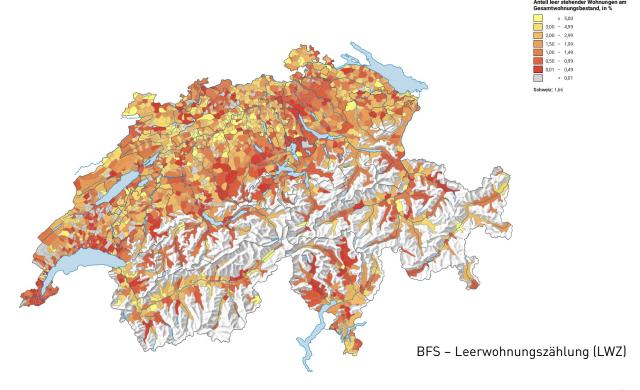
datahouse A Crawling Application with R

## What about Real Estate?



#### Real Estate Development

- Know the supply and demand
- Increasing investment volumes
- Single-point vs. development-over-time
- Required data:
  - Supply of real estate
  - Demand / absorption
  - Influencing factors (price, location ...)

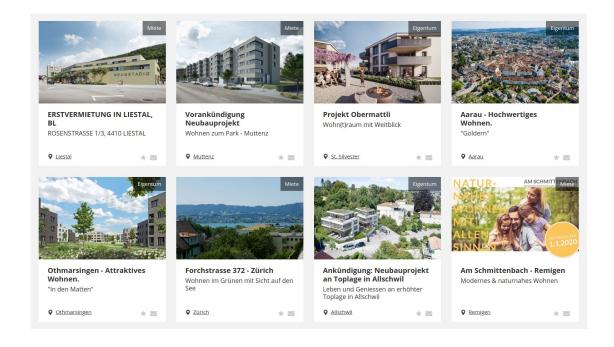




#### **New Real Estate Projects**



- Currently 2111 listings (September 2019)
- From single housings to big projects
- Existing workflow at Wüest Partner



# A Crawling App with R

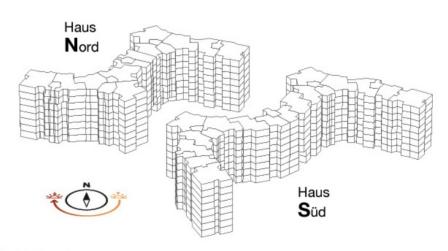
LAGE

WOHNUNGEN

GALERIE

KONTAKT **D** E

Haus: Nord Süd Zimmer: 1-1.5 2.5 3.5 4.5 5.5 Nur verfügbare Wohnungen



Frei = Jetzt bewerben

Reserviert = Mietvertrag ist unterwegs

**Vermietet** = Wohnung ist vergeben

Tiefgaragenparkplätze und Elektroparkplätze à CHF 180.00 können mit den 4.5 und 5.5 Zimmerwohnungen dazu gemietet werden.

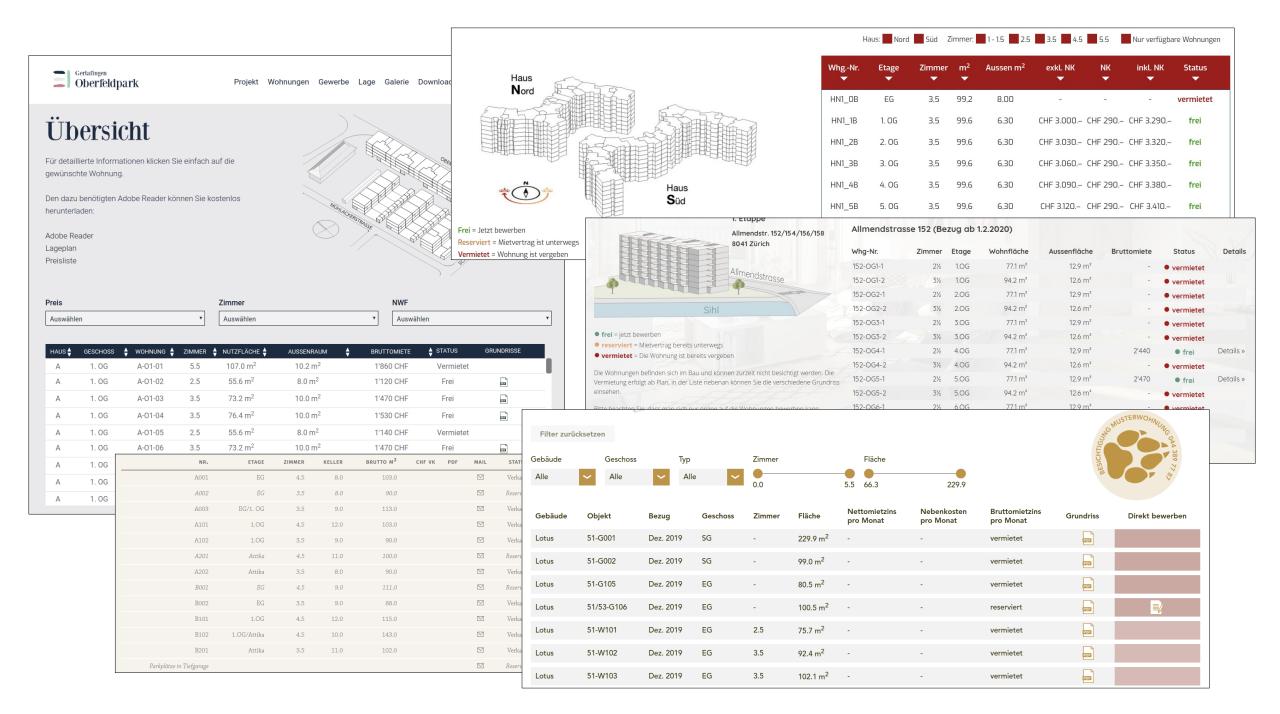
SCHRITTE ZUR ANMELDUNG

MIETZINSLISTE

KURZBAUBESCHRIEB

UMGEBUNGSPLAN

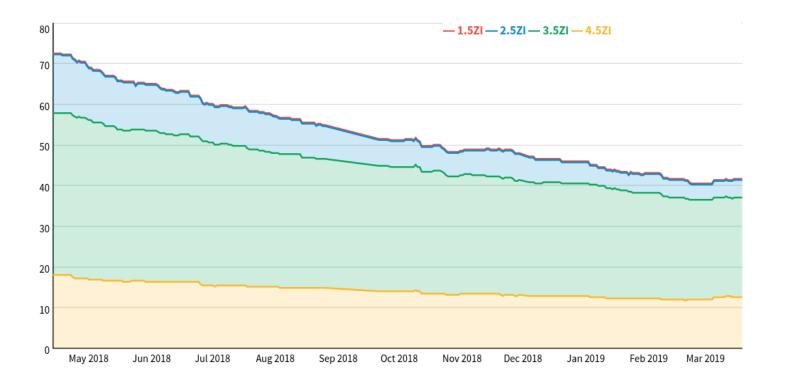
WhgNr. ▼	Etage ▼	Zimmer •	m² ▼	Aussen m <sup>2</sup>	exkl. NK ▼	NK ▼	inkl. NK ▼	Status ▼
HN1_OB	EG	3,5	99,2	8,00	-	-	-	vermietet
HN1_1B	1. OG	3,5	99,6	6,30	CHF 3.000	CHF 290	CHF 3.290	frei
HN1_2B	2. OG	3,5	99,6	6,30	CHF 3.030	CHF 290	CHF 3.320	frei
HN1_3B	3. OG	3,5	99,6	6,30	CHF 3.060	CHF 290	CHF 3.350	frei
HN1_4B	4. OG	3,5	99,6	6,30	CHF 3.090	CHF 290	CHF 3.380	frei
HN1_5B	5. OG	3,5	99,6	6,30	CHF 3.120	CHF 290	CHF 3.410	frei
HN1_6B	6. OG	3,5	99,6	6,30	CHF 3.150	CHF 290	CHF 3.440	frei
HN1_7B	7. OG	3,5	99,6	6,30	-	-	-	reserviert
HN1_1A	1. OG	4,5	119,5	8.1 + 3.3	CHF 3.880	CHF 350	CHF 4.230	frei
HN1_2A	2. OG	4,5	119,1	8.1 + 3.3	CHF 3.910	CHF 350	CHF 4.260	frei
HN1_3A	3. OG	4,5	119,5	8.1 + 3.3	CHF 3.940	CHF 350	CHF 4.290	frei
HN1_4A	4. OG	4,5	119,5	8.1 + 3.3	CHF 3.970	CHF 350	CHF 4.320	frei



## Requirement Measure the Absorption Rate of Real Estate



#### **Example: Absorption Rate of Flats**

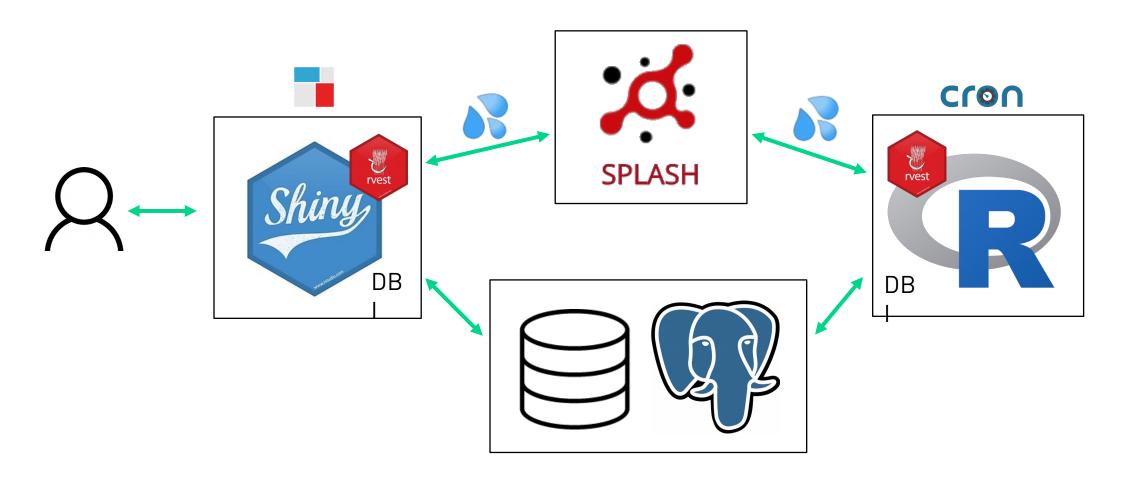


## App Demo

### Architecture Overview

#### Architecture





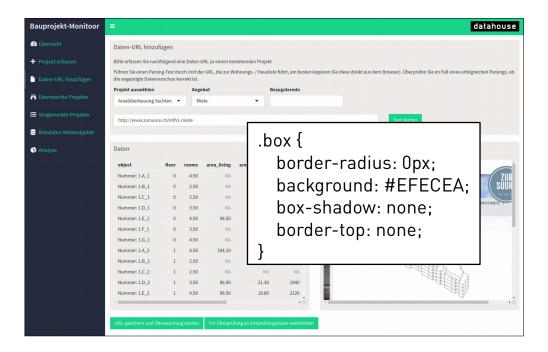
# Architecture Details





### **Shiny Dashboard**

- Shiny Proxy: deployment and user management
- Shiny Dashboard with custom CSS
- From custom Shiny-Proxy-Verse Image
- DataTable, Leaflet

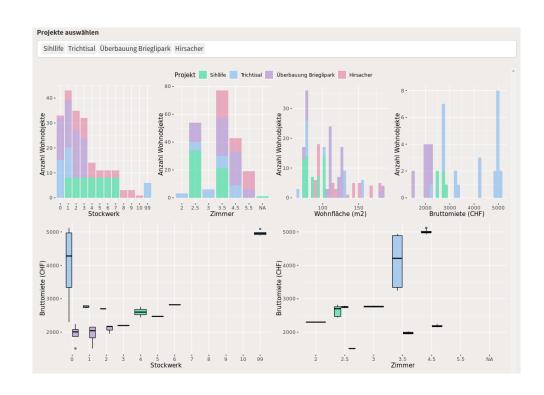




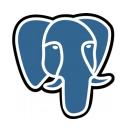


#### **Shiny Dashboard**

- Enter new projects
- Register URLs with data
- Test automatic extraction and parsing
- Survey projects statuses
- Analyze data







#### **Postgres Database**

- Tables: projects, urls, dwellings
- Connection from R with DBI
- Load / append data frames

```
db <- dbPool(
  odbc(),
  Driver = "PostgreSQL Unicode",
  Database = Sys.getenv("POSTGRES_DB"),
  UserName = Sys.getenv("POSTGRES_USER"),
  Password = Sys.getenv("POSTGRES_PASSWORD"),
  Servername = Sys.getenv("POSTGRES_SERVER"),
  Port = Sys.getenv("POSTGRES_PORT")
}

urls_to_crawl <- dbReadTable(db, "urls") %>%
  filter(url_parse_status == 1)

dbWriteTable(db, "dwellings", parsed_data, append = TRUE)
```





### Splash









#### Splash

- Headless browser
- Renders URL (data often loaded via JS)
- Out-of-the-box splash image
- Accessed from R with splashr



Returns full HTML / Screenshot

```
library(splashr)

my_splash <- splash(host="splash", port=8050)

html <- render_html(
    splash_obj = my_splash,
    url = input$url
)

screenshot <- render_png(
    splash_obj = my_splash,
    url = input$url,
}</pre>
```





### **Auto Extraction / Parsing Logic**

- Parse HTML with rvest
- Extract tables to data frames
  - html table()
  - html\_nodes() %>% html\_text()
  - Difficult to automate!
- Parse data frames with dplyr
  - Search Columns (key words, numbers)
  - Mutate
  - Validate

```
extract rooms <- possibly(function(df) {
 rooms <- tryCatch({</pre>
    df %>%
      select(matches(or("zimmer", "zi\\."))) %>%
     pull(1)
 }, error = function(e) {
   df %>%
     select_if(~all(!is.na(.))) %>%
      select_if(~any(str_detect(tolower(as.character(.)), "zi"))) %>%
     pull(1)
 })
  rooms %>%
   str_replace(",", ".") %>%
   str_replace("½", ".5") %>%
   str replace("Studio", "1") %>%
   str_remove_all(" ") %>%
   str_extract("\\d(\\.5)?") %>% # 3 or 3.5
    as.numeric()
}, NA_real_)
```





#### **Crawling Daemon**

- Separate R container
- Scheduled script with cron job
- Get URLs from db, get HTML, parse data
- One request per day  $\rightarrow$  small server load

```
db <- dbPool(odbc(), Driver = , Database = ..., ...)
my_splash <- splash(host="splash", port=8050)

urls_to_crawl <- dbReadTable(db, "urls") %>%
  filter(url_parse_status == 1)

html <- render_html(
    splash_obj = my_splash,
    url = urls_to_crawl
)

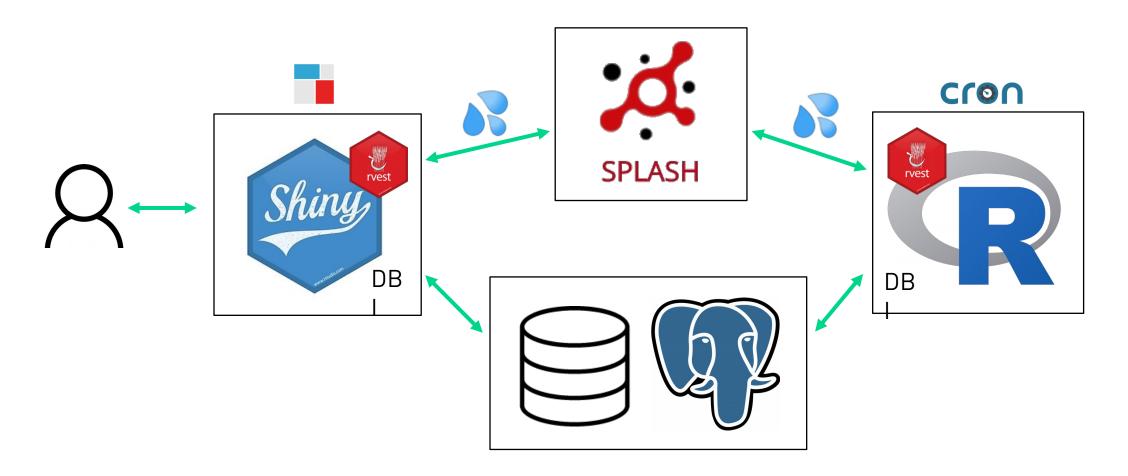
parsed_data <- return_table(html, urls_to_crawl)

dbWriteTable(db, "dwellings", parsed_data, append = TRUE)</pre>
```



#### Architecture





### Current State

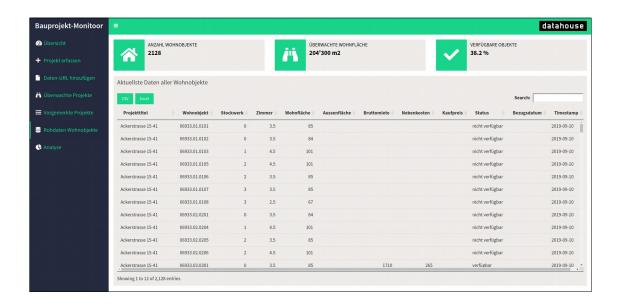


#### **Current State**

- Included in Wüest Partner weekly workflow
- Significant portion of large projects tracked
- Daily data tracking

#### **Next steps:**

- Model absorption rates
- Productizing analysis dashboard
- Automatic URL finder



# Challenges Help me:-)



#### **Auto Extract Data Frames from HTML**

```
▼
 <caption>Haus A (Bezug ab 1. September 2019)</caption>
▼<thead>
 ▶ ...
 </thead>
▼ 
 ▼
 class="rooms4 odd">
   A1.01
   1. 0G
   4
   84.5 m²
   10 m²
   vermietet
 ▶<tr data-id="W A1.02" data-rooms="3" data-status="vermietet"
 class="rooms3 even">...
 ▶<tr data-id="W A1.03" data-rooms="4" data-status="vermietet"
 class="rooms4 odd">...
 >
 class="rooms2 even">...
```

vs. <div/>

Table geometry

```
▶ <div class="row hover status3" data-over="http://
www.zursuure.ch/mm/1D 1.jpg" data-out="http://www.zursuure.ch/mm/
MFH 1.jpg">...</div>
▼<div class="row hover status1" data-over="http://
www.zursuure.ch/mm/1E 1.jpg" data-out="http://www.zursuure.ch/mm/
MFH 1.jpg">
  ▶ <div class="col spalte8 left">...</div>
  ▶ <div class="col spalte1 left">...</div>
  ▶ <div class="col spalte3 left">...</div>
  ▼<div class="col spalte4 left">
     <span class="hidden">Zimmer:</span>
     "4 %"
   </div>
  ▼<div class="col spalte5 left">
     <span class="hidden">Wohnfl.:</span>
     "98.5 m2"
   </div>
  ▼<div class="col spalte6 left">
     <span class="hidden">Aussenfl.:</span>
     "15.0 m2"
   </div>
  ▼<div class="col spalte9 left">
     <span class="hidden">Keller:</span>
     "7.2 m2"
   </div>
  ▼<div class="col spalte7 left">
     <span class="hidden">Miete:</span>
     "CHF 1'890. -"
   </div>
```

# Web Crawling in General



### **Large Crawling Projects**

- Performance, automation
- Easy form submission
- Extended session management capabilities
- Quick deployment, testing

More software development Less data analysis



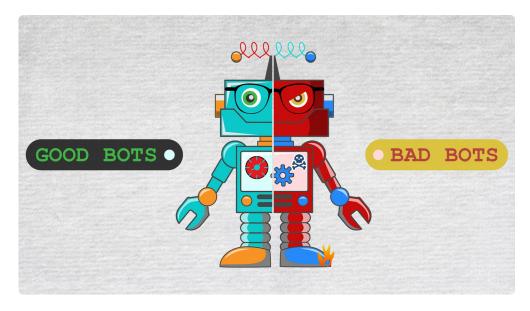
An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.





#### Things to consider

- Only collect as much as you really need
- Do not overload sites
- Respect data privacy
- Consider social implications

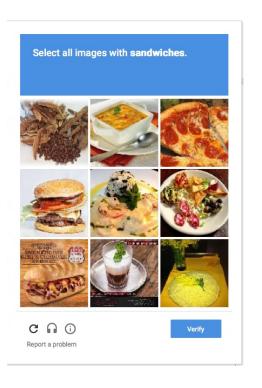


https://www.shieldsquare.com/good-bots-and-bad-bots/



#### How not to get crawled

- Provide an API
- Use a CAPTCHA (medium protection)
- Limiting requests via cookies (low protection)
- Limiting requests via IP (low protection)
- Think like a crawler (medium to high protection)



#### datahouse

September 16<sup>th</sup>, 2019

#### Thank you for your attention!

**Thomas Maier** 

+41 44 289 92 63

thomas.maier@datahouse.ch

Datahouse AG

Bleicherweg 5 8001 Zürich

www.datahouse.ch

**Daniel Meister** 

+41 44 289 92 30

daniel.meister@datahouse.ch

**Datahouse AG** 

Bleicherweg 5 8001 Zürich

www.datahouse.ch