Mako: I've been doing this for many years. I started in 2008 and have done this almost every single year since.

This began as an excuse for me to make sure I was up to date on Wikimedia Research.
“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From my Wikimania 2008 Submission

Back at Wikimania 2008, I set out to run a session that would provide a comprehensive literature review of articles in Wikipedia published in the last year.

“This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project.”

– From my Wikimania 2008 Submission

Then, about two weeks before Wikimania, I did the scholar search so I could build the literature.
"This talk will try to [provide] a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a birds eye view, the academic landscape as it is shaping up around our project."

– From my Wikimania 2008 Submission

I tried to import the whole list into Zotero and managed to get banned for abusing Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper...

And believe it or not, this year is even bigger.

And this talk is even shorter.
“This talk will try to provide a quick tour – a literature review in the scholarly parlance – of the last year’s academic landscape around Wikimedia and its projects geared at non-academic editors and readers. It will try to categorize, distill, and describe, from a bird’s eye view, the academic landscape as it is shaping up around our project.”

– From my Wikimania 2008 Submission

I tried to import the whole list into Zotero and managed to get banned for abusing Google Scholar because they thought that no human being could realistically consume the amount of material published on Wikipedia that year.

So anyway, I had a 45 minute talk so it worked out to 3.45 seconds to per paper...

And believe it or not, this year is even bigger.

And this talk is even shorter.
Academics have written a lot of papers about Wikipedia. There are more than 500 papers published about Wikipedia each year and although we’ve reached and moved past a peak it seems, it’s not slowing by much.
• 7,828 Wikipedia-related publications in the Scopus database as of yesterday (July 20, 2018)
• 109 recent publications covered in the 8 issues of the Wikimedia Research Newsletter from June 2017 to June 2018 (and hundreds more on our list!)

The newsletter aims to be comprehensive, but mostly ignores papers that use Wikipedia as a corpus only (which is popular e.g. in NLP research).
In selecting papers for this session, the goal is always to choose examples of work that:

- Represent **important themes** from Wikipedia in the last year.
- Research that is likely to be of **interest** to Wikimedians.
- Research by people who are **not at Wikimania**.
- ...with a bias towards **peer-reviewed** publications

Within these goals, the selections are **incomplete**, and **wrong**.

Mako

This paper is by a team at the University of Michigan and Northwestern University and it looks as image use.

Image use is something that has historically been studied very little. This year, it sort of exploded in popularity and there were a series of papers on the topic.
This paper really focused on understanding “image diversity” and it looks at it in the biggest 25 language editions of Wikipedia. This is what they mean by image diversity is that they found articles on the same topic (from inter-wiki links stored in WikiData and in the individual wikis) and then they looked at overlap in terms of images in commons.
Here is an example from the article on happiness. German shows a gorilla. Some images show up in a few. But—in general—there's a ton diversity.
They found that 67% of images appear in only one of the 25 editions. Some concepts—like wiki—have a tone of overlap. Other concepts—like science—have a huge amount of diversity.
Previous work has shown that text is very diverse in Wikipedia in the sense that different languages use different concepts to talk about a particular topic.

There was reason to believe that there might be less in image since they are hosted in commons and don’t need to be localized.

Every dot on this graph is a language pair. Things below the red line have more image diversity than text diversity.

As you can see, there is generally much more image diversity than text diversity.
Whose suggestions/opinions make it to the article and do not get reverted? 53k+ instances of interaction on talk pages paired with edit actions were analyzed.
Winning or losing depends on...

- Language (inviting, requesting, demanding an answer, promising something etc.)
- How many times you talk
- Who starts/ends the talk
- Your style (???? or !!!! etc)
- How authoritative you are
- How emotional your language is
You are most likely to win if you...

• Talk in detail about content
• Give examples
• Cite sources
• Do word work (spelling, word choice and order, etc)
You are most likely to lose if you...

- Talk about policies
- Moderate the talk
- Talk about page formatting
Multilingual Comparisons
Knowldege gaps are the theme of this Wikimania, and in (one form or the other) they have been a big theme in research this year too.

Some of this research is already being presented elsewhere here, so it’s out of scope for this talk. E.g. yesterday’s keynote by Martin Dittus about geographical imbalances, the “Wikipedia Cultural Diversity Observatory” (which goes beyond geolocation to incorporate other data for a fuller picture of diversity), and the Wikimedia Foundation’s own research and technology development to bridge such knowledge gaps.
Construct a common quality metric to compare over 28 million articles in 44 language Wikipedias, based on:

- article length
- number of references
- number of images
- number of first- and second-level headers
- ratio of references to the article length
- the number of quality flaw templates (e.g. lack of sources, NPOV violation)

These are combined into a single number.

Popularity is measured via pageviews.

As the authors point out, more sophisticated quality metrics exist, including the Wikimedia Foundation's ORES service, which is machine learning based. They didn't use it because it was only available for three languages.
These five metrics are positively correlated with the quality grades that editors assign manually on the English Wikipedia.

E.g., on the left you can see that there are almost no featured articles (blue) with less than 15000 bytes length. But more than half of the articles over 250k have featured article status.
Articles were grouped into 12 topic areas (e.g. “film”, “person”, “university”) based on infoboxes and interwiki links. This Venn diagram shows the overlap of articles about universities in the English, German and French Wikipedias. (Online tool: http://data.lewoniewski.info/informatics2017/vn/)

Categories were deliberately not used, and Wikidata isn’t mentioned in the paper at all.
This results in a detailed comparison of average quality and popularity across 12 topics and 44 languages. E.g.:

- The German Wikipedia’s articles about albums and video games have the highest average quality score (among the 44 languages).
- However, its footballer biographies only rank 10 in quality.
- Quality and popularity (measured via pageviews) correlate positively - but more strongly for some topics and languages than for others. Most strongly for the topic “company”, most weakly for the topic “settlements”.

NB: This result does not necessarily mean that the German Wikipedia has the best experts about albums and video games among its editors. More likely, this is because its overage of these topics is much more limited due to stricter notability criteria. (Some quick comparisons via the Venn diagram tool seem to confirm that other major languages have many more articles about these topics.)

The authors wisely refrain from calculating an overall quality score for each Wikipedia. I myself was less prudent and couldn't resist playing around with their data to (rather unscientifically) calculate the average of all topic averages for each language. By that measure, the German Wikipedia would come out on top - but only narrowly, closely followed by the English, Greek, Hindi and Chinese Wikipedia ;)

Presentation Title
Paper Summaries
Lewoniewski et al.: Multilingual quality and popularity

Lewoniewski et al.: Multilingual quality and popularity
This results in a detailed comparison of average quality and popularity across 12 topics and 44 languages. E.g.:

- The German Wikipedia’s articles about albums and video games have the highest average quality score (among the 44 languages).
- However, its footballer biographies only rank 10 in quality.
- Quality and popularity (measured via pageviews) correlate positively - but more strongly for some topics and languages than for others. Most strongly for the topic “company”, most weakly for the topic “settlements”.

NB: This result does not necessarily mean that the German Wikipedia has the best experts about albums and video games among its editors. More likely, this is because its overage of these topics is much more limited due to stricter notability criteria. (Some quick comparisons via the Venn diagram tool seem to confirm that other major languages have many more articles about these topics.)

The authors wisely refrain from calculating an overall quality score for each Wikipedia. I myself was less prudent and couldn't resist playing around with their data to (rather unscientifically) calculate the average of all topic averages for each language. By that measure, the German Wikipedia would come out on top - but only narrowly, closely followed by the English, Greek, Hindi and Chinese Wikipedia ;)
Nonparticipation: Who is not contributing?

Once again, an important theme this year—related to knowledge equity—is why do internet users from different social groups contribute differently to Wikipedia?
This paper explored the factors and processes that influence these ‘participation gaps.’
Analysing survey data collected from 1512 adults in the US in 2016, the authors used logistic regression to model the activity of online knowledge production as a step-by-step process that internet users who contribute to Wikipedia go through.

They conceptualized a pipeline that anticipates leaks at the different stages of the knowledge production process so that fewer contributors remain at each subsequent step, beginning from a cohort of internet users.

Most work on the participation gap has focused on the final stage about whether or not people contribute. The authors of this paper show that there are gaps at many earlier stages such as whether or not people know that Wikipedia is editable, whether they have been on the site, or whether they know it even exists.
Participation increased at all stages of the pipeline when respondents’
• Had high education
• Had high internet skills and
• Were younger in age

*So?* Support interventions that reduce technical and knowledge-based” entry barriers

Participation divides emerge at early stages of the pipeline according to respondents’
• Income
• Employment status
• Racial / ethnic background

*So?* Address early participation gaps in minorities and lower income classes by reducing internet experience and autonomy obstacles

The results showed that: (At all stages of the pipeline): Education levels, internet literacy levels, and age; significantly influenced levels of activity at each step of the pipeline.

(Recommendation): With this information, the authors recommend the “support to interventions that reduce technical and knowledge-based” entry barriers as a means to increase participation at all the levels of knowledge production.

(At the early stages of the pipeline): Income, employment and race are significant factors that influence levels of activity in that stage of knowledge production.

(Recommendation): “This suggests the need for interventions addressing early participation gaps in minorities and lower income classes by reducing internet experience and autonomy obstacles”.
Participation divides are again visible in the two later stages of the pipeline with less activity recorded for females.

**Recommendations:**

- Create awareness especially among females that Wikipedia is a crowdsourced project.
- Provide continued support for gendergap campaigns and initiatives that seek to recruit more female contributors.

(At the later stage of the pipeline): Gender played a role to determine that, compared with males, fewer people who identify as female know that “Wikipedia is editable” and actually go beyond that awareness to contribute to Wikipedia. (Recommendation): The results therefore suggest two things; the need to 1. Create awareness among females that Wikipedia is a crowdsourced project that anybody can edit. 2. To continue support for gendergap campaigns and initiatives that seek to recruit more female contributors.
Mako

Perhaps the only topic that we've covered ever year is studies that use Wikipedia as source of data because there are loads and loads of these papers—every year.

Once again, this year saw a new crop of these.
One of these papers was a paper led by Mohamed Medhi at Concordia University in Montréal that uses papers that use Wikipedia as a source of data as... wait for it... a source of data.

This paper is a systematic review of work meaning that it doesn't present new work. It presents a summary of a large body of other work. In this case, 132 papers that use Wikipedia as a data source.
In addition summarizing papers, they break things down very systematically into 10 tables that categorize papers along a set of dimensions.

For example, they categorize most papers in this space as in the broad area of information retrieval — a body of computer and information science focused around giving people good answers to queries.

The other big area is natural language processing. In this case, Wikipedia contains data which can help systems that seek to understand language. This might include studies that use wikidata and inter-language links as a source of translation data.

Each has a bunch of subareas.
### Table 4

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Ch</th>
<th>Du</th>
<th>En</th>
<th>Fr</th>
<th>Ge</th>
<th>Ja</th>
<th>ko</th>
<th>NS</th>
<th>MU</th>
<th>Pe</th>
<th>Ru</th>
<th>Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information retrieval</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-language IR</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic IR</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multimedia IR</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other IR topics</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query processing</td>
<td>11</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ranking and clustering systems</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text classification</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Textual IR</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Natural language processing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computational linguistics</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information extraction</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other natural language processing topics</td>
<td>6</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic relatedness</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontology building</td>
<td>1</td>
<td>12</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other corpus topics</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total number of distinct studies</strong></td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>76</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>48</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
The paper also describes:

- Derivative datasets created from Wikipedia data
- Tools that can be used to study Wikipedia
- The dataset of papers used to create the paper (https://wikilit.referata.com)

Published in 2017 but the paper has been a long time coming. The first version of this paper was submitted in 2014! The speed of academic publishing.

The big change has been a push toward WikiData
Those are our eight exemplary studies from the past year. There has been just tons and tons of work in this area. Trying to talk about this in 40 minutes strikes me as increasingly crazy every year we try to do it.

The most important source is the Wikimedia Research Newsletter which has since 2011 been published monthly in the (English) Signpost and syndicated on the Wikimedia Research space on Meta-Wiki. (Special thanks to Dario Taraborelli and User:Masssly for finding and cataloguing new publications throughout the year!)

But there are other resources as well. And I encourage you to get involved.