Women in Machine Learning and Data Science Boston #1

March 26, 2019

WiMLDS Mission

WiMLDS is an international organization for women+ interested in Machine Learning and Data Science.

We host local events where we discuss machine learning and data science in an informal setting with the purpose of building a community around women and gender minorities in these fields.

WiMLDS History

2005: Hanna Wallach, Jennifer Wortman Vaughan, Lisa Wainer and Angela Yu shared a room at NuerIPS

2006: Annual WiML Workshop begins

2013: After inspiration from WiML Workshop, Erin Ledell creates Bay Area WiMLDS meetup

2014: WiMLDS NYC created

2014: WiMLDS board created

2018: WiMLDS chapters across 6 continents

WiMLDS Numbers

27,234 members

62 chapters

27 countries

WiMLDS Resources

Twitter: @wimlds & @wimlds_boston

Slack: To join, email slack@wimlds.org

Job board: http://wimlds.org/jobs/

Email: boston@wimlds.org

What should Boston be?

“The point is to show that our meetup is not only about putting women front and center, it is also a recognized scientific and technical meetup in its own right!”
- Caroline Chavier, founder of Paris WiMLDS chapter

What should Boston be?

- How often should we meet?
- Event types:
  - Hands-on workshops
  - Speakers
  - Hack nights
  - Chapter challenges
  - Community service
  - New speaker nights
  - Panels

Kira Tebbe

Twitter: @k_tebbe

Email: kira@tebbe.com

Stories move people.

How do I create a story from data?

- Start with a question – then think of more questions!
- Get your data – the more the merrier!
- During data exploration:
  - Start with one of your initial questions
  - Keep an open eye for any surprises
  - Go beyond summary statistics
  - Don’t lose the forest for the trees!
- Turning results into a story:
  - Think big – what were you trying to answer?
  - Find a logical progression in your results
  - Know your audience
  - Make your slides/report/visualization easy to interpret

Phoebe Wong

Twitter: @phoebewong2012

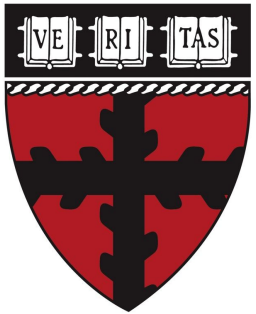
Email: wong@g.harvard.edu

Phoebe Wong

WiMLDS Boston
Mar 26, 2019

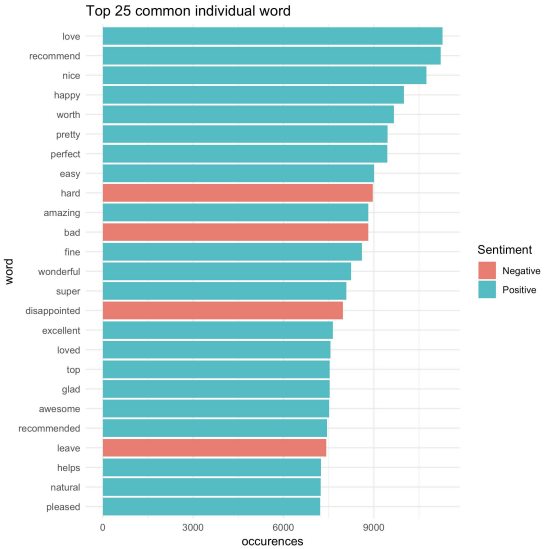
Who am I?

- MSc in Data Science at Harvard SEAS
- Freelance Data Journalist at Fractl
- Previously, Primary Research Analyst at Legendary Pictures
- BA in Psychology from UC Berkeley



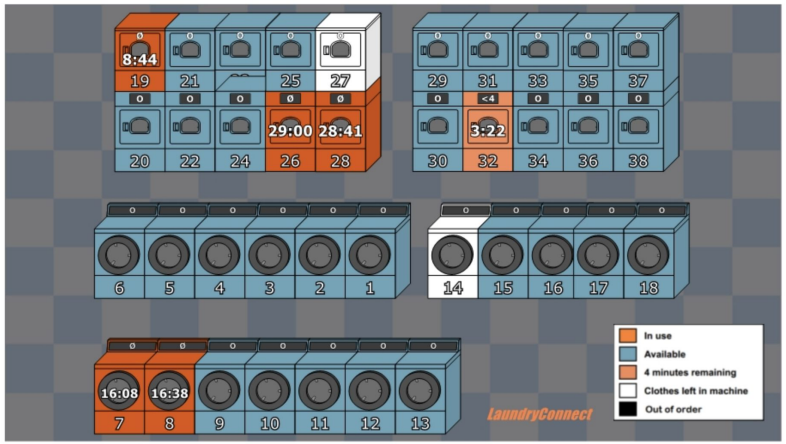
Data Science Interest

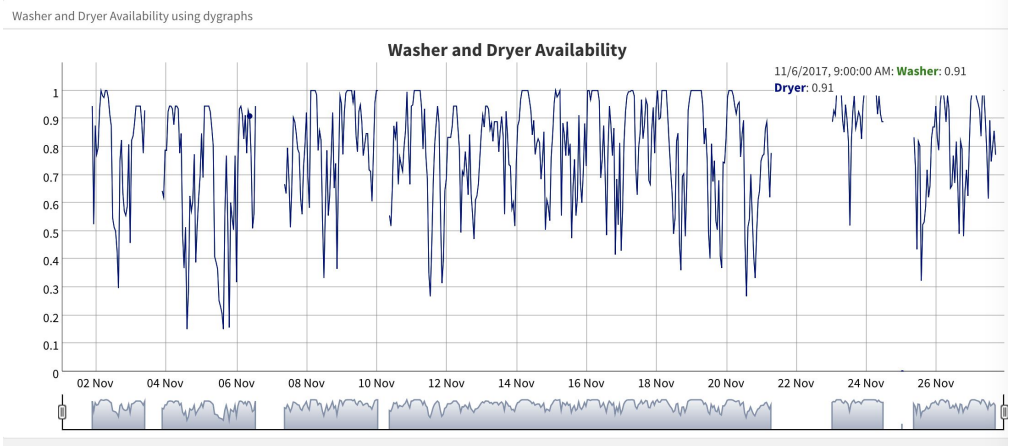
- Recommendation System
  - Song recommendation using Spotify 1 million playlist dataset
  - https://phoebewong.github.io/music-recommendation-teamNPK/
- Statistical Inferences
  - 142.8 million Amazon product rating and reviews (May 1996-July 2014)
  - https://nzstern.com/projects/amazon-product-rating-review-statistics/
- Natural Language Processing
  - Sentiment analysis of Amazon product review
- Interactive data visualization (R Shiny dashboards)



Personal Projects

- Laundry Dashboard: https://smalldatabigfindings.com/featured-projects/laundry_flexdashboard/
- MS Outlook Meeting History Analysis: https://phoebewong.shinyapps.io/calendar_shinyio/





R-Ladies Boston

- Co-organizer of R-Ladies Boston chapter
- Meet once every month to talk about applications of R in academia and industry
- High-level talks, hands-on workshops, project nights and social nights
  - Feb: how to build a R package, March: NLP analysis in R
- Join us on Meetup: https://meetup.com/rladies-boston/
- Follow us on Twitter: https://twitter.com/RLadiesBoston

Love to meet you all!

Twitter: phoebewong2012

LinkedIn: wphoebe

Email: wong@g.harvard.edu

Website: https://smalldatabigfindings.com

Xiaoying Shi

Email: xiaoying.shih@gmail.com

San Wang

Twitter: @SanwangSan

Email: San_Wang@hms.harvard.edu

Image Forensics

WiMLDS lightning talk
3.26.2019
San Wang @ HMS

About Me

Education

Bachelor: Mathematics & Statistics

Master: Data Science

Data Science Journey

ML class -> NLP research assistant -> independent projects (movie recommendation system, fashion explore platform, Kaggle)

Now

Computer Vision Research Associate @ HMS

https://san-wang.github.io/

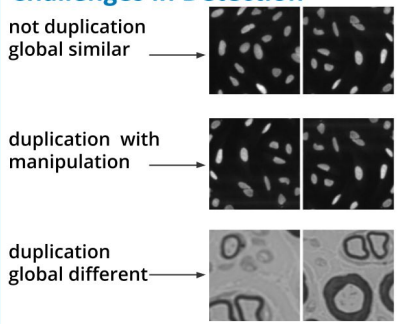
Now let’s talk about image forensics

Automated Quantitative Assessment of Inappropriate Image Reuse

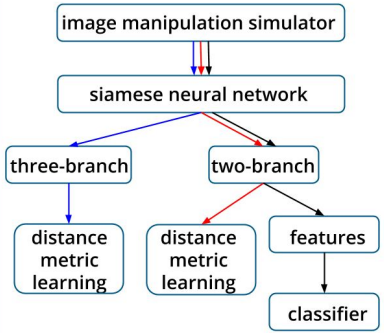
San Wang
Research Associate

home department: **Harvard Medical School(HMS)** • supervisor: Marcelo Cicconet, Mary Walsh
research group: **HMS Office For Academic and Research Integrity(ARI) & HMS Image Data and Analysis Core(IDAC)**

Challenges in Detection

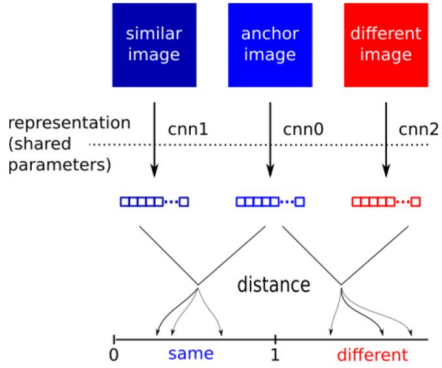


Methods in Action

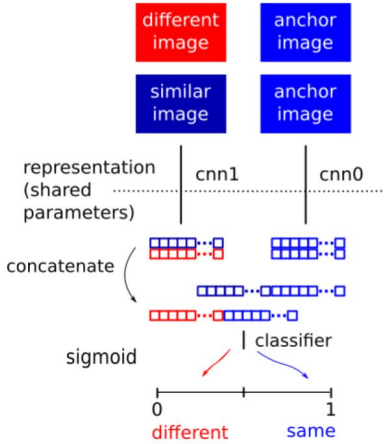


Model Diagram Snapshot

Three-Branch Siamese



Two-Branch Siamese



cnn: Convolutional Neural Network

Next Steps

- → Combine with local image features methods
  - ◆ bag of local features
  - ◆ SIFT, SURF, edges
- → Train model with cases of different complexity under hierarchical data structure
- → Use attention and feature visualization to explain learned model and provide specific coordinates to indicate duplicated region
- → Build manipulation verification model that can reproduce manipulation process and provide manipulation type and metric

Result

Model Name	Discipline	Test Accuracy
triplets siamese	similarity distance	0.89
twins siamese	similarity distance	0.84
	feature extraction + classification	0.90

Note: test accuracy is the average of 10 runs on real world dataset using model trained on synthetic dataset, which simulates real world manipulation patterns

Acknowledgements

Authors: San Wang[2,3], Hunter L. Elliott[2], Gretchen Brodnicki[1], Jen Ryan[1], Rachel Fouche[1], Blake Talbot[1], Daniel H. Wainstock[1], Marcelo Cicconet[2,3] and Mary C. Walsh[1,3]
[1] Office of Academic and Research Integrity, and
[2] Image and Data Analysis Core, Harvard Medical School. Boston, MA, USA
[3] Corresponding authors

Harvard Medical School

Questions?

Reach out:

Email, LinkedIn, Github@
https://san-wang.github.io/

Lydia Skrabonja

Twitter: @lydium90

Email: lskrabonja@gmail.com

Finding the Optimal Length of Stay in Hospice

Lydia Skrabonja
Cyft

Cyft

- Healthcare
- Performance Improvement
- This project: end of life care

Soft Problems

- Stigma around end of life conversations
- Many people don’t get goals concordant care

Data Problems

- No agreed upon optimal length of stay in hospice
  - Max 6 months
- Known to save money, unknown exactly how much

Hypothesis: For each person, there exists a cost curve along “# days in hospice” with a minima = Optimal Length of Stay in Hospice

Proposed Solution: Step 1

- Cohort: eligible plan members who died since 1/1/2017
- Build a model that predicts Total Medical Expense (TME) in the last 6 months of life
- 1 feature = % of last 6 months enrolled in hospice
- Current step

Proposed Solution: Step 2

- For each member
- For all values of “% of last 6 months enrolled in hospice”

- Generate the expected TME in the last 6 months
- Find the minima/inflection point/etc
- Output the difference in both # days and TME

Proposed Solution: Expected Outcomes

- Overall
  - Savings of the current hospice program
  - How aligned the hospice program is timing-wise

- Individually
  - Optimal length of stay for each member
  - Expected cost for no hospice
  - Expected savings with optimal hospice

Thank you!

@lydium90
lskrabonja@ cyft.io / gmail.com

Sarah Rich

Twitter: @sarahjrich

Email: sarah@canopy.cr

Making Great Private Recommendations

Sarah Rich (@sarahjrich)
Machine Learning at Canopy (@ourcanopy)

How Do “We” “Usually” Do ML For Personalization?

- Log everything (geo-tracking, in-app behavior, in-app history, device IDs, browser history, co-occurring app installs)
- Store everything (write pipelines to parse this data and make it readily accessible at scale)
- Train models on all of this data to predict interest/engagement/demographics
- Use this data for analytics and advertising:
  - Who do we think is using our product and how can we improve it based on that information?
  - Who do we think is using our product and how can we sell their attention to advertisers?
- (Maybe even?) Sell this data

What's Wrong With This Picture?

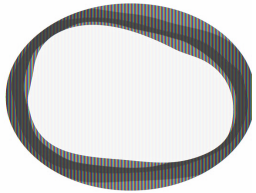
- Possible data breaches leave sensitive data vulnerable
- The app is not the product; the person using the app is the product
- The customer is advertisers or data brokers
- Fundamental mismatch between what we're doing and what we say we're doing
- Feels disingenuous and erodes trust between the person using the app and the company

What We're Doing Instead at Canopy

- Content recommendation (articles and podcasts right now)
- Want to know as little about you as possible
- Store zero raw data on our servers
- Train an ML model online that is differentially private
- We are focused on making private recommendations as delightful as possible

What This Means for Me

- Analytics: have to be private and online, suitably aggregated
- Training ML models: We don’t store any raw training data!
- Getting to solve super interesting new problems and help build a better Internet!



Thank you!

@sarahjrich @ourcanopy

1. What was the last movie you saw?
2. What was the most frustrating bug you worked on lately?
3. Share a shocking statistic.
4. Favorite place to eat in Boston?
5. What is something cool you have programmed lately?
6. Where were you living in 2012?