The Adolph C. and Mary Sprague
MILLER INSTITUTE
for Basic Research in Science

Newsletter
Winter 2020

What Happens to Algorithms When Data Get Big?

Miller Fellow Focus: Sam Hopkins

Algorithms, machine learning, big data: we've all heard the buzzwords. We can gather and harness much larger data sets than in decades past, across tech, medicine, the sciences, and beyond. The scale of this data revolution demands that we address questions ranging from practical to social to mathematical. (How can we use data to improve peoples' lives and make scientific discoveries? How do we protect the privacy of individuals represented in large data sets? How can we reason rigorously about the conclusions we extract from them?)

I study the mathematical underpinnings of big data analysis, using tools from statistics and computer science. For most of the 1900s, the field of statistics developed tools to analyze small data sets: usually, the amount of data available was the limiting factor in what conclusions could be drawn. In the big data era, a different limiting factor emerges: computation.

In this article I'll discuss two more specific topics: *algorithmic phase transitions* and *high-dimensional medians.*

Beginning with algorithmic phase transitions, let's start with an example. Imagine you are faced with a network data set, like **(FIGURE 1)**. It could represent people, where lines denote friendships, or proteins, where lines denote joint participation in a biological pathway, or any other situation where entities ("nodes") experience pairwise interactions ("edges").

A common goal is to identify *communities* in the network: groups of nodes which share a more-than-typical number of edges. Suppose we want to split the network into k groups so that the number of in-group edges is as large as possible. How should we find the best grouping? The most naive approach is to try every possible grouping and take the best. But here is our computational roadblock: if the network contains n nodes, then there are more than 2^n possible groupings. 2^{300} already exceeds the number of atoms in the universe, but modern network data sets frequently contain thousands of nodes and sometimes hundreds of millions. The *brute-force* approach requires an exponential amount of computation: it is fundamentally intractable.

"I really appreciated the atmosphere of discussion and the culture of scientific and personal exchange at UC Berkeley and in particular at the Miller Institute. The opportunity to hear about exciting research in other disciplines was exciting and often eye-opening. The Miller Institute is in fact a place that beautifully combines the intellectual, collaborative, and social aspects of science."

- Bernd Abel
Visiting Miller Professor Fall 2018, Chair for Chemical Engineering of Polymers (Technische Chemie der Polymere), Ostwald-Institut für Physikalische und Theoretische Chemie and Institut für Technische Chemie, Head of the Chemical Department of the Leibniz Institute of Surface Engineering (IOM)

This *community detection* problem displays many of the hallmarks of big data problems: the data set is large, and we want to extract a similarly-large amount of

INSIDE THIS EDITION

CONTINUED ON PAGE 2 >

> CONTINUED FROM PAGE 1 [Miller Fellow Focus]

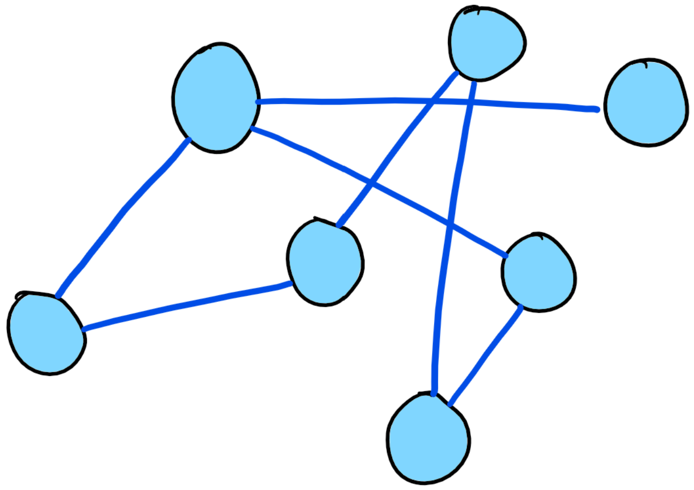

Figure 1: a network

information from it – a grouping of all the nodes at once, not just an answer to a single yes/no question. It displays another signature feature: even though the brute-force algorithm is intractable for even a modestly large network, *efficient algorithms do exist for "nice-enough" data.*

Let us illustrate this phenomenon in a simple mathematical model. Imagine that our network data is generated in the following way: first, each of n nodes randomly chooses one of k colors. Then, for each pair of nodes a,b, we flip a coin and add an edge to the network if the coin comes up heads. To capture the group structure from the colors, the coin is *biased*: if a,b have the same color, the coin is heads with probability p, and otherwise the coin is heads with probability q, with $p \geq q$. Now p-q is a mathematical measure of "niceness". When $p=q$, the network contains no information about the underlying groups. At the opposite extreme, if $p=1,q=0$ then the network can be split into k sub-networks with no edges going between them **(FIGURE 2)**. What happens between the extremes?

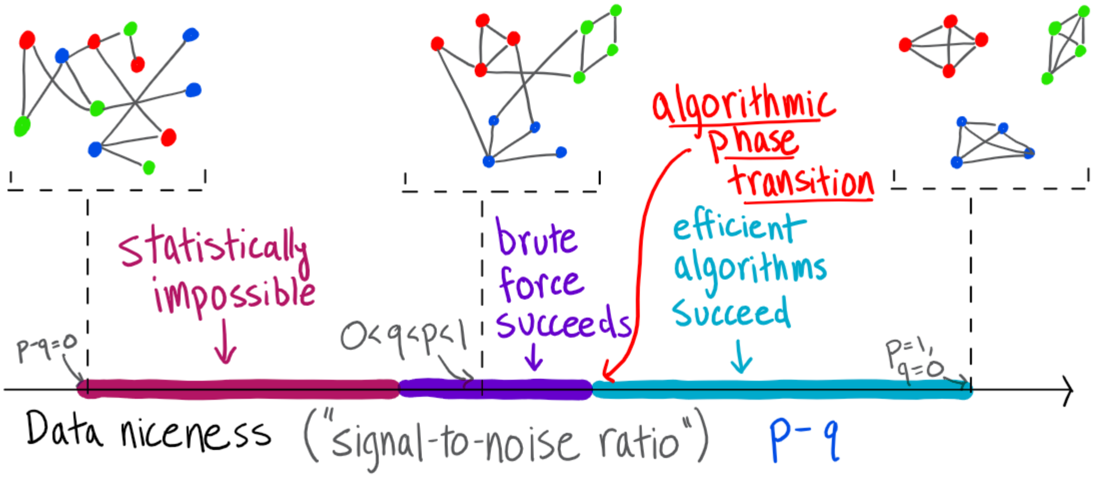

Figure 2: information-theoretic and algorithmic phase transitions as signal-to-noise ratio increases.

Remarkably, there are two distinct changes in most big data problems as niceness increases. First, the data become nice enough to extract information (e.g. a community) by brute force. Then, as data get nicer, computationally-efficient algorithms emerge – in this case based on eigenvectors of matrices associated to the network. This is the *algorithmic phase transition*.

Algorithmic phase transitions are poorly understood. For most big data problems, we have little rigorous evidence that the algorithmic phase transition even exists – typically, all we know is that all the algorithms we can devise cease to recover meaningful information at the same threshold of niceness.

During my PhD, I studied algorithmic phase transitions across numerous big data problems. My collaborators and I gave the strongest mathematically rigorous evidence to date for non-existence of efficient algorithms for some key network and component analysis problems which do admit brute-force solutions. We largely ruled out the possibility that any approach using known techniques for computationally efficient algorithms could tolerate less nice data than current algorithms do.

Let's turn to medians in high dimensional data. "High-dimensional" means that many measurements have been recorded for each sample in the data. For example, in a data set of images, each image might consist of ten thousand pixels – a pixel is a single measurement, so this data set would be ten-thousand dimensional. Large dimensionality is another hallmark of big data problems.

The median is familiar in one dimension: it is frequently used in place of the mean because of its robustness to outliers and skewed data. In high dimensions, however, the median is difficult to *define* and more so to *compute*. Natural approaches like computing the median in each coordinate of the data individually do not to live up to the standard of robustness set by the one-dimensional median.

In the 2010s, statisticians found a new definition of high-dimensional median, building on foundational work going back to the 1960s. In a precise sense, this new median is the first to bring the robustness of the one-dimensional

> CONTINUED FROM PAGE 2 [Miller Fellow Focus]

median to the high-dimensional setting. The key is that it accounts for the need to be near the middle of the data *no matter what coordinates you use.*

Imagine a 2-dimensional *data* set $(x_1,y_1),\ldots,(x_n,y_n)$. Naturally, if (a,b) is its median, then a should be near the middle of $x_1,\ldots,x_n$ and similarly for b and the y's. But to achieve robustness, something stronger is needed: $a+b$ should be near the middle of $x_1+y_1,\ldots,x_n+y_n$, and $0.3a-1.2b$ near the middle of $0.3x_1-1.2y_1,\ldots,0.3x_n-1.2y_n$, and so on for any (linear) way of putting together the coordinates **(FIGURE 3)**.

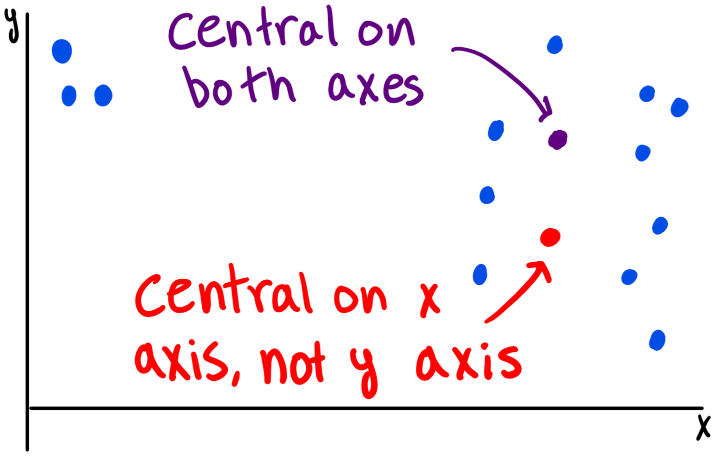

Figure 3: a good median is near the middle of the data in every direction.

This creates a computational problem: in a data set with d dimensions, there are about 2^d ways to put the coordinates together. The exponential growth again means that naive algorithms to compute a median are intractable. In my first year as a Miller Fellow, I developed the first computationally efficient algorithm to compute this new high-dimensional median.

The rise of big data in the last decade has been largely engineering-driven – new algorithms and applications are often discovered by "hacking until it works". The expanding breadth and gravity of its impacts demand that we tackle a grand challenge: creating a rigorous and foundational theory of algorithms for big data, to help make principled decisions about what algorithms to use in what circumstances, and to reason with mathematical confidence about the consequences.

Sam Hopkins is a second-year Miller Fellow in the Electrical Engineering and Computer Science Department, advised jointly by Prof. Luca Trevisan and Prof. Prasad Raghavendra. He received his PhD in Computer Science from Cornell University, and his BS in Math and Computer Science from the University of Washington in his hometown of Seattle. He likes whiteboards and bicycles.

Contact: hopkins@berkeley.edu

In the News

(see more current & past Miller Institute news: **miller.berkeley.edu/news**)

James Peebles (Visiting Miller Professor 1987) was awarded the **2019 Nobel Prize in Physics** *"for theoretical discoveries in physical cosmology."*

Jennifer Doudna (Miller Senior Fellow 2017) was honored with the **2019 Life Sciences Leadership Award** as California's most innovative and dedicated life sciences leader for her ongoing contributions to California's life sciences sector. She was also recognized with the **2019 Welfare Betterment Prize,** a relatively new Hong Kong-based prize, for her pioneering discovery of CRISPR-Cas9 gene editing.

Birgitta Whaley (Miller Professor 2002 - 2003) was among **seven new advisers appointed to the U.S. President's Council of Advisors on Science and Technology (PCAST)** and is a foremost expert in the fields of quantum information, quantum physics, molecular quantum mechanics and quantum biology.

Kam-Biu Luk (Miller Professor Fall 2001) won the **2019 Future Science Prize** for his role in the neutrino discoveries of the past two decades, especially his leadership of the Daya Bay experiment.

Norman Yao (Miller Fellow 2014-2017) was awarded the **2020 George E. Valley, Jr. Prize** by the American Physical Society *"For the elucidation of non-equilibrium quantum phases of matter, in particular time crystalline order, and for enabling the realization of these phases in quantum optical systems."*

Jerry Mitrovica (Visiting Miller Professor 2004) was named a **2019 MacArthur Fellow** for *"revising our understanding of the dynamics and structure of Earth's interior and developing models to better predict the geometry and sources of sea level change in the modern world and the geological past."*

Miller Members named winners of the **2020 Breakthrough Prize:**

- **Xie Chen** (Miller Fellow 2012-2014) was recognized among other physicists with the **2020 New Horizons in Physics Prize** *"For incisive contributions to the understanding of topological states of matter and the relationships between them."*
- **Sergio Ferrara** (Visiting Miller Professor 2008) was honored with a **Special Breakthrough Prize in Fundamental Physics** to recognize the discovery of the theory of supergravity.
- **Feryal Özel** (Visiting Miller Professor 2014, Advisory Board Member 2017-Present) was recognized as one of **The Event Horizon Telescope Collaboration Prizewinners.**

NEWS CONTINUED ON PAGE 5 >

Miller Research Competitions: Awards

The Advisory Board

On December 2, 2019, the Advisory Board of the Miller Institute met to select next year's Professorship awards. The Board is comprised of four advisors external to UCB: Steven Block (Physics, Stanford University), Luis Caffarelli (Mathematics, University of Texas, Austin), Feryal Özel (Astronomy & Physics, University of Arizona) and Tim Stearns (Biology, Stanford University); and four internal Executive Committee members: Executive Director Marla Feller (Molecular & Cell Biology), Stephen Leone (Chemistry/Physics), Roland Burgmann (Earth & Planetary Science) and Yun Song (EECS/Statistics/IB). The Board is chaired by Chancellor Carol Christ.

The Miller Institute is proud to announce the awards for Professorship terms during the Academic Year 2020-2021. These outstanding scientists pursue their research, following promising leads as they develop. The Visiting Miller Professors join faculty hosts on the Berkeley campus for collaborative research interactions.

Miller Professorship Awards

Steven Beissinger
Environmental Science, Policy and Management

William Boos
Earth & Planetary Science

Suncica Canic
Mathematics

Abby Dernburg
Molecular & Cell Biology

Oskar Hallatschek
Physics

Holger Mueller
Physics

James Olzmann
Nutritional Sciences & Toxicology

Sug Woo Shin
Mathematics

Feng Wang
Physics

Visiting Miller Professorship Awards

Immanuel Felix Bloch

Physics

Host: Dan Stamper-Kurn

Home Institution: Ludwig Maximilian University & Max Planck Institute of Quantum Optics, Germany

David Fisher

Mathematics

Host: Ian Agol

Home Institution: Indiana University, Bloomington

.

Zeljko Ivezic

Astronomy

Host: Joshua Bloom

Home Institution: University of Washington

Astrid Kiendler-Scharr

ESPM

Host: Allen Goldstein

Home Institution: Institute for Energy and Climate Research (IEK-8: Troposphere), Germany

John Thomas Lis

Molecular & Cell Biology

Host: Xavier Darzacq

Home Institution: Cornell University

Lorraine Pillus

Molecular & Cell Biology

Host:Jasper Rine

Home Institution: UC San Diego

Katharine Nash Suding

Integrative Biology

Host: Todd Dawson

Home Institution: University of Colorado, Boulder

David Wales

Chemistry

Host: Richard Saykally

Home Institution: University of Cambridge

David Weinberg

Astronomy

Host: Daniel Weisz

Home Institution:
The Ohio State University

Roland Wester

Somorjai Visiting Miller Professor

Chemistry

Host: Daniel Neumark

Home Institution: University of Innsbruck

> CONTINUED FROM PAGE 3 [In the News]

In the News

Ehud Isacoff (Miller Professor 2013) received the **2020–22 McKnight Award** for his work on photo-activation of dopamine receptors in models of Parkinson's Disease.

Nicole King (Miller Professor 2018-2019) was named a **2019 Pew Trust Innovation Fund Investigator** to explore the innate immune response in aquatic, unicellular organisms called choanoflagellates.

Nicola Spaldin (Miller Professor 2007) was awarded the **2019 Swiss Science Prize by the Marcel Benoist Foundation** *"for her ground-breaking research in multiferroic materials, with which she has laid the foundations for new ultrafast and energy-efficient data storage technologies."*

Paul Alivisatos (Miller Professor 2001-2002) was named a recipient of the prestigious **2019 Robert A. Welch Award in Chemistry** "*for his important research contributions in the fields of nanoscience and nanotechnology which have had a significant, positive impact on humankind."*

Arash Komeili (Miller Professor 2016-2017) was named one of the **2019-2020 Bakar Fellows** for engineering bacteria to efficiently isolate metals from minerals, thereby minimizing the environmental damage typical of traditional mining.

JoAnne Stubbe (Visiting Miller Professor 2020) was named **2020 Priestley Medalist** as the top mechanistic biochemist of her generation with the American Chemical Society's highest honor.

Ehud Altman (Visiting Miller Professor 2012) and **Dung-Hai Lee** (Miller Professor 1999) received **Moore Foundation Award titled *"Emergent Phenomena in Quantum Systems Theory Center."***

Daniel Fletcher's (Miller Professor 2019-2020) bioengineering laboratory **was awarded a $1.9 million Gates Foundation grant** to support the scaled-up production of the LoaScope.

Kenneth Ribet (Miller Professor 1990) **was honored by Brown University in a new plaque celebrating his contribution to the University's Open Curriculum.**

Yu He (Miller Fellow 2019 - 2022) **co-authored a paper,** *"Electronic map reveals 'rules of the road' in superconductor"*, published in the American Physical Society Journal.

Doug Hemingway (Miller Fellow 2015 - 2018) **co-authored the article** *"Cascading parallel fractures on Enceladus"* published in Nature Astronomy.

Please email news to: miller_adm@berkeley.edu.

Annual Fall Dinner 2019

Miller Fellow Yu He, Tang Tang, Miller Fellow Alumna Becca Tarvin, Miller Fellows Alex Turner & Thibault de Poyferre, Miller Fellow Alumnus Simone Ferraro, Miller Fellows Georgios Moschidis & Ruby Fu, Miller Fellow Alumna Sarah Slotznick and Miller Fellow Naomi Latorraca

Miller Senior Fellow Raymond Jeanloz, Miller Fellow Sho Takatori, retired Chief Administrative Officer Kathy Day & Chancellor Carol Christ

Professor Chris Hoofnagle, Miller Professor Linda Wilbrecht, Miller Fellow Louis Kang, Miller Professor Hartmut Haeffner & Professor Mike DeWeese

Miller Senior Fellow Susan Marqusee & Miller Fellow Alumnus Chris Lemon

Gifts to the Miller Institute

The Miller Institute gratefully acknowledges the following contributors to the Miller Institute programs in 2019. With your generosity, the Miller Institute is able to continue to support basic research in science at UC Berkeley.

Kathryn A. Day Miller Postdoctoral Fellowship Fund

The Kathryn A. Day Miller Postdoctoral Fellowship was established with a generous gift by Nobel Laureate Professor Randy Schekman and Professor Sabeeha Merchant to honor Kathy Day, who served as the Chief Administrative Officer at the Miller Institute for Basic Research in Science from 1989 - 2019. The purpose of the Fund is to provide an annual stipend, benefits and a research fund to a postdoctoral researcher at the Miller Institute who has demonstrated efforts towards community building and outreach in support of science.

Anonymous (2)
Rachel Akeson
David Aldous
Andreas Bausch
Robert Bergman
Douglas Black
Roger Blandford
Roland Burgmann
Mary (Beth) Burnside
Kathleen Collins
Rebekah Dawson
Kathy Day
Dmitry Dolgopyat
Jennifer Doudna and Jamie Cate
Christopher Douglas
August Evrard
Marla Feller
Alexei Filippenko - in memory of Judey Miller
Mary K Gaillard
Brooke Gardner
Britt Glaunsinger
Richard Harland
Cassandra Hunt
Sharon Inkelas
Russell Jones
Thomas Juenger
Judith Klinman
Tsit-Yuen Lam
Stephen Leone

Michael Manga
Susan Marqusee
Frederick Matsen
Sébastien Merkel
Nancy Missert
Akinao Nose
Sarah Otto
Vijay Pande
Ingrid Parker
Jonas Peters
Catherine Pfister and J. Timothy Wootton
Thomas Pollard
Anne Pringle
Jessica Ray
Christine Read
Adam Retchless
Jasper Rine
Dustin Rubenstein
Richard Saykally and Christine Read
Joshua Shaevitz
Terence Speed
Stephen Stearns
Jesse Thaler
Jeremy Thorner
Jeffrey Townsend
Danqing Wang
Kenneth Watcher
Rebecca White
Patricia Zambryski
Yuanbo Zhang

> CONTINUED FROM PAGE 6 [Gifts to the Miller Institute]

Gabor A. and Judith K. Somorjai Visiting Miller Professorship Award Fund

The purpose of the Somorjai Visiting Miller Professorship Award is to support the collaborative research of an early-career visiting scientist within the broad field of chemical sciences. Initiated by a generous gift from Professor and Mrs. Somorjai, this fund supports a visiting professor for a one-month term on campus.

X A Markenscoff
Emmanouil Mavrikakis

Miller Institute for Basic Research in Science General Fund

The Miller Institute for Basic Research in Science is dedicated to the encouragement of creative thought and the conduct of research and investigation in the field of pure science. Contributions to this fund will support the four programs of the Miller Institute: the Miller Research Fellowship, the Miller Professorship, the Visiting Miller Professorship, and the Miller Senior Fellowship.

Roger and Elizabeth Blandford
Justin Brown
William Carter
Joel Ellis
Marla Feller
Stephen Glickman
Gilbert Hawkins
Arash Komeili
Barbara J. Meyer (The Byron R Meyer Living Trust)
Charles Nicoll
Andrew Ogg
Eve Ostriker
Richard Roberts
Stephen Suh - in honor of Jonathan Suh
Salil Vadhan
Fausta Segre-Walsby and Anthony Walsby
Norman Yao

Miller Fellowship Program Development Fund

The Miller Fellowship Program Development Fund provides an annual stipend, benefits, and research support to young researchers at Berkeley. The program gives researchers the chance to explore ideas in a stimulating and supportive environment.

Anonymous
Dmitry Dolgopyat
Jiaxing Huang - in honor of Peidong Yang
Ronald Johnson
Yasuyuki Kawahigashi
Takahiro Kawai
Nancy Steinhaus - in memory of Edward A. Steinhaus

University of California, Berkeley
Miller Institute for Basic Research in Science
468 Donner Lab
Berkeley, CA 94720-5190
510.642.4088
miller.berkeley.edu

Miller Institute News : Winter 2020
Please send address corrections to:
miller_adm@berkeley.edu

Where in the World Is Your Fleece?

Please share a photo of you wearing your Miller Institute fleece and we will post these photos on our website, highlighting the worldwide reach of the Miller Institute!

Please email photos to ***millerinstitute@berkeley.edu***

Sho Takatori (Miller Fellow 2017-2020), Eva Schmid (Miller Fellow 2008 - 2011) and Dan Fletcher (Miller Professor 2019-2020) showing off their Miller pride!

Emeritus Miller Institute Members!

Do you wish to receive the Miller Institute newsletter at your home address? Please email millerinstitute@berkeley.edu with your updated address information so as not to miss a single issue.

Next Steps

Rebecca (Becky) Jensen-Clem
Assistant Professor
Department of Astronomy and Astrophysics
UC Santa Cruz

Sho Takatori
Assistant Professor
Chemical Engineering Department
UC Santa Barbara (starting March 2020)

Make a Gift

Private donations are becoming an increasingly significant resource for the Miller Institute. Your personal investment in support of the future of the Miller Institute will be greatly appreciated.

Join Miller friends and alumni in contributing to this important endeavor by logging on to ***miller.berkeley.edu/gift*** to help support the independent research of the Miller Institute members.