Ranking in the Age of Algorithms and Curated News

News is produced and consumed in significantly different fashion today compared to the previous decade. Three fundamental entities are responsible for facilitating this new consumption paradigm - media professionals, the interested online population and algorithms. Whereas journalists and editors have always **picked** which news stories their audience should be informed about, readers and consumers can now **participate** on what becomes "news" by upvoting, sharing or retweeting content wherever they see it. Algorithmic systems have the power to **analyze** billions of signals to highlight popular stories or breaking news, which in turn helps news producer learn about what's happening.

Sophisticated ranking techniques assist all three entities. For example, Digg uses internal algorithmic tools that help editors pick certain stories that display unique characteristics. Similarly, crowd sourced news services such as Reddit deploy special rankings for controversial or 'hotness' of posts. Purely algorithmic news feeds like Facebook or Google news that mine through a plethora of data (social & web respectively) also need to rely on ranking techniques to decide what items are worth surfacing. Ideally, this decision (what to surface or suppress) should reflect the **voice of all three entities - human expertise, crowd opinion and algorithmic correctness** in varying ratios. How can ranking algorithms maintain a sweet spot of balance between all three?

When services whose job is to inform the public stands on the shoulder of ranking algorithms, it is imperative to truly understand how ranking works, what goes into designing these algorithmic systems, what signals they may use or reject, how they are maintained and updated with new information and finally, how to effectively test and critique them.

During this panel we will highlight different aspects of ranking algorithms in news production and consumption, from the ground up to more abstract issues including (but not limited to) these three topics:

1. **Data, Model Parameters & Design Strategies**

Without knowledge of parameters, data and design choices, it becomes increasingly hard to understand, test or critique algorithms. Here are some interesting discussion points:

- In most cases, variation of a single parameter in ranking algorithms can encode a commitment to a specific viewpoint (e.g., PageRank at low alpha parameter is "one person, one vote", whereas at high alpha it is "more power to the powerful"). How do scientists **assign parameter values**?
- How should we think about "trending", "hot" or "controversial" articles, and **what signals** can we use to identify them?
- Who decides whether I should see more posts in my feed from my friends who post more often? How do signals from social cues help rank content? What are the implications of using these?
- Should we leverage algorithmically powered **alerts**/**notifications** - given that an alert may interrupt a user? How many notifications should we send and when? Should we personalize these?

- **Ephemerality** - should the decay factor in ranking change based on how fast new stories are incoming? This revolves around what is the ground truth for trending. What if we want a more permanent aspect of the web?

2. **Deeper factors in Ranking**

Currently, certain qualities are hard to model computationally at large scale, but could be very interesting if we can chalk up the metrics and design philosophies of these aspects:

- **Story Explainers** - how do algorithms detect/understand if some article is a good explainer? Think of the Wikipedia model, which performs one function of the article topic brilliantly - bringing the reader up to speed. Wikipedia is itself constantly updated by its users.
- **Original Content** - Should stories about the same event copied from elsewhere be suppressed in search? In other words, should original content get priority?
- **Opinions** - Today's social networks might inadvertently suppress new, different, and challenging ideas because their ranking strategies prioritize the popular and habitual. How can we change that?
- **Serendipity & Personalization** - How do we balance between the long tail/niche/divergent vs. a more general approach to surfacing stories? This is hard because the general and specific never move quite in unison. In fact for news, the opposite of personalization may be true - i.e. it is the unpredictability that keeps us interested and urges us to scroll the feed.

3. **What are we optimizing for? Informed public vs. profit, entertainment**

- News is a business after all, so the right ranking is one that produces an **optimal equilibrium** between user satisfactions vs. value extraction through advertising. When information produced is both intelligent and entertaining, it can have high attention value (like Snapchat stories).
- This is also where the question of **Trust** comes in - what might audiences feel if they realize that native advertising driven articles are pushed to them more often than others? There are also ethical questions - will Apple News highlight stories about working conditions in China? Will Twitter rank a trend lower which discusses the exit of its CEO? It is understandable that ranking algorithms might prioritize something new, funny, revelatory, or delightful so that users feel compelled to share it and it goes viral. But is it fair to use click bait, listicles and 'curiosity gap' in capturing attention but not delivering enough compelling content.
- The **reliability** of ranking algorithms is a big question mark, especially compared to the judgment of editorial expertise. Here we can talk about media hacking - ways in which you can game the ranking to surface unsuitable information. The professional journalist is trained in ethical norms and rules of practice - how can we build reliable ranking algorithms which can replicate this? And who is to blame if the ranking result is unexpected or unworthy.

In this panel we will avoid the philosophical issues of whether or not algorithms should be used by media outlets. Instead, we argue that these algorithmic systems are not really black boxes as media hype suggests, but rather humans design them. Deliberate choices are made on data signals, models, user behavior and computational capability with mathematical justifications.

The goal of this panel is to discuss what factors should (or should not) be considered in ranking news content, how do we find signals in data that can authentically represent these factors and what are the best practices of implementing/modeling such factors computationally. We anticipate having a mixture of professional working in old or new media/startups as panelists, who have experience in conceptualizing, building or designing algorithmic content scoring systems or curating news.

List of Panelists:

Catherine D'Ignazio
Assistant Professor, Department of Journalism, Emerson College; previously with MIT Media Lab
dignazio@mit.edu

Mike Dewar
New York Times R&D Lab
mikedewar@gmail.com

Anthony De Rosa
The Daily Show, previously Editor-in-Chief at Circa
anthony@gmail.com

Suman Deb Roy
Lead Data Scientist at Betaworks
suman@betaworks.com