Platforms, Place and Your Profession

Prof. Paul Groth | @pgroth | pgroth.com | indelab.org

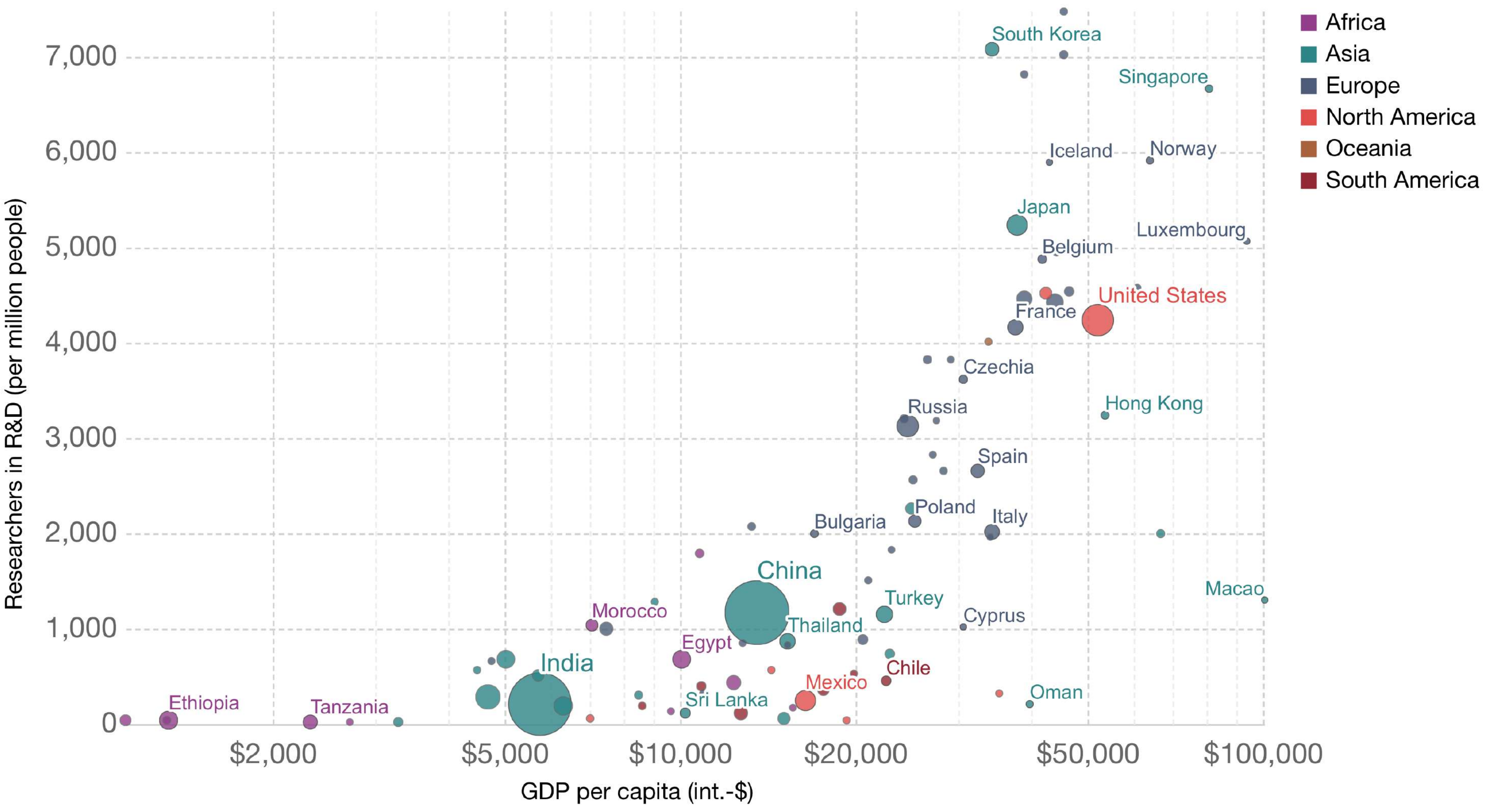
ESWC 2021 PhD Symposium
June 7, 2021

INDE lab
INtelligent Data Engineering
University of Amsterdam

The world needs you!

Researchers in Research and Development vs GDP per capita, 2015

Researchers in Research & Development (R&D) are professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects concerned. Postgraduate PhD students engaged in R&D are included.



Source: World Bank

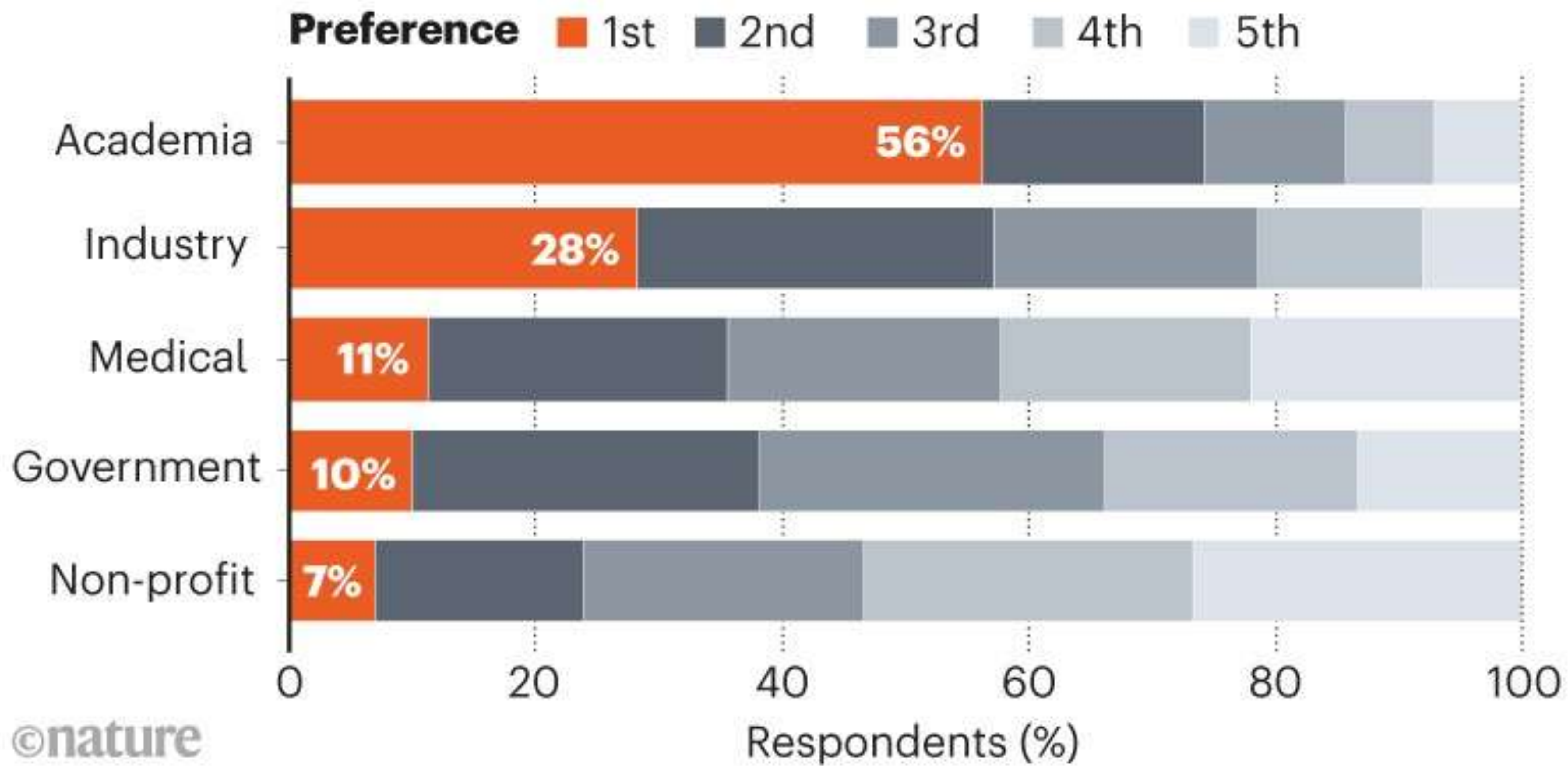
See also the importance of ideas: https://www.nobelprize.org/prizes/economic-sciences/1995/lucas/biographical/

What do you want to be when you grow up?

ACADEMIC DREAMS

PhD students around the world continue to aspire to careers in academia despite a global job crunch. Industry — a growing job sector for PhD scientists — rates a distant second.

Q: Which of the following sectors would you most like to work in (beyond a postdoc) when you complete your degree?



Source: https://www.nature.com/articles/d41586-019-03459-7

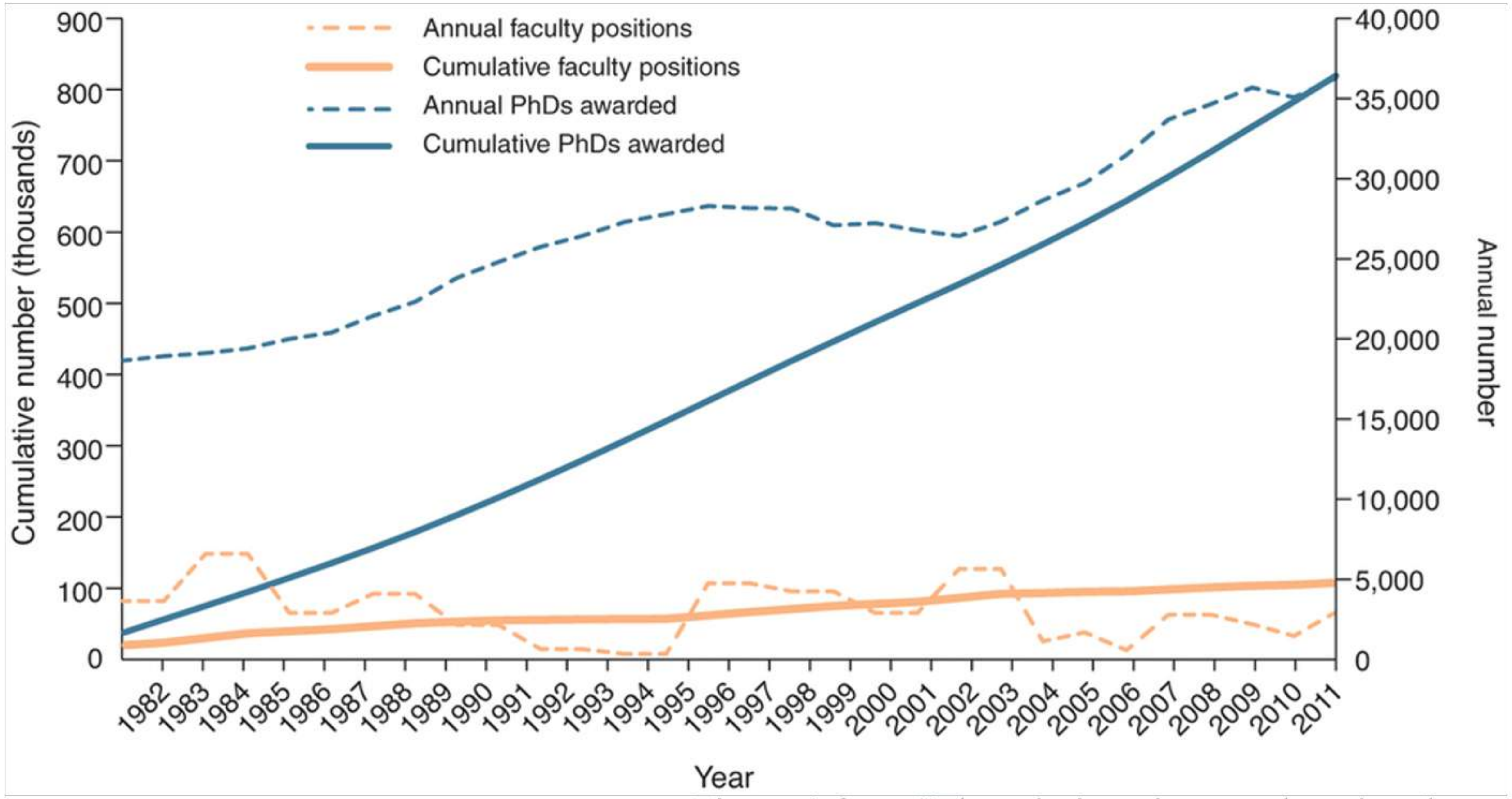
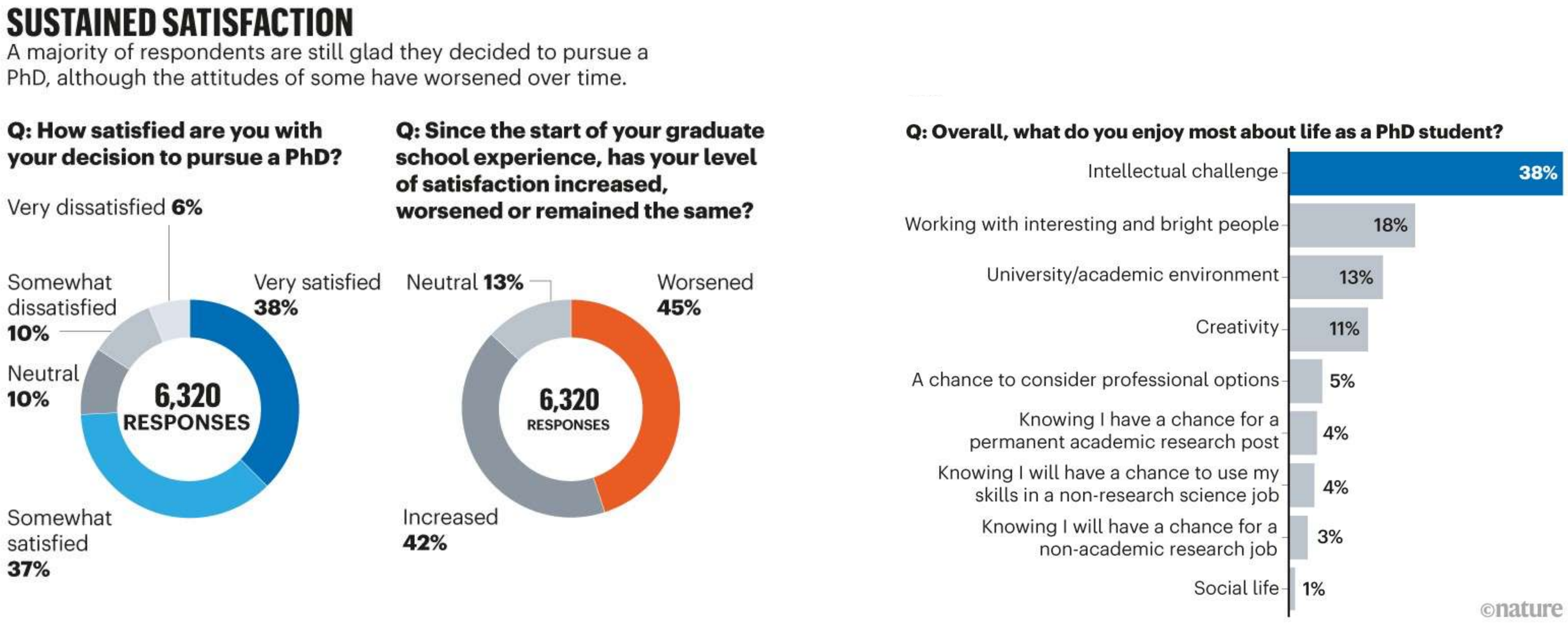


Figure 1 from "The missing piece to changing the university culture."
M. Schillebeeckx *et al.* Nature Biotechnology 31, 938–941 (2013)



Source: https://www.nature.com/articles/d41586-019-03459-7

What you do is different than

Where you do it

Things that a PhD can help you do from PhDs not in academia

Startup Founder
Product Manager
VP Business Development
Communication
Industry Researcher
Enterprise Architect
Head/VP of Data Science
Consultant

See a nice list of other positions for phd holders:
https://medium.com/bits-and-behavior/most-ph-d-s-arent-professors-13a741ef6868

Problem Solving

- Taught me how to identify a problem and define the steps to tackle it
- Learned the ability to go deep and also think high level
- It helped me gain experience in picking up complex new ideas from research literature quickly
- Critical thinking and evidence-based decision making, ability to know when to seek confirmation or alternative sources of information
- Use data in a meaningful way
- Exposed me to complex areas of mathematics I wouldn’t have been exposed to as an undergrad. Math is transferable. A Phd gives you time and freedom to explore these tools that you wouldn’t get in industry.
- Ability to engage in many different areas of science, identify what I don’t yet know/understand
- Engage in targeted learning

Work Habits

- PhD is much deeper than any general project one might encounter, but the attention to detail that companies' demand almost matches that of PhD. Use this to your advantage, especially in interviews, but also in day to day work.

- Combine your PhD with work experience where possible. Real world results from your theoretical experiments are arguably more valuable validation signals than peer-review.

- Be able to follow through on long-term projects

- Be able to think critically

- Be able to plan and provide proof point

- Not having to be managed but becoming more pro-active

Communication

- Taught me how to clearly communicate my ideas and findings to my colleagues, both through talks and papers
- In the course of a PhD, you have time to perfect your message. Not so in work environment. Be prepared to let go of your perfectionism.
- Be able to structure arguments
- It helped me gain experience in forming theories on complex processes where understanding is incomplete subsequently defend that position with data, logic and experimental results. In the current evolution of century old business' morphing into data driven organisations, this is becoming increasingly valuable
- PhDs that have experience navigating cross-disciplinary domain can often help resolve a lot of the communication challenges that plague large companies – e.g. by teaching other teams to adopt existing techniques

Entrepreneur Mindset

- A PhD positions you to be at the forefront of your field. Be on the lookout for any business opportunities at every step.
- Startups – especially the ones that are both in the hype curve and achieved significant vc-funding – often hire PhDs because they bring in the field-expertise of working with novel technologies
- Confidence of managers in the communication skills and work habits above
- Gives me credibility in my projects with academics in the field
- The connections I made in academia are still relevant
- A PhD positions you as an expert in your field. The confidence this brings is extremely helpful in the work environment. Remember to position yourself as an expert, where appropriate.

What do you like to do?

Build your platform

Help others

Do (mostly) what I like

Support my life priorities

Connect to a community

Expand my skills

Reach my goals

Your position

Build your platform

??????????????????????

Help others | Do (mostly) what I like | Support my life priorities

Connect to a community | Expand my skills | Reach my goals

Your position

Paul Groth

Professor of Data Science

Amsterdam, North Holland, Netherlands · Contact info

Professor of Algorithmic Data Science
University of Amsterdam
Nov 2018 – Present · 2 yrs 8 mos
Amsterdam Area, Netherlands

I lead the Intelligent Data Engineering Lab (http://indelab.org). Our lab investigates intelligent systems that support people in their work with data and information from diverse sources.

Disruptive Technology Director, Elsevier Labs
Elsevier
Jan 2015 – Sep 2018 · 3 yrs 9 mos
Amsterdam, Netherlands

In this role, I did research and advanced prototyping focused on how technology can improve science. I advised product teams and senior management on technology decisions and trends. In addition, I regularly engaged with external groups and collaborators around these themes. ...see more

Vrije Universiteit Amsterdam
5 yrs

Assistant Professor
Jun 2011 – Dec 2014 · 3 yrs 7 mos
Amsterdam Zuidoost, Provincie Noord-Holland, Netherlands

Researching new methods for dealing with diverse contextual data. Including data integration, knowledge integration, data science and data provenance.

Highlights ...see more

Open Data: Ready Set Go

Altmetrics Overview

Postdoc
2009 – Jul 2011 · 2 yrs

- Led the interdisciplinary Semantically Mapping Science project where I developed novel network analysis based methods applied to knowledge about scientific activity extracted from the Web.
- I developed the technical underpinnings for nanopublications - a way of repres ...see more

Total Impact - aggregating measure...

Thoughts on A First EU Proposal

Postdoctoral Research Associate
Information Sciences Institute, University of Southern California
Oct 2007 – Aug 2009 · 1 yr 11 mos

University of Southampton
Ph.D., Computer Science
2004 – 2007
Activities and Societies: School of Electronics and Computer Science Graduate School Board, Treasurer and Player Tennis Team, Table Tennis Team

Worked on both the PASOA (http://www.pasoa.org/) and EU Provenance (http://www.gridprovenance.org) projects.

University of West Florida
B.Sc. (Hon), Computer Science
1999 – 2003

INDE lab

INtelligent Data Engineering

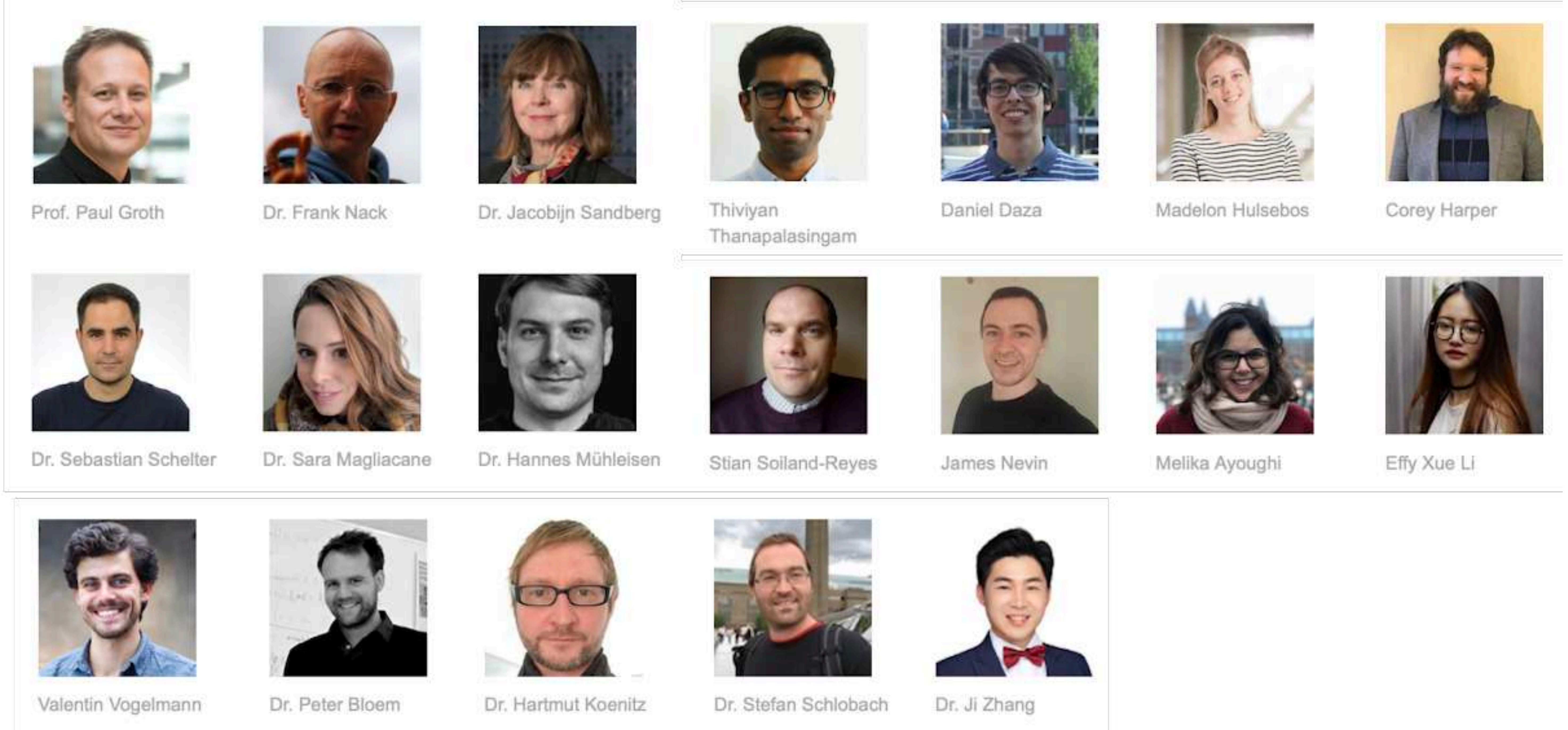
University of Amsterdam

We investigate intelligent systems that support people in their work with data and information from diverse sources.

In this area, we perform applied and fundamental research informed by empirical insights into data science practice.

Current topics:

- Automated Knowledge Base Construction
- Data Search + Data Provenance
- Data Management for Machine Learning
- Causality for machine learning on messy data



indelab.org

Sara Magliacane · 1st

Assistant professor at University of Amsterdam & Research Scientist at MIT-IBM Watson AI Lab

Experience

Assistant Professor
University of Amsterdam
Nov 2020 – Present · 8 mos
Amsterdam, North Holland, Netherlands

Research Scientist
MIT-IBM Watson AI Lab
Apr 2019 – Present · 2 yrs 3 mos

- Co-PI on exploratory MIT-IBM project with Armando Solar-Lezema (MIT) and Nathan Fulton (MIT-IBM) on safe AI approaches and program synthesis
- Continuing work on core MIT-IBM project on causality

Postdoctoral Researcher in AI Foundations team
IBM Thomas J. Watson Research Center · Full-time
Nov 2017 – Apr 2019 · 1 yr 6 mos
Yorktown Heights, New York

- Part of core MIT-IBM project on learning causal graphs from data, experiment/intervention design, causal transfer learning with Caroline Uhler (MIT) and Guy Bresler (MIT)
- Part of core MIT-IBM project on neuro-symbolic approaches, learning logic r ...see more

Researcher in th Causality Group
University of Amsterdam
Mar 2016 – Nov 2017 · 1 yr 9 mos

- Research on causal transfer learning and causal structure learning from noisy data in different experimental settings with Joris Mooij
- Designed and taught a new joint UvA/VU course on neuro-symbolic approaches: "Combining symbolic and statistical representations in AI" with Frank van Harr ...see more

Software Engineer Intern
Google Research
May 2014 – Aug 2014 · 4 mos
New York

Extracting information from semi-structured data in the WebTables team (hosts: Cong Yu, There’s lots of opportunity at the border

Twin Win Model of Research

Ben Shneiderman. (2019) Twin-Win Research: Breakthrough Theories and Validated Solutions for Societal Benefit, Second Edition

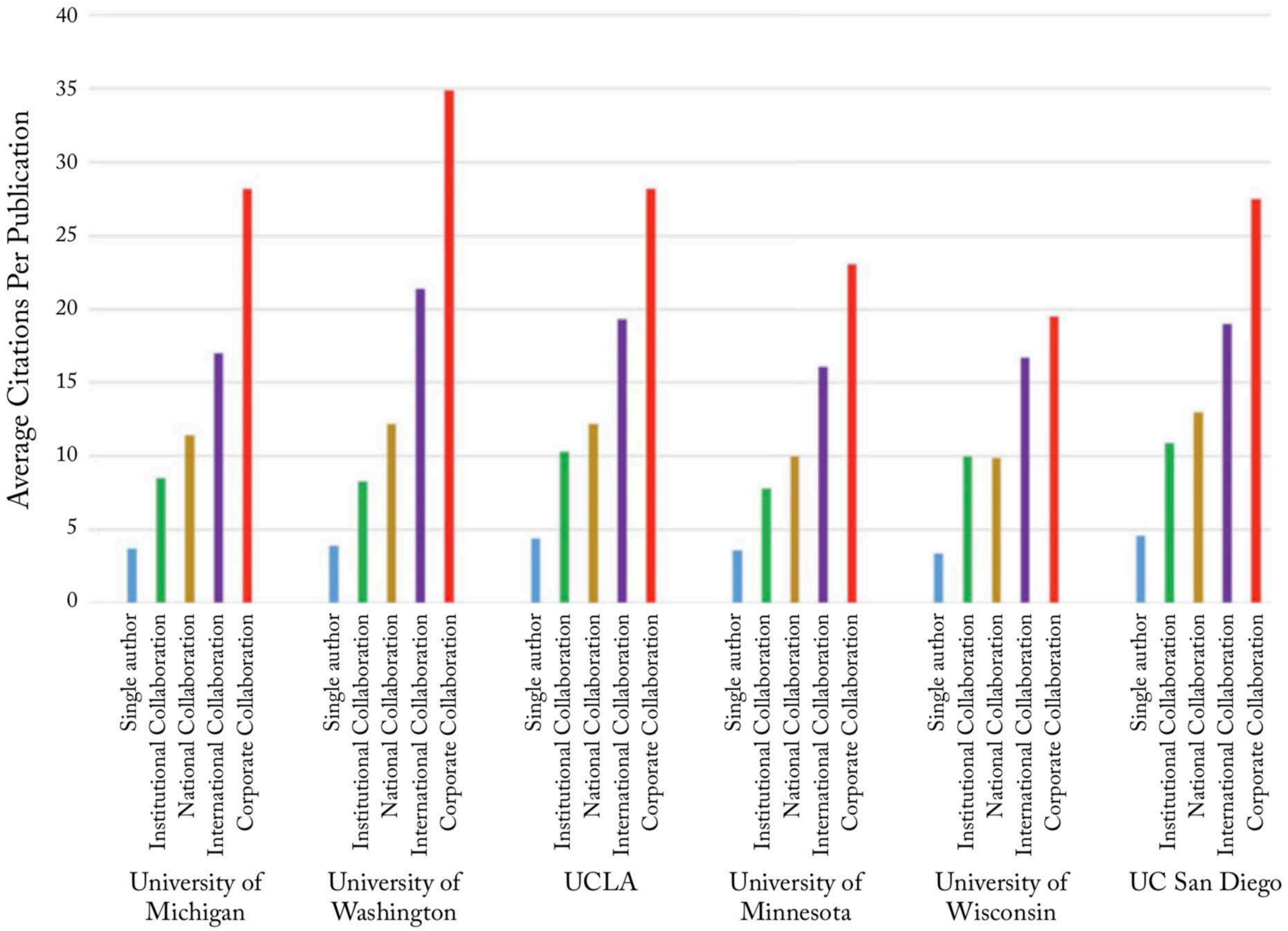
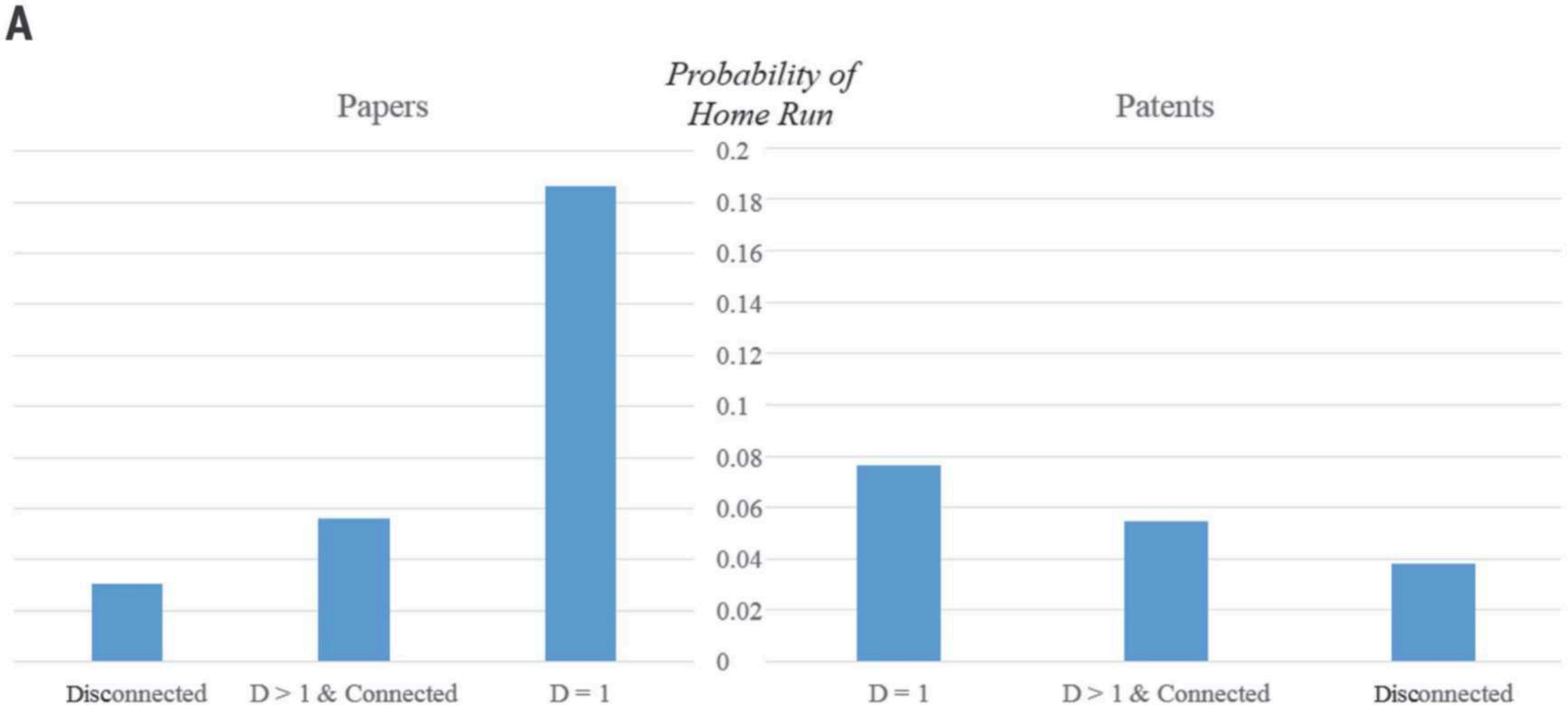


Figure 1.2: Effect of co-authorship on citations: top six public universities in the U.S.

Hitting a home run

Ahmadpoor et al., Science 357, 583–587 (2017)



B

	Distance where Home Run Rate is Largest, by Field or Class		
	$D = 1$	$D > 1$	Disconnected
Papers (WOS fields)	99%	1%	0%
Patents (USPTO classes)	86%	14%	0%

Fig. 3. Distance and impact. (**A**) Impact close to and far from the paper-patent boundary. A "home run" is defined as being in the upper 5% of citations received in that field and year, for a patent or a research paper. (**B**) Home-run outcomes relative to distance for each field, when each fielc is analyzed separately. The supplementary materials examine alternative impact measures, including methods based on patent-renewal payments.

Sebastian Schelter · 1st

Assistant Professor at University of Amsterdam

Assistant Professor
University of Amsterdam
Apr 2020 – Present · 1 yr 3 mos
Amsterdam, North Holland, Netherlands

Assistant Professor with the University of Amsterdam, conducting research at the intersection of data management and machine learning. I manage the "AI for Retail Lab" Amsterdam. Visit my homepage at https://ssc.io for more details.

Research Fellow
Ahold Delhaize · Part-time
Apr 2020 – Present · 1 yr 3 mos
Amsterdam, North Holland, Netherlands

Member of the Apache Software Foundation
The Apache Software Foundation
Jul 2010 – Present · 11 yrs

Elected member of the Apache Software Foundation, where I have been involved in a variety of projects: Apache Mahout (machine learning), Apache Giraph (graph processing), Apache Flink (stream processing). I have additionally helped to start the the MXNet & TVM projects (a deep learning engine & compiler).

New York University
1 yr 6 mos

Assistant Professor ("Faculty Fellow")
Sep 2019 – Apr 2020 · 8 mos
New York

Faculty Fellow with the Center for Data Science, conducting independent research on the intersection of data management and machine learning, with the interdisciplinary application to computational social science. Design and teaching of a master's course on "Data Engineering for Machine Learning".

Moore-Sloan Data Science Fellow
Nov 2018 – Sep 2019 · 11 mos
New York

Amazon
4 yrs 7 mos

Senior Applied Scientist
Part-time
Nov 2018 – Apr 2020 · 1 yr 6 mos
New York

Scalable data validation used in SageMaker Model Monitor https://aws.amazon.com/blogs/aws/amazon-sagemaker-model-monitor-fully-managed-automatic-monitoring-for-your-machine-learning-models/

Applied Scientist
Part-time
Oct 2015 – Oct 2018 · 3 yrs 1 mo
Berlin und Umgebung, Deutschland

Applied Scientist in Amazon's Core Machine Learning team in Berlin, with a focus on data management issues of end-to-end machine learning applications.

Senior Researcher & Guest Lecturer
Technische Universität Berlin · Part-time
Oct 2015 – Oct 2018 · 3 yrs 1 mo
Berlin und Umgebung, Deutschland

Senior Researcher / Guest lecturer with the Database Systems and Information Management Group of TU Berlin.

Research Associate / PhD student
Technische Universität Berlin
May 2011 – Oct 2015 · 4 yrs 6 mos
Berlin Area, Germany

Research in the area of large scale data analysis and parallel processing platforms at the Database Systems and Information Management group (DIMA). Implemented the runtime for iterative batch computations in Apache Flink. PhD on "Scaling Data Mining in Massively Parallel Dataflow Systems" with "summa cum laude" (best possible grade).

Software Engineering Intern
Twitter · Internship
Jul 2014 – Sep 2014 · 3 mos
San Francisco Bay Area

Learnings from a Retail Recommendation System on Billions of Interactions at bol.com

Barrie Kersbergen Sebastian Schelter
Ahold Delhaize Research & AIRLab, University of Amsterdam
bkersbergen@bol.com s.schelter@uva.nl

***Abstract*—Recommender systems are ubiquitous in the modern internet, where they help users find items they might like. We discuss the design of a large-scale recommender system handling billions of interactions on a European e-commerce platform.**

We present two studies on enhancing the predictive performance of this system with both algorithmic and systems-related approaches. First, we evaluate neural network-based approaches on proprietary data from our e-commerce platform, and confirm recent results outlining that the benefits of these methods with respect to predictive performance are limited, while they exhibit severe scalability bottlenecks. Next, we investigate the impact of a reduction of the response latency of our serving system, and conduct an A/B test on the live platform with more than 19 million user sessions, which confirms that the latency reduction of the recommender system correlates with a significant increase in business-relevant metrics. We discuss the implications of our findings with respect to real world recommendation systems and future research on scalable session-based recommendation.

I. Introduction

Today's internet users face an ever increasing amount of information.This situation has triggered the development of recommender systems: intelligent filters that learn about the users' preferences and suggest relevant information for them. With rapidly growing data sizes, the predictive performance, processing efficiency and scalability of machine learning-based recommendations systems and their underlying computations becomes a major concern.

In this paper, we describe the architecture of a real world recommender system ABO for bol.com, a large European e-commerce platform which handles billions of interactions on several dozen million items every day in Section II. The ABO ('Anderen bekeken ook', Dutch for 'others also viewed') recommendations are shown on the product detail page[1] to enable customers to discover other products that are relevant to them, such as different versions of the same product, similar products or products that are complementary to the displayed item. We describe the individual components of our system, which are backed by cloud infrastructure from the Google Cloud Platform such as BigTable and BigQuery. In addition, we detail our nearest-neighbor-based recommendation approach, we discuss how we conduct distributed offline model training, and how we efficiently serve the recommendations online with low latency.

A natural question when operating such a real world recommendation system is how to improve its predictive performance. In this work, we explore two directions for improvement, and present the results of two corresponding studies. First, we investigate the potential of *algorithmic improvements* in Section III. Neural networks have shown outstanding performance in computer vision [1] and natural language processing tasks [2], and we therefore evaluate recently proposed neural network-based approaches [3]–[6] for session-based recommendation on real data from our platform, based on an existing academic study [7]. We evaluate the predictive performance of these neural networks, as well as their deployability for production settings, in terms of training time, cost of hyperparameter search, prediction latency and scalability. Next, we study a *system-specific improvement*: we do not change the recommendation algorithm itself, but optimise our serving infrastructure to drastically reduce its response latency (Section IV). We describe how we control the insertion rate of bulk updates into the production database of our recommendation system, in order to adhere to a latency SLA (service-level agreement) of 50ms for recommendation responses. We run a large-scale online A/B test on 19 million user session to investigate the impact of this response latency reduction on the predictive performance of our recommender system. In summary, we provide the following contributions:

- We discuss the design of a large-scale recommender system handling billions of interactions on a European e-commerce platform (Section II).
- We present two studies on enhancing the predictive performance of this system: (*i*) We evaluate recent neural network-based approaches on proprietary data from our e-commerce platform, and confirm recent results outlining that the benefits of these methods with respect to predictive performance are limited, while they exhibit severe scalability bottlenecks (Section III); (*ii*) We optimise the response latency of our serving system, and conduct an A/B test on the live platform with more than 19 million user sessions, which confirms that the latency reduction correlates with a significant increase in metrics based on purchases and revenue (Section IV).
- We discuss the implications of our findings with respect to real world recommendation systems, as well as future research on session-based recommendation (Section VI)

[1] https://www.bol.com/nl/p/-/9200000104430048

Elastic Machine Learning Algorithms in Amazon SageMaker

Edo Liberty, Zohar Karnin, Bing Xiang, Laurence Rouesnel, Baris Coskun, Ramesh Nallapati, Julio Delgado, Amir Sadoughi, Yury Astashonok, Piali Das, Can Balioglu, Saswata Chakravarty, Madhav Jha, Philip Gautier, David Arpin, Tim Januschowski, Valentin Flunkert, Yuyang Wang, Jan Gasthaus, Lorenzo Stella, Syama Rangapuram, David Salinas, Sebastian Schelter, Alex Smola
Amazon AI
{libertye,zkarnin,bxiang,rouesne,barisco,rnallapa,juliod,sadoughi,yastasho,pialidas,balioglu,saswatac, madhavjh,gautierp,arpin,tjnsch,flunkert,yuyawang,gasthaus,stellalo,rangapur,dsalina,sseb,smola}@amazon.com

ABSTRACT

There is a large body of research on scalable machine learning (ML). Nevertheless, training ML models on large, continuously evolving datasets is still a difficult and costly undertaking for many companies and institutions. We discuss such challenges and derive requirements for an industrial-scale ML platform. Next, we describe the computational model behind Amazon *SageMaker* which is designed to meet such challenges. *SageMaker* is an ML platform provided as part of Amazon Web Services (AWS), and supports incremental training, resumable and elastic learning as well as automatic hyperparameter optimization. We detail how to adapt several popular ML algorithms to its computational model. Finally, we present an experimental evaluation on large datasets, comparing *SageMaker* to several scalable, JVM-based implementations of ML algorithms, which we significantly outperform with regard to computation time and cost.

ACM Reference Format:
Liberty et al.. 2020. Elastic Machine Learning Algorithms in Amazon SageMaker. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD'20), June 14–19, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3318464.3386126

ACM ISBN 978-1-4503-6735-6/20/06…$15.00
https://doi.org/10.1145/3318464.3386126

1 INTRODUCTION

Machine learning (ML) has become an integral part of modern software systems. Unfortunately, training ML models on large, continuously evolving datasets is still a significant undertaking for many companies and institutions, especially if ML is not their core competency. Building an industrial-scale model training platform for such cases involves a set of challenges, many of which are not addressed by current systems available in academia and open source.

(*i*) Support for *incremental training and model freshness:* It is highly uncommon to encounter large static datasets. In most cases, data keeps being generated constantly, which is often addressed with an unwelcome trade-off between training cost and accuracy. Training on a large subset of the data produces accurate models but can become extremely costly and slow, while training on new, small updates of the data (e.g., the last day) is cheaper but might not lead to very accurate results. Therefore, industrial ML platforms have to support incremental model training to regularly and cost-efficiently update existing models and to quickly provide accurate and 'fresh' models.

(*ii*) *Predictability of training costs:* for large amounts of data, customers need to be able to roughly estimate in advance how much a training job would cost and how long it would run for. It is difficult to estimate the cost ahead of time for many scalable systems, which do not support incremental learning with linear update times or have irregular performance drops for high-dimensional models [5].

(*iii*) *Elasticity and support for pausing and resuming training jobs:* large-scale ML scenarios often result in imbalanced workloads, where data scientists spend several days without running a single job (while they are collecting data or writing code) and then they launch several large concurrent training jobs on hundreds of machines. They also may want to pause and resume such jobs, e.g., for hyperparameter tuning or if

Conclusion

- The world needs you!
- Use your PhD period to learn - lots of transferable skills
- Think about your positions as helping to build a platform to achieve your life goals
- There's lots of opportunity at the border
  - real problems often poise extremely interesting research problems

Paul Groth | @pgroth | pgroth.com | indelab.org

INDE lab
INtelligent Data Engineering

UNIVERSITY OF AMSTERDAM