

# Web Scraping and APIs

DSIER [/di'zi:ɪər/] — Summer 2023

Irene Iodice

Bielefeld University

## What skills are demanded in Germany ?

You can use **conventional database**, for example:

- Employer-employee database (access is limited)
- ONET database for skills content
- maps German occupations classes (KdB) to ONET classes

And look at the skill-content of newly created jobs

# Job offers online

You can create your own data starting from **job posting online**

The screenshot shows a web browser displaying a job listing on Indeed.com. The page is split into two main columns. The left column contains a list of job offers, and the right column shows the details for the selected 'Global Access Risk Analyst' position at Meta.

**Job Listing Summary (Left Column):**

- Global Access Risk Analyst**  
Meta 4.1 ★  
Home Office  
- This position will be responsible for understanding and supporting the design of Facebook's organizational, procedural and technological security controls within the context of security risk to our...  
vor 2 Tagen · Für Analyst
- Data Analyst (m/f/diverse) - Continental Business Consulting**  
Continental 4.8 ★  
Regensburg  
- Driving challenging data science projects in a cross-functional business environment and developing pragmatic solutions to solve critical business problems.  
vor 27 Tagen · Für Analyst
- Financial Advisor / Analyst (f/m/d)**  
EY 4.8 ★  
Eschborn  
- Analyses of project and financing structures of large projects.  
- Preparation of solutions concepts, presentations, expert opinions, and reports.  
- Coaching and reviewing of junior team members with...  
Heute · Für Analyst
- Machine learning engineer (Remote: Intern)**  
Konfido GmbH  
Home Office  
- JDZ direkt bewerben  
- You will help the team to maintain and improve the current machine learning models and create new solutions based on the company's needs.  
- Building machine learning pipelines and models for pricing...

**Job Details (Right Column):**

**Global Access Risk Analyst**  
Meta 4.1 ★ (652 Bewertungen)  
Home Office · Homeoffice  
Erstellen Sie ein Indeed-Konto, bevor Sie zur Website des Unternehmens weitergeleitet werden.

[Weiter zur Bewerbung](#)

- 3+ years of working experience in access management, and/or information security capacity.
- Experience in information security concepts and applying them at a global scale.
- Experience independently leading projects to completion.
- Experience working with leadership and engineers.
- Experience working independently and collaboratively across various levels and teams.
- Communication, presentation, and interpersonal experience.
- Experience working across cross-functional and global teams.
- Experience managing competing priorities and simultaneous projects.

**Preferred Qualifications:**

- B.A.B.S in Computer Science or equivalent quantitative field, or International Relations.
- Strong desire to learn and continuously develop and deepen technical skills.
- Certifications in one or more of the following areas: CISSP, CISA, CISM, CISO, SCRM, CPM.
- Strong track record of understanding and interest in current and emerging technologies demonstrated through training, job experience and/or industry activities.
- Independent worker and motivated self-starter; thrives on ambiguity.
- Change-oriented – proactively generates process improvements, suggests and drives change, and confronts difficult circumstances in creative ways.

Facebook  
vor 2 Tagen  
[Diesen Job melden](#)

Right click the page and "Inspect" the page

# Type of occupations

## HTML source of this link

```
<div class="heading4 color-text-primary singleLineTitle tapItem-gutter">
  <h2 class="jobTitle jobTitle-color-purple">
    <a aria-label="full details of RN Specialty Practice Clinic" class="jcs-JobTitle" data-hide-spinner="true" data-hiri
      <span title="RN Specialty Practice Clinic">
        RN Specialty Practice Clinic
      </span>
    </a>
  </h2>
</div>
<div class="heading6 company_location tapItem-gutter companyInfo">
  <span class="companyName">
    Androscoggin Valley Hospital - NURSING - PHYS...
  </span>
<div class="companyLocation">
  Berlin, NH 03570
</div>
<div class="heading6 tapItem-gutter metadataContainer noJEMChips salaryOnly">
  <div class="metadata estimated-salary-container">
    <span class="estimated-salary">
      <svg aria-hidden="true" aria-label="Estimated $63.9K to $80.9K a year" fill="none" role="presentation" viewBox="0 0
    </span>
    <defs>
  </defs>
```

Look across divs (generic block element) and within "h2" match the attribute "class":"jobTitle jobTitle-color-purple" and extract content in the "span" (in-block element)

```
> jobs %>% head()  
[1] "RN Specialty Practice Clinic" "Communications Assistant"  
[3] "Stocker" "Director of Rehabilitation/DOR - Gorham, NH"  
[5] "Nursing Unit Aide" "Administrative Assistant to the Superintendent of School"
```

# New Data vs Conventional Sources

Strengths

Weakness

# New Data vs Conventional Sources

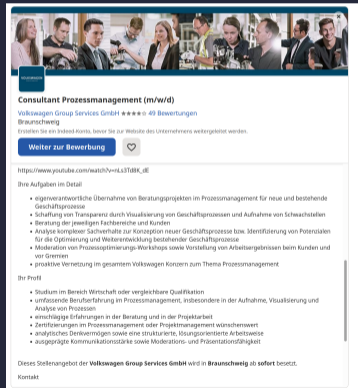
## Strengths

- You can better identify demand (matches confounds supply characteristics)


# New Data vs Conventional Sources

## Strengths

- You can better identify demand (matches confounds supply characteristics)
- Rich and customizable



**Consultant Prozessmanagement (m/w/d)**  
Volkswagen Group Services GmbH ★★★★★ 49 Bewertungen  
Braunschweig  
Erstellen Sie ein Indeed-Konto, bevor Sie zur Website des Unternehmens weitergeleitet werden.

[Weiter zur Bewerbung](#) 

[https://www.youtube.com/watch?v=rLSTd8K\\_dE](https://www.youtube.com/watch?v=rLSTd8K_dE)

**Ihre Aufgaben im Detail**

- eigenverantwortliche Übernahme von Beratungsprojekten im Prozessmanagement für neue und bestehende Geschäftsprozesse
- Schaffung von Transparenz durch Visualisierung von Geschäftsprozessen und Aufnahme von Schwachstellen
- Beratung der jeweiligen Fachbereiche und Kunden
- Analyse komplexer Sachverhalte zur Konzeption neuer Geschäftsprozesse bzw. Identifizierung von Potenzialen für die Optimierung und Weiterentwicklung bestehender Geschäftsprozesse
- Moderation von Prozessoptimierungs-Workshops sowie Vorstellung von Arbeitsergebnissen beim Kunden und vor Gremien
- proaktive Vernetzung im gesamten Volkswagen Konzern zum Thema Prozessmanagement

**Ihr Profil**

- Studium im Bereich Wirtschaft oder vergleichbare Qualifikation
- umfassende Berufserfahrung im Prozessmanagement, insbesondere in der Aufnahme, Visualisierung und Analyse von Prozessen
- einschlägige Erfahrungen in der Beratung und in der Projektarbeit
- Zertifizierungen im Prozessmanagement oder Projektmanagement wünschenswert
- analytisches Denkvermögen sowie eine strukturierte, lösungsorientierte Arbeitsweise
- ausgeprägte Kommunikationsstärke sowie Moderations- und Präsentationsfähigkeit

Dieses Stellenangebot der **Volkswagen Group Services GmbH** wird in **Braunschweig** ab sofort besetzt.

[Kontakt](#)



# New Data vs Conventional Sources

## Weakness

- Incomplete and Non-Representative

The screenshot shows a job listing on the Indeed website. The search criteria are 'Was: Schullehrer' and 'Wo: Stadt, Bundesland, Postleitzahl oder "Homeoffice"'. The job title is 'Nachhilfelehrer:in (m/w/d) in Deutsch, Mathematik oder Englisch' by 'Dustin Maszutt | Nachhilfe & Lernförderung Delmenhorst'. The job is categorized as 'Freie Mitarbeit' and is marked as 'Dringend gesucht'. The job description includes details about the role as a family-run tutoring service and lists tasks such as organizing the tutoring service, consulting with parents, and conducting lessons. It also mentions that the position is a full-time role with flexible working hours and no fixed working times.

**Lebenslauf anlegen - Einfache Bewerbung auf tausende Jobs.**

Schullehrer jobs  
Sortieren nach: Relevanz - Datum Seite 1 von 5 Jobs

**Nachhilfelehrer:in (m/w/d) in Deutsch, Mathematik oder Engl...**  
Dustin Maszutt | Nachhilfe & Lernförderung  
Delmenhorst • 4.0/5

Jetzt direkt bewerben • Aktiver Arbeitgeber • Dringend gesucht

Wir sind ein inhabergeführtes, mittelständisches Familienunternehmen, welches an mehreren Standorten qualifizierte hochwertigen Nachhilfeunterricht für Schüler...

Wir haben 4 Stellenangebote, ähnlich der bereits angezeigten anfordern wir eine monatliche Gehaltsangabe von ca. 1000€, wendest du dir über diese und bewirbst du die ausgeschriebene Stelle gerne wir.

Laden Sie Ihren Lebenslauf mühelos hoch

**Nachhilfelehrer:in (m/w/d) in Deutsch, Mathematik oder Englisch**  
Dustin Maszutt | Nachhilfe & Lernförderung  
Delmenhorst  
Freie Mitarbeit

Wie in den letzten 30 Tagen auf 51-74 % der Bewerbungen geantwortet, dauert üblicherweise bis zu 7 Tagen.

Schnellbewerbung

**Dringend gesucht**

**Stellenbeschreibung**

**Anstellungsart**  
Freie Mitarbeit

**Vollständige Stellenbeschreibung**

Wir sind ein inhabergeführtes, mittelständisches Familienunternehmen, welches an mehreren Standorten qualifizierte hochwertigen Nachhilfeunterricht für Schüler:innen organisiert.

Wir sind schon mehrere Jahre in der Region aktiv und suchen ab sofort eine geeignete **Nachhilfelehrkraft (m/w/d)**, die unser Lehrkräfte-Team unterstützen und unsere Schüler:innen in den Fächern Deutsch, Mathematik und/oder Englisch (Biswas bis 10. Klasse) weiterbringen möchte.

**Aufgabengebiet:**

- Organisation und Durchführung des Nachhilfeunterrichts
- Nachbesprechung des Unterrichtsstoffs mit dem Eltern
- Leistungsbesprechung mit dem Schullehrer:innen

**Wir bieten:**

- eine abwechslungsreiche Tätigkeit
- ein hohes Maß an Handlungsspielraum bei der Arbeit
- flexible Arbeitszeiten (frei einstellbar, keine fest vorgegebenen Arbeitszeiten)

# New Data vs Conventional Sources

## Weakness

- Incomplete and Non-Representative
- Time dimension
- Unless you are Google, might not be easy to web scrape

## When web scraping online info?

Find data that does not exist elsewhere

1. Data is generated by contributions from large user base (Uber, Amazon)
2. Data is itself just measurement of activity on website (forums like Reddit) or network (Facebook, LinkedIn, WeiBo)

## Why web scraping online info?

Re-arrange data in more convenient format

1. Data from many sources aggregated on one site (Wiki on Civil Wars)
2. Parsing techniques of webscraping can also be used when data provider gives you data in inefficient form (ex: max 1000 spreadsheets)

# Literature using online data I

- **Job posting:**

- Kuhn, P., Shen, K. (2013). Gender discrimination in job ads: Evidence from China. QJE
- Deming et al. (2018). Skill requirements across firms and labor markets: Evidence from job postings for professionals.
- Javorcik et al. (2019). The Brexit vote and labour demand
- Acemoglu et al. (2020). AI and jobs: Evidence from online vacancies.

- **Rental and real estate:**

- Halket and Pignatti (2015): scrape Craigslist to study US rental market
- Horn, K., Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. Journal of Housing Economics
- Yilmaz, O., Talavera, O., Jia, J. (2020). Liquidity, seasonality, and distance to universities: The case of UK rental markets

## Literature using online data II

- **Online vs offline prices:**

- Chevalier et al. (2003). Measuring prices and price competition online: Amazon.com and BarnesandNoble.com.
- Ellison, G., Ellison, S. F. (2009). Search, obfuscation, and price elasticities on the internet. *Econometrica*
- Cavallo and Rigobon (2015): "Billion Prices Project"
- Cavallo, A. (2017). Are online and offline prices similar? Evidence from large multi-channel retailers. (*AER*)
- Gorodnichenko, Y., Talavera, O. (2017). Price setting in online markets: Basic facts, international comparisons, and cross-border integration. *American Economic Review*,
- Cavallo, A. (2018). Scraped data and sticky prices. *RESTAT*

- **Consumption behaviors:**

- Baye, M. R., Morgan, J. (2009). Brand and price advertising in online markets. *Management Science*
- Davis and Dingell (2016): use Yelp to look ethnic segregation in consumption

## Where to start

- Does this data already exist?
- Would the site be willing to give you the data or partner with you?
- Do they have an API?
  - .. if not, extract website data yourself!

## How to continue

- If the data is user-contributed, who are the users?
  - Is selection bias going to be a big problem? (eg: which houses are rented on craigslist, who lists on eBay, who posts on social media?)
- Does the site customize the data based on characteristics of the browser? Can you deal with this ?
  - location of IP address, time of day, frequency of visits, ..)? Fake cookies?
- Do you need a panel?
  - If the website changes or is pulled down, could you write a paper with just a few periods?
- How much measurement error can you tolerate in your research design?
  - if you assemble a panel and a few observations per period fail due to impartial scrapes, is this manageable?



## Overview of today

1. Main economic data sources available
2. How to use an API to access data
3. Basics of web scraping

## Applications

1. The billions Prices Project

## Conventional Sources of Data in Economics

- **IMF data** World Economic Outlook database, time series on a wide number of countries for macro aggregates.
- **World Bank WDI** the World Development Indicators
- **FRED** US economy, such as GDP, interest rates, financial indicators and monetary aggregate
- **OECD**: data from the Organization for Economic Cooperation and Development.
- **UNCTAD**: trade and development data, including COMTRADE
- **Eurostat** aggregates data from European stats offices and more
- **CEPII**: trade related database: BACI better version of COMTRADE, MACMAP for tariffs, Geography variables (gravity etc)

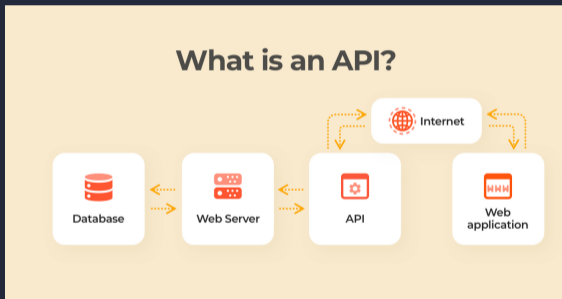
- **EUKLEMS**: it is a project that provides industry-level time series data for developed countries in a consistent accounting framework.
- **Penn World Table**: Real national accounts converted in dollars to provide unified framework and comparability.
- **Policy Uncertainty**: the Economic Policy Uncertainty index for developed countries, for US it provides also components of the series, which is both news- and data-based. Authored and maintained by Baker, Bloom and Davis.
- **Real-time data**: the Philadelphia Fed provides real-time data for a number of policy-relevant variables, good for estimations of policy rules in the US economy.

- **BIS**: the Bank of International Settlements provides financial data on a set of countries, along a wide range of dimensions.
- **BLS**: data on the US economy, also providing granular breakdowns at occupational and frequency levels
- **Groningen Data**: The GGDC 10-Sector Database provides a long-run internationally comparable dataset on sectoral productivity performance in Africa, Asia, and Latin America.

API

# API

Application Programming Interface: online tool to access info or download raw data



## API Data format

1. XML
2. JSON
3. CSV, RSS, etc.

# API Data format

## 1. XML (Extensible Markup Language) - Node structure

```
<job-offers>
  <job>
    <job-title> Data Analyst </job-title>
    <location> Berlin </location>
    <benefits>
      <salary> 50k </salary>
      <remote> yes </remote>
      <type> full-time </type>
    </benefits>
  </job>
</job-offers>
```

<job-title>   Data Analyst   </job-title>  
Node opening tag   value   Node closing tag

## 2. JSON



## API Data format

1. XML
2. JSON - Key-value pairs structure

```
{ "job-offers": [{  
  "job-title": "Data Analyst",  
  "location": "Berlin",  
  "benefits": {  
    "salary": "50k",  
    "remote": "yes",  
    "type": "full time"  
  }  
}] }
```

"job-title": "Data Analyst"  
Key value

"benefits": {"salary": "50k"}  
Key Key value  
value

# APIs

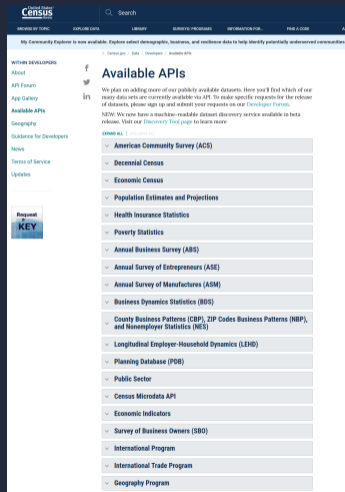
- List of Economic Institutions providing APIs
- Financial data: Tiingo, Bloomberg
- Github page with all sort of free (at least partially) APIs

API	Description	Auth	HTTPS	CORS
API Title(Link to API documentation)	Description of API	Does this API require authentication? *	Does the API support HTTPS?	Does the API support CORS? *

- Auth = OAuth, apiKey, no
- CORS = Without proper CORS configuration an API will only be usable server side.

# US Census API

R Consortium Census Working Group makes it easier via packages on CRAN (here)



The screenshot shows the US Census API website. The header includes the US Census logo, a search bar, and navigation links: "EXPLORE BY TOPIC", "EXPLORE DATA", "LIBRARY", "SUBJECTS PROGRAMS", "INFORMATION FOR...", and "PICK A CENSUS". Below the header, a message states: "My Community Explorer is now available. Explore select demographics, business, and residence data to help identify potentially underserved communities." The main content area is titled "Available APIs" and includes a paragraph: "We plan on adding more of our publicly available datasets. Here you'll find which of our many data sets are currently available via API. To make specific requests for the release of datasets, please sign up and submit your requests on our [Developer Forum](#)." Below this, a "NEW" notice says: "We now have a machine-readable dataset discovery service available in beta release. Visit our [Discovery Tool](#) page to learn more." A list of APIs is shown with expandable sections:

- AMERICAN COMMUNITY SURVEY (ACS)
  - Decennial Census
  - Economic Census
  - Population Estimates and Projections
  - Health Insurance Statistics
  - Poverty Statistics
- ANNUAL BUSINESS SURVEY (ABS)
- ANNUAL SURVEY OF ENTREPRENEURS (ASE)
- ANNUAL SURVEY OF MANUFACTURES (ASM)
- BUSINESS DYNAMICS STATISTICS (BDS)
- COUNTY BUSINESS PATTERNS (CBP), ZIP CODES BUSINESS PATTERNS (NBP), and Nonemployer Statistics (NES)
- LONGITUDINAL EMPLOYER-HOUSEHOLD DYNAMICS (LEHD)
- PLANNING DATABASE (POB)
- PUBLIC SECTOR
- CENSUS MICRODATA API
- ECONOMIC INDICATORS
- SURVEY OF BUSINESS OWNERS (SBO)
- INTERNATIONAL PROGRAM
- INTERNATIONAL TRADE PROGRAM
- GEOGRAPHY PROGRAM

On the left sidebar, there are links for "WITHIN DEVELOPERS" (About, API Forum, App Gallery), "Available APIs" (Geography), "Guidance for Developers", "News", "Terms of Service", and "Updates". At the bottom of the sidebar, there is a "Request a KEY" button.

## API key setup

Many institutions provide an Open API or access via a Key (here key for the Census)

```
# Add key to .Renviron
Sys.setenv(CENSUS_KEY=yourkey)
# Reload .Renviron
readRenviron("~/Renviron")
# Check to see that the expected key is output in your R console
Sys.getenv("CENSUS_KEY")
```

## Listing available datasets

censusapi. A wrapper for the U.S. Census Bureau APIs that returns data frames of Census data and metadata

```
> censusapi::listCensusApis()  
title                               name  vintage  
Economic Surveys: Annual Business Survey  abscb  2017  
Economic Surveys: Annual Business Survey  abscb  2018  
Economic Surveys: Annual Business Survey  abscb  2019
```

## Trade data in the US Census

```
exp_monthly_hs <- "timeseries/intltrade/exports/hs" #this is the name
censusapi::listCensusMetadata(name = exp_monthly_hs, type = "variables")
```

name	label
NT_WGT_YR	15-digit Year-to-Date Containerized Vessel Shipping Weight
DF	2-character Domestic or Foreign Code
MONTH	2-character Month
QTY_1_MO	15-digit Quantity 1
COMM_LVL	4-character aggregation levels for commodity code.
UNIT_QY2	3-character Export Unit of Quantity 2
UNIT_QY1	3-character Export Unit of Quantity 1
CNT_VAL_MO	15-digit Containerized Vessel Value
CTY_CODE	4-character Country Code
DIST_NAME	50-character District name
VES_VAL_YR	15-digit Year-to-Date Vessel Value

## Trade data in the US Census

```
> exports <- censusapi::getCensus(  
+   name = "timeseries/intltrade/exports/hs",  
+   key = Sys.getenv("CENSUS_KEY"),  
+   vars = c("E_COMMODITY", "CTY_CODE", "CTY_NAME", "ALL_VAL_MO", "ALL_VAL_YR"),  
+   YEAR=2020,  
+   MONTH="05",  
+   COMM_LVL="HS6")  
> head(exports)
```

	E_COMMODITY	CTY_CODE	CTY_NAME	ALL_VAL_MO	ALL_VAL_YR	YEAR	MONTH	COMM_LVL
1	091012	0025	EURO AREA	0	47317	2020	05	HS6
2	401694	0025	EURO AREA	34429	110219	2020	05	HS6
3	610210	0025	EURO AREA	0	571110	2020	05	HS6
4	030692	0025	EURO AREA	0	596724	2020	05	HS6
5	160419	0025	EURO AREA	0	10133	2020	05	HS6
6	843930	0025	EURO AREA	0	19500	2020	05	HS6

# WEB SCARPING



# Definitions

- Scraping: Using tools to gather data you can see on a webpage
  - Parsing: The act of analyzing the text (HTML, ...) to collect the data you need
  - Crawling: Moving across or through a website to gather data from multiple URL/pages
- 2 main technologies to build a webpage:
  - HTML (the Hypertext Markup Language) structure of the page
  - CSS (Cascading Style Sheets) visual and aural layout, for a variety of devices.

# Web Crawling

`https://de.indeed.com/Jobs?q=Analyst&l=Berlin&start=10`

protocol                      host                      path

- "q=" query for type of job, separating search terms with "+" (i.e. "school+teacher" jobs)
- "&l=" begins the string for location, again "North+Westphalia"
- "&start=" index the number of item you want to see

# HTML

- HTML: similar to XML, both use tags and node structure
- different functions: HTML displays content on a web page, XML represents data in a hierarchical structure
- XML is case-sensitive while HTML is not
- [Link here to tags in html](#)

`<p>` `this is a paragraph` `</p>`  
start tag      HTML element      close tag

- in HTML you can leave tags open, in XML not

# HTML vs CSS

Save the following in a file .html and open via the browser

```
<!DOCTYPE html>
<html>
<head>
<!-- We define the layout of paragraph using CSS -->
<style>
p {
  color: purple;
  text-align: center;
}
</style>
</head>
<body>

<p>Hello World!</p>
<p>These paragraphs are styled with CSS.</p>

</body>
</html>
```

## Web Scraping without getting blocked

- Check website's ToS: limit the number of requests? size?
- try to emulate human-like browsing behavior by adding random pauses
  - scraping tools/ library: Scrapy, BeautifulSoup, rvest, Selenium
- use proxies from different IP addresses, making it harder for the website to track your activity.

# THE BILLION PRICES PROJECT

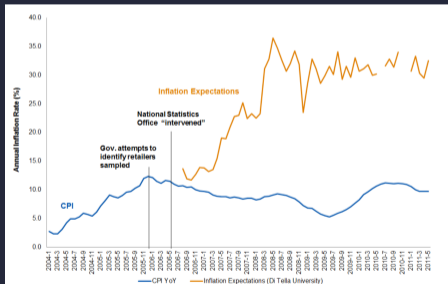
*“By 2010, we were collecting 5 millions prices every day.”*

*Alberto Cavallo and Roberto Rigobon 2016*

- Institutions: Inflacion Verdadera, MIT, PriceStats
- Daily price data since 2008
- From hundreds of large multi-channel retailers
- In over 60 countries

## How did it start?

- Argentina's inflation data (2007-2015) was widely questioned
- CPI inflation remained below 10% for many years
- Survey inflation expectations consistently above 25%





# CPI

- Calculating the CPI for a single item

$$\text{consumer price index} = \frac{\text{market basket of desired year}}{\text{market basket of base year}} \times 100$$

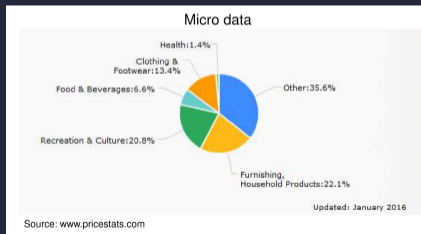
- Calculating the CPI for multiple items  $i = 1, \dots, n$

$$\text{CPI} = \frac{\sum_{i=1}^n \text{CPI}_i \times \text{weight}_i}{\sum_{i=1}^n \text{weight}_i}$$

Ideally, the weights would relate to the composition of expenditure during the time between the price-reference period and the current period.

## How representative is the online basket?

- On average, prices from about 60% of CPI expenditure weights can be found online
- All kinds of goods, including food and fuel
- Relatively few services (improving)
- Housing/rents information online, but not directly included the indices



# Comparing Data Sources

*Table 1*  
**Alternative Micro-Price Data Sources**

	<i>Online data</i>	<i>Scanner data</i>	<i>CPI data</i>
Cost per observation	Low	Medium	High
Data frequency	Daily	Weekly	Monthly
All products in retailer (Census)	Yes	No	No
Uncensored price spells	Yes	Yes	No
Countries with research data	~60	<10	~20
Comparable across countries	Yes	Limited	Limited
Real-time availability	Yes	No	No
Product categories covered	Few	Few	Many
Retailers covered	Few	Few	Many
Quantities or expenditure weights	No	Yes	Yes

*Source:* Table 1 from Cavallo (2015).

*Notes:* The Billion Prices Project ([bpp.mit.edu](http://bpp.mit.edu)) datasets contain information from over 60 countries with varying degrees of sector coverage. Nielsen US scanner datasets are available at the Kilts Center for Marketing at the University of Chicago. Klenow and Malin (2010) provide stickiness results with Consumer Price Index data sources from 27 papers in 23 countries. See Cavallo (2013) for more details.

# Data Collection

- a random set of goods
- at the website and any physical store of a given retailer
- in a similar time span (7-day window)
- 10 countries, 5-10 retailers in each
- Retailers included:
  - top 20 largest retailers by market share
  - Sell both online and offline
  - Availability of a product id (UPC)

# Data Validation

Short-run discrepancies (mainly in developing c.) but medium/long-term co-movement

## Developing vs Developed Countries

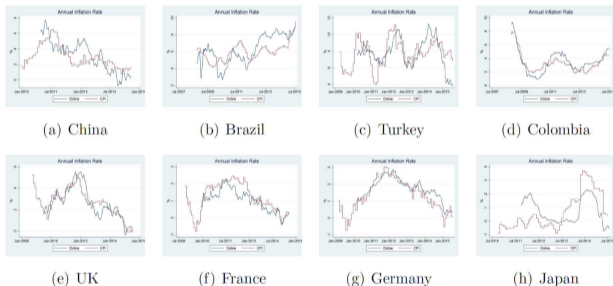
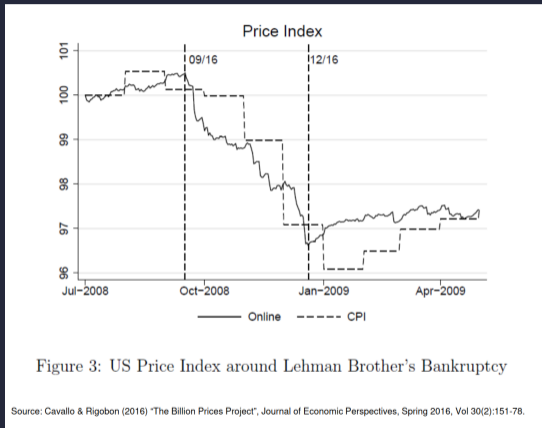


Figure 5: Online vs CPI Annual Inflation Rates

Source: Cavallo & Rigobon (2016) "The Billion Prices Project", Journal of Economic Perspectives, Spring 2016, Vol 30(2):151-78.

# What is new with Online Data?

## Anticipation of changes in inflation trends

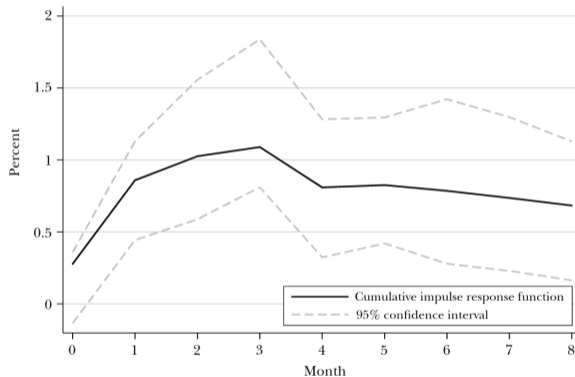


It took more than two months after Lehman's disaster for the official Consumer Price Index numbers to reflect the full impact on price levels!

Figure 7

**Cumulative Impulse Response of the US Consumer Price Index (CPI) to an Online Price Index Shock**

(response to a 1% shock in the online index)



Source: Authors using online data computed by PriceStates and US Consumer Price Index.

Notes: The Consumer Price Index is a US city average, all items non-seasonally adjusted from the Bureau of Labor Statistics. Data from July 2008 to January 2015.

## What is new with Online Data?

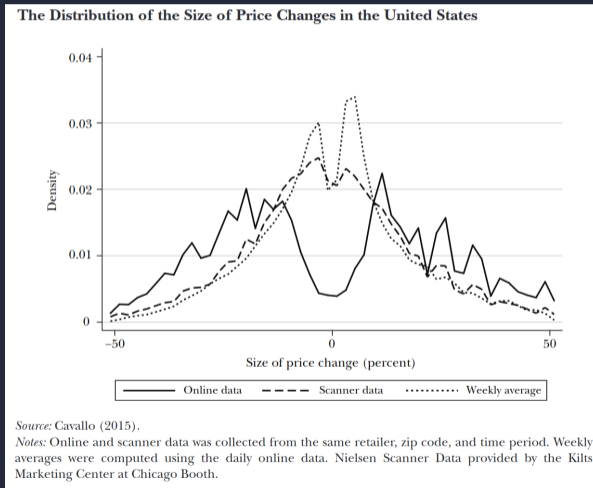
Apart from measurement issues, online prices allowed to offer new light on

- price stickiness
- Law of One Price (LOOP)



## Price Stickiness?

With offline data, no evidence of menu pricing... price stickiness (Woodfrod 2009, Midrigan 2011)?



## Law of One Price

The law of one price is the theory that an economic good or asset will have the same price in different markets (controlling for transaction costs).

Issue: comparing the same good in different countries (barcodes differ, id names).

Solution: global retailers IKEA, Zara etc.

→ LOOP is valid within countries that use the same currency

Inflation with Covid Consumption Baskets\*

Alberto Cavallo  
Harvard Business School & NBER

December 2020

**Abstract**

The Covid-19 Pandemic led to changes in expenditure patterns that can introduce significant bias in the measurement of Consumer Price Index (CPI) inflation. Using publicly-available data on card transactions, I update the official CPI weights and re-calculate inflation with Covid consumption baskets. I find that the US CPI underestimated the Covid inflation rate, as consumers spent relatively more on food with positive inflation, and less on transportation and categories experiencing deflation. The bias peaked in May, when US Covid annual inflation was 0.95% compared to just 0.13% in the CPI and low-income households were experiencing nearly twice as much inflation as those at the top of the income distribution. I find similar evidence of higher Covid inflation in 12 of 19 additional countries.

**JEL-Code:** C43;E31;E32

**Keywords:** COVID, consumer expenditures, CPI, inflation.

## Inflation with Covid Consumption Baskets

Alberto Cavallo

NBER Working Paper Series, No. 27352

[Download Paper](#) | [Download Data](#)

The Covid-19 Pandemic led to changes in expenditure patterns that can introduce significant bias in the measurement of Consumer Price Index (CPI) inflation. Using publicly-available data on card transactions, I update the official CPI weights and re-calculate inflation with Covid consumption baskets. I find that the US CPI underestimated the Covid inflation rate, as consumers spent relatively more on food with positive inflation, and less on transportation and categories experiencing deflation. The bias peaked in May, when US Covid annual inflation was 0.95% compared to just 0.13% in the CPI and low-income households were experiencing nearly twice as much inflation as those at the top of the income distribution. I find similar evidence of higher Covid inflation in 12 of 19 additional countries.

\*I am grateful to Francesco Furler for excellent research assistance, to John Friedman for sharing the Opportunity Insights data, to Olivier Cochet and Edward Bertsch for help with the CPI data, and to Robert DS Yoho, Michael Tomzard, Dan Siskel, Tarek Jorjani, and Dan Kilian for helpful comments and suggestions. Financial support for this paper was provided by Harvard Business School.

## Tariff Passthrough at the Border and at the Store: Evidence from US Trade Policy\*

Alberto Cavallo  
Harvard University

Gita Gopinath  
Harvard University and IMF

Brent Neiman  
University of Chicago

Jenny Tang  
Federal Reserve Bank of Boston

October 2019

### Abstract

We use micro data collected at the border and at retailers to characterize the effects brought by recent changes in US trade policy -- particularly the tariffs placed on imports from China -- on importers, consumers, and exporters. We start by documenting that the tariffs were almost fully passed through to total prices paid by importers, suggesting the tariffs' incidence has fallen largely on the United States. Since we estimate the response of prices to exchange rates to be far more muted, the recent depreciation of the Chinese renminbi is unlikely to alter this conclusion. Next, using product-level data from several large multi-national retailers, we demonstrate that the impact of the tariffs on retail prices is more mixed. Some affected product categories have seen sharp price increases, but the difference between affected and unaffected products is generally quite modest, suggesting that retail margins have fallen. These retailers' imports increased after the initial announcement of possible tariffs, but before their full implementation, as the intermediate passthrough of tariffs to their prices may not persist. Finally, in contrast to the case of foreign exporters facing US tariffs, we show that US exporters lowered their prices on goods subjected to foreign retaliatory tariffs compared to exports of non-targeted goods.

JEL-Codes: F11, F13, F14, F16

Keywords: trade policy, tariffs, exchange rate passthrough.

\*This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed herein are those of the authors and do not necessarily reflect the views of the BLS, the Federal Reserve Bank of Boston, the Federal Reserve System, or those of the IMF, the Executive Board, or Management. We are grateful to Alan Ales for his substantial efforts as BLS project coordinator, to Florence Wu, Keith Burdette, Douglas No, and Augustus Olinde for excellent research assistance, and to Chad Stone and Ulrich for his helpful comments and suggestions. Alberto Cavallo is a shareholder of Palantir LLC, a private company that provides proprietary data used in this paper without any requirements to revise the findings prior to their release.

# Tariff Passthrough at the Border and at the Store: Evidence from U.S. Trade Policy

Alberto Cavallo, Gita Gopinath, Brent Neiman, and Jenny Tang

American Economic Review: Insights, Vol 3 Issue 1. March 2021

[Download Paper](#) | [Download Data](#)

We use micro data collected at the border and at retailers to characterize the effects brought by recent changes in US trade policy -- particularly the tariffs placed on imports from China -- on importers, consumers, and exporters. We start by documenting that the tariffs were almost fully passed through to total prices paid by importers, suggesting the tariffs' incidence has fallen largely on the United States. Since we estimate the response of prices to exchange rates to be far more muted, the recent depreciation of the Chinese renminbi is unlikely to alter this conclusion. Next, using product-level data from several large multi-national retailers, we demonstrate that the impact of the tariffs on retail prices is more mixed. Some affected product categories have seen sharp price increases, but the difference between affected and unaffected products is generally quite modest, suggesting that retail margins have fallen. These retailers' imports increased after the initial announcement of possible tariffs, but before their full implementation, so the intermediate passthrough of tariffs to their prices may not persist. Finally, in contrast to the case of foreign exporters facing US tariffs, we show that US exporters lowered their prices on goods subjected to foreign retaliatory tariffs compared to exports of non-targeted goods.

## Access to the project

- data and results available at the link webpage ([bpp.mit.edu](http://bpp.mit.edu))
- ..also script to replicate the results!
- data are not pretreated
- The US and Argentina inflation indexes used in this paper are published with a 30-day lag on the Billion Prices Project website
- PPP exchange rate information discussed in the previous section are currently published with a one-year lag on the PriceStats website
- raw micro data collected by PriceStats are not publicly available but can be shared with academic researchers who collaborate with the Billion Prices Project and sign a data-access agreement.