George Fletcher¹, Paul Groth², Juan Sequeda³

¹Eindhoven University of Technology
²University of Amsterdam
³data.world
Finding digital truth—that is, identifying and combining data that accurately represent reality—is becoming more difficult and more important.

More difficult because data and their sources are multiplying.

And more important because firms need to get their data house in order to benefit from AI, which they must to stay competitive.

-- The Economist, February 2020
Data interoperability and quality, as well as their structure, authenticity and integrity are key for the exploitation of the data value, especially in the context of AI deployment.

-- European Commission, “A European strategy for data”, February 2020
Less than half of companies noted success in these key metrics of progress: Driving innovation with data, Competing on data and analytics, Managing data as a business asset, Forging a data culture, Creating a data-driven organization.”

Big Data and AI Executive Survey 2021.
The next big wave is about KNOWLEDGE

Thesis. Big Data and Data Science have been the rave for the past years, with tremendous successes. However, organizations need to next focus on data and knowledge culture.

This requires new roles (and their technologies) who understand, manage and are responsible for the meaning of the data in a data team!

Towards a paradigm shift to realize the full potential of data- and knowledge-driven science, industry, and society.
# Recent data history

<table>
<thead>
<tr>
<th>Context</th>
<th>Trend</th>
<th>Organizational Need</th>
<th>Technology</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web + Moore’s Law</td>
<td>Big Data</td>
<td>Harness and collect data</td>
<td>Commodity distributed computing platforms (e.g., Hadoop)</td>
<td>Data Engineer</td>
</tr>
<tr>
<td>Big Data + GPU Compute</td>
<td>AI Revolution</td>
<td>Draw value from data</td>
<td>Commodity machine learning (e.g., TensorFlow, SciPy)</td>
<td>Data Scientist</td>
</tr>
<tr>
<td>AI Revolution + Cloud Computing</td>
<td>Data-Driven Organization, Digital Transformation</td>
<td><strong>Trustworthy data</strong></td>
<td>Clean, meaningful, beautiful data technologies (e.g., knowledge graphs, data wrangling systems, data catalog platforms)</td>
<td></td>
</tr>
</tbody>
</table>
Towards trustworthy data

Data wrangling (up to 80% of a data scientist’s work) is not just grunt work.

The work of reliable data is about

- **understanding the ecosystem between people, data and tasks in an organization** and
- **communicating, documenting, and maintaining that knowledge.**

This is vital work **at the heart of any data-driven organization**
## Recent data history

<table>
<thead>
<tr>
<th>Context</th>
<th>Trend</th>
<th>Organizational Need</th>
<th>Technology</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web + Moore’s Law</td>
<td>Big Data</td>
<td>Harness and collect data</td>
<td>Commodity distributed computing platforms (e.g., Hadoop)</td>
<td>Data Engineer</td>
</tr>
<tr>
<td>Big Data + GPU Compute</td>
<td>Al Revolution</td>
<td>Draw value from data</td>
<td>Commodity machine learning (e.g., TensorFlow, SciPy)</td>
<td>Data Scientist</td>
</tr>
<tr>
<td>AI Revolution + Cloud Computing</td>
<td>Data-Driven Organization, Digital Transformation</td>
<td>Rely on data</td>
<td>Clean, meaningful, beautiful data technologies (e.g., knowledge graphs, data wrangling systems, data catalog platforms)</td>
<td>Knowledge Scientist</td>
</tr>
</tbody>
</table>

The **Knowledge Scientist** is responsible for reliable trustworthy data.
Who is the Knowledge Scientist?

**Knowledge Scientist:** the bridge builder between data, people, and their organizational requirements, questions, and needs.

**Role:** understand and document knowledge from *Users, Data Scientists, Data Engineers,* and their environment.

**Goal:** reliable data for the data-driven organization.
Who is the Knowledge Scientist?

Knowledge Science is **people work:**
data in context, grounded in shared meaning

- Working together with data consumers and producers for explainable meaningful data

Knowledge Science is **technical work:**
maintained, accessible, clean data

- Safeguarding actionable, trustworthy, and timely data
Some Implications

**For students:** Where are there opportunities to be a leader and make impact?

**For researchers:** New critical research challenges in knowledge at extreme scale in data-driven organizations, in human context

**For universities:** Are we preparing our students for the next data wave?

**For organizations:** Are you treating data as a product? Will you be able to grow as an organization without the best data?
Today’s program

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:00 - 3:10 PM</td>
<td>Opening</td>
<td></td>
</tr>
<tr>
<td>3:10 - 3:40 PM</td>
<td>University session</td>
<td>Elena Simperl</td>
</tr>
<tr>
<td></td>
<td>Chaired by Paul Groth</td>
<td></td>
</tr>
<tr>
<td>3:40 - 4:40 PM</td>
<td>Knowledge Science in action</td>
<td>Pieter Colpaert</td>
</tr>
<tr>
<td></td>
<td>Chaired by George Fletcher</td>
<td>Knowledge Science for mobility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Catia Pesquita</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Knowledge Science for AI-based biomedical and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>clinical applications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ilaria Tiddi</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Knowledge-aware robots</td>
</tr>
<tr>
<td>4:40 - 5:20 PM</td>
<td>Industry session</td>
<td>Mohammed Aaser</td>
</tr>
<tr>
<td></td>
<td>Chaired by Juan Sequeda</td>
<td>Tim Gasper</td>
</tr>
<tr>
<td>5:20 - 5:30 PM</td>
<td>Closing remarks</td>
<td></td>
</tr>
</tbody>
</table>

Today’s goals

- International experts share their views on knowledge in academia, research, and industry
- Start conversations around implications and future of knowledge science
Takeaways

● Bring software engineering practices to data and knowledge
● Language/ML Models are new forms of knowledge representation. How are these maintained and evaluated?
● Regulatory compliance of data and knowledge
● What is reliable? What is trustworthy?
● Courses on Semantic Web technologies and complemented with how to stop using 80% of your time on finding and cleaning data
● Bias and explainability, in ML models... work for knowledge science?
● We need context. Need to understand users
● Knowledge Science part of CS curriculum, but also in any general Science curriculum to train the next generation of scientist
● Complex/Long ML/data pipelines emphasize the need for knowledge science
● Knowledge Discovery Process 1996, what has really changed? More of everything, but ...do we have more interpretation?
● Reality gap! Knowledge fills the gap
Industry Takeaways

- the knowledge world and modern data world are on a collision course
  - analytics engineering --> semantic layer <-- knowledge engineering/science
- importance of data has increased to the executive level, data is a competitive advantage, build a view of customers, suppliers, employees.
- no more silos, bring together to unlock insights
- Data as a Service makes it even more necessary to have roles to curate/manage data supply?
- they haven't invested in common models. If you want to scale, you need to invest in those efforts.
- what are the most common requests that we get from our biz stakeholders. how do we standardize that.
- Role or a Skill that other roles should have?
- Knowledge Scientist or a Data Product Manager or Analytics Engineer?
- Skills: Data Modeling, Business (Accounting, Marketing), SQL, Knowledge Graph
- Knowledge science is truly interdisciplinary: CS, Data Sci, Industrial Design, Economics, … ..
- How do you scale data usage in your organization? Get ontology/models right, you can save 50%-80% of effort
- (Knowledge) Graph technology is still not well known. Need to make it easier
The great barrier to AI adoption in healthcare and biomedical research is lack of trust.

Knowledge Science is the answer.

https://www.knowledgescientist.org/

Continue the conversation on Slack
https://knowledgescientist.slack.com/