Real world data science at scale

Diverse Panelists

Moderator	Panelist 1	Panelist 2	Panelist 3	Panelist 4
Company1	Company 2	Company3	Company4	Company 5
Email	Email	Email	Email	Email

Abstract

In this panel, we look beyond the textbook and discuss how to apply data science techniques to industry problems. Our diverse panel with 30+ years of combined experience will discuss topics such as working with real-world data, scaling algorithms and infrastructure, and evaluation. The advice is intended to inspire the audience and give them a head-start on common challenges when beginning their data science careers.

Tells reviewers upfront what this submission is about

Audience

Students, industry professionals, and academics with relevant academic background in computer science, statistics or machine learning, looking to understand what it's like to work on search, recommendations, and discovery systems in industry. Beginner level.

Tells reviewers that this is an Overview 101 type of submission

Introduction

We'll cover a breadth of topics related to the differences between theoretical concepts taught in academic coursework and practicing data science in industry. These will include designing production stacks for large-scale data analysis, feature engineering, evaluation metrics, which algorithms are widely used for training predictive models with massive amounts of input, and how to work with noisy and constantly evolving data. The panelists bring perspective from a wide variety of data science problems, including traditional, marketplace and real time search, data mining on social and professional networks, and recommendations of products or services. We will end with a discussion of why we love data science, and hope to inspire women interested in steering their careers towards this exciting field.

Any panel submission must have a specific core point of contention, and a clear indication of who will be taking which sides of that issue.

Plan of Action

Reviewers can clearly visualize how this session will play out on event day.

What are some questions that you intend to ask?

- Short "Introduce Yourself": *We have a diverse set of experiences here: from traditional, marketplace and real time search to building a social or professional network. Areas we've worked on include search, recommendations, discovery and experimentation.*
- What data architecture do we use and why? *From Hadoop to AWS for storing the data and hive to cascading for data processing, we've used a wide variety of open source and home grown technologies.*
- Machine learning vs heuristics based approaches *Heuristic approaches to data mining are still very common in industry as they can often take us quite far without much effort. Machine learning approaches may require initial investment but allow us to add new features in a principled way. The panel discusses when the different approaches are appropriate.*

- How do you deal with the data evolving? *Real world data doesn't sit still. This implies not only that we need to frequently re-train our models, but also that incorporating real-time information can make a big difference.*
- Any advice on scaling up algorithms for data mining? *Build simple but equally performing algorithms, instead of complex ones. Don't overscale. Fail fast! Follow Minimum Viable Product (MVP).*
- What is the most important part of model building? *Feature engineering - garbage in, garbage out! With lots of data, the choice of the model itself is usually not as important. Good features can drastically improve the performance.*
- How do you evaluate changes? *A/B Testing! This is challenging for several reasons, including what metrics to measure and how to get enough statistical power.*
- What is the best part of your job as a data scientist? *The panel shares their favorite parts of the job - we all love what we do!*

What unique perspective would these set of panelists bring?

Naz / Twitter: Real time processing, storage and retrieval of data. Extracting personalization signals from text and urls, working with exceptionally fragmented data where the limitation of 140 characters makes it highly challenging to infer the context.
Alex / Amazon: Recommender systems, retail sales data, A/B testing.
Surabhi / Airbnb: Search, discovery, personalization and working on traditional search versus a two sided marketplace.
Ya / LinkedIn Large professional social network, large scale A/B testing platform.
Cristina / Facebook: Large amounts of social data with trillions of edges. Variety of user intentions, from people search to restaurant recommendations, to finding needle in the haystack in posts seen in the news feed.

Any competing points of view amongst the panelists?

Machine learning versus heuristics: Depending on the company age and product stage, one approach might be more suitable than the other.
Different company structures - vertical or horizontal data science teams: Some companies prefer to have a dedicated team which supports multiple products (horizontal), others have smaller teams working for specific products (vertical). (Twitter, Linkedin vs Facebook)

Description of timing of the session and the format of interaction among participants and with the audience

Flow of session is well thought out. Audience engagement is built in.

During the panel, the moderator will get the panelists perspectives on real world data science at scale. Once the floor is open, members from the audience can ask questions to all the panelists.

Timeline breakdown of how you intend to run this session.

This 60-minute panel will have 40 minutes of panelists giving their perspectives on the questions outlined above, with 2-3 panelists speaking for 2-3 minutes per question. The last 20 minutes will be devoted to Q&A with the audience.

Outcomes / Conclusion

The goal of this panel is to present what data science is like in practice and provide an understanding of how we work with data at scale. The audience should leave with a better idea

of the role of a data science professional in the industry. They should get an understanding of the most common challenges we face and some methodologies used to solve them. Female data engineers and scientists are highly underrepresented in the today's tech world; therefore, we hope to encourage women to join the field. We hope to encourage the audience to get their hands dirty with the real, challenging, and fun problems in this exciting and growing field.

Key Takeaways and purpose of this session is articulated.

Participant Statement

We, the five contributors to this proposal have committed to attending the Grace Hopper 2014 conference in Phoenix, Arizona and will serve on this panel if accepted.

Panelists (P) and Moderator (M)

Bio 1 (M)

Bio 2 (P)

Bio 3 (P)

Bio 4 (P)

Bio 5 (P)

Bios should not be generic but must pertain to the session. Let us know if the moderator has had any prior experience moderating a panel.