How Data Scientists Review the Scholarly Literature

Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, Narges Mahyar
Data science is booming
Data science is booming

- Papers double every two years\(^1\)

\(^1\) Krenn, Mario, et al. "Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network." arXiv, 2022
Data science is booming

- Papers double every two years\(^1\)
- A rise in tools intended to help data scientists keep up with the literature

---

\(^1\) Krenn, Mario, et al. "Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network." arXiv, 2022
Data science is booming

- Papers double every two years\(^1\)
- A rise in tools intended to help data scientists keep up with the literature
- Competitions to predict the research data science research frontier\(^1\)

---

\(^1\) Krenn, Mario, et al. "Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network." arXiv, 2022
Data scientists are overwhelmed!
Data scientists are overwhelmed!

[D] Giving Up on Staying Up to Date and Splitting the Field

Does anyone else feel completely unable to keep up with machine learning and AI in general? I have my sub-sub-field and I do my work in (applied, mostly) and I read those papers, but I at least try to keep somewhat up to date on the entire topic of machine learning.

I mean, at this point I understand Transformers and related, and I kind of understand Latent Diffusion Models and Graph Neural Networks but not enough to use them, but I've lost the bubble on what's happening in deep reinforcement learning. I'm sure AlphaTensor is great, but I just don't have the time and energy.

I'm dreading NeurIPS and trying to figure out what people are talking about. I am wondering if ML needs to do what physics did a while ago, and just give up on trying to understand all of it.

I have a relative who does physics of solar cells (something about hot carriers and hyperfine states???) who doesn't understand what the relativity people he went to undergraduate with are talking about. They go to different conferences now.
A consequent slowdown in scientific progress (?)
A consequent slowdown in scientific progress (?)

Peer reviewers struggle to recognize and understand novel ideas
Exploring data scientists reviewing the literature
Exploring data scientists reviewing the literature

✦ Data scientists

✦ Literature reviews

1 Crisan et al. "Passing the data baton: A retrospective analysis on data science work and workers." IEEE Transactions on Visualization and Computer Graphics, 2020

Exploring data scientists reviewing the literature

✧ Data scientists¹

✧ Individuals trained in computer science, statistics, and application specific disciplines e.g. economics or biology

✧ Engaged in data work, applied engineering, and research

✧ Literature reviews²

¹Crisan et. al. "Passing the data baton: A retrospective analysis on data science work and workers." IEEE Transactions on Visualization and Computer Graphics, 2020

Exploring data scientists reviewing the literature

❖ Data scientists\(^1\)
  ❖ Individuals trained in computer science, statistics, and application specific disciplines e.g. economics or biology
  ❖ Engaged in data work, applied engineering, and research

❖ Literature reviews\(^2\)
  ❖ A learning process spanning information seeking, sensemaking, and composition
  ❖ Obtaining research literature, forming a synthesized understanding of the gathered data, and presentation of this information

\(^1\)Crisan et. al. "Passing the data baton: A retrospective analysis on data science work and workers." IEEE Transactions on Visualization and Computer Graphics, 2020

Research Goals
Research Goals

✦ Data scientists practices and challenges in:
Research Goals

✦ Data scientists practices and challenges in:

Search
Research Goals

✨ Data scientists practices and challenges in:

- Search
- Selecting sources
Research Goals

✦ Data scientists practices and challenges in:

- Search
- Selecting sources
- Interacting with sources
- Synthesis & presentation
Research Goals

- Data scientists practices and challenges in:

- Search
- Selecting sources
- Interacting with sources
- Synthesis & presentation
Research Goals

✦ Data scientists practices and challenges in:

✦ Examining the activities around search; skimming, reading, and synthesis
Remainder of the talk

- Study goals and motivations
- Our participants
- Methods for the study & analysis
- Results
- Implications
Study Participants

- 20 participants; self-id as data scientists
- Recruited from university lists + social media
Study Participants

- 20 participants; self-id as data scientists
- Recruited from university lists + social media
- Workplace: 13 university, 7 industry & non-profit
Study Participants

- 20 participants; self-id as data scientists
- Recruited from university lists + social media
- Workplace: 13 university, 7 industry & non-profit
- Average published papers: 4
Study Participants

- 20 participants; self-id as data scientists
- Recruited from university lists + social media
- Workplace: 13 university, 7 industry & non-profit
- Average published papers: 4
- Noted pronouns: 11 he/him, 9 she/hers, 2 they/them
Study & Analysis Methods
Study & Analysis Methods

- Semi-structured interviews
- Think-aloud observation
- Encouraged to discuss all interactions with literature
- Participant explores literature

1 hour x 20
Study & Analysis Methods

✦ Semi-structured interviews
✦ Encouraged to discuss all interactions with literature
✦ Think-aloud observation
✦ Participant explores literature

1 hour x 20
✦ After a pilot; 15 participants
Study & Analysis Methods

- Semi-structured interviews
- Encouraged to discuss all interactions with literature
- Think-aloud observation
- Participant explores literature

Axial coding; 3 authors
Agreement $\alpha$: 0.92
Thematic analysis

1 hour x 20
After a pilot; 15 participants
3 months
Study & Analysis Methods

- Semi-structured interviews
- Encouraged to discuss all interactions with literature
- Think-aloud observation
- Participant explores literature
- After a pilot; 15 participants
Study & Analysis Methods

- Semi-structured interviews
- Encouraged to discuss all interactions with literature
- After a pilot; 15 participants
- Think-aloud observation
- Participant explores literature

“Recall a literature review you conducted in the past. Imagine you were re-starting this process and show us how you went through the literature review.”

OR

“Imagine you are interested in finding and documenting the latest work on a topic of your interest, show us how you go about this process.”

OR

“Imagine you are planning future work on a problem you are interested in, conduct the literature review to help plan your future work”
Remainder of the talk

✧ Study goals and motivations
✧ Our participants
✧ Methods for the study & analysis
✧ Results
✧ Implications
Results

✦ Why do data scientists access the literature?
✦ How do they access the literature?
✦ How do they select papers?
✦ What challenges do they face in reading papers?
✦ How do they lean on social ties?
Results

✦ Why do data scientists access the literature?
✦ How do they access the literature?
✦ How do they select papers?
✦ What challenges do they face in reading papers?
✦ How do they lean on social ties?
Why do data scientists access the literature?

How do they access the literature?

How do they select papers?

What challenges do they face in reading papers?

How do they lean on social ties?
Results

✦ Why do data scientists access the literature?
✦ How do they access the literature?
✦ How do they select papers?
✦ What challenges do they face in reading papers?
✦ How do they lean on social ties?
Why do data scientists access the scientific literature?
Why do data scientists access the scientific literature?

✦ Desire to understand disciplinary norms

“When I’m starting work in a problem ... I’m not sufficiently familiar with to know what the typical approaches are, how is this evaluated, what kinds of approaches are falling out of favor versus becoming more accepted by the community.” - P15
Why do data scientists access the scientific literature?

✦ Desire to understand disciplinary norms

✦ Passively following a discipline

“Where the community is going, or what people that I have previously followed the works of are up to right now” - P10

“When I’m starting work in a problem ... I’m not sufficiently familiar with to know what the typical approaches are, how is this evaluated, what kinds of approaches are falling out of favor versus becoming more accepted by the community.” - P15
Why do data scientists access the scientific literature?

✦ Desire to understand disciplinary norms

✦ Passively following a discipline

✦ Brainstorming solutions

“When I’m starting work in a problem ... I’m not sufficiently familiar with to know what the typical approaches are, how is this evaluated, what kinds of approaches are falling out of favor versus becoming more accepted by the community.” - P15

“Where the community is going, or what people that I have previously followed the works of are up to right now” - P10

“After you figure out [the problem], it’s like I have an idea for what you could do better, and then it’s seeing if others have done something similar before” - P5
Why do data scientists access the scientific literature?

- Desire to understand disciplinary norms
- Passively following a discipline
- Brainstorming solutions

“Where the community is going, or what people that I have previously followed the works of are up to right now” - P10

“When I’m starting work in a problem ... I’m not sufficiently familiar with to know what the typical approaches are, how is this evaluated, what kinds of approaches are falling out of favor versus becoming more accepted by the community.” - P15

“After you figure out [the problem], it’s like I have an idea for what you could do better, and then it’s seeing if others have done something similar before” - P5
How do data scientists access the scientific literature?
How do data scientists access the scientific literature?

- Data scientists seeking the literature with search

“I had an idea in my head, but I was not sure how to map this into normative terms used by communities … Ultimately it just took trial and error, finding some papers and coming back to it over several weeks, and eventually I kind of started to find things that actually matched.” - P15
How do data scientists access the scientific literature?

✦ Data scientists seeking the literature with search
✦ The literature finding data scientists with automated + personal recommendations

“I had an idea in my head, but I was not sure how to map this into normative terms used by communities … Ultimately it just took trial and error, finding some papers and coming back to it over several weeks, and eventually I kind of started to find things that actually matched.” - P15
How do data scientists access the scientific literature?

- Data scientists seeking the literature with search
- The literature finding data scientists with automated + personal recommendations
- Trapped in a disciplinary bubble

“I had an idea in my head, but I was not sure how to map this into normative terms used by communities … Ultimately it just took trial and error, finding some papers and coming back to it over several weeks, and eventually I kind of started to find things that actually matched.” - P15

“I’m probably heavily in my own bubble of papers … if I’m working on hate speech, most of my recommendations will be very computer science based but maybe there’s relevant stuff in social science that I’m probably never going to come across.” - P11.
How do data scientists select papers?
How do data scientists select papers?

✦ Understanding salient differences between similar items
How do data scientists select papers?

- Understanding salient differences between similar items

“If the field is very crowded - sometimes I find RL, and the problems I am focusing on to be crowded, then it becomes frustrating and you’re always finding papers that do the same thing.” - P19

“In grad school a professor synthesizes these things and says hey, this is the main theme of all these papers. When that information is there for you, it tells you what to expect otherwise you’re spending a lot of time and don’t understand how different it is from previous papers” - P7.
How do data scientists select papers?

- Understanding salient differences between similar items
- Users understand similar item variants in terms of their differences or aligned to time e.g. code snippets

“If the field is very crowded - sometimes I find RL, and the problems I am focusing on to be crowded, then it becomes frustrating and you’re always finding papers that do the same thing.” - P19

“In grad school a professor synthesizes these things and says hey, this is the main theme of all these papers. When that information is there for you, it tells you what to expect otherwise you’re spending a lot of time and don’t understand how different it is from previous papers” - P7.

How do data scientists select papers?

- Understanding the salient differences between similar items
- Establishing the credibility of papers with the knowledge context
How do data scientists select papers?

- Understanding the salient differences between similar items
- Establishing the credibility of papers with the knowledge context

“One thing is that its hard to figure the credibility of a paper, so it’s sort of trying to figure it out based on discussions by online forums like Twitter, Reddit or Openreview. Even if this is highly reviewed what do other people who have worked in similar domains think about it” - P14
How do data scientists select papers?

- Understanding the salient differences between similar items
- Establishing the credibility of papers with the knowledge context
- A mismatch from information scent

“One thing is that its hard to figure the credibility of a paper, so it’s sort of trying to figure it out based on discussions by online forums like Twitter, Reddit or Openreview. Even if this is highly reviewed what do other people who have worked in similar domains think about it” - P14

People do a lot of re-branding, sometimes a lot of ideas are not very new but the motivation section is like poetry and when you read the details you feel [its] not what they are claiming they do. ... [or] exaggerating their contribution and not meeting the expectation in their experiments. So identifying those trends from papers is very important.” - P19
What challenges do data scientists face in reading papers?
What challenges do data scientists face in reading papers?

✧ Understanding the hidden details of papers through code

[I ask authors if] there is any publicly available code for what you’re doing. Because many of these papers look well on paper but then it’s unclear how to implement them. Or its unclear which specific hyperparameter choices they made. - P16
What challenges do data scientists face in reading papers?

- Understanding the hidden details of papers through code
- Understanding the math on display in papers with blogs, code, and talks

[I ask authors if] there is any publicly available code for what you’re doing. Because many of these papers look well on paper but then it's unclear how to implement them. Or its unclear which specific hyperparameter choices they made. - P16
What challenges do data scientists face in reading papers?

- Understanding the hidden details of papers through code
- Understanding the math on display in papers with blogs, code, and talks
- Problem with disciplinary norms

[I ask authors if] there is any publicly available code for what you’re doing. Because many of these papers look well on paper but then its unclear how to implement them. Or its unclear which specific hyperparameter choices they made. - P16

In writing for niche audiences it requires having to show that [an idea] is important or useful and often that means that they will add equations or theorems [for an idea] that really is not as complicated ... if there’s a lot of math or if it’s hard to understand it must be impressive. - P5
How do data scientists lean on social ties?

✦ Seeking recommendations, collaboratively brainstorming, and making sense of papers with
  ✦ Peers in person
  ✦ Peers in online forums
✦ Leveraging engagement with authors
  ✦ In direct communication
  ✦ Passively through talks and forum posts
Results

- Why do data scientists access the literature?
- How do they access the literature?
- How do they select papers?
- What challenges do they face in reading papers?
- How do they lean on social ties?
Implications

✦ Support cross-disciplinary access
✦ Facilitate reliance on close peers
✦ Leverage the knowledge context of papers
Implications

✦ Support cross-disciplinary access
✦ Facilitate reliance on close peers
✦ Leverage the knowledge context of papers

- Example prior work → possible future work

- Likely to span work in several related fields (IR, HCI, NLP, CSCW …)
Implications

Support cross-disciplinary access
Implications

Support cross-disciplinary access

✦ Data science is a rapidly evolving and interdisciplinary field

✦ Data scientists operate in their own knowledge silos
Implications

Support cross-disciplinary access

✦ Data science is a rapidly evolving and interdisciplinary field

✦ Data scientists operate in their own knowledge silos

✦ Some ways forward:

✦ Verbose, interactive, conversational searches for cross domain exploration

✦ Skimming aids such as adaptive document layouts, document level FAQs, and QA - perhaps with personalization to readers

✦ Reading aids like paraphrasing documents toward different disciplinary audiences
Implications

Support cross-disciplinary access

✧ Data science is a rapidly evolving and interdisciplinary field

✧ Data scientists operate in their own knowledge silos

✧ Some ways forward:

✧Verbose, interactive, conversational searches for cross domain exploration

✧Skimming aids such as adaptive document layouts, document level FAQs, and QA - perhaps with personalization to readers

✧Reading aids like paraphrasing documents toward different disciplinary audiences

✧But, people learn when a task is perceived as challenging\(^1,2\)

\(^1\) Vakkari, Perti, and Saila Huuskonen. "Search effort degrades search output but improves task outcome." JASIST, 2012

Implications

Facilitate reliance on close peers
Implications

Facilitate reliance on close peers

- With peers, data scientists did:
  - Received recommendations, brainstormed, established credibility, read papers
Implications

Facilitate reliance on close peers

✦ With peers, data scientists did:
  ✦ Received recommendations, brainstormed, established credibility, read papers
  ✦ Some ways forward:
    ✦ Sensemaking, reading, discovery in collaborative feed-readers
    ✦ Collaborative conversational agents aiding brainstorming

Implications

Leverage the knowledge context of papers
Implications

Leverage the knowledge context of papers

✦ With forum discussions, recorded talks, videos, blogs, code repositories data scientists did:

✦ Discovery, establish credibility, aided reading.
Implications

Leverage the knowledge context of papers

✦ With forum discussions, recorded talks, videos, blogs, code repositories data scientists did:
  ✦ Discovery, establish credibility, aided reading.

✦ Some ways forward:
  ✦ Knowledge context in SERPs
  ✦ The knowledge context as a reading and skimming aid

---

1 Smith, Catherine L., and Soo Young Rieh. "Knowledge-context in search systems: Toward information-literate actions.”, CHIIR, 2019
2 Rachatasumrit, Napol, et al. "CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading." IUI, 2022
Implications

Leverage the knowledge context of papers

✦ With forum discussions, recorded talks, videos, blogs, code repositories data scientists did:

✦ Discovery, establish credibility, aided reading.

✦ Some ways forward:

✦ Knowledge context in SERPs¹

✦ The knowledge context as a reading and skimming aid²

✦ Provider fairness and the knowledge context³

---

¹ Smith, Catherine L., and Soo Young Rieh. "Knowledge-context in search systems: Toward information-literate actions.", CHIIR, 2019

² Rachatasumrit, Napol, et al. "CiteRead: Integrating Localized Citation Contexts into Scientific Paper Reading." IUI, 2022

³ McDonald, Graham, Craig Macdonald, and Iadh Ounis. "Search results diversification for effective fair ranking in academic search." IRJ, 2022
Summary

- 20 participants
- University, industry/non-profit
Summary

- 20 participants
- University, industry/non-profit
Summary

- 20 participants
- University, industry/non-profit

Why do data scientists access the literature?
- How do they access the literature?
- How do they select papers?
- What challenges do they face in reading papers?
- How do they lean on social ties?
# How Data Scientists Review the Scholarly Literature

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Institution</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheshera Mysore</td>
<td><a href="mailto:smysore@cs.umass.edu">smysore@cs.umass.edu</a></td>
<td>University of Massachusetts, Amherst</td>
<td>USA</td>
</tr>
<tr>
<td>Mahmood Jasim</td>
<td><a href="mailto:mjasim@cs.umass.edu">mjasim@cs.umass.edu</a></td>
<td>University of Massachusetts, Amherst</td>
<td>USA</td>
</tr>
<tr>
<td>Sarah Akbar</td>
<td><a href="mailto:sakbar@umass.edu">sakbar@umass.edu</a></td>
<td>University of Massachusetts, Amherst</td>
<td>USA</td>
</tr>
<tr>
<td>Andre Kenneth Chase</td>
<td><a href="mailto:andreken@umass.edu">andreken@umass.edu</a></td>
<td>University of Massachusetts, Amherst</td>
<td>USA</td>
</tr>
<tr>
<td>Randall Narges Mahyar</td>
<td><a href="mailto:nmahyar@cs.umass.edu">nmahyar@cs.umass.edu</a></td>
<td>University of Massachusetts, Amherst</td>
<td>USA</td>
</tr>
</tbody>
</table>

**MSheshera/dslitreview-study**