

Recommendation Systems and Knowledge Gaps in Wikipedia

Leila Zia



Research Showcase
2017-12-13

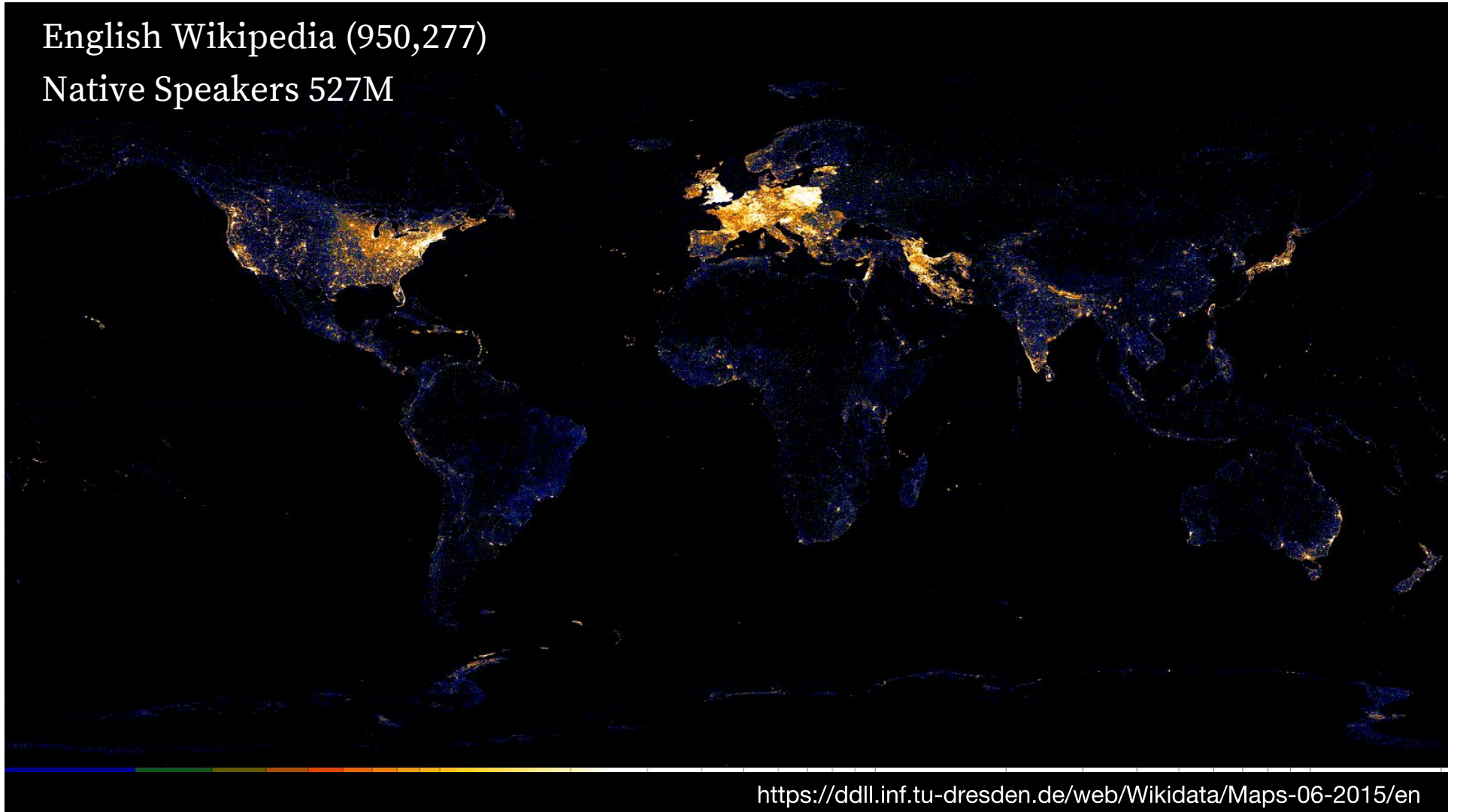
Hi! It's been a while!

Let's rewind.

2015

English Wikipedia (950,277)

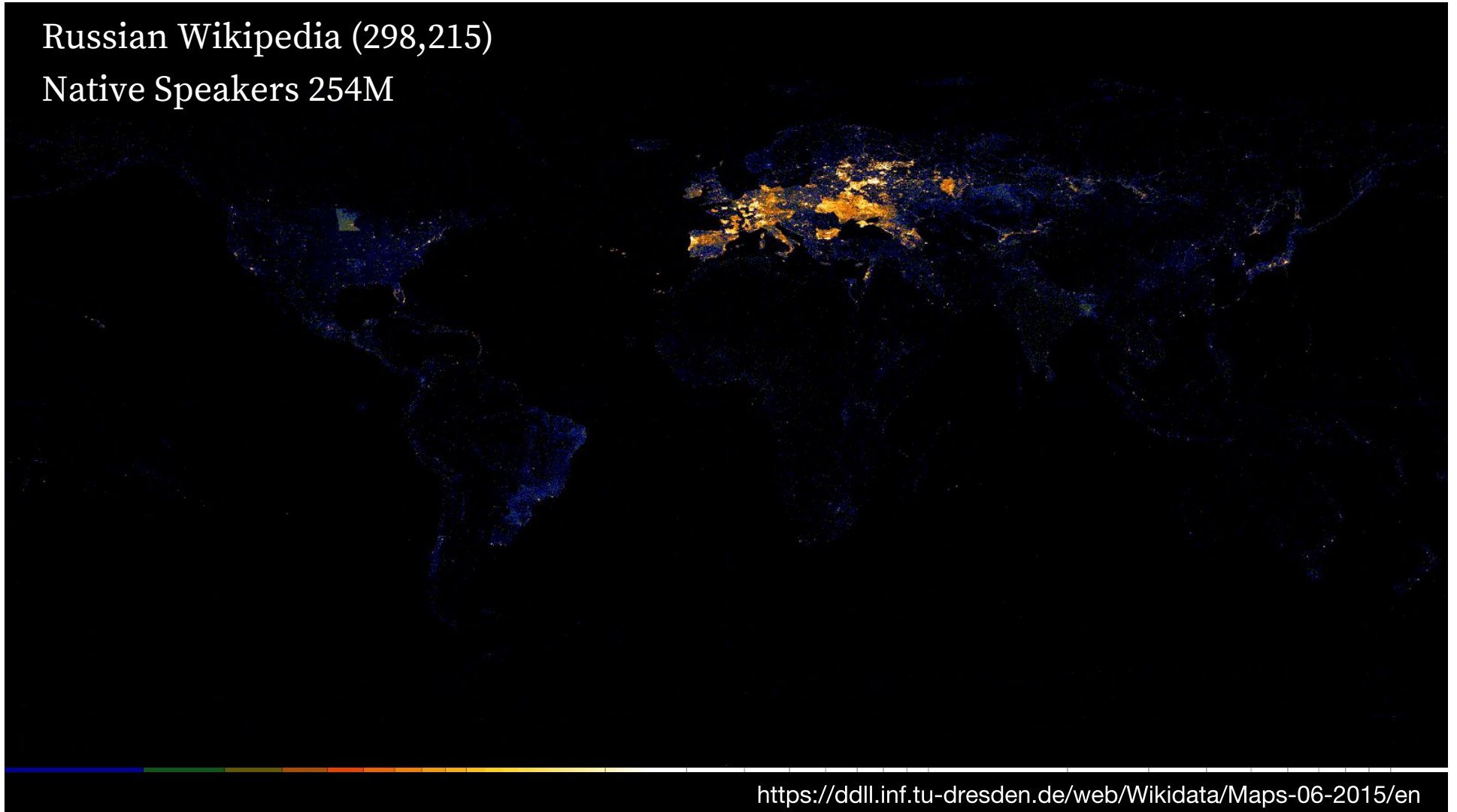
Native Speakers 527M



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

Russian Wikipedia (298,215)

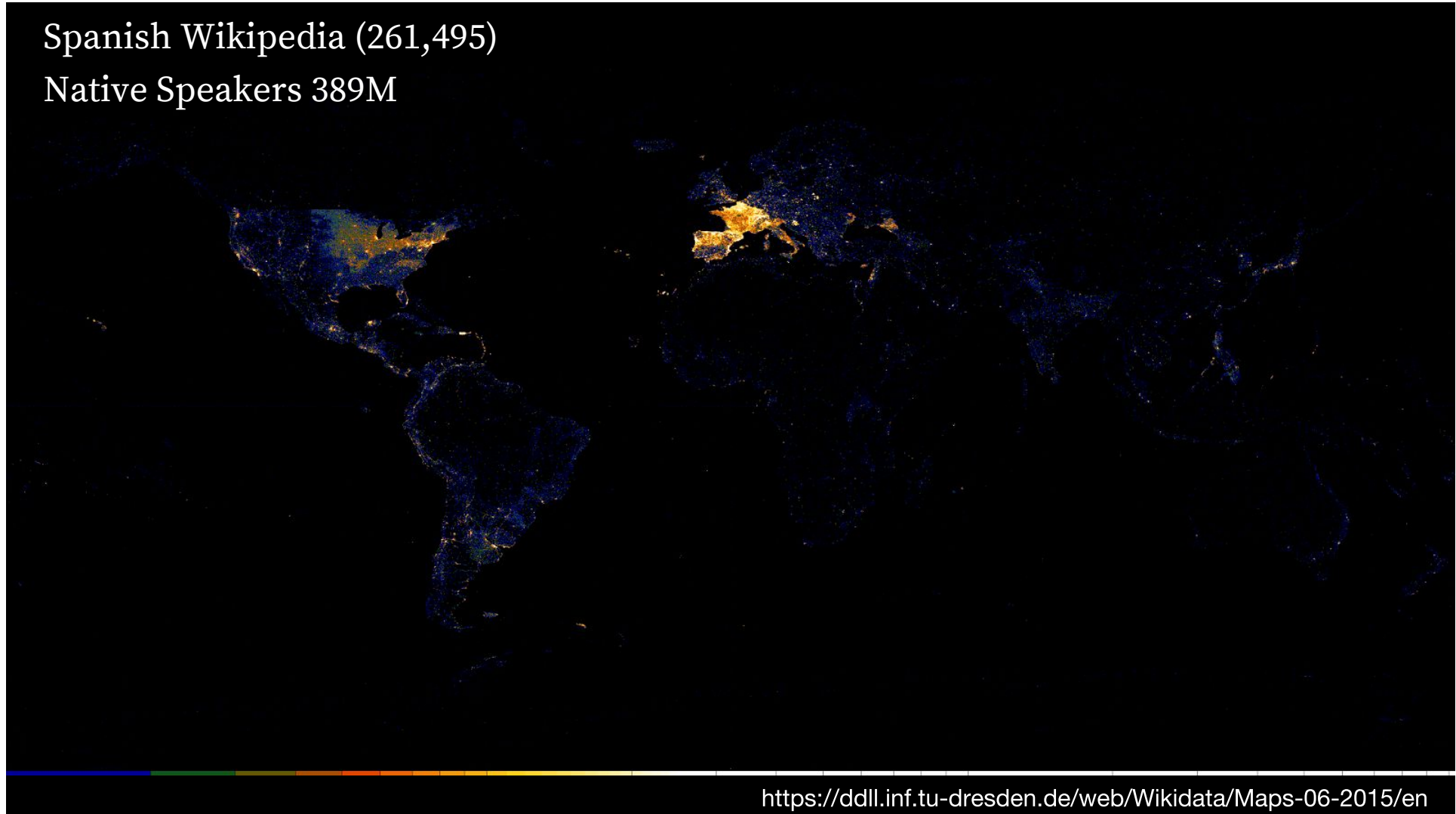
Native Speakers 254M



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

Spanish Wikipedia (261,495)

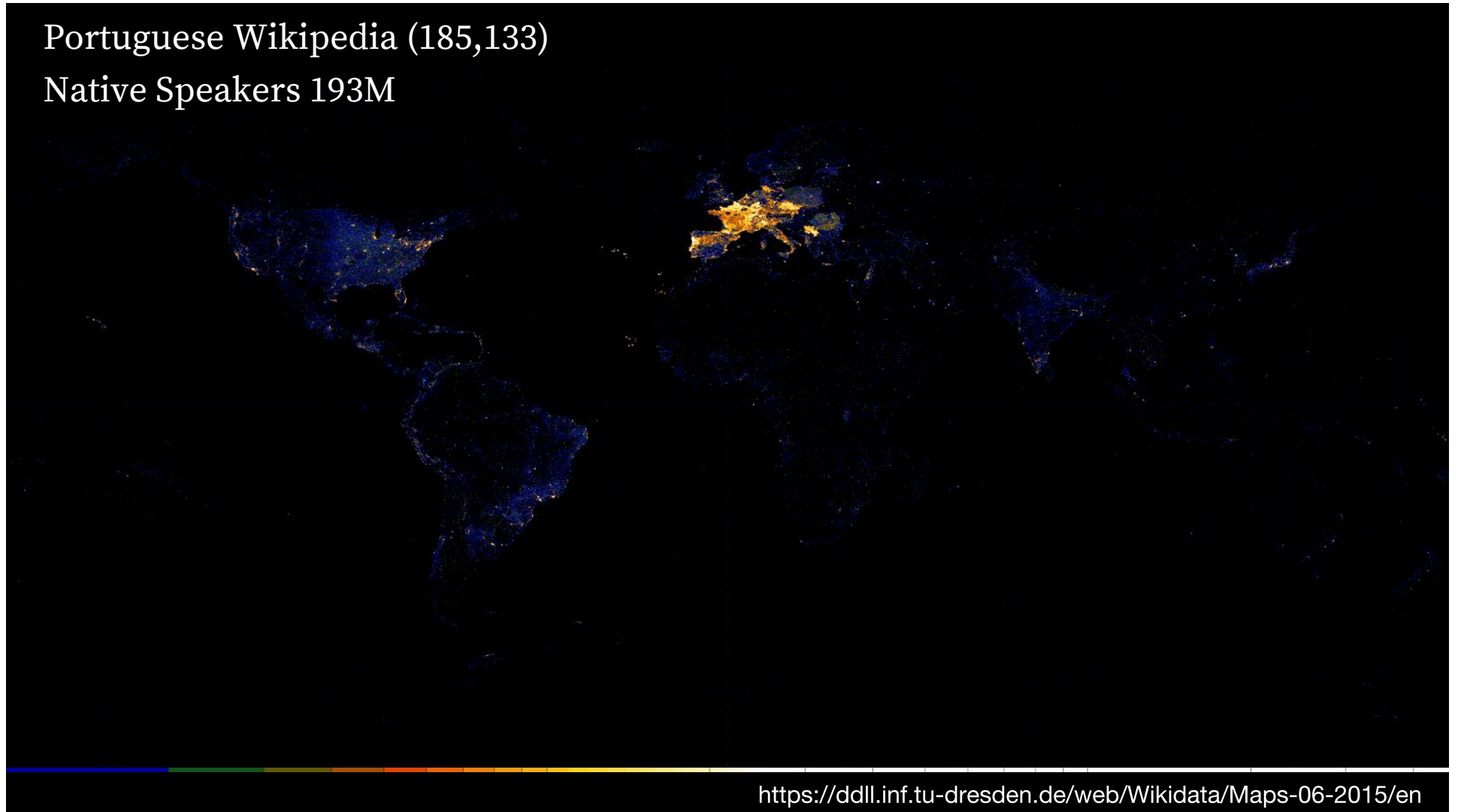
Native Speakers 389M



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

Portuguese Wikipedia (185,133)

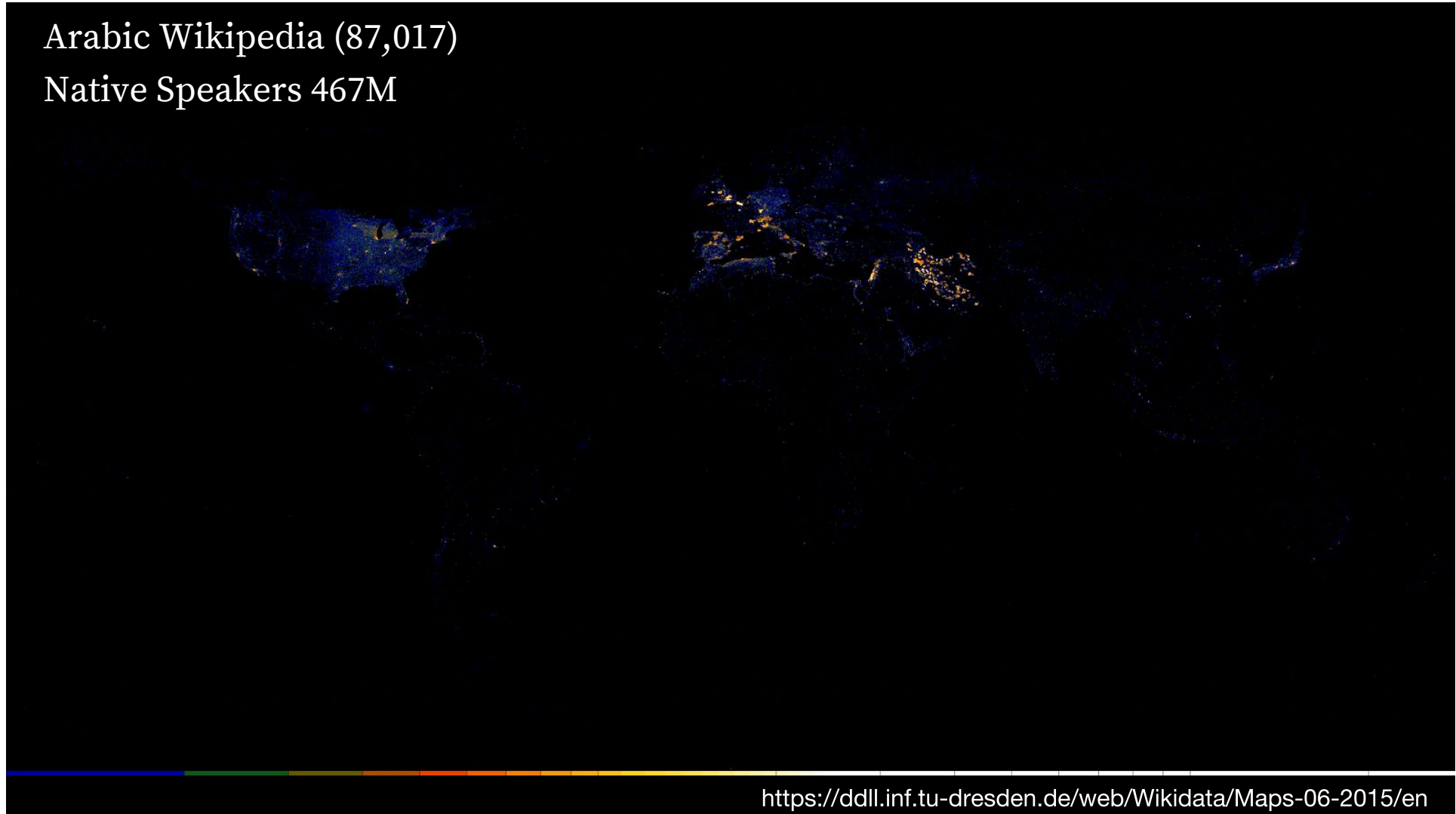
Native Speakers 193M



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

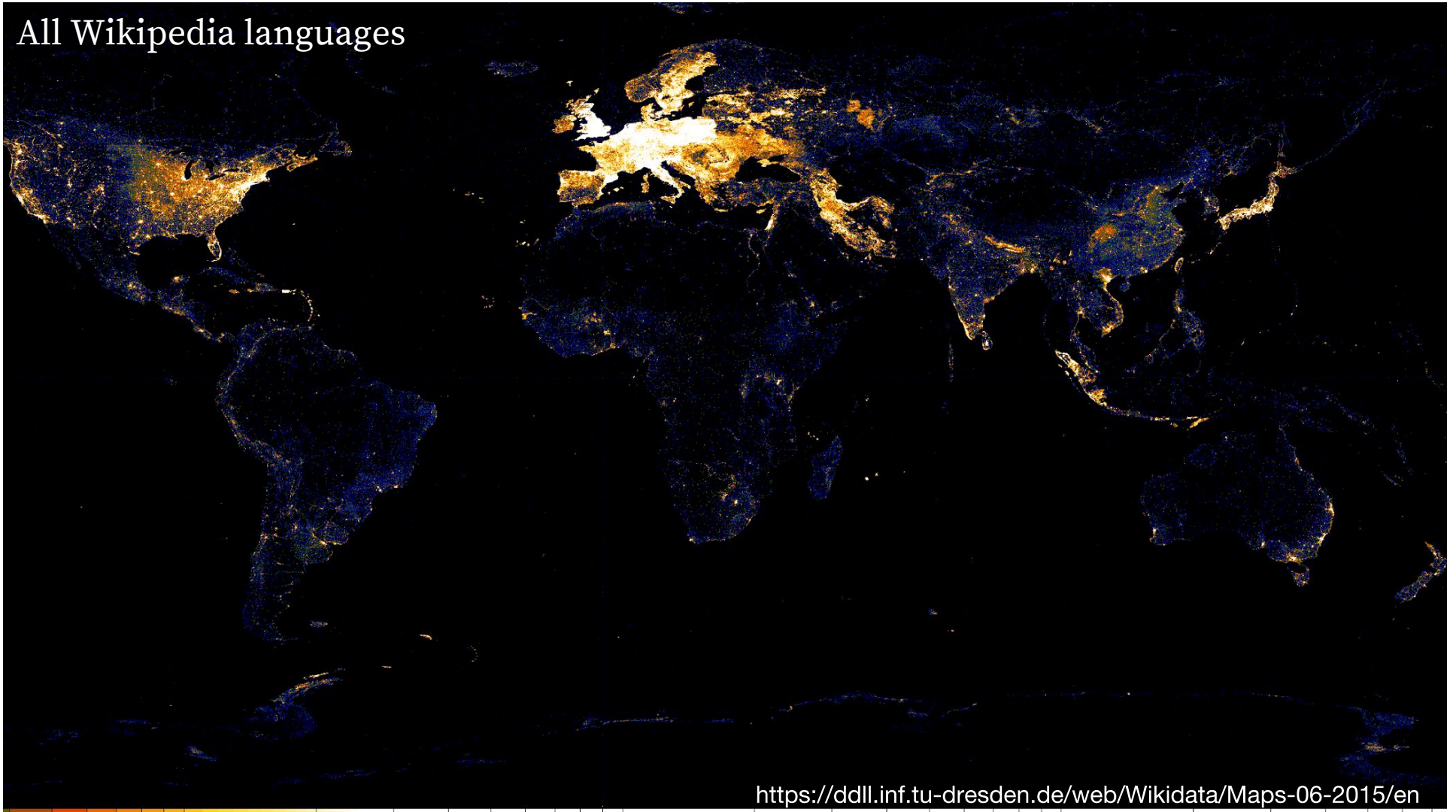
Arabic Wikipedia (87,017)

Native Speakers 467M



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

All Wikipedia languages



<https://dill.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en>

Insights

- We have a lot of missing articles across languages
 - Not all of these missing articles are available in at least one Wikipedia language
- ⇒ Wikipedia alone won't be the solution for all knowledge gaps.

Goal

Increase article coverage in terms of the number of articles in different languages and the contents of the articles within a language by **identifying and prioritizing** missing content and **routing attention** where it's needed.

Growing Wikipedia Across Languages by Article Creation recommendations

[Growing Wikipedia Across Languages via Recommendation](#), Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. International World Wide Web (WWW) Conference, Montréal, Qué., 2016.

The system



Articles



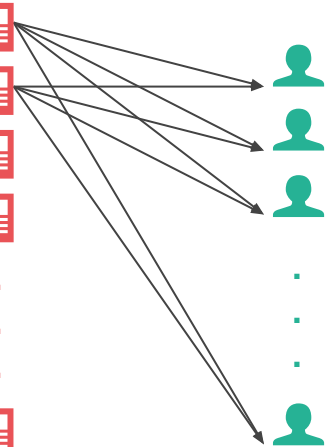
Articles



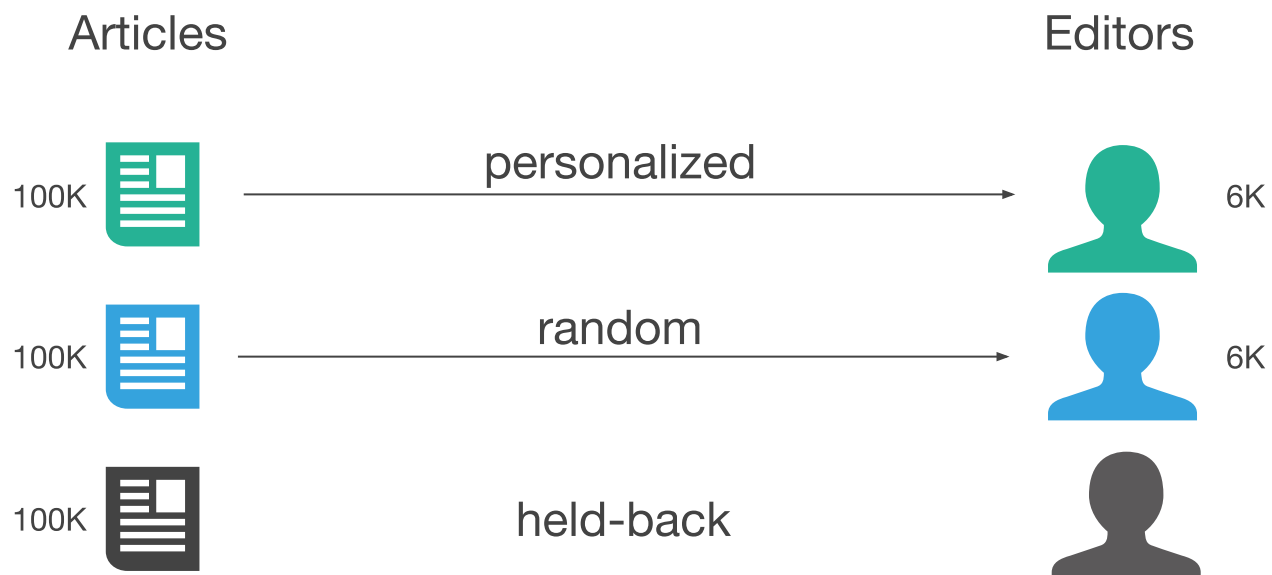
Articles



Editors



Experiment



Results

- **3.2x** growth in article creation rate with no loss in article quality
- **2x** growth in activation rate if the recommended article is personalized for the editor based on their interests

On the ground

- **GapFinder**

<https://www.mediawiki.org/wiki/GapFinder>


- **ContentTranslation** Suggestions
- Wherever else you're using the API :)

GapFinder


 Wikipedia **GapFinder**


English ▾ فارسی ▾

agriculture|





Environmental impac...
agriculture's impact on the environment

5041 recent views 




Crop rotation

10836 recent views 




Monoculture

5068 recent views 




Intensive animal far...

6039 recent views 



Agricultural wastewa...

1568 recent views 




Integrated pest man...

7365 recent views 




Renewable resource
a natural resource which can replenish with the passage of time,

12915 recent views 



Environmental impac...

5749 recent views 

Crop rotation



"Fallow" redirects here. For other uses, see *Fallow (disambiguation)*.

 This article includes a list of references, but **its sources remain unclear** because it has **insufficient inline citations**. Please help to improve this article by **introducing** more precise citations. (April 2009)
(Learn how and when to remove this template message)

Crop rotation is the practice of growing a series of dissimilar or different types of *crops* in the same area in sequenced *seasons*. It is done so that the *soil* of farms is not used for only one set of nutrients. It helps in reducing *soil erosion* and increases *soil fertility* and *crop yield*.

Growing the same *crop* in the same place for many years in a row disproportionately depletes the *soil* of certain *nutrients*. With rotation, a crop that leaches the soil of one kind of nutrient is followed during the next growing season by a dissimilar crop that returns that nutrient to the soil or draws a different ratio of nutrients. In addition, crop rotation mitigates the buildup of *pathogens* and pests that often occurs when one species is continuously cropped, and can also improve *soil structure* and *fertility* by increasing biomass from varied *root structures*.

Crop cycle is used in both conventional and *organic farming* systems.



Agriculture

History

- History of organic farming
- Arab Agricultural Revolution
- British Agricultural Revolution
- Green Revolution
- Neolithic Revolution

On land

- Animal husbandry (cattle · pig · poultry · sheep) · Dairy · Dryland · Extensive · Free-range · Grazing · Hobby · Intensive (animal · crop) · Natural · Orchard · Organic · Ranching · Sharecropping · Slash-and-burn

In water

- Aquaculture · Aquaponics · Hydroponics

Related

- Agribusiness · Agricultural engineering · Agricultural science · Agroecology · Agroforestry · Agronomy · Animal-free · Crop diversity · Ecology · Livestock · Mechanisation · Permaculture · Sustainable · Urban

Lists

- Government ministries

Contents	
1	History
1.1	Two-field system
1.2	Three-field system
1.3	Four-field rotation
1.4	Modern developments
2	Crop choice
2.1	Row crops
2.2	Legumes
2.3	Grasses and cereals
2.4	Green manure
3	Planning a rotation
4	Implementation
4.1	Incorporation of livestock
4.2	Organic farming
4.3	Intercropping
5	Benefits
5.1	Soil organic matter
5.2	Carbon sequestration
5.3	Nitrogen fixing
5.4	Pathogen and pest control
5.5	Weed management
5.6	Preventing soil erosion



Create from scratch

Translate

Climate change in Malagasy

Wikipedia **GapFinder**

English ▾

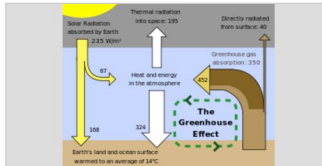
Malagasy ▾

climate change



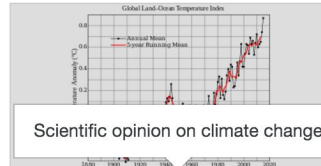
Intergovernmental P...
scientific intergovernmental body

10546 recent views



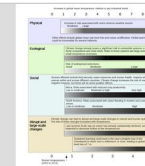
Greenhouse gas
gas in an atmosphere that
absorbs and emits radiation within

51057 recent views



Scientific opinion on ...

8047 recent views



Effects of global war...

22475 recent views



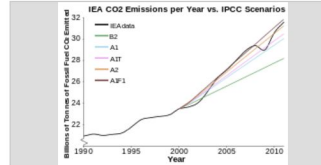
Climate change
significant time variation in long-
term weather patterns

0 recent views



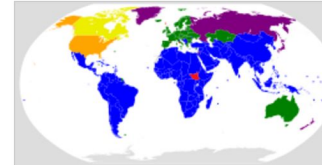
Tropical cyclone
storm system

27758 recent views



Climate change mitig...
actions to limit climate change in
order to reduce the risks of global

4595 recent views



Kyoto Protocol
International Treaty to reduce
greenhouse gas emissions

37088 recent views

2016 to now

A series of revelations

Stubs

We were reminded, over and over, that there are many articles already in Wikipedia that are stubs.

“We have enough low-quality stub articles that need human effort to improve and we're not really interested in more unless either (a) they demonstrably combat some of the systematic biases we're struggling with or (b) they demonstrably attract new cohorts users to do that improvement.” -- stuart

Onboarding newcomers

- We learned about **The Africa Destubathon** project.
- We reached out to User:Anthere to see what kind of research would help them in this kind of project. We learned:

⇒ A lot of manual work is going into building **templates** to help newcomers learn *how* to expand an already existing article in Wikipedia


Existing initiatives

Secure | https://www.wikidaheim.at

WikiDaheim

Gemeinde hier suchen...

Stell deine Heimat in der Wikipedia vor!




WikiDaheim ist ein Projekt von Freiwilligen der Wikimedia-Projekte wie Wikipedia, das sich mit dem Sammeln von Informationen über Gemeinden in ganz Österreich beschäftigt. Gerade in Österreich sind die Möglichkeiten, den Charakter eines Ortes mit Bildern oder Texten zu zeigen, noch kaum erschlossen worden.

Ergebnisse des Fotowettbewerbs

Die [Gewinner des Fotowettbewerbs 2017](#) stehen fest! Vielen Dank an alle Teilnehmer des Wettbewerbs! Die Preisverleihung findet am 24. Jänner 2018 um 17 Uhr im Ahnensaal in der Wiener Hofburg statt.

Das Siegerbild stammt von Friedrich Böhlinger aus Vorarlberg:



© Böhlinger Friedrich, Madeldorferstraße 9 Meiningen, Interior 03. CC BY-SA 3.0 AT

https://ma-commune.wikipédia.fr/Q21965/

Ma Commune Wikipédia


Mainvillers

🕒 Dernière modification de la page : 12/11/2017, 3:52:20 AM
📄 Taille de la page : 5 632 octets
👤 Nombre de contributeurs enregistrés : 50
👤 Nombre de contributeurs anonymes : 8

Wikipédia
Wikimedia Commons
Wikidata

Illustrations

L'article compte actuellement 4 images, dont :



Importer les vôtres !

Idées d'amélioration

- 🔗 Votre aide est la bienvenue pour corriger les liens présents dans l'article, vers les pages d'homonymie Cure → Quelques explications pour effectuer ces corrections. -- 20 octobre 2015 à 10:26 (CET) [Voir sur Wikipédia](#)

Avancement

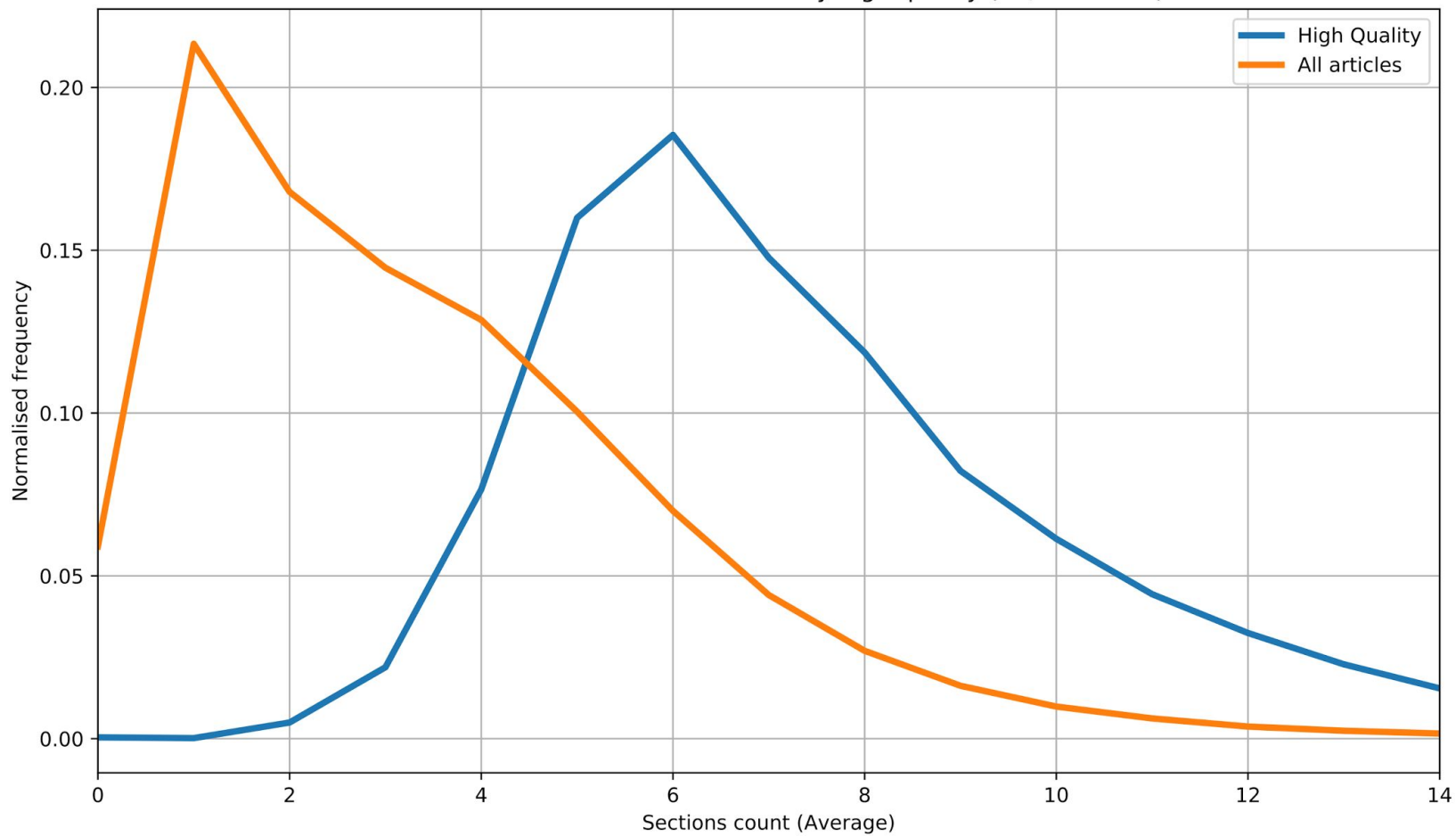
🕒 Dernière vérification de l'avancement il y a 28 jours.

- Urbanisme**
Beaucoup d'informations peuvent manquer
La section est très brève par rapport aux autres articles comparables. N'hésitez pas à la compléter.
 - Conseils de rédaction pour la section UrbanismeTaille : 0 octets (0% de la moyenne)
- Économie**
Beaucoup d'informations peuvent manquer
- Culture locale et patrimoine**
Beaucoup d'informations peuvent manquer
- Politique et administration**
Beaucoup d'informations peuvent manquer
- Population et société**
Beaucoup d'informations peuvent manquer
- Géographie**
Beaucoup d'informations peuvent manquer

The numbers

- **Stubs:** 37% of enwiki articles have a stub template
- **Good or better:** Only 1% of enwiki articles have quality class label of Good or better
- **Inconsistency:** 80% of the sections created in enwiki are used only in one article
- **Potential:** 14K accounts get created every month

Sections count for: all articles vs. only high quality (FA, GA and A)



Take-away

- We have **a lot of missing content** in Wikipedia articles
- Editathon organizers are spending **a lot of time** on manually creating templates that can help onboard newcomers
- There are initiatives that have identified these needs and are working on this problem and the challenge for them is how to **scale** the recommendations beyond very specific topics/types of articles

Let's imagine!

Imagine

... we have a system that breaks down the article structure into different pieces (lead paragraph, infobox, sections, images, links, references, ...), provides recommendation for each of these pieces to the editors, and provides more in-depth information about the pieces (such as: “early life section in biographies in enwiki on average contains x characters.”).

**Let's start
with sections**

Growing Wikipedia Across Languages by Section Recommendations

https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages, Tiziano Piccardi, Michele Catasta, Diego Saez-Trumper, Robert West, Leila Zia

← <https://en.wikipedia.org/wiki/Sanandaj>



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes

Article **Sanandaj** Talk

From Wikipedia, the free encyclopedia

For the administrative subdivision, see Sanandaj County.

Sanandaj pronunciation (help·info) (Persian: سنندج) is a city with a population of 373,987 (2016 census), its population was 373,987 [1] and is the capital of **Kordestan** province at Iran. Sanandaj is the twenty-third largest city in Iran. Sanandaj is a city founded about 200 years ago, yet under its name it has grown to become a center of Kurdish culture.

Contents [hide]

- Society
- Famous people connected to Sanandaj
- References
- External links

Intra-language Information extraction

← <https://en.wikipedia.org/wiki/Tehran>

Arabic: طهران
Aragonés: Asturiano
Asturianu
Azərbaycanca
Башҡортса
Boarisch
Žemaitėška
Bikol Central
Беларуская
Беларуская (тарашкевіца)
বাংলা
Български
Brezhoneg
Bosanski
Ming-dêṅ-gŭ
Cebuano
کوردی، ناوەندی
Qırımtatarca
ЧӀавашла
Cymraeg
Zazaki
Ελληνικά
Esperanto
Eesti
Euskara
Estremeñu
Võro
Frysk
Gaeilge
Gàidhlig
Galego
हिन्दी
Fiji Hindi
Hrvatski
Hornjoserbsce
Kreyòl ayisyen
Հայերեն
Interlingua
Interlingue
Ilokano
Ido
Íslenska

Contents [hide]

- History
 - Classical era
 - Medieval period
 - Early modern era
 - Late modern era
- Geography
 - Location and subdivisions
 - Climate
 - Environmental issues
- Demographics
 - Religion
- Economy
 - Shopping
 - Tourism
- Infrastructure
 - Transport
 - Highways and streets
 - Cars
 - Buses
 - Railway and subway
 - Airport
 - Parks and gardens
- Education
- Culture
 - Architecture
 - Graffiti
 - Cuisine and restaurants
 - Performing arts
 - Sports
 - Events
- Twin towns and partner cities
- Panoramic view
- See also
- References
- External links

← <https://en.wikipedia.org/wiki/Sanandaj>



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes

Article Talk

Sanandaj

From Wikipedia, the free encyclopedia

For the administrative subdivision, see Sanandaj County.

Sanandaj ⓘ ⓘ pronunciation (help·info) (Persian: سانداج) is a city in the northwestern part of Iran. In the 2016 census, its population was 373,987 ^[1] and it is the capital of **Kordestan** province at Iran. Sanandaj is the twenty-third largest city in Iran. Sanandaj is a city founded about 200 years ago, yet under its name it has grown to become a center of Kurdish culture.

Contents [hide]

- Society
- Famous people connected to Sanandaj
- References
- External links

Inter-language Information extraction

← <https://fa.wikipedia.org/wiki/سنندج> Search

تغییرات مرتبط
بارگذاری پرونده
صفحه‌های ویژه
پیوند پایدار
اطلاعات صفحه
آیتم ویکی‌داده
یادکرد پیوند این مقاله

زبان‌های دیگر
Deutsch
English
Español
Français
हिन्दी
Italiano
한국어
Русский
中文

۳۵ مورد دیگر
✎ ویرایش پیوند

محتویات [نهفتن]

- نام
- پیشینه تاریخی
 - تاریخچه‌ای از منطقه سنندج
 - سنه‌دز
- وضعیت طبیعی
 - جغرافیا
 - موقعیت جغرافیایی
 - آب و هوا
- مردم
 - زبان
 - جمعیت
 - مذهب
 - لباس
- فرهنگ و هنر
 - مشاهیر
 - جشن‌ها
 - سینماها
 - سینماهای فعال
 - کتابخانه‌های عمومی
- شهادت‌های عالی
 - فردوسی
 - هه زار
 - امام رضا
 - خاتم الانبیاء
 - مستوره اردلان
- جاذبه‌های تاریخی و مذهبی
 - عمارت و خانه‌های تاریخی
 - بازارها
 - حمام‌ها
 - دیگر آثار تاریخی
 - جاذبه‌های مذهبی
 - مساجد
 - امامزاده‌ها
 - مکان‌های مذهبی دیگر ادیان
 - جشن‌های مذهبی
- جاذبه‌های طبیعی
 - قلل مرتفع
 - مجموعه پارک تفریحی آبیدر
 - بزرگترین سینمای رویار جهان
- فهرست هتل‌های سنندج

Inter-language recommendations

Problem statement

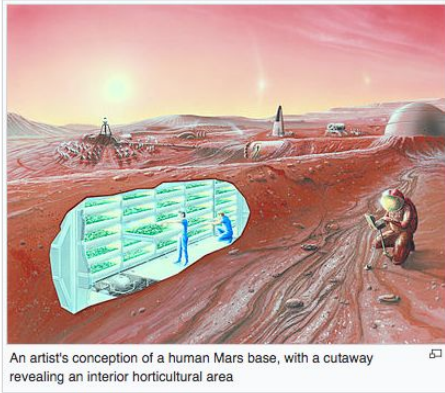
Given an article A in language L, recommend the list of sections to be added to the article considering similar articles to article A in language L.

Colonization of Mars

From Wikipedia, the free encyclopedia

Mars is the focus of much scientific study about possible [human colonization](#). Its surface conditions and the presence of [water on Mars](#) make it arguably the most [habitable of the planets](#) in the [Solar System](#), other than [Earth](#). Mars requires less energy per unit mass ([delta-v](#)) to reach from Earth than any planet except [Venus](#).

Permanent human habitation on a planetary body other than the Earth is one of science fiction's most prevalent themes. As technology has advanced, and concerns about the [future of humanity on Earth](#) have increased, the argument that [space colonization](#) is an achievable and worthwhile goal has gained momentum.^{[1][2]} Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.



An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area

Sections you can add

- [Relative similarity to Earth](#)
- [Differences from Earth](#)
- [Conditions for human habitation](#)
- [Radiation](#)
- [Transportation](#)
- [Equipment needed for colonization](#)
- [Robotic precursors](#)
- [Mission concepts](#)
- [Economics](#)
- [Possible locations for settlements](#)
- [Planetary protection](#)

< >

Edit to add new sections

How to define similarity?

Article
text

categories

Secure | https://en.wikipedia.org/wiki/Colonization_of_Mars

Not logged in | [Talk](#) | [Contributions](#) | [Create account](#) | [Log in](#)

Article | [Talk](#) | [Read](#) | [Edit](#) | [View history](#) |

Colonization of Mars

From Wikipedia, the free encyclopedia

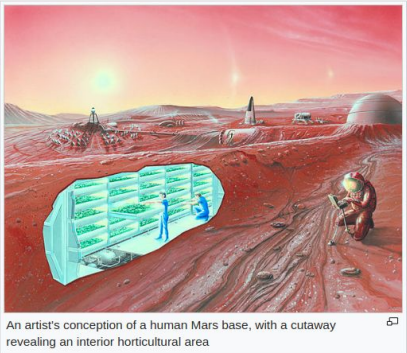
Mars is the focus of much scientific study about possible **human colonization**. Mars' surface conditions and past presence of **water**, make it arguably the most **hospitable planet** in the **Solar System** besides **Earth**. Mars requires less energy per unit mass (**delta-v**) to reach from Earth than any planet, except **Venus**.

Permanent human habitation on other planets is one of science fiction's most prevalent themes. As technology advances, and concerns about **humanity's future on Earth** increase, arguments favoring **space colonization** gain momentum.^{[1][2]} Other reasons for colonizing space include economic interests, long-term scientific research best carried out by humans as opposed to robotic probes, and sheer curiosity.

One of **Elon Musk's** stated goals, through his company **SpaceX**, is to make such colonization possible by providing transportation.

Contents [hide]

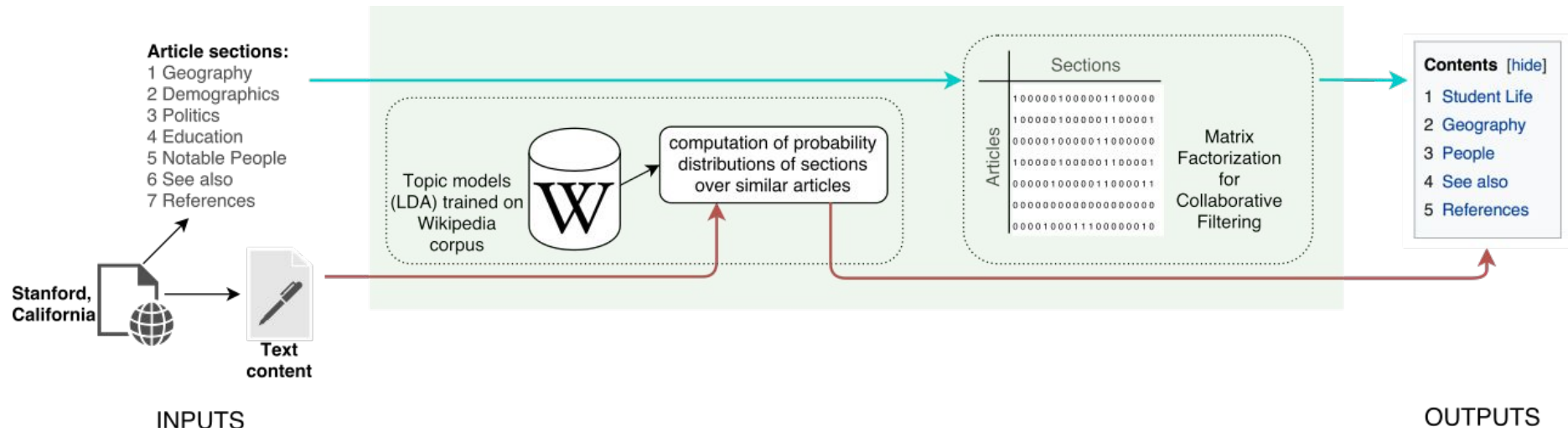
- Mission concepts and timelines
 - Mars One (colonization)
 - US Government (non-colonizing return trip)
 - Russian Government (non-colonizing return trip)
- Relative similarity to Earth
- Differences from Earth
- Conditions for human habitation
 - Effects on human health
 - Physical effects
 - Psychological effects
 - Terraforming



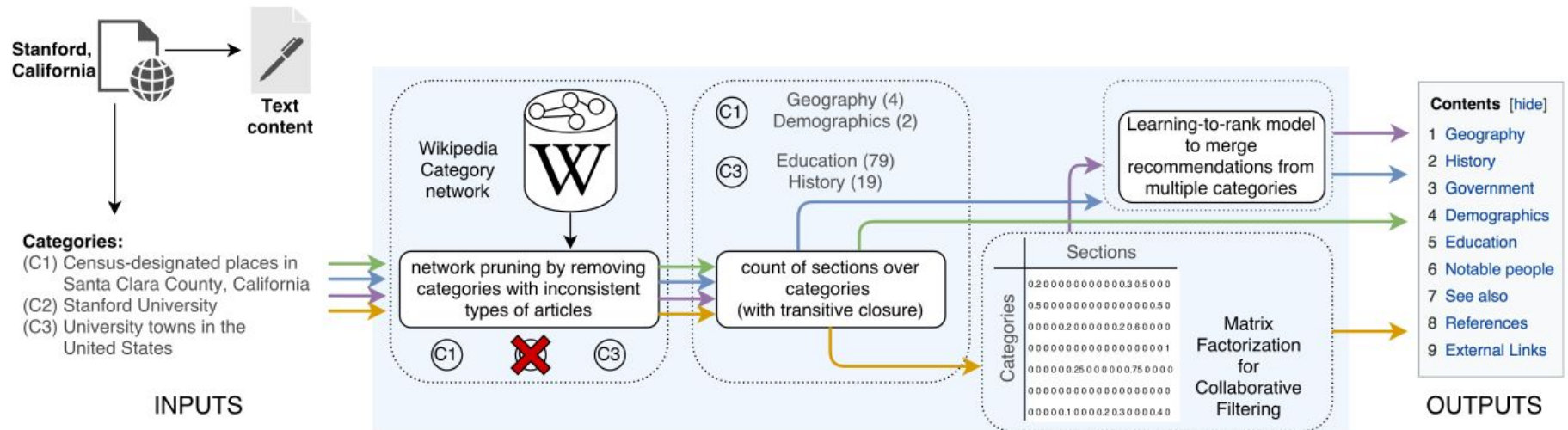
An artist's conception of a human Mars base, with a cutaway revealing an interior horticultural area

Categories: [Colonization of Mars](#) | [Exploration of Mars](#) | [Manned missions to Mars](#) | [Space colonization](#) | [Mars Society](#)

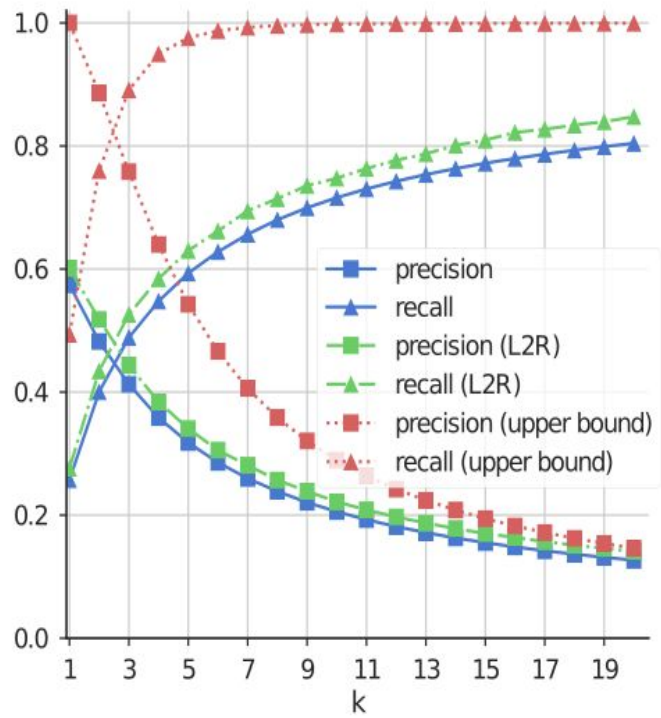
Article-based recommendations



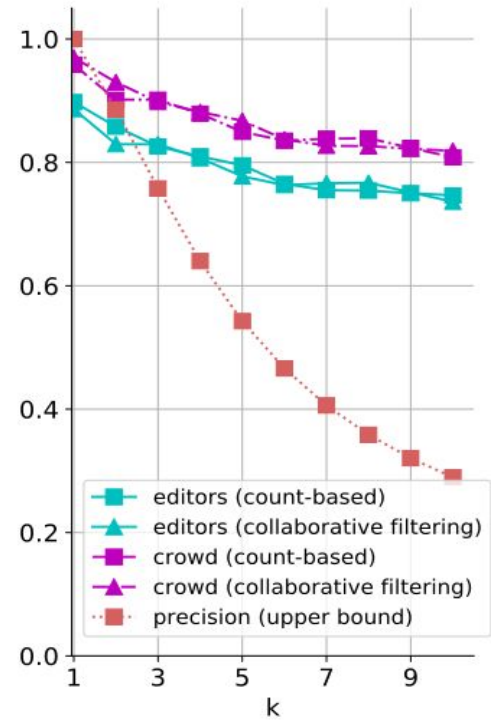
Category-based recommendations



Evaluation



Automatic



Human

Conclusion

- Introduced the section recommendation problem
- Explored several methods using
 - features derived from the raw input article
 - Wikipedia's category network
- Showed that the category centric count-based approach does best (precision @10 at 80%) for a Wikipedia language with a large category network such as enwiki.
- Learned that enwiki's category network is key in offering useful recommendations, but the network needs pruning (we developed a methodology for doing that)

Discussion

- **What about non-enwiki?**
 - The methods used are language independent though they (will most likely) rely on the size of the category network and the number of articles in these categories in a language
 - We expect the methodology to work in languages with large enough category network. We will test it next in frwiki
 - What if the language is small?

Discussion

- **Improving section recommendations**
 - Sourcing signal from other languages (inter-language)
 - Semantically related sections
 - Providing more in-depth information
 - The order of sections
 - Including less frequent and the long tail of sections
- **Entry point**
 - Article versus category

What's next?

- **More languages**
 - Build the tool/API for testing further
 - Test the models in frwiki: are they good?
 - What if the language is much smaller than fr?
 - Test in a smaller language
 - Develop a different set of technologies for small languages
 - Ma Commune
- **Experiment with less-experienced editors**

What's next?

- **Other types of recommendations**
 - Image recommendation
 - Info-box recommendation (a simpler problem)
 -

Thanks! :)

Documentation at

https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_articles_across_languages