

# Leveraging the Power of the Crowd to Save the Web

Vishwajeet Pattanaik\*

Shweta Suran

Dirk Draheim

Tallinn University of Technology

Estonia

## WORLD WIDE WEB

The WorldWideWeb (W3) is a wide-area hypermedia[1] information retrieval initiative aiming to give universal access to a large universe of documents.

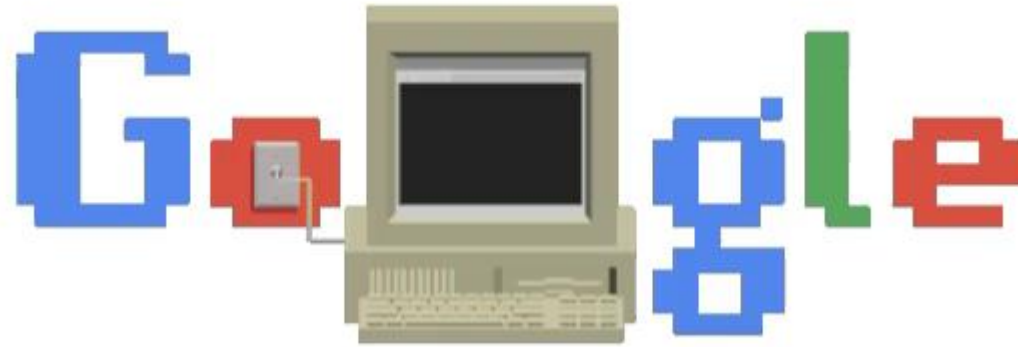
Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary[2] of the project, Mailing lists[3] , Policy[4] , November's W3 news[5] , Frequently Asked Questions[6] .

What's out there?[7]Pointers to the world's online information,  
subjects[8] , W3 servers[9], etc.

Help[10] on the browser you are using

Software Products[11] A list of W3 project components and their current state. (e.g. Line Model[12] ,X11 Viola[13] , NeXTStep[14] , Servers[15] , Tools[16] , Mail robot[17] , Library[18] )

Technical[19] Details of protocols, formats, program internals etc



On 12<sup>th</sup> March this year, the Web turned 30!

Tim Berners-Lee wrote his memo “Information Management: A Proposal” which outlined the World Wide Web.

\*Source: [Google Doodles Achieve](#)

“The Web is starting to wane in the face of a ‘nasty storm’ of issues”

– Tim Berners-Lee\*

\*[Tim Berners-Lee on the future of the web: 'The system is failing'](#), Olivia Solon, The Guardian, November' 2017

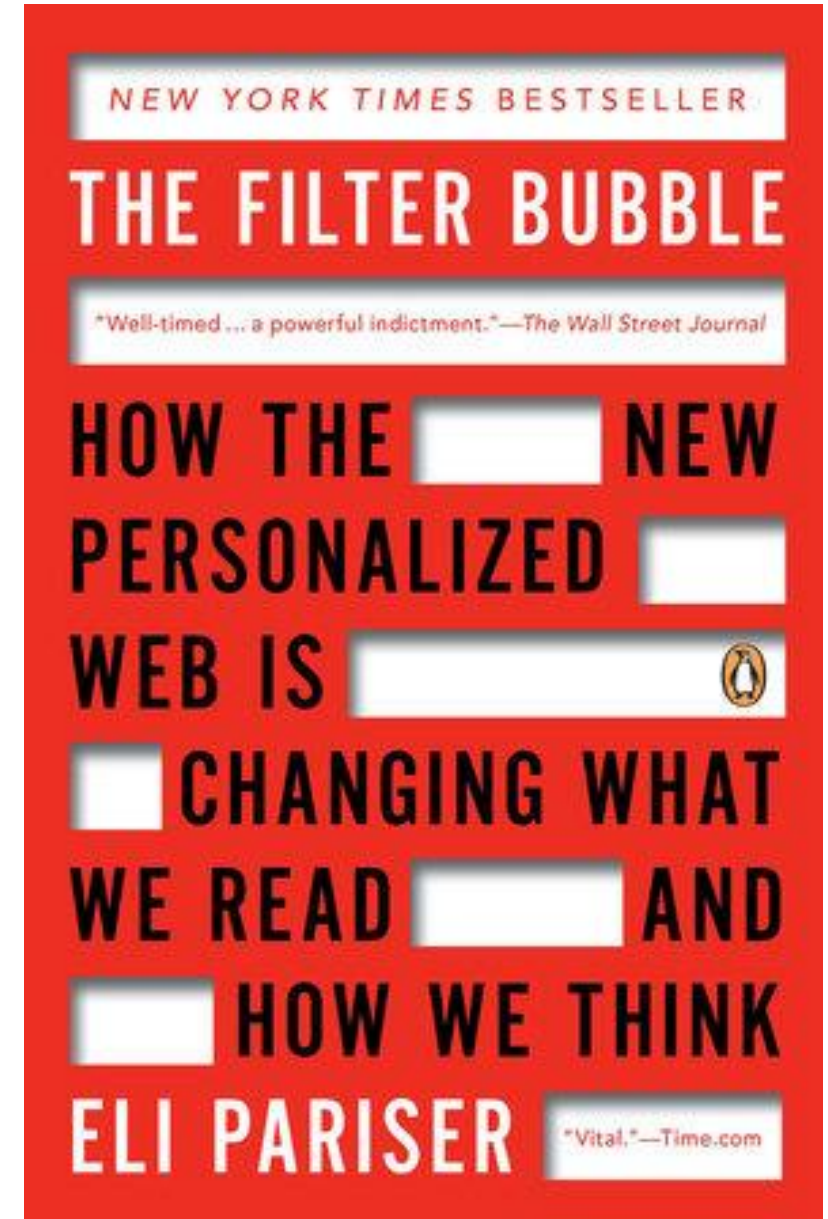
# Threats Facing the Web

- filter bubble [[978-1-59-420300-8](#)]
- clickbait [[10.1007/978-3-319-63751-8](#)]
- link rot (or, web page decay) [[10.1007/s00799-016-0171-9](#)]
- fake news [[10.1126/science.aao2998](#)]
- weaponised AI propaganda (or, behavioural microtargeting) [[10.1353/jod.2017.0025](#)]

# Filter Bubble

*“... refers to the concept that a website’s personalization algorithm selectively predicts the information that users will find of most interest based on data about each individual – including signals such as their history of Likes, search history, and other past online behavior – and that this creates a form of online isolation from a diversity of opinions ...” i.e., **echo chambers***

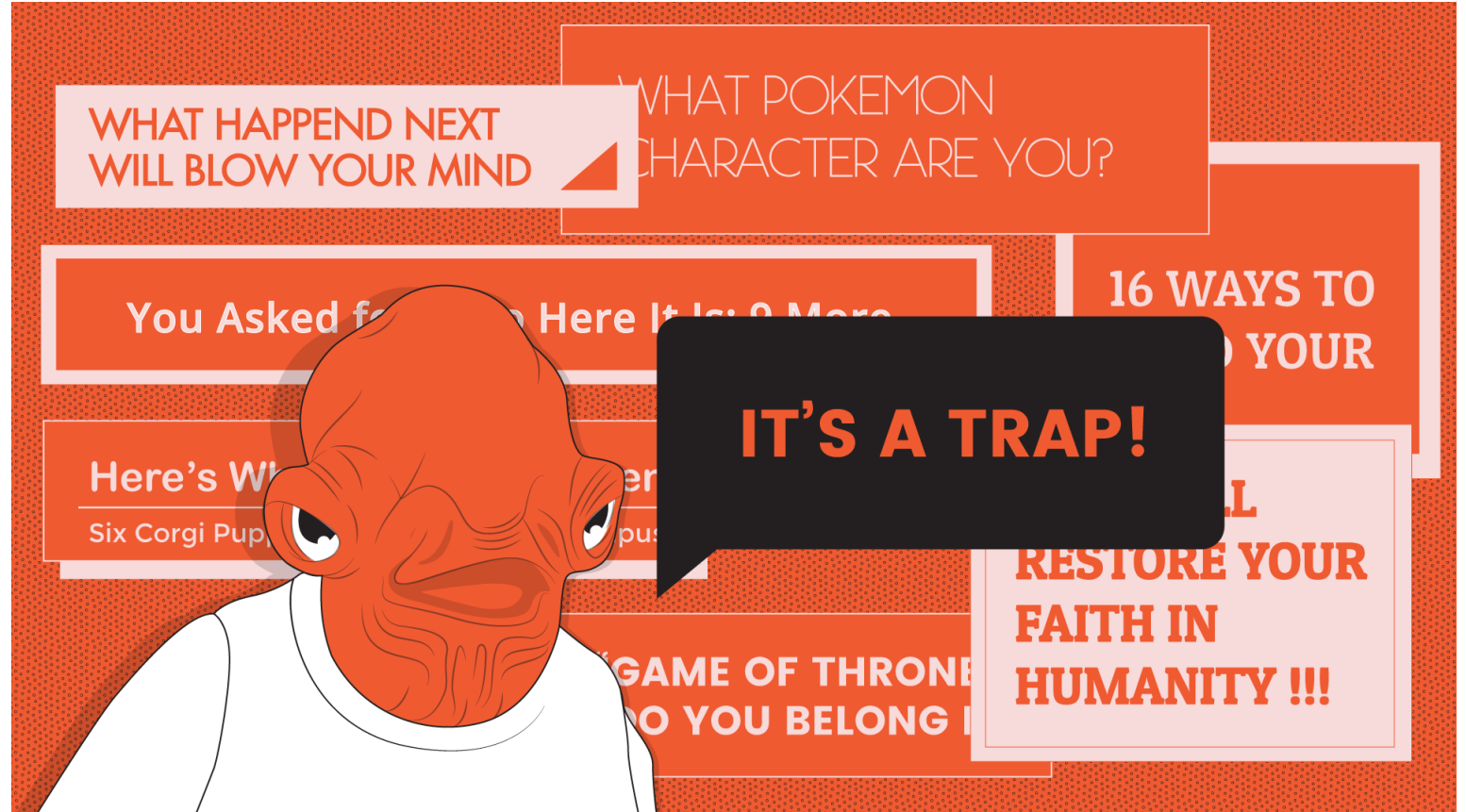
[[10.1016/j.dcm.2018.03.005](https://doi.org/10.1016/j.dcm.2018.03.005)]



# Clickbait

*“... refers to social media messages that are foremost designed to entice their readers into clicking an accompanying link to the posters’ website, at the expense of informativeness and objectiveness ...”*

[\[arXiv:1812.10847v1\]](https://arxiv.org/abs/1812.10847v1)





# Fake News

*... refers to “fabricated information that mimics news media content in form but not in organizational process or intent”*

[\[10.1126/science.aao2998\]](https://doi.org/10.1126/science.aao2998)



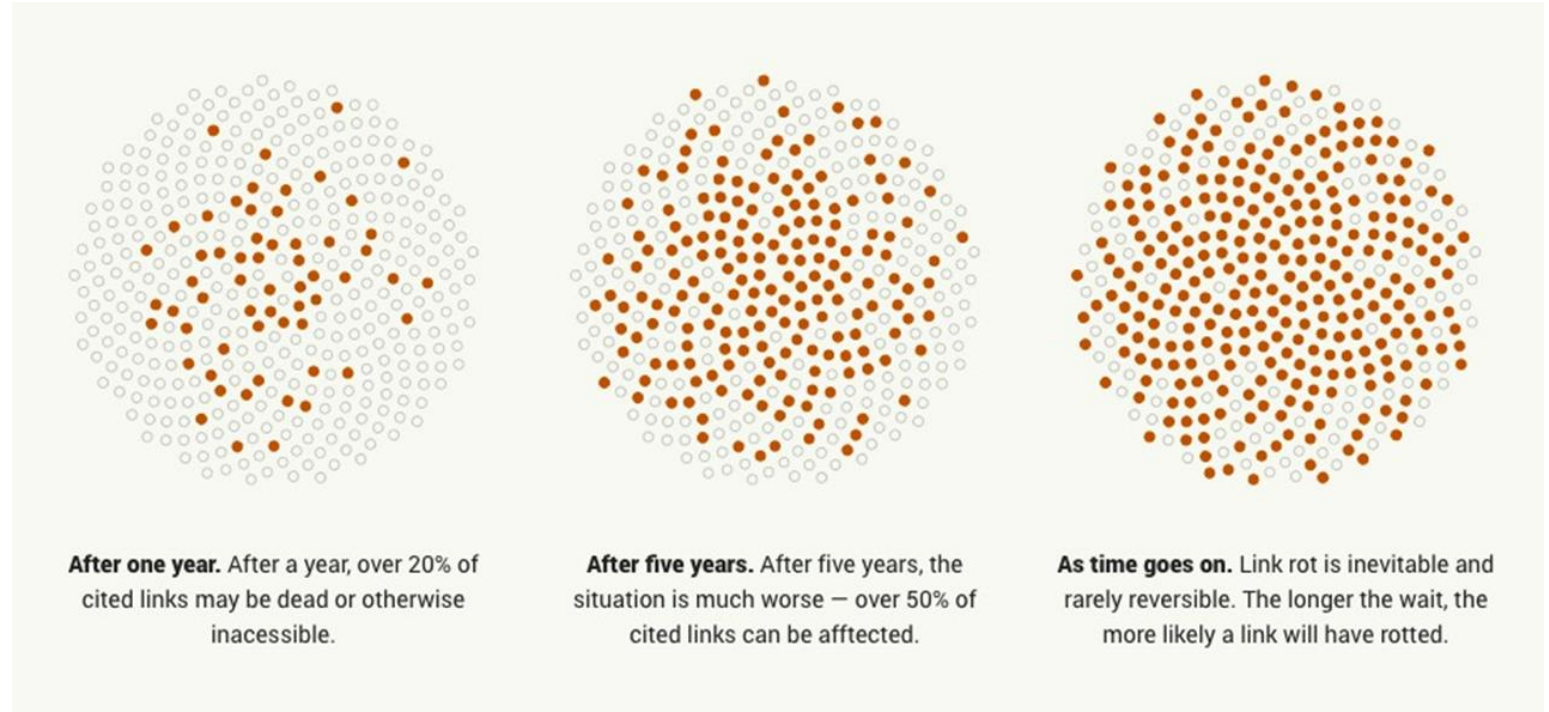


# Link rot

*... refers to “broken or altered links, and web content which has changed, disappeared or moved”*

[\[10.6084/m9.figshare.7090694.v1\]](https://doi.org/10.6084/m9.figshare.7090694.v1)

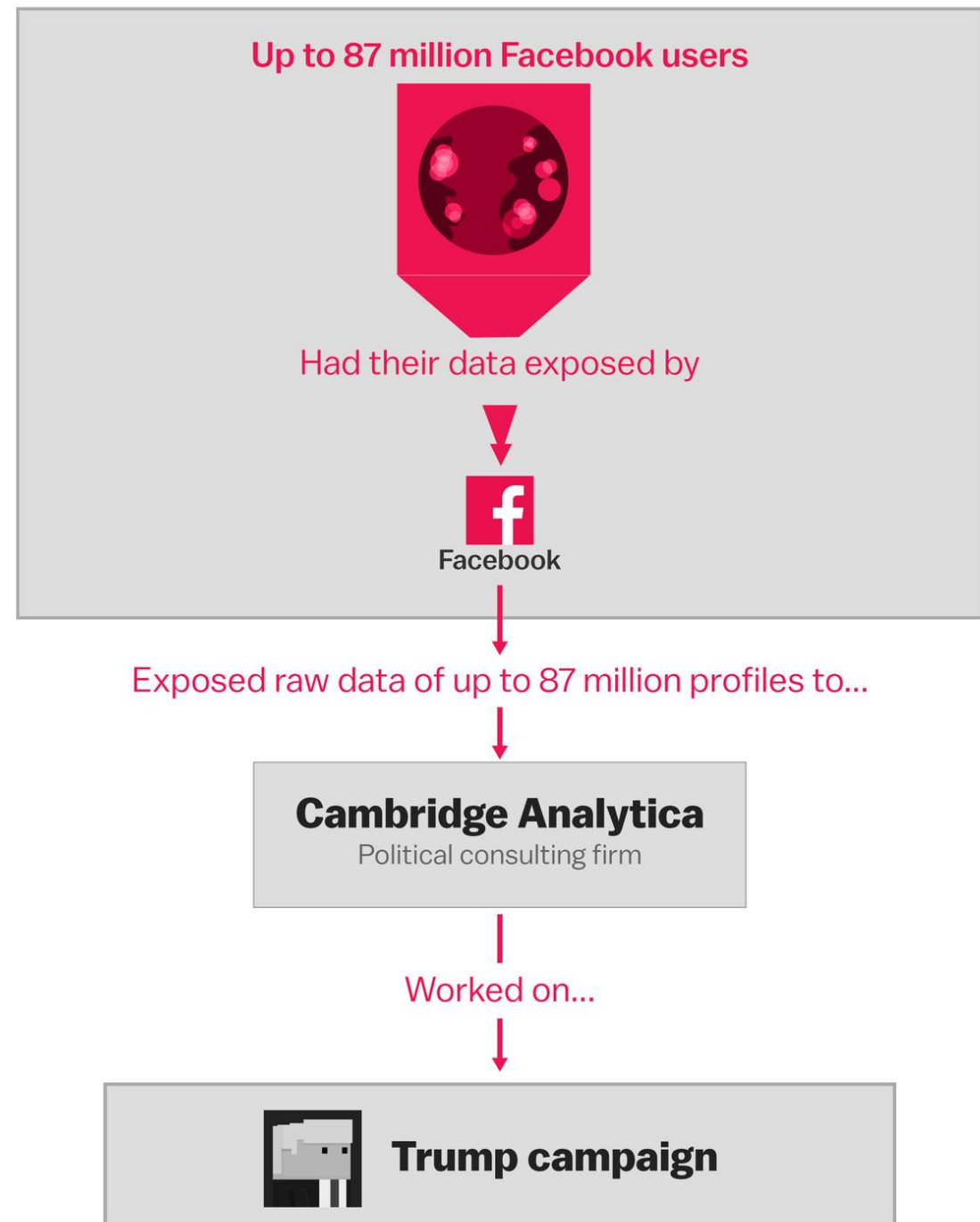
- more than **69%** web pages change within days [\[10.1145/1326561.1326566\]](https://doi.org/10.1145/1326561.1326566)
- **11%** of the shared content on social media are completely lost within a year [\[10.1007/978-3-642-33290-6\\_14\]](https://doi.org/10.1007/978-3-642-33290-6_14)
- the decay rate of web documents has dropped to **nearly two years** [\[10.1002/asi.23561\]](https://doi.org/10.1002/asi.23561)



# Behavioural Microtargeting

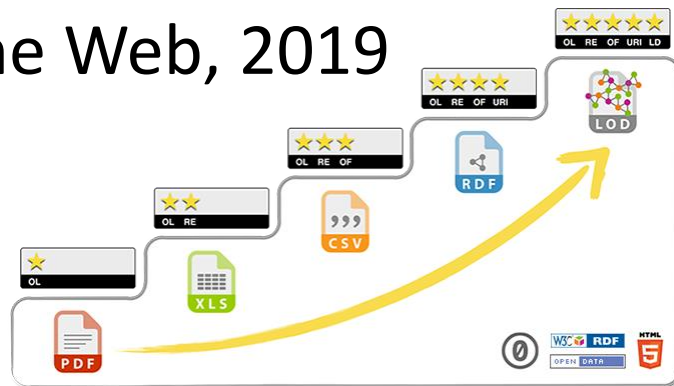


Monday, 03 February 2020



# Recent Initiatives by Tim Berners-Lee

- 5 ★ Open Data, 2012
- ‘Magna Carta’ for the Web, 2014
- Solid (web decentralization project), 2016
- Contract for the Web, 2019



Part of the British Library's 'My Digital Rights' project to celebrate 800 years of Magna Carta. Created by young people and voted for by the public.

**LIBRARY**  
**HSJLIRB**

### THE WEB WE WANT WILL...

- 1 not let companies pay to control it, and not let governments restrict our right to information.
- 2 allow freedom of speech.
- 3 be free from government censors in all countries.
- 4 not allow any kind of government censorship.
- 5 be available for all those who wish to use it.
- 6 be free from censorship and mass surveillance.
- 7 allow equal access to knowledge, information and current news worldwide.
- 8 have freedom of speech.
- 9 not be censored by the government.
- 10 not sell our personal information and preferences for money, and will make it clearer if the company /Website intends to do so.

Magna Carta for the digital age – 15 June 2015

# Recent Research Artefacts

A word cloud of research artefacts and their associated features. The artefacts are: Dokieli, CIMBA, MakingSense, SOLID, and Musubi. The features are: MicroBlogging, Mobile, Social, Linked Data, Summarization, Editor, Network, Integrated, Tagging, Client-Side, and Collaborative.

Dokieli  
MicroBlogging  
Mobile  
Social Linked Data  
Summarization  
Editor  
Network  
Integrated  
CIMBA  
MakingSense  
Tagging  
Client-Side  
Client  
SOLID  
Collaborative  
Musubi

*“If we leave the web as it is, there’s a very large number of things that will go wrong. We could end up with a **digital dystopia** if we don’t turn things around. It’s not that we need a 10-year plan for the web, **we need to turn the web around now.**”*

- Tim Berners-Lee @ launch of [“Contract for the Web”](#)

Can we solve the 'nasty storm' of  
issues with Web, using the  
*wisdom of the crowd?*

...while not relying on developers and content providers...

# Annotation

“... is a note added to a book, drawing or any other kind of text as a comment or explanation.” [\[NYT, 2015\]](#)

*Web Annotations have emerged as a First-Class Object.*  
[\[10.1109/MIC.2013.123\]](#)

Web annotation tools are gaining tremendous interest among academicians [\[10.1038/528153a\]](#), [\[10.1038/d41586-019-01427-9\]](#)

**Circle** important words.  
Add a synonym or 2-3 word explanation in the margin.

**!\* Mark new and/or big ideas.**  
Summarize the idea in 2-3 words within the margin.

Draw arrows **< >** to show related ideas. | Label the connection in 2-3 words noted in the margin.

Number **1.** steps  
**2.** lists  
**3.** details | Note in 2-3 words what all the numbers represent.

**LOL** Mark humorous ideas.

**?** Jot questions and confusions in the margin.

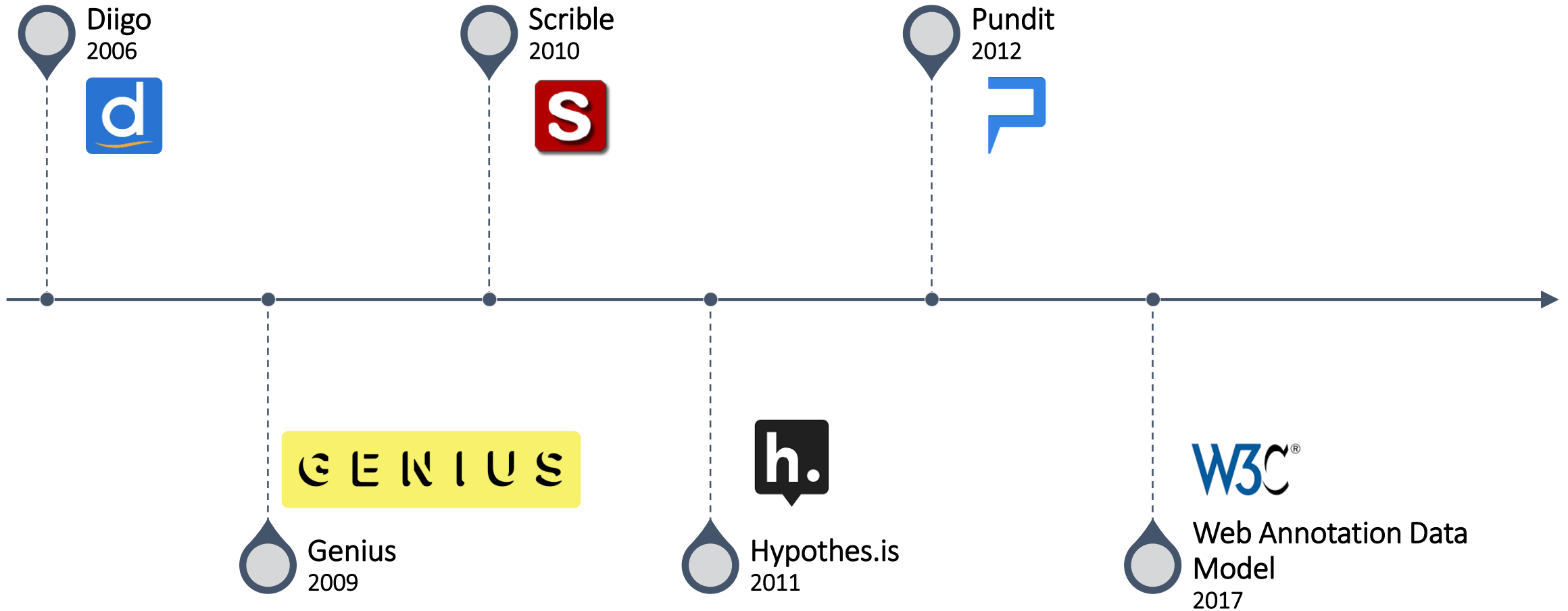
**+** Identify additional information learned about a previously-mentioned idea.

**↔** Mark ideas/opinions that contradict what was previously read or learned.  
Summarize the alternative viewpoint within the margin.

Image source: [Smekenseducation.com](#)



# Popular Web Annotation Systems



# Hypothes.is

- free, open, non-profit, neutral, 100% community moderated, merit based, pseudonymous, and more...
- aims “to enable a conversation over the world’s knowledge”
- It’s **215,000 users** have added more than **5 million comments** on scholarly sites [[10.1038/d41586-019-01427-9](https://doi.org/10.1038/d41586-019-01427-9)]

## COMMENT COUNTS

Users of the software tool Hypothesis have collectively posted more than 5 million annotations since 2015.

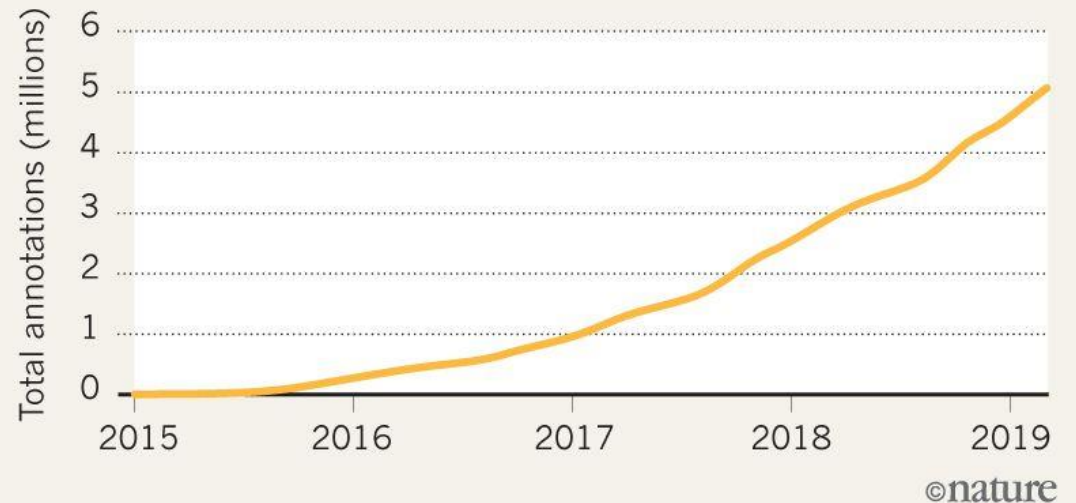


Image Source: [Nature](#)

# Before Hypothes.is' Fuzzy Anchoring

- XPath (XML Path Language) [e.g. `/html/body/div[3]/div[3]/div[4]/div/p[2]/b[3]`] Matching

The image shows a browser window displaying a Wikipedia article about the World Wide Web. On the left, a DOM tree diagram illustrates the document's structure, starting from the root element `<html>` and branching into `<head>` and `<body>`. The `<body>` contains a `<title>` element with the text "My title", a `<h1>` element with the text "A heading", and an `<a>` element with the text "Link text" and an attribute `href`. The browser's developer tools are open, showing the HTML source code with the following snippet highlighted:

```
<b>World Wide Web</b>
(
<b>WWW</b>
), also called
<b>the Web</b>
, is an
<a href="/wiki/Information_space" title="Information space">information space</a>
where documents and other
<a href="/wiki/Web_resource" title="Web resource">web resources</a>
```

On the right, a global map of the web index is shown, with a legend indicating the index values for various countries. The map is color-coded, with red representing the highest index values and blue representing the lowest.

# After Hypothes.is' Fuzzy Anchoring [2013]

- Robustly anchoring annotations using keywords [Brush et al. [2001](#) *Microsoft Research*]
  - Robust anchoring of annotations to content [Brush et al. [2010](#) *Patent*]
  - uses a modified version of Google's [diff-match-patch](#)
  - Bitap matching [[10.1145/135239.135244](#)] *for text matching*
  - Myers diff [[10.1007/BF01840446](#)] *for text comparison*
- } Levenshtein distance [[mathnet.ru/dan31411](#)]

# How does Fuzzy Anchoring work?

- Selectors

- *RangeSelector*
- *TextPositionSelector*
- *TextQuoteSelector*

- Strategies

- *From Range Selector*
- *From Position Selector*
- *Context-first Fuzzy Matching*
- *Selector-only Fuzzy Matching*

# How does Fuzzy Anchoring work? *(example)*

“... new Lecture Hall Complex (Neues Institutgebäude, NIG), the lecture hall complex Althanstraße (UZA), the campus on the premises of the [Historical General Hospital of Vienna](#), the Faculty of Law (Juridicum) and others. The [Botanical Garden of the University of Vienna](#) is housed in the Third District, as are the Department of Biochemistry and related research centres...”

- [Wikipedia - University of Vienna](#)

**RangeSelector:** `//*[@id="mw-content-text"]/div/p[9]`

**TextPositionSelector:** String offsets (i.e., position) of first and last character in the selected text (with respect to the whole document)

**TextQuoteSelector:** `exact`, `prefix` and `suffix`

# What's wrong with Fuzzy Anchoring?

- In 2015, Aturban et al. analyzed **6281** highlighted text **annotations** from Hypothes.is [[10.1007/978-3-319-24592-8\\_2](https://doi.org/10.1007/978-3-319-24592-8_2)]
- **27%** annotations were completely **orphaned**
- only **3.5 %** of orphans **could be reattached** using public web archives
- ...and **61%** were at **risk of being orphaned** due page decay



# Our Goal

- Design and evaluate a web-based Crowdsourcing Information System (CIS)
  - that acts as conversation layer over the Web
  - is interoperable
  - supports activities on-the-fly
  - provides a social environment that promotes co-creation
  - provides a stable and robust approach for tracking textual contextual
  - is based on the principles for Collective Intelligence

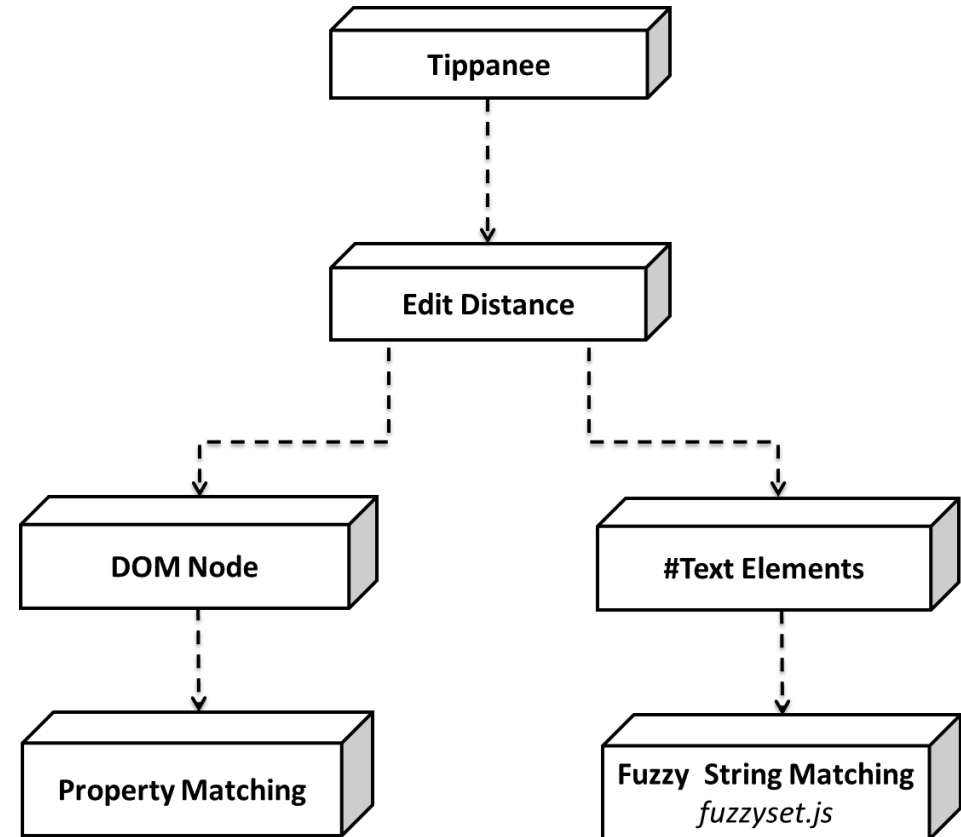
# Proposed Anchoring Approach

- Selectors

- *TextSelector*
- *DOMSelector* (in prefix order)

- Strategies

- Edit (i.e., Levenshtein) Distance
  - *Fuzzy String Matching*
  - *DOM Property Matching*



# Edit (i.e., Levenshtein) distance

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

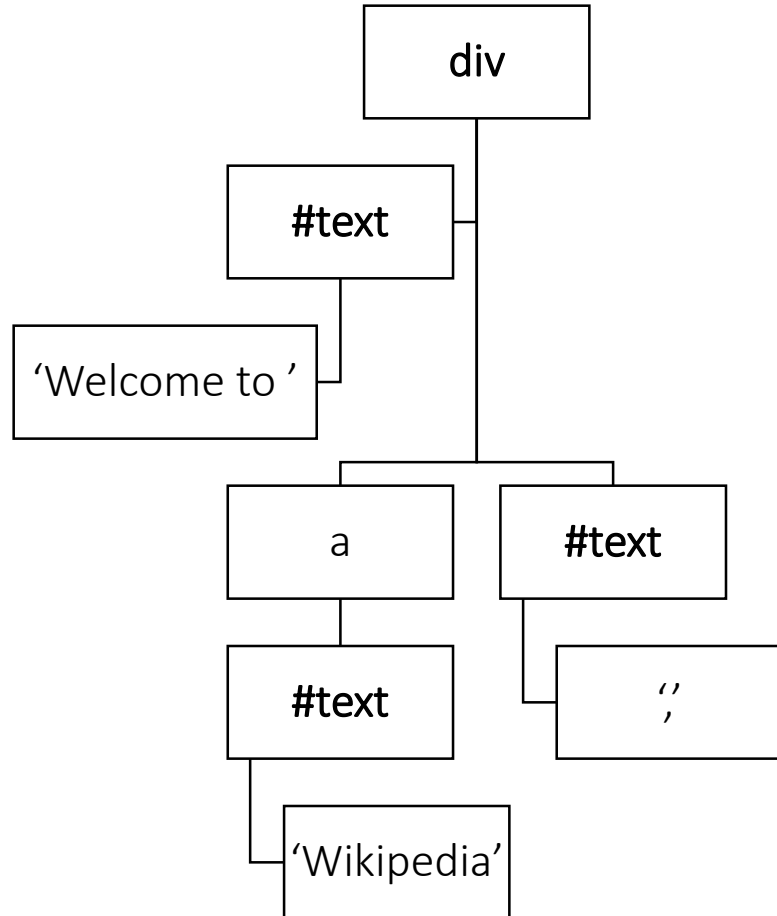
		S	A	T	U	R	D	A	Y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
U	2	1	1	2	2	3	4	5	6
N	3	2	2	2	3	3	4	5	6
D	4	3	3	3	3	4	3	4	5
A	5	4	3	3	4	4	4	3	4
Y	6	5	4	4	5	5	5	4	3

S A T U R D A Y

|    add    add    |    replace    |    |    |

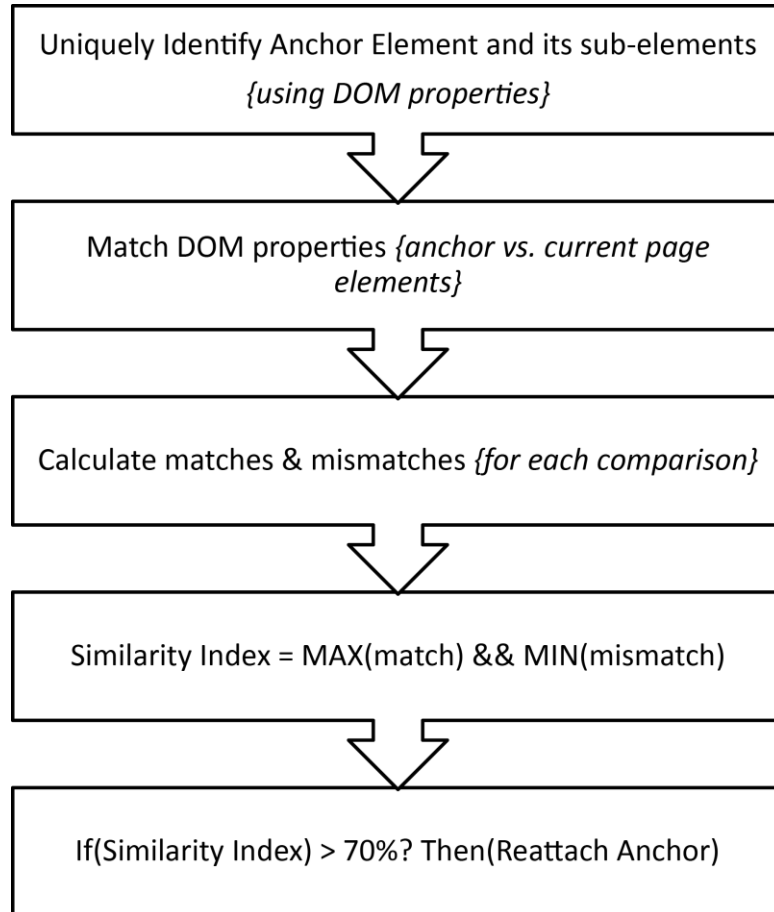
S    \_    \_    U    N    D    A    Y

# Anchors



```
"ww-1540806054738-a2ca6765" : {  
  "addedon" : "October 29, 2018 11:40 AM",  
  "anchor" : [ {  
    "nodeDepth" : 0,  
    "nodeName" : "DIV"  
  }, {  
    "annotated" : true,  
    "endOffset" : 3,  
    "nodeDepth" : 1,  
    "nodeName" : "#text",  
    "nodeValue" : "Welcome to ",  
    "startOffset" : 0  
  }, {  
    "href" : "https://en.wikipedia.org/wiki/Wikipedia",  
    "nodeDepth" : 1,  
    "nodeName" : "A"  
  }, {  
    "nodeDepth" : 2,  
    "nodeName" : "#text",  
    "nodeValue" : "Wikipedia"  
  }, {  
    "nodeDepth" : 1,  
    "nodeName" : "#text",  
    "nodeValue" : ","  
  } ],  
  "owner" : "abc@abc.com",  
  "selectedtext" : "Welcome ",  
  "sharedwith" : "LSS",  
  "transclusion" : "ww-1540823638993-7270b4f1",  
  "urlHost" : "en.wikipedia.org",  
  "urlParameter" : "",  
  "urlPathname" : "/wiki/Main_Page",  
  "urlProtocol" : "https:"  
}
```

# Similarity Index

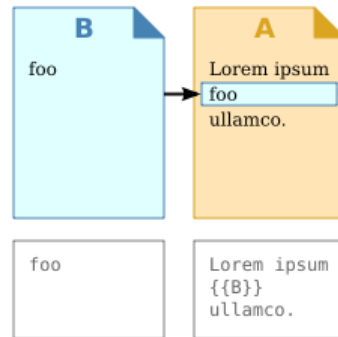


```
{  
  "strIdxMAXmat": 6,  
  "strMat": [  
    4.98,  
    4.98,  
    4.98,  
    4.98,  
    4.98,  
    4.98,  
    4.98  
  ],  
  "strMis": [  
    75.89999999999999,  
    48.940000000000005,  
    9.959999999999999,  
    9.51,  
    9.06,  
    4.53,  
    0  
  ],  
  "strSimIdx": 0.9960000000000001  
}
```

# Advantages over Fuzzy Anchoring

- *new* **robust** anchoring approach
  - resilient to content or structure change
- preserves both the annotated content and it's surrounding content

- enables transclusions



- support knowledge/information exchange by enabling “web of annotations”



# Tippane Chrome Extension

... as well as demonstration of innovative syst...  
... sessions at the CAiSE conference will f...  
discussion, and exchange of ideas among presenters and participar...  
CAiSE'18 Forum are welcome to address any of the CAiSE'18 confere...  
particular the theme of this year's conference: *Information Systems i*

Important Dates	
Paper submission deadline:	4 March 2018
Notification of acceptance:	6 April 2018
Camera-ready deadline:	13 April 2018

Notes

April 13, 2018 3:30 PM  
Paper submission deadline: 4 March 2018  
Strict Deadline!!!  
Apr 13, 2018 3:35 PM

Reconstruct Anchor

Link Annotations

Add semantic description





# Similarity Index

Main Page [Talk](#) [Read](#) [View source](#)

Welcome to **Wikipedia**,  
the free encyclopedia that anyone can edit.  
5,569,971 articles in English

From today's featured article

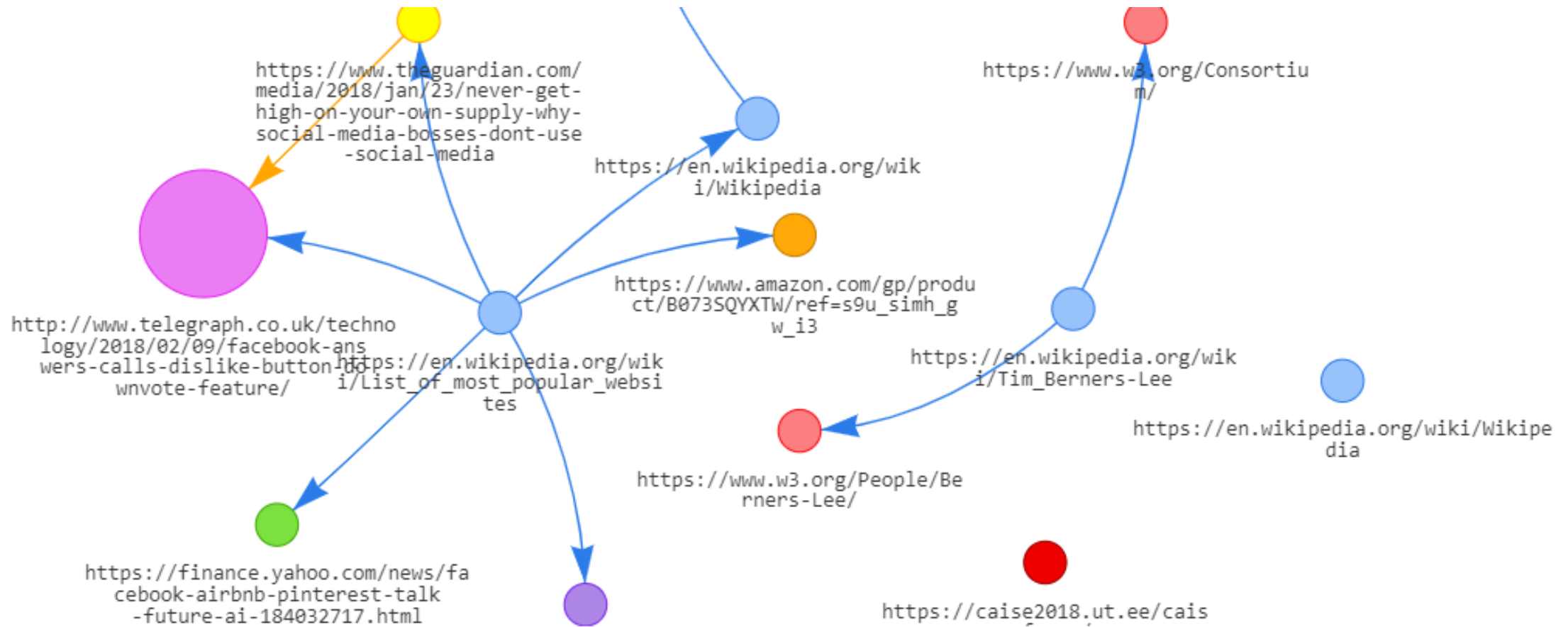
 The **Indian**  
two coin  
series struck

February 10, 2018 9:11 PM  
5,568,753 articles in English  
Add a note...

Similarity Index: 0.9266666666666666

Similarity Index: 0.9266666666666666

# Web of Annotations



Hypothes.is

Victor\_013 (Public) Sep 13

5,710,618

Wikipedia article count (numbers only)

Show replies (1)

Annotations on page: 3, 2

Page content: Welcome to Wikipedia, the free encyclopedia that anyone can edit. 5,745,710 articles in English. Today's featured article: The Eurasian tree sparrow (*Passer*).

vs.

Tippanee

5,714,536

Add a note...

Similarity Index: 0.89

Annotations on page: 1

Page content: encyclopedia that anyone can edit. 5,745,730 articles in English. Today's featured article: The Eurasian tree sparrow (*Passer*).

# Preliminary Evaluation

- Experiment 1:
  - replicated *735 (Hypothes.is) annotations from more 650 different websites*
  - observed annotations over 3 months (*expecting some web page decay*)
  - **91.41%** annotations were successfully attached
  - **12.41%** over Hypothes.is' [79%](#) expected success
- Experiment 2:
  - presented the tool to 25 candidates
    - found the tool useful and easy to use
    - users preferred the tool for social interactions, expression of opinion and information sharing
    - helped identify bugs and suggested additional UI features

# Tippaneer's Features

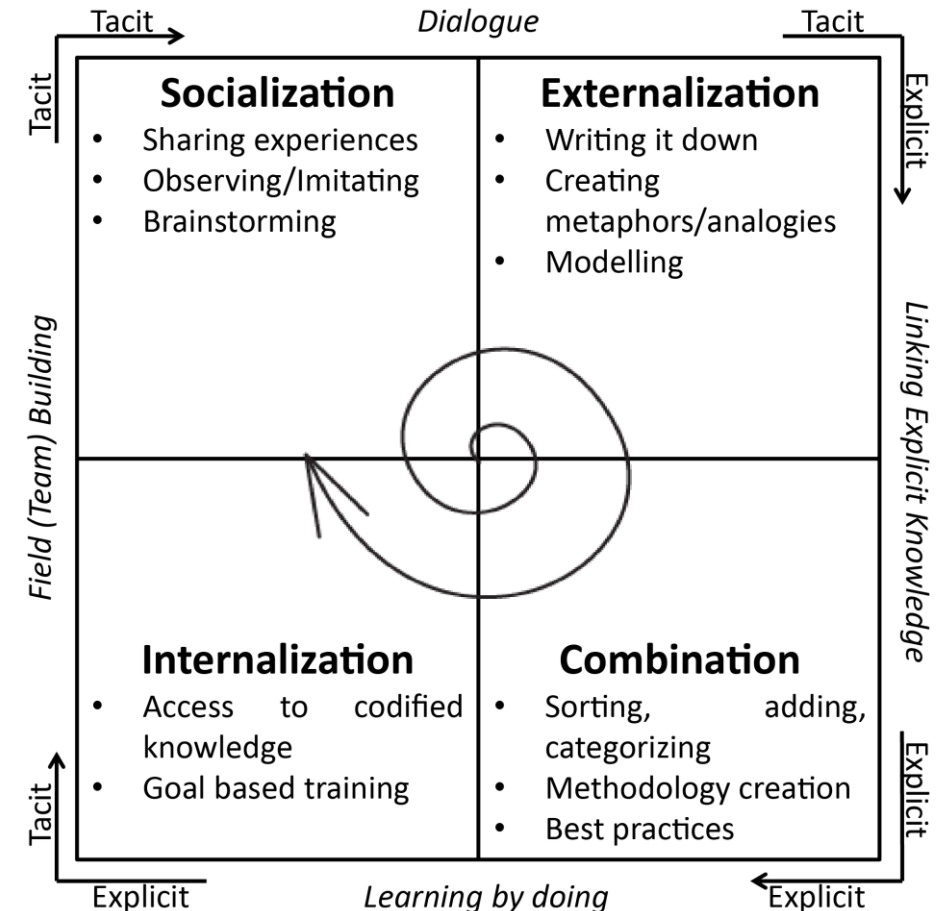
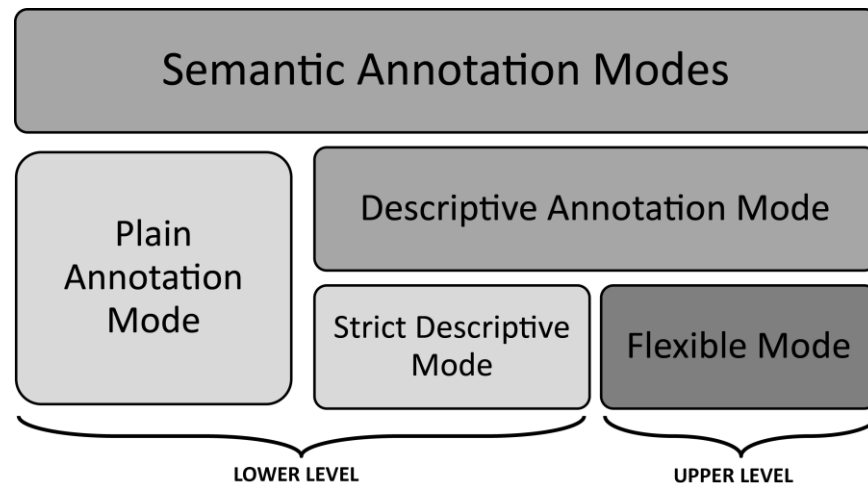
- Novel anchoring approach
  - stable and robust
  - works both online-offline
- End-user oriented features
  - data critiquing and content quality monitoring
  - personalized archival of textual content
  - social knowledge management
    - Linking and visualizing annotated content (*i.e.*, [knowledge graph](#))
  - enriching web content with semantic metadata
    - allows for creation of new semantic vocabularies\* work in progress

# Some More Motivation (but from Organizations)

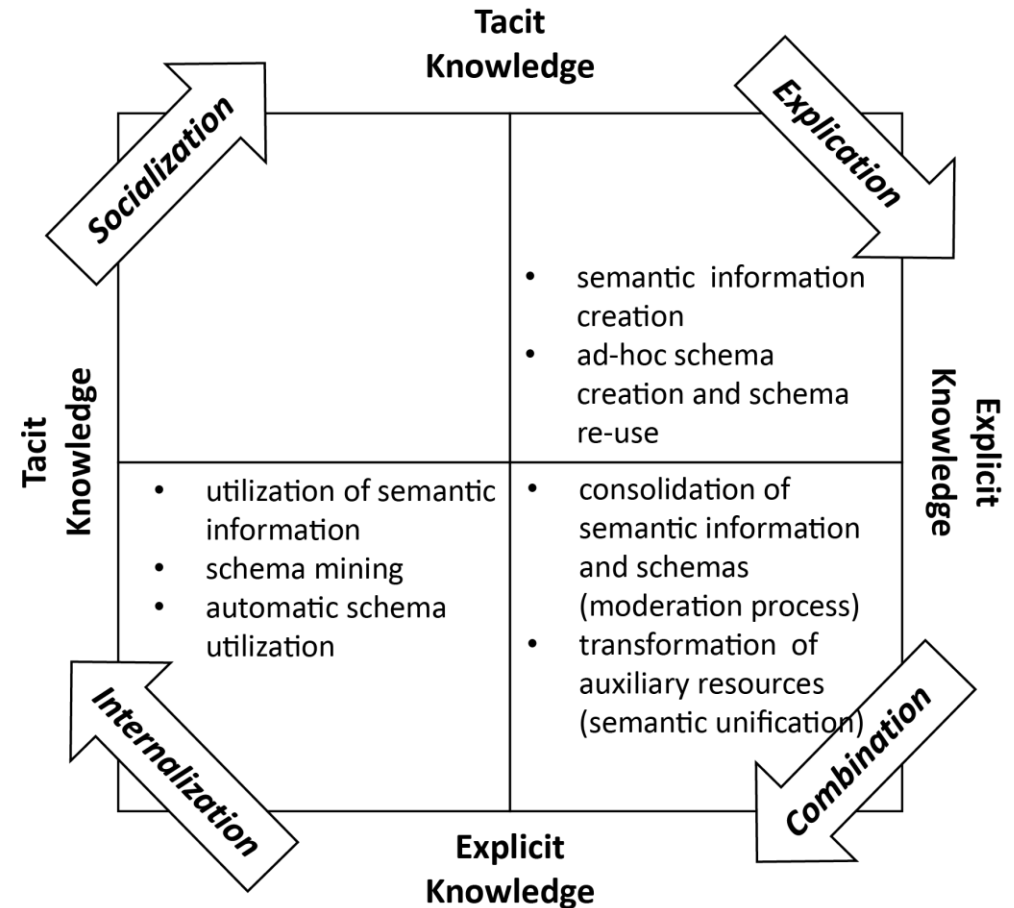
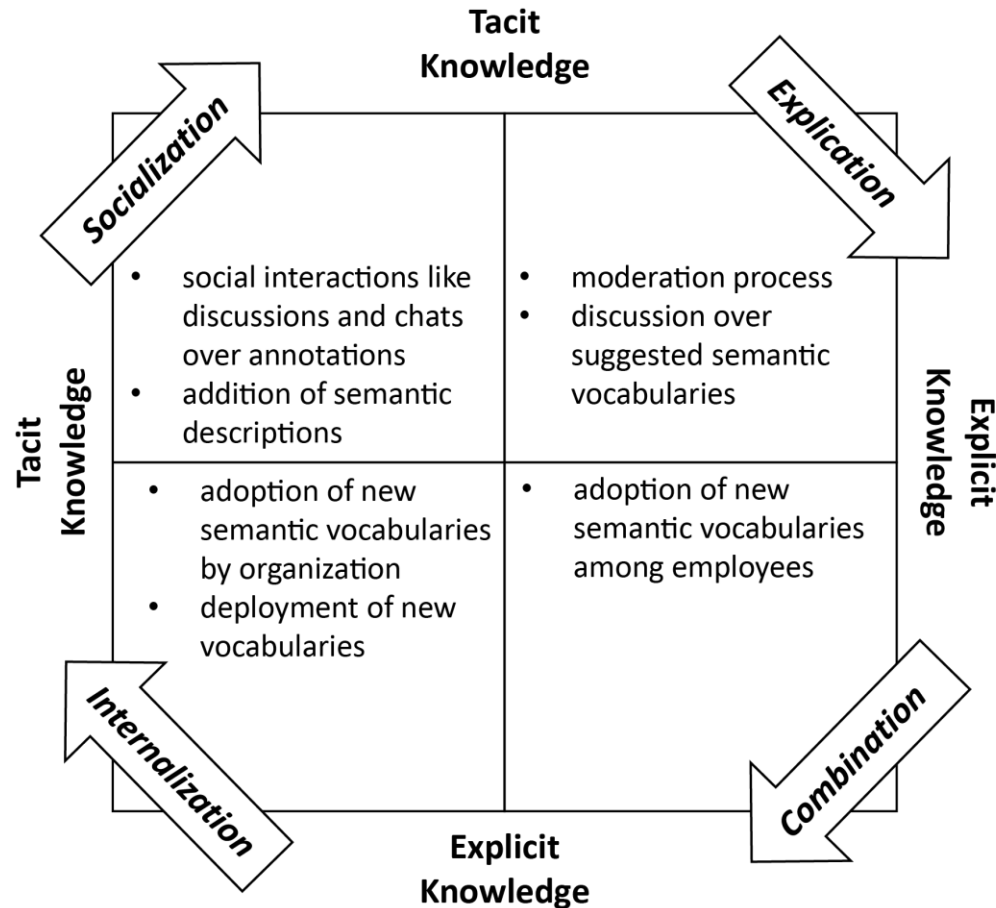
- Knowledge Management in organizations is a challenging task  
[\[10.1080/23311975.2015.1127744\]](https://doi.org/10.1080/23311975.2015.1127744)
  - heterogeneous environments
  - lack of knowledge sharing
  - tacit knowledge transfer
  
- ... especially in today's Social Media Landscape

# SECI through Web Annotations

- based on “Nonaka’s Knowledge Spiral”
- for “Knowledge Creating Companies”

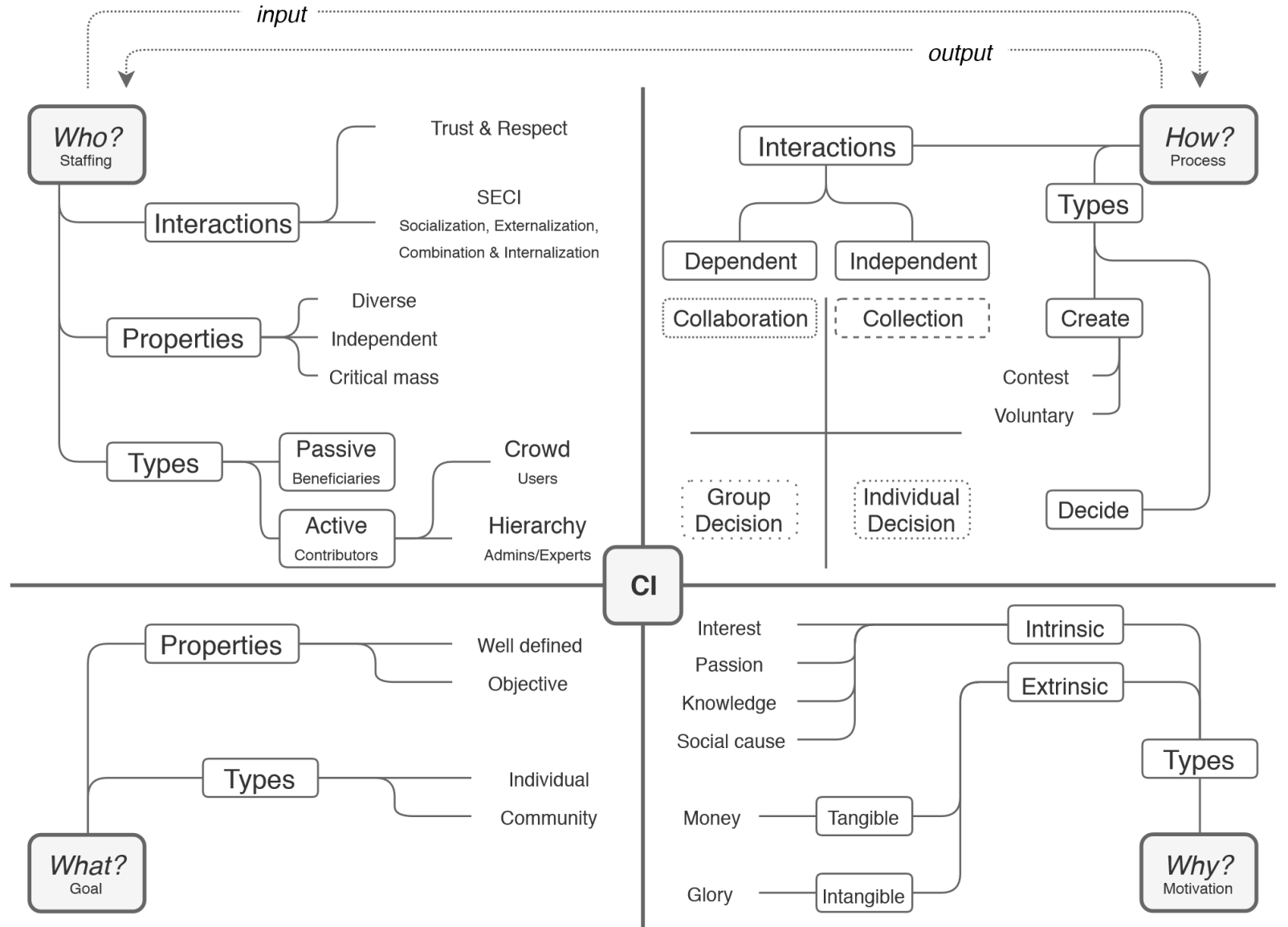


# Lower (left) and Higher (right) Level Annotation Activities





# 'Generic' framework for CI Systems



# Other Ongoing Work

- Anchoring approach test bench:
  - 50 websites
  - 120 webpages
  - 9 annotations per page
  - 96 variants per annotation
  - *103,680 data points for evaluation*
- Implement & evaluation of “SECI through Web Annotations”
- Develop a novel user reputation model *[less prone to bias]*

