ATYPON

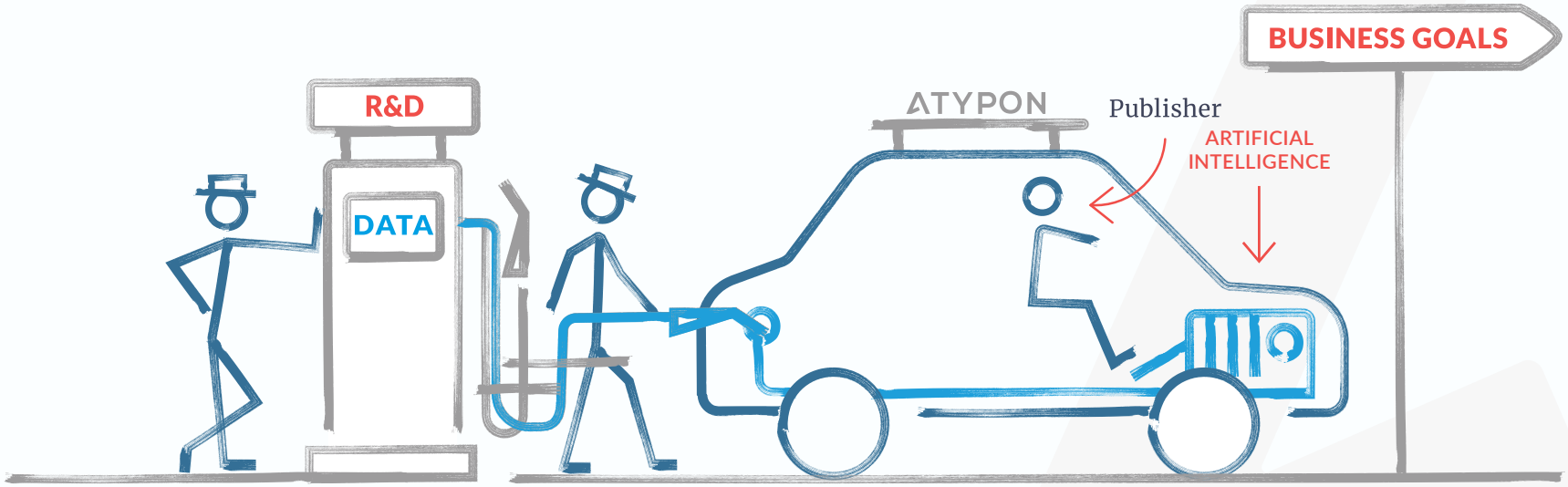
WebinarSeries

Accelerating Content Discovery

Atypon's R&D in artificial intelligence and machine learning

Hong Zhou
Senior Product Manager of AI Products & R&D
July 10, 2019

NOTE: SOME SLIDES CONTAIN ANIMATIONS THAT PREVENT THEM FROM RENDERING PROPERLY AS A PDF.


BUSINESS GOALS
R&D
DATA
ATYPON
Publisher
ARTIFICIAL
INTELLIGENCE

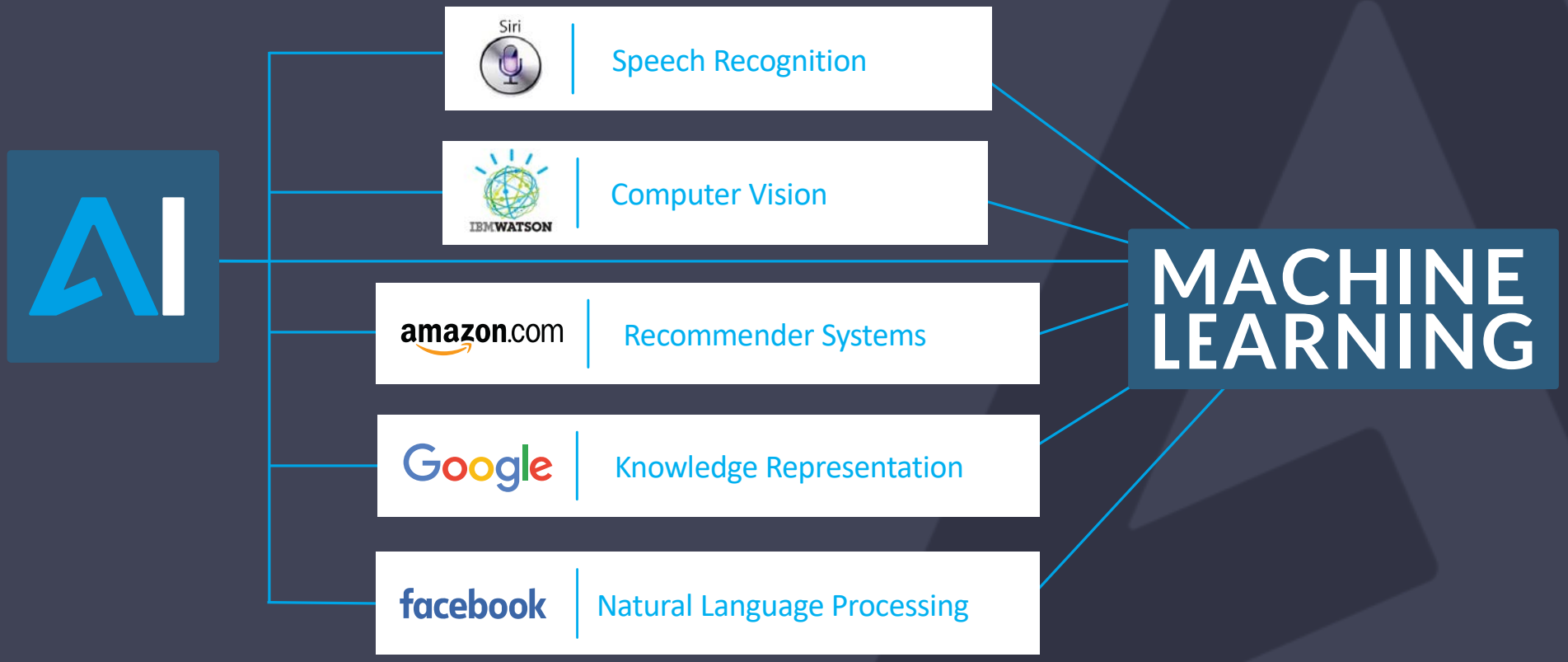
Outline

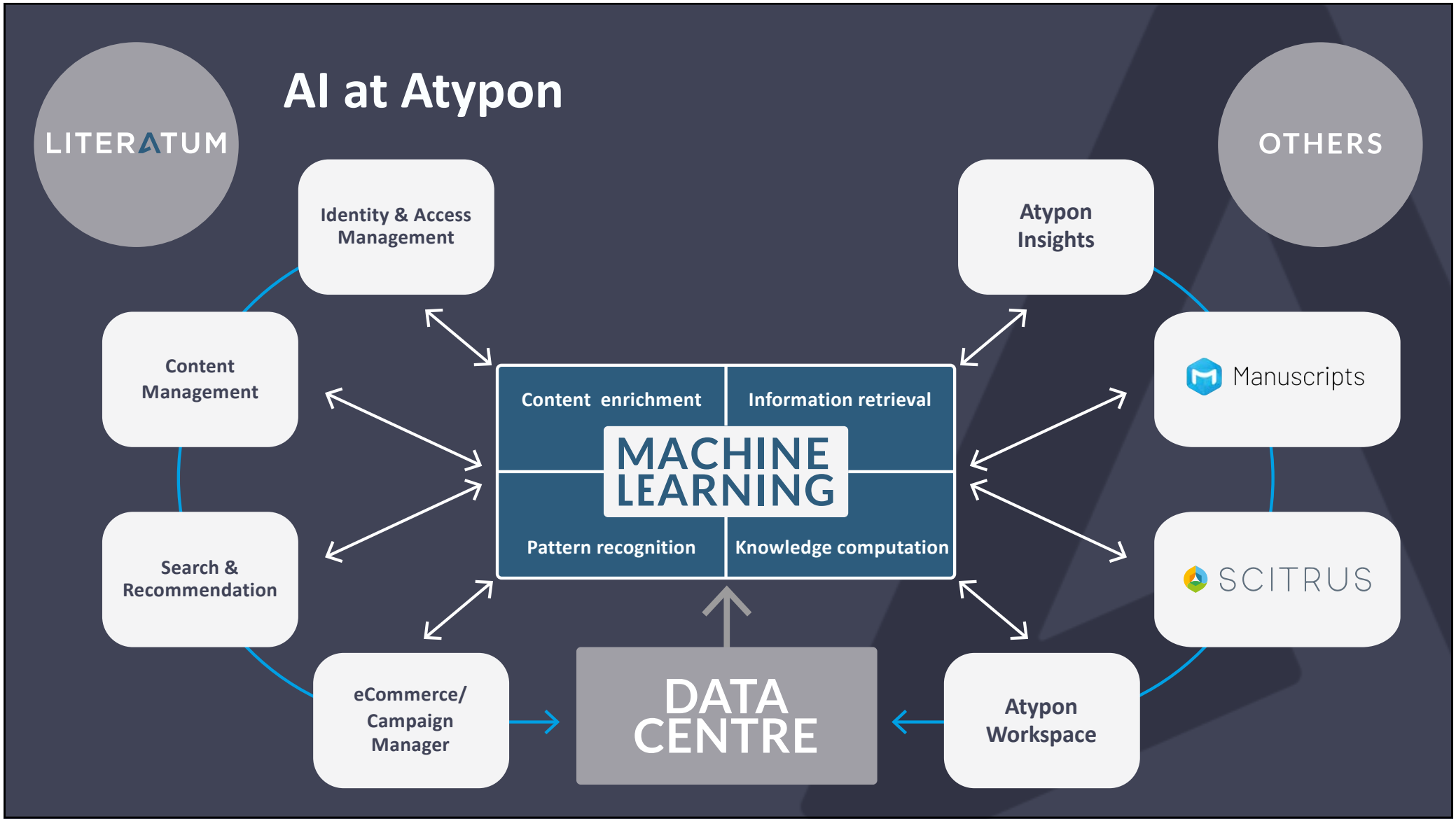
- An overview of Atypon AI strategies & capabilities
- Examples of Atypon AI applications
  1. Content enrichment
  2. Information extraction
  3. Information retrieval
  4. User profiling
  5. Trend/impact analysis
  6. Knowledge representation
- AI product roadmap

“AI in All” strategy

- Driven by the needs of—and input from—our clients
- [illegible] learners to **know** more, **[illegible] and achieve** more
- [illegible] product functionality
- Collaboration with the world’s best research institutions

What is AI?


AI
Siri
Speech Recognition
IBM WATSON
Computer Vision
amazon.com
Recommender Systems
Google
Knowledge Representation
facebook
Natural Language Processing
MACHINE LEARNING


AI at Atypon
LITERATUM
OTHERS
Identity & Access Management
Content Management
Search & Recommendation
eCommerce/ Campaign Manager
Atypon Insights
Manuscripts
SCITRUS
Atypon Workspace
Content enrichment
Information retrieval
MACHINE LEARNING
Pattern recognition
Knowledge computation
DATA CENTRE

AI applications

Content enrichment

- Semantic annotation & tagging on text and images
- Content summarization
- Topic/concept identification
- Entity disambiguation

AI applications

Information extraction

- Key metadata extraction from PDFs: keywords, grants, figures, tables, captions, citations, titles, affiliations, abstracts, etc.
- Compound-figure segmentation
- Knowledge extraction

AI applications

Information retrieval

- Intelligent recommendations for papers, venues, reviewers, and users
- Cross-publisher recommendations
- Personalized search
- Global search
- Smart query suggestion
- Near real-time search
- Power search for complex queries
- Automatic Q&A (under development)

AI applications

User profiling

- Predictive analytics on usage data
- Users clustering
- User intention/interests detection

AI applications

Trend/impact analysis

- Discover trending topics
- Impact prediction
- Influence scores of entities

1. Content enrichment

- Tagging & semantic annotation
- Content summarization

Literatum's content auto-tagger

Stage 1: Training classifier

Stage 2: Auto-tagging

Tagged doc 1

Tag1: Computer

Tag2: Education

Tag3: IT technology

Tagged doc

g1: Design

g2: Education

g3: Project Management

g4: Software Maintenance

Tagged doc N

Tag1: Computer

Tag2: Education

Tag3: IT technology

Tag4: Software engineering

Tag5: Software Quality

Training Data

Auto Classifier

Model

TAG	Confidence Score
Education	0.75
Software Design	0.68
Computer	0.60
IT Technology	0.58

Literatum's content auto-tagger: Example

Existing capability

- 111 tags in Taxonomy
- ~37.5K tagged documents for training purposes
- ~11 minutes to train the model
- ~10 minutes to tag 45K new publications
- ~85% accuracy

Client Testimonial

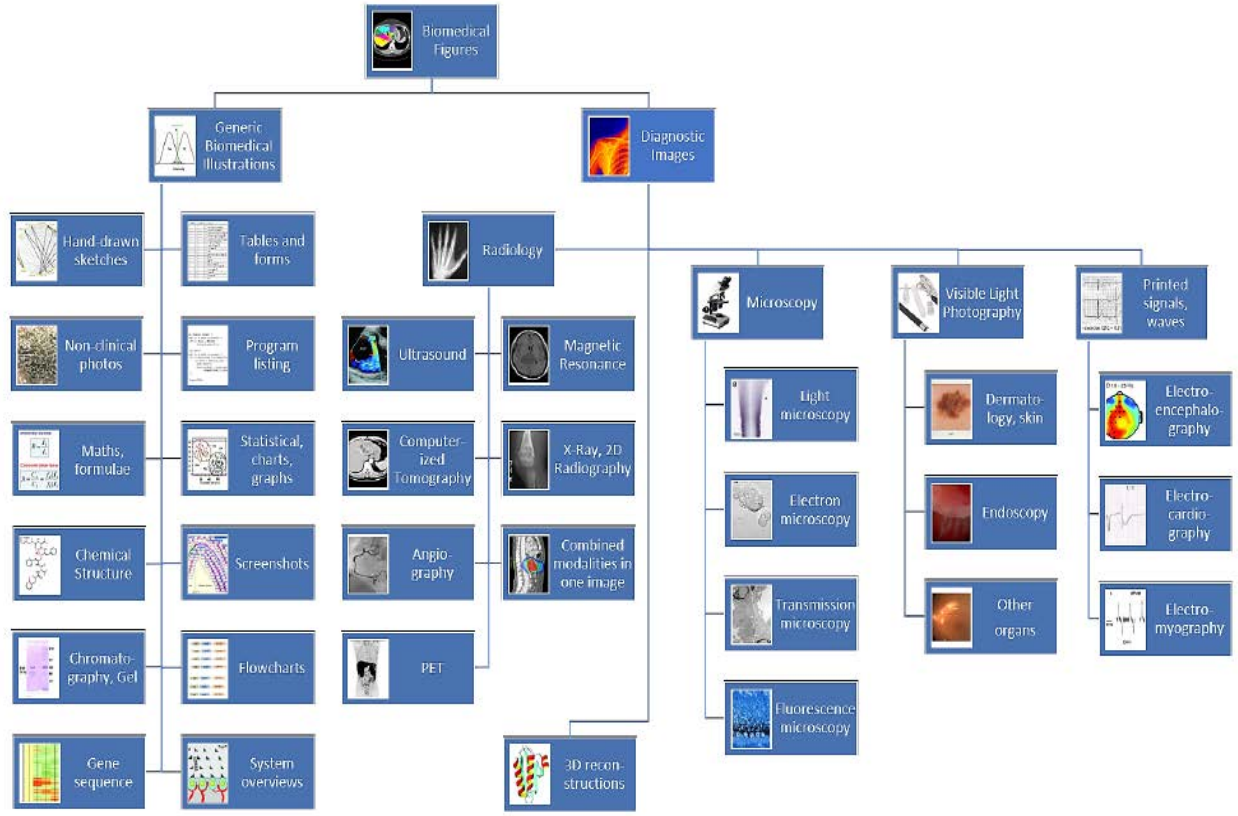
"We use auto-tagger to tag articles for our 'micro-sites'. It works really well for this purpose! We just load the content, the system tags it and then it shows up on the 'micro-sites'!"

Literatum's image auto-tagger

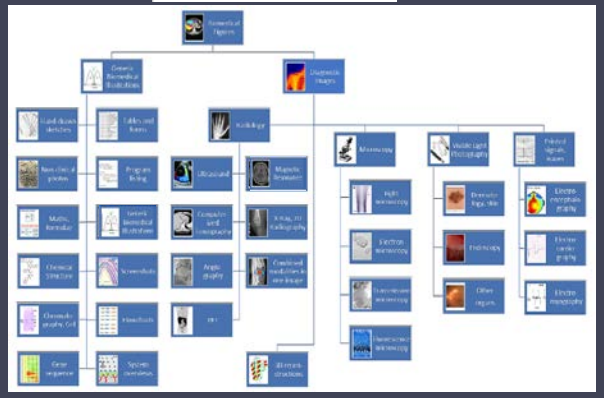
Figure taxonomy

- Ultrasound images
- 3D images
- Graphs and charts
- PET scans
- Microscopies
- Gene sequences

- 4M+ figures only in PubMed
- Biomedical research, education, and clinical decision-making
- Lack of metadata

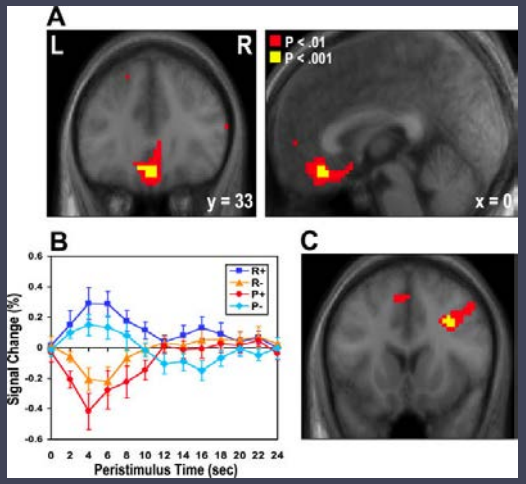


Literatum's image auto-tagger: Example



Magnetic Resonance

Statistical, charts, graphs



CAPTION

Figure 1. A compound figure comprising 3 sub-figures: one graph and two obtained via magnetic resonance imaging

RELEVANT INFORMATION

Figure 1 is an example of compound figure comprising 3 different sub figures and it is used to explain the novel method in this paper

Content summarization tool

Background: Transurethral resection (tur) with cystoscopy is the current main treatment for non-muscle-invasive bladder cancer (nmibc), but residual tumour was found in 30%–44% patients after initial treatment. White light cystoscopy (wlc) was considered the current standard method for detecting tumours during tur, however, its sensitivity and specificity was not entirely satisfactory. This study aimed to conduct a meta-analysis of evidence from randomized controlled trials (rcts).

Materials and methods: An electronic database search of medline, embase and the cochrane database was systematically undertaken to identify studies conducted between 1996 and october 2012. Rcts that assessed the clinical efficacy of fc and compared it with that of wlc in patients with suspected or proven nmibc were included.

Results: The recurrence rate was significantly lower in the fc group than in the wlc group (OR: 0.5; 95% ci: 0.4–0.62; p<0.00001). In the pooled estimates, a statistically significant difference in favour of fc was observed at 1 (HR: 0.69; 95% ci: 0.59–0.81; p>0.00001) and 2 years (p<0.0001). Subgroup analysis also detected a statistically difference between the fc and wlc groups (MD: 7.39 week; 95% ci: 3.87–10.91; p≤0.0001), and the time to first recurrence was delayed significantly (7.39).

Conclusions: Compared with wlc, fc guided tur could significantly decrease recurrence rates, prolong the time to first recurrence after initial tur and improve rfs at 1 and 2 years. Therefore fc was demonstrated to be an effective procedure for delaying recurrence of nmibc. Further studies are required to explore possible reasons.

- Summarize key points of the article
  - Help the users quickly understand the content
  - Can be indexed for improved IR results
  - Especially useful for articles without abstract
  - Automatically suggest an abstract for a manuscript
- Produce a structured summary with key sections
  - Improve readability and comprehension
  - Common practice for many articles especially in life sciences

Content summarization tool: Less technical option

Original Abstract

Methods: Patients with moderate to severe knee osteoarthritis viewed a video about knee osteoarthritis treatments options, including total knee arthroplasty, and received a personalized arthritis report. An adapted version of the western Ontario and McMaster universities osteoarthritis index was used to assess pain and physical function expectations following total knee arthroplasty before/after the intervention. These scores were compared to an age- and gender-adjusted means for a cohort of patients who had undergone total knee arthroplasty. Decision readiness and conflict were also measured.

- Part of the original abstract of the same article
- Highlighted are the overlapping parts

Technical Summary

Methods: Male and female patients with moderate to severe knee oa, defined as having a score of >39 on the western Ontario and McMaster universities osteoarthritis index (*WOMAC*), were invited to participate in this study. Participants completed baseline questionnaires and were invited to a group meeting to attend the study intervention.

- More detailed and technically sound
- Usually is meant for experts

Less Technical Summary

Methods: Male and female patients were recruited prospectively using a computer-assisted survey. Participants were invited to a group meeting to attend the research assistant and were screened using a western Ontario and McMaster universities osteoarthritis index (*WOMAC*).

- Easier to understand if not an expert
- Usually more compact and less detailed

2. Information extraction

Text-based

Key metadata extraction from PDFs: Keywords, grant info, figures, tables, captions, citations, titles, affiliations, abstracts, etc.

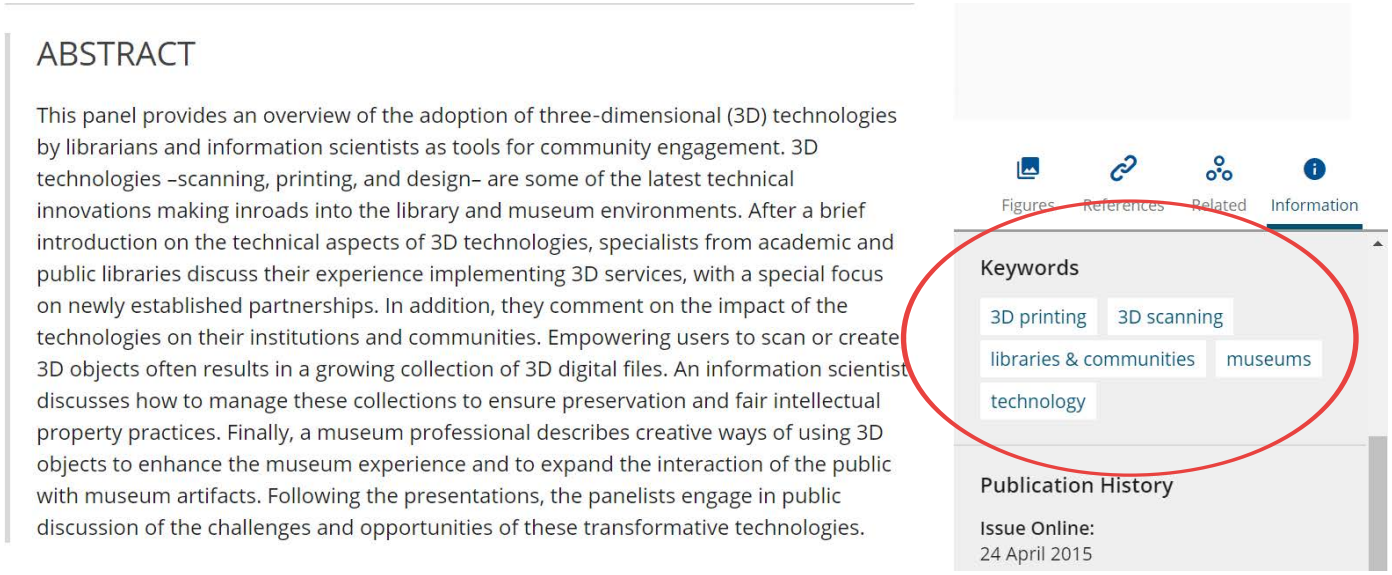
Image-based

Compound figure segmentation

Automatic key phrase extraction

Use: Improve content readability and discoverability and search results quality

Example: Extract article's actual keywords vs. relying on what author submits

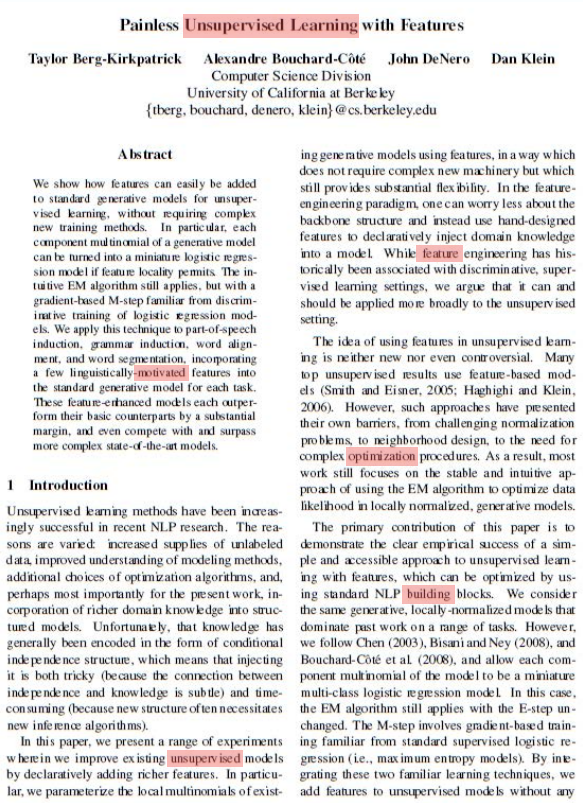


ABSTRACT

This panel provides an overview of the adoption of three-dimensional (3D) technologies by librarians and information scientists as tools for community engagement. 3D technologies –scanning, printing, and design– are some of the latest technical innovations making inroads into the library and museum environments. After a brief introduction on the technical aspects of 3D technologies, specialists from academic and public libraries discuss their experience implementing 3D services, with a special focus on newly established partnerships. In addition, they comment on the impact of the technologies on their institutions and communities. Empowering users to scan or create 3D objects often results in a growing collection of 3D digital files. An information scientist discusses how to manage these collections to ensure preservation and fair intellectual property practices. Finally, a museum professional describes creative ways of using 3D objects to enhance the museum experience and to expand the interaction of the public with museum artifacts. Following the presentations, the panelists engage in public discussion of the challenges and opportunities of these transformative technologies.

Automatic key phrase extraction: Example

Available now



Painless Unsupervised Learning with Features

Taylor Berg-Kirkpatrick Alexandre Bouchard-Côté John DeNero Dan Klein
Computer Science Division
University of California at Berkeley
{tberg, bouchard, denero, klein}@cs.berkeley.edu

Abstract

We show how features can easily be added to standard generative models for unsupervised learning, without requiring complex new training methods. In particular, each component multinomial of a generative model can be turned into a miniature logistic regression model if feature locality permits. The intuitive EM algorithm still applies, but with a gradient-based M-step familiar from discriminative training of logistic regression models. We apply this technique to part-of-speech induction, grammar induction, word alignment, and word segmentation, incorporating a few linguistically-motivated features into the standard generative model for each task. These feature-enhanced models each outperform their basic counterparts by a substantial margin, and even compete with and surpass more complex state-of-the-art models.

1 Introduction

Unsupervised learning methods have been increasingly successful in recent NLP research. The reasons are varied: increased supplies of unlabeled data, improved understanding of modeling methods, additional choices of optimization algorithms, and, perhaps most importantly for the present work, incorporation of richer domain knowledge into structured models. Unfortunately, that knowledge has generally been encoded in the form of conditional independence structure, which means that injecting it is both tricky (because the connection between independence and knowledge is subtle) and time-consuming (because new structure often necessitates new inference algorithms).

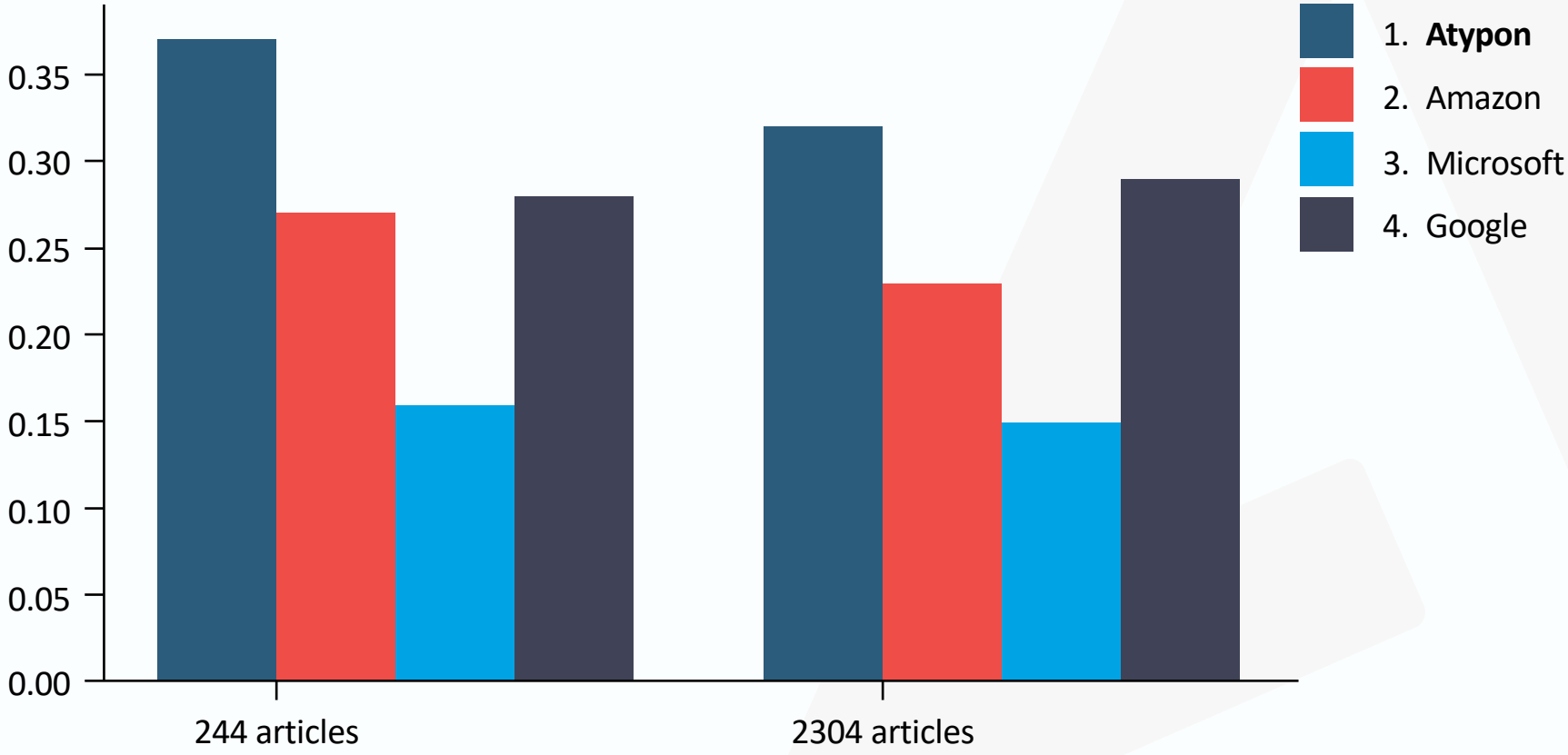
In this paper, we present a range of experiments wherein we improve existing unsupervised models by declaratively adding richer features. In particular, we parameterize the local multinomials of existing generative models using features, in a way which does not require complex new machinery but which still provides substantial flexibility. In the feature-engineering paradigm, one can worry less about the backbone structure and instead use hand-designed features to declaratively inject domain knowledge into a model. While feature engineering has historically been associated with discriminative, supervised learning settings, we argue that it can and should be applied more broadly to the unsupervised setting.

The idea of using features in unsupervised learning is neither new nor even controversial. Many top unsupervised results use feature-based models (Smith and Eisner, 2005; Haghighi and Klein, 2006). However, such approaches have presented their own barriers, from challenging normalization problems, to neighborhood design, to the need for complex optimization procedures. As a result, most work still focuses on the stable and intuitive approach of using the EM algorithm to optimize data likelihood in locally normalized, generative models.

The primary contribution of this paper is to demonstrate the clear empirical success of a simple and accessible approach to unsupervised learning with features, which can be optimized by using standard NLP building blocks. We consider the same generative, locally-normalized models that dominate past work on a range of tasks. However, we follow Chen (2003), Bisani and Ney (2008), and Bouchard-Côté et al. (2008), and allow each component multinomial of the model to be a miniature multi-class logistic regression model. In this case, the EM algorithm still applies with the E-step unchanged. The M-step involves gradient-based training familiar from standard supervised logistic regression (i.e., maximum entropy models). By integrating these two familiar learning techniques, we add features to unsupervised models without any

Rank	Key Phrase
1	Word segmentation
2	Unsupervised learning
3	EM algorithm
4	NLP
5	Features
6	Standard generative models

Performance of Atypon's automatic key phrase extractor



Key metadata extraction

KP-Miner: Participation in SemEval-2

Samhaa R. El-Beltagy
Cairo University
Giza, Egypt
samhaa@computer.org

Ahmed Rafea
The American University in Cairo
New Cairo, Egypt
rafea@aucegypt.edu

Abstract

This paper briefly describes the KP-Miner system which is a system developed for the extraction of keyphrases from English and Arabic documents, irrespective of their nature. The paper also outlines the performance of the system in the "Automatic Keyphrase Extraction from Scientific Articles" task which is part of SemEval-2.

1 Introduction

KP-Miner (El-Beltagy, 2006) (El-Beltagy, 2009) is a system for the extraction of keyphrases from English and Arabic documents. When developing the system, the goal was to build a general purpose keyphrase extraction system that can be easily configured by users based on their understanding of the documents from which keyphrases are to be extracted and without the need for any training documents or the use of any sophisticated natural language processing or linguistic tools. As such, the keyphrase extraction process in KP-Miner is an un-supervised one. When building a general purpose keyphrase extraction system, this was an important objective, as training data is not always readily available for any type of data. The goal of entering the KP-Miner system into the SemEval-2 competition, was to see how well it will perform on a specific task, without making any changes in its default parameters.

2 System Overview

Keyphrase extraction in the KP-Miner system is a three step process: candidate keyphrase selection, candidate keyphrase weight calculation and finally keyphrase refinement. Each of these steps, is explained in the following sub-sections. More details about the employed algorithm, and justification for using certain values for selected parameters, can be found in (El-Beltagy, 2009).

2.1 Candidate keyphrase selection

In KP-Miner, a set of rules is employed in order to elicit candidate keyphrases. As a phrase will never be separated by punctuation marks within some given text and will rarely have stop words within it, the first condition a sequence of words has to display in order to be considered a candidate keyphrase, is that it is not be separated by punctuation marks or stop words. A total of 187 common stopwords (the, then, in, above, etc) are used in the candidate keyphrase extraction step. After applying this first condition on any given document, too many candidates will be generated; some of which will make no sense to a human reader. To filter these out, two further conditions are applied. The first condition states that a phrase has to have appeared at least n times in the document from which keyphrases are to be extracted, in order to be considered a candidate keyphrase. This is called the least allowable seen frequency(lasf) factor and in the English version of the system, this is set to 3. However, if a document is short, n is decremented depending on the length of the document.

The second condition is related to the position where a candidate keyphrase first appears within an input document. Through observation as well as experimentation, it was found that in long documents, phrases occurring for the first time after a given threshold, are very rarely keyphrases. So a cutoff constant CutOff is defined in terms of a number of words after which if a phrase appears for the first time, it is filtered out and ignored. The initial prototype of the KP-Miner system (El-Beltagy, 2006), set this cutoff value to a constant (850). Further experimentation carried out in (El-Beltagy, 2009) revealed that an optimum value for this constant is 400. In

Available now

Paper submission form

Add Article / Poster PDF: Drag and Drop PDF File Or Select One

Content Type: (•) Article/Manuscript ( ) Poster

Title: KP-Miner: Participation in SemEval-2

Abstract: This paper briefly describes the KP-Miner system which is a system developed for the extraction of key phrases from English and Arabic documents, irrespective of their nature. The paper also outlines the performance of the system

Taxonomy Tags: AI | Key Extraction | Machine Learning | NLP

Funders:

Funders	Grand Number	Recipient
The Wellcome Trust	206545/Z/17/Z	Cairo University

Authors:

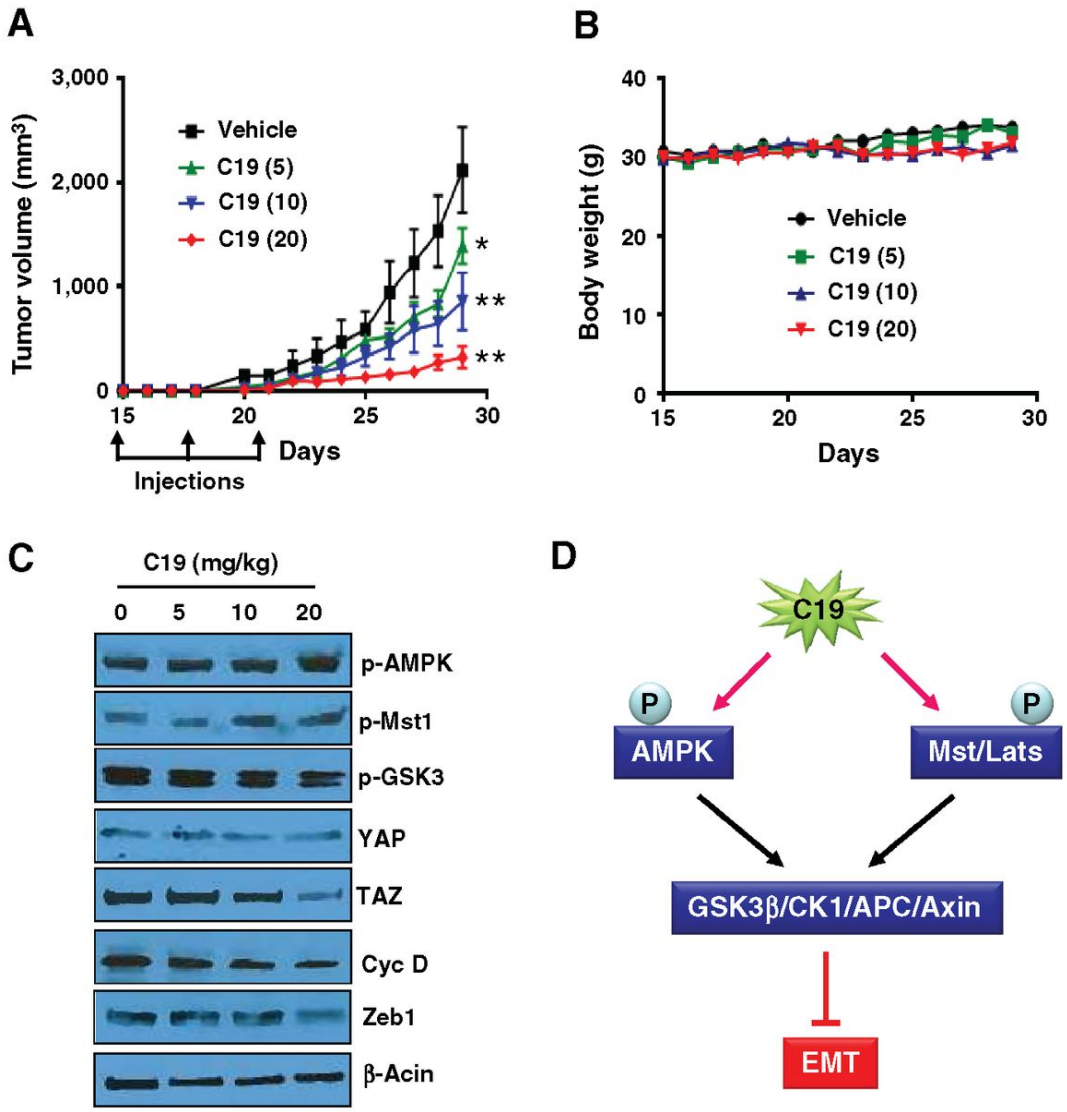
First name	Last Name	Affiliation	Contact
Samhaa	Beltagy	Computer Science, Cario University	samhaa@computer.org
Ahmed	Rafea	American University in Cairo	rafea@aucegypt.edu

Compound figure extraction

Challenges

- Compare compound figures with different info
- Manually segment compound figures

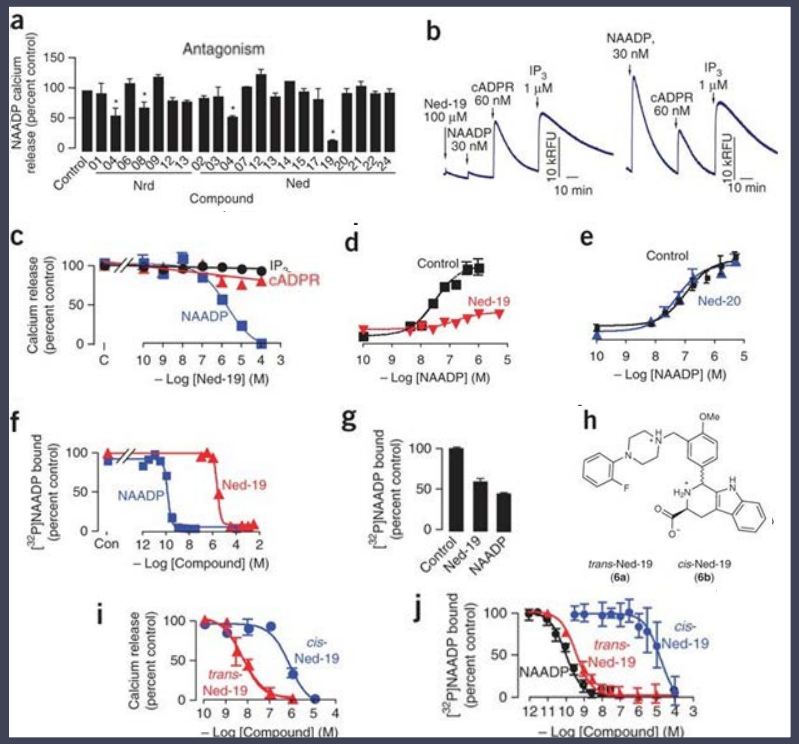
Available now



Compound figure extraction: Example

Antagonism

cADPR



a. NAADP compound b. NAADP IP3 c. Calcium release

d. NAADP Log Ned-19 e. NAADP Log Ned-20

f. NAADP bound g. Control Ned-19 h. cis-Ned-19

i. Log compound M j. Percent control bound

3. Information retrieval

Search

- Query
- Ranking

Recommendation

- Intelligent recommendation for paper, venue, users
- Cross-publisher recommendation

Intelligent Google-like auto-suggest

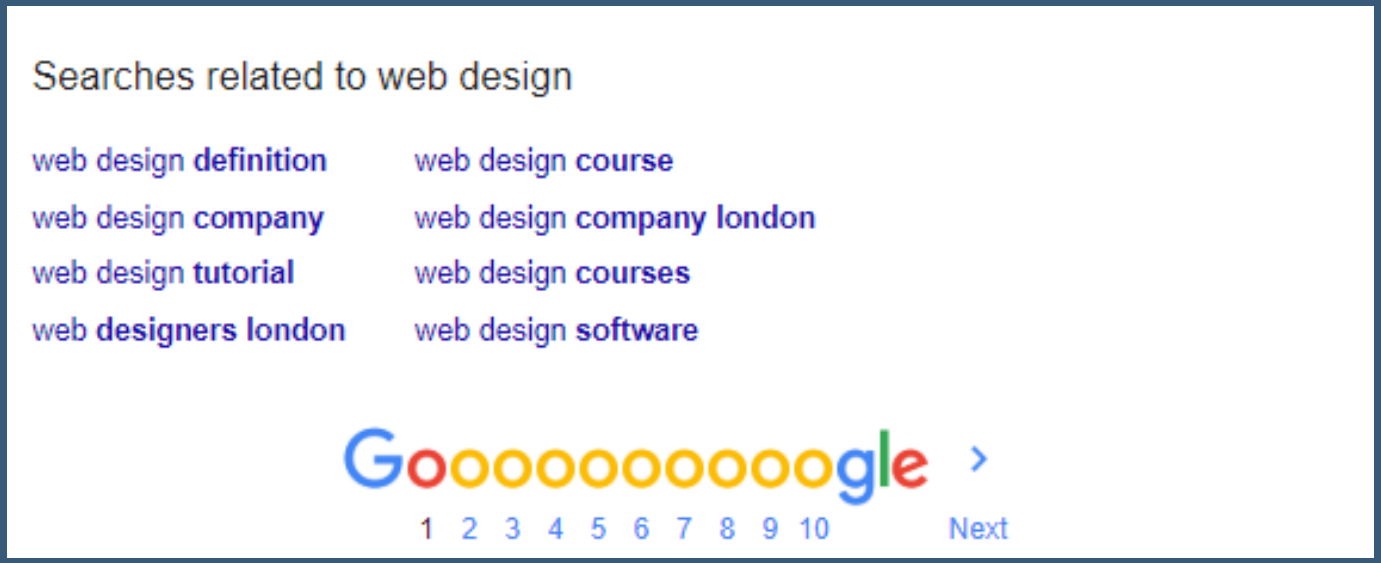
Available now as an upgrade

Search Term = “waste”

Before

After

Query suggestion

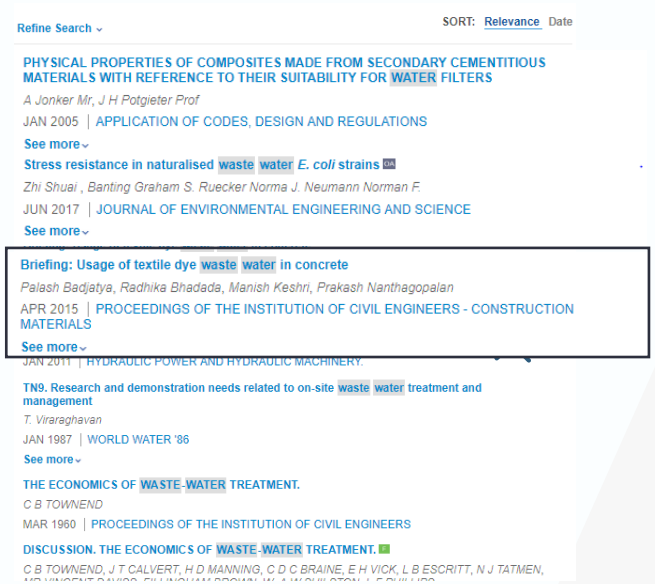


Under evaluation

Personalized search
Keyword-based search

Launching in early 2020

Waste Water



SEARCH ENGINE

- Browsing history
- Click history
- Popularity
- Recency
- Relevance

MACHINE LEARNING

RERANK

Automated natural language Q&A: The future of search

Semantic search

- What drugs reduce anxiety?
- Are there common genes between comorbid diseases?
- What are factors for liver cancer among diabetics?

Causal analytics

- How does a protein interact with a disease?
- What is the comorbidity of two diseases?

New recommendation engine

Available now




DOWNLOAD

FAVOURITE

VIEW

CLICK



MACHINE LEARNING

User	Interest 1	Interest 2	Interest 3	...
	Computer science	Machine learning	Nature science	...
	Biomedical	Machine learning	Health care	...
	Biomedical	Health care	Nature science	...
	Health policy	Health care	Nature science	...

Content	Attribute 1	Attribute 2	Attribute 3	...
	Computer	Science	Biomedical	...
	AI	Machine learning	Recommend-ation	...
	Biomedical	Health	Policy	...
	Health care	Policy	Science	...

Atypon's New Recommendation Engine

Available now

New algorithms improve speed and accuracy of discovery

- Scalable for millions of users and items
- **Accuracy**: Average 60% more accurate— and often much more accurate
- **Speed**: Average 60% faster on a single machine, and can be many times faster on a cluster

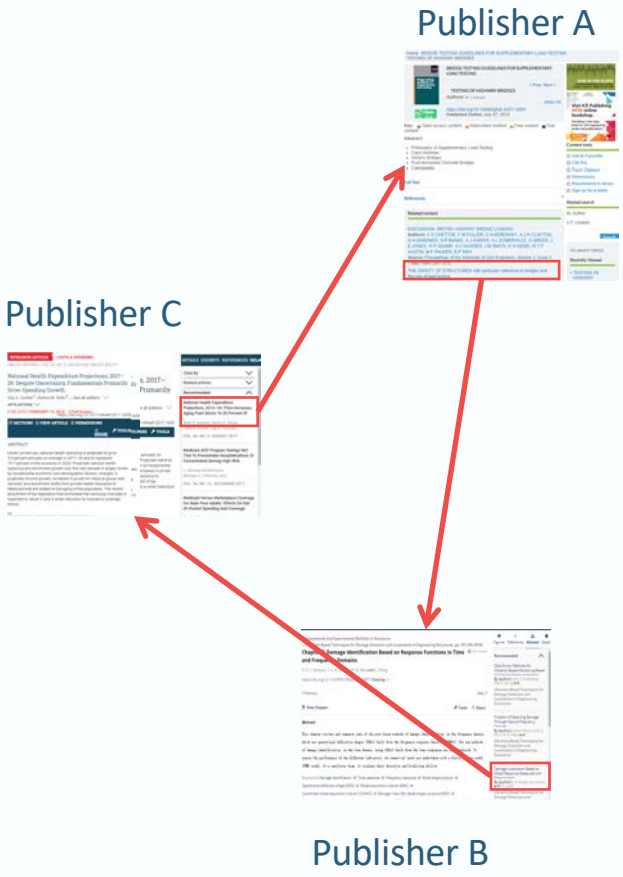
Atypon's recommendation engine: Example applications

Available now

Recommend content *and* people:

- Changing the list of articles that appears on your home page for visitors you know
- Identify users who read an article on a specific topic and then promote a conference on the same topic to them
- Help authors identify collaborators from around the world with similar interests

Cross-publisher recommendations (like TrendMD)



Atypon's advantages:

- Quality information
- Seamless integration
- Trustable content
- Various types of content
- New channels/readers

Journal & conference finder

- Personalized recommendations
- Search all indexed journals and conferences
- Find the best place for publication

Reviewer/collaborator finder

Pre-publication challenges:

- Draft (Researcher)
  - Who can provide a preliminary review of my draft?
  - Who are appropriate collaborators in my field of specialty?
  - Who can help me improve my rejected publication?
- Review (Publisher)
  - Effortless review allocation?
  - Most suitable reviewer profile?

A valuable tool for:

Publishers

Speed-up reviewer identification process

Identify best candidate reviewers

Researchers

Identify collaborators for manuscripts

Request informal peer-review for increased research quality

Can we address these challenges?

- Atypon's "Reviewer/Collaborator finder" tool

Systematic article review tool

Title: Galactomannan detection for invasive aspergillosis in immunocompromised patients .
Objective: Our primary objective was to assess the diagnostic accuracy of galactomannan detection in serum for the diagnosis of invasive aspergillosis in immunocompromised patients, at different cut-off values for test positivity.
Type of Study, Participants, Target Conditions, Reference Standards

"Aspergillus"[MeSH], "Aspergillosis"[MeSH] "Pulmonary Aspergillosis"[MeSH], aspergill*[tiab] fungal infection[tw], (invasive[tiab] AND fungal[tiab])
#1 OR #2 OR #3 OR #4 OR #5 OR 6,
"Serology"[MeSH]
Serology"[MeSH],
(serology[tiab] OR serodiagnosis[tiab] OR serologic[tiab]), "Immunoassay"[MeSH]
(immunoassay[tiab] OR immunoassays[tiab])

Text Mining

PubMed

Efficiently ranked documents

Feedback

Learning-to-rank Model

4.
User profiling

- User clustering/targeting
- User intention/interests detection
- Predictive analytics on usage data

User grouping & intent detection

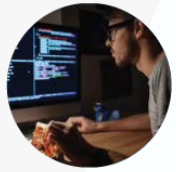
Biomedical Science
Computational chemistry

Business management
Accounting & finance

Software development
Python programming

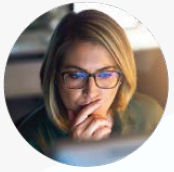






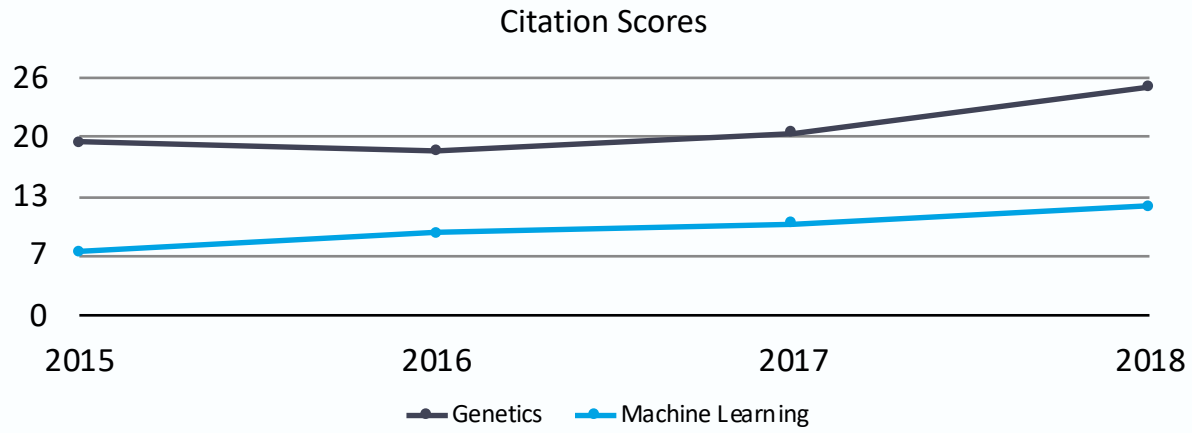


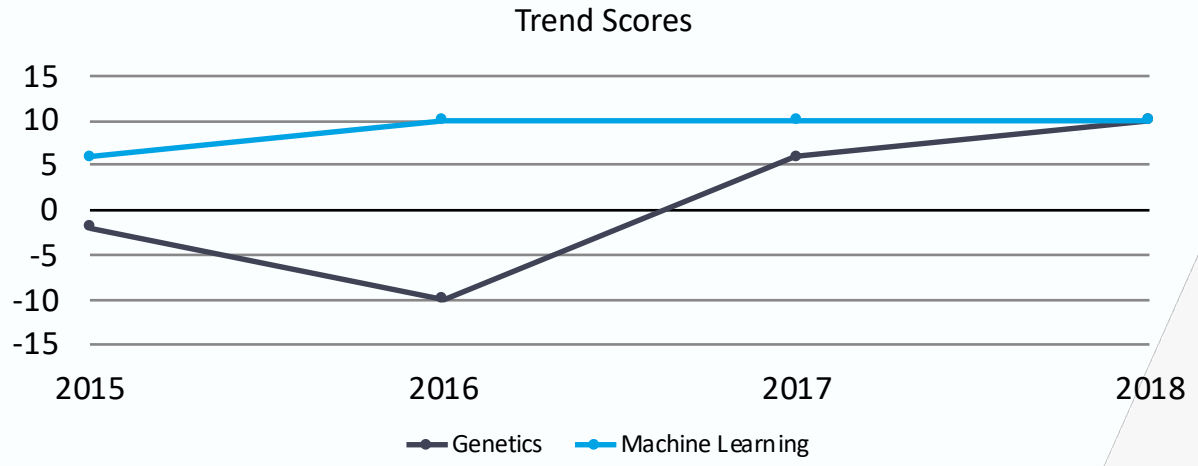


5. Trend/Impact analysis

- Impact prediction
- Discover trending topics
- Influence scores of entities

Impact/Trend Analysis





- **Citation score:** article citations by topic
  - Some topics get more citations than others
  - Useful when trying to prioritize articles
  - Changes over time

- **Trend score:** Increase/decrease in popularity
  - Change and adapt over time
  - Articles on trending topics get more attention

Article citation prediction

How is academic success measured?

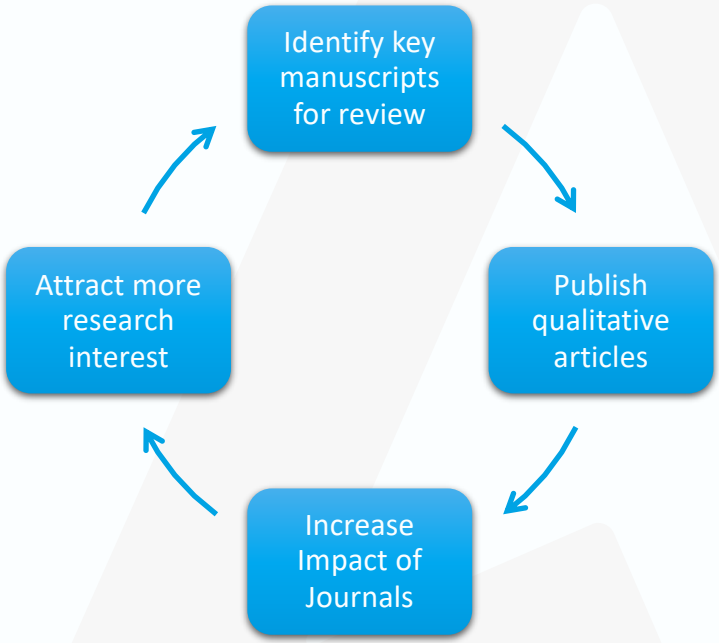
- Citation-based metrics
  - Author h/i-index
  - Citation counters
  - Journal impact factor
- Altmetrics
  - Social media
  - Online dissemination

Limitations of current metrics:

- Unfair to compare publications from
  - Different years
  - Distinct scientific disciplines
- Diversity across scientific fields in
  - Citation velocity
  - Citation accumulation frequency

Article Citation prediction

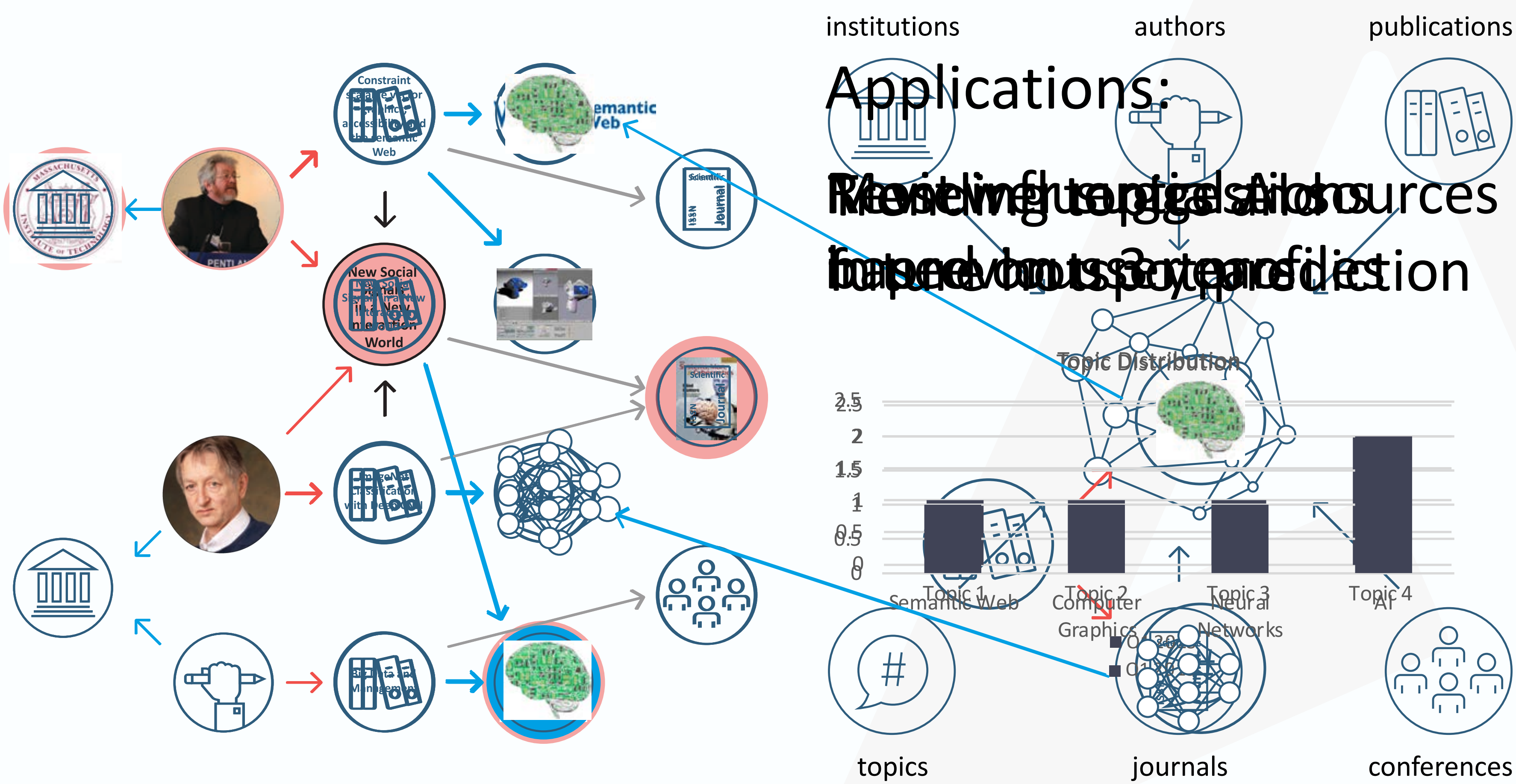
- Existing success metrics refer to *already* published material
- How can we help publishers prioritize submissions?
- **Atypon's "article citation prediction" tool**
  - Estimates number of citations an article is likely to receive
  - Addresses limitations of current success metrics



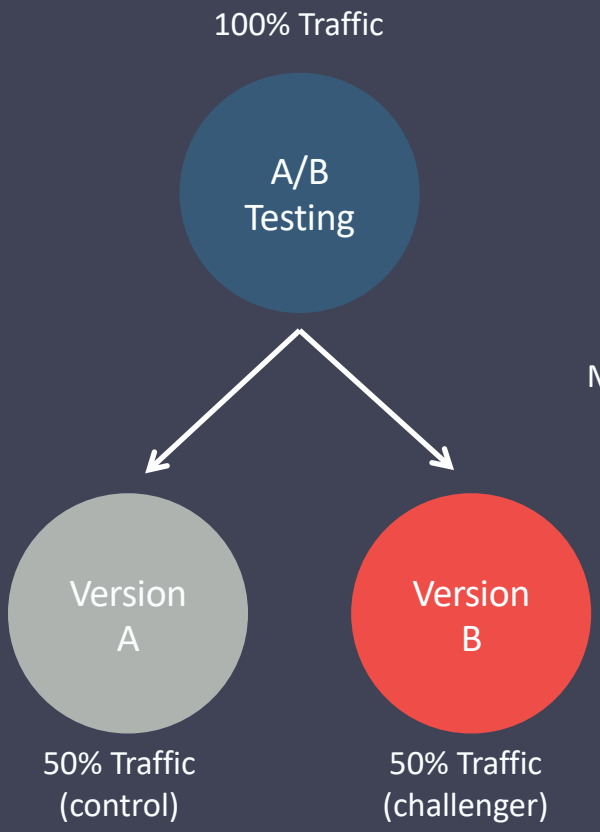
6. Knowledge representation

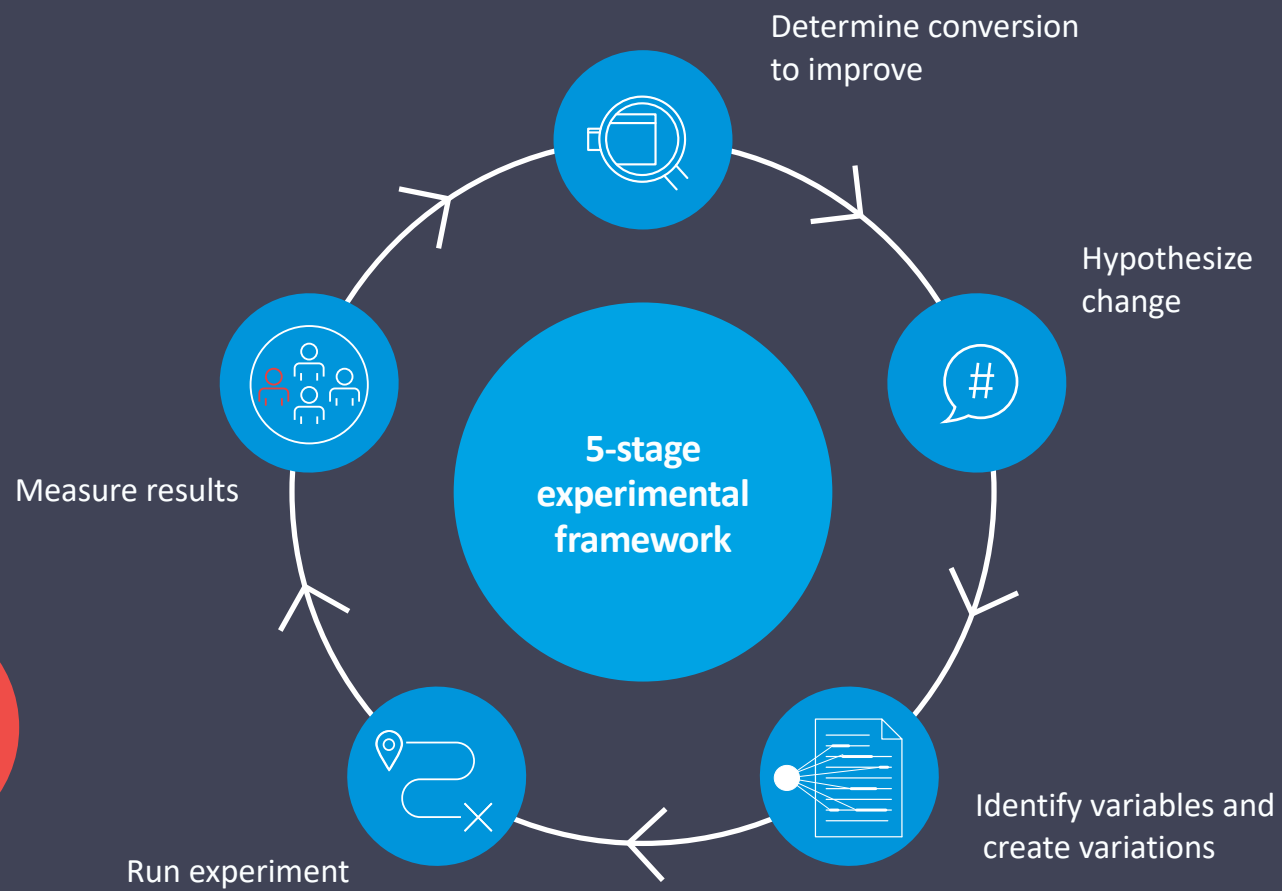
- Publication knowledge graphs
- Domain specific knowledge graphs

Connect and organize information: Content and people



Atypon's A/B testing framework





A/B testing monitoring example

AB Experiment Administration

4	100003	Relevance Ranking	Relevance Feedback query expansion	Tue May 08 00:00:00 PDT 2018	100 %	N/A
5	100006	Relevance Ranking	A/A ranking	Fri Jul 13 00:00:00 PDT 2018	100 %	N/A
6	100011	Relevance Ranking	Evaluating updated configuration for PRF	Mon Oct 08 00:00:00 PDT 2018	100 %	2019-01-23 03:59:22

Evaluating updated configuration for PRF

Evaluating different configuration for relevance feedback query expansion (please check LIT-235866)

Creation Date: Mon Oct 08 00:00:00 PDT 2018

Type: Relevance Ranking

Incoming Traffic: 1.0

Update evaluation metrics

Experiment variations

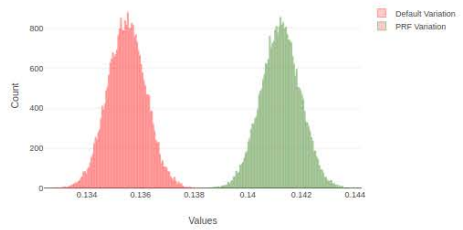
ID	Variation name	Class	Percentage split	Configuration	Constraints
100020	Default Variation	DefaultVariation	50 %	{}	{}
100021	PRF Variation	SearchParamVariation	50 %	{relevanceFeedbackEnabled=true}	{}

Display metrics

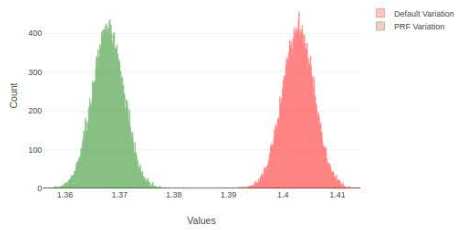
- Click Through Rate (ctr)
- Number of distinct sessions for each user (engagement)
- Average distance of query chain (search_distance)
- Items viewed per query chain (viewed_items)
- Ratio of queries that did not lead to a click (abandonment)

Aggregated Statistics	Default Variation	PRF Variation
Average number of queries per query chain	5.354	4.98152
Total number of users	101435.0	101454.0
Median number of distinct session per user	1.0	1.0
Median number of clicks per query chain	0.0	0.0
Viewed items per query chain	0.24350	0.26104
Median number of queries per query chain	1.0	1.0

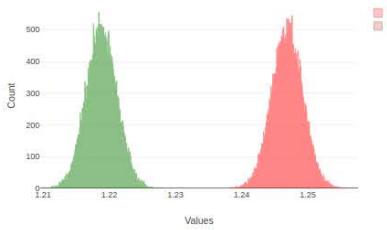
Click Through Rate



Number of distinct sessions for each user



Average distance of query chain



Atypon's AI-based feature roadmap
by release date

2017. 1 – 2018.3

Faster ATM

Relevance feedback

New auto-tagger

Compound figure segmentation

New ATM-based recommendations

Intelligent auto-suggest

2019.1

Query suggestion

Journal finder

Publication knowledge graph

New CF-based recommendations

Key entity extraction

Author disambiguation

2020.1

Global search

Reviewer/collaborator finder

Impact factor prediction for new publication

Figure/Table snippet extraction

Authority measurement

Grant information extraction

Personalized search

Cross-publisher recommendations

Future

Image search

Information retrieval powered by deep learning

Figure modality classification

Question answering system with knowledge graph in different domains

Auto classification without examples

Content summarization

Automated taxonomy creation

Automated poster generation for scientific papers

Automatic Q&A

Conclusion

- Atypon AI research published in ***Journal of Machine Learning Research*** and ***Biomedical Semantics***
- **1st and 2nd place** in BioASQ Challenge 2018 and CLEF eHealth Challenge 2018
- Long-term, continuous investment in R&D
- We welcome more use cases, problems and ideas!


ATYPON
WebinarSeries
Get More Out of Literatum
A new free 10-webinar series
Register at atypon.com/webinars
UP NEXT:
WEDNESDAY, AUGUST 7
10:30 AM ET
Deepening readers' engagement with your content
TUESDAY, SEPTEMBER 10, 10:30 AM ET
Attract targeted audiences, extend their time onsite, and monetize their visits— through content marketing
TUESDAY, SEPTEMBER 26,
10:30 AM ET
Creating multimedia content for your website— a hands-on demo of Digital Objects


Q&A


ATYPON
WebinarSeries
Register at atypon.com/webinars
@atypon
atypon.com
info@atypon