Retrieval and Mining as Scientific Tools

Benno Stein
Bauhaus-Universität Weimar
www.webis.de

7. KSWD, February 9th 2017

Outline

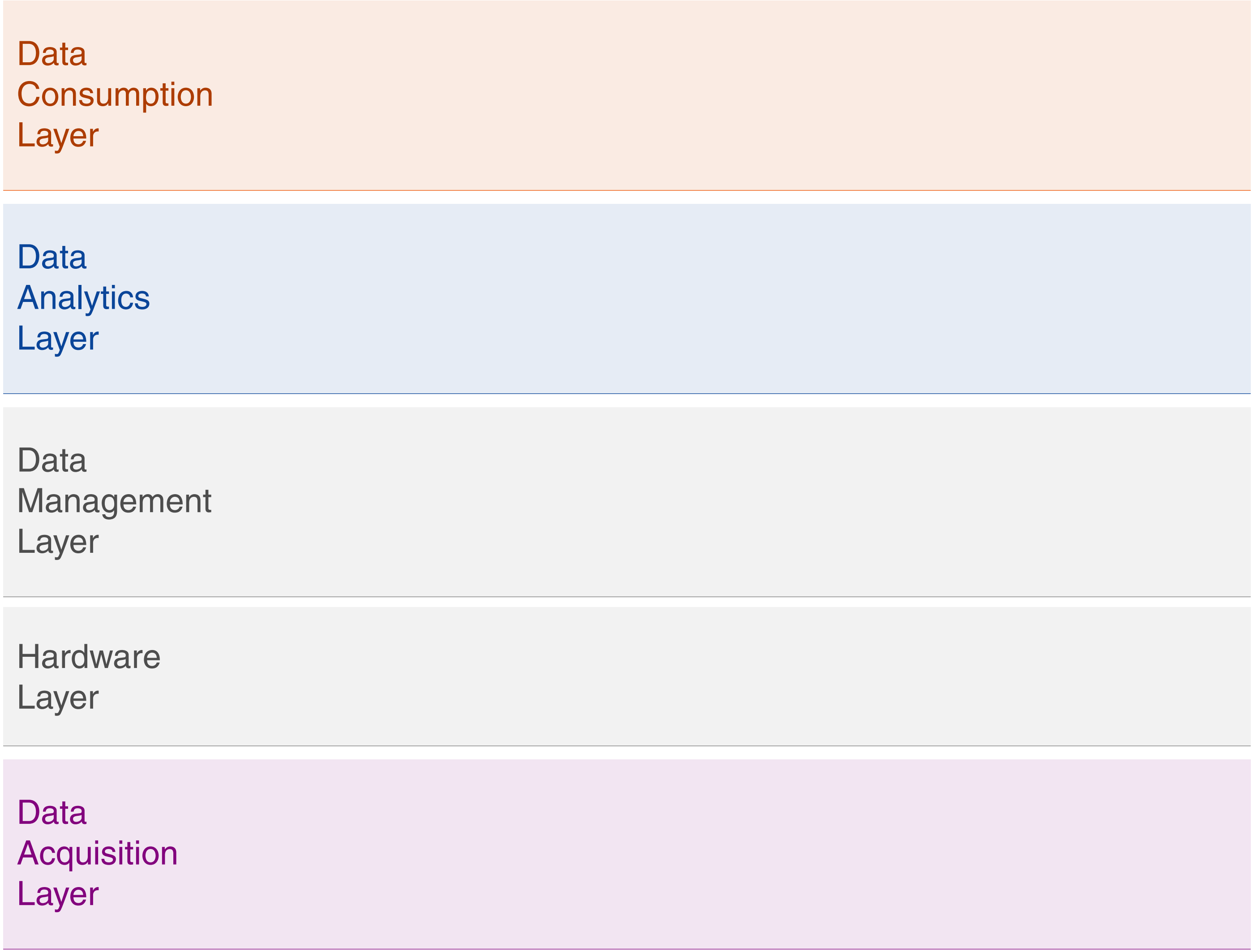
❑ The Big Data Setup

❑ Vandalism in Wikipedia

❑ The BSI Password Creation Advice

❑ Outlook: The “Why” Search Engine

(My) Big Data Setup



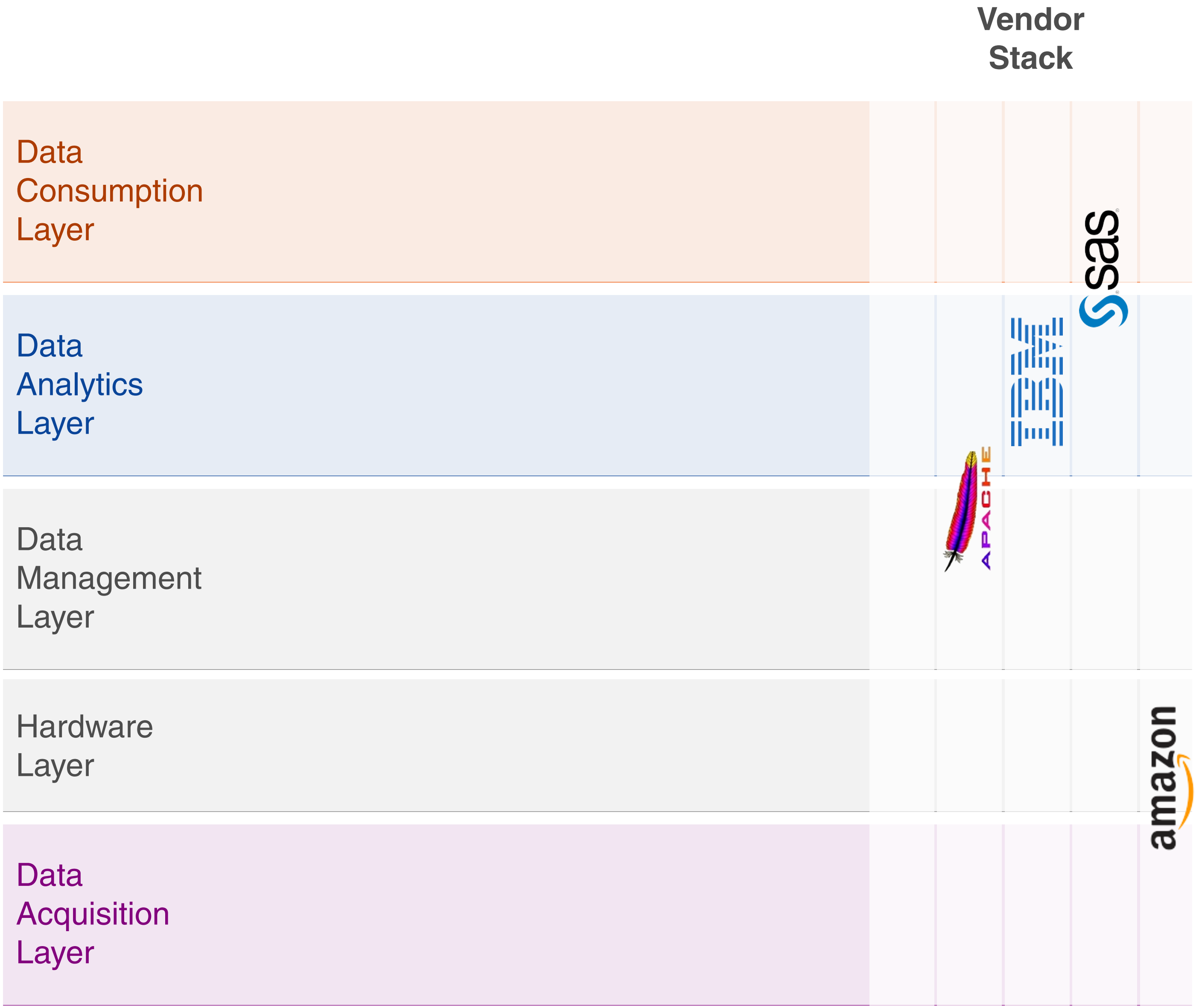
Data Consumption Layer

Data Analytics Layer

Data Management Layer

Hardware Layer

Data Acquisition Layer



Vendor Stack

Data Consumption Layer

Data Analytics Layer

Data Management Layer

Hardware Layer

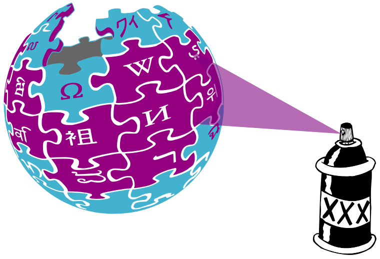
Data Acquisition Layer

	Technology Stack	Vendor Stack
Data Consumption Layer	- Level-of-detail management - Multi-touch	sas
Data Analytics Layer	- MapReduce - State-space search	IBM
Data Management Layer	- Key-value store - RDF triple store	APACHE
Hardware Layer	- Cloud vs. inhome	amazon
Data Acquisition Layer	- Crowdsourcing	

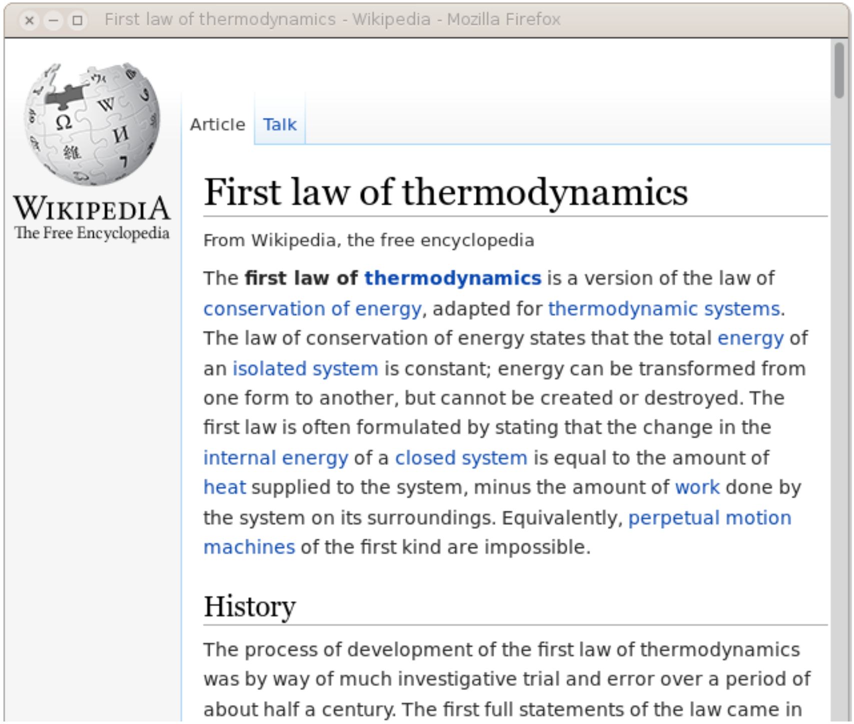
	Abstract Task Stack	Technology Stack	Vendor Stack
Data Consumption Layer	- Explore - Query - Interact	- Level-of-detail management - Multi-touch	sas
Data Analytics Layer	- Reason - Structure identificat. - Structure verification	- MapReduce - State-space search	IBM
Data Management Layer	- Store - Normalize	- Key-value store - RDF triple store	APACHE
Hardware Layer	- Virtualization	- Cloud vs. inhome	amazon
Data Acquisition Layer	- Collect - Log	- Crowdsourcing	

	Abstract Task Stack	Technology Stack	Vendor Stack	Roles
Data Consumption Layer	- Explore - Query - Interact	- Level-of-detail management - Multi-touch	sas	Data scientist
Data Analytics Layer	- Reason - Structure identificat. - Structure verification	- MapReduce - State-space search	IBM	
Data Management Layer	- Store - Normalize	- Key-value store - RDF triple store	APACHE	System architect
Hardware Layer	- Virtualization	- Cloud vs. inhome	amazon	System admin
Data Acquisition Layer	- Collect - Log	- Crowdsourcing		Data scientist

Vandalism in Wikipedia



Vandalism in Wikipedia (Example: wrong facts, nonsense)



First law of thermodynamics - Wikipedia - Mozilla Firefox

First law of thermodynamics

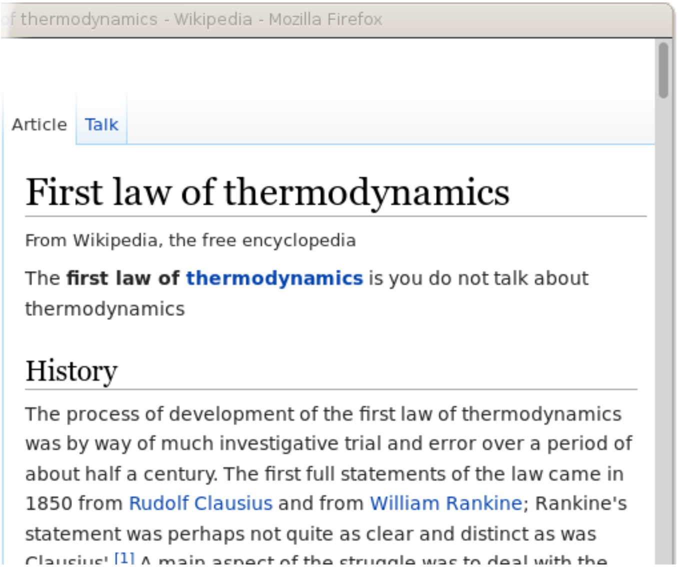
From Wikipedia, the free encyclopedia

The **first law of thermodynamics** is a version of the law of conservation of energy, adapted for thermodynamic systems. The law of conservation of energy states that the total energy of an isolated system is constant; energy can be transformed from one form to another, but cannot be created or destroyed. The first law is often formulated by stating that the change in the internal energy of a closed system is equal to the amount of heat supplied to the system, minus the amount of work done by the system on its surroundings. Equivalently, perpetual motion machines of the first kind are impossible.

History

The process of development of the first law of thermodynamics was by way of much investigative trial and error over a period of about half a century. The first full statements of the law came in

→



f thermodynamics - Wikipedia - Mozilla Firefox

First law of thermodynamics

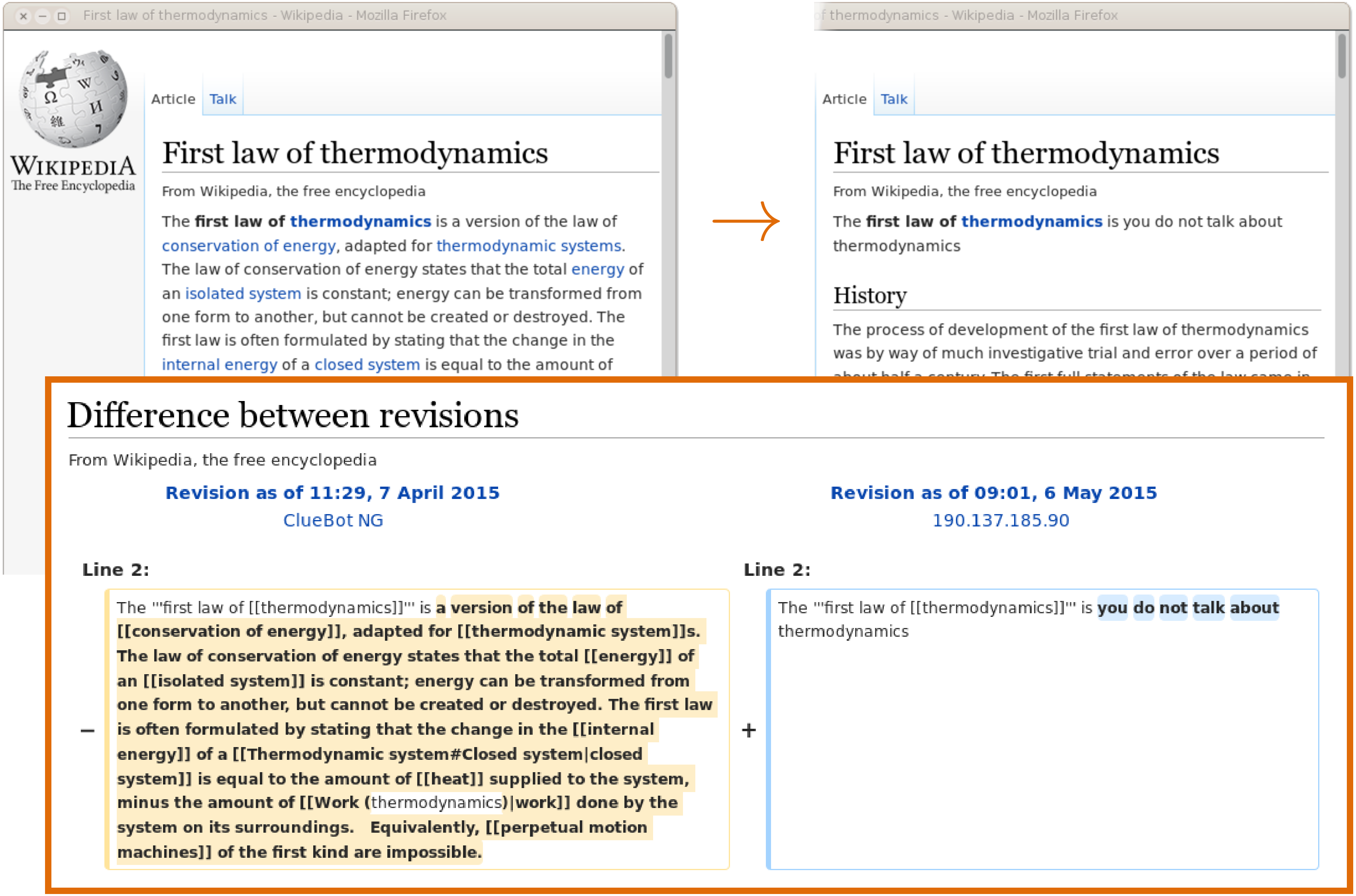
From Wikipedia, the free encyclopedia

The **first law of thermodynamics** is you do not talk about thermodynamics

History

The process of development of the first law of thermodynamics was by way of much investigative trial and error over a period of about half a century. The first full statements of the law came in 1850 from Rudolf Clausius and from William Rankine; Rankine's statement was perhaps not quite as clear and distinct as was Clausius'.[1] A main aspect of the struggle was to deal with the

Vandalism in Wikipedia (Example: wrong facts, nonsense)



First law of thermodynamics

From Wikipedia, the free encyclopedia

The **first law of thermodynamics** is a version of the law of conservation of energy, adapted for thermodynamic systems. The law of conservation of energy states that the total energy of an isolated system is constant; energy can be transformed from one form to another, but cannot be created or destroyed. The first law is often formulated by stating that the change in the internal energy of a closed system is equal to the amount of

First law of thermodynamics

From Wikipedia, the free encyclopedia

The **first law of thermodynamics** is you do not talk about thermodynamics

History

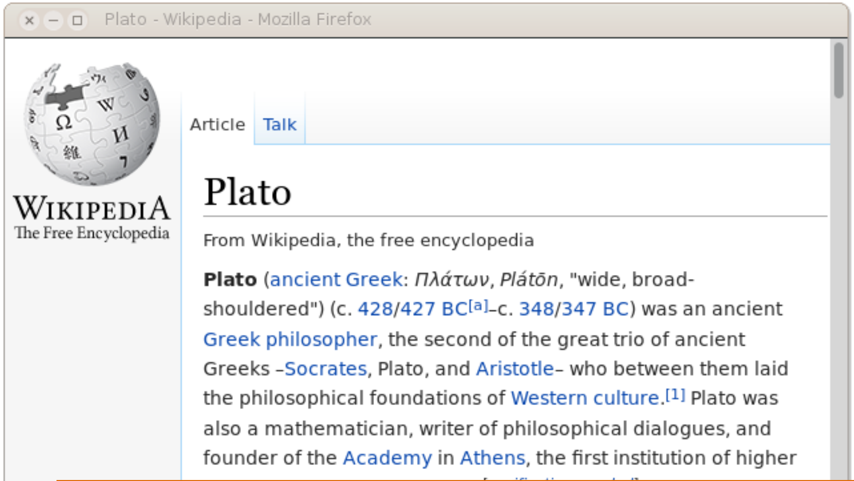
The process of development of the first law of thermodynamics was by way of much investigative trial and error over a period of

Difference between revisions

From Wikipedia, the free encyclopedia

Revision as of 11:29, 7 April 2015 ClueBot NG	Revision as of 09:01, 6 May 2015 190.137.185.90
Line 2:	**Line 2:**
− The '''first law of [[thermodynamics]]''' is **a version of the law of [[conservation of energy]], adapted for [[thermodynamic system]]s. The law of conservation of energy states that the total [[energy]] of an [[isolated system]] is constant; energy can be transformed from one form to another, but cannot be created or destroyed. The first law is often formulated by stating that the change in the [[internal energy]] of a [[Thermodynamic system#Closed system|closed system]] is equal to the amount of [[heat]] supplied to the system, minus the amount of [[Work (**thermodynamics**)|work]] done by the system on its surroundings. Equivalently, [[perpetual motion machines]] of the first kind are impossible.**	+ The '''first law of [[thermodynamics]]''' is **you do not talk about** thermodynamics

Vandalism in Wikipedia (Example: wrong facts, nonsense)



Plato - Wikipedia - Mozilla Firefox

WIKIPEDIA
The Free Encyclopedia

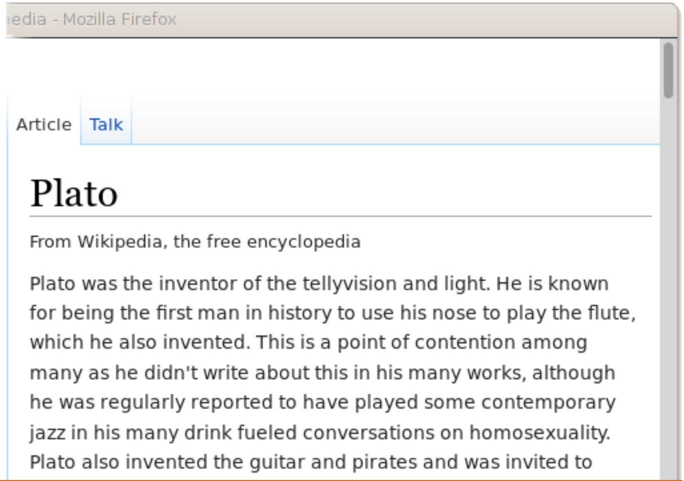
Article Talk

Plato

From Wikipedia, the free encyclopedia

Plato (ancient Greek: *Πλάτων*, *Plátōn*, "wide, broad-shouldered") (c. 428/427 BC[a]–c. 348/347 BC) was an ancient Greek philosopher, the second of the great trio of ancient Greeks –Socrates, Plato, and Aristotle– who between them laid the philosophical foundations of Western culture.[1] Plato was also a mathematician, writer of philosophical dialogues, and founder of the Academy in Athens, the first institution of higher

→



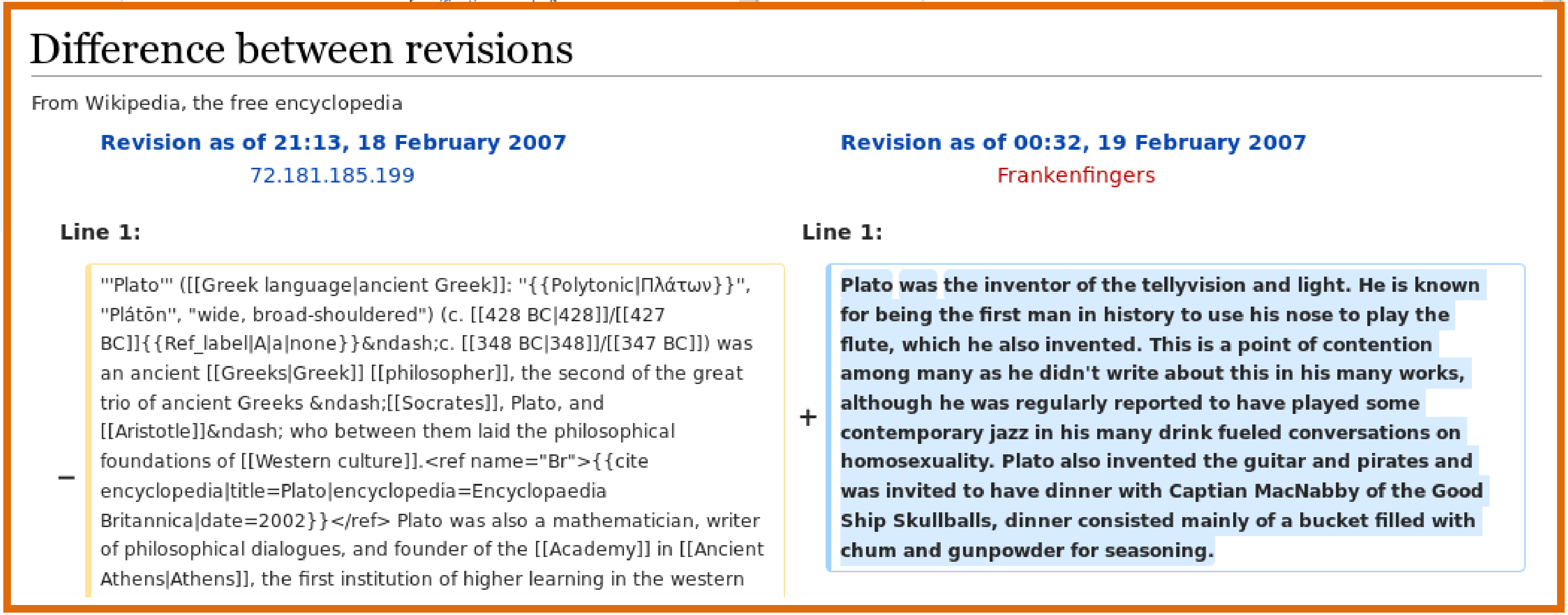
edia - Mozilla Firefox

Article Talk

Plato

From Wikipedia, the free encyclopedia

Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a point of contention among many as he didn't write about this in his many works, although he was regularly reported to have played some contemporary jazz in his many drink fueled conversations on homosexuality. Plato also invented the guitar and pirates and was invited to



Difference between revisions

From Wikipedia, the free encyclopedia

Revision as of 21:13, 18 February 2007 72.181.185.199	Revision as of 00:32, 19 February 2007 Frankenfingers
Line 1:	**Line 1:**
− '''Plato''' ([[Greek language|ancient Greek]]: "{{Polytonic|Πλάτων}}", "Plátōn", "wide, broad-shouldered") (c. [[428 BC|428]]/[[427 BC]]{{Ref_label|A|a|none}}&ndash;c. [[348 BC|348]]/[[347 BC]]) was an ancient [[Greeks|Greek]] [[philosopher]], the second of the great trio of ancient Greeks &ndash;[[Socrates]], Plato, and [[Aristotle]]&ndash; who between them laid the philosophical foundations of [[Western culture]].<ref name="Br">{{cite encyclopedia|title=Plato|encyclopedia=Encyclopaedia Britannica|date=2002}}</ref> Plato was also a mathematician, writer of philosophical dialogues, and founder of the [[Academy]] in [[Ancient Athens|Athens]], the first institution of higher learning in the western	+ **Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a point of contention among many as he didn't write about this in his many works, although he was regularly reported to have played some contemporary jazz in his many drink fueled conversations on homosexuality. Plato also invented the guitar and pirates and was invited to have dinner with Captian MacNabby of the Good Ship Skullballs, dinner consisted mainly of a bucket filled with chum and gunpowder for seasoning.**

Vandalism in Wikipedia

❑ 470 million article edits since 2003

❑ 40 million edits (9.5%) are vandalism

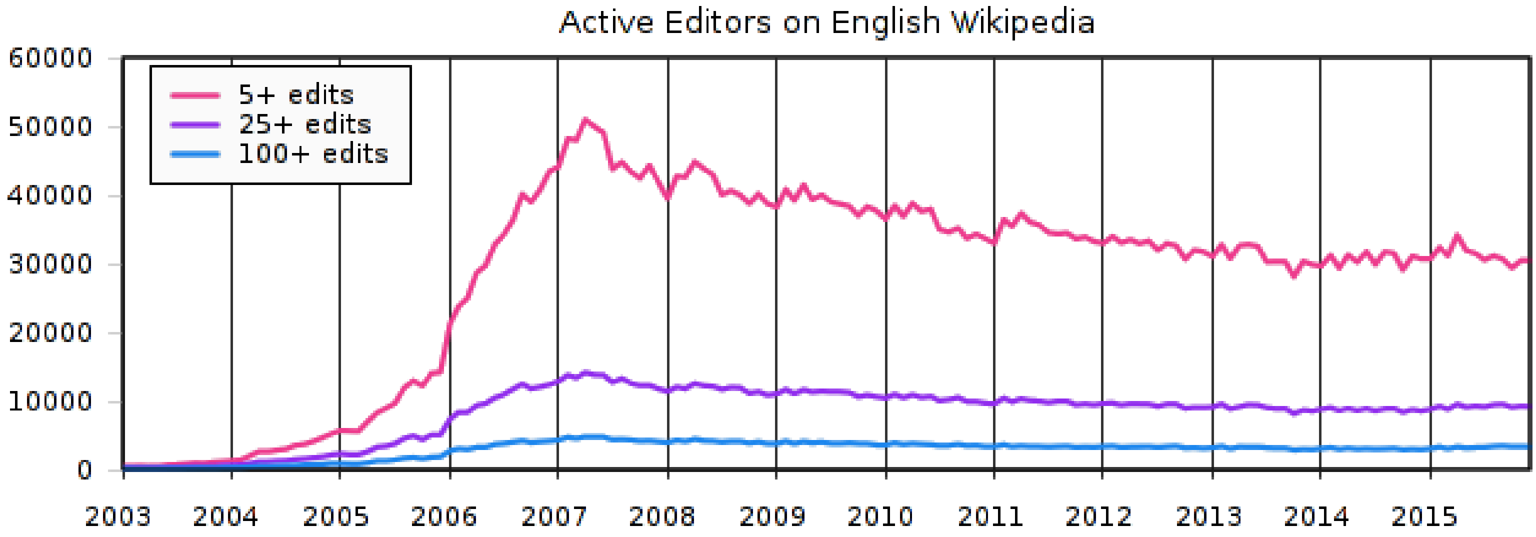
→ a vandalism exploit every 10s

Vandalism in Wikipedia

- ❏ 470 million article edits since 2003
- ❏ 40 million edits (9.5%) are vandalism
- → a vandalism exploit every 10s

Countermeasure: Vandalism detection bots that revert certain edits

Problem: High false positives rates of the bots discourage new editors



Vandalism in Wikipedia

How to build better vandalism detection technology?

→ Understand why people vandalize in Wikipedia.

  → Analyze *when* people vandalize.

  → Analyze *where* these people are.

Vandalism in Wikipedia

How to build better vandalism detection technology?

→ Understand why people vandalize in Wikipedia.

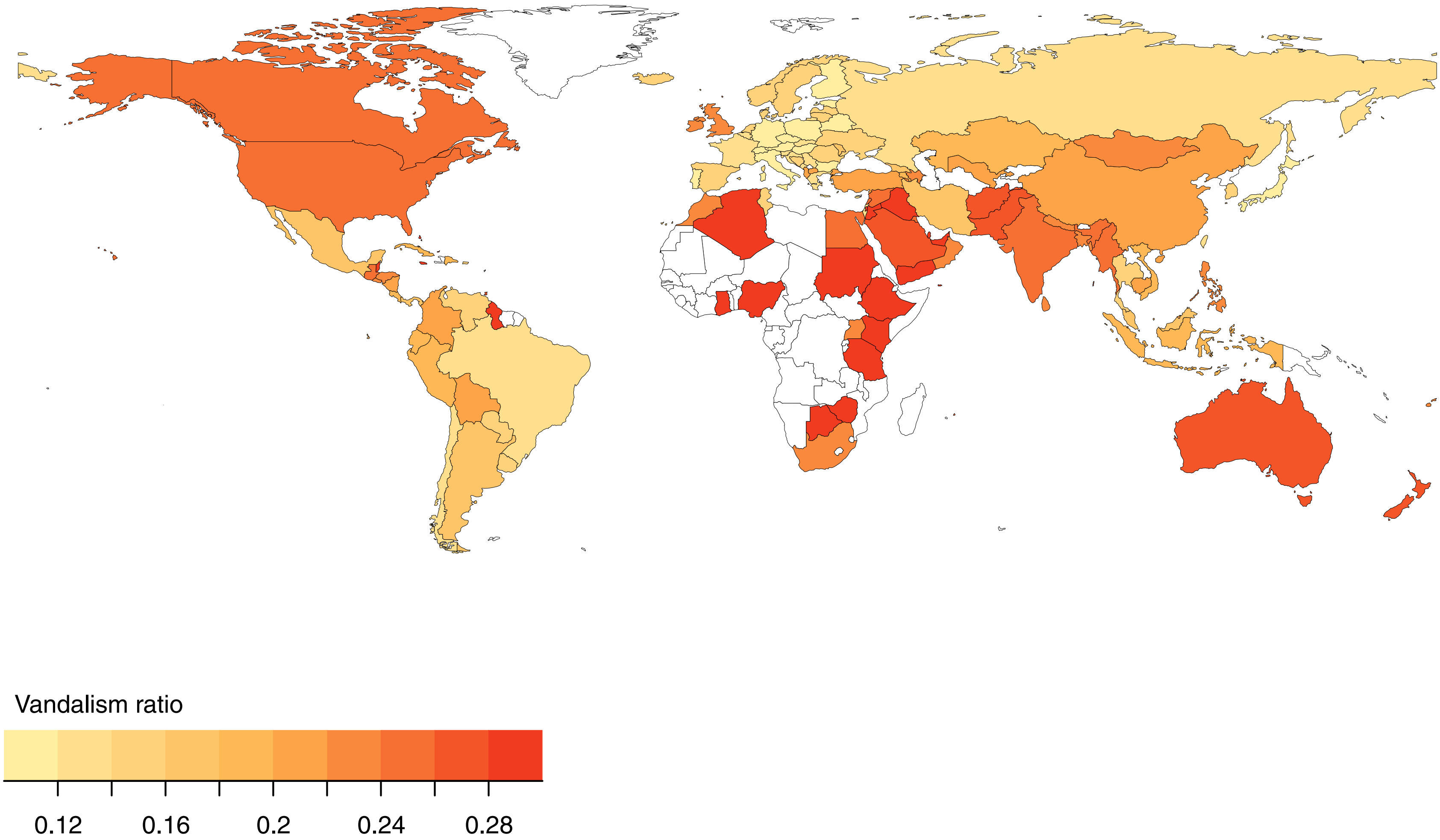
→ Analyze *when* people vandalize.

→ Analyze *where* these people are.

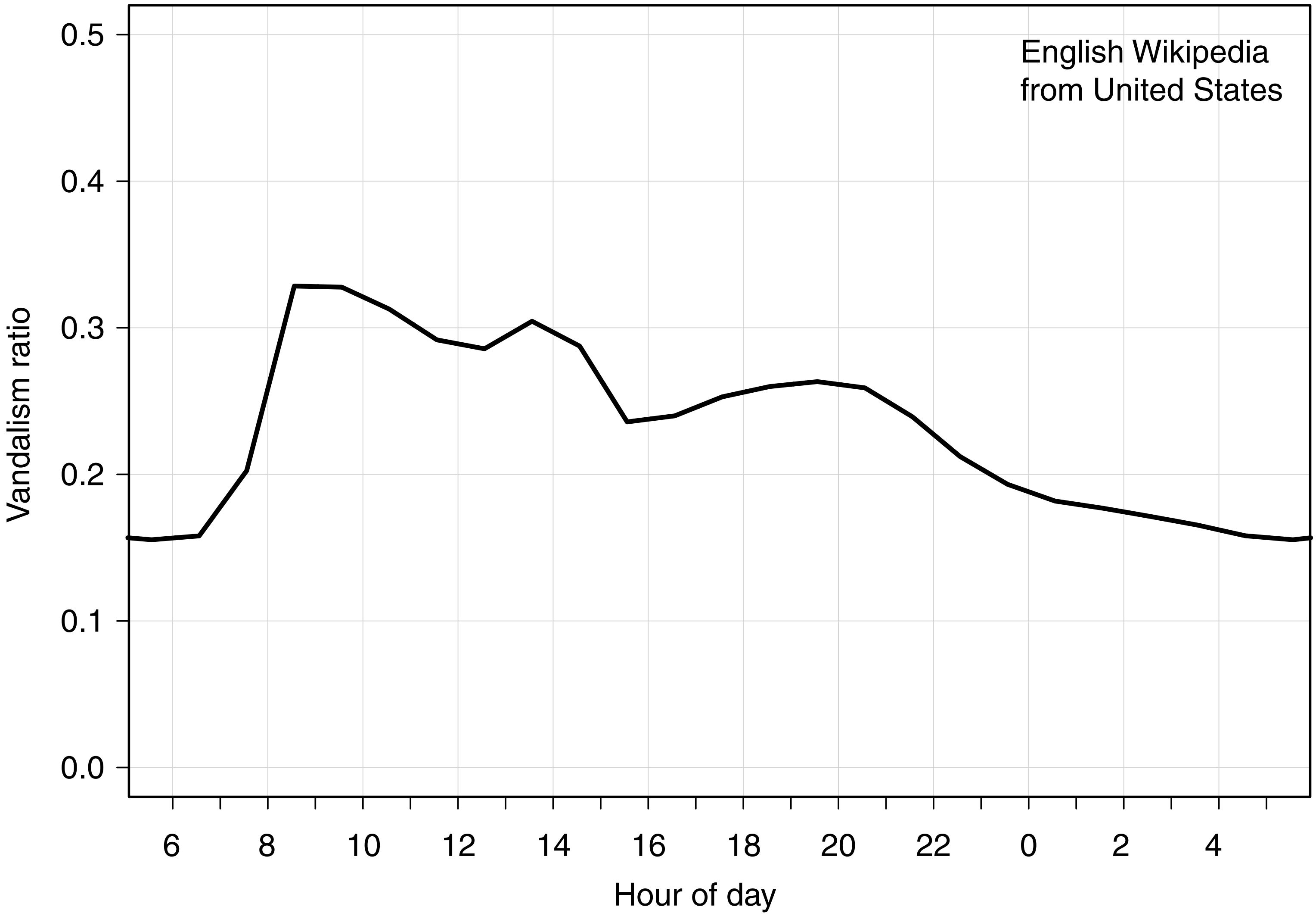
We analyzed in this regard . . .

- ≈ 30 Million articles
- ≈ 1.2 Billion page and article edits
- all involved IP addresses of the last 13 years
- the individual geolocations and time zones at the days of the edits

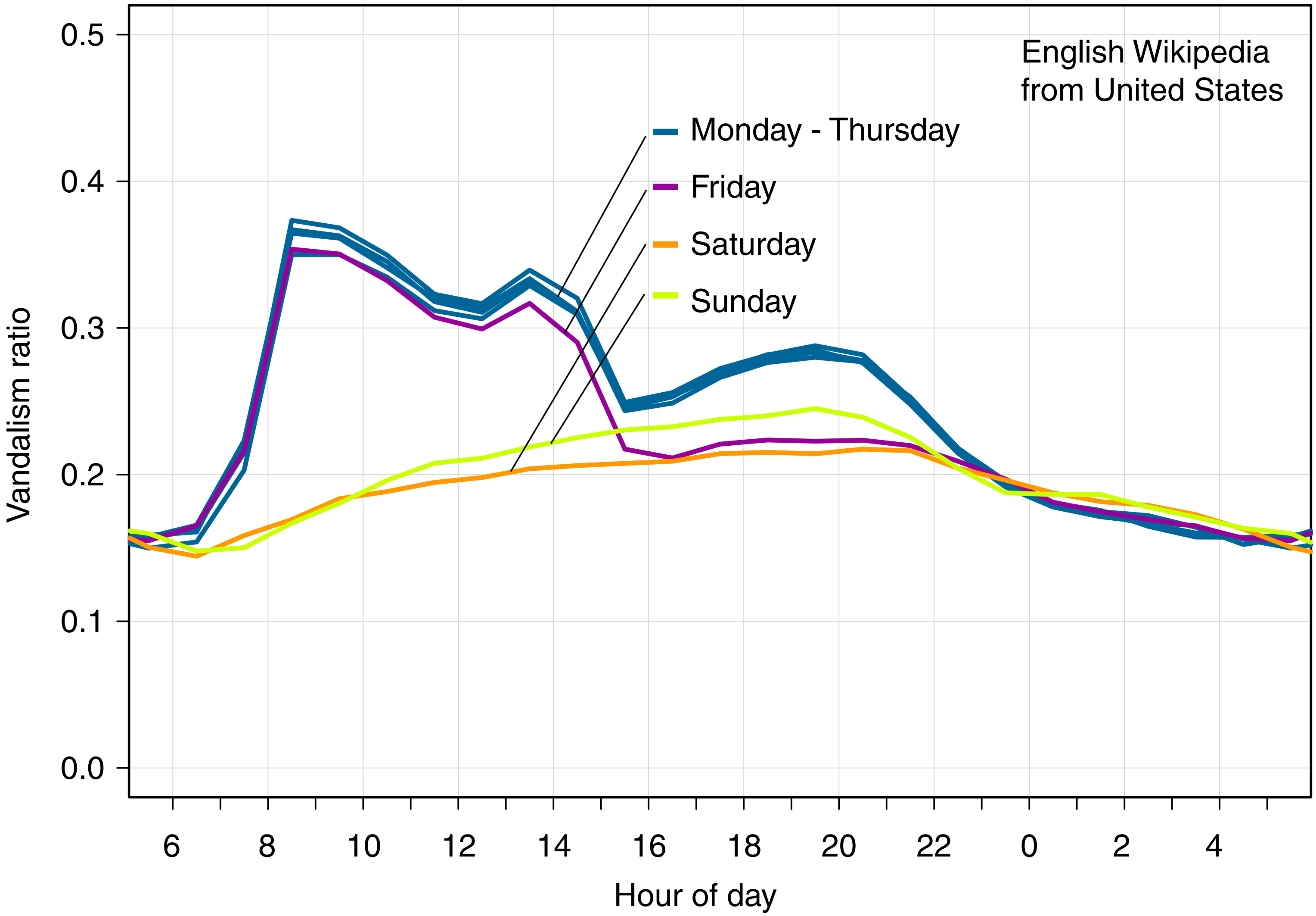
Vandalism in Wikipedia (Geolocated Activity)



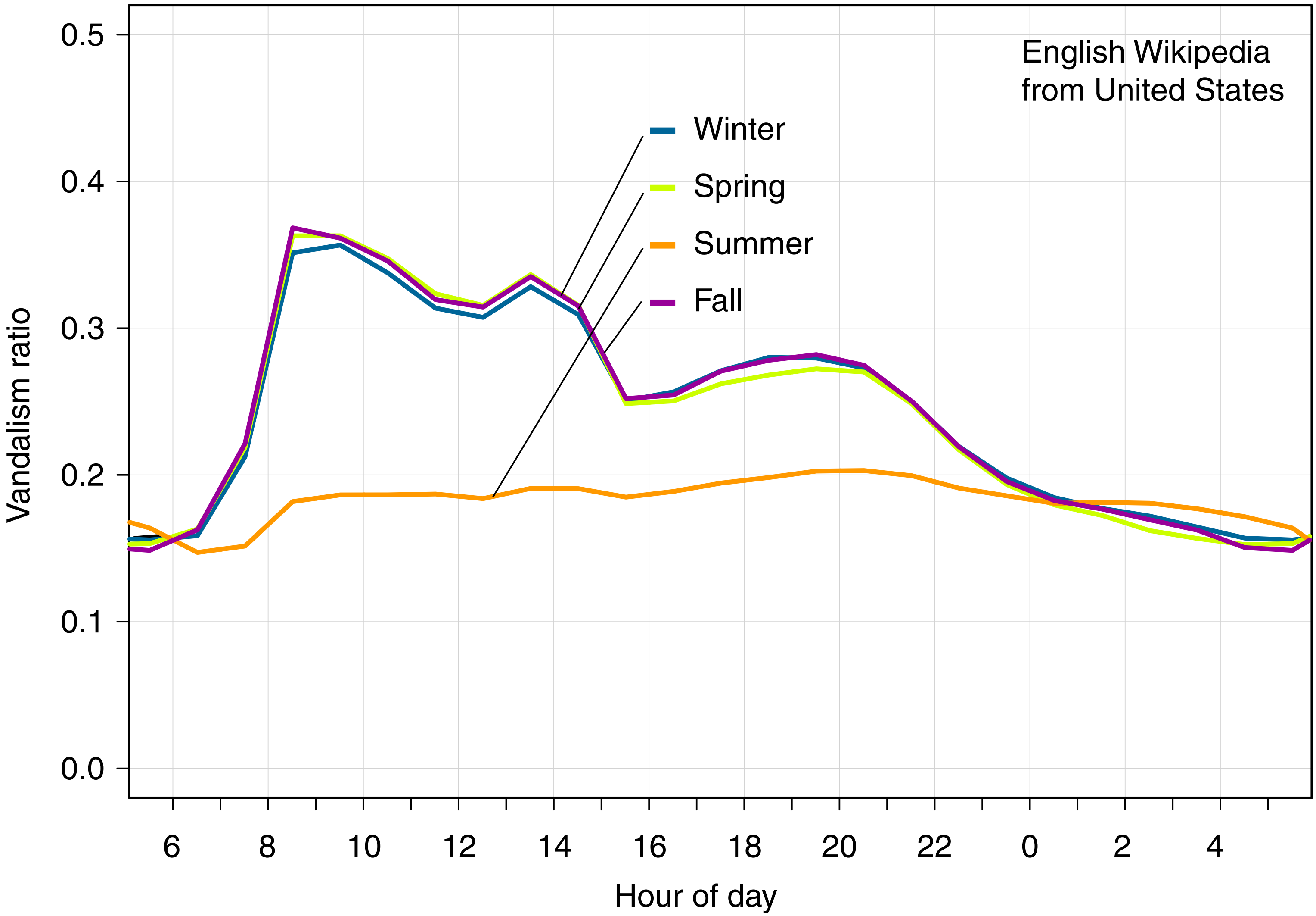
Vandalism in Wikipedia (USA)



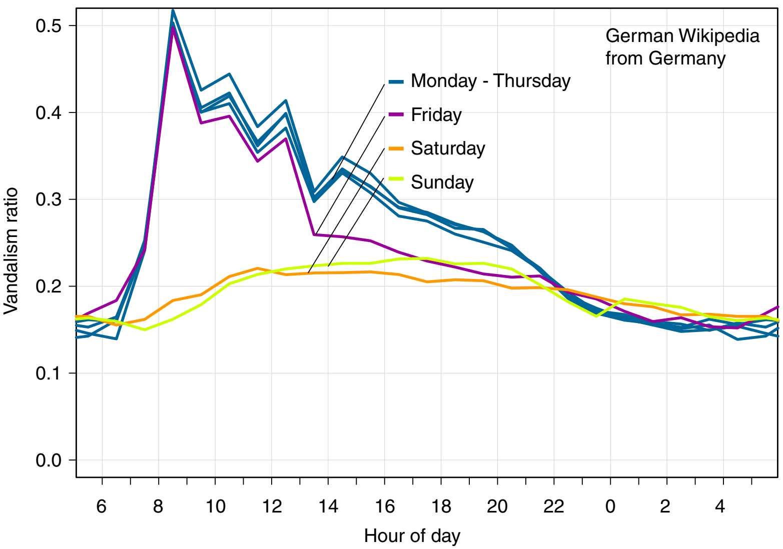
Vandalism in Wikipedia (USA)

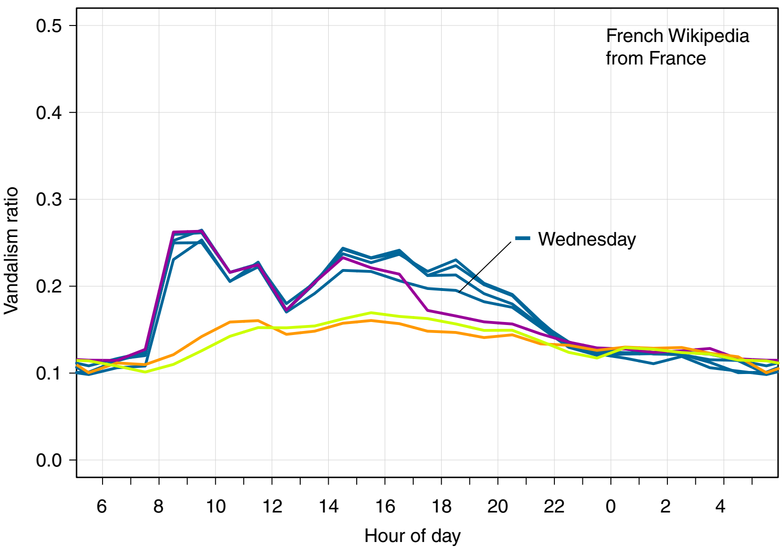


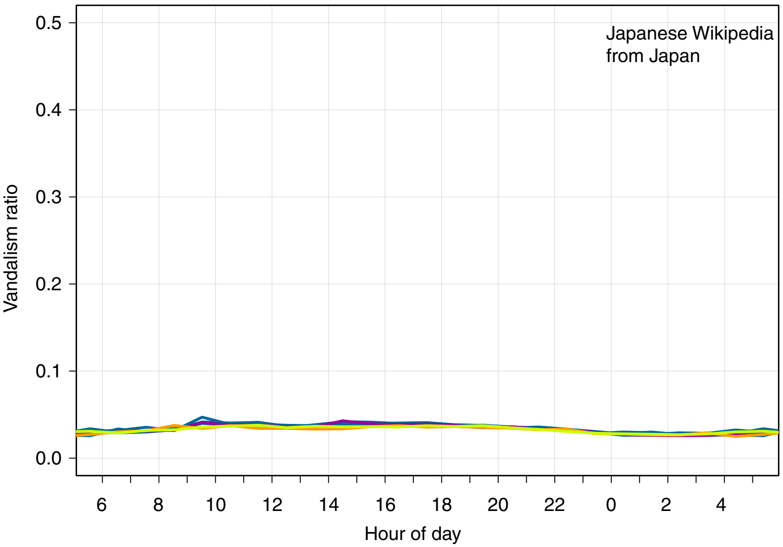
Vandalism in Wikipedia (USA)



Vandalism in Wikipedia (other countries)





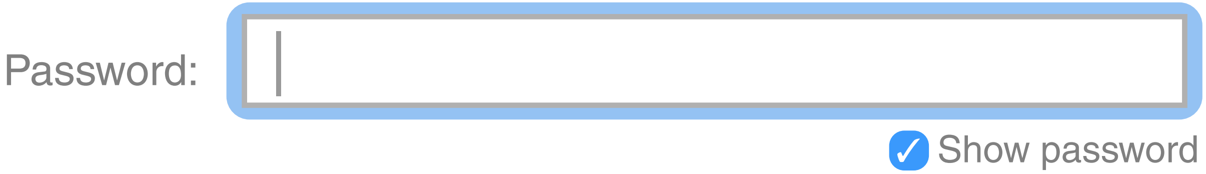


The BSI* Password Creation Advice**

* Bundesamt für Sicherheit in der Informationstechnik.

** J. Kiesel, B. Stein, S. Lucks. A Large-scale Analysis of the Mnemonic Password Advice. 24th Annual Network and Distributed System Security Symposium, NDSS 17, Feb. 2017.

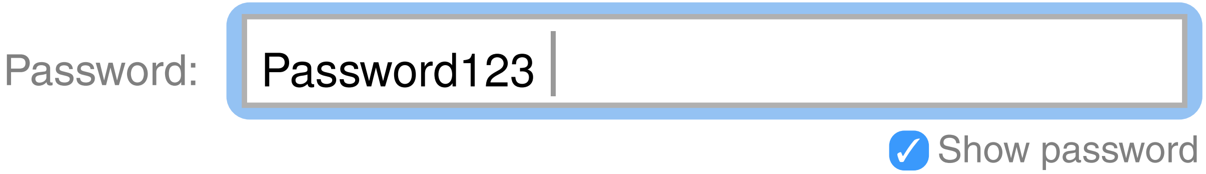
Mnemonic Password Creation



Password:

☑ Show password

Mnemonic Password Creation

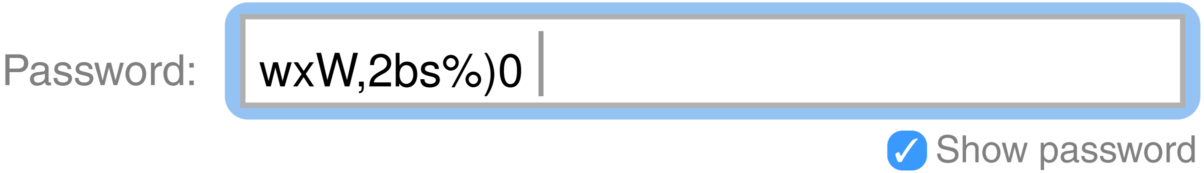


Password: Password123

☑ Show password

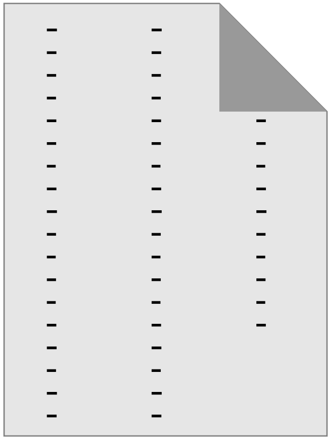
	Ad-hoc guessing
Uniformly distributed chances	10^8

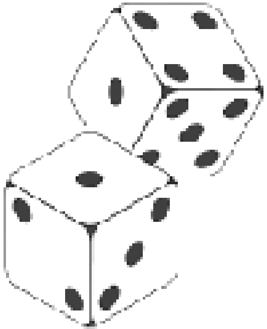
Mnemonic Password Creation



Password: wxW,2bs%)0

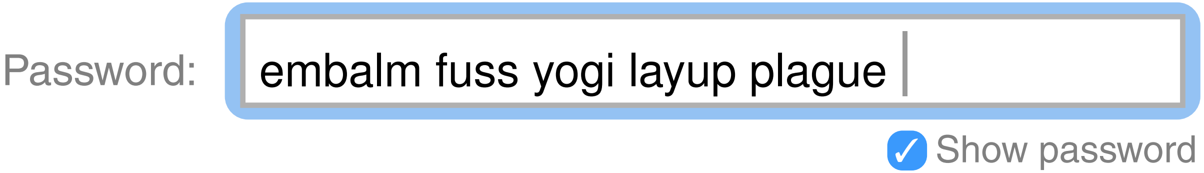
☑ Show password





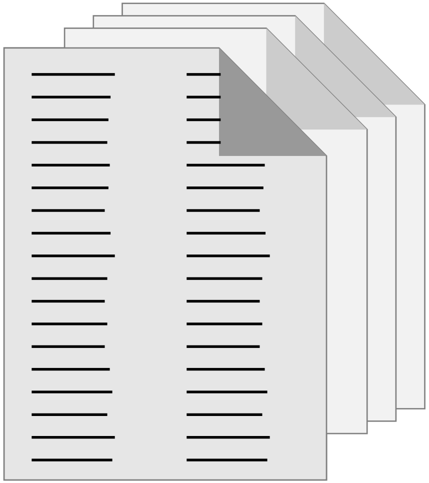
	Ad-hoc guessing	Random characters (10 chars out of 96)
Uniformly distributed chances	10^8	10^{19}

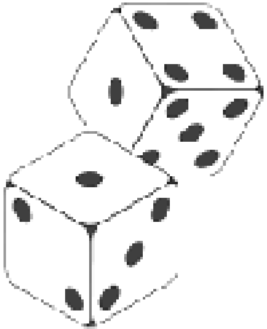
Mnemonic Password Creation



Password: embalm fuss yogi layup plague

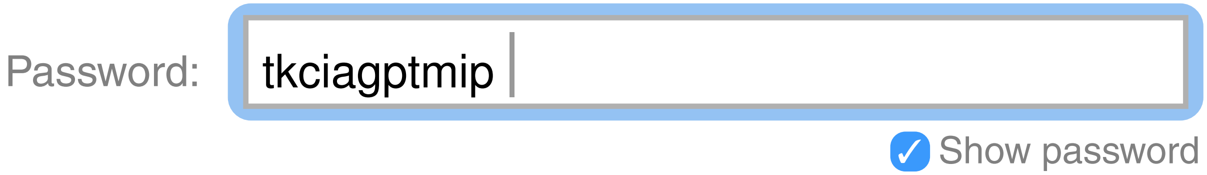
☑ Show password





	Ad-hoc guessing	Random characters (10 chars out of 96)	Random words (5 words out of 7776)
Uniformly distributed chances	10^8	10^{19}	10^{19}

Mnemonic Password Creation



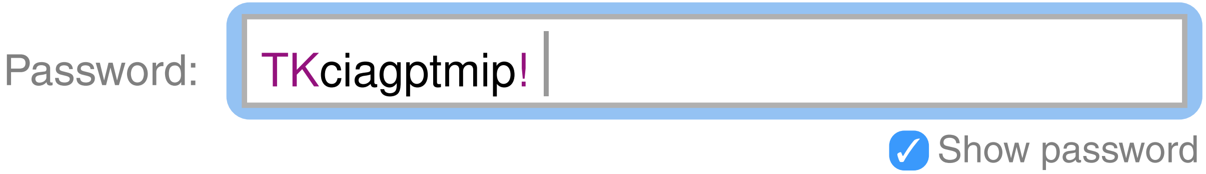
Password: tkciagptmip

☑ Show password

“The KSWD conference is a great place to meet interesting people!”

	Ad-hoc guessing	Random characters (10 chars out of 96)	Random words (5 words out of 7776)	Mnemonic sentence
Uniformly distributed chances	10^8	10^{19}	10^{19}	?

Mnemonic Password Creation



Password: TKciagptmip!

☑ Show password

“The KSWD conference is a great place to meet interesting people!”

	Ad-hoc guessing	Random characters (10 chars out of 96)	Random words (5 words out of 7776)	Mnemonic sentence
Uniformly distributed chances	10^8	10^{19}	10^{19}	?

Correlation and Frequency in Natural Language

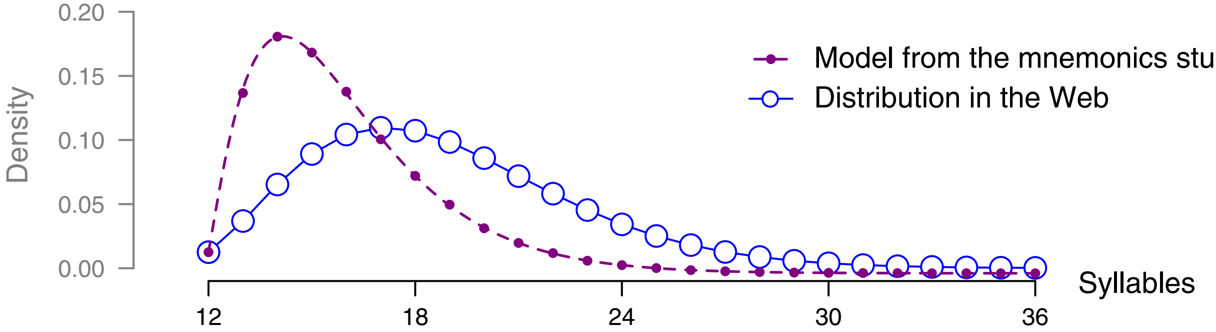
		successor																					
		a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
predecessor	a	.0972	.0380	.0527	.0305	.0265	.0439	.0205	.0471	.0575	.0053	.0048	.0334	.0453	.0234	.0498	.0458	.0028	.0295	.0756	.1561	.0106	.0086
	b	.1304	.0308	.0435	.0301	.0221	.0337	.0139	.0527	.0682	.0057	.0041	.0187	.0363	.0185	.0693	.0333	.0018	.0262	.0654	.1630	.0146	.0048
	c	.1233	.0549	.0368	.0253	.0190	.0418	.0122	.0405	.0702	.0041	.0026	.0155	.0261	.0161	.1061	.0279	.0011	.0207	.0512	.1144	.0124	.0044
	d	.1154	.0449	.0362	.0188	.0163	.0400	.0097	.0400	.0741	.0045	.0059	.0153	.0254	.0393	.0942	.0289	.0014	.0174	.0497	.1405	.0101	.0035
	e	.1115	.0385	.0406	.0242	.0207	.0432	.0130	.0320	.0817	.0030	.0039	.0164	.0311	.0134	.1072	.0354	.0013	.0203	.0570	.1370	.0091	.0048
	f	.1291	.0345	.0397	.0250	.0268	.0330	.0139	.0479	.0626	.0044	.0026	.0186	.0386	.0121	.0721	.0340	.0013	.0214	.0544	.1884	.0090	.0054
	g	.1355	.0417	.0405	.0282	.0184	.0368	.0116	.0484	.0656	.0049	.0037	.0166	.0356	.0129	.0852	.0319	.0012	.0221	.0564	.1269	.0178	.0055
	h	.1023	.0751	.0489	.0338	.0241	.0431	.0173	.0671	.0477	.0049	.0072	.0285	.0351	.0260	.0495	.0343	.0014	.0270	.0792	.0993	.0087	.0056
	i	.1215	.0289	.0441	.0282	.0211	.0303	.0152	.0495	.0697	.0043	.0048	.0182	.0337	.0268	.0517	.0318	.0017	.0225	.0582	.2051	.0087	.0056
	j	.1456	.0583	.0524	.0291	.0155	.0408	.0214	.0524	.0427	.0117	.0078	.0233	.0350	.0117	.0505	.0272	.0019	.0252	.0621	.0990	.0058	.0039
	k	.1201	.0311	.0311	.0207	.0145	.0290	.0104	.0766	.0704	.0062	.0062	.0145	.0248	.0145	.1139	.0166	.0021	.0166	.0497	.1470	.0104	.0041
	l	.1425	.0392	.0365	.0254	.0187	.0401	.0138	.0450	.0677	.0062	.0031	.0209	.0312	.0151	.0864	.0316	.0022	.0200	.0535	.1332	.0151	.0053
	m	.1169	.0606	.0428	.0282	.0218	.0388	.0146	.0486	.0598	.0069	.0056	.0242	.0313	.0170	.0933	.0391	.0019	.0236	.0614	.1044	.0114	.0058
	n	.1029	.0485	.0485	.0311	.0301	.0408	.0175	.0417	.0509	.0058	.0078	.0306	.0471	.0165	.0888	.0388	.0034	.0301	.0626	.1033	.0097	.0073
	o	.1042	.0285	.0490	.0245	.0260	.0310	.0158	.0497	.0428	.0056	.0032	.0209	.0422	.0165	.0562	.0398	.0012	.0217	.0607	.2765	.0084	.0068
	p	.1257	.0367	.0439	.0243	.0203	.0428	.0118	.0369	.0787	.0040	.0051	.0152	.0268	.0104	.1207	.0286	.0013	.0195	.0500	.1140	.0126	.0051
	q	.1640	.0317	.0423	.0265	.0212	.0370	.0106	.0317	.0635	.0053	.0000	.0159	.0265	.0159	.1164	.0265	.0000	.0212	.0688	.0794	.0106	.0053
	r	.1236	.0399	.0346	.0217	.0185	.0560	.0101	.0382	.0749	.0032	.0020	.0145	.0250	.0125	.1027	.0258	.0016	.0185	.0483	.1610	.0109	.0064
	s	.1233	.0441	.0392	.0274	.0189	.0408	.0131	.0468	.0669	.0044	.0030	.0195	.0322	.0140	.0988	.0306	.0021	.0201	.0570	.1224	.0126	.0055
	t	.0791	.0496	.0648	.0389	.0307	.0439	.0242	.0489	.0491	.0056	.0071	.0295	.0512	.0220	.0458	.0592	.0028	.0370	.0860	.1107	.0143	.0091
	u	.1245	.0325	.0334	.0171	.0171	.0397	.0108	.0424	.0704	.0036	.0027	.0171	.0226	.0153	.0695	.0262	.0018	.0190	.0875	.1995	.0081	.0063
	v	.1065	.0403	.0452	.0306	.0242	.0452	.0145	.0339	.0677	.0032	.0032	.0274	.0339	.0161	.1242	.0403	.0016	.0242	.0613	.0855	.0097	.0081
	w	.1272	.0510	.0429	.0310	.0210	.0307	.0179	.0774	.0641	.0052	.0050	.0193	.0355	.0336	.0467	.0298	.0016	.0255	.0676	.1474	.0107	.0052
	x	.3333	.0000	.0000	.3333	.0000	.0000	.0000	.0000	.3333	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	y	.1140	.0421	.0706	.0322	.0198	.0335	.0297	.0607	.0446	.0062	.0161	.0273	.0545	.0223	.0496	.0310	.0012	.0161	.0582	.0855	.0074	.0050
	z	.1667	.0556	.0556	.0000	.0000	.0556	.0556	.0000	.0556	.0000	.0000	.0000	.0556	.0000	.0556	.0556	.0000	.0000	.0556	.1111	.0000	.0000

➜ Position-dependent, higher-order language model learning on Big data.

Challenge: Building a Corpus for Mnemonic Analyses

Q. What characterizes sentences that humans use for mnemonic passwords?

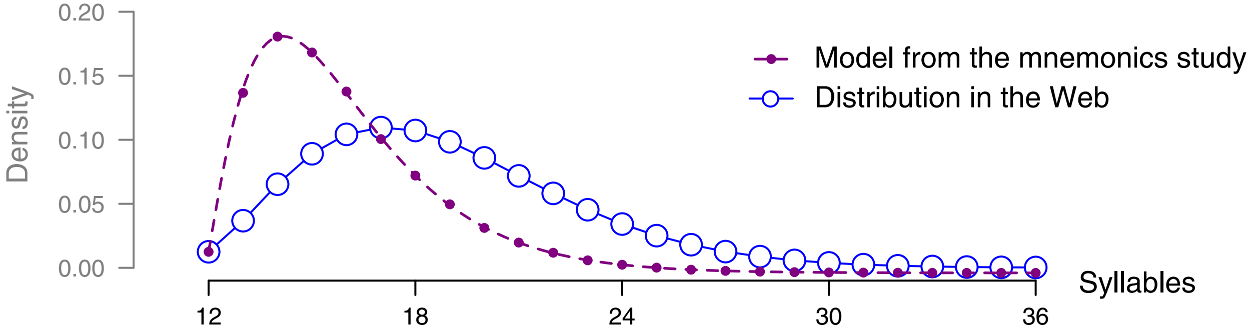
A. Ask 1,000 people via a study ⇝ language complexity, readability, lengths



Challenge: Building a Corpus for Mnemonic Analyses

Q. What characterizes sentences that humans use for mnemonic passwords?

A. Ask 1,000 people via a study ⤳ language complexity, readability, lengths



Q. How many sentences do we need?

A. The more the better: training 7th order model requires 10^8 examples

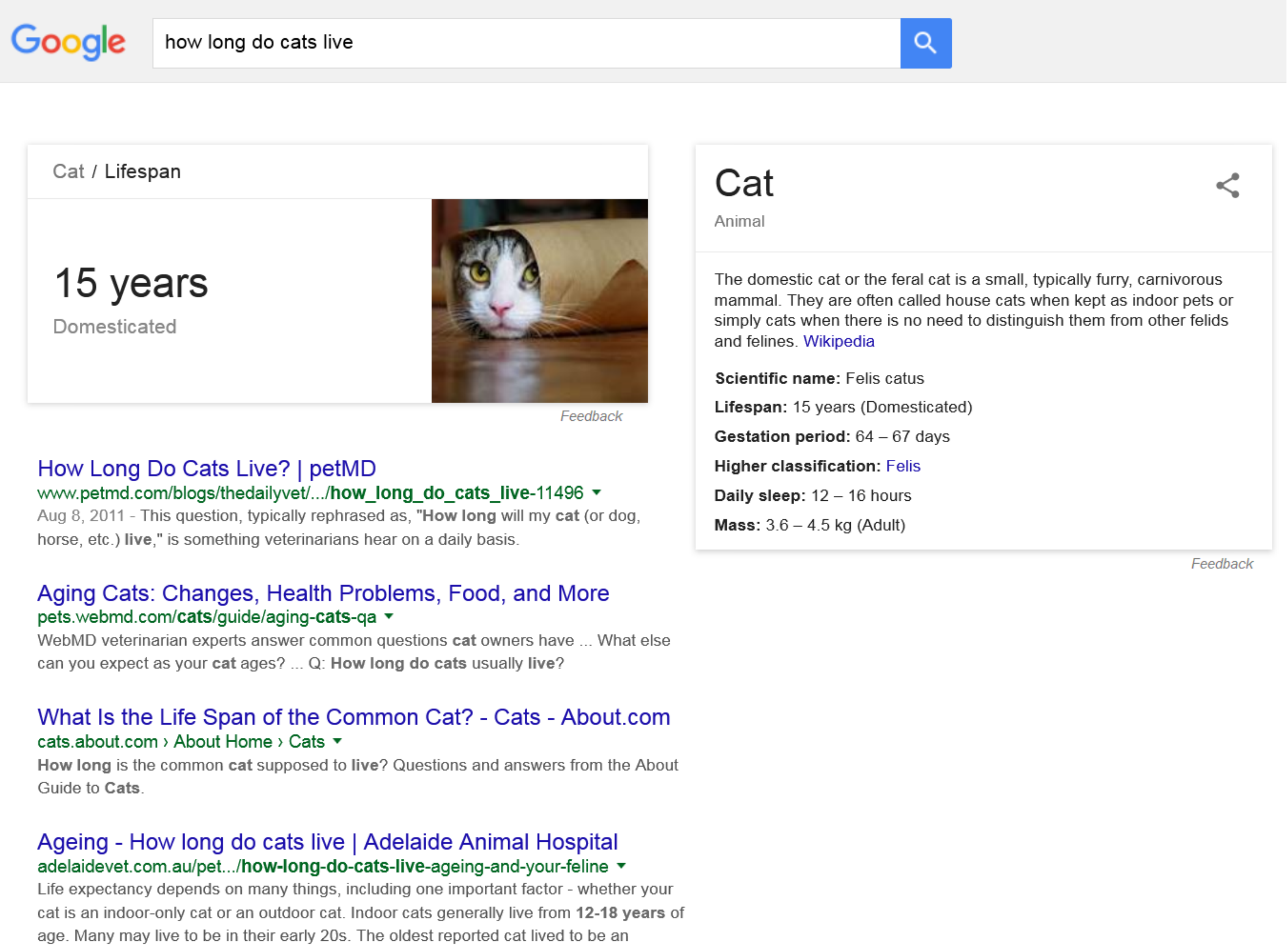
≈ 80,000	Sentences	The Bible
≈ 5,000,000	Sentences	Encyclopedia Britannica
730,000,000	Web pages	ClueWeb12, 27.3 TB
3,400,000,000	Sentences	extracted and filtered
25,000,000,000	Passwords	18 password generation rules

Mnemonic Password Creation: Selected Results

	Ad-hoc guessing	Random characters	Random sentence	Mnemonic sentence
		(10 chars out of 96)	(5 words out of 7776)	(9.5 words)
Uniformly distributed chances	10^8	10^{19}	10^{19}	10^{11}

- Large character sets (%, !, @, #, etc.) add only 3.3 Bit $(\cdot 10^{1.0})$
- Word replacements (for → 4, at → @) add only 1.6 Bit $(\cdot 10^{0.5})$
- Complex sentences (rich vocabulary) add only 1.2 Bit $(\cdot 10^{0.4})$
- Different generation rules often decrease the entropy.

Outlook: The “Why” Search Engine



how long do cats live

Cat / Lifespan

15 years

Domesticated

Feedback

How Long Do Cats Live? | petMD
www.petmd.com/blogs/thedailyvet/.../**how_long_do_cats_live**-11496 ▾
Aug 8, 2011 - This question, typically rephrased as, "**How long** will my **cat** (or dog, horse, etc.) **live**," is something veterinarians hear on a daily basis.

Aging Cats: Changes, Health Problems, Food, and More
pets.webmd.com/**cats**/guide/aging-**cats**-qa ▾
WebMD veterinarian experts answer common questions **cat** owners have ... What else can you expect as your **cat** ages? ... Q: **How long do cats** usually **live**?

What Is the Life Span of the Common Cat? - Cats - About.com
cats.about.com › About Home › Cats ▾
How long is the common **cat** supposed to **live**? Questions and answers from the About Guide to **Cats**.

Ageing - How long do cats live | Adelaide Animal Hospital
adelaidevet.com.au/pet.../**how-long-do-cats-live**-ageing-and-your-feline ▾
Life expectancy depends on many things, including one important factor - whether your cat is an indoor-only cat or an outdoor cat. Indoor cats generally live from **12-18 years** of age. Many may live to be in their early 20s. The oldest reported cat lived to be an

Cat

Animal

The domestic cat or the feral cat is a small, typically furry, carnivorous mammal. They are often called house cats when kept as indoor pets or simply cats when there is no need to distinguish them from other felids and felines. Wikipedia

Scientific name: Felis catus

Lifespan: 15 years (Domesticated)

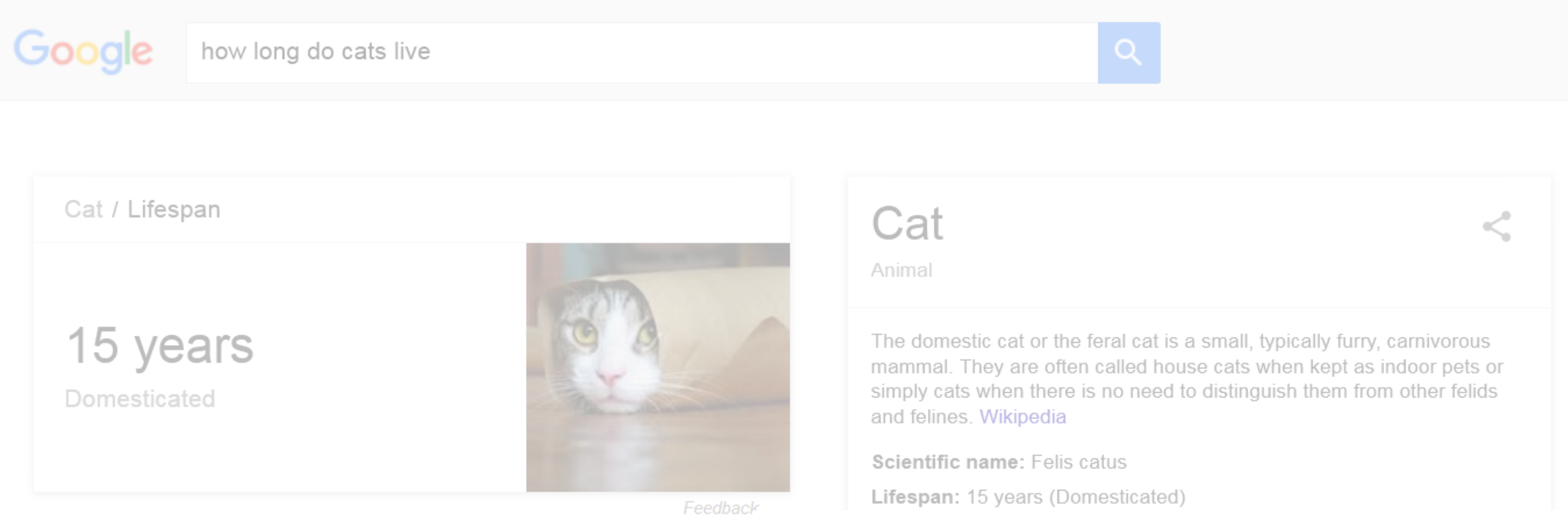
Gestation period: 64 – 67 days

Higher classification: Felis

Daily sleep: 12 – 16 hours

Mass: 3.6 – 4.5 kg (Adult)

Feedback



how long do cats live

Cat / Lifespan

15 years

Domesticated

Cat

Animal

The domestic cat or the feral cat is a small, typically furry, carnivorous mammal. They are often called house cats when kept as indoor pets or simply cats when there is no need to distinguish them from other felids and felines. Wikipedia

Scientific name: Felis catus

Lifespan: 15 years (Domesticated)

Feedback

How Long Do Cats Live? | petMD
www.petmd.com/blogs/thedailyvet/.../how_long_do_cats_live-11496 ▾
Aug 8, 2011 - This question, typically rephrased as, "**How long** will my **cat** (or d
horse, etc.) **live**," is something veterinarians hear on a daily basis.

Aging Cats: Changes, Health Problems, Food, and More
pets.webmd.com/**cats**/guide/aging-**cats**-qa ▾
WebMD veterinarian experts answer common questions **cat** owners have ... Wh
can you expect as your **cat** ages? ... Q: **How long do cats** usually **live**?

What Is the Life Span of the Common Cat? - Cats - About
cats.about.com › About Home › Cats ▾
How long is the common **cat** supposed to **live**? Questions and answers from th
Guide to **Cats**.

Ageing - How long do cats live | Adelaide Animal Hospital
adelaidevet.com.au/pet.../**how-long-do-cats-live**-ageing-and-your-feline
Life expectancy depends on many things, including one important factor - wheth
cat is an indoor-only cat or an outdoor cat. Indoor cats generally live from **12-18**
age. Many may live to be in their early 20s. The oldest reported cat lived to be a

Konrad
Lischka

How does Google know when my cat will die?

23. September 2015 by Konrad Lischka, in Blog @en

How long do cats live? Exactly 15 years says Google.com. Not “10 to 15”, not “about 15 years”, but “15 years”. That sounds like a definitive answer. It’s Google’s answer to the search query “How long do cats live”.

Thank you!

Matthias Hagen

Johannes Kiesel

Henning Wachsmuth

Martin Potthast

Tim Gollub

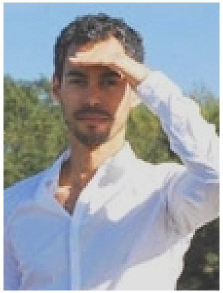
Thank you!

Tsvetomira Palakarska

Nadin Glaser

Michael Völske

Khalid Al-Khatib


Nedim Lipka