COMP7630 – Web Intelligence and its Applications

Web Intelligence in a Nutshell

Valentino Santucci

(valentino.santucci@unistrapg.it)

Outline

- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

WEB … connected world-wide!

Why analyzing the Web?

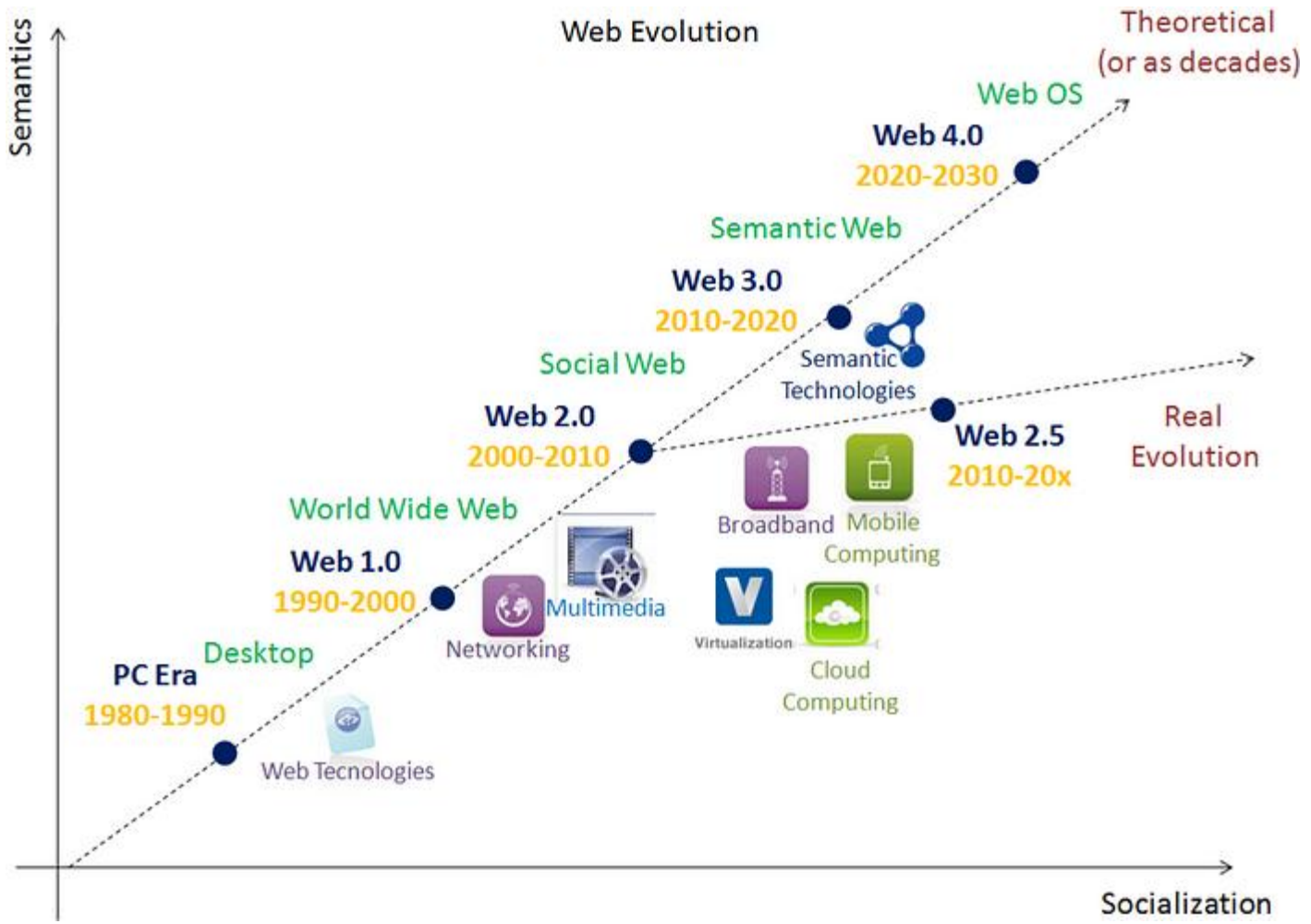
- Web = a huge pool of hyperlinked information (href) in hypertext format (HTML) sitting on the Internet (http over servers and clients) contributed/shared by different individuals and organizations.

- Web as a "place" for commerce (eCommerce)
- Web as a "media" for political campaigns
- Web as a survey and opinion collecting "tool"
- …

Technologies/ideas making possible the Web

- Internet
- TCP/IP
- Client-Server architecture
- Browser
- Hypertext
- Hyperlinks
- Hypermedia
- HTTP
- HTML
- URL
- Search engines (e.g. Google)
- …

Continuous and end-less evolution of the Web

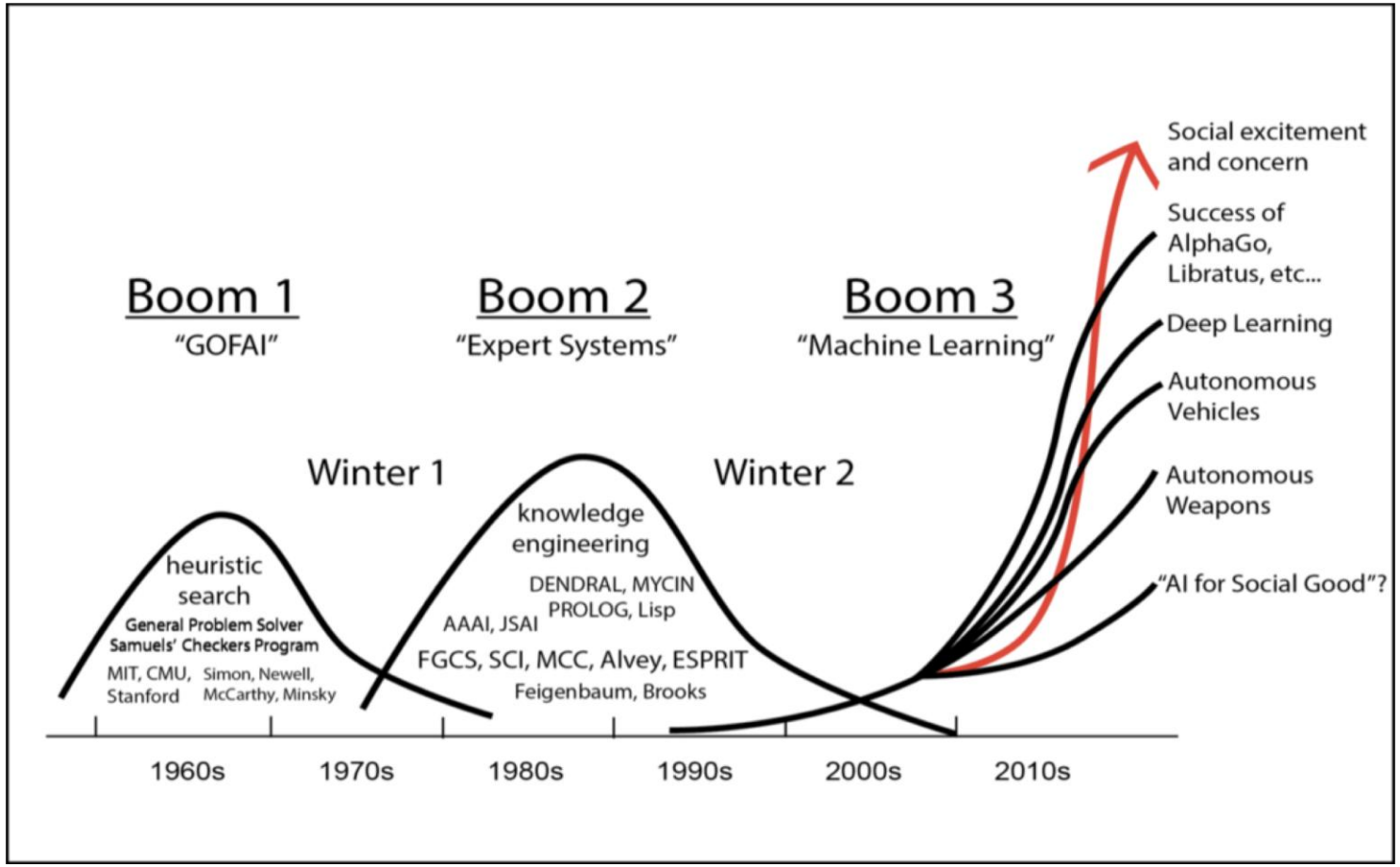


Web Intelligence

=

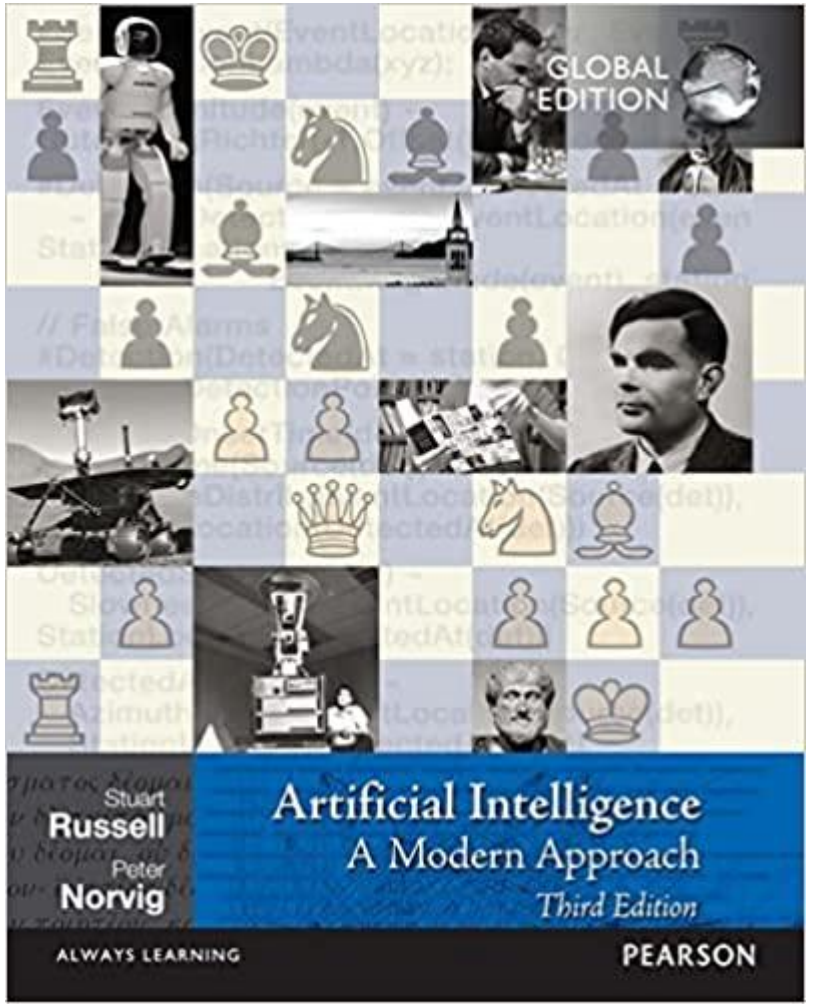
Web + Artificial Intelligence

AI? … summer, winter, summer, winter, Summer!



AI means …

- Knowledge representation and reasoning
- Automated Planning and Scheduling
- Constraint Programming
- Machine Learning, Data Mining, Knowledge Discovery
- Multi-agent Systems
- Heuristics and Meta-heuristics
- Evolutionary Computation and Artificial Life
- Computer Vision
- Natural Language Processing
- …



Web Intelligence

- http://wi-consortium.org/

[…] scientific research and development to explore the fundamental roles as well as practical impacts of Artificial Intelligence (AI) (e.g., knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence) and advanced Information Technology (IT) (e.g., wireless networks, ubiquitous devices, social networks, wisdom Web, and data/knowledge grids) on the next generation of Web-empowered products, systems, services, and activities. It is one of the most important as well as promising IT research fields in the era of Web and agent intelligence.

Outline

- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

The Web is Distributed

- Web and Internet allow for distributed computing
- Cooperative Distributed Problem Solving (CDPS) is a very important algorithmic topic in distributed computing
- Multi-agent systems and Evolutionary Computing systems are two prominent examples of CDPS

Evolutionary Computation (EC)

- EC studies Evolutionary and Swarm Intelligence-based Algorithms
- EC algorithms can be applied to virtually any optimisation problem
- EC algorithms are made up by a population of computational entities:
  - Each entity maintains a solution to the problem at hand
  - Each entity iteratively update its solution
  - Entities may cooperate and communicate among them
  - Altogether the entities implement a distributed problem solving approach

How many Evolutionary Algorithms are there!?

Outline

- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

Characteristics of Web Data

- Amount of information on the Web is huge and growing
- Content of heterogeneous types exist on the Web:
  - structured tables
  - semi-structured pages (e.g., HTML, XML, JSON)
  - unstructured texts
  - multimedia files (images, audios, and videos)
- Even for the same type, information on the Web is heterogeneous due to diverse authorships
  - Using different wordings and formats
  - Wide vs long tables
  - …

Subject	Time1	Time2	Time3
A	5	3	4
B	2	6	8
C	7	5	1

Subject	Time	Value
A	1	5
A	2	3
A	3	4
B	1	2
B	2	6
B	3	8

Characteristics of Web Data

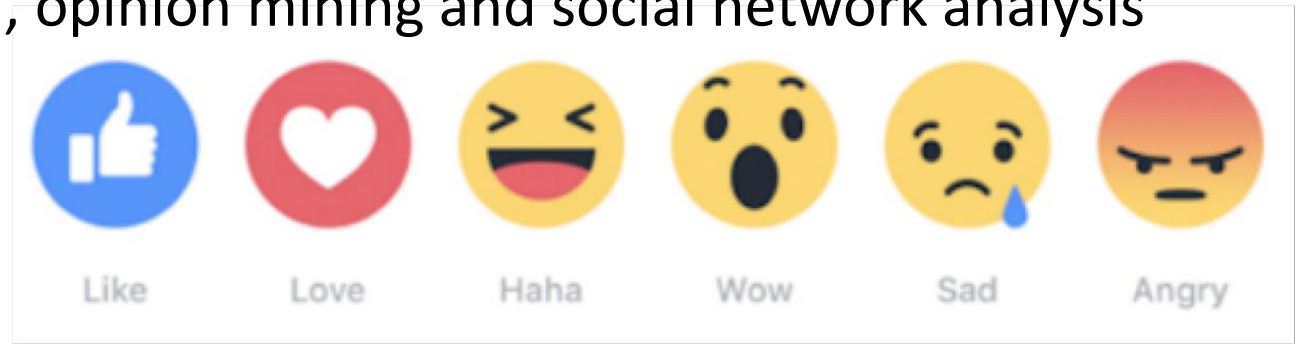
- Hyperlinks exist among Web pages
  - Within a site (for information organization)
  - Across different sites (indicating authority)

- The information on the Web is NOISY
  - Only part of a page is useful!
  - Navigation links, advertisements, copyright notices, privacy policies – useful?

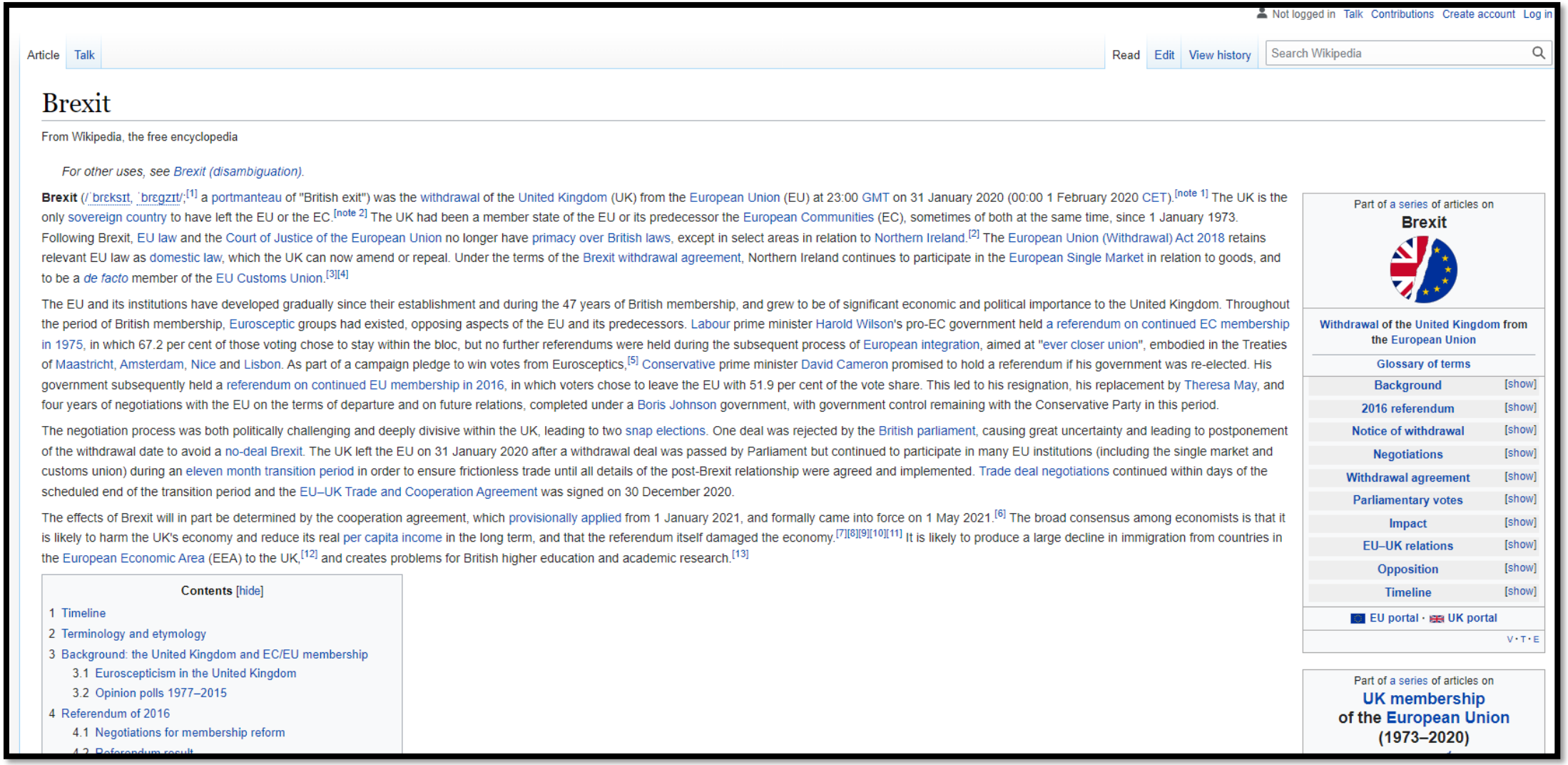
- Web is of low quality, erroneous, or even misleading
  - Anyone can write anything!

Characteristics of Web Data

- The Web supports e-commerce
  - People click through to browse, purchase, pay, …
  - Automated Web services (APIs) are needed (let think to Paypal or similars)
- The Web is a virtual society
  - People can communicate anywhere in the world
  - Express views and opinions on anything in Internet forums, blogs, review sites and social network sites
  - New new mining tasks, e.g., opinion mining and social network analysis



Web 2.0: Example - Wikipedia



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

Brexit

From Wikipedia, the free encyclopedia

For other uses, see Brexit (disambiguation).

Brexit (/ˈbrɛksɪt, ˈbrɛgzɪt/;[1] a portmanteau of "British exit") was the withdrawal of the United Kingdom (UK) from the European Union (EU) at 23:00 GMT on 31 January 2020 (00:00 1 February 2020 CET).[note 1] The UK is the only sovereign country to have left the EU or the EC.[note 2] The UK had been a member state of the EU or its predecessor the European Communities (EC), sometimes of both at the same time, since 1 January 1973. Following Brexit, EU law and the Court of Justice of the European Union no longer have primacy over British laws, except in select areas in relation to Northern Ireland.[2] The European Union (Withdrawal) Act 2018 retains relevant EU law as domestic law, which the UK can now amend or repeal. Under the terms of the Brexit withdrawal agreement, Northern Ireland continues to participate in the European Single Market in relation to goods, and to be a *de facto* member of the EU Customs Union.[3][4]

The EU and its institutions have developed gradually since their establishment and during the 47 years of British membership, and grew to be of significant economic and political importance to the United Kingdom. Throughout the period of British membership, Eurosceptic groups had existed, opposing aspects of the EU and its predecessors. Labour prime minister Harold Wilson's pro-EC government held a referendum on continued EC membership in 1975, in which 67.2 per cent of those voting chose to stay within the bloc, but no further referendums were held during the subsequent process of European integration, aimed at "ever closer union", embodied in the Treaties of Maastricht, Amsterdam, Nice and Lisbon. As part of a campaign pledge to win votes from Eurosceptics,[5] Conservative prime minister David Cameron promised to hold a referendum if his government was re-elected. His government subsequently held a referendum on continued EU membership in 2016, in which voters chose to leave the EU with 51.9 per cent of the vote share. This led to his resignation, his replacement by Theresa May, and four years of negotiations with the EU on the terms of departure and on future relations, completed under a Boris Johnson government, with government control remaining with the Conservative Party in this period.

The negotiation process was both politically challenging and deeply divisive within the UK, leading to two snap elections. One deal was rejected by the British parliament, causing great uncertainty and leading to postponement of the withdrawal date to avoid a no-deal Brexit. The UK left the EU on 31 January 2020 after a withdrawal deal was passed by Parliament but continued to participate in many EU institutions (including the single market and customs union) during an eleven month transition period in order to ensure frictionless trade until all details of the post-Brexit relationship were agreed and implemented. Trade deal negotiations continued within days of the scheduled end of the transition period and the EU–UK Trade and Cooperation Agreement was signed on 30 December 2020.

The effects of Brexit will in part be determined by the cooperation agreement, which provisionally applied from 1 January 2021, and formally came into force on 1 May 2021.[6] The broad consensus among economists is that it is likely to harm the UK's economy and reduce its real per capita income in the long term, and that the referendum itself damaged the economy.[7][8][9][10][11] It is likely to produce a large decline in immigration from countries in the European Economic Area (EEA) to the UK,[12] and creates problems for British higher education and academic research.[13]

Contents [hide]

1 Timeline
2 Terminology and etymology
3 Background: the United Kingdom and EC/EU membership
  3.1 Euroscepticism in the United Kingdom
  3.2 Opinion polls 1977–2015
4 Referendum of 2016
  4.1 Negotiations for membership reform
  4.2 Referendum result

Part of a series of articles on
Brexit

Withdrawal of the United Kingdom from the European Union

Glossary of terms

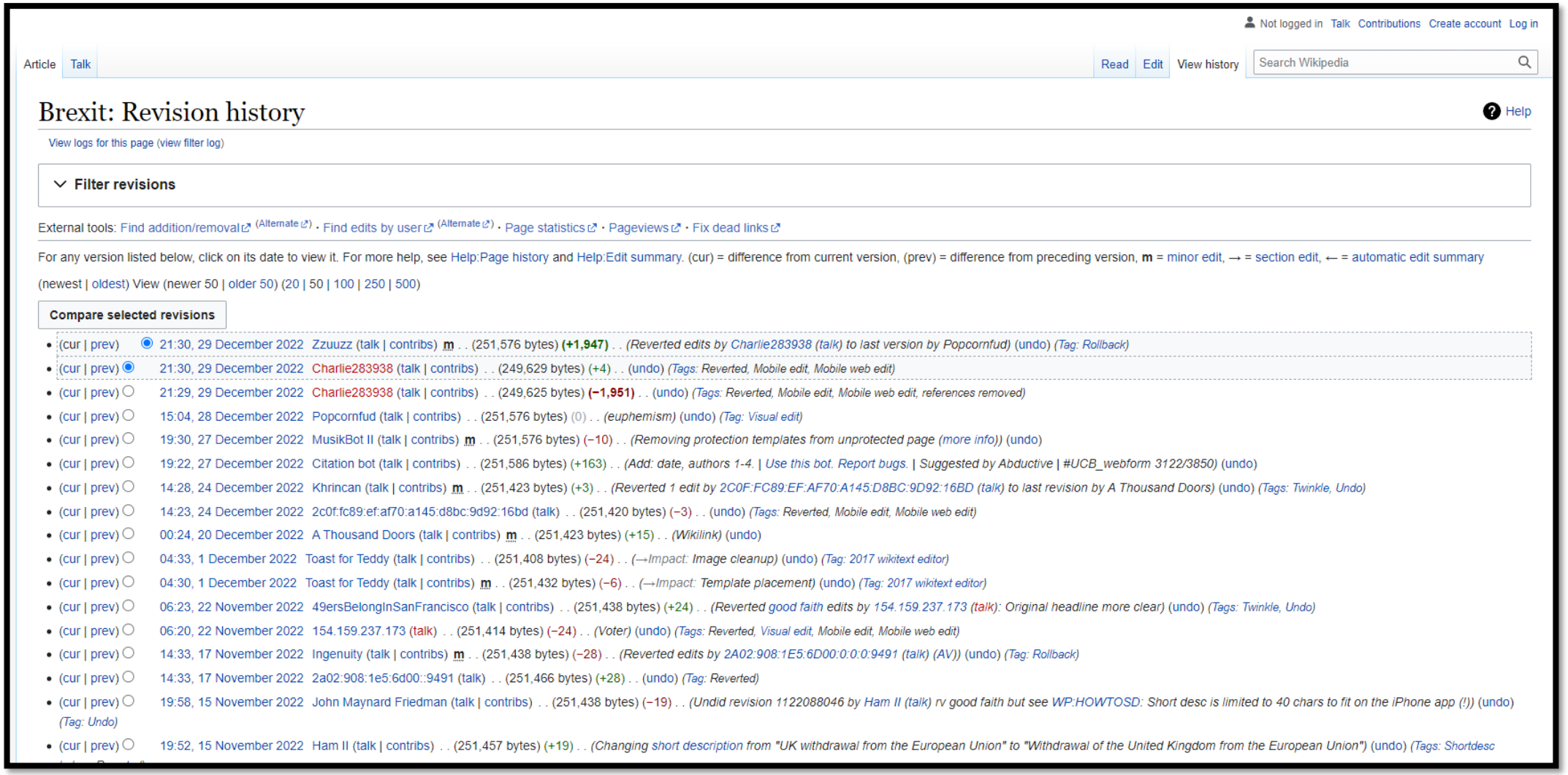
Background	[show]
2016 referendum	[show]
Notice of withdrawal	[show]
Negotiations	[show]
Withdrawal agreement	[show]
Parliamentary votes	[show]
Impact	[show]
EU–UK relations	[show]
Opposition	[show]
Timeline	[show]

EU portal · UK portal

V · T · E

Part of a series of articles on
UK membership of the European Union (1973–2020)

Web 2.0: Example - Wikipedia



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Brexit: Revision history

Help

View logs for this page (view filter log)

Filter revisions

External tools: Find addition/removal (Alternate) · Find edits by user (Alternate) · Page statistics · Pageviews · Fix dead links

For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary. (cur) = difference from current version, (prev) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary

(newest | oldest) View (newer 50 | older 50) (20 | 50 | 100 | 250 | 500)

Compare selected revisions

- (cur | prev) 21:30, 29 December 2022 Zzuuzz (talk | contribs) **m** . . (251,576 bytes) **(+1,947)** . . *(Reverted edits by Charlie283938 (talk) to last version by Popcornfud)* (undo) *(Tag: Rollback)*
- (cur | prev) 21:30, 29 December 2022 Charlie283938 (talk | contribs) . . (249,629 bytes) (+4) . . (undo) *(Tags: Reverted, Mobile edit, Mobile web edit)*
- (cur | prev) 21:29, 29 December 2022 Charlie283938 (talk | contribs) . . (249,625 bytes) **(−1,951)** . . (undo) *(Tags: Reverted, Mobile edit, Mobile web edit, references removed)*
- (cur | prev) 15:04, 28 December 2022 Popcornfud (talk | contribs) . . (251,576 bytes) (0) . . *(euphemism)* (undo) *(Tag: Visual edit)*
- (cur | prev) 19:30, 27 December 2022 MusikBot II (talk | contribs) **m** . . (251,576 bytes) (−10) . . *(Removing protection templates from unprotected page (more info))* (undo)
- (cur | prev) 19:22, 27 December 2022 Citation bot (talk | contribs) . . (251,586 bytes) (+163) . . *(Add: date, authors 1-4. | Use this bot. Report bugs. | Suggested by Abductive | #UCB_webform 3122/3850)* (undo)
- (cur | prev) 14:28, 24 December 2022 Khrincan (talk | contribs) **m** . . (251,423 bytes) (+3) . . *(Reverted 1 edit by 2C0F:FC89:EF:AF70:A145:D8BC:9D92:16BD (talk) to last revision by A Thousand Doors)* (undo) *(Tags: Twinkle, Undo)*
- (cur | prev) 14:23, 24 December 2022 2c0f:fc89:ef:af70:a145:d8bc:9d92:16bd (talk) . . (251,420 bytes) (−3) . . (undo) *(Tags: Reverted, Mobile edit, Mobile web edit)*
- (cur | prev) 00:24, 20 December 2022 A Thousand Doors (talk | contribs) **m** . . (251,423 bytes) (+15) . . *(Wikilink)* (undo)
- (cur | prev) 04:33, 1 December 2022 Toast for Teddy (talk | contribs) . . (251,408 bytes) (−24) . . *(→Impact: Image cleanup)* (undo) *(Tag: 2017 wikitext editor)*
- (cur | prev) 04:30, 1 December 2022 Toast for Teddy (talk | contribs) **m** . . (251,432 bytes) (−6) . . *(→Impact: Template placement)* (undo) *(Tag: 2017 wikitext editor)*
- (cur | prev) 06:23, 22 November 2022 49ersBelongInSanFrancisco (talk | contribs) . . (251,438 bytes) (+24) . . *(Reverted good faith edits by 154.159.237.173 (talk): Original headline more clear)* (undo) *(Tags: Twinkle, Undo)*
- (cur | prev) 06:20, 22 November 2022 154.159.237.173 (talk) . . (251,414 bytes) (−24) . . *(Voter)* (undo) *(Tags: Reverted, Visual edit, Mobile edit, Mobile web edit)*
- (cur | prev) 14:33, 17 November 2022 Ingenuity (talk | contribs) **m** . . (251,438 bytes) (−28) . . *(Reverted edits by 2A02:908:1E5:6D00:0:0:0:9491 (talk) (AV))* (undo) *(Tag: Rollback)*
- (cur | prev) 14:33, 17 November 2022 2a02:908:1e5:6d00::9491 (talk) . . (251,466 bytes) (+28) . . (undo) *(Tag: Reverted)*
- (cur | prev) 19:58, 15 November 2022 John Maynard Friedman (talk | contribs) . . (251,438 bytes) (−19) . . *(Undid revision 1122088046 by Ham II (talk) rv good faith but see WP:HOWTOSD: Short desc is limited to 40 chars to fit on the iPhone app (!))* (undo) *(Tag: Undo)*
- (cur | prev) 19:52, 15 November 2022 Ham II (talk | contribs) . . (251,457 bytes) (+19) . . *(Changing short description from "UK withdrawal from the European Union" to "Withdrawal of the United Kingdom from the European Union")* (undo) *(Tags: Shortdesc*

Web Data Mining

- Just applying data mining techniques to Web Data to discover knowledge?

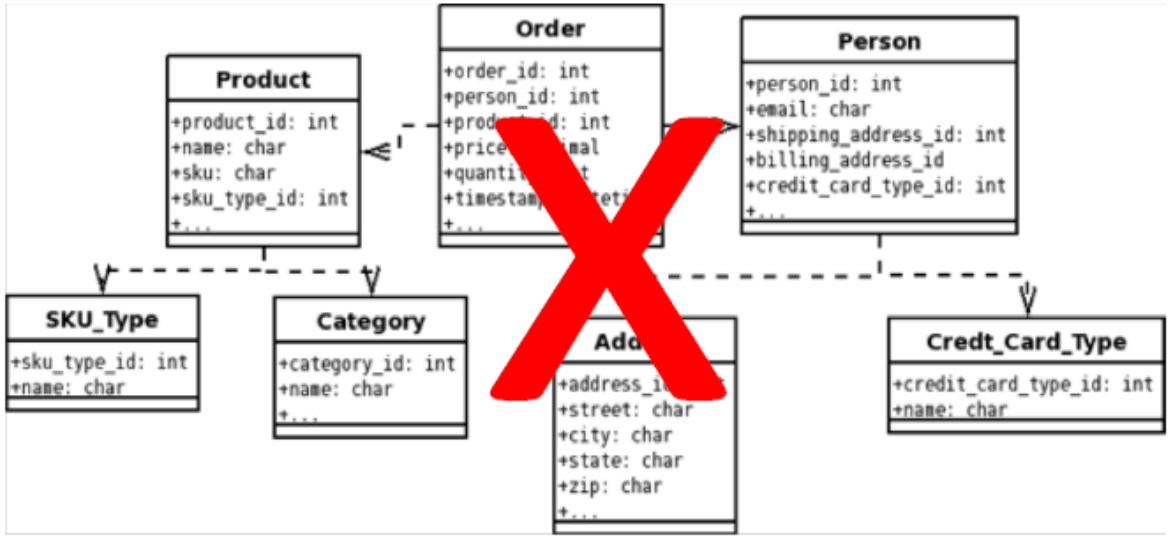
- In general, data acquisition and pre-processing steps are more challenging and (unfortunately) time-consuming.

Outline

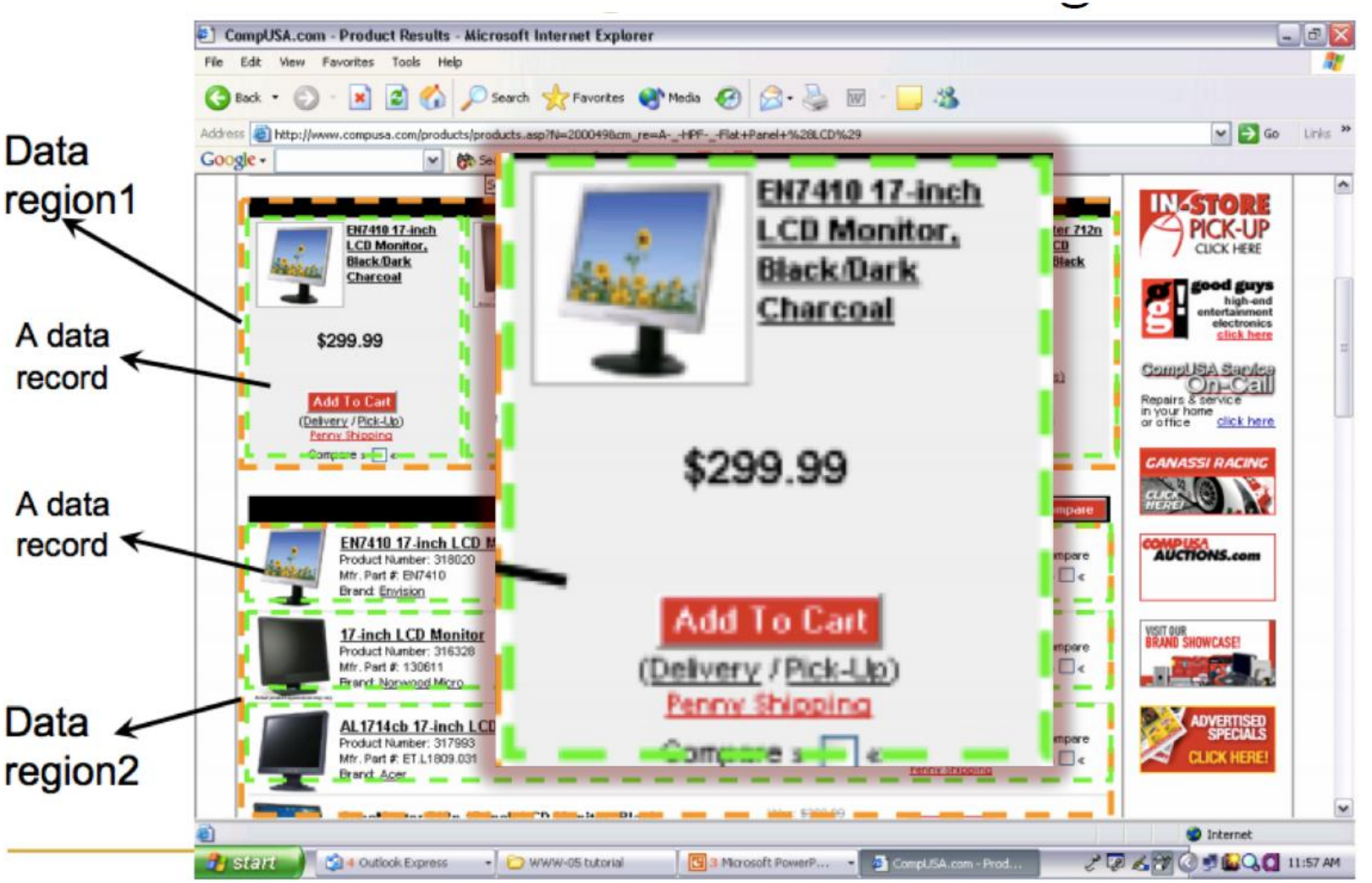
- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Recommender Systems
- Semantic Web

Web Content Mining

- To extract or mine useful information or knowledge from Web page contents
- No longer structured data (most of the cases)



Content shown on Web Page



https://www.cs.uic.edu/~liub/Web-Content-Mining-2.pdf

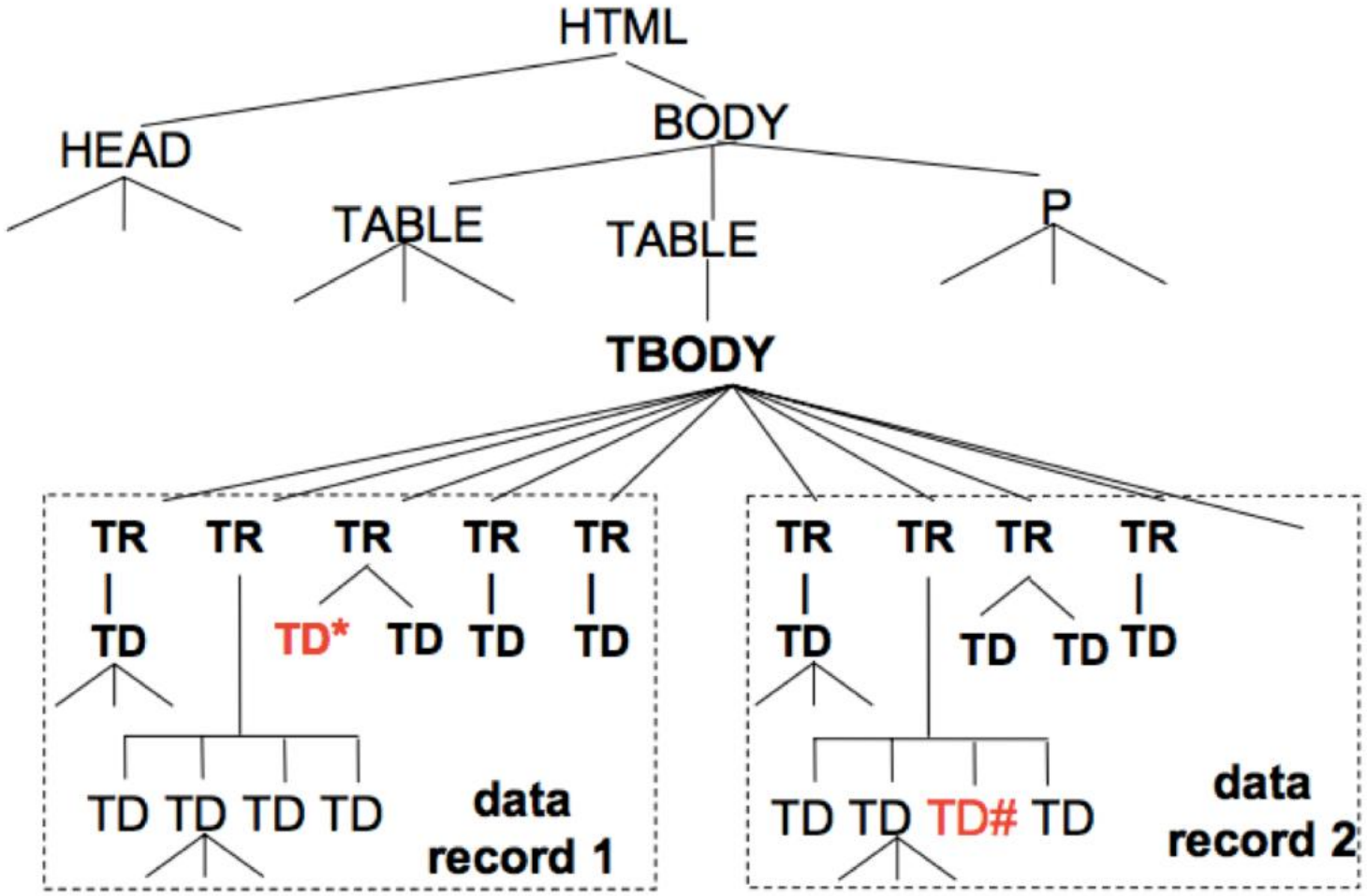
Structured Data to be obtained

image 1	EN7410 17-inch LCD Monitor Black/Dark charcoal		$299.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image 2	17-inch LCD Monitor		$249.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image 3	AL1714 17-inch LCD Monitor, Black		$269.99		Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare
image 4	SyncMaster 712n 17-inch LCD Monitor, Black	Was: $369.99	$299.99	Save $70 After: $70 mail-in-rebate(s)	Add to Cart	(Delivery / Pick-Up)	Penny Shopping	Compare

HOW?

https://www.cs.uic.edu/~liub/Web-Content-Mining-2.pdf

Web Page Layout Structure (tree of HTML elements + others)



https://www.cs.uic.edu/~liub/Web-Content-Mining-2.pdf

How to extract?

- Some examples:
  - Get the text content of all <h1> elements
  - Get the text content of all <p> elements with CSS class "item"
  - Get the text content of the first <p> element which appear inside a element whose CSS class is "movie"
  - Do the same as before, but skip all the content before the
 tag
  - etc...
- Libraries for navigating HTML tree may help (Beatiful Soup is one of these)
- Scripts and softwares for data acquisition have usually a short life …

Extract From Data in XML and JSON Obtained via Web APIs

XML

```xml
<?xml version="1.0" standalone="yes"?>
<BankAccount>
	<Number>1234</Number>
	<Type>Checking</Type>
	<OpenDate>11/04/1974</OpenDate>
	<Balance>25382.20</Balance>
	<AccountHolder>
		<LastName>Singh</LastName>
		<FirstName>Darshan</FirstName>
	</AccountHolder>
</BankAccount>
```

JSON

```json
{
 "query": {
  "count": 1,
  "created": "2014-08-22T03:02:17Z",
  "lang": "en-US",
  "results": {
   "quote": {
    "symbol": "MSFT",
    "Ask": null,
    "Bid": "43.00"
   }
  }
 }
}
```

From Web Data to Numerical Features

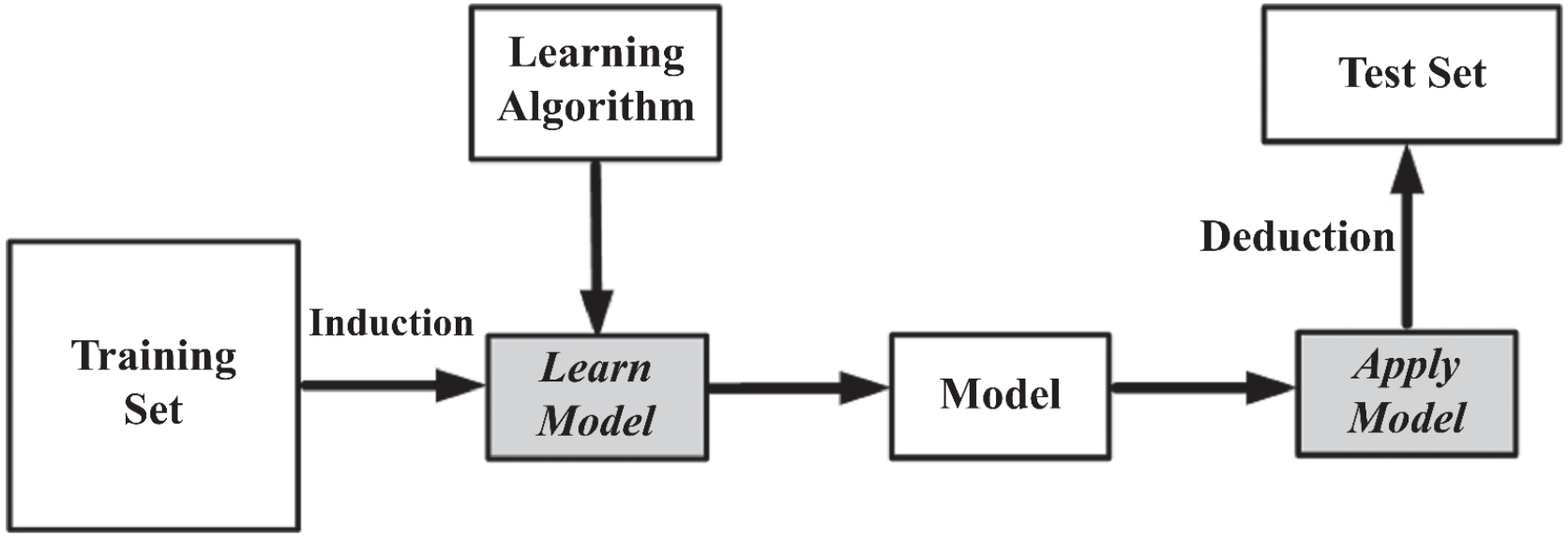
- Contents extracted from Web can be represented as some Feature Vectors so that some content mining tasks can be carried out

- Transforming Web Textual Data into numerical feature vectors allow to use a large variety of data mining and machine learning tools

Web Content Mining Tasks

- Web page classification
- Web page clustering
- Web information retrieval

Classification



- Given a set of labeled records/instances $\{(\boldsymbol{x}, y)\}$
  - $\boldsymbol{x}$ is a feature vector and y is the class label
- **[Training]** Find a mapping function m such that $m(x) = y$
- **[Testing]** Given an unlabeled instance $(\boldsymbol{x}', ?)$, compute $m(x')$ to predict the output label.

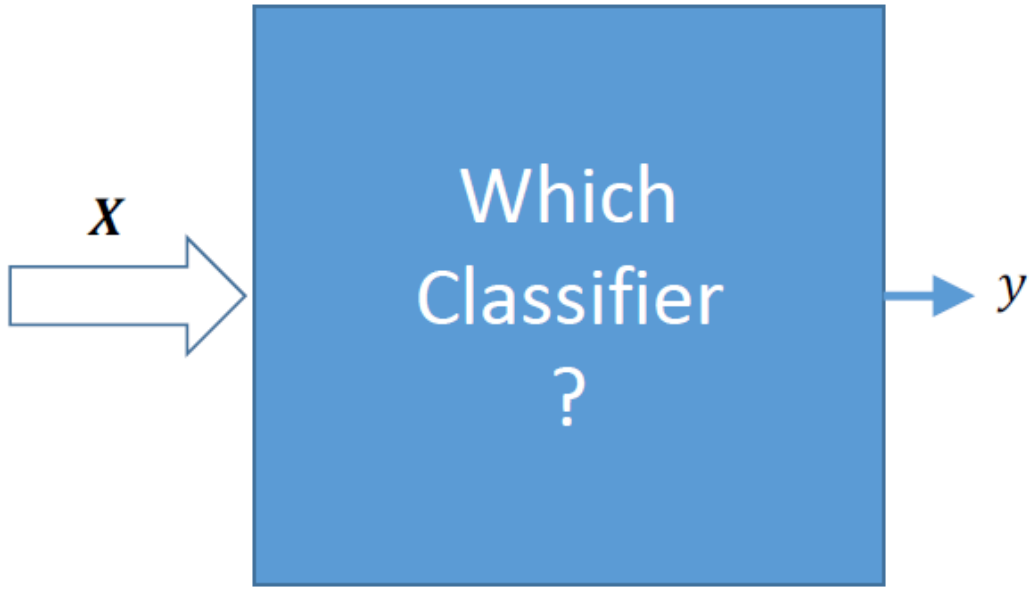
Classification Algorithms

- Many have been proposed:
  - Neural Networks
  - K-nearest Neighbor
  - Support Vector Machine
  - Decision Tree
  - Random Forest
  - XGBoost Tree
  - Naive Bayes
  - …

- Many are available in the Scikit Learn library:
  - https://scikit-learn.org/stable/supervised_learning.html

Classification Algorithms

- Many have been proposed:
  - Neural Networks
  - K-nearest Neighbor
  - Support Vector Machine
  - Decision Tree
  - Random Forest
  - XGBoost Tree
  - Naive Bayes
  - …



- Many are available in the Scikit Learn library:
  - https://scikit-learn.org/stable/supervised_learning.html

Be curious ☺

sklearn.neural_network.MLPClassifier

class sklearn.neural_network. **MLPClassifier** (*hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10*) [source]

If you want to know more about the implement ation ..

Multi-layer Perceptron classifier.

This model optimizes the log-loss function using LBFGS or stochastic gradient descent.

New in version 0.18.

Parameters:

hidden_layer_sizes : *tuple, length = n_layers - 2, default (100,)*

The ith element represents the number of neurons in the ith hidden layer.

activation : *{'identity', 'logistic', 'tanh', 'relu'}, default 'relu'*

Activation function for the hidden layer.

- 'identity', no-op activation, useful to implement linear bottleneck, returns f(x) = x
- 'logistic', the logistic sigmoid function, returns f(x) = 1 / (1 + exp(-x)).
- 'tanh', the hyperbolic tan function, returns f(x) = tanh(x).
- 'relu', the rectified linear unit function, returns f(x) = max(0, x)

solver : *{'lbfgs', 'sgd', 'adam'}, default 'adam'*

The solver for weight optimization.

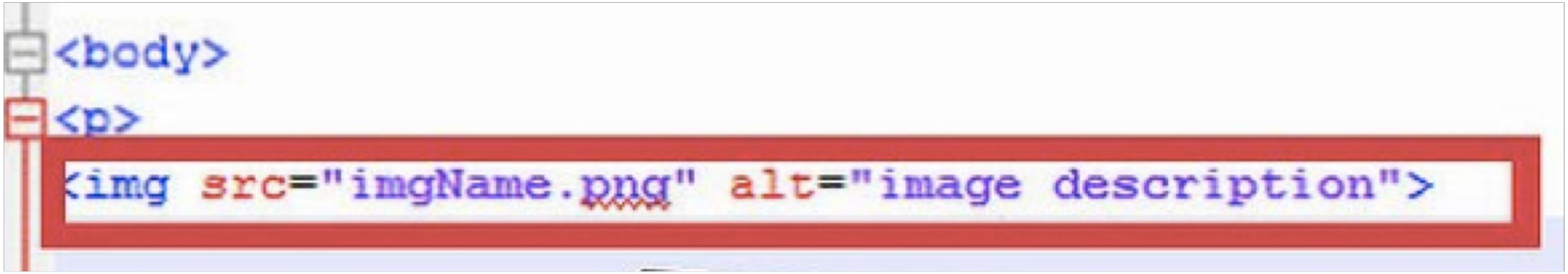
- 'lbfgs' is an optimizer in the family of quasi-Newton methods.

Ref: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Web Page Classification

- Textual content can be classified
- Layouts can be classified
- Hyperlink structures can be classified

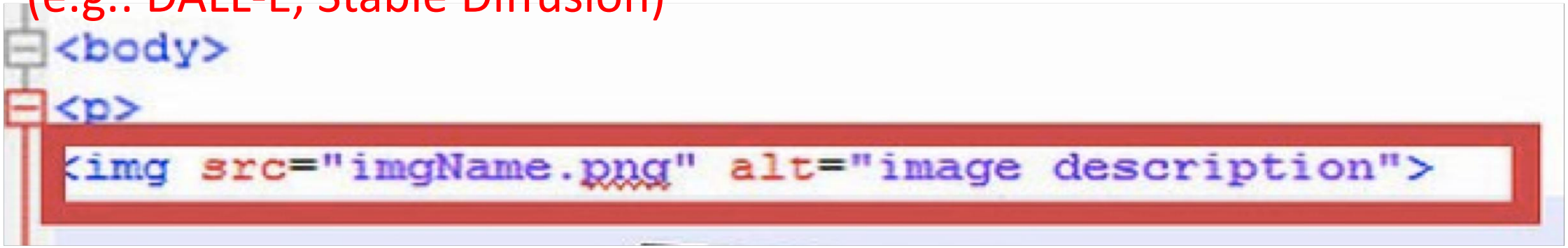
- Alternative text of images?



Web Page Classification

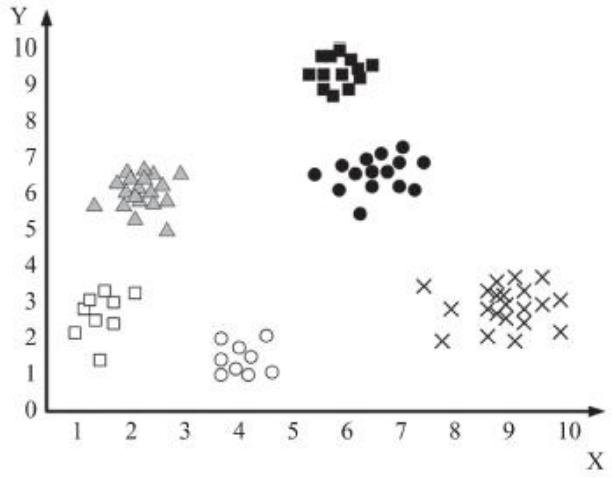
- Textual content can be classified
- Layouts can be classified
- Hyperlink structures can be classified

- Alternative text of images? Used to train the automatic image captioning systems and also the very recent "text to image" tools (e.g.: DALL-E, Stable Diffusion)

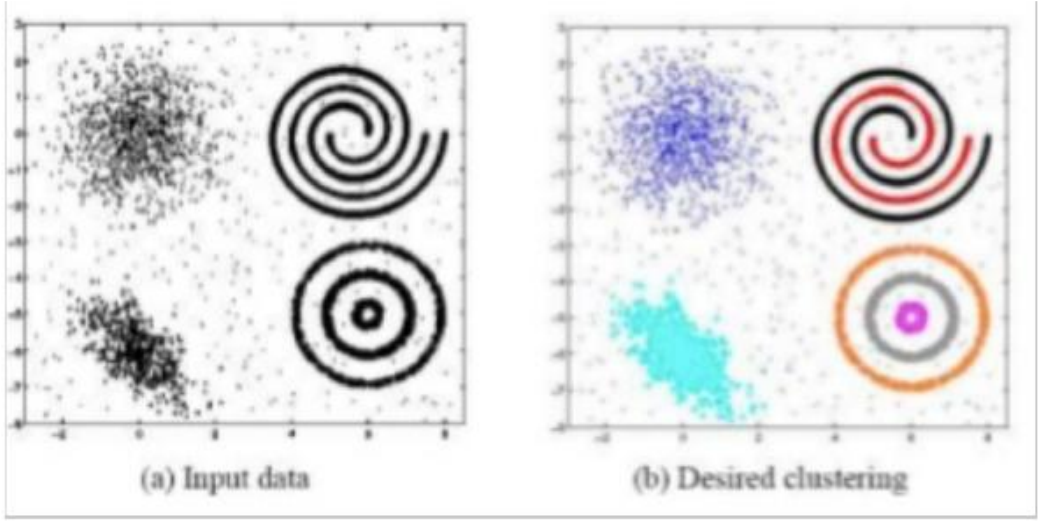


Clustering

- Clustering algorithms group together similar items
- No Label: The algorithm does not have examples showing how the samples should be grouped together (Unsupervised Learning)
- Again feature vectors are useful!



Easy Clustering Tasks

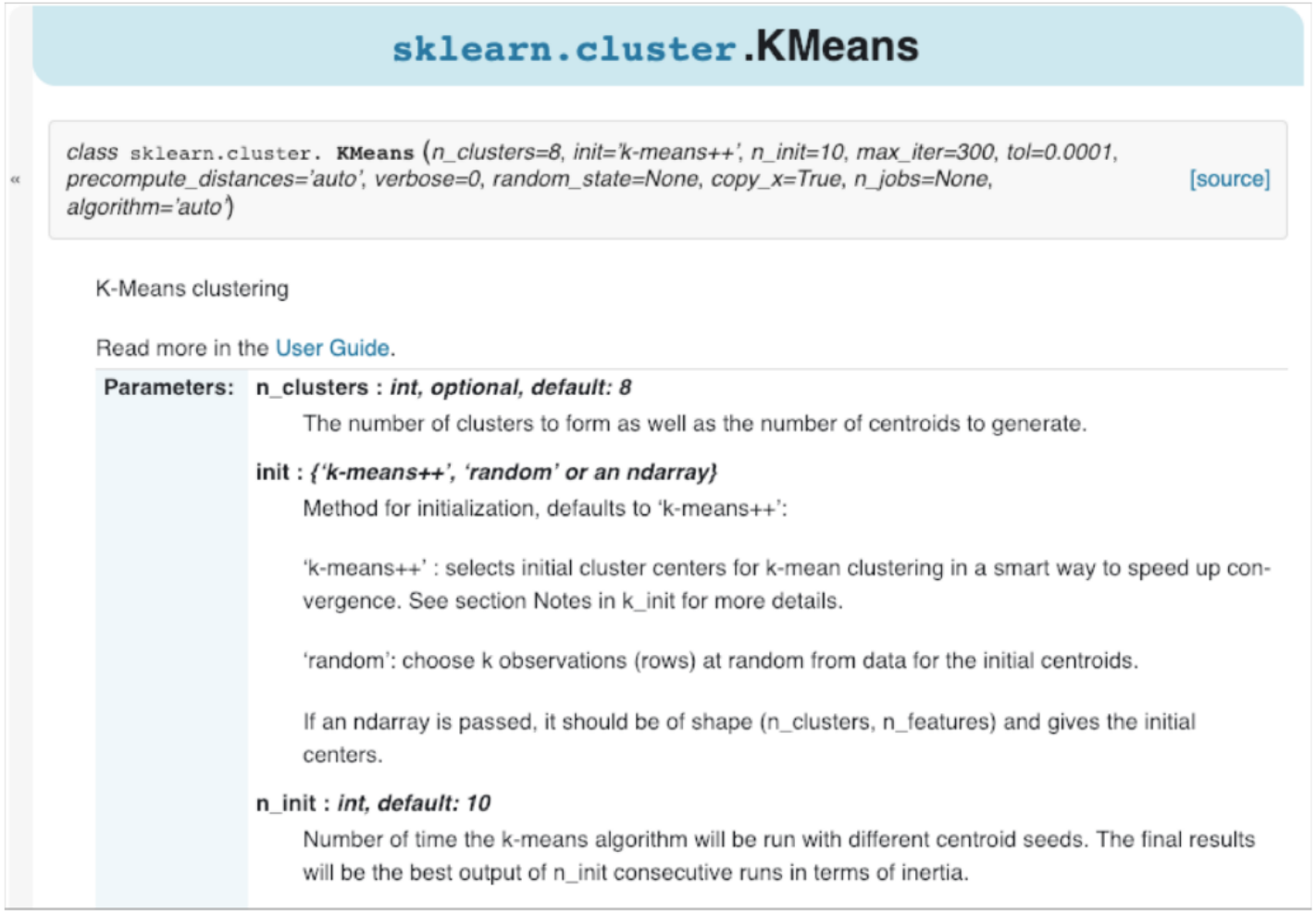


Challenging Clustering Tasks

Clustering Algorithms

- Many have been proposed
  - k-means
  - Agglomerative Hierarchical Clustering
  - Affinity Propagation
  - DBScan
  - …
- Many are available in Scikit Learn library:
  - https://scikit-learn.org/stable/modules/clustering.html

Reading documentation before using is important!



sklearn.cluster.KMeans

class sklearn.cluster. **KMeans** (*n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto'*) [source]

K-Means clustering

Read more in the User Guide.

Parameters:

n_clusters : *int, optional, default: 8*

The number of clusters to form as well as the number of centroids to generate.

init : *{'k-means++', 'random' or an ndarray}*

Method for initialization, defaults to 'k-means++':

'k-means++' : selects initial cluster centers for k-mean clustering in a smart way to speed up convergence. See section Notes in k_init for more details.

'random': choose k observations (rows) at random from data for the initial centroids.

If an ndarray is passed, it should be of shape (n_clusters, n_features) and gives the initial centers.

n_init : *int, default: 10*

Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia.

Ref: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Web Page Clustering

- Two pages are similar because
  - Contents are similar? or
  - Many hyperlinks between them? or
  - Their in-links and out-links are similar? or
  - …

- A meaningful similarity measure need to be defined!

Web Information Retrieval

- Problem
  - Search for web pages which are similar to a query and rank them
- Same (numerical) representation can be used for both pages and queries
- Output rankings may be customized for different users
  - Relevance feedback can allow that

Outline

- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

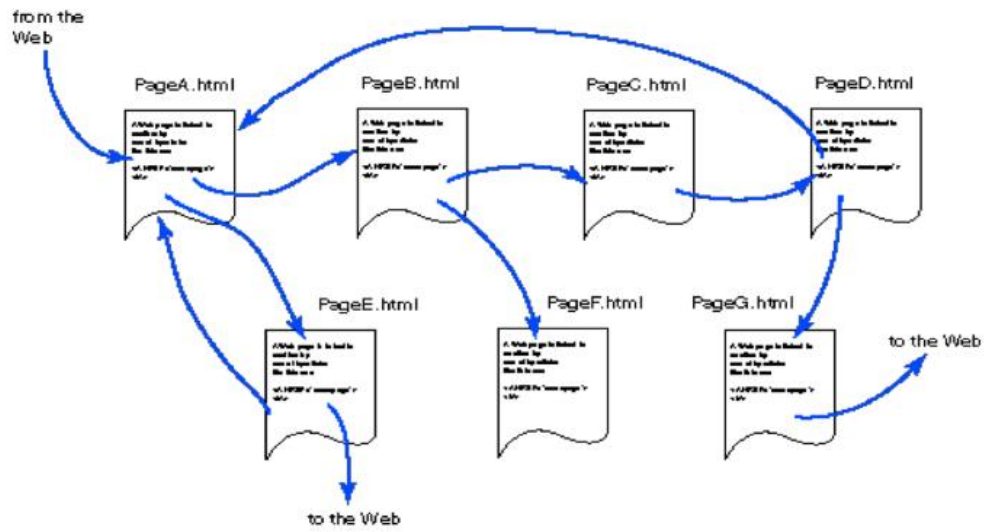
Web Structure Data

English Wikipedia

From Wikipedia, the free encyclopedia

The **English Wikipedia** is the English-language edit
encyclopedia Wikipedia. Founded on 15 Januar 20
Wikipedia and, as of November 2017, has the most
editions.[2] As of January 2018, 12% of articles in all
English-language edition. This share has gradually c
percent in 2003, due to the growth of Wikipedias in c
are 5,551,715 articles on the site (live count).[4] In O
text of the English Wikipedia's articles totalled 11.5 g
compressed.[5] On 1 November 2015, the English W

```
</table>
<p>The <b>English Wikipedia</b> is the <a href="
title="English language">English-language</a> ed
encyclopedia <a href="/wiki/Wikipedia" title="Wi
Founded on 15 January 2001, it is the first edit
November 2017<sup class="plainlinks noexcerpt no
style="display:none;"><a class="external text"
href="//en.wikipedia.org/w/index.php?title=Engli
[update]</a></sup>, has <a href="/wiki/List_of_W
Wikipedias">the most articles of any of the edit
class="reference"><a href="#cite_note-2">[2]</a>
<span style="display:none" class="sortkey">70011
articles in all Wikipedias belong to the English
has gradually declined from more than 50 percent
```

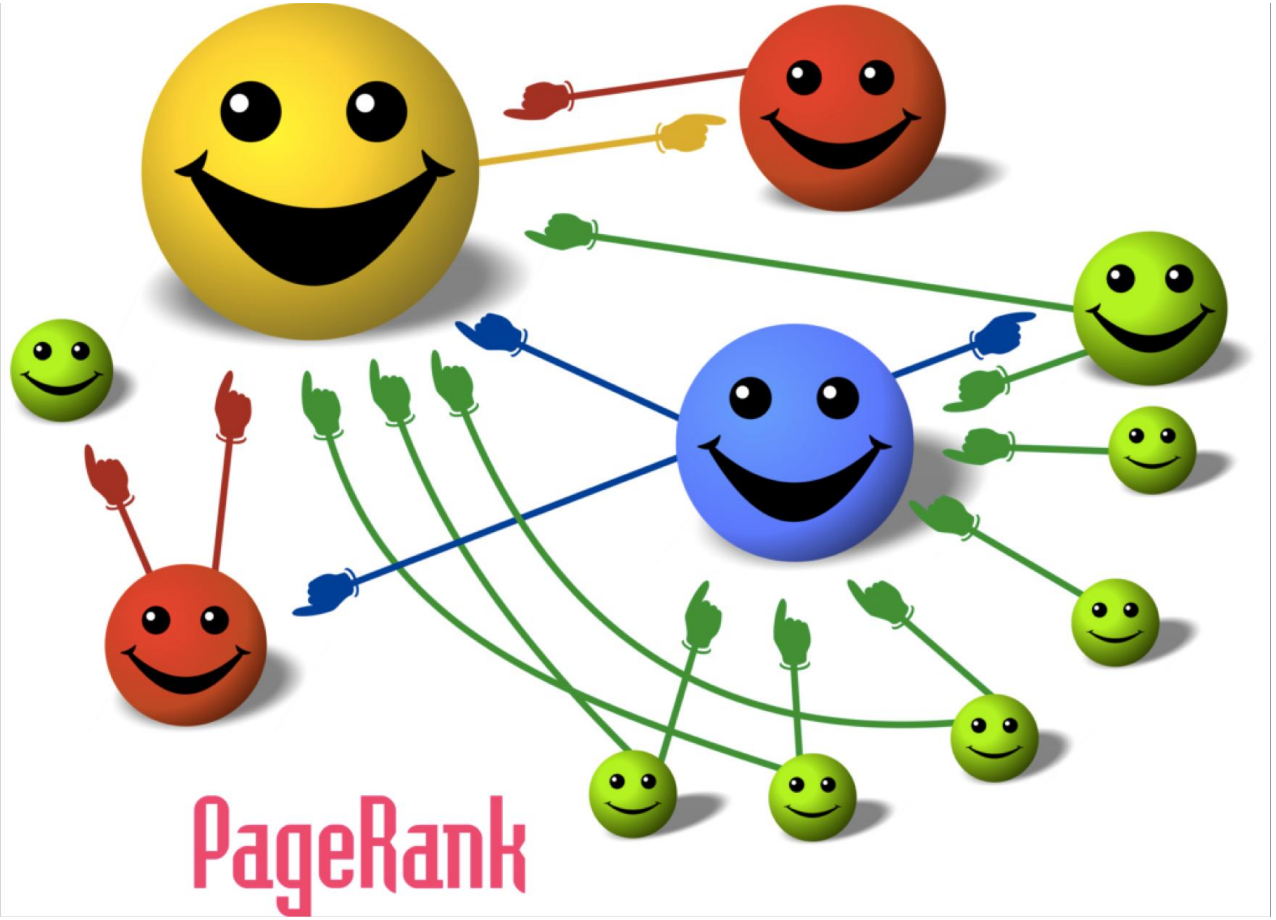


Crawling

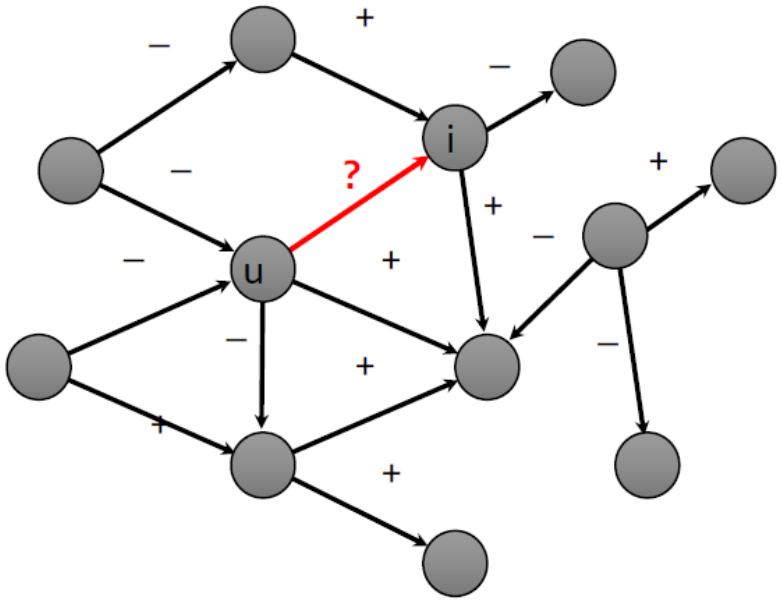
- The crawler starts at an initial web page
- The web page (in HTML) is downloaded and parsed
- The links within the web page are extracted
- All the links are candidates for the next web pages to be requested
- Crawling may continue following different strategies:
  - Breadth-first
  - Depth-first
  - how about crawling for pages under some focused topics?

Web Structure Mining

- Discover important Web pages from link structure (key technology used in search engines)
- Discover communities of users/pages who are closely linked with each others



Link Prediction in Social Networks



- The problem is to predict what will be the new connections

Outline

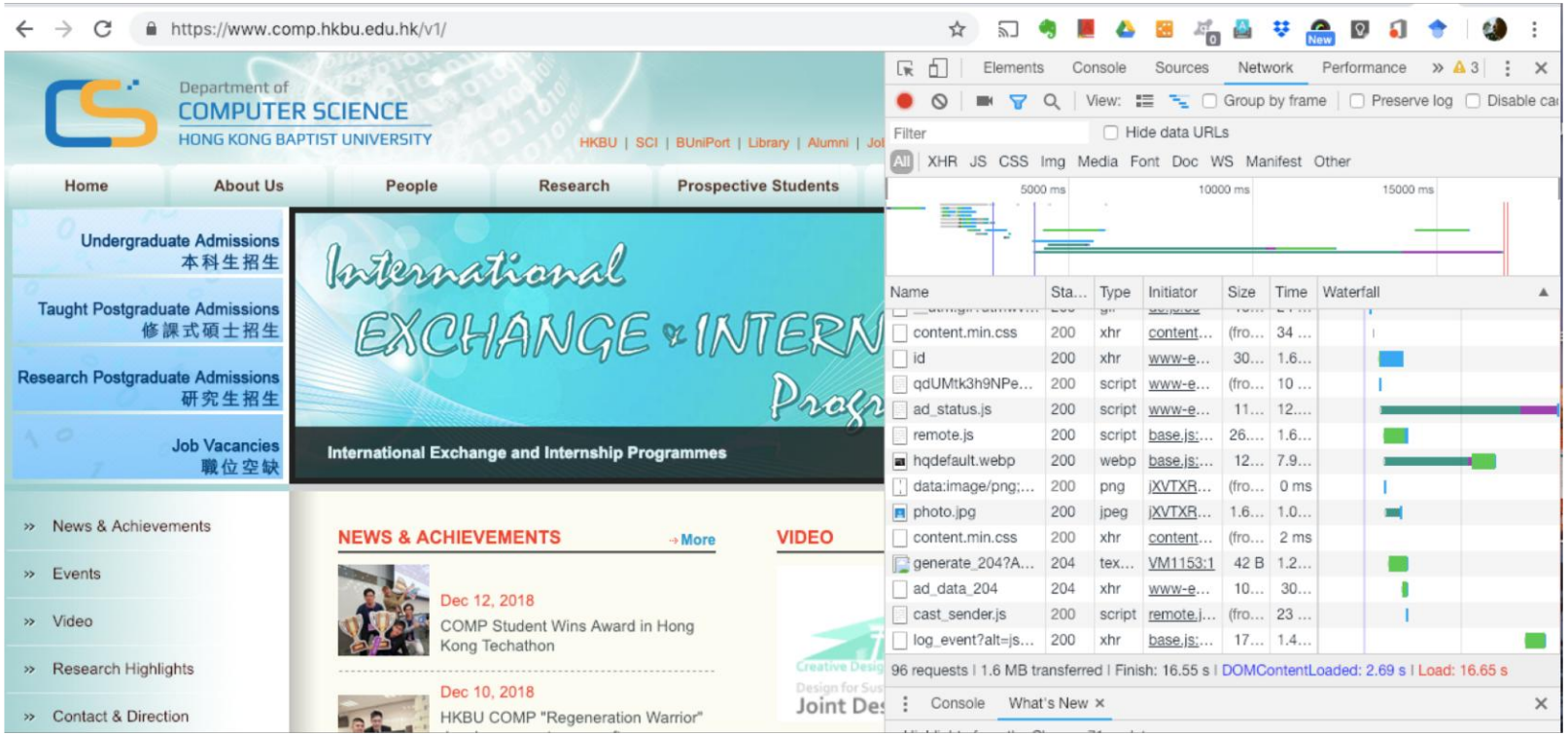
- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

Web Usage Mining and Recommendation

- To discover user access patterns from Web usage logs, which record every click made by each user
  - Page/product recommendation (people click this also click that).
  - User intention prediction (people clicking through up to this “status” is more likely to buy)
  - Web surfing regularity characterization (rational users, random users, or recurrent users)
  - …

Clickstream Data

- Clickstream data obtained from web logs need to be preprocessed
  - When a user clicks on a hyperlink on a web page, how many http requests will be generated?
  - Also, how many users are surfing a web server concurrently?



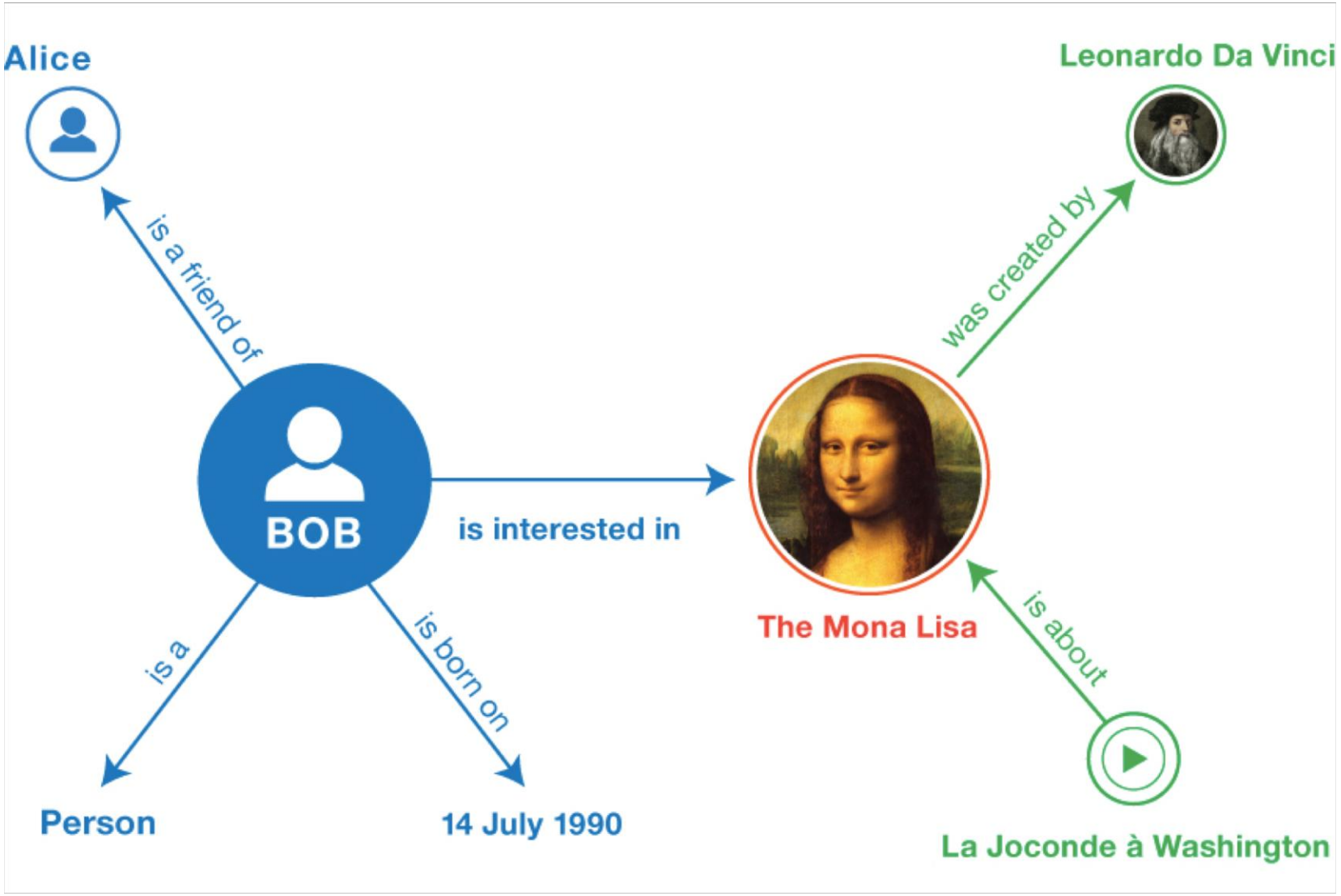
Web Mining (again)

- Conventional data mining - Data is often already collected and stored in a data warehouse
- Web mining - Data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages.
- Once the data is collected, data pre-processing is still needed.
- However, the techniques used for each step can be quite different from those used in conventional data mining.

Outline

- Definition of Web Intelligence
- Distributed Problem Solving
- Characteristics of Web Data
- Web Content Mining and Retrieval
- Web Structure / Social Network Analysis
- Web Usage Mining / Collaborative Filtering
- Semantic Web

Semantic Web (RDF, OWL, SPARQL)



Knoledge Representation and Reasoning

- The semantic web is formed by a set of standard languages which allow:
  - representing information about the world in a form that a computer system can utilize to solve complex tasks such as diagnosing a medical condition or having a dialog in a natural language;
  - incorporates findings from logic to automate various kinds of reasoning, such as the application of rules or the relations of sets and subsets.
- Knowledge graphs may be built
  - Useful to fight fake-news!!!
- Sir Tim Berners Lee is behind the Semantic Web, but we are still far ...
- AI techniques on unstructured and semi-structured data are the main technology applied right now!

REFERENCES

- Liu Bing. Web Data Mining – Exploring Hyperlinks, Contents and Usage Data, Springer, 2011 [Ch 1]

- Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. Social Media Mining: An Introduction, Cambridge University Press, 2014 [Ch 5]