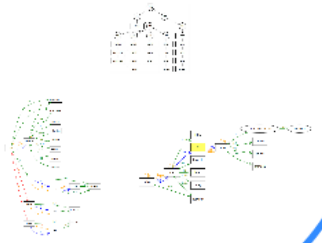
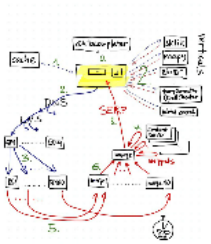
Roman Rožník
seznam fulltext revealed

query processor



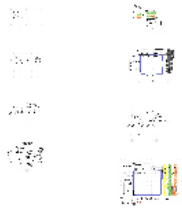
crawler

the Search

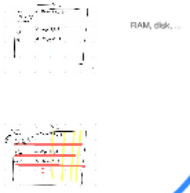


indexing

machine learning



reverse index

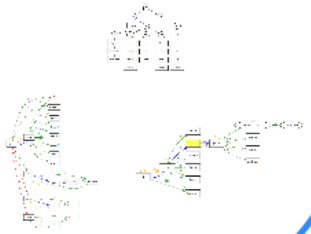


конец

romik@gmail.com
roman.roznik@linmus.seznam.cz
fulltext.sblog.cz
https://www.youtube.com/user/DusanJaninsky
http://cyber.felk.cvut.cz/research/theses/papers/471.pdf

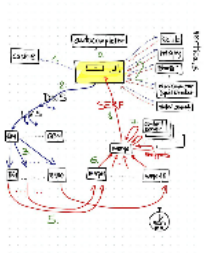
Roman Rožník
seznam fulltext revealed

query processor



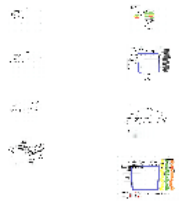
crawler

the Search



indexing

machine learning

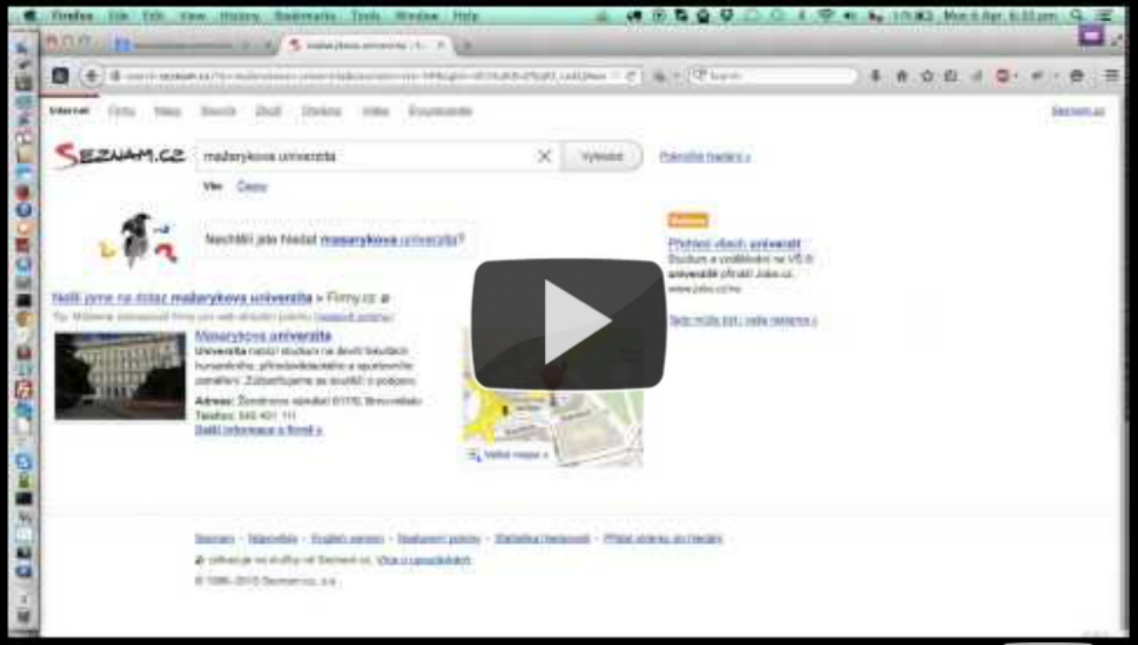


reverse index

RAM, disk, ...

конец

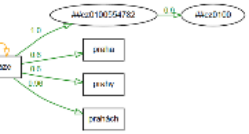
rozrik@gmail.com
roman.roznik@firma.seznam.cz
fulltext.sblog.cz
https://www.youtube.com/user/DusanJanousky
http://cyber.felk.cvut.cz/research/theses/papers/472.pdf

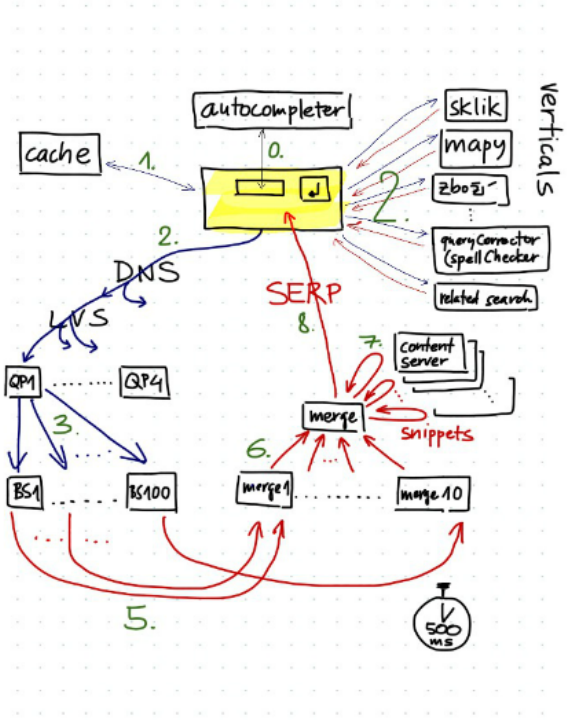


YouTube

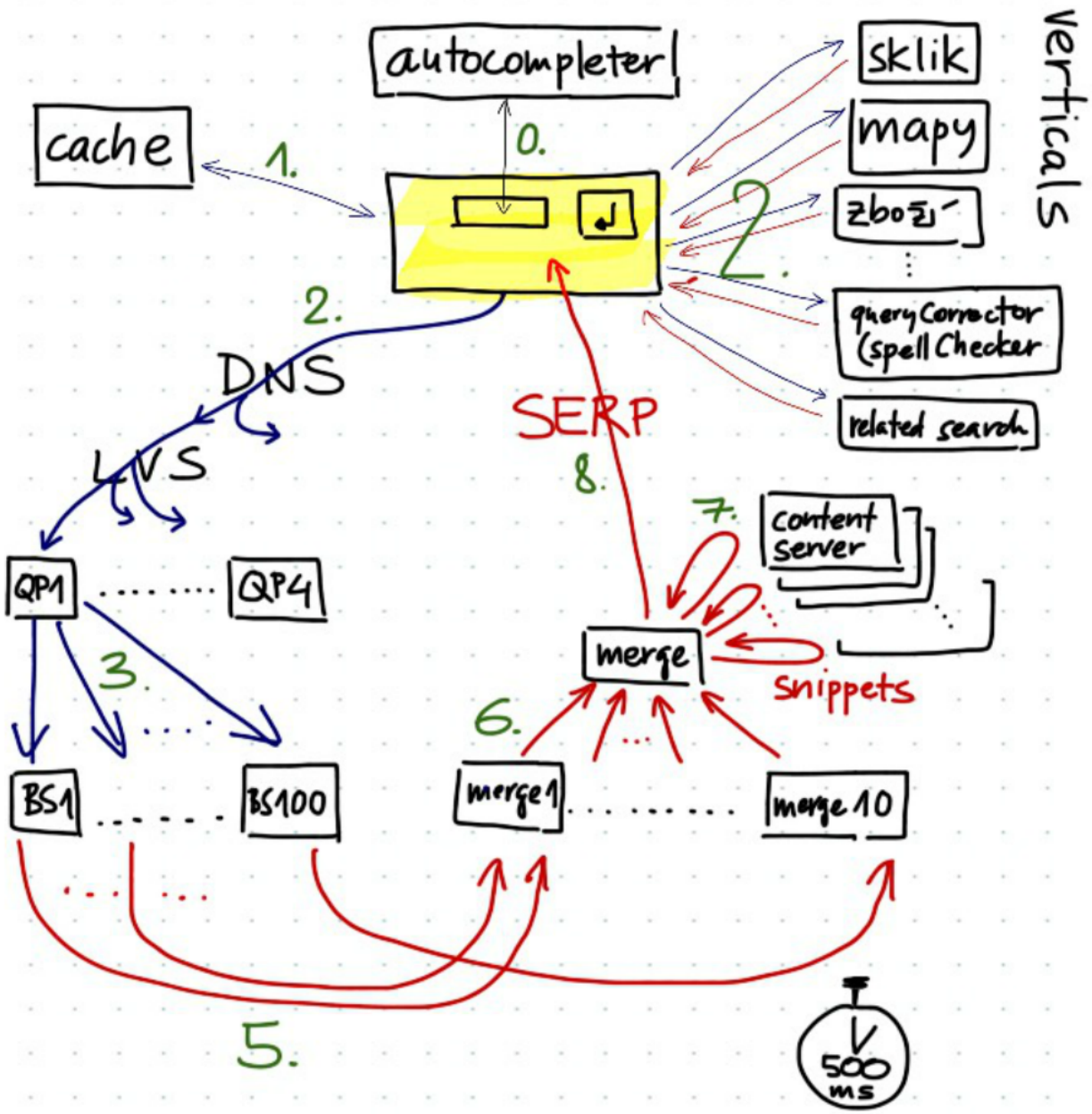
Prezi

the Search

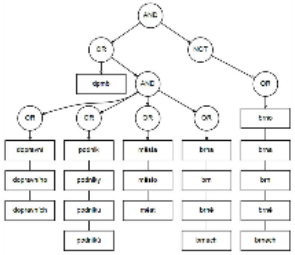


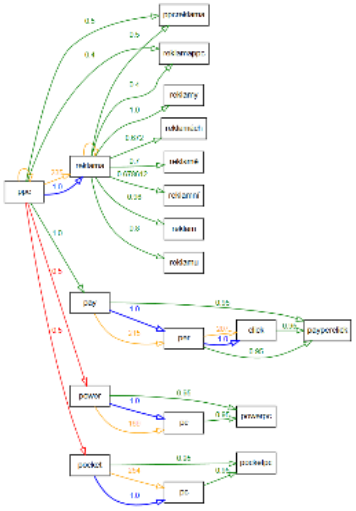


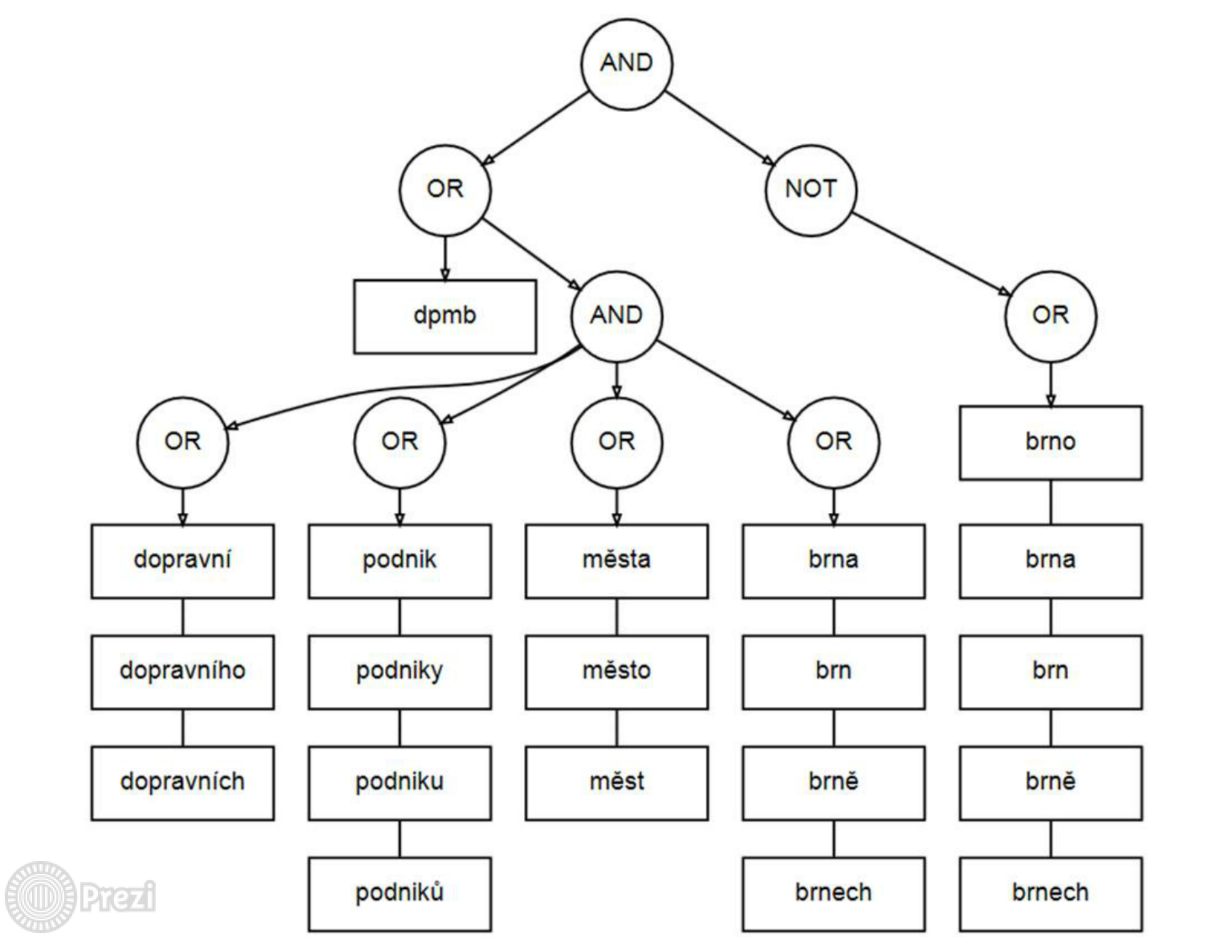
arning

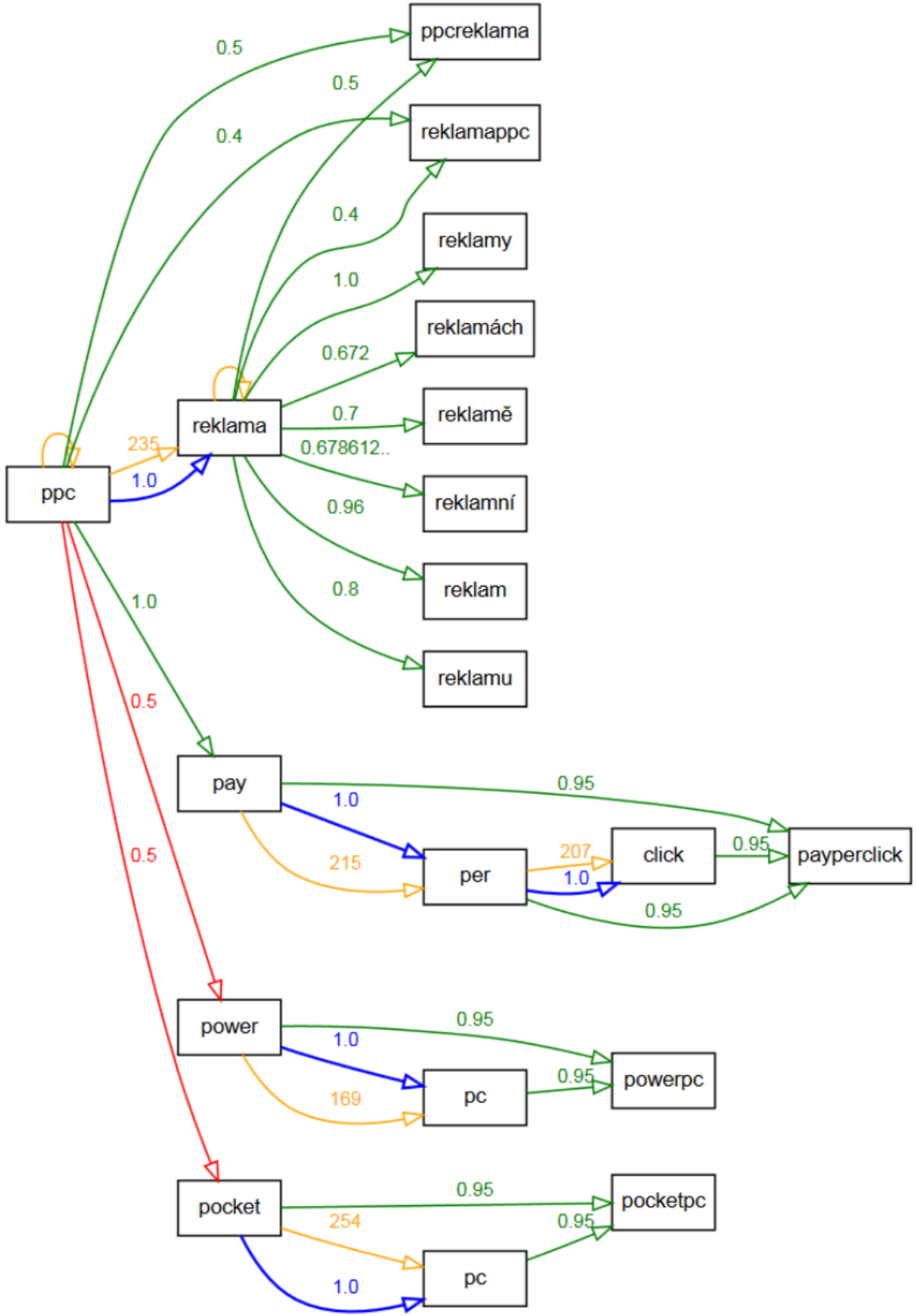

autocompleter
cache
1.
0.
sklik
mapy
zboží
2.
verticals
queryCorrector
(spell Checker
related search
2.
DNS
SERP
8.
LVS
7.
content
server
QP1
QP4
merge
snippets
3.
6.
BS1
BS100
merge1
merge 10
5.
500
ms

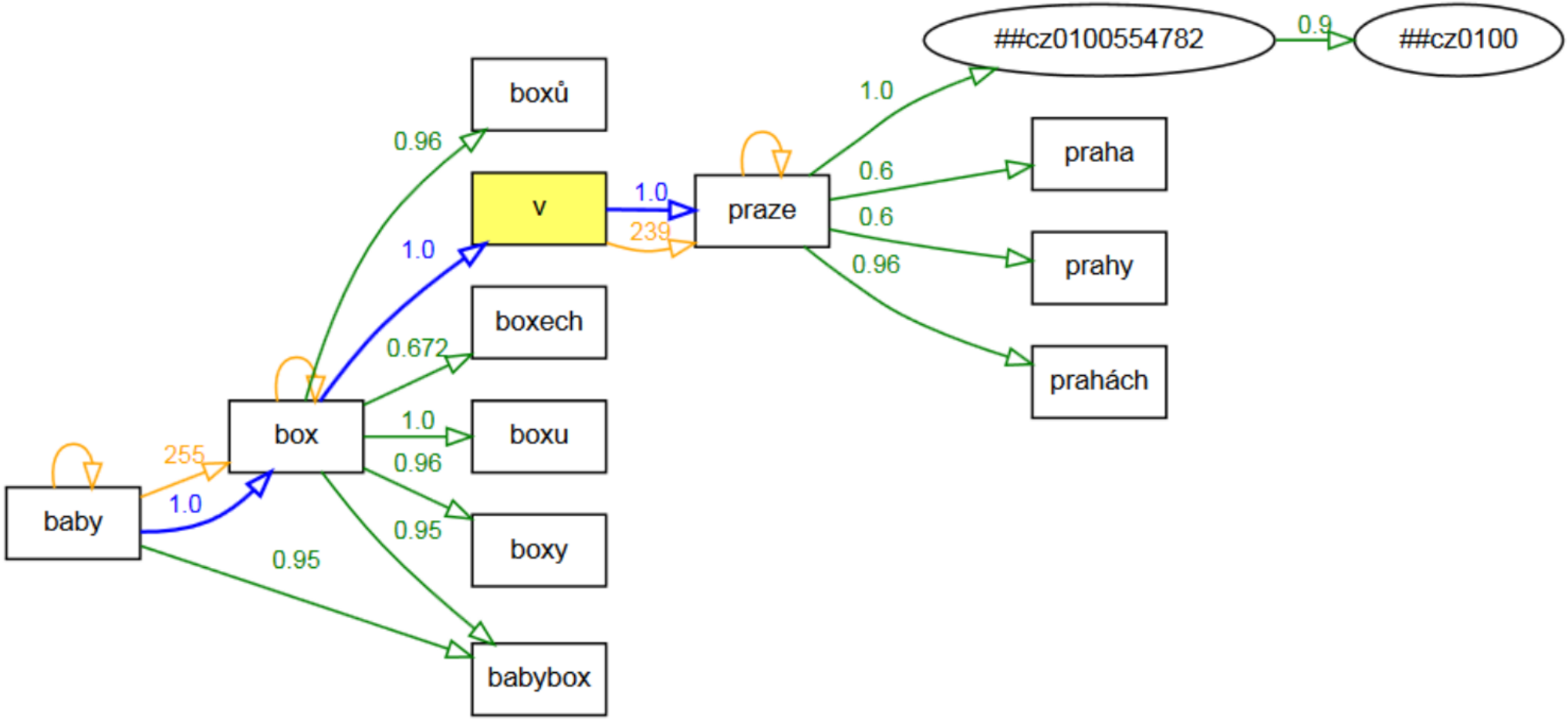
query processor

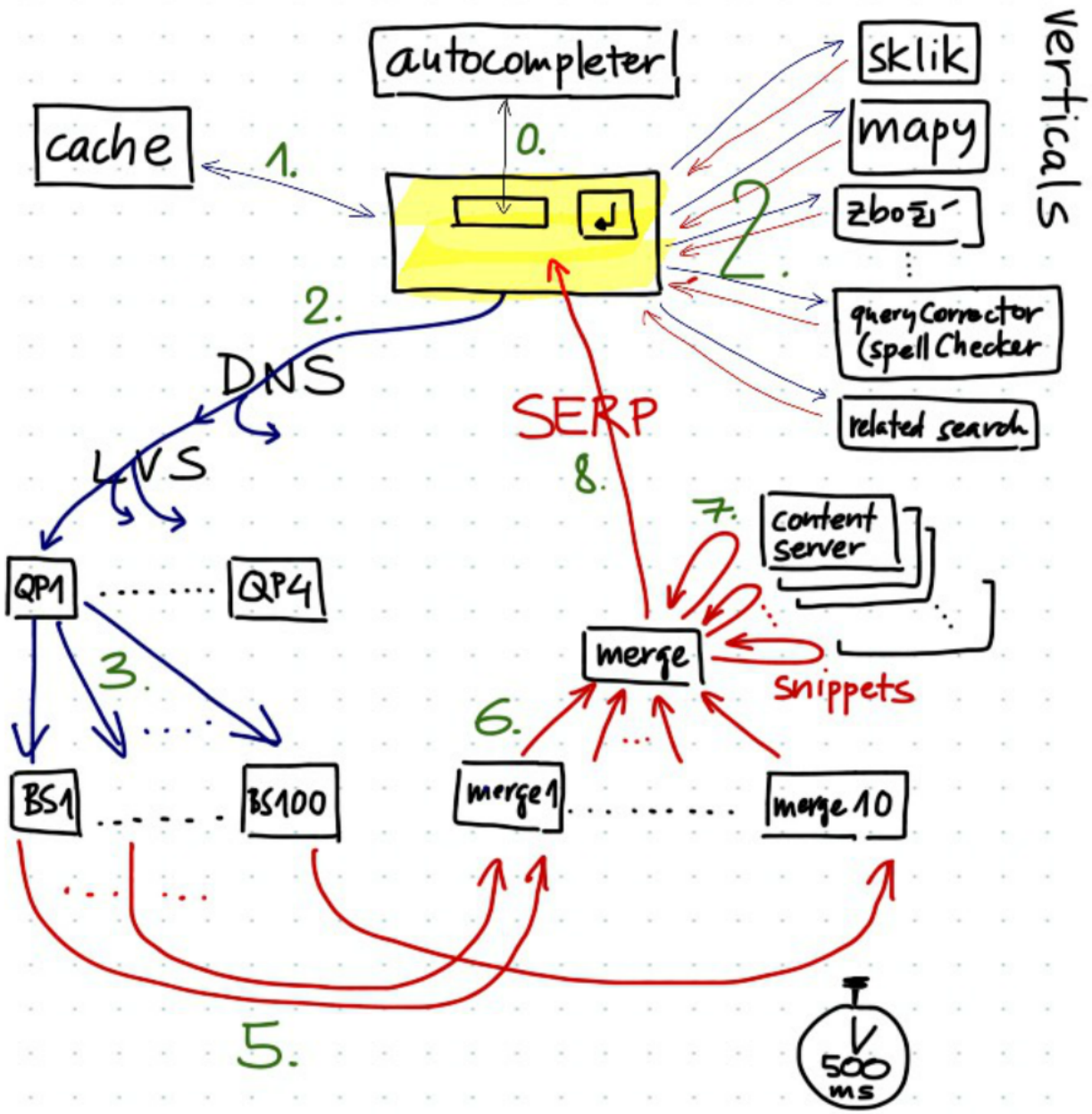




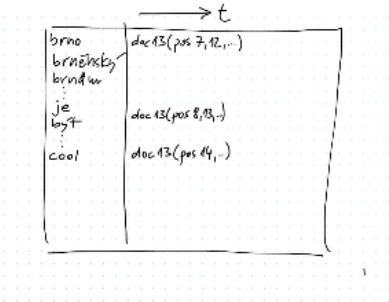

AND
OR
NOT
dpmb
AND
OR
OR
OR
OR
OR
brno
dopravní
podnik
města
brna
brna
dopravního
podniky
město
brn
brn
dopravních
podniku
měst
brně
brně
podniků
brnech
brnech


ppcreklama
reklamappc
reklamy
reklamách
reklamě
reklamní
reklam
reklamu
reklama
ppc
pay
per
click
payperclick
power
pc
powerpc
pocket
pc
pocketpc
0.5
0.5
0.4
0.4
1.0
0.672
0.7
0.678612..
0.96
0.8
235
1.0
1.0
0.5
0.5
0.95
1.0
215
207
1.0
0.95
0.95
0.95
1.0
169
0.95
0.95
254
1.0

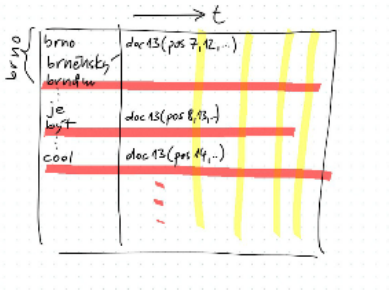

baby
box
boxů
v
boxech
boxu
boxy
babybox
praze
##cz0100554782
##cz0100
praha
prahy
prahách
255
1.0
0.95
0.96
1.0
0.672
1.0
0.96
0.95
1.0
239
1.0
0.6
0.6
0.96
0.9

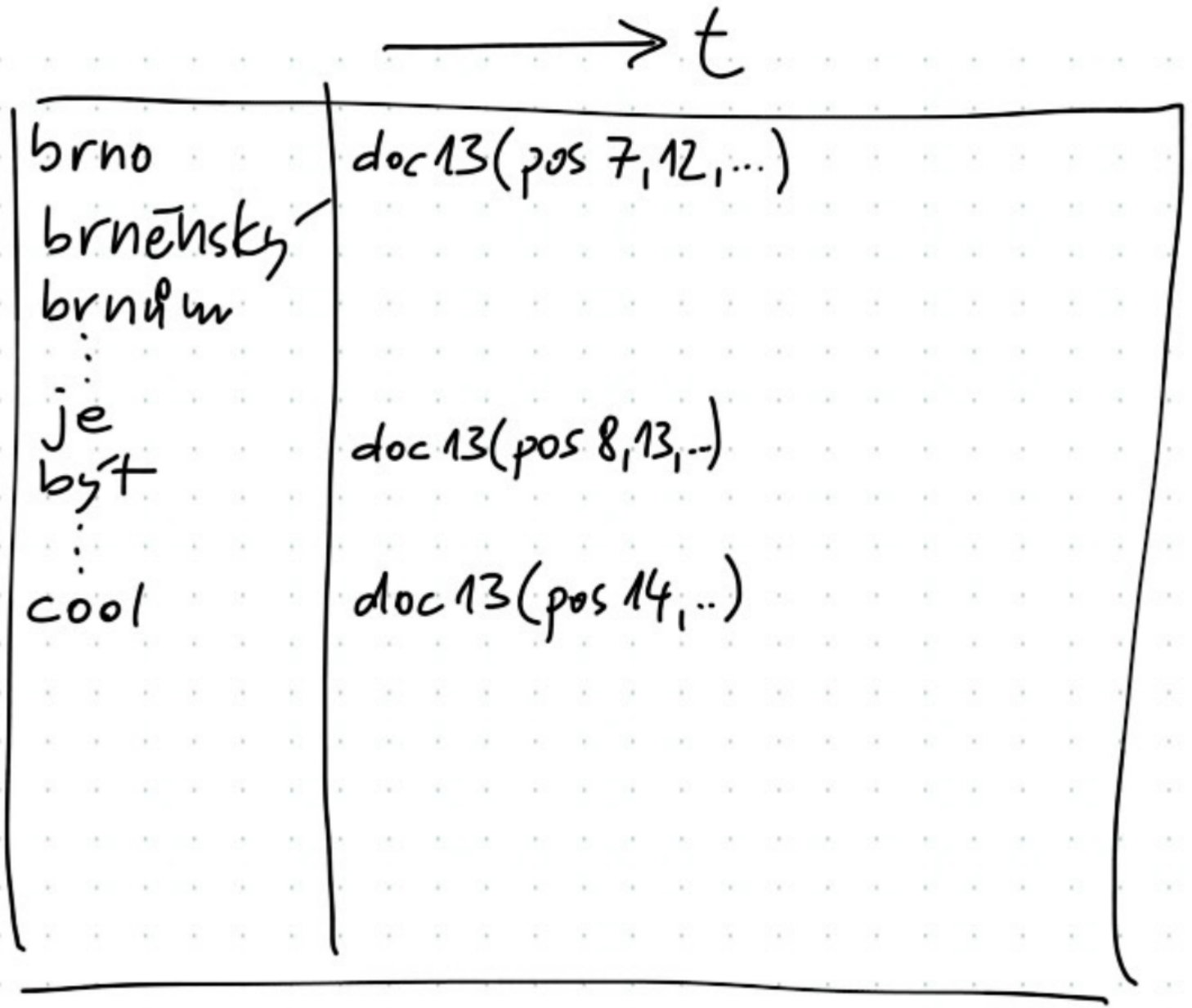

autocompleter
cache
1.
0.
sklik
mapy
zboží
2.
verticals
queryCorrector
(spell Checker
related search
2.
DNS
SERP
8.
LVS
7.
content
server
QP1
QP4
merge
snippets
3.
6.
BS1
BS100
merge1
merge 10
5.
500
ms

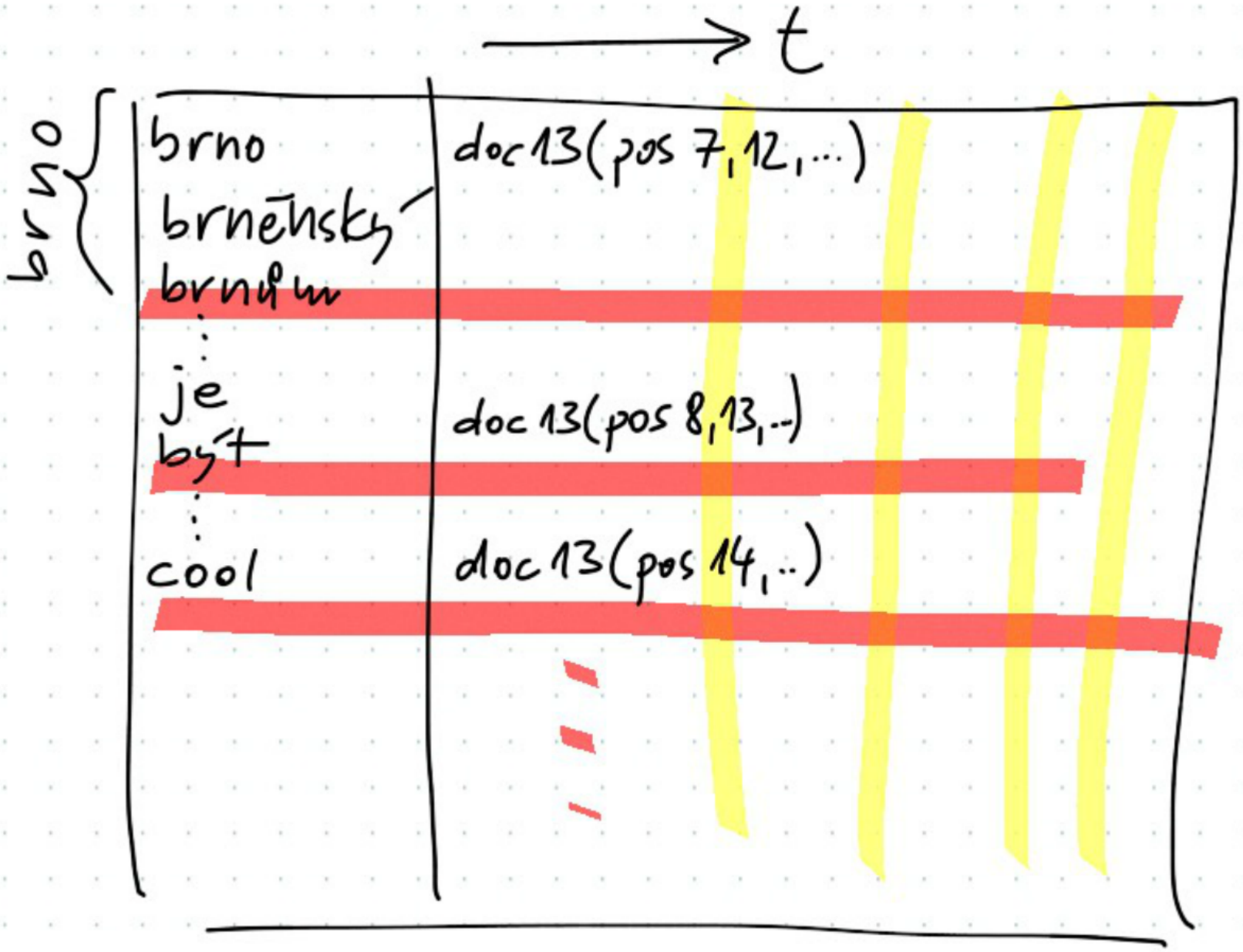
reverse index



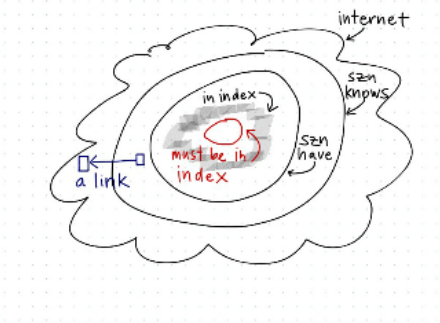
RAM, disk, ...




t
brno
brněnský
brněnům
je
být
cool
doc13(pos 7,12,...)
doc13(pos 8,13,..)
doc13(pos 14,..)

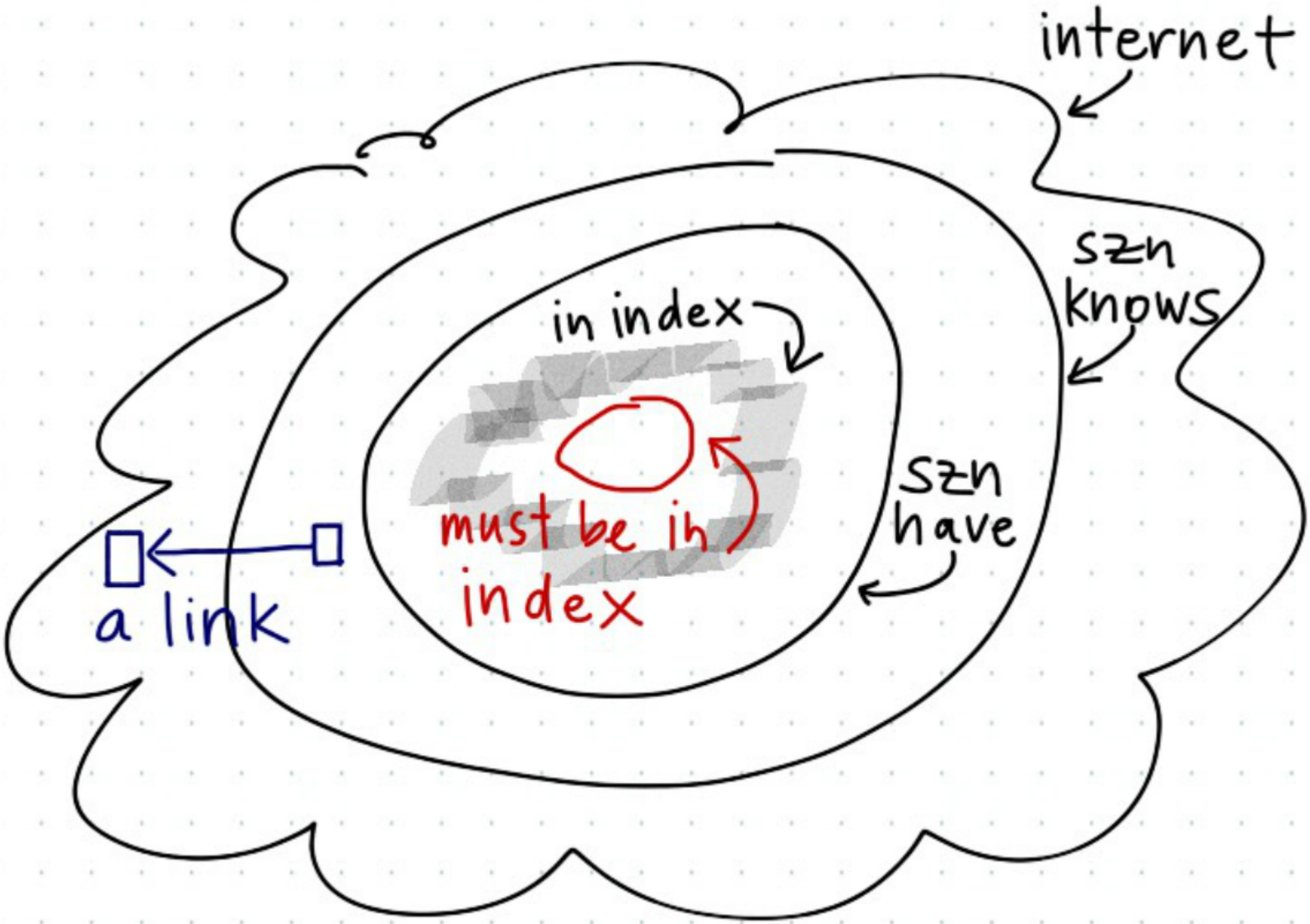

t
brno
brno
brněnský
brněnům
je
být
cool
doc 13 (pos 7, 12, ...)
doc 13 (pos 8, 13, ..)
doc 13 (pos 14, ..)

crawler



www.fi.muni.cz/usr/pelikan

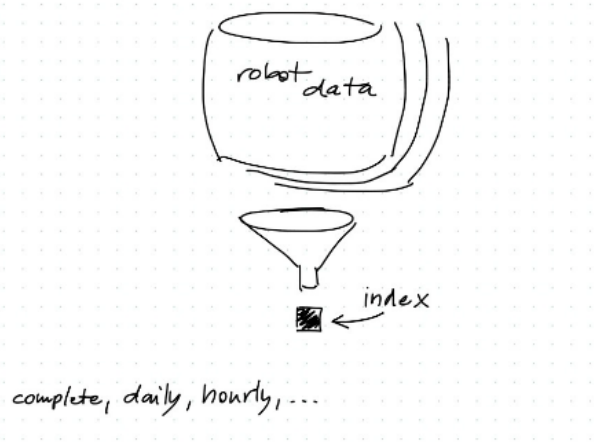
1.1.1970	500
2.3.1998	200
2.9.1998	304
2.9.1999	304
2.9.2002	304
2.9.2008	304
≈2020	planned

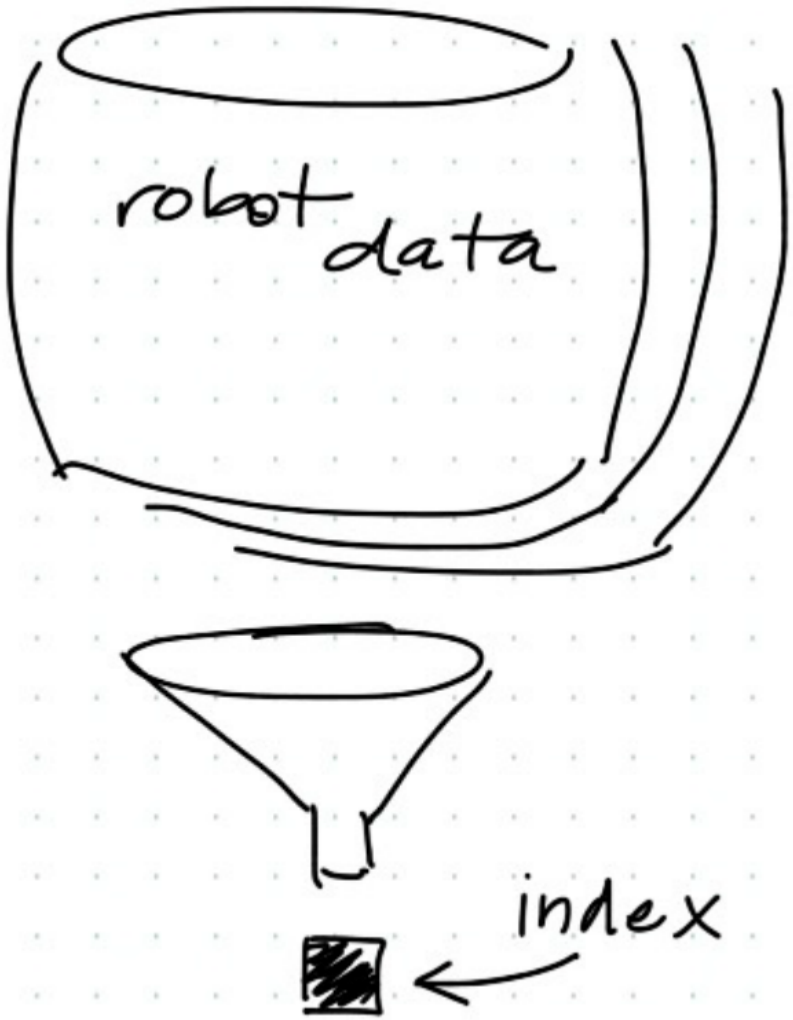

internet
szn knows
in index
must be in index
szn have
a link

www.fi.muni.cz/usr/pelikan

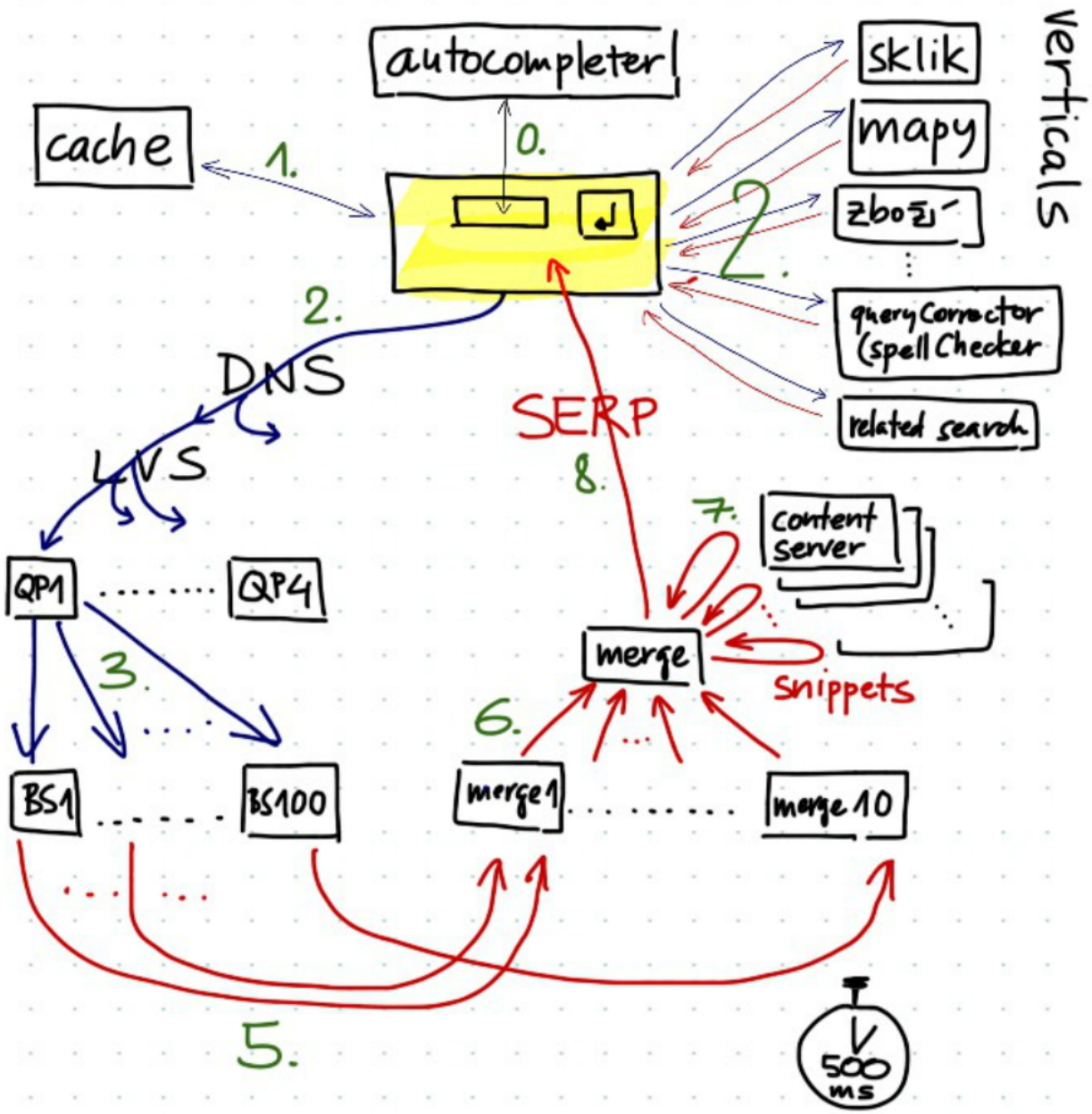
1.1.1970	500
2.3.1998	200
2.9.1998	304
2.9.1999	304
2.9.2002	304
2.9.2008	304
≈2020	planned

indexing

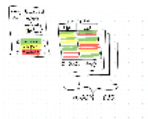


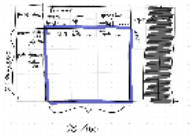
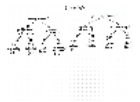


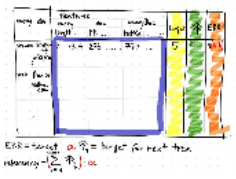
complete, daily, hourly, ...


autocompleter
cache
1.
0.
sklik
mapy
zboží
2.
verticals
queryCorrector
(spell Checker
related search
2.
DNS
SERP
8.
LVS
7.
content
server
QP1
QP4
merge
snippets
3.
6.
BS1
BS100
merge1
merge 10
5.
500
ms

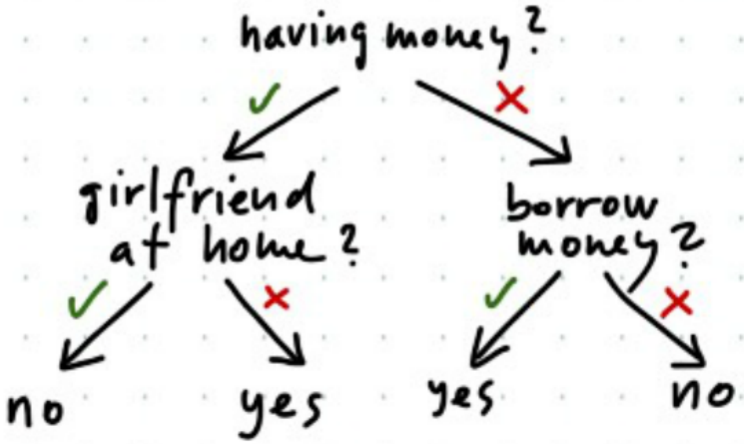
machine learning

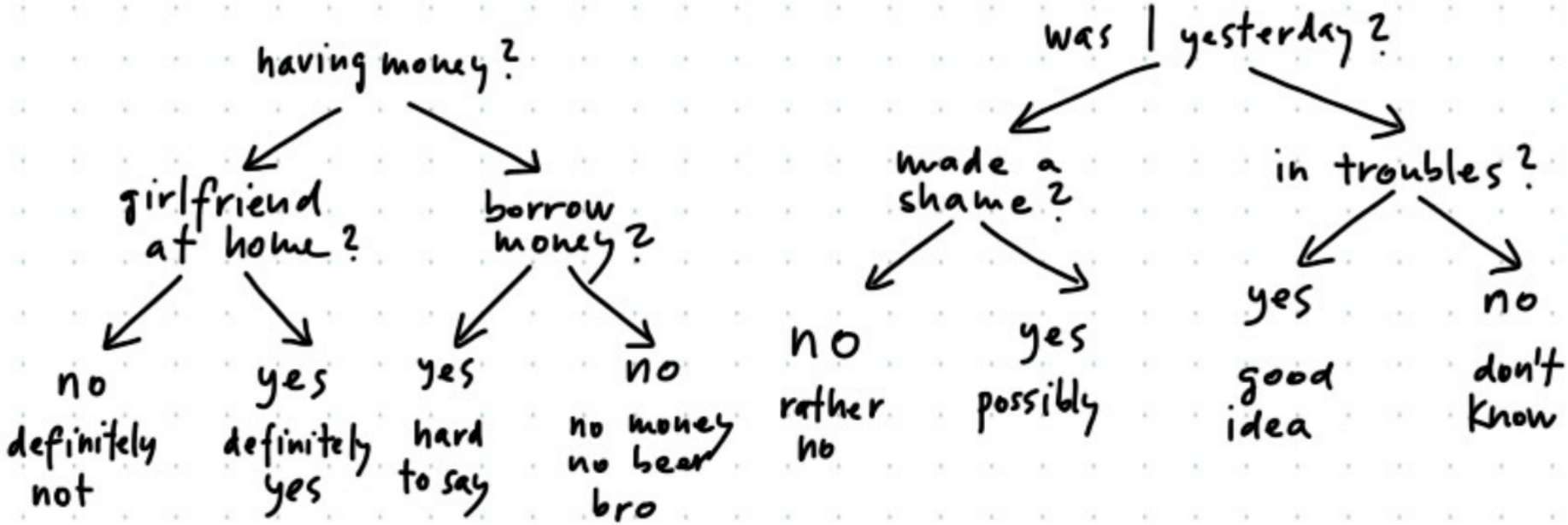



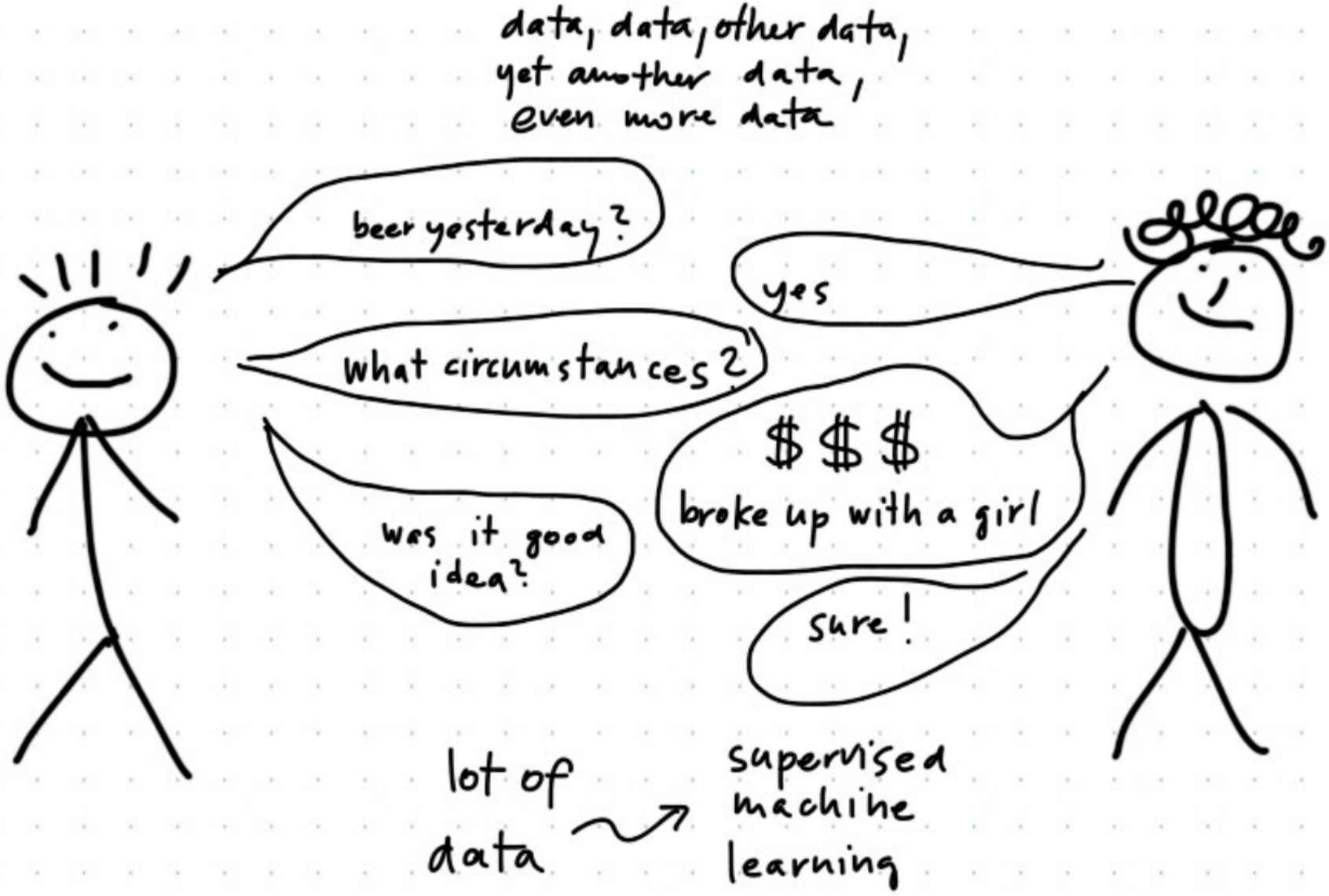
beer today?

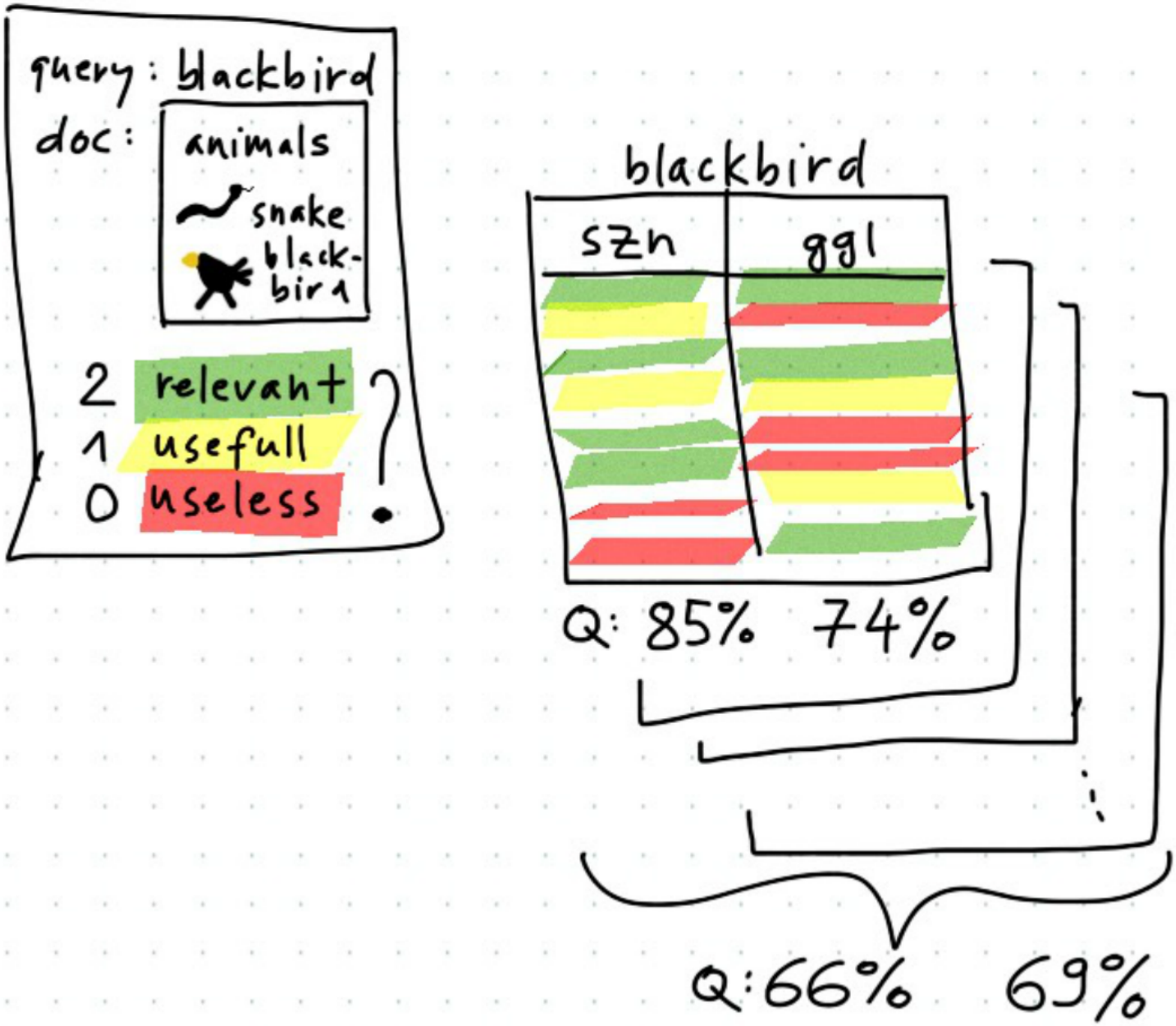


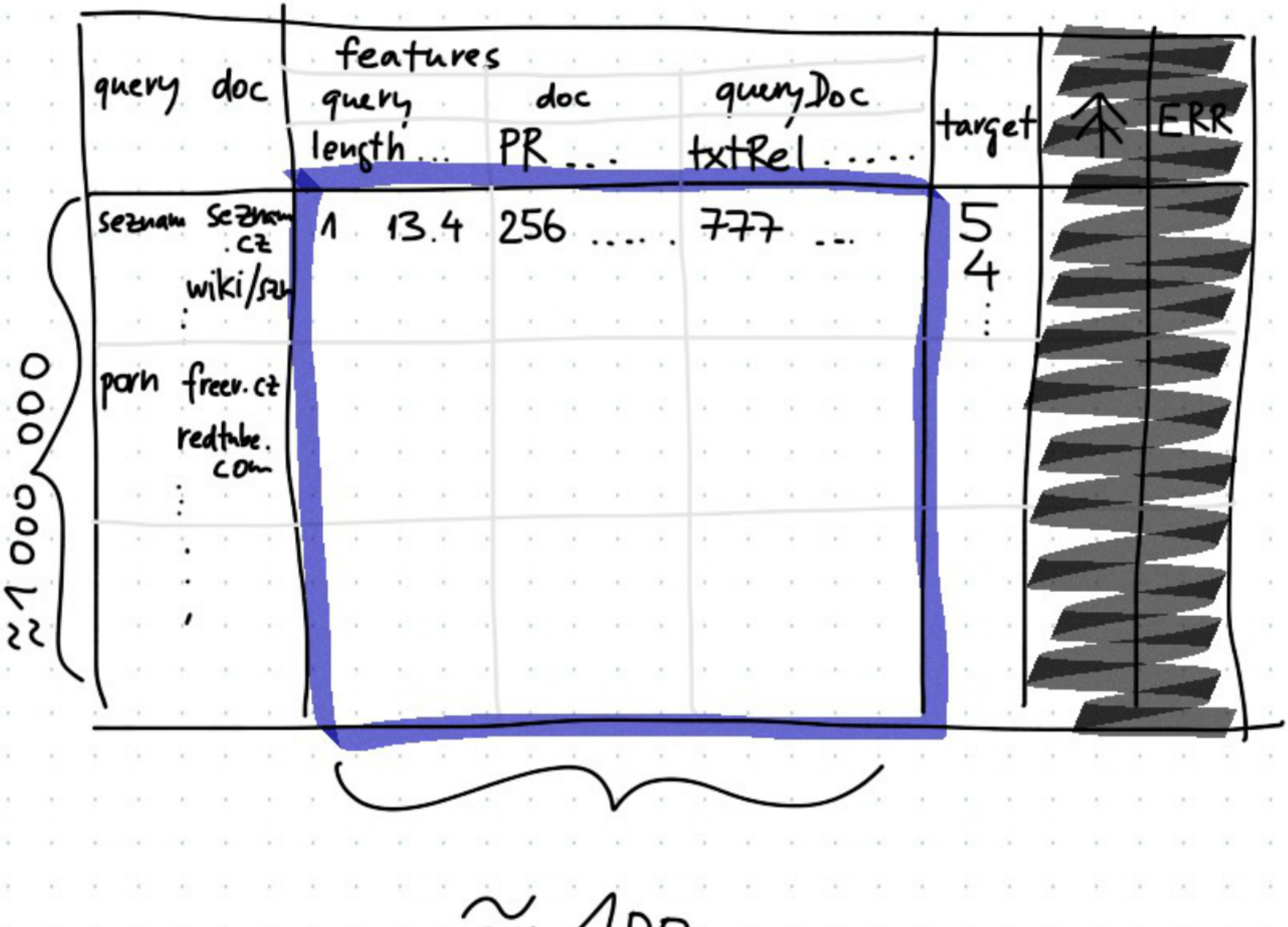

beer today?
having money?
girlfriend at home?
borrow money?
no
yes
yes
no
was I yesterday?
made a shame?
in troubles?
no
yes
yes
no

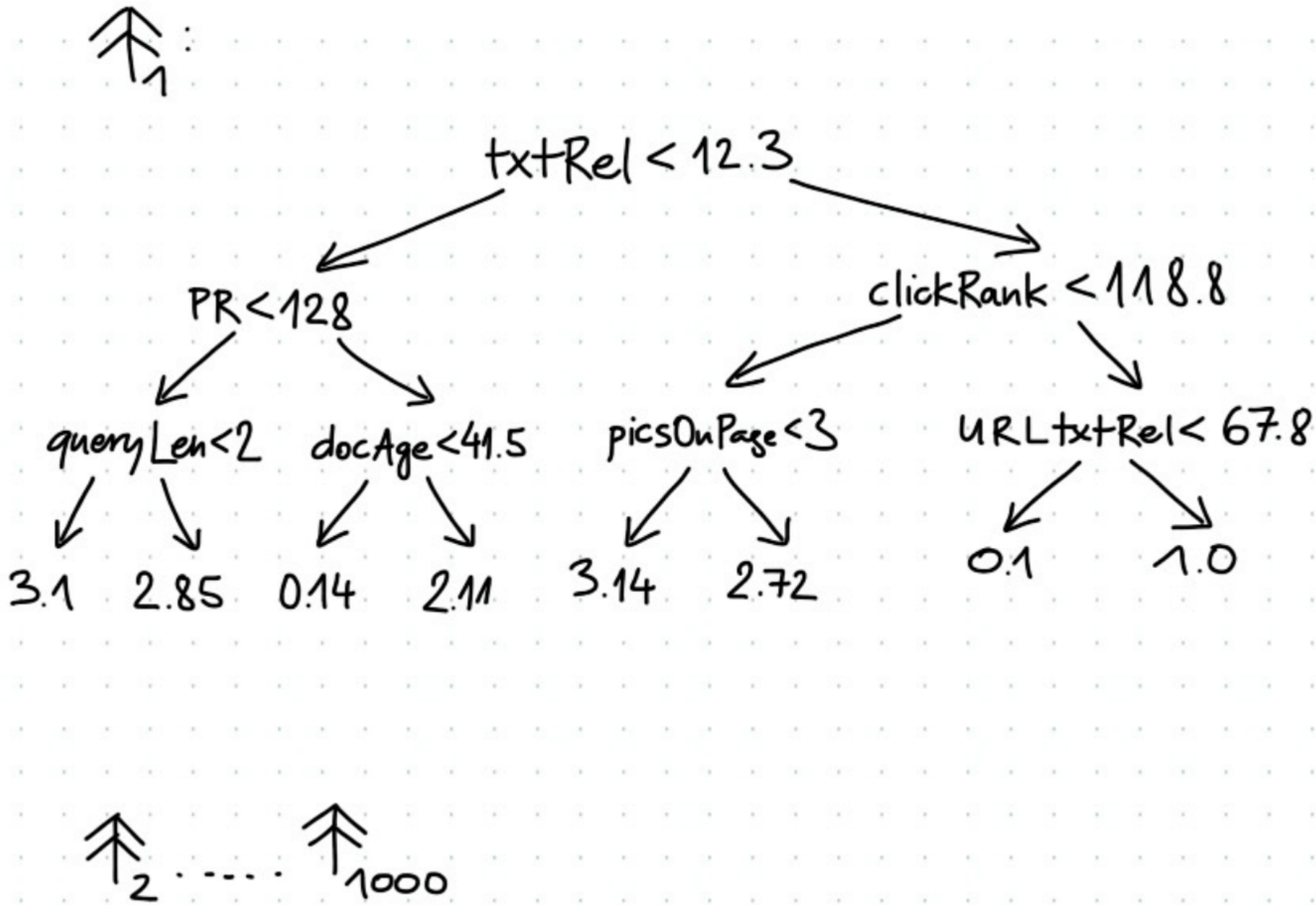
beer today?

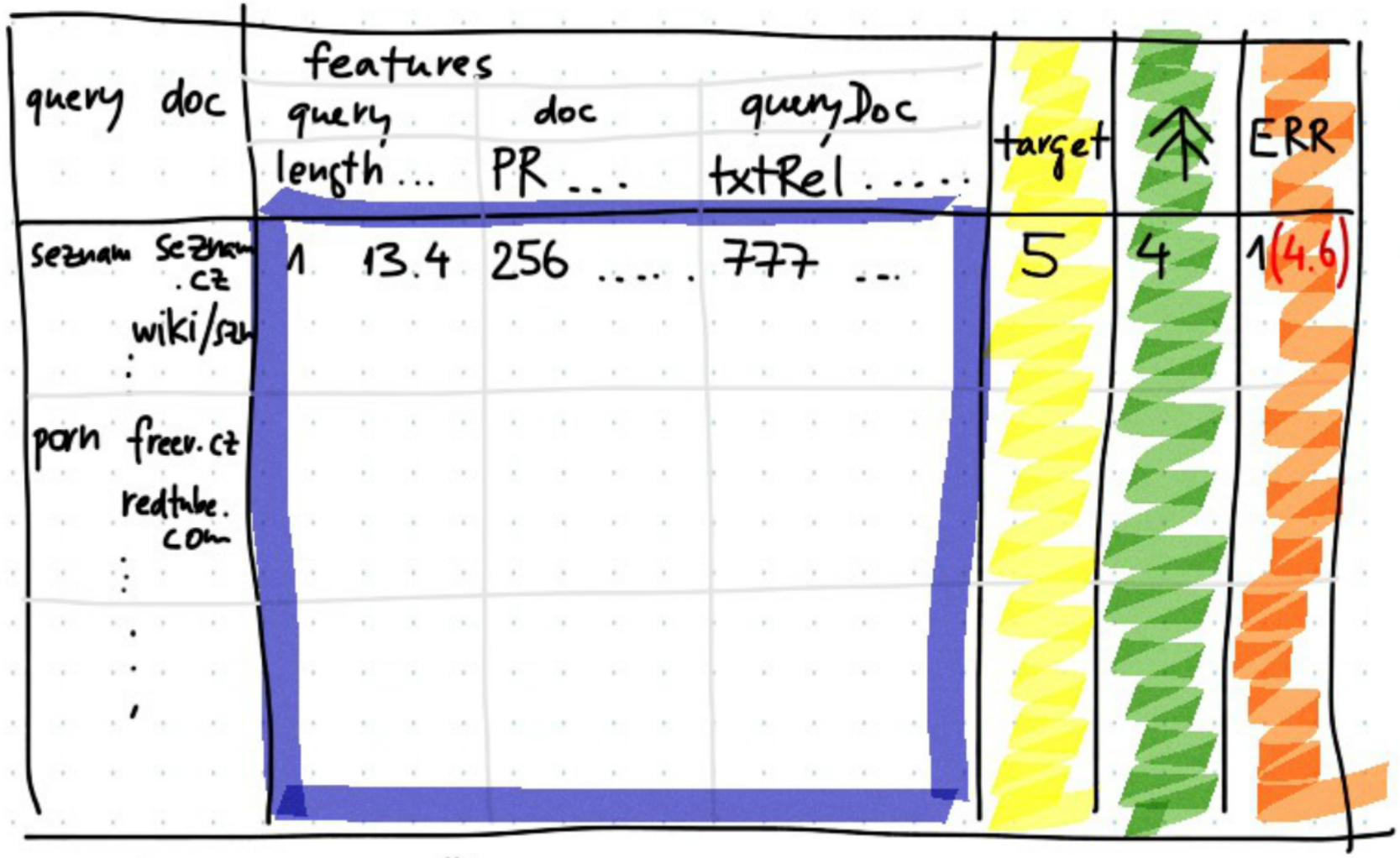



data, data, other data,
yet another data,
even more data
beer yesterday?
yes
what circumstances?
$ $ $
broke up with a girl
was it good idea?
sure!
lot of data
supervised machine learning


query : blackbird
doc :
animals
snake
black-
bird
2 relevant
1 usefull
0 useless
?
blackbird
szn
ggl
Q: 85% 74%
Q: 66% 69%

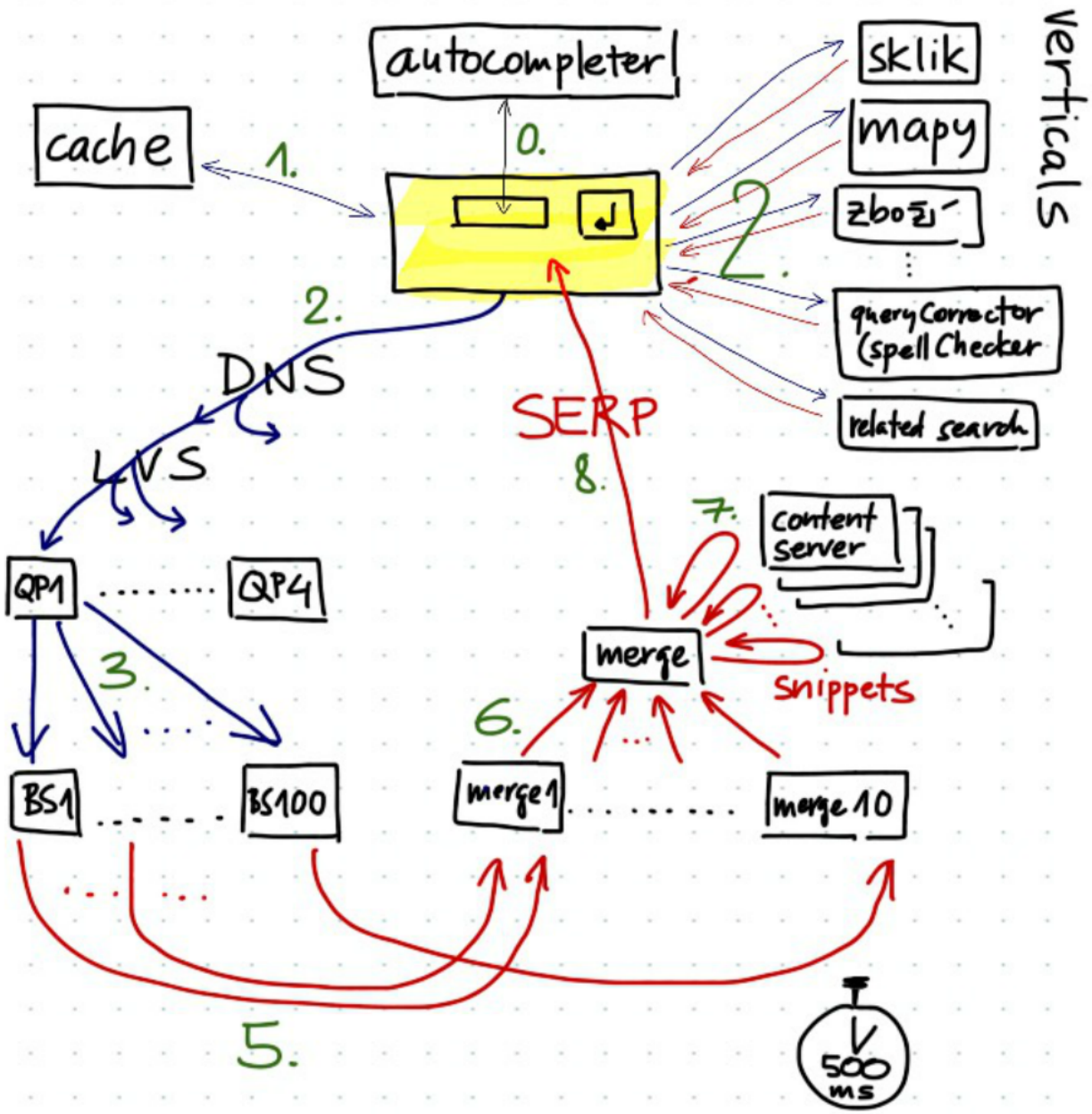

query
doc
features
query
doc
queryDoc
length ...
PR ...
txtRel .....
target
ERR
seznam
Seznam .cz
wiki/sz
porn
freer.cz
redtube. com
1
13.4
256 .....
777 ...
5
4
≈1 000 000
≈ 100


1
txtRel < 12.3
PR<128
clickRank < 118.8
queryLen<2
docAge<41.5
picsOnPage<3
URLtxtRel< 67.8
3.1
2.85
0.14
2.11
3.14
2.72
0.1
1.0
2
1000



$$ERR = target - \alpha \cdot \Uparrow_1 = target\ for\ next\ tree$$

$$relevancy = \left(\sum_{i=1}^{1000} \Uparrow_i\right) \cdot \alpha$$


autocompleter
cache
1.
0.
sklik
mapy
zboží
2.
verticals
queryCorrector
(spell Checker
related search
2.
DNS
SERP
8.
LVS
7.
content
server
QP1
QP4
merge
snippets
3.
6.
BS1
BS100
merge1
merge 10
5.
500
ms

конец

roznik@gmail.com
roman.roznik@firma.seznam.cz
fulltext.sblog.cz
https://www.youtube.com/user/DusanJanovsky
http://cyber.felk.cvut.cz/research/theses/papers/471.pdf