

Basic classification of second-level domains in .CZ.

IT20 conference

Maciej Andziński • maciej.andzinski@nic.cz • 12. 11. 2020

Motivation

- There are **1.36M+** .CZ domains, but how many remain „active“?



Motivation

- There are **1.36M+** .CZ domains, but how many remain „active“?
- What is an „active domain“? Our assumption:
 - Hosts a non-parking website
or
 - Runs a mail server



Motivation

- There are **1.36M+** .CZ domains, but how many remain „active“?
- What is an „active domain“? Our assumption:
 - Hosts a non-parking website
or
 - Runs a mail server
- Let's use DNS crawler* data

*<https://pypi.org/project/dns-crawler/>



DNS crawler results

- DNS crawler scan on **23 October 2020**
 - Scanned **1,365,753** .CZ domains
 - Domain list as of **22 Oct 2020 23:59:59 UTC**
 - Gathered **200 GB** of raw data




Web content scan

- For each domain we collected web content for various combinations of:
 - IP version (**IPv4/IPv6**)
 - Web server ports (**80/443**)
 - Prefix labels (empty/**www**)
- We followed all HTTP redirections
- The longest content was used for further analysis



Detecting parking websites



Doména je zaregistrována


Tato doména je zaregistrována a funkční. Na www.WEDOS.com si můžete objednat webhosting. V zákaznickém centru na adrese client.wedos.com je možné doménu spravovat a případně si zdarma aktivovat minisweb.

Další informace a návody nalezte ve [znalostní bázi WEDOS](#).

This domain is registered

[Hosting WEDOS](#) - [registrace domén](#) [webhosting](#) [serverhosting](#)

This page is parked in Subreg.CZ
Tato stránka je zaparkována u Subreg.CZ



Nabízíme

- Registrace domén celého světa
- Plně automatická administrace
- Zajištění lokálních kontaktů
- Prodej domén přes síť AfterNIC
- Aukce domén, Escrow služby
- Reseller a Registrar Panel

- Nízké ceny domén na trhu
- Speciální registrace CZ, COM a NET domén
- Rozšířené API - SOAP, EPP
- 10 MB FTP, DNS, Přesměrování
- Parking domén
- Whitelabel systém



Webová prezentace k této doméně nebyla zatím vytvořena.


Webhosting k této doméně poskytl společnost CZECH MULTIMEDIA INTERACTIVE.

Všechny a databázový server je spravován [webem Panelu, Panelu, Panelu](#) celkověovou jedižkou na poli hostingového softwaru.

Tato stránka však, protože je doménou pouze zastupována a není k ní vytvořena žádná prezentace.

Malový server je obsluhován poslední věcí: [SmartMail Enterprise serveru](#) - jedižkou v oblasti správy emailových schránek.

Více informací o [CZECH MULTIMEDIA INTERACTIVE](#) naleznete na našich stránkách www.cmi.cz



active24.CZ

Tato doména je zaregistrována prostřednictvím doménového portálu **ACTIVE24**.

REGISTRACE DOMÉNY Ověřte si dostupnost domény

Parkování

CZ Tato doména je parkována u hostingových služeb gigaserver.cz. Majitel se ještě nerozhodl využít této domény a zatím zde není ani k nábízení.

SK Tato doména je parkována u hostingových služeb gigaserver.cz. Majitel se zatím nerozhodl využít této domény, a tak je momentálně neaktivní.

EN This domain is located at the hosting services of gigaserver.cz. Its owner has not decided to exploit this domain yet and for the time being there is nothing to be found here.

DE Diese Domäne ist reserviert für Host Leistungen bei „gigaserver.cz“. Der Nutzer hat über Verwertung und Inhalt der Domäne noch nichts entschieden.

GIGASERVER
» [výkonajet](#)
» [gigaserver.cz](#)
» [server.cz](#)

© GIGASERVER.CZ Webhosting provozovaný společností Server Multimedia s.r.o.



Doména [redacted].CZ
je parkována u CZECHIA.COM

Registrujte si doménu

Zadejte název domény...

Domény v akci **CZ 169 Kč | EU 69 Kč | COM 269 Kč | INFO 129 Kč | WEBSITE 99 Kč**

Ceny jsou uvedeny bez DPH a platí na první rok registrace.



THE DOMAIN NAME IS REGISTERED

☒ DOMÉNA JE ZAREGISTROVÁNA
☒ DOMÉNA JE ZAREGISTROVÁNA
☒ DOMÉNA JE ZAREGISTROVÁNA
☒ DOMÉNA JE ZAREGISTROVÁNA
☒ DOMÉNA JE ZAREGISTROVÁNA

Prohlédněte si nabídku domén a webhostingu na stránce: <https://www.forpsi.com/en/offer/>

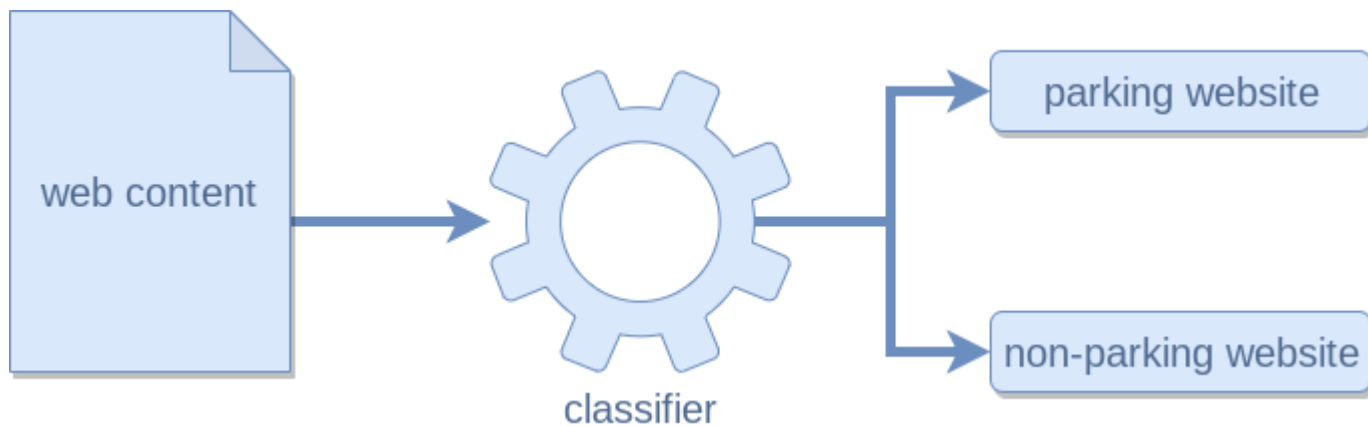
DOMÉNA JE REGISTROVÁNA

[Kontakt](#)



Detecting parking websites

- Machine Learning methods



Detecting parking websites

- **Classifier input**
 - Preprocessed web content
 - Visible text extracted from HTML
 - No JS rendering



Detecting parking websites

- **Classifier output**

- Class label

- **Parking web** - web content was classified as parking website
 - **Active web** - web content was classified as non-parking website
 - **No content** - web server was unreachable or if the web content was empty
 - **HTTP error** - web server answered with 4xx or 5xx HTTP status code



Detecting parking websites

- **Classifier**
 - **Random forest** based on **TF-IDF**
 - Term frequency in a document versus term frequency in corpus
 - Facilitates identifying important terms
 - Removed *Czech*, *Slovak*, *English* and *Polish* stop words
 - 1-3 level n-grams of words
 - Trained using data from domain report
 - ~600 manually classified domains
 - Balanced dataset



Detecting parking websites

- **Classifier performance**

- F score = **0.92**

- Accuracy **92%**

- Web content was classified correctly for **92** out of **100** domains in a batch (parking website versus non-parking website)



Detecting parking websites

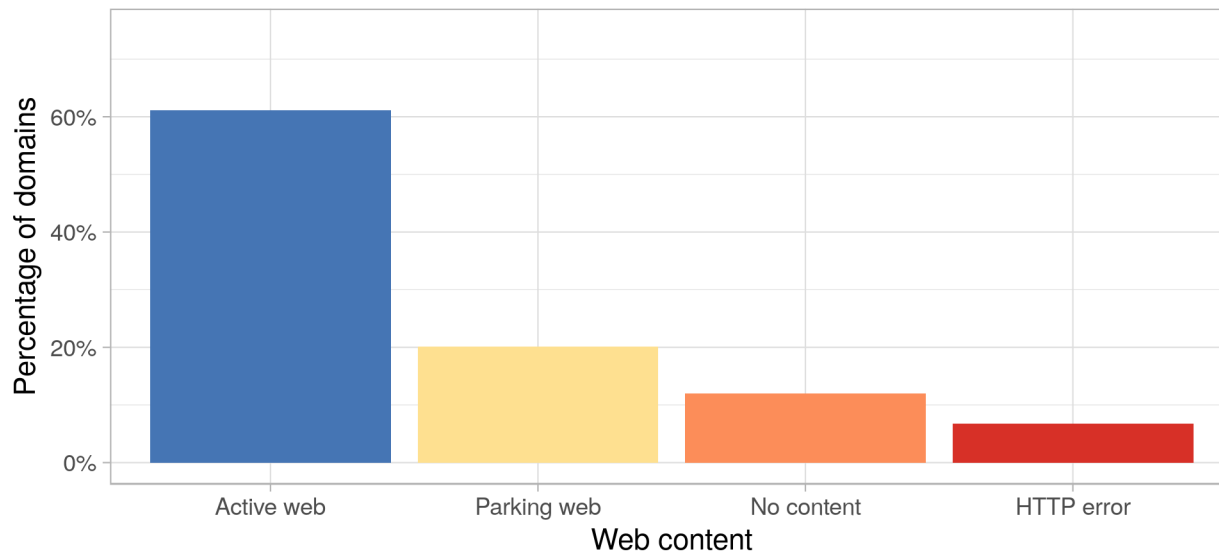
- **Classifier findings**

- Popular terms

- Parking websites: *“domena zaregistrovana”* (Czech: *“domain registered”*)
 - Non-parking websites: *“vsechna prava vyhrazena”* (Czech: *“all rights reserved”*)



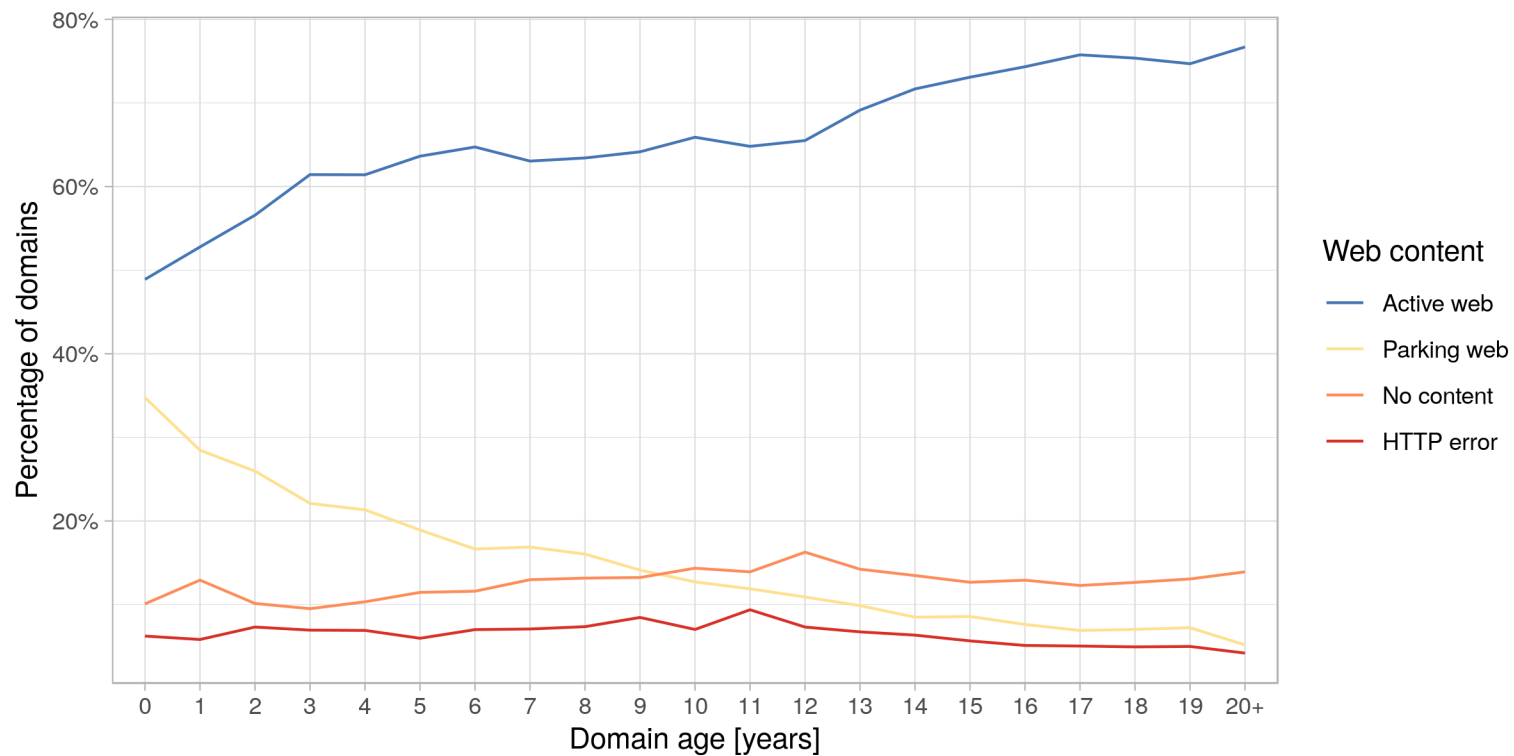
Results – web content



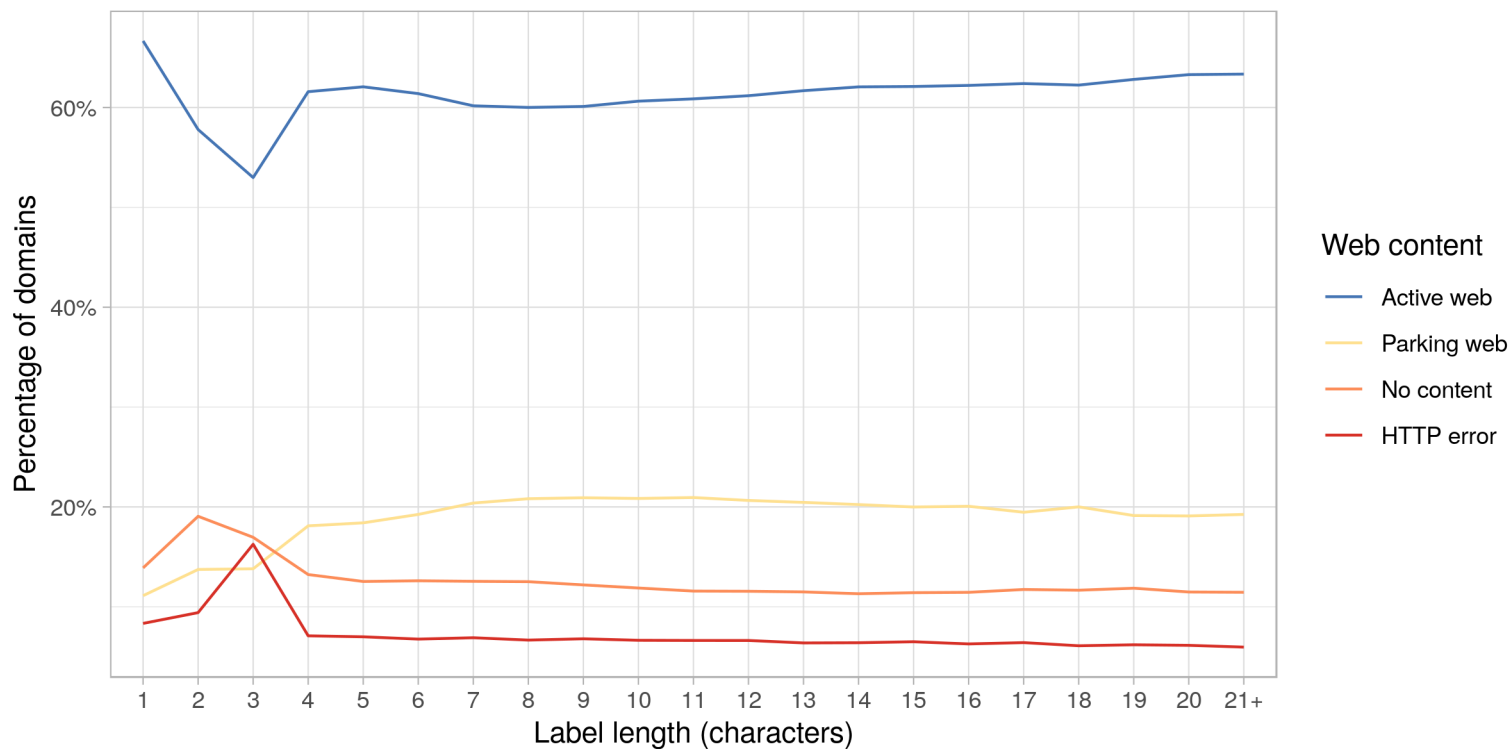
- **61.1%** domains have active web (host a non-parking website)



Results – web content



Results – web content

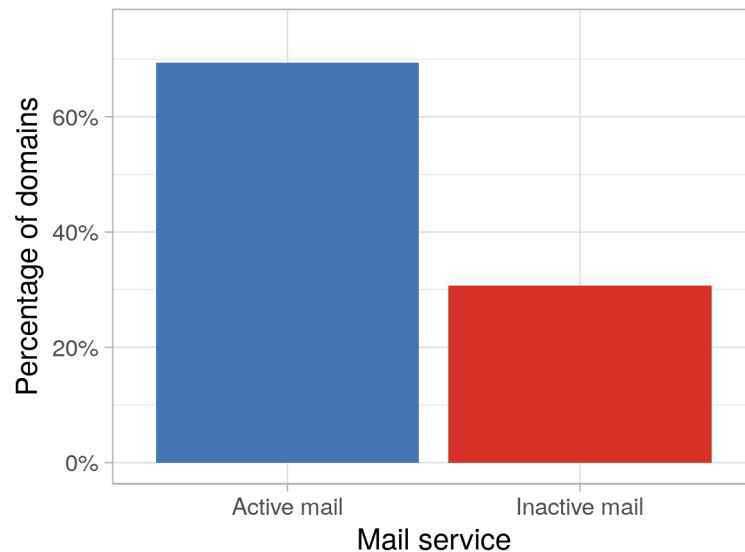


Detecting mail service

- **Active mail** - if there was a reachable mail server for a domain
- **Inactive mail** - if there was no mail server for a domain or it was unreachable
- For each domain:
 - DNS crawler connected to port **25**, **465** or **587** on its mail server
 - Mail server indicated in MX or A/AAAA records (see RFC5321)



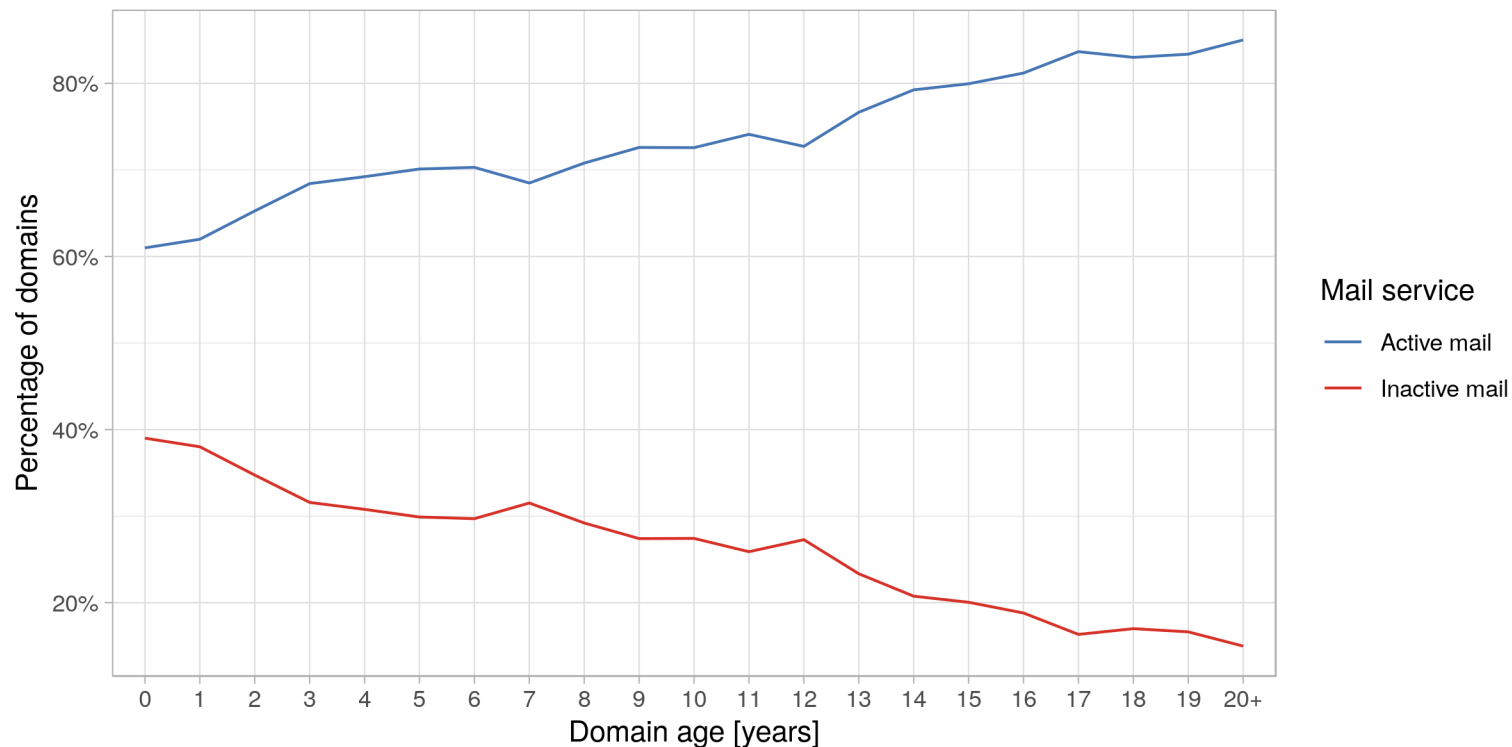
Results – mail service



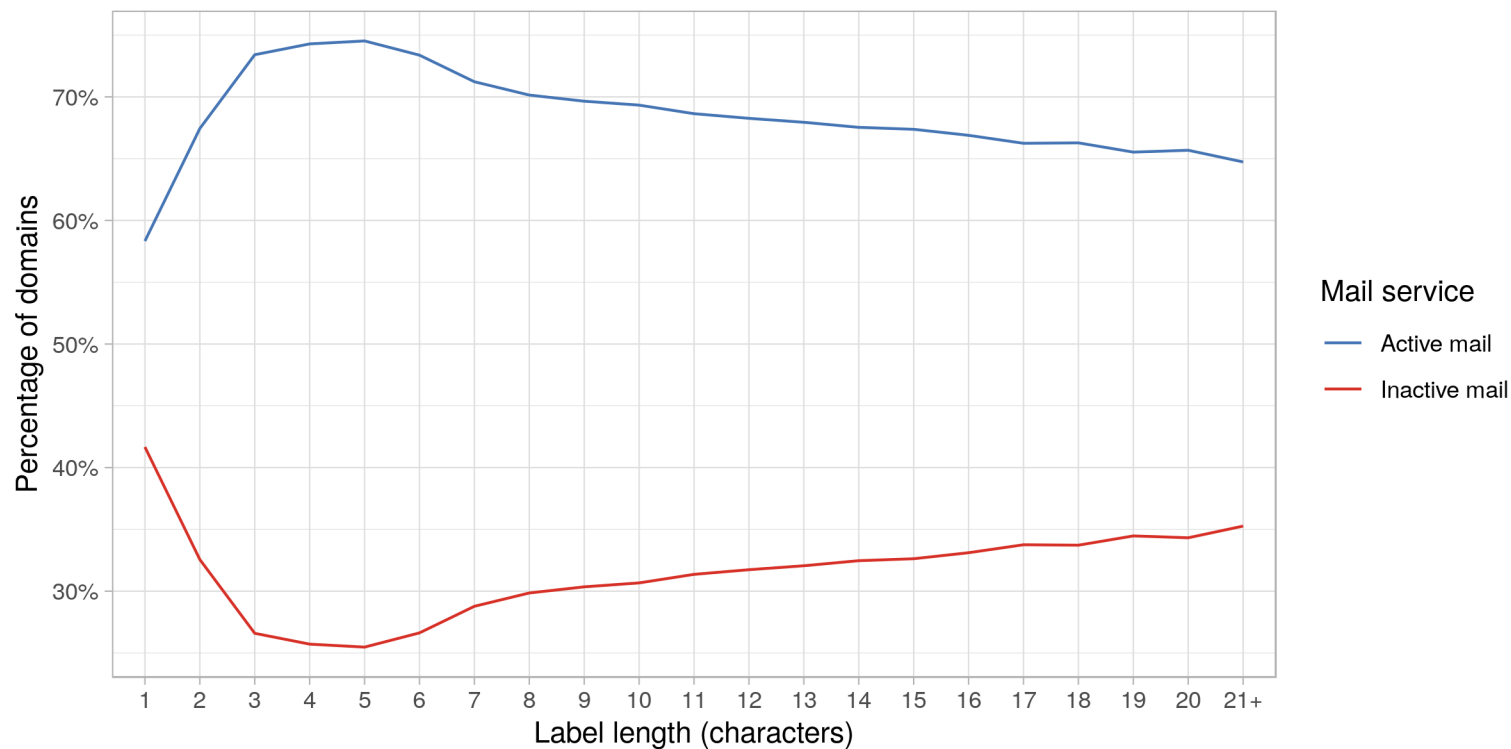
- **69.3%** domains have active mail server



Results – mail service



Results – mail service

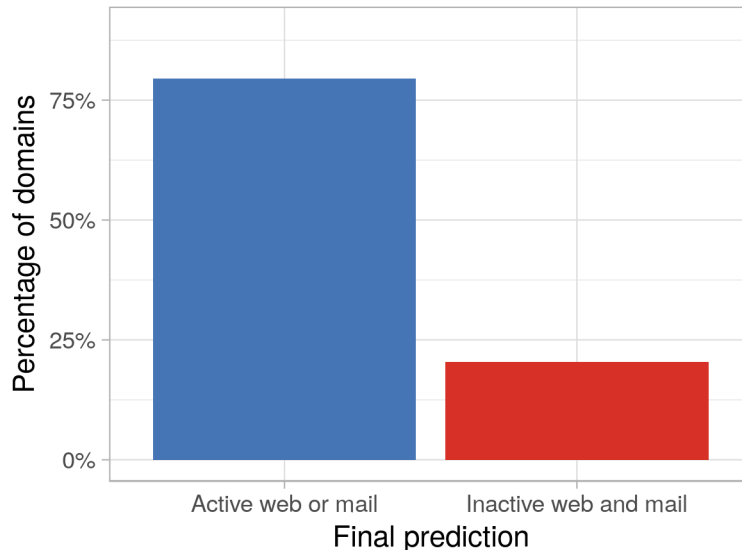


Final prediction

- **Active web or mail** – if a domain hosted a non-parking website or had a working mail server)
- **Inactive web and mail** - if there was neither a non-parking website nor an active mail server for this domain



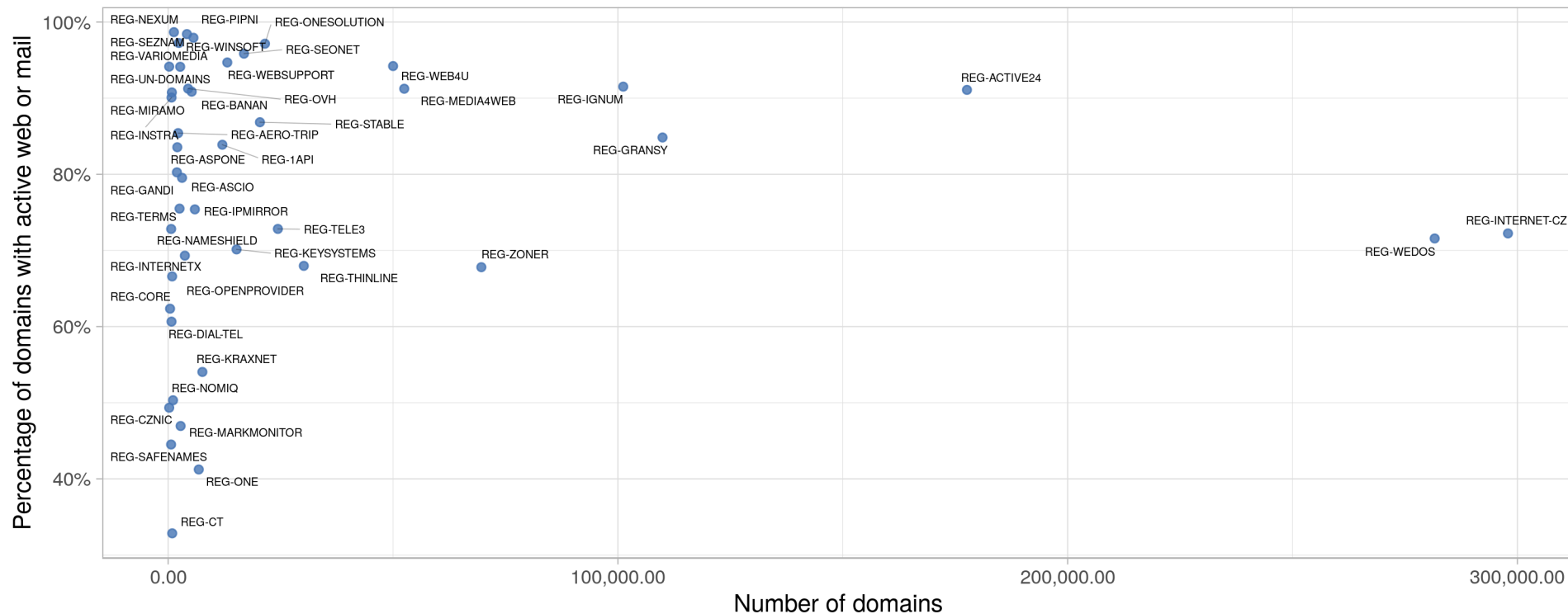
Results – final prediction



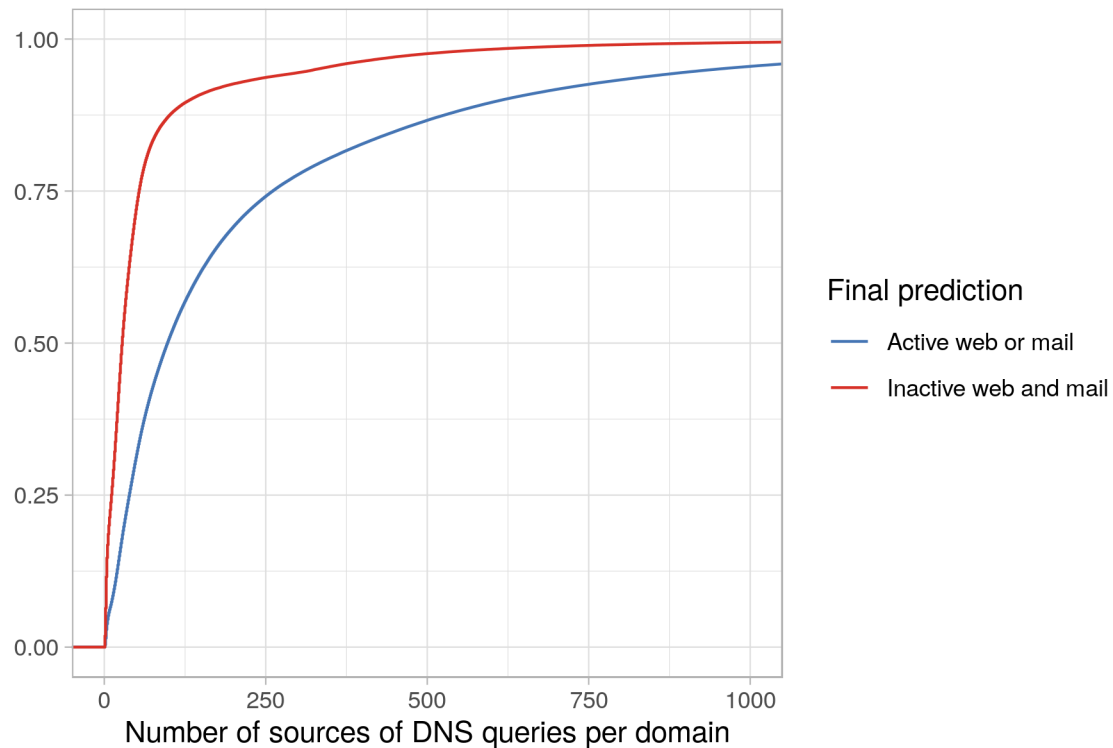
- **79.6%** domains have active web or mail server



Results – final prediction



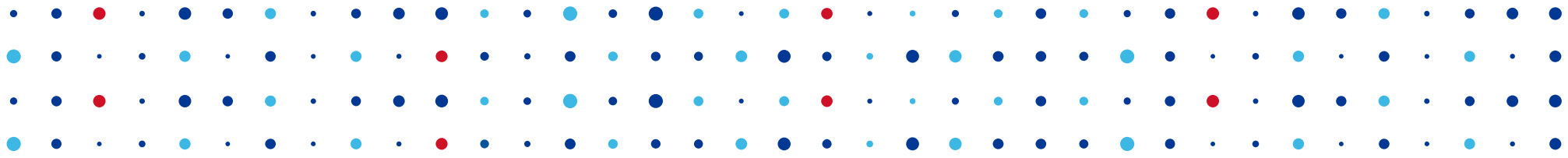
Results – final prediction



Conclusions

- Out of **1,365,753** .CZ domains:
 - ~**60%** host a non-parking website
 - ~**70%** run a mail server
 - ~**80%** host a non-parking website or run a mail server
- Domain age and label length matters
- Remark: a domain can be used for a different purpose than mail or www





Thank You

Maciej Andziński • maciej.andzinski@nic.cz

