it.
    - Fine-tune hyperparameters for out-of-sample fit.
    - Interpret predictions using model explanation methods.

Objectives

1. **What is the policy problem or research question?**
2. Corpus and Data:
    - obtain, clean, preprocess, and link.
    - Produce descriptive visuals and statistics on the text and metadata
3. Machine learning:
    - Select a model and train it.
    - Fine-tune hyperparameters for out-of-sample fit.
    - Interpret predictions using model explanation methods.
4. Empirical analysis
    - Produce statistics or predictions with the trained model.
    - **Solve the problem / Answer the research question.**

Outline

Writing Good Code

Three commandments:

1. **Comments**
   - Don't write documentation you will not maintain.
   - Code should be self-documenting.
2. Reuse functions (no copy & paste) $\rightarrow$ **automation**
   - Automate everything that can be automated
   - Write a single script that executes all code from beginning to end
3. Use **version control** (no doc_final_v2.py)
   - Store code and data under version control.
   - Run the whole directory before checking it back in.

More on Project Management

- Naming
- PEP8
- Directories' organization
  - Separate directories by function.
  - Separate files into inputs and outputs.
  - Make directories portable.

Homework for next time

- install anaconda and packages
  - sklearn, gensim, xgboost
- run jupyter notebook
  - Play with the R-stata-cookbook-for-python notebook (on the github page of the class)
- register in the twitter API
  - For installation guidelines, see this (paragraph "An Example with Twitter's API)