Some Stuff About My Lab

Hanna M. Wallach
University of Massachusetts Amherst
wallach@cs.umass.edu
Who is Hanna Wallach?
This Talk

- Background: science and innovation policy
- Methodological approach
- Ongoing and future projects
- My advising style
“Whether it's improving our health or harnessing clean energy, protecting our security or succeeding in the global economy, our future depends on reaffirming America's role as the world's engine of scientific discovery and technological innovation.”

— President Barack Obama
“The public has generally treated this progress as something that just happened, without recognizing that it is, in fact, largely the result of a sustained federal commitment to support science through science policies.”

— http://science-policy.net
Science and Innovation Policy

- Goal: identify administrative, financial, political actions
- Actions chosen to have impact on, e.g.,
  - Stimulating breakthrough research
  - Increasing economic prosperity
  - Broadening participation
- Government, private sector, education
- This talk: statistical models for facilitating efficient, data-driven science policy decisions
Examples of Policy Actions

- Funding actions:
  - Using federal funds for research on human stem cells
  - “People not projects” vs. pre-defined deliverables

- Patenting actions:
  - Granting software patents

- Educational actions:
  - Running high school outreach activities
  - Providing mentoring programs
Data-Driven Policy Decisions

- Discovery: identifying possible policy actions
- Prediction: estimating expected impact
- Evaluation: assessing observed outcomes

⇒ Automated data analysis
“Scientific information is both the basic raw material for, and one of the principal products of, scientific research [...] Scientists find out what other scientists are accomplishing through [...] journals, books, abstracts and indexes, bibliographies, reviews.”

― NSF Brochure, 1962
My Research Goal

To develop new statistical models and computational tools for representing and analyzing large quantities of complex data in order to better enable scientific policy-makers to identify and evaluate high-impact policy actions and advance the study of science and innovation policy.
Collaborate to Study Collaboration

“There needs to be a greater focus on what these [science interaction] data mean [...] This requires the input of social scientists, rather than just those more traditionally involved in data capture, such as computer scientists.”

— Julia Lane, NSF, 24 March 2010
This Talk

- Background: science and innovation policy
- **Methodological approach**
- Ongoing and future projects
- My advising style
Statistical Models

• Modeling challenges:
  - Aggregating and representing large data sets
  - Handling data from sources with disparate emphases
  - Reasoning under uncertain information
  - Performing efficient inference

• Bayesian latent (hidden) variable models:
  - Powerful and flexible [Wallach et al. & Adams et al., AISTATS '10]
  - In particular: statistical topic models
Generative Statistical Modeling

- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data into estimated model
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

----

Topics and Words

- human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
...

- evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
...

- disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
...

- computer
models
information
data
computers
system
network
systems
model
parallel
...

probability
Generative Process
Choose a Distribution Over Topics

- probability

- computer
  - models
  - information
  - data

- disease
  - host
  - bacteria
  - resistance
  - bacterial
  - strains

- evolution
  - species
  - organisms
  - life

- human
  - genome
  - dna
  - sequence
  - gene
  - molecular
  - sequencing

- parallel
  - network
  - systems
  - model
  - control

probability
Choose a Topic

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 150 genes, and that the earliest life forms required a mere 128.
Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 150 genes, and that the earliest life forms required a mere 126 genes.
... And So On

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 150 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, there's a prediction...
Real Data: Statistical Inference

Seeking Life’s Bare (Genetic) Necessities

Cold Spring Harbor, New York—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, these predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Stiivid Andrensen of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Aracdy Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland.

“Stripping down,” he suggests, “is a way to assess the minimum modern and ancient genomes.”


Science • Vol. 272 • 24 May 1996
This Talk

- Background: science and innovation policy
- Methodological approach
- Ongoing and future projects
- My advising style
Help! All my topics consist of “the, and of, to, a ...”

Now they all consist of “invention, present, thereof ...”

Wait, but how do I choose the right number of topics?

Preprocess your data to remove stop words...

Make a domain-specific list of stop words...

Evaluate the probability of unseen data for different numbers...
Utilizing Existing Knowledge

● Many human-curated ontologies, e.g., MeSH
● Many, many problems:
  – Expensive to construct and maintain
  – Inter-annotator agreement is low
● But! They represent human constructions of knowledge
● Goal: incorporate existing human knowledge into large-scale automated tools for textual pattern discovery
Detecting Scientific Emergence

- T36 = carbon nanotube transistors
- T44 = (carbon nanotube electronics)
- T49 = (carbon nanotube electronics)
Understanding Diversity of Science

- Policy actions shape the diversity of science:
  - Idea diversity: array of different ideas
  - Individual diversity: variety of people and organizations
- Goal: develop new methods and tools for:
  - Quantifying the diversity of science
  - Assessing impact of policy actions on diversity
Studying FOSS Development

- Free & open source software (FOSS):
  - Complex technological, legal, social structures
  - Collaboration on a massive scale
- Most communication is online and publicly available
  - Informal documents: messy, unstructured
- Goal: use these data to study organizational and social processes underlying FOSS development
Analyzing Debian Mailing Lists

- women, men, debian, debian-women, men women, women men
- packages, package, debian, maintainers, maintainer, upstream
- dd, upload rights, people, dds, voting rights, rights
- debian, packages, package, sponsor, nm process, process
Modeling Politics

- Most modeling work in political science ignores text
- Many, many goals:
  - Discovering issue-based voting coalitions in the senate
  - Analyzing senators' representational style
  - Characterizing persuasiveness of emails
  - Predicting when to declassify documents
  - ... and more!
So You Publish Where…?

- NIPS
- ICML
- AISTATS
- EMNLP
- JITP
- ICWSM
- ...

Hanna M. Wallach :: UMass Amherst :: 31
This Talk

- Background: science and innovation policy
- Methodological approach
- Ongoing and future projects
- My advising style
Advising Style

- Work with students to help them become researchers
  - Talk about research methods
  - Treat students as collaborators
- Encourage students to become part of the wider machine learning and CSS communities
- Encourage students to visit other labs
- Lab meetings (brief updates plus reading group)
Thanks!

If you would like to get involved, email me! wallach@cs.umass.edu