Machine-Learning Evaluation of Author Ethnicity in Lattice Publications

Department of *Physics and Astronomy* & *Computational Mathematics, Science and Engineering*
Michigan State University, East Lansing, MI 48824
Contact: Prof. Huey-Wen Lin (hwlin@pa.msu.edu)

Motivation

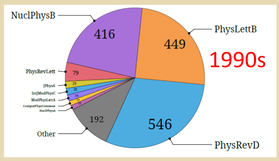
Nationwide growing effort to improve diversity in STEM

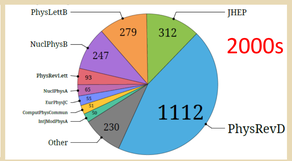
- Both NSF and DOE had grant opportunities in 2020
  - Improve diversity and retention at the doctoral level
- Many higher-education institutions and labs are also investing in outreach efforts targeting the pipeline issue in STEM
  - Funding K-12 activities locally throughout the year
  - REU programs
  - Summer internships
- One can expect a wave of new-generation diverse scientists with unique ways of thinking and problem-solving abilities
- In the late stages of the pipeline within academic jobs, scientists are subject to evaluation based on their publications
  - Affects students getting postdocs; postdocs to faculty
  - Faculty to get grants and achieve tenure, and so on
- Any significant gender or racial could cause years of effort bringing up a diverse workforce to be in vain
  - We must study and monitor any potential bias and make sure that there is no bottleneck in these final stages
  - In this poster, we focus on hep-lat primary publication only

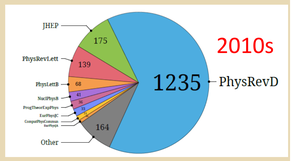
Data Collection and Processing

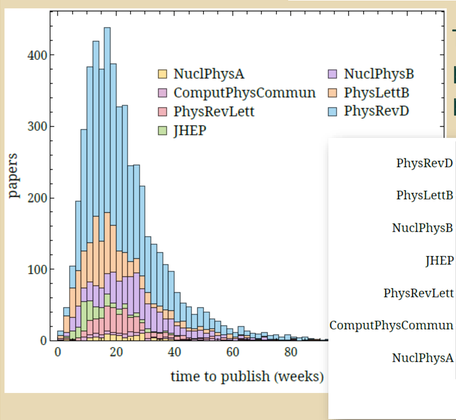
Starting with hep-lat arXiv metadata using API

- Removed all papers cross-listed to other fields
- Removed proceedings (*Nucl. Phys. Proc. Suppl.*, *PoS*, *J of Phys Conf. Series*, etc.).
- Results below sorted by decade

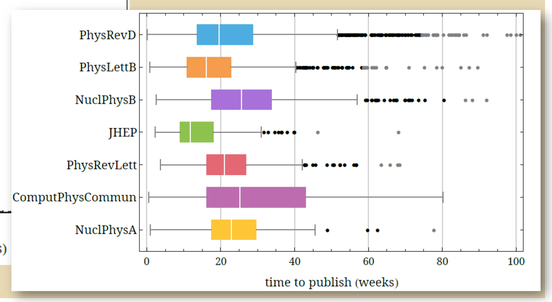








Time in weeks between the arXiv and publication dates of non-proceeding papers in top lattice journals



Author Ethnicity via Machine Learning

Use TensorFlow to perform the NN training with GPU acceleration via NVIDIA's cuDNN deep-neural-network library.

US Census Training Data

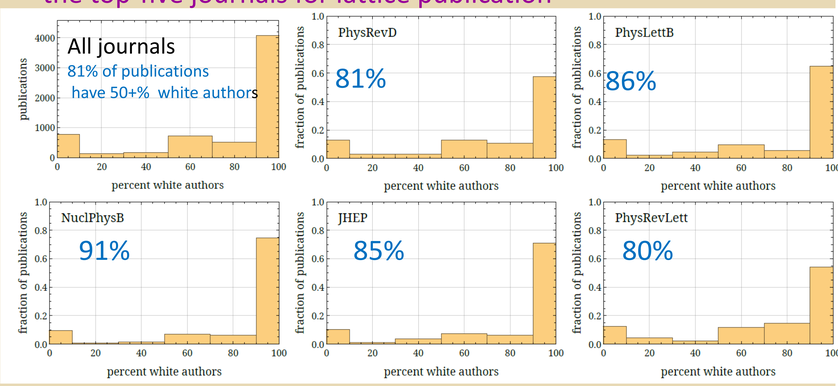
- Known for high accuracy and amount of data: 162k surnames
  - Breaks down surnames by race categories: “white”, “black”, “Hispanic”, “Asian/Pacific islander”, “American Indian”, “mixed race”
- Prepare the surnames for input into the DNN
  - Convert each name into a vector of reals in the range [0,1] with uniform length, padded to the left to width 20 by zeros
  - Each letter is converted to its position in the alphabet: A → 0, Z → 1
  - Model uses 4 dense layers with widths 200, 64, 32 and 4 trained with a *mean-squared* loss function and the *Adadelta* optimizer over 300 epochs
- Predictions of “white”, “black”, “Asian” and “Hispanic”
  - Recognizes “AOKI”, “LIU” and “HASHIMOTO” as “Asian”, and “JANSEN”, “LEINWEBER” and “HELLER” as “white”
  - Predicts “LIN” as “white”, possibly due to the fact that “-LIN” is a common ending in surnames such as “FRANKLIN”
  - Far more authors in this data tagged as “black” than our field has
    This can be understood due to American history; black and white Americans simply share many surnames, and there is a mixture of surnames in the training data. For example, “SMITH” is about 50% white and black according to the census data. These American trends do not well match the demographics of the lattice authors data set.

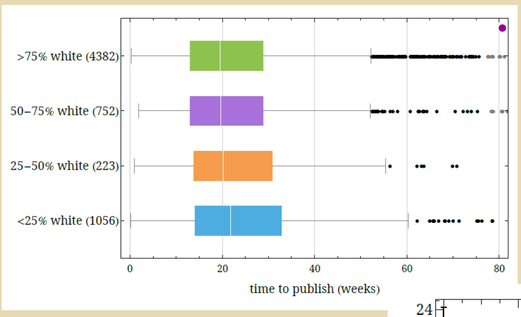
Identifying the ethnic origin of the surnames

- Training a network with the most common surnames from a diverse selection of countries
  - Most common 1000 surnames from 27 countries:
    Spain, Mexico, Brazil, Argentina, Italy, France, England, Scotland, Ireland, Germany, Greece, Sweden, Russia, Poland, Hungary, Turkey, Israel, Egypt, Syria, Iran, Pakistan, India, Vietnam, China, Taiwan, Japan and South Korea
  - Expanded our alphabet to 42 characters and padded length to 21, yielding an input dimension of 882
- Four-layer model with different widths (1764, 441, 216 and 27)
  - Final layer has sigmoid activation to constrain it to [0,1]
  - The output should be a unit vector (left it unnormalized)
  - Loss function to *categorical cross-entropy*; optimizer: *Adam*
  - New model has 2.4M parameters
- Final results
  - 3% errors on the overall training
  - Output the top-3 predictions for each last name
  - Resolved many previously problematic surnames; it identified “JANSEN” as Swedish, “HELLER” as German, and “LIN” as Taiwanese.
  - Hand-checked the top-323 surnames, covering all authors with 20 or more lattice papers, and found 77% accuracy at the country level and 92% accuracy discerning white versus Asian

Results and Discussion

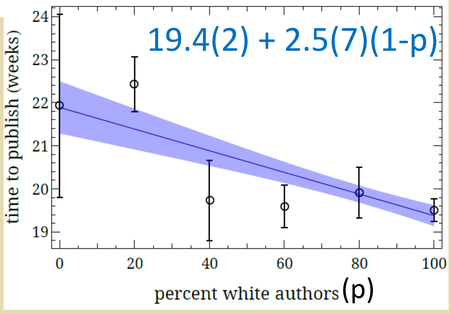
- Histograms of the author-list composition for all published papers recorded on arXiv (upper left) and the breakdowns for the top-five journals for lattice publication





Time to publication as a function of author ethnicity fraction
- Increase of about 10% in the median for papers with few white authors

- Publication as a function of author ethnicity fraction fit to a linear trend
  - An all-nonwhite authored paper expects to be tied up in the editorial process for 2.5 weeks longer than all-white author group



Conclusion & Outlook

- The machine-learning evaluation of the author ethnicity suggests a 10% extra time for non-white authors to get their hep-lat papers published
- Big Caveat: **Survivor bias**
  - These public data are for all published papers.
    People may have some difficulties, but they make it.
  - We cannot measure how many papers are blocked from the publication system entirely
- We should minimize the (un)conscious bias
  - Double blind system? Unified referee data base to keep track of the unconscious bias training? Other suggestions?