Wikidata Statistics: What, Where, and How?

Goran S. Milovanović, Phd

Data Scientist for Wikidata and Wiktionary
Wikimedia Deutschland

WIKI DATA CON

Wikidata
& languages 2019

Wikidata Statistics:
What, Where, and How?

All data products mentioned in this talk are available from

http://wmdeanalytics.wmflabs.org/

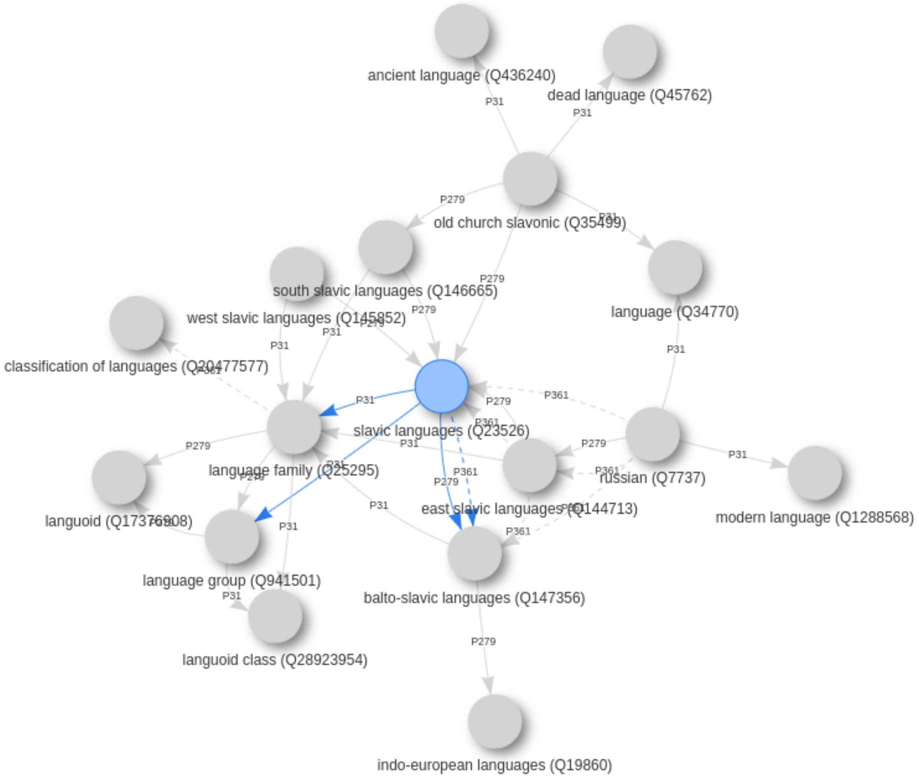
WIKI DATA CON

Wikidata
& languages 2019

Wikidata Statistics: What, Where, and How?

Goals

- An Overview of **Wikidata Statistics & Analytics systems**
- Exemplify **the usage of our analytics** in several domains (Wikidata items, languages, external identifiers, item quality)
- Go just a bit under the hood to illustrate **how we are doing it**



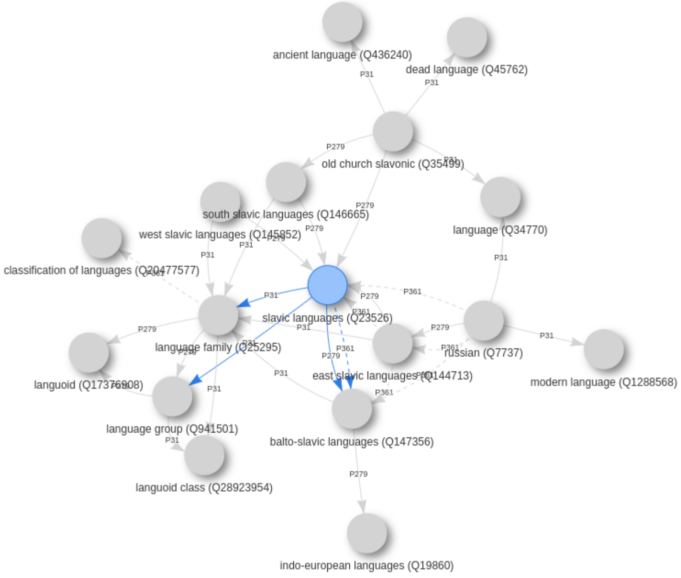
ancient language (Q436240)
dead language (Q45762)
old church slavonic (Q35499)
south slavic languages (Q146665)
west slavic languages (Q145852)
language (Q34770)
classification of languages (Q20477577)
slavic languages (Q23526)
language family (Q25295)
russian (Q7737)
modern language (Q1288568)
languoid (Q17376908)
east slavic languages (Q144713)
language group (Q941501)
balto-slavic languages (Q147356)
languoid class (Q28923954)
indo-european languages (Q19860)
P31
P279
P361

http://wmdeanalytics.wmflabs.org/WD_LanguagesLandscape/

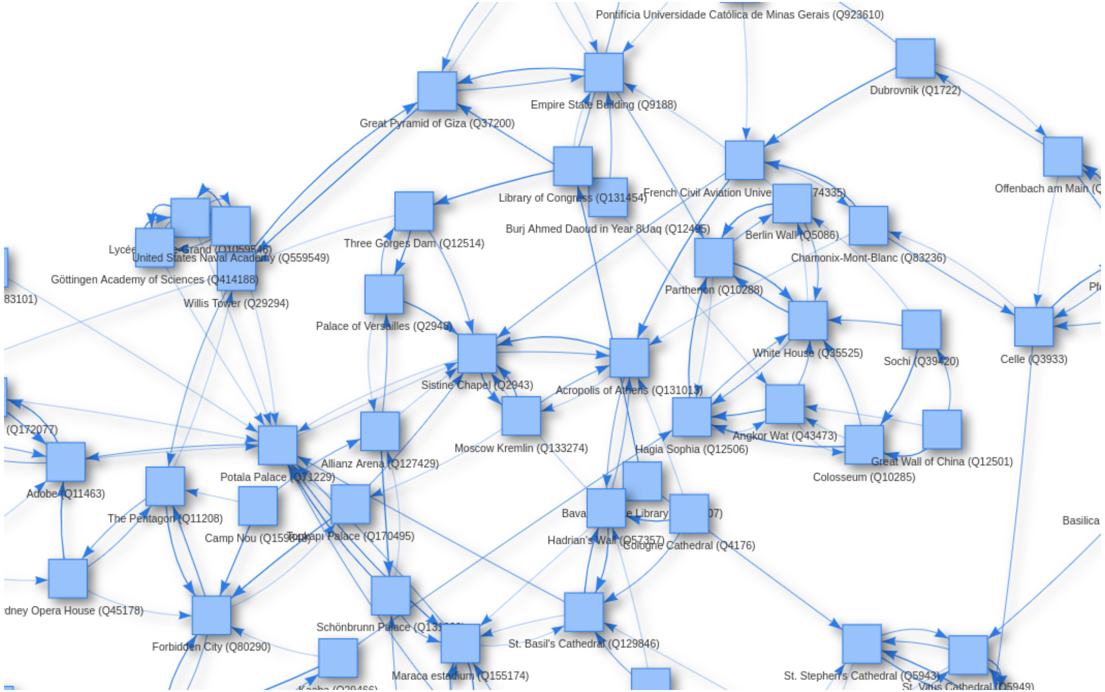
WIKI DATA CON
Wikidata & languages 2019

In 2017: we need an analytical system that will give us an insight into the ways the Wikidata items are reused across the Wikimedia projects (Wikipedia, Wikivoyage, Wikisource, etc)

Wikidata Concepts Monitor
→item reuse similarity structures
= items frequently used together across the Wikimedia projects are connected.



ancient language (Q436240)
dead language (Q45762)
old church slavonic (Q35499)
south slavic languages (Q146665)
west slavic languages (Q145852)
language (Q34770)
classification of languages (Q20477577)
slavic languages (Q23526)
language family (Q25295)
russian (Q7737)
modern language (Q1288568)
languoid (Q17376908)
east slavic languages (Q144713)
language group (Q941501)
balto-slavic languages (Q147356)
languoid class (Q28923954)
indo-european languages (Q19860)
P31
P279
P361

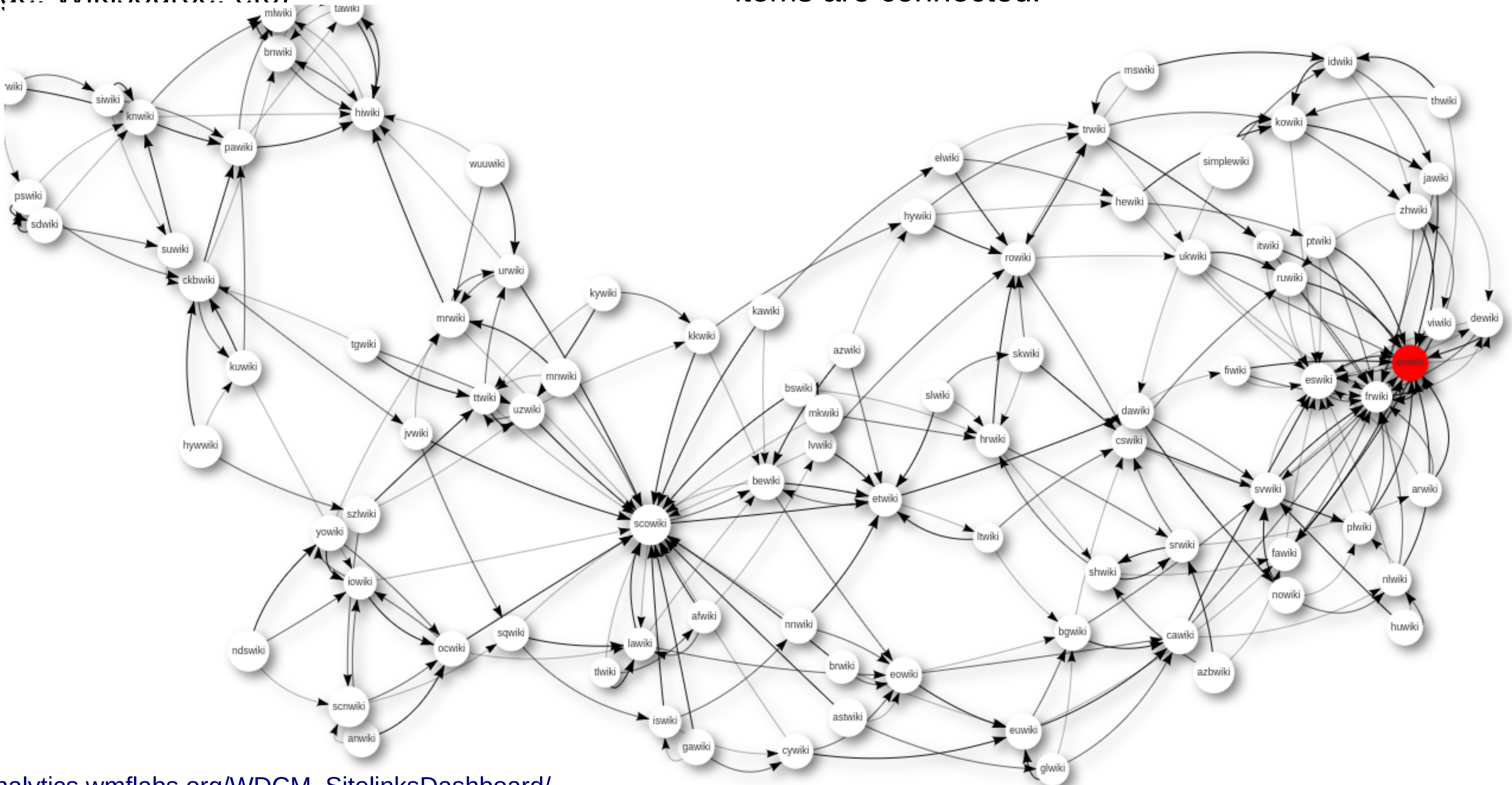


Pontifícia Universidade Católica de Minas Gerais (Q923610)
Dubrovnik (Q1722)
Empire State Building (Q9188)
Great Pyramid of Giza (Q37200)
Offenbach am Main (Q
Library of Congress (Q131454)
French Civil Aviation Unive
Burj Ahmed Daoud in Year 8Uaq (Q12495)
Berlin Wall (Q5086)
Three Gorges Dam (Q12514)
Chamonix-Mont-Blanc (Q83236)
United States Naval Academy (Q559549)
Göttingen Academy of Sciences (Q414188)
Parthenon (Q10288)
Willis Tower (Q29294)
Palace of Versailles (Q2946)
White House (Q35525)
Sochi (Q39420)
Celle (Q3933)
Sistine Chapel (Q2943)
Acropolis of Athens (Q131013)
(Q172077)
Moscow Kremlin (Q133274)
Hagia Sophia (Q12506)
Angkor Wat (Q43473)
Great Wall of China (Q12501)
Allianz Arena (Q127429)
Potala Palace (Q71229)
Colosseum (Q10285)
Adobe (Q11463)
The Pentagon (Q11208)
Camp Nou (Q15
Topkapı Palace (Q170495)
Hadrian's Wall (Q57357)
Cologne Cathedral (Q4176)
Basilica
dney Opera House (Q45178)
Schönbrunn Palace (Q13
St. Basil's Cathedral (Q129846)
Forbidden City (Q80290)
Maracã estadium (Q155174)
St. Stephen's Cathedral (Q5943)
St. Vitus Cathedral (Q5949)

http://wmdeanalytics.wmflabs.org/WDCM_SitelinksDashboard/

In 2017: we need an analytical system that will give us an insight into the ways the Wikidata items are reused across the Wikimedia projects (Wikipedia, Wikivoyage, Wikisource, etc)

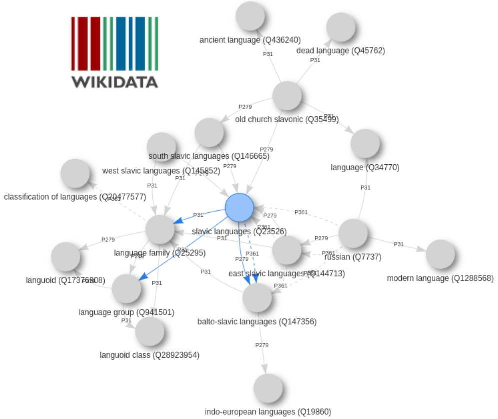
Wikidata Concepts Monitor
→Wikipedia similarity structures
= projects that reuse the similar Wikidata items are connected.

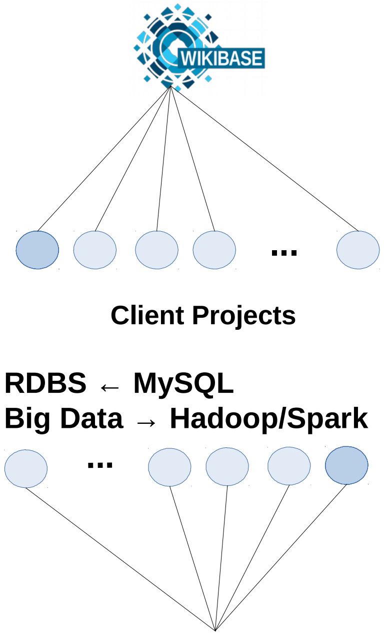


http://wmdeanalytics.wmflabs.org/WDCM_SitelinksDashboard/

In 2017: we need an analytical system that will give us an insight into the ways the Wikidata items are reused across the Wikimedia projects (Wikipedia, Wikivoyage, Wikisource, etc) →
Wikidata Concepts Monitor

RDF Graph ← SPARQL
WDQS or WD Dump processing


WIKIDATA
ancient language (Q436240)
dead language (Q45762)
old church slavonic (Q35499)
south slavic languages (Q146665)
west slavic languages (Q145852)
language (Q34770)
classification of languages (Q20477577)
slavic languages (Q23526)
language family (Q25295)
russian (Q7737)
modern language (Q1288568)
languoid (Q17376908)
east slavic languages (Q144713)
language group (Q941501)
balto-slavic languages (Q147356)
languoid class (Q28923954)
indo-european languages (Q19860)
P31
P279
P361



Client Projects

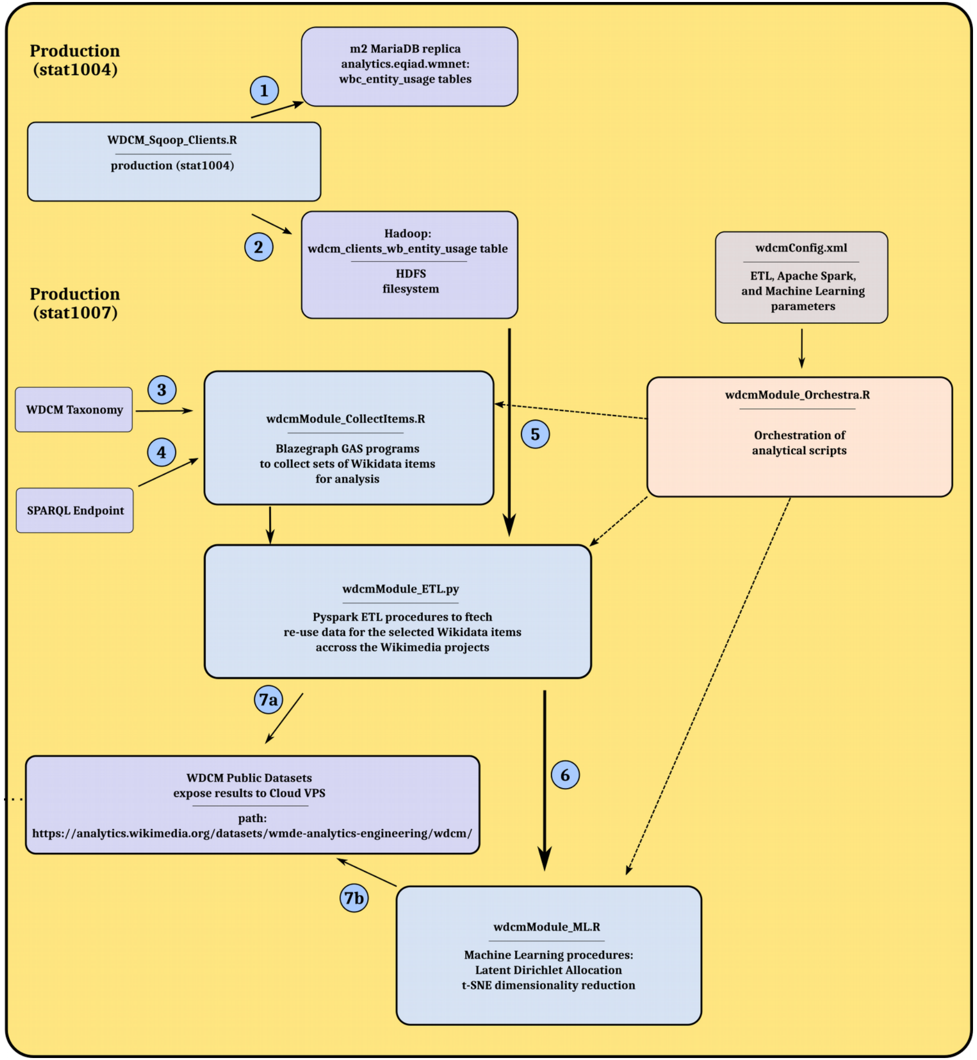
RDBS ← MySQL
Big Data → Hadoop/Spark

Wikimedia Foundation
Data Lake

Data Model
&
Structures

Motivation/Goals
→
Questions

http://wmdeanalytics.wmflabs.org/WDCM_UsageDashboard/



Production (stat1004)

m2 MariaDB replica
analytics.eqiad.wmnet:
wbc_entity_usage tables

1

WDCM_Sqoop_Clients.R

production (stat1004)

2

Hadoop:
wdcm_clients_wb_entity_usage table

HDFS
filesystem

Production (stat1007)

wdcmConfig.xml

ETL, Apache Spark,
and Machine Learning
parameters

WDCM Taxonomy

3

wdcmModule_CollectItems.R

Blazegraph GAS programs
to collect sets of Wikidata items
for analysis

4

SPARQL Endpoint

5

wdcmModule_Orchestra.R

Orchestration of
analytical scripts

wdcmModule_ETL.py

Pyspark ETL procedures to ftech
re-use data for the selected Wikidata items
accross the Wikimedia projects

7a

6

WDCM Public Datasets
expose results to Cloud VPS

path:
https://analytics.wikimedia.org/datasets/wmde-analytics-engineering/wdcm/

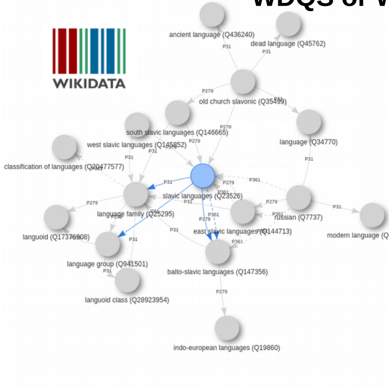
7b

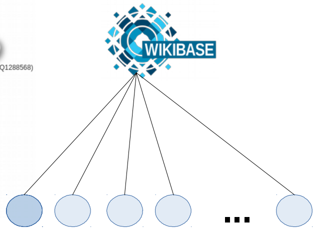
wdcmModule_ML.R

Machine Learning procedures:
Latent Dirichlet Allocation
t-SNE dimensionality reduction

Many systems need to work together just in order for you to get your data and organize the data model appropriately...

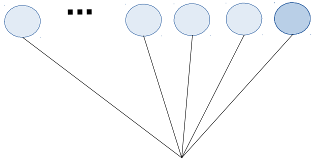
RDF Graph ← SPARQL
WDQS or WD Dump processing


WIKIDATA
ancient language (Q436240)
dead language (Q45762)
old church slavonic (Q35499)
south slavic languages (Q146665)
west slavic languages (Q145852)
language (Q34770)
classification of languages (Q20477577)
slavic languages (Q23526)
language family (Q25295)
russian (Q7737)
languoid (Q17376908)
east slavic languages (Q144713)
modern language (Q1288568)
language group (Q941501)
balto-slavic languages (Q147356)
languoid class (Q28923954)
indo-european languages (Q19860)
P31
P279
P361



Client Projects

RDBS ← MySQL
Big Data → Hadoop/Spark



Wikimedia Foundation
Data Lake

External Teams
&
Communication

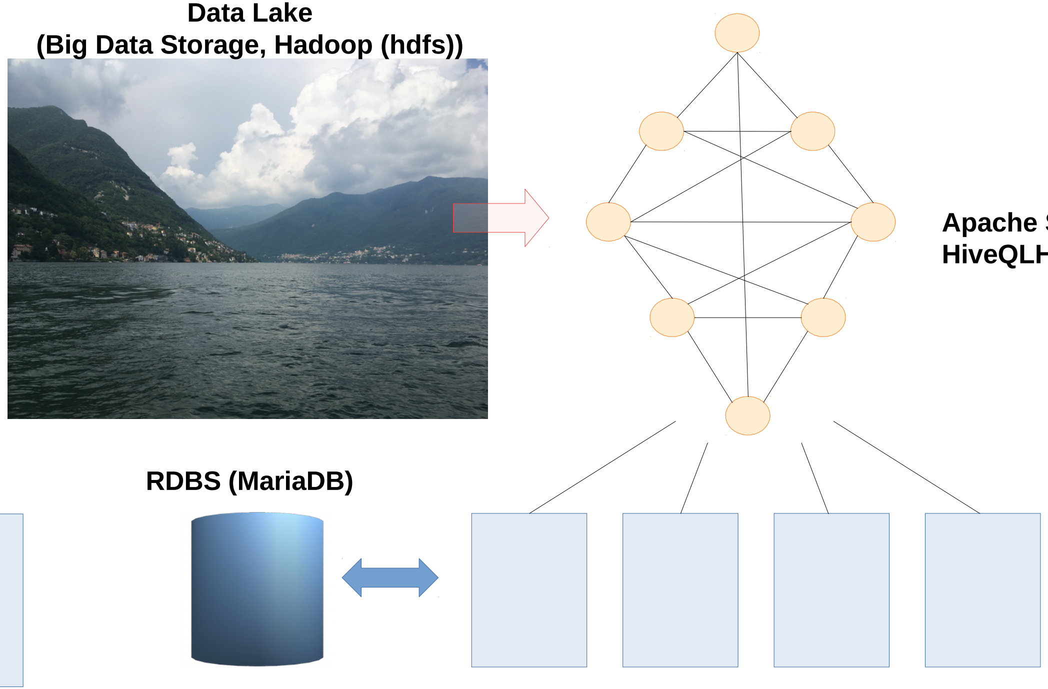
Analytics
Infrastructure

Data Model
&
Structures

Motivation/Goals
→
Questions

Analytics Cluster
The WMF Analytics Engineering Team

Data Lake
(Big Data Storage, Hadoop (hdfs))



Apache Spark,
HiveQLHadoop

RDBS (MariaDB)

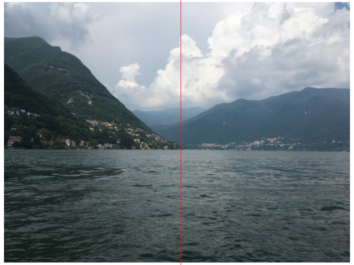
R, Python,
SQL, etc.

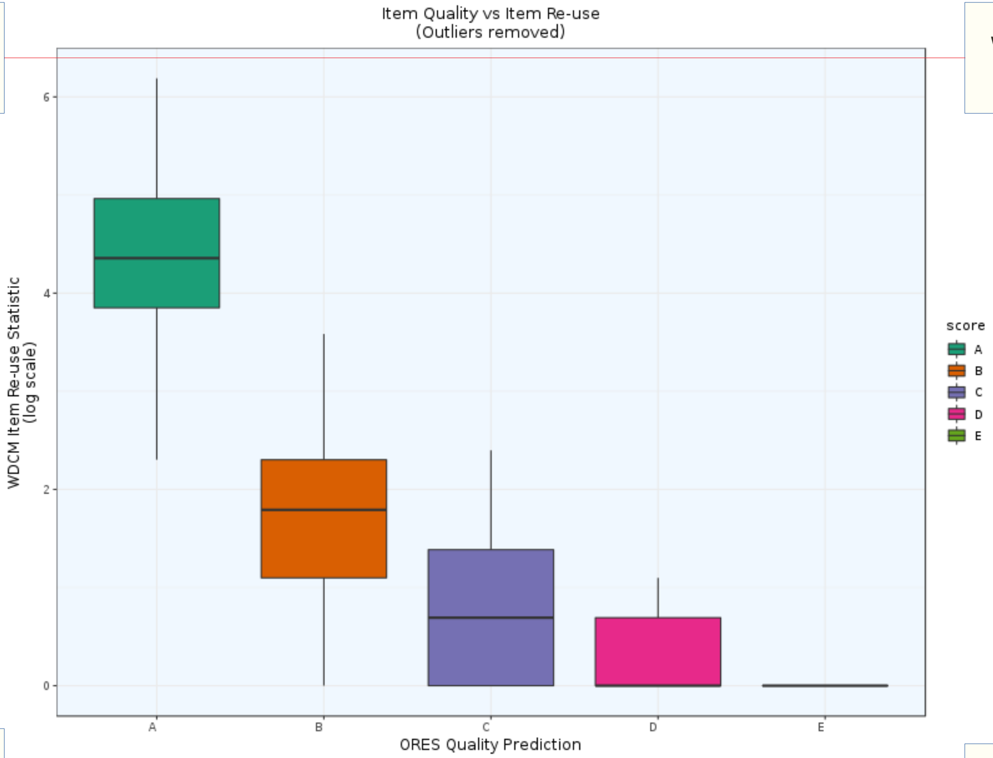
Stats servers (stat100*)

Data + Systems Synchronization:
it can get really nasty...

wmf.mediawiki_history table (huge)

Wikidata Concepts Monitor Reuse Statistics





WD JSON Dump in HDFS

ORES ML System Item Quality Predictions

http://wmdeanalytics.wmflabs.org/Wikidata%20Quality%20Report.nb.html

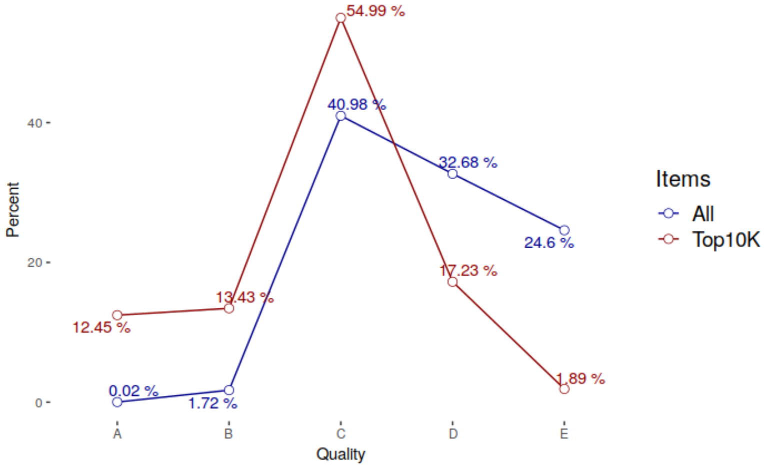
Data + Systems Synchronization:
it can get really nasty...

wmf.mediawiki_history table (huge)

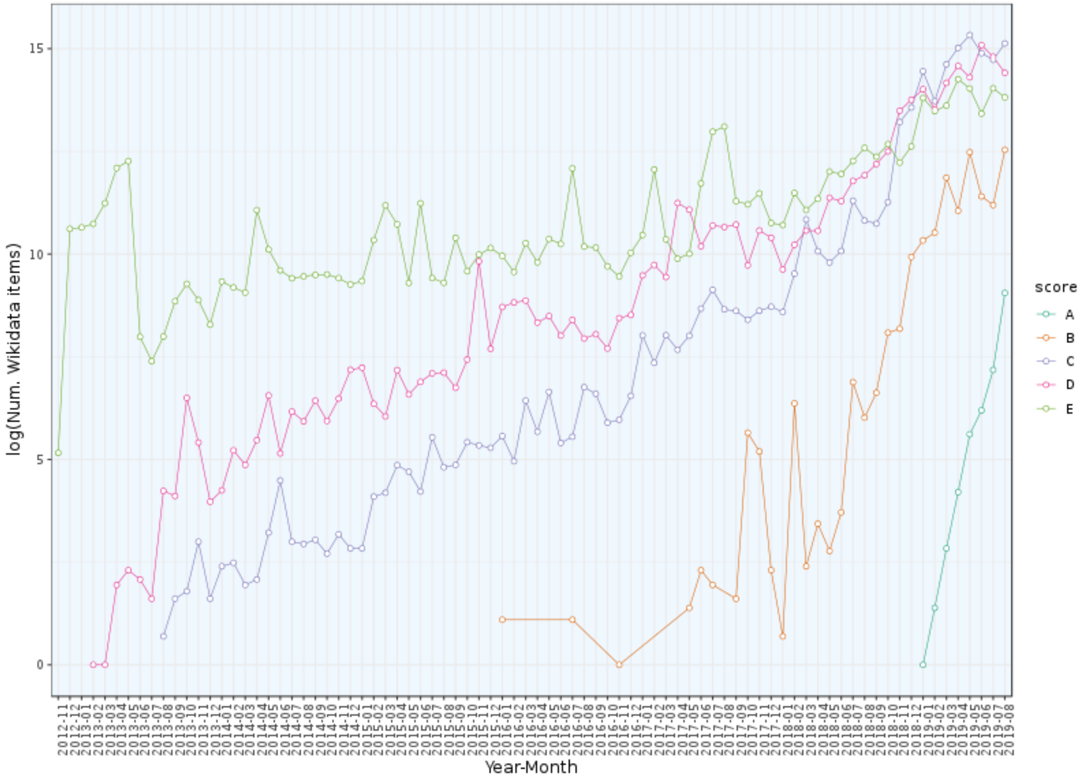
Wikidata Concepts Monitor Reuse Statistics

WD JSON Dump in HDFS

ORES ML System Item Quality Predictions

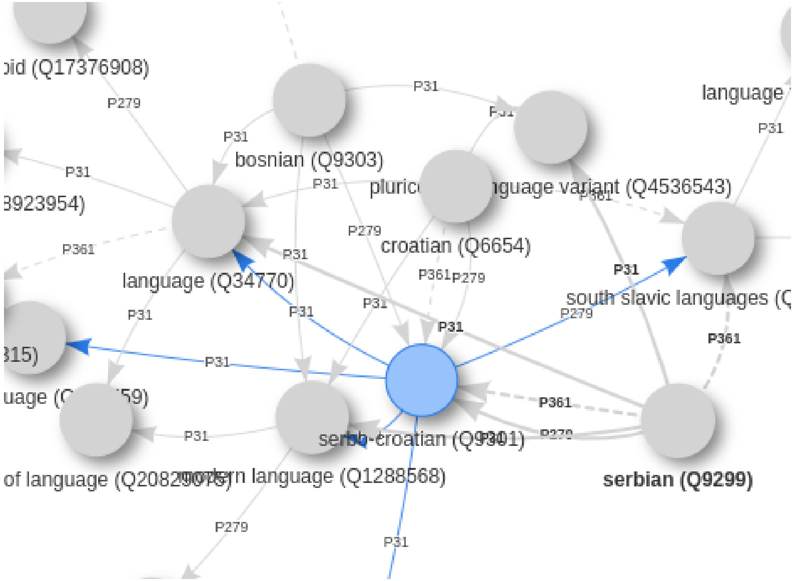


When did the item receive its latest revision?
(NOTE. Only items with the ORES quality class prediction are considered)



http://wmdeanalytics.wmflabs.org/Wikidata%20Quality%20Report.nb.html

While we play with large and complex datasets, we try to make use of the byproducts of our work…
<u>And you should too!</u>



http://wmdeanalytics.wmflabs.org/WD_LanguagesLandscape/

Fix the Ontology!

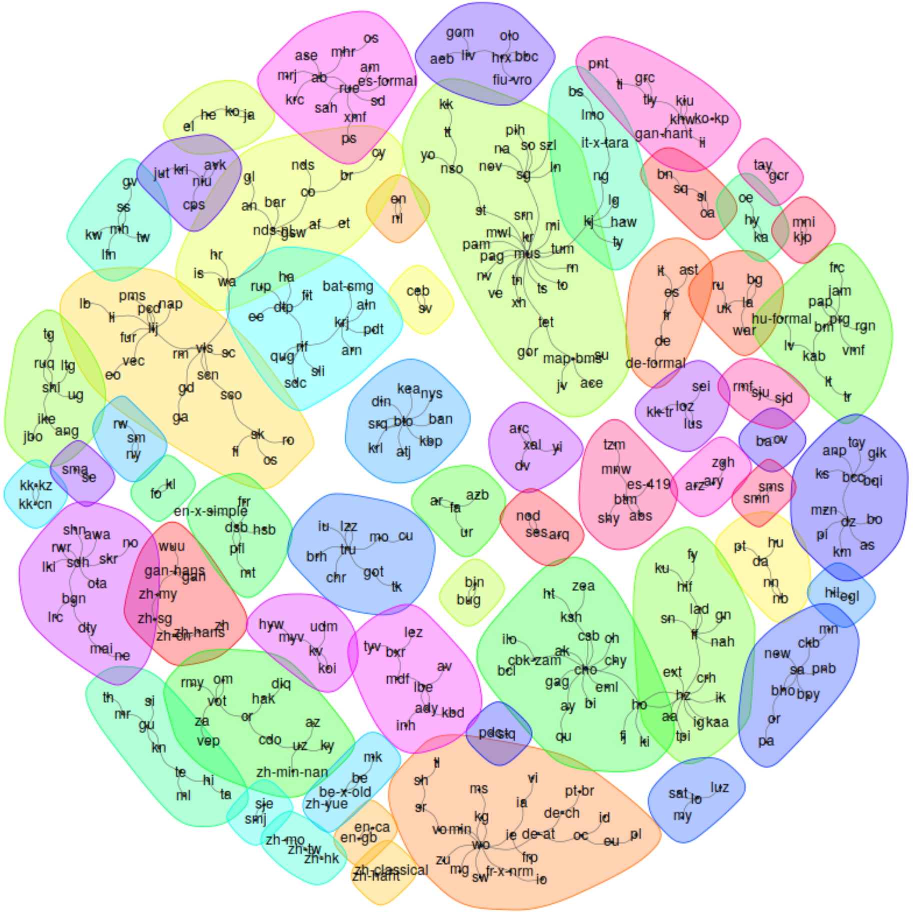
Example: languages that are a part of (P361) and a subclass of (P279) something at the same time (e.g. both Serbian (Q9299) and Croatian (Q6654) are parts of and subclasses of Serbo-Croatian (Q9301) at the same time; so is Serbo-Croatian a language, or a set of languages (besides being a pluricentric language (Q250858))?

Mereological and set-theoretic relations are not the same.

Speaking of languages...

The Wikidata Languages Landscape

… relies on different data sources to provide a comprehensive picture of how different languages are used in Wikidata and - via the entities that they refer to - how they are mapped across the universe of Wikimedia project

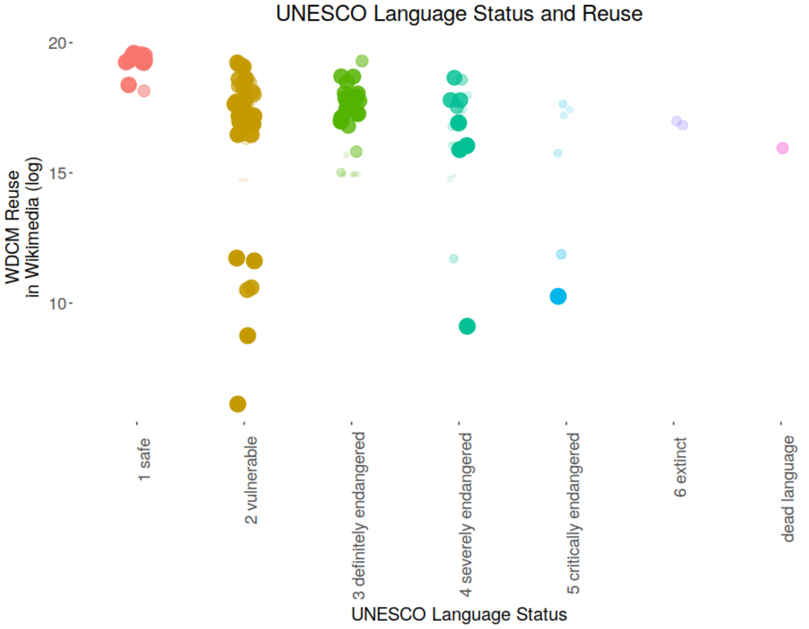


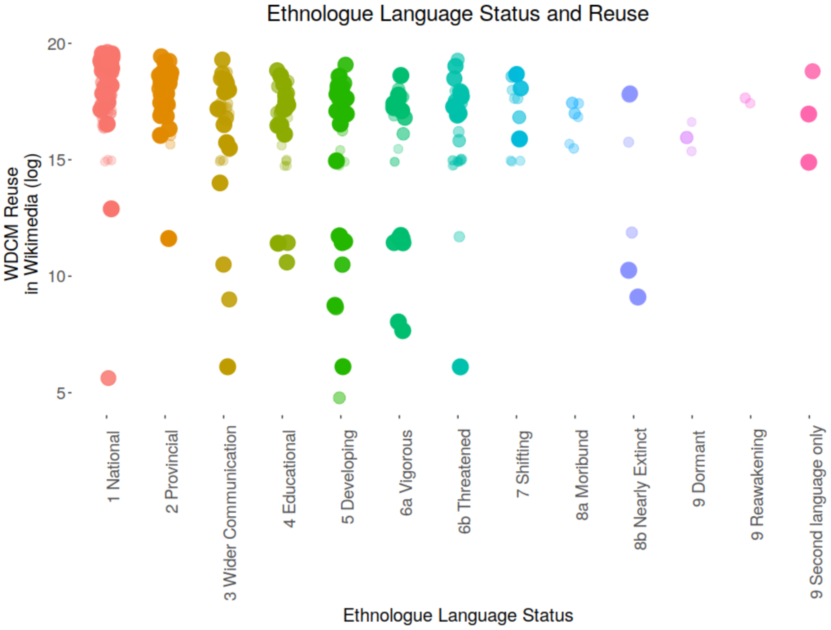
http://wmdeanalytics.wmflabs.org/WD_LanguagesLandscape/

Speaking of languages...

The Wikidata Languages Landscape

… relies on external data found in Wikidata to make relevant assessments of the way languages are used...


UNESCO Language Status and Reuse

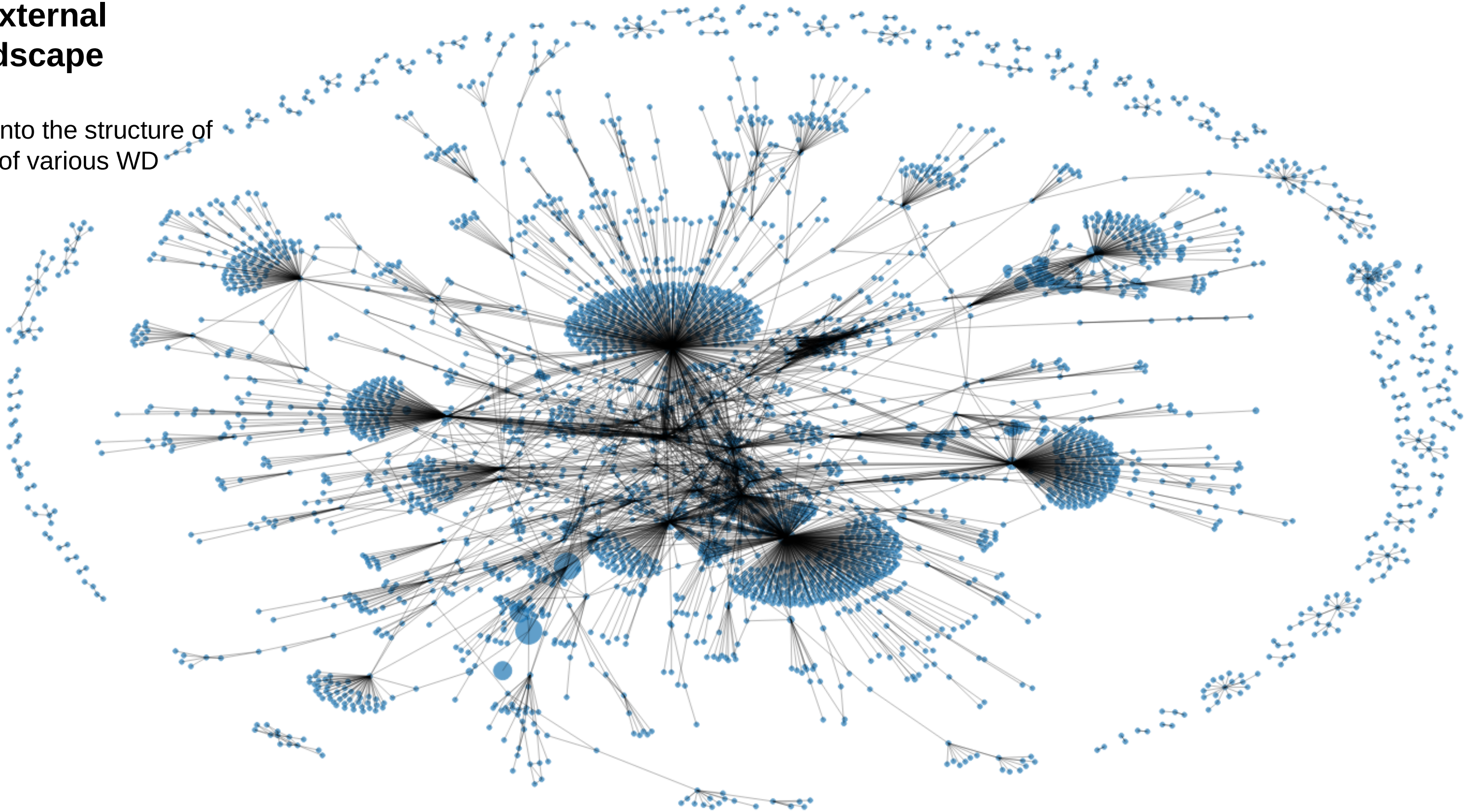

Ethnologue Language Status and Reuse

http://wmdeanalytics.wmflabs.org/WD_LanguagesLandscape/

Speaking of external resources...

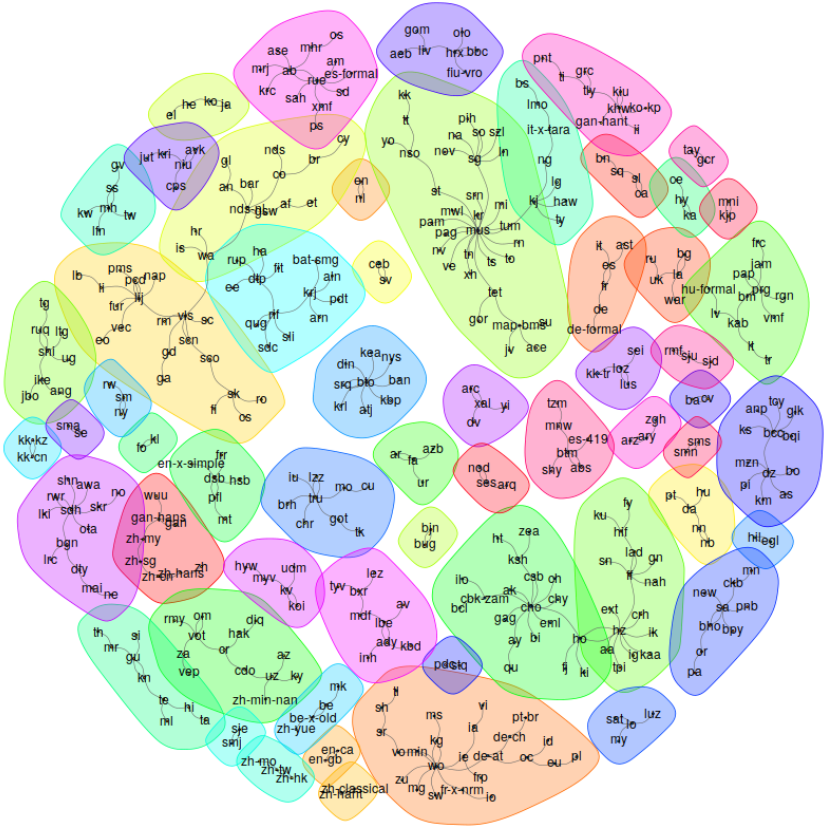
The Wikidata External Identifiers Landscape

… to provide insight into the structure of the overlap in usage of various WD external identifiers.

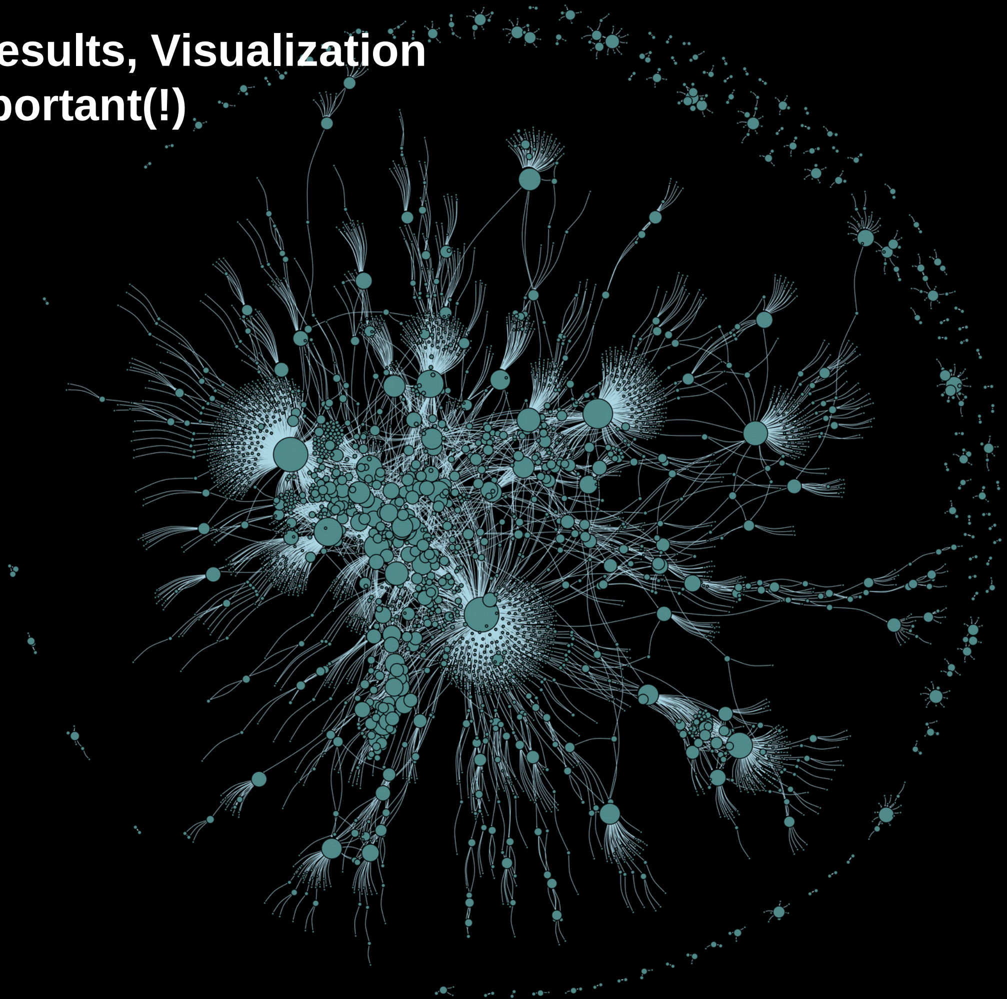


http://wmdeanalytics.wmflabs.org/WD_ExternalIdentifiersDashboard/

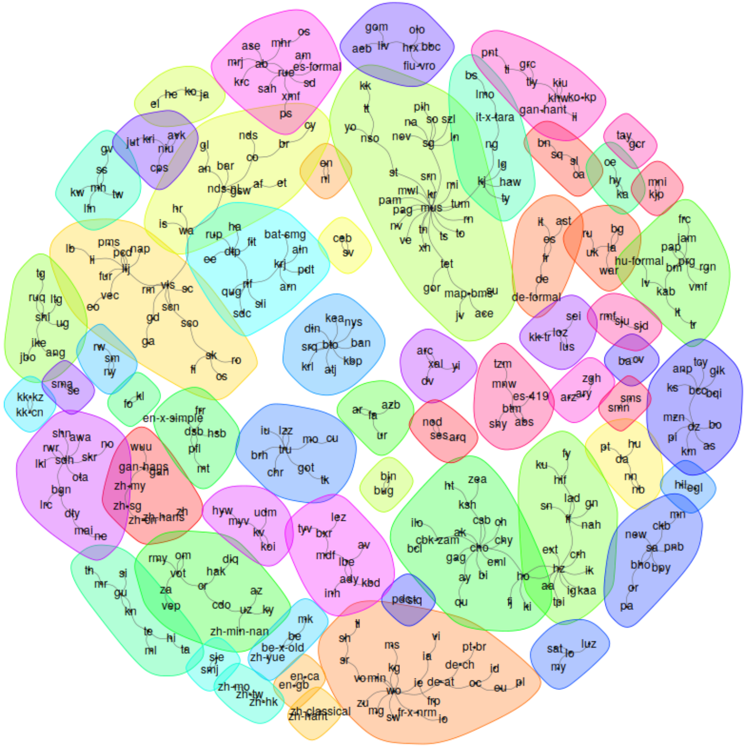
Communication of our results, Visualization & Aesthetics → Very Important(!)



Communication of our results, Visualization & Aesthetics → Very Important(!)



But one just does not get interesting results for free: **Machine Learning and Statistics**

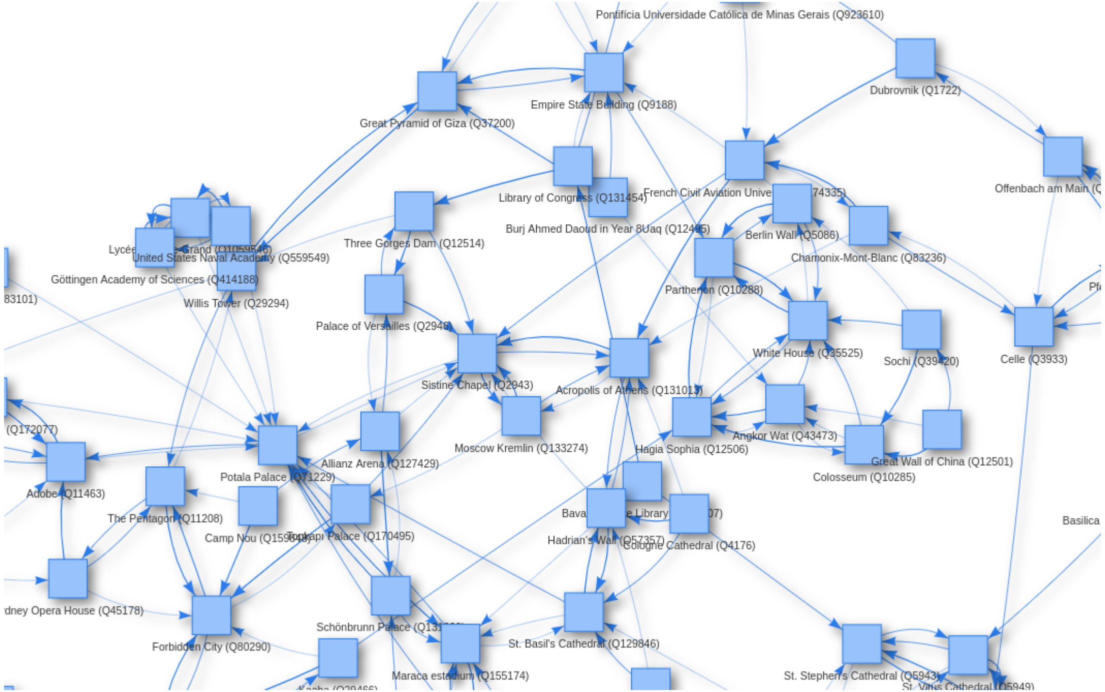


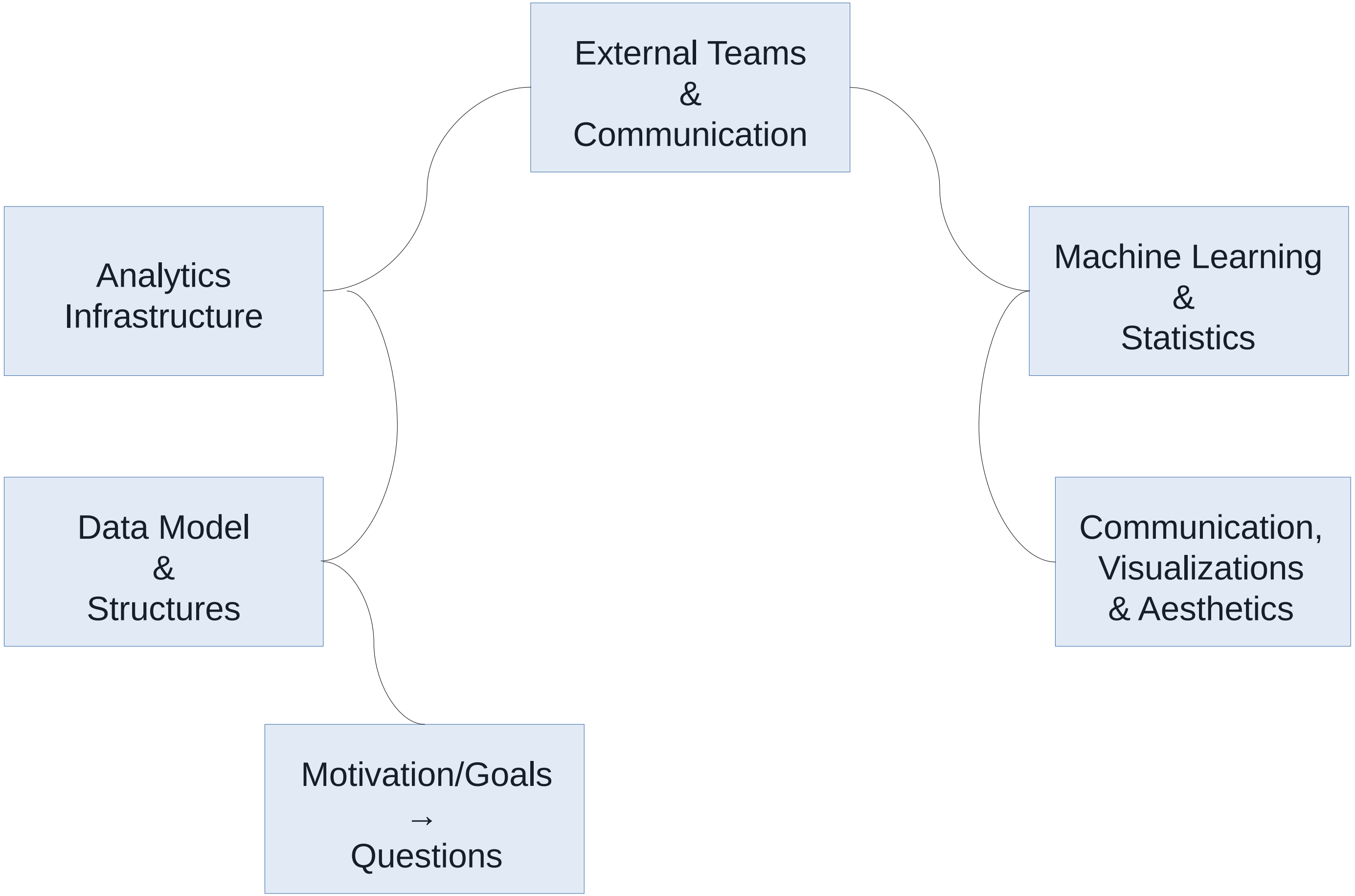
ase mhr os am mrj ab rue es-formal krc sah sd xmf ps cy
gom olo aeb liv hrx bbc fiu-vro
pnt ti grc tly kiu khw ko-kp gan-hant ii
el he ko ja
jut kri avk niu cps
gv se kw nrh tw lm
gl nds an bar co br af et nds-gsw
en nl
kk tt yo nso nev pih na so szl sg ln st srn mwl kr pam pag mus tum rw tn ts to ve xh mi rn tet gor map-bms su jv ace
bs lmo it-x-tara ng lg kj haw ty
bn sq sl oa
tay gcr
oe hy ka
mni kjp
hr is wa rup ha fit bat-smg ain ee dtp krj pdt quq nf arn sdc sli
ceb sv
lb pms pcd nap li lij fur vec rm vls sc eo ssn gd sco ga sk ro fi os
tg ruq ltg shi ug ike jbo ang
rw sm ny
sma se
kk-kz kk-cn
fo kl
it ast es fr de de-formal
ru uk bg la war
frc jam pap hu-formal bm prg rgn lv kab vmf lt tr
din kea nys srq bto ban kri atj kbp
arc xal yi dv
tzm mnw es-419 btm shy abs
sei kk-tr loz lus
rmf sju sjd
ba ov
zgh arz ary
sms smn
anp tay gik ks bcc bqi mzn dz bo pi km as
en-x-simple fr dsb hsb pfl mt
iu lzz brh mo cu tru chr got tk
ar fa azb ur
nod ses arq
bjn bug
shn awa rwr no lki sdh skr ota bgn lrc dty mai ne
wuu gan-hans gan zh-my zh-sg zh zh-cn zh-hans
hyw myv udm kv koi
tyv lez bxr av mdf lbe ady kbd inh
ht zea ksh ilo csb oh cbk-zam ak bcl chy cho gag eml ay bi qu fj
ku fy hif sn lad gn nah ext crh hz ik ae ig kaa tpi
pt hu da nn nb
hil egl
new ckb mn sa pnb bho bpy or pa
th si mr gu kn te hi ml ta
rmy om vot diq hak za or az vep cdo uz ky zh-min-nan
sje smj
mk be be-x-old zh-yue
zh-mo zh-tw zh-hk
en-ca en-gb zh-classical zh-hant
pdc stq
lt sh sr ms vi pt-br ia kg de-ch id vo min ie de-at oc eu pl wo zu mg sw fr-x-nrm frp io
sat lo luz my

← **Obtained by running a clustering algorithm against a Jaccard distance matrix derived from *408 languages* x *~60M* items contigency matrix...**

Machine Learning and Statistics

- t-distributed stochastic neighbor (t-SNE) embedding for dimensionality reduction
- Latent Dirichlet Allocation (LDA) for extracting semantic themes in projects
- Various Clustering algorithms
- Various distance metrics
- etc.


Pontifícia Universidade Católica de Minas Gerais (Q923610)
Dubrovnik (Q1722)
Empire State Building (Q9188)
Great Pyramid of Giza (Q37200)
Offenbach am Main (Q
Library of Congress (Q131454)
French Civil Aviation Unive
74335)
Burj Ahmed Daoud in Year 8Uaq (Q12495)
Berlin Wall (Q5086)
Three Gorges Dam (Q12514)
Lycée
Grand (Q1059546)
United States Naval Academy (Q559549)
Chamonix-Mont-Blanc (Q83236)
Göttingen Academy of Sciences (Q414188)
Pf
Parthenon (Q10288)
83101)
Willis Tower (Q29294)
Palace of Versailles (Q2948)
White House (Q35525)
Sochi (Q39420)
Celle (Q3933)
Sistine Chapel (Q2943)
Acropolis of Athens (Q131013)
(Q172077)
Angkor Wat (Q43473)
Moscow Kremlin (Q133274)
Hagia Sophia (Q12506)
Allianz Arena (Q127429)
Great Wall of China (Q12501)
Potala Palace (Q71229)
Colosseum (Q10285)
Adobe (Q11463)
Bava
e Library
07)
Basilica
The Pentagon (Q11208)
Camp Nou (Q159848)
Topkapı Palace (Q170495)
Hadrian's Wall (Q57357)
Cologne Cathedral (Q4176)
dney Opera House (Q45178)
Schönbrunn Palace (Q131
Forbidden City (Q80290)
St. Basil's Cathedral (Q129846)
Maraca estadium (Q155174)
St. Stephen's Cathedral (Q5943)
Kaaba (Q29466)
St. Vitus Cathedral (Q5949)



External Teams
&
Communication

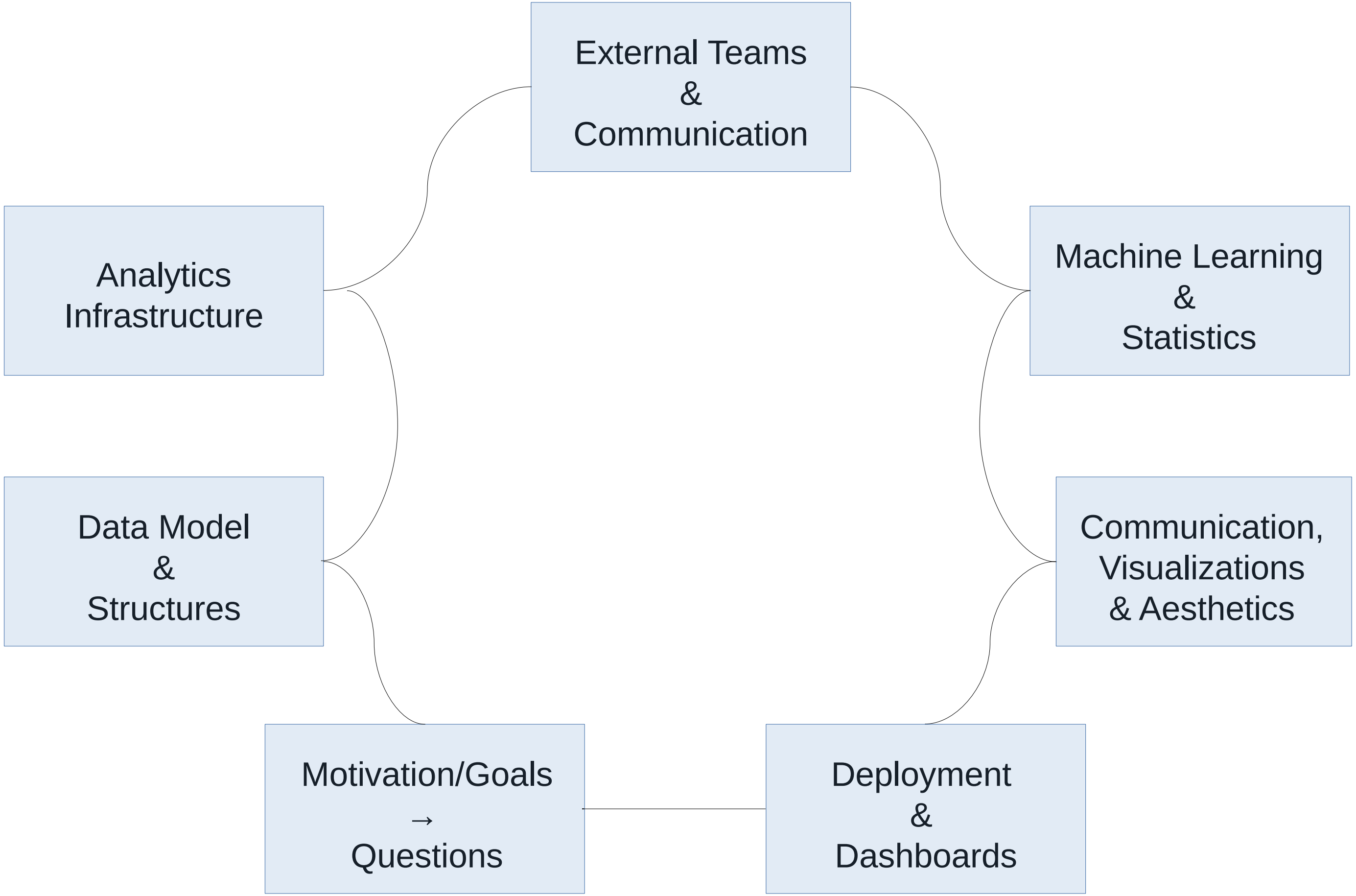
Analytics
Infrastructure

Data Model
&
Structures

Motivation/Goals
→
Questions

Machine Learning
&
Statistics

Communication,
Visualizations
& Aesthetics



External Teams
&
Communication

Machine Learning
&
Statistics

Communication,
Visualizations
& Aesthetics

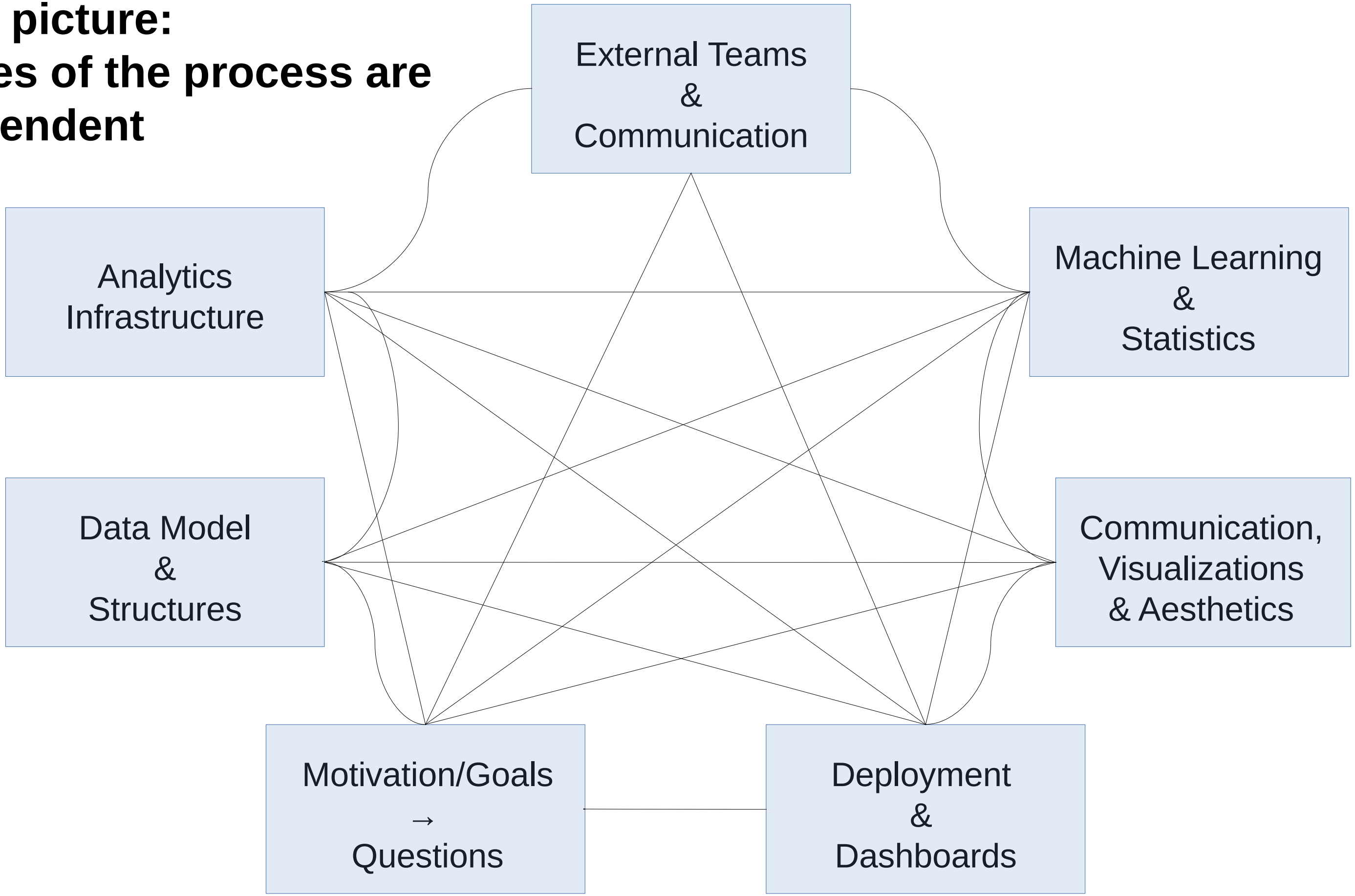
Deployment
&
Dashboards

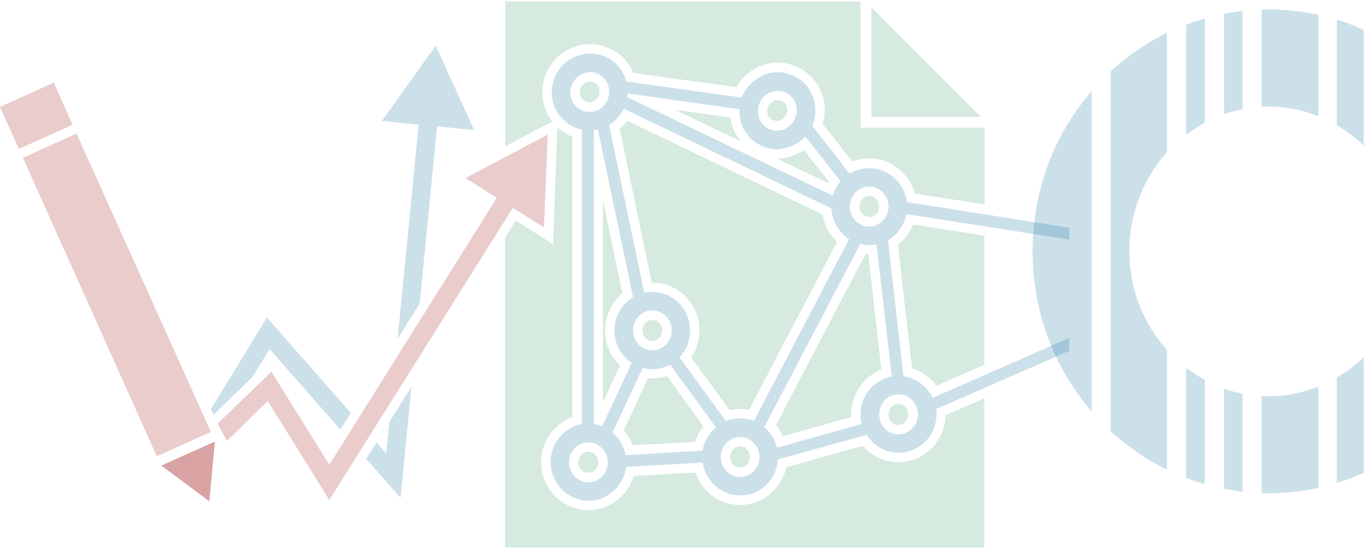
Motivation/Goals
→
Questions

Data Model
&
Structures

Analytics
Infrastructure

The true picture:
all phases of the process are interdependent


External Teams & Communication
Analytics Infrastructure
Machine Learning & Statistics
Data Model & Structures
Communication, Visualizations & Aesthetics
Motivation/Goals → Questions
Deployment & Dashboards



WIKI DATA CON

Wikidata & languages

2019