Agenda

- Poll: What’s your main focus area?
- Data Science vs Information Science (Dave Clarke)
- How does Knowledge Management relate? (Patrick Lambe)
- Commonalities and differences (Dave and Patrick)
- Poll on alignment/coordination between DS, IS, KM
- When we fail to “mind the gap”: issues around exploiting common ground and respecting differences (discussion)
Poll 1: What’s your background?

Mainly information management/ information science

Mainly data management data science

Mainly knowledge management

My role involves trying to integrate DS, IS and KM

Other
Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyse actual phenomena" with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. However, data science is different from computer science and information science. Turing Award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.
Information science

From Wikipedia, the free encyclopedia

Information science (also known as information studies) is an academic field which is primarily concerned with analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information. Practitioners within and outside the field study the application and the usage of knowledge in organizations in addition to the interaction between people, organizations, and any existing information systems with the aim of creating, replacing, improving, or understanding information systems.

Historically, information science is associated with computer science, data science, psychology, technology, library science, healthcare, and intelligence agencies. However, information science also incorporates aspects of diverse fields such as archival science, cognitive science, commerce, law, linguistics, museology, management, mathematics, philosophy, public policy, and social sciences.
Defined knowledge models (taxonomies) support extraction and tagging.
Entity and relationship extraction enriches taxonomies.
Data comes in many forms, all of it messy. Whether we’re talking about missing data, unstructured data, or data that lacks regular structure, you need methods to cleanse data before you can process it to improve its quality. This tutorial explores some of the key problems of working with real-world data and the methods you can apply.

This is the first installment in a three-part series that explores the problem of messy data and the methods available to increase the quality of your data sets (see Figure 1). Parts 2 and 3 will continue with data analysis using machine learning and data visualization.

Figure 1. Data processing pipeline, from cleansing to learning and visualization

Data science and its algorithms are clean and precise, but the data on which they operate come from the real world, which means that they’re messy and require some preparation before you can use them effectively. The quality of insights you derive from data depends on the validity of that data, so some preparation is required.

Data cleaning has a long history in databases and is a key step in what’s known as extract, transform, load (ETL). ETL is commonly used in data warehouses, where data is extracted from one or more sources, transformed into its proper format and structure, including cleansing of the data, and finally loaded into a final target location, such as a

https://lod-cloud.net

Information Extraction

NLP coupled with big knowledge graphs can help with the identification and disambiguation of entities.

This NOW news demo uses NLP and Wikidata.

NLP most suited to extraction of novel named entities (people, places, organizations, etc.).

Extraction of abstract topical concepts is harder to achieve.

Finding specific things mentioned in text is easier than determining overall document aboutness.

Best ever image of the universe revealed by Nasa’s James Webb telescope

Nasa and the James Webb Space Telescope just peered billions of years further into the past than ever before, and brought photos home for all to see.

Around 6:20pm EDT, US President Joe Biden revealed the very first public image taken by the Webb telescope, a sea of distant galaxies and brilliant spiky stars, the most distant view yet of the Cosmos. The image was presented on a screen at the White House along with comments by Nasa Administrator Bill Nelson.

“Mr President, if you held a grain of sand on the tip of your finger at arms length, that is the part of the universe you are seeing.” Mr Nelson said, a tiny portion of the sky magnified by Webb to reveal thousands of galaxies. “That light you are seeing on one of those little specks, has been traveling for over 13 billion years.”

Nasa’s Twitter account published a higher resolution version of the image than was projected at the White House.

Stars in the foreground appear as brilliant blue-white spikes, while the most distant galaxies appear as orange-red specks.

Information Extraction

NLP most suited to extraction of novel named entities (people, places, organizations, etc.).

Extraction of abstract topical concepts is harder to achieve.
Bridging the Human v. Machine-Readable Gap

Explicit disambiguator in parentheses

Hierarchy provides category context

Both are human and machine readable for semantic contextualization

Natural language descriptions / definitions are readable by humans and by NLP ML semantic similarity algorithms

To disambiguate ambiguous concepts simple NOT words and phrases can be added to help the machine eliminate false matches
Synaptica example of an Info-Science / Data Science feedback loop

- Collaboratively curate enterprise taxonomies and ontologies
- Using SKOS, OWL or your own semantic schema
- Store ontologies in an RDF graph database
- Data reconciliation analytics and visualization
- Entity extraction of candidate concepts and names
- Perform inline semantic tagging and document level classification
- Manage transparent tagging rules
- Test tagging and classification quality against evaluation corpora
- Entity extraction of candidate concepts and names
Knowledge management

From Wikipedia, the free encyclopedia

Not to be confused with Content management or Information management.

Knowledge management (KM) is the collection of methods relating to creating, sharing, using and managing the knowledge and information of an organization.\(^1\) It refers to a multidisciplinary approach to achieve organisational objectives by making the best use of knowledge.\(^2\)

An established discipline since 1991,\(^3\) KM includes courses taught in the fields of business administration, information systems, management, library, and information science.\(^3\)\(^4\) Other fields may contribute to KM research, including information and media, computer science, public health and public policy.\(^5\) Several universities offer dedicated master’s degrees in knowledge management.

Many large companies, public institutions and non-profit organisations have resources dedicated to internal KM efforts, often as a part of their business strategy, IT, or human resource management departments.\(^6\) Several consulting companies provide advice regarding KM to these organisations.\(^6\)

Knowledge management efforts typically focus on organisational objectives such as improved performance, competitive advantage, innovation, the sharing of lessons learned, integration and continuous improvement of the organisation.\(^7\) These efforts overlap with organisational learning and may be distinguished from that by a greater focus on the management of knowledge as a strategic asset and on encouraging the sharing of knowledge.\(^2\)\(^8\) KM is an enabler of organisational learning.\(^9\)\(^10\)

https://en.wikipedia.org/wiki/Knowledge_management

Processes for creating, using, sharing

Aligned with organizational objectives

Interdisciplinarity – should orchestrate IS, IM and DS

Enables org learning
The Big Picture

KNOWLEDGE MANAGEMENT
Organisation/user context & purpose

INFORMATION SCIENCE
Information structure, standards, rules

DATA SCIENCE
Machine learning, data processing & analytics
KNOWLEDGE MANAGEMENT
Organisation/user context & purpose

Learning, tacit knowledge
People interactions
Application of data, information & knowledge in activities and decisions
Supporting work processes and outcomes

Strategic alignment/Purpose
User needs analysis
Governance
Systems integration
Communications & change management
Competencies & capacity building

Knowledge discovery
Access to knowledge resources

Taxonomies
Ontologies
Knowledge Graphs
Controlled Vocabularies

Search technologies
Text analytics
Metadata

Decision support
Generating insight
Sensemaking
Experimentation

Big data
Data management
Data processing, quality, integration
Data analytics, dashboards

INFORMATION SCIENCE
Information structure, standards, rules

DATA SCIENCE
Machine learning, data processing & analytics

Dave Clarke, Synaptica
Patrick Lambe, Straits Knowledge

International Society for Knowledge Organization
Singapore Chapter

ISKOSpace Singapore
July 15, 2022
Poll 2: Where is your organization at in relation to managing DS/IS/KM orchestration?

0 2 4 6 8 10 12

Not aware of the need, completely siloed
Aware of the need but still siloed
Starting to coordinate
Effectively coordinated
DS/IS/KM represent competing centres of influence/resources
Discussion: “Mind the Gap” – Orchestrating but Respecting Differences

• Some horror stories
• Where have you seen orchestration work well?
• Let’s talk about weaknesses:
  • KM – Tends to open up too many fronts for change – progress is slow
  • IS – Tends to be overly conceptual – high resource to implement and govern well
  • DS – Tends to rush to technology, not always easy to validate accuracy of insights
Additional Resources

Critique of the DIKW pyramid for being too simplistic


Explainability of AI/ML/ data science outputs


Contact

Dave Clarke
• www.synaptica.com
• dave.clarke@synaptica.com

Patrick Lambe
• www.straitsknowledge.com
• plambe@straitsknowledge.com