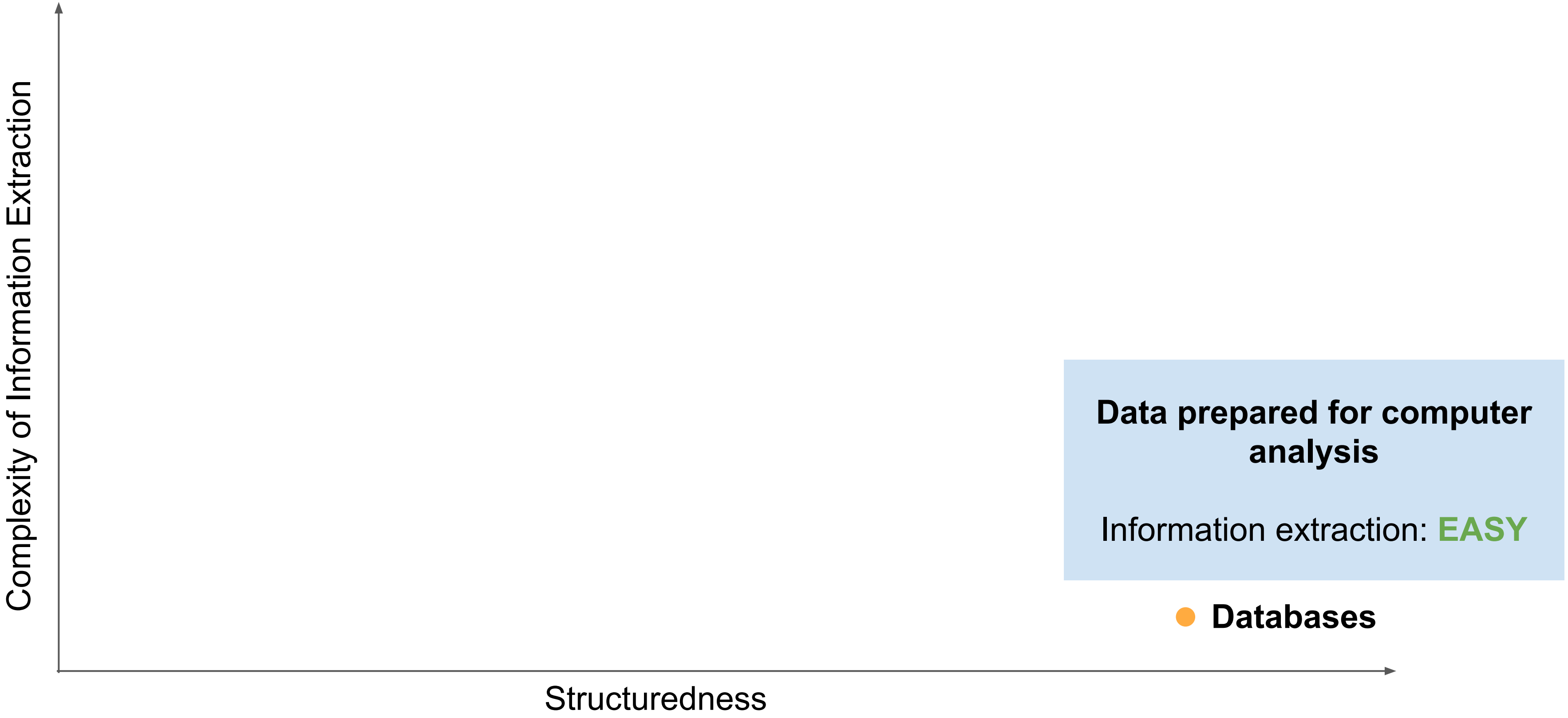
Deep Neural Networks for Information Extraction

Tomáš Gogár, Petr Baudiš

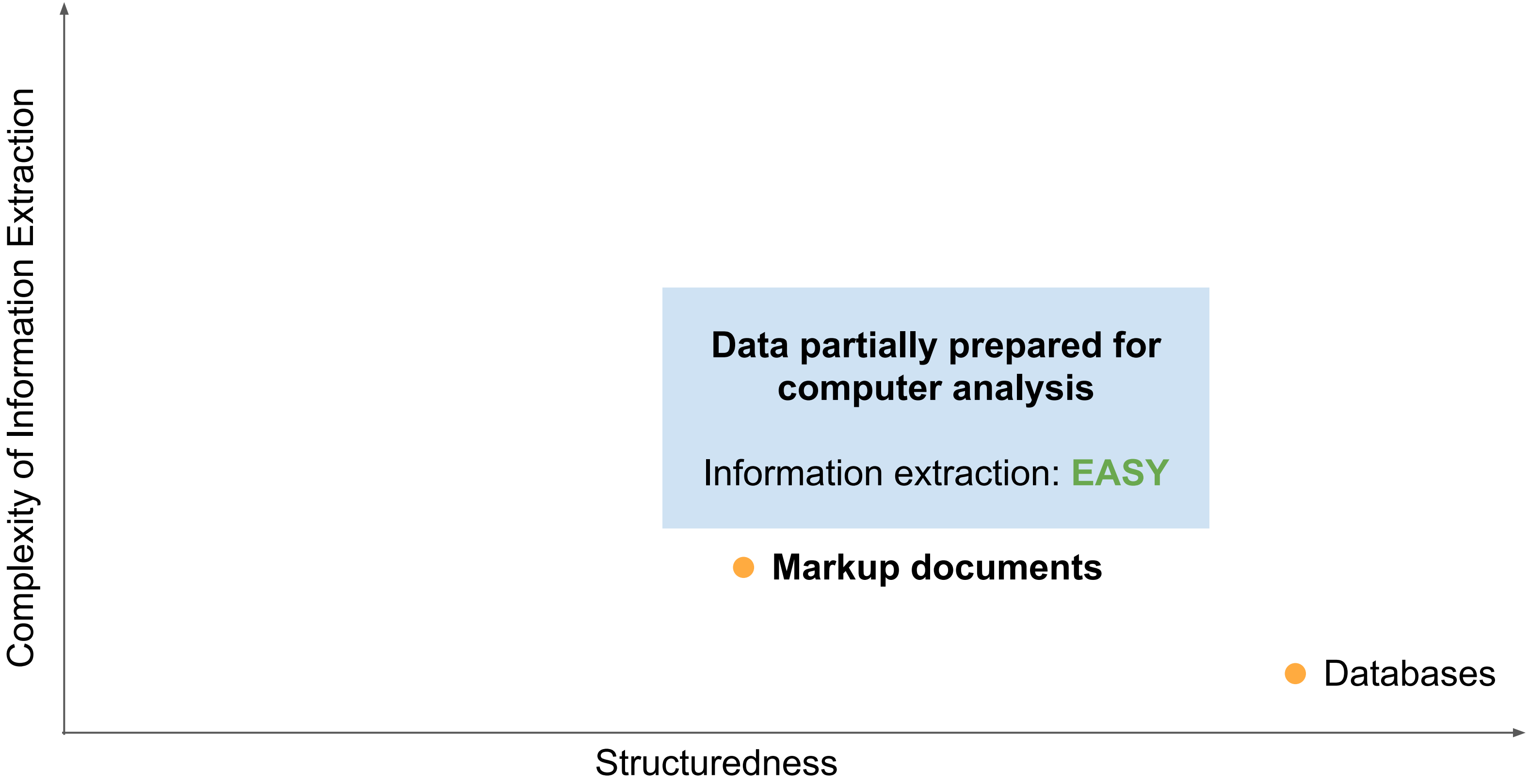
Information Extraction



Data prepared for computer analysis

Information extraction: **EASY**

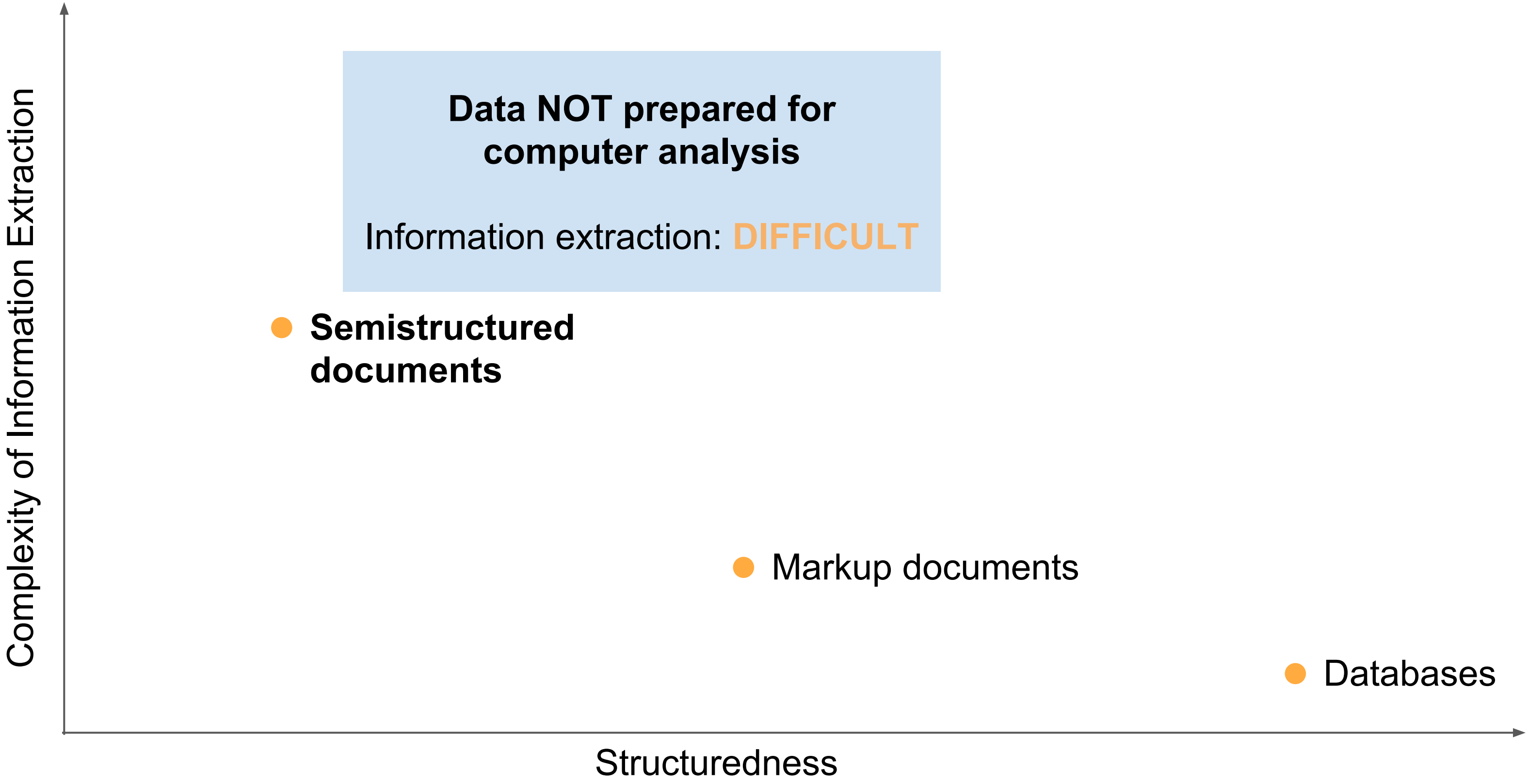
Information Extraction



Data partially prepared for computer analysis

Information extraction: **EASY**

Information Extraction



Data NOT prepared for computer analysis

Information extraction: **DIFFICULT**

● **Semistructured documents**

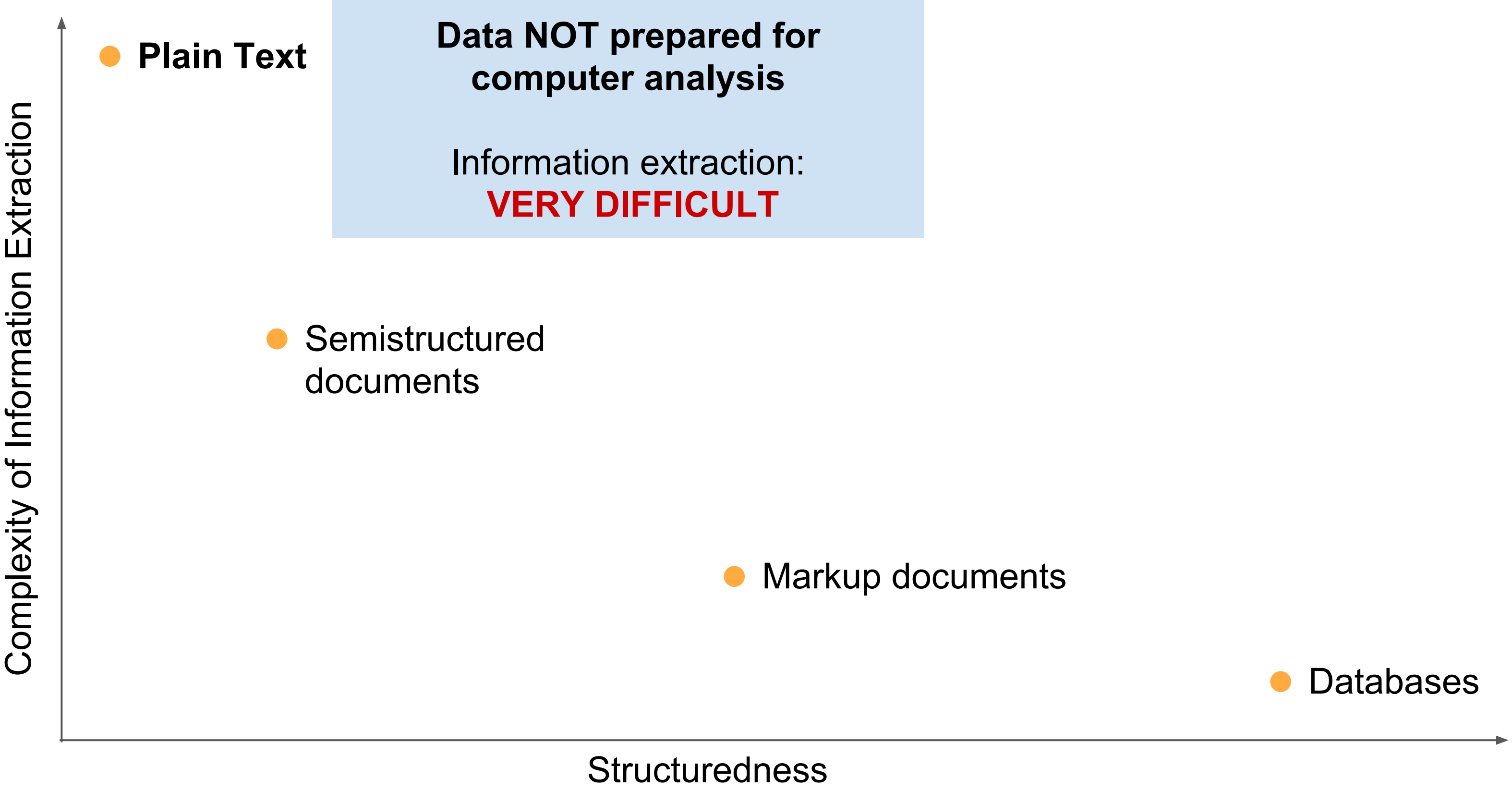
● Markup documents

● Databases

Complexity of Information Extraction

Structuredness

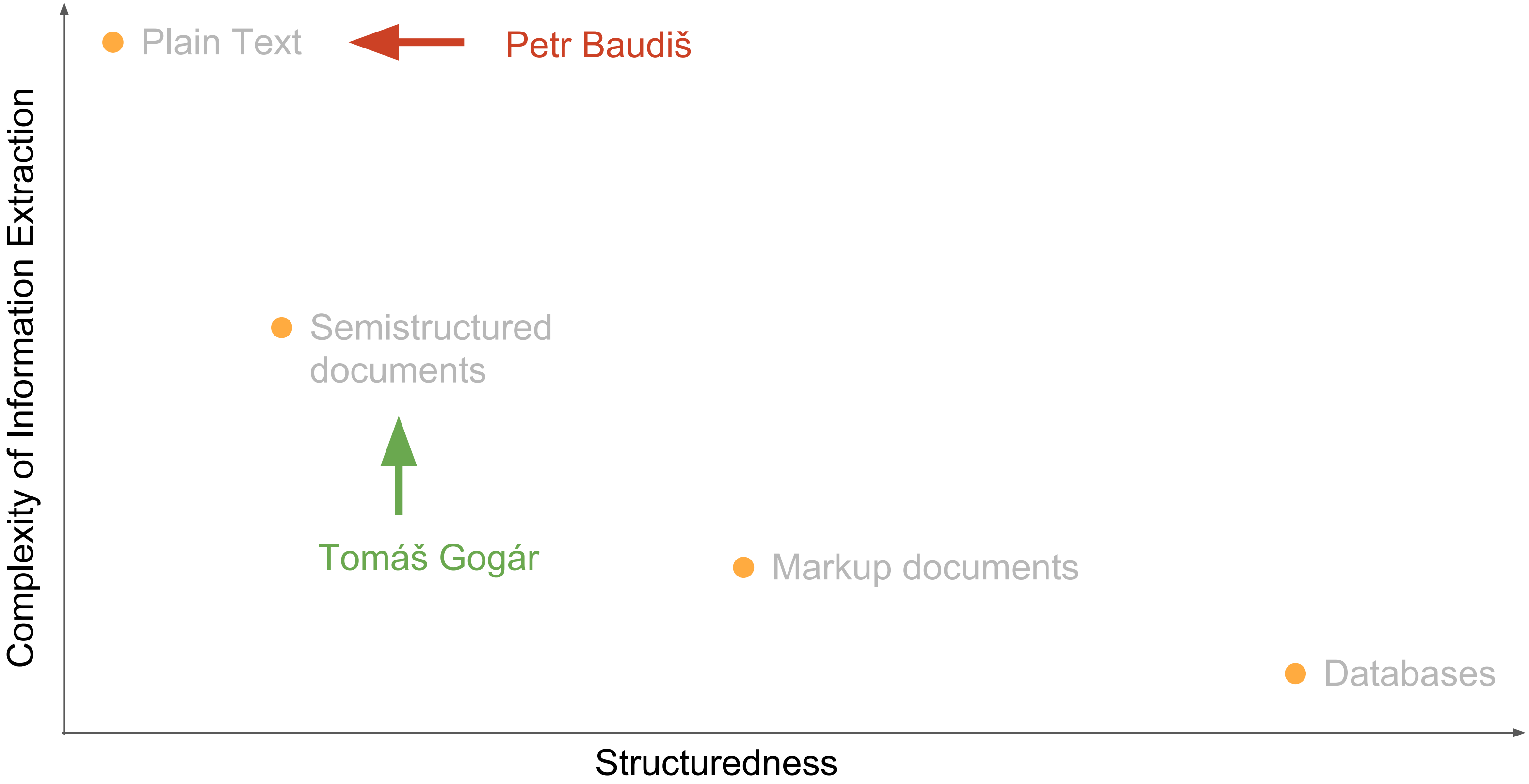
Information Extraction


● **Plain Text**
● Semistructured documents
● Markup documents
● Databases
Structuredness
Complexity of Information Extraction

Data NOT prepared for computer analysis

Information extraction:
VERY DIFFICULT

Information Extraction



Plain Text

Petr Baudiš

Semistructured documents

Tomáš Gogár

Markup documents

Databases

Complexity of Information Extraction

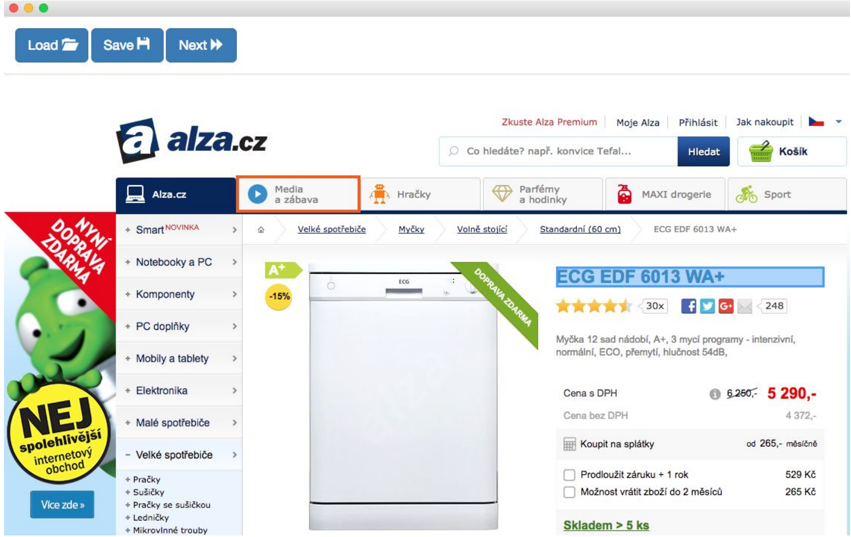
Structuredness

Web Information Extraction - Current systems

- Web pages are created from Templates
- **Learn template structure** ⇨ **Extract Information**

Web Information Extraction - Current systems

- Web pages are created from Templates
- **Learn template structure ⇨ Extract Information**
- **Template learning:**
  - **Manual annotation - Scraping**



ECG EDF 6013 WA+

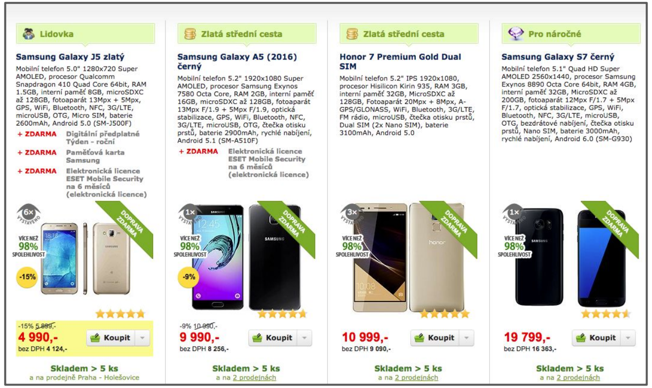
Myčka 12 sad nádobí, A+, 3 mycí programy - intenzivní, normální, ECO, přemytí, hlučnost 54dB,

Cena s DPH	~~6 250,-~~ 5 290,-
Cena bez DPH	4 372,-
Koupit na splátky	od 265,- měsíčně
□ Prodloužit záruku + 1 rok	529 Kč
□ Možnost vrátit zboží do 2 měsíců	265 Kč

Skladem > 5 ks

Web Information Extraction - Current systems

- Web pages are created from Templates
- **Learn template structure ⇨ Extract Information**
- **Template learning:**
  - **Manual annotation**
  - **Automatic learning - repeated patterns**



Lidovka

Samsung Galaxy J5 zlatý

Mobilní telefon 5.0" 1280x720 Super AMOLED, procesor Qualcomm Snapdragon 410 Quad Core 64bit, RAM 1.5GB, interní paměť 8GB, microSDXC až 128GB, fotoaparát 13Mpx + 5Mpx, GPS, WiFi, Bluetooth, NFC, 3G/LTE, microUSB, OTG, Micro SIM, baterie 2600mAh, Android 5.0 (SM-J500F)

+ ZDARMA Digitální předplatné Týden - roční

+ ZDARMA Paměťová karta Samsung

+ ZDARMA Elektronická licence ESET Mobile Security na 6 měsíců (elektronická licence)

-15% ~~5 899,-~~

4 990,-

bez DPH 4 124,-

Koupit

Skladem > 5 ks

a na prodejně Praha - Holešovice

Zlatá střední cesta

Samsung Galaxy A5 (2016) černý

Mobilní telefon 5.2" 1920x1080 Super AMOLED, procesor Samsung Exynos 7580 Octa Core, RAM 2GB, interní paměť 16GB, microSDXC až 128GB, fotoaparát 13Mpx F/1.9 + 5Mpx F/1.9, optická stabilizace, GPS, WiFi, Bluetooth, NFC, 3G/LTE, microUSB, OTG, čtečka otisku prstů, baterie 2900mAh, rychlé nabíjení, Android 5.1 (SM-A510F)

+ ZDARMA Elektronická licence ESET Mobile Security na 6 měsíců (elektronická licence)

-9% ~~10 990,-~~

9 990,-

bez DPH 8 256,-

Koupit

Skladem > 5 ks

a na 2 prodejnách

Zlatá střední cesta

Honor 7 Premium Gold Dual SIM

Mobilní telefon 5.2" IPS 1920x1080, procesor Hisilicon Kirin 935, RAM 3GB, interní paměť 32GB, MicroSDXC až 128GB, Fotoaparát 20Mpx + 8Mpx, A-GPS/GLONASS, WiFi, Bluetooth, 3G/LTE, FM rádio, microUSB, čtečka otisku prstů, Dual SIM (2x Nano SIM), baterie 3100mAh, Android 5.0

10 999,-

bez DPH 9 090,-

Koupit

Skladem > 5 ks

a na 2 prodejnách

Pro náročné

Samsung Galaxy S7 černý

Mobilní telefon 5.1" Quad HD Super AMOLED 2560x1440, procesor Samsung Exynos 8890 Octa Core 64bit, RAM 4GB, interní paměť 32GB, MicroSDXC až 200GB, fotoaparát 12Mpx F/1.7 + 5Mpx F/1.7, optická stabilizace, GPS, WiFi, Bluetooth, NFC, 3G/LTE, microUSB, OTG, bezdrátové nabíjení, čtečka otisku prstů, Nano SIM, baterie 3000mAh, rychlé nabíjení, Android 6.0 (SM-G930)

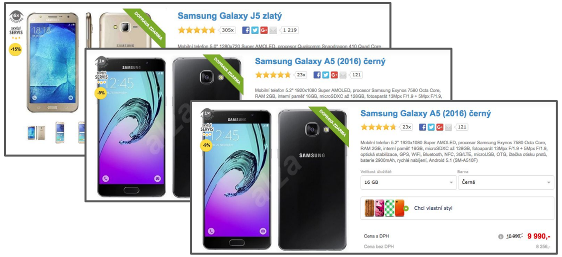
19 799,-

bez DPH 16 363,-

Koupit

Skladem > 5 ks

a na 2 prodejnách



Samsung Galaxy J5 zlatý

305x 1 219

Mobilní telefon 5.0" 1280x720 Super AMOLED, procesor Qualcomm Snapdragon 410 Quad Core

Samsung Galaxy A5 (2016) černý

23x 121

Mobilní telefon 5.2" 1920x1080 Super AMOLED, procesor Samsung Exynos 7580 Octa Core, RAM 2GB, interní paměť 16GB, microSDXC až 128GB, fotoaparát 13Mpx F/1.9 + 5Mpx F/1.9,

Samsung Galaxy A5 (2016) černý

23x 121

Mobilní telefon 5.2" 1920x1080 Super AMOLED, procesor Samsung Exynos 7580 Octa Core, RAM 2GB, interní paměť 16GB, microSDXC až 128GB, fotoaparát 13Mpx F/1.9 + 5Mpx F/1.9, optická stabilizace, GPS, WiFi, Bluetooth, NFC, 3G/LTE, microUSB, OTG, čtečka otisku prstů, baterie 2900mAh, rychlé nabíjení, Android 5.1 (SM-A510F)

Velikost úložiště: 16 GB

Barva: Černá

Chci vlastní styl

Cena s DPH ~~10 990,-~~ 9 990,-

Cena bez DPH 8 256,-

In a page

Across website

Web Information Extraction - Current systems

- Web pages are created from Templates
- **Learn template structure ⇨ Extract Information**
- **Template learning:**
  - **Manual annotation**
  - **Automatic learning - repeated patterns**

It’s just a hack!

What matters in Information Extraction

What is written?

Where it is written?

How it is written?

What matters in Information Extraction

What is written?

Where it is written?

How it is written? **[Screenshot]**

What matters in Information Extraction

What is written?

Where it is written?

[SPATIAL BAG-OF-WORDS]

How it is written? **[Screenshot]**

Intro: Bag-of-Words

Text representation often used in NLP:

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

Vocabulary:

"John"	1
"likes"	2
"to"	3
"watch"	4
"movies"	5
"also"	6
"football"	7
"games"	8
"Mary"	9
"too"	10

Vectors:

Document1 = [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]

Document2 = [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

Intro: Hashing Trick

Text representation often used in NLP:

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

~~**Vocabulary:**~~

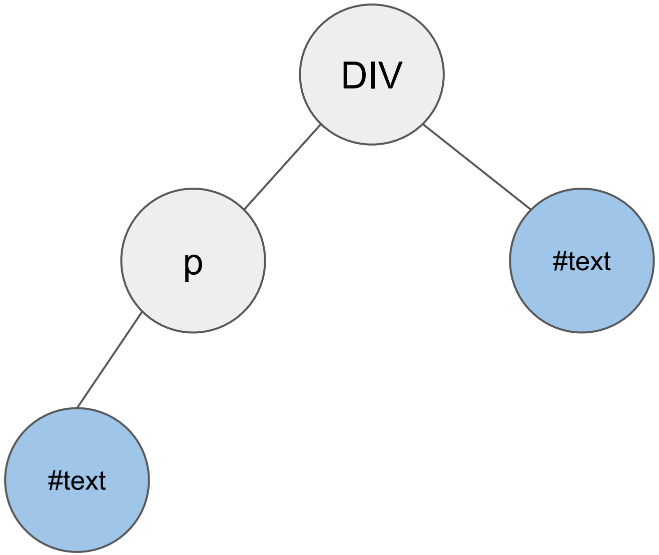
Hashing function

h(john**) = 3**
h(likes**) = 1**
…

- Does not need vocabulary :-)
- Arbitrary size of result vector :-)
- Collisions :-(

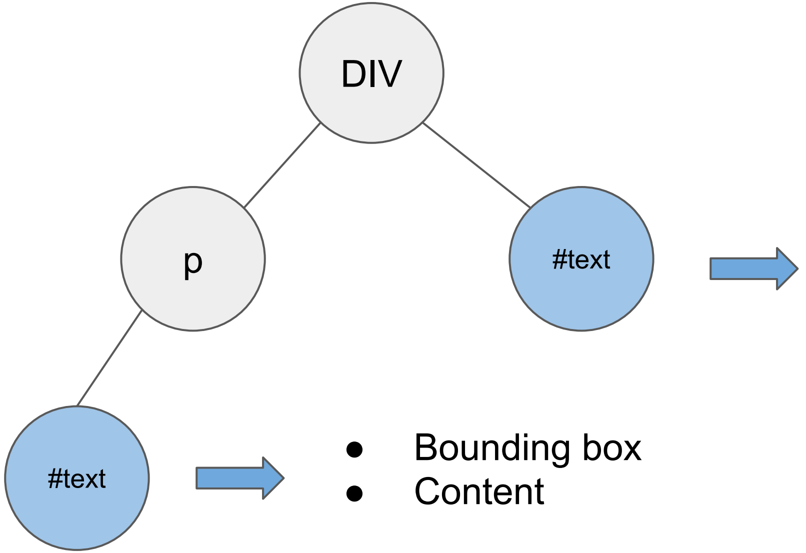
Spatial Bag-of-Words

- We do not process text as a whole
- We process each TEXT NODE individually



Spatial Bag-of-Words

- We do not process text as a whole
- We process each TEXT NODE individually

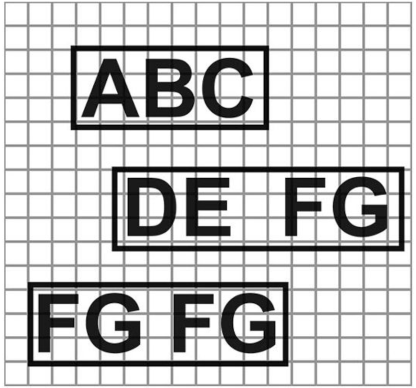


- Bounding box
- Content

- Bounding box
- Content

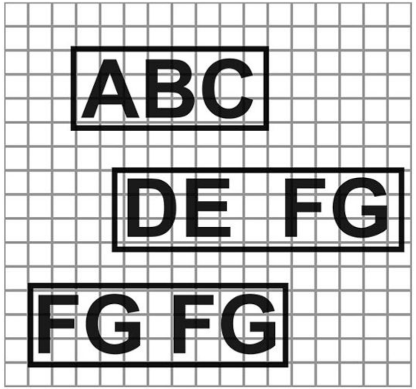
Spatial Bag-of-Words

- We do not process text as a whole
- We process each TEXT NODE individually



Spatial Bag-of-Words

- We do not process text as a whole
- We process each TEXT NODE individually



```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0
0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

h(ABC)

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

h(DE)

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 2 2 2 2 2 2 2 2 2 2 2 0 0 0 0 0
0 2 2 2 2 2 2 2 2 2 2 2 0 0 0 0 0
0 2 2 2 2 2 2 2 2 2 2 2 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

h(FG)

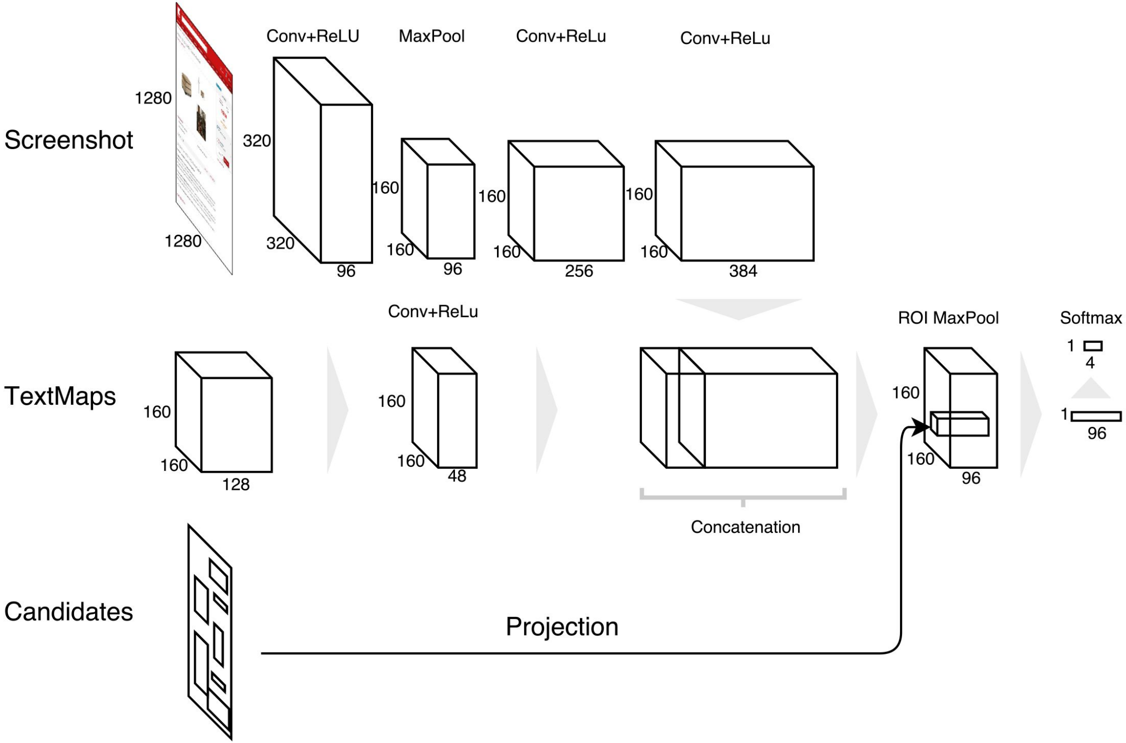
Spatial Bag-of-Words

- We do not process text as a whole
- We process each TEXT NODE individually

TEXT ENCODED IN TENSOR

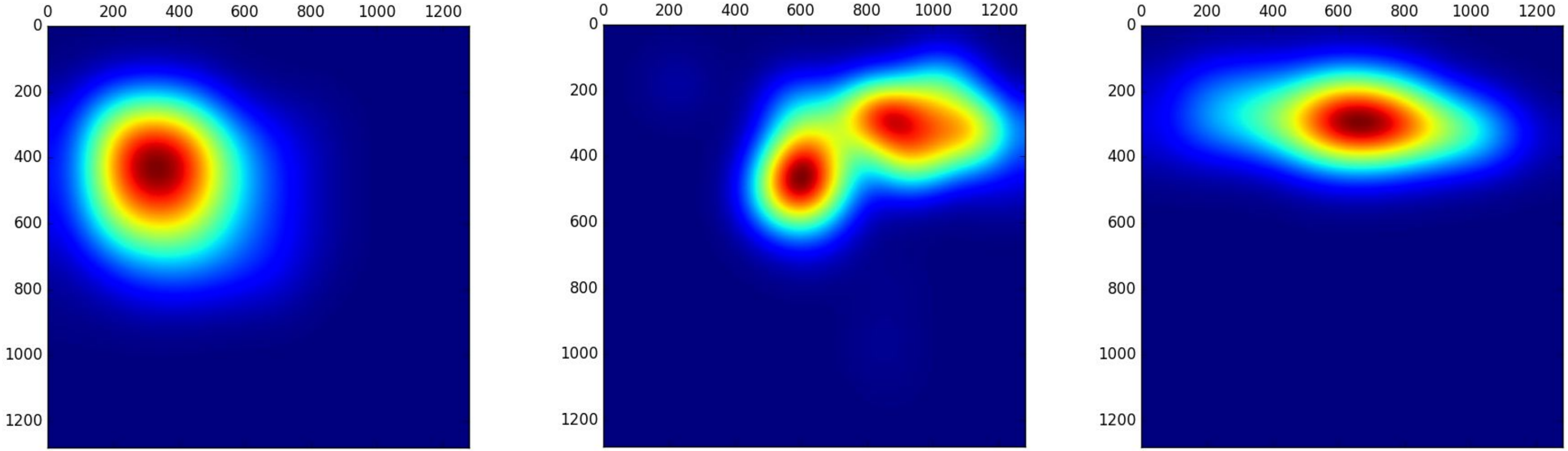
(SAME AS IMAGE)

Net architecture

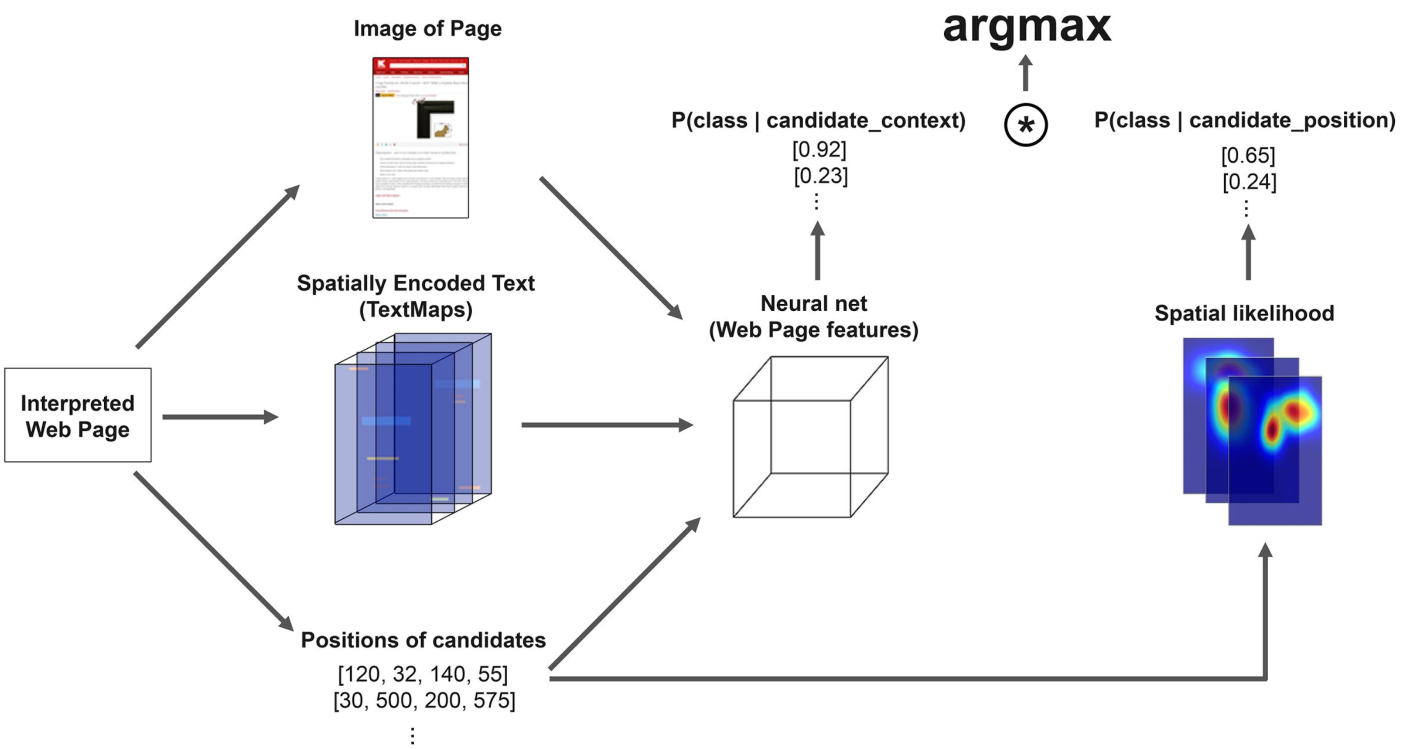


Problem: Uses only local information!

Spatial likelihood



Final system



Interpreted Web Page

Image of Page

Spatially Encoded Text (TextMaps)

Positions of candidates

[120, 32, 140, 55]
[30, 500, 200, 575]
⋮

Neural net (Web Page features)

P(class | candidate_context)

[0.92]
[0.23]
⋮

⊛

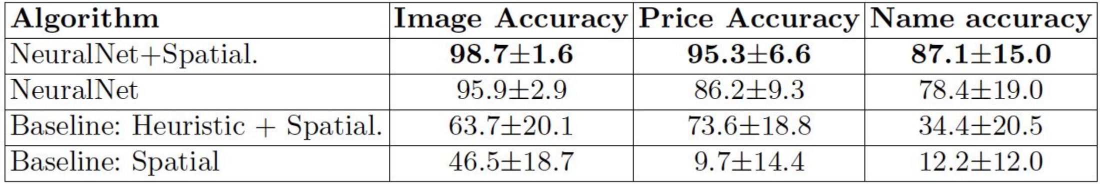
argmax

Spatial likelihood

P(class | candidate_position)

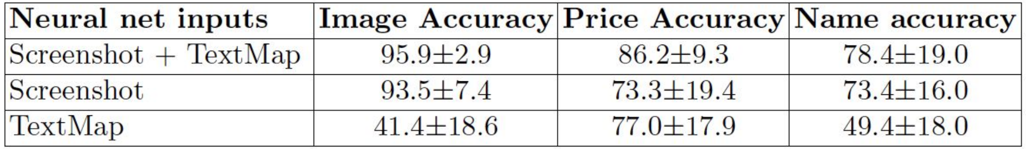
[0.65]
[0.24]
⋮

Results



Algorithm	Image Accuracy	Price Accuracy	Name accuracy
NeuralNet+Spatial.	**98.7±1.6**	**95.3±6.6**	**87.1±15.0**
NeuralNet	95.9±2.9	86.2±9.3	78.4±19.0
Baseline: Heuristic + Spatial.	63.7±20.1	73.6±18.8	34.4±20.5
Baseline: Spatial	46.5±18.7	9.7±14.4	12.2±12.0

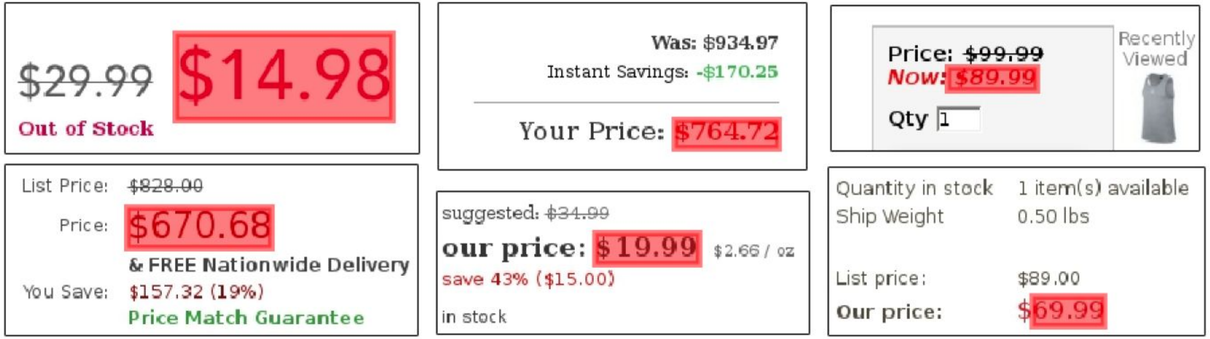
Table 3. Comparison of algorithms: mean and standard deviation of accuracy across 10 splits (in percents).



Neural net inputs	Image Accuracy	Price Accuracy	Name accuracy
Screenshot + TextMap	95.9±2.9	86.2±9.3	78.4±19.0
Screenshot	93.5±7.4	73.3±19.4	73.4±16.0
TextMap	41.4±18.6	77.0±17.9	49.4±18.0

Table 4. Neural Net with different input data: mean and standard deviation of accuracy across 10 splits (in percents).

Results



~~$29.99~~ $14.98

Out of Stock

Was: $934.97

Instant Savings: -$170.25

Your Price: $764.72

Price: ~~$99.99~~

Now: $89.99

Qty 1

Recently Viewed

List Price: ~~$828.00~~

Price: $670.68

& FREE Nationwide Delivery

You Save: $157.32 (19%)

Price Match Guarantee

suggested: ~~$34.99~~

our price: $19.99 $2.66 / oz

save 43% ($15.00)

in stock

Quantity in stock 1 item(s) available

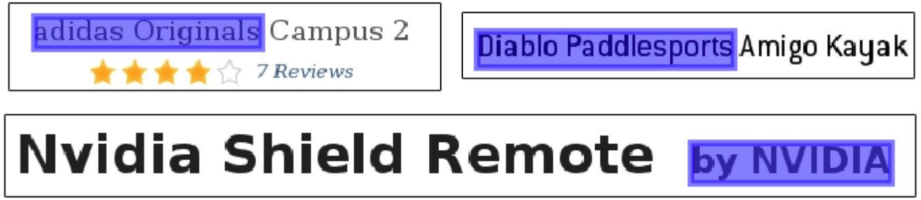
Ship Weight 0.50 lbs

List price: $89.00

Our price: $69.99

Fig. 5. Examples of *current price* detection.

Results



adidas Originals Campus 2

7 Reviews

Diablo Paddlesports Amigo Kayak

Nvidia Shield Remote by NVIDIA

Fig. 6. Examples of product names divided into two parts (manufacturer + model).

Future work

Machine learning:

- Solve global position problem: Attention network?
- Try to learn text features
- Try other similar tasks: ex. classification

Practical problems:

- Popup windows
- Information distributed in multiple DOM-Elements

Source code:

github.com/gogartom/TextMaps