Platforms, Place and Your Profession

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

ESWC 2021 PhD Symposium
June 7, 2021
The world needs you!

See also the importance of ideas:

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
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?

- **Academia**: 56%
- **Industry**: 28%
- **Medical**: 11%
- **Government**: 10%
- **Non-profit**: 7%

Source: [Link](https://www.nature.com/articles/d41586-019-03459-7)
Figure 1 from “The missing piece to changing the university culture.”
**SUSTAINED SATISFACTION**
A majority of respondents are still glad they decided to pursue a PhD, although the attitudes of some have worsened over time.

**Q: How satisfied are you with your decision to pursue a PhD?**
- Very dissatisfied: 6%
- Somewhat dissatisfied: 10%
- Neutral: 10%
- Somewhat satisfied: 37%
- Very satisfied: 38%

**Q: Since the start of your graduate school experience, has your level of satisfaction increased, worsened or remained the same?**
- Increased: 42%
- Worsened: 45%
- Neutral: 13%

**Q: Overall, what do you enjoy most about life as a PhD student?**

- Intellectual challenge: 38%
- Working with interesting and bright people: 18%
- University/academic environment: 13%
- Creativity: 11%
- A chance to consider professional options: 5%
- Knowing I have a chance for a permanent academic research post: 4%
- Knowing I will have a chance to use my skills in a non-research science job: 4%
- Knowing I will have a chance for a non-academic research job: 3%
- Social life: 1%

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

<table>
<thead>
<tr>
<th>Help others</th>
<th>Do (mostly) what I like</th>
<th>Support my life priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect to a community</td>
<td>Expand my skills</td>
<td>Reach my goals</td>
</tr>
</tbody>
</table>

**Your position**
# Build your platform

<table>
<thead>
<tr>
<th>Help others</th>
<th>Do (mostly) what I like</th>
<th>Support my life priorities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connect to a community</td>
<td>Expand my skills</td>
<td>Reach my goals</td>
</tr>
</tbody>
</table>

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, Anea, Netherlands
I lead the Intelligent Data Engineering Lab (http://ideelab.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
Assistant Professor
Jun 2011 – Dec 2014 • 3 yrs 7 mos
Amsterdam Zuinoost, Provincie Noord-Holland, Netherlands
Researching new methods for dealing with diverse contextual data. Including data integration, knowledge integration, data science and data provenance.

Highlights
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

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

University of South Hampton
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
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
### 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-Lezama (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...

**Researcher in the 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 Harm...  

**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
Figure 1.2: Effect of co-authorship on citations: top six public universities in the U.S.
**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 field is analyzed separately. The supplementary materials examine alternative impact measures, including methods based on patent-renewal payments.
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 MXNet & Flink 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

Applied Scientist
Part-time
Oct 2015 – Oct 2016 · 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 2018 · 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 (DIMAG). 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, CA, Area
Learnings from a Retail Recommendation System on Billions of Interactions at bol.com

Bart van Kerkwijk, Alois Delbue Research & AMALiA, University of Amsterdam
Sebastian Schelter

ABSTRACT—Recommender systems are ubiquitous in the modern internet, where they help users find items they might like. We present an overview of a large-scale recommender system handling billions of interactions on a European e-commerce platform. We discuss the implications of our findings with respect to predictive performance, system scalability and cost. Next, we investigate the impact of current state-of-the-art recommender systems on real data from an e-commerce platform, based on an existing academic study [7]. We evaluate the predictive performance of those neural networks, as well as their deployability for production setups, in terms of training time, cost of hyperparameter search, prediction latency and scalability. Next, we study a scenario-specific improvement: we do not change the recommendation algorithm itself, but instead introduce setting parameters to dynamically adjust the recommendation latency. Finally, we evaluate the performance impact of this large-scale recommender system on real interactions on an e-commerce platform. In summary, we provide the following contributions:

• We present two studies on evaluating the predictive performance of this system. (I) We evaluate several neural network-based approaches on a large dataset of interactions, and benchmark the results against state-of-the-art recommender systems. (II) We train a large-scale online A/B test on 19 million user sessions to investigate the impact of the expected 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 an e-commerce platform (Section II).
  • We present two studies on evaluating the predictive performance of this system. (I) We evaluate several neural network-based approaches on a large dataset of interactions, and benchmark the results against state-of-the-art recommender systems. (II) We train a large-scale online A/B test on 19 million user sessions to investigate the impact of the expected 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 an e-commerce platform (Section II).

1 INTRODUCTION

There is a large body of research on scalable machine learning (ML). Nevertheless, 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 covered by current systems available in academia and open source.

1.1 Motivation

There is a large body of research on scalable machine learning (ML). Nevertheless, 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.

1.1 Motivation

Training large-scale ML models 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 covered by current systems available in academia and open source.
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