CENTER ON RURAL INNOVATION

BEYOND RAW DATA

Streamlining Insights with Curated Data Packages

Olivier Leroy
Center On Rural Innovation

2024-10-20

SLIDES ARE ONLINE:

- Created in Quarto (RevealJS)
- Running R and cori.data.fcc
- Hosted and deployed on GitHub

  https://ruralinnovation.github.io/prez_auber_2024/

GOALS:

1. Primary data is great! (but can be a pain...)
2. Use code (other pain?) to work with primary data
3. Packaging it will make live easier!

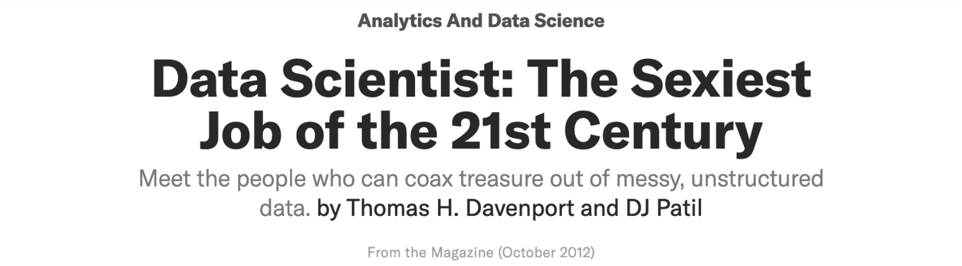
Center on Rural Innovation

PLAN: RETRACE OUR JOURNEY

[TODO: rework at the end]

👋HI, I AM OLIVIER SENIOR DATA ENGINEER

- Working with primary data: “messy, unstructured data” is upon us



Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Source: Harvard Business Review

- Socials: LinkedIn | Mastodon | Personal Website

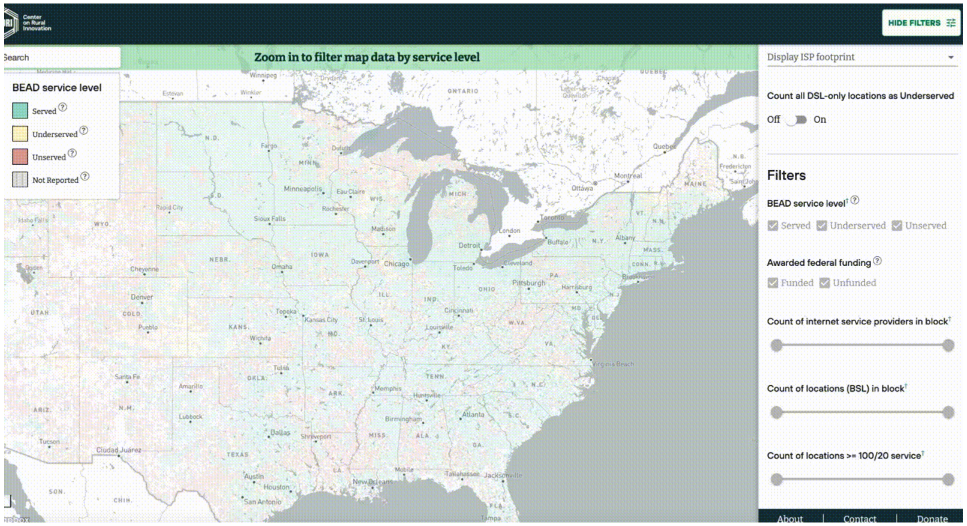
DATA AT CENTER ON RURAL INNOVATION (CORI)

1. Rural data: we can’t work with aggregate and need to go deeper

I plan to delete this unless good points against

BROADBAND: AN IMPORTANT PART OF OUR WORKS

PUBLIC-FACING APPS:



Center on Rural Innovation

HIDE FILTERS

Search

Zoom in to filter map data by service level

BEAD service level

- 🟩 Served
- 🟨 Underserved
- 🟥 Unserved
- ▨ Not Reported

Display ISP footprint

Count all DSL-only locations as Underserved

Off ● On

Filters

BEAD service level[†]

☑ Served ☑ Underserved ☑ Unserved

Awarded federal funding

☑ Funded ☑ Unfunded

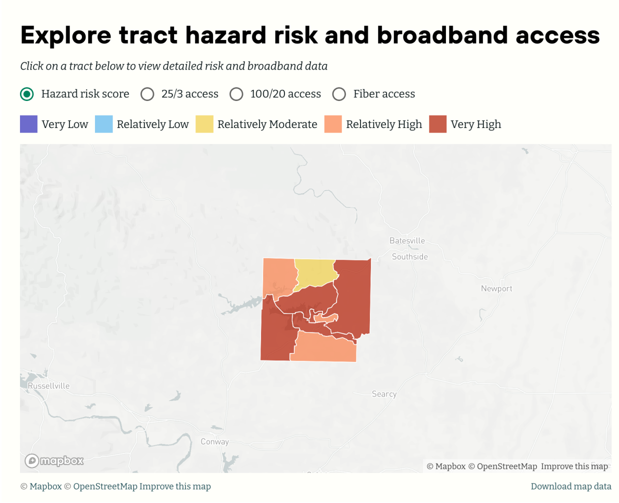
Count of internet service providers in block[†]

Count of locations (BSL) in block[†]

Count of locations >= 100/20 service[†]

About | Contact | Donate

Rural Broadband Mapping Tool



Explore tract hazard risk and broadband access

Click on a tract below to view detailed risk and broadband data

● Hazard risk score ○ 25/3 access ○ 100/20 access ○ Fiber access

- 🟪 Very Low
- 🟦 Relatively Low
- 🟨 Relatively Moderate
- 🟧 Relatively High
- 🟥 Very High

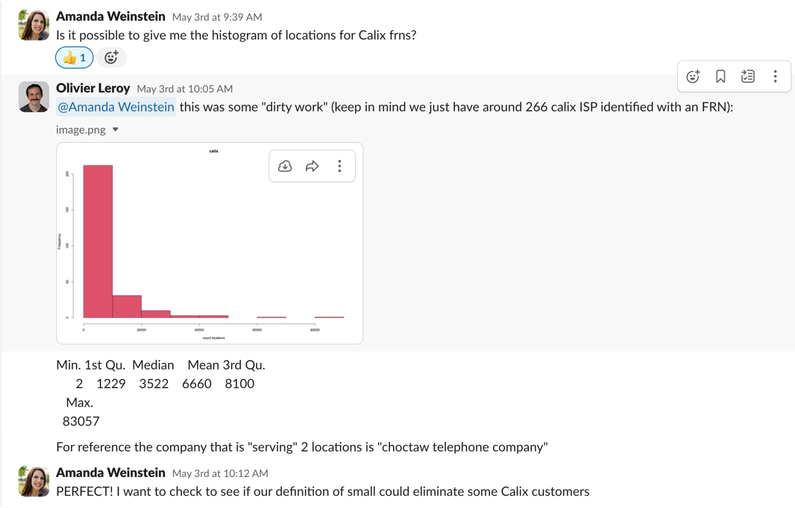
© Mapbox © OpenStreetMap Improve this map

© Mapbox © OpenStreetMap Improve this map

Download map data

Broadband Climate Risk Mitigation Tool

RESEARCH WORK: FROM IDEATION TO REPORT



Amanda Weinstein May 3rd at 9:39 AM
Is it possible to give me the histogram of locations for Calix frns?

Olivier Leroy May 3rd at 10:05 AM
@Amanda Weinstein this was some "dirty work" (keep in mind we just have around 266 calix ISP identified with an FRN):

```
Min. 1st Qu.  Median    Mean 3rd Qu. 
   2    1229    3522    6660    8100 
  Max. 
 83057 
```

For reference the company that is "serving" 2 locations is "choctaw telephone company"

Amanda Weinstein May 3rd at 10:12 AM
PERFECT! I want to check to see if our definition of small could eliminate some Calix customers

Beyond Connectivity:

The Role of Broadband in Rural Economic Growth and Resilience

Dr. Amanda Weinstein
Director of Research, Knowledge, and Evaluation
Center on Rural Innovation

Dr. Adam Dewbury
Researcher
Center on Rural Innovation

CORI Center on Rural Innovation

BROADBAND RESEARCH IS EVOLVING!

FEDERAL COMMUNICATIONS COMMISSION (FCC)

- Agency in charge of regulating "everything" communications
- Their data landscape is evolving fast (now they have a platform!)

CHALLENGE: HOW TO WORK WITH CONSTANT CHANGE UPSTREAM?

FEDERAL COMMUNICATIONS COMMISSION (FCC)

	Form 477	National Broadband Map
US Census Boundaries	2010	2020
Type of recording	self declarative	self declarative
Granularity	Census blocks	Locations
Services	Mobile/ Fixed	Mobile/ Fixed
Timeframe	2014-2021	2022 - *Ongoing*
Releases	**twice a year**	**twice a year**[1]

[1] One release can have multiple versions

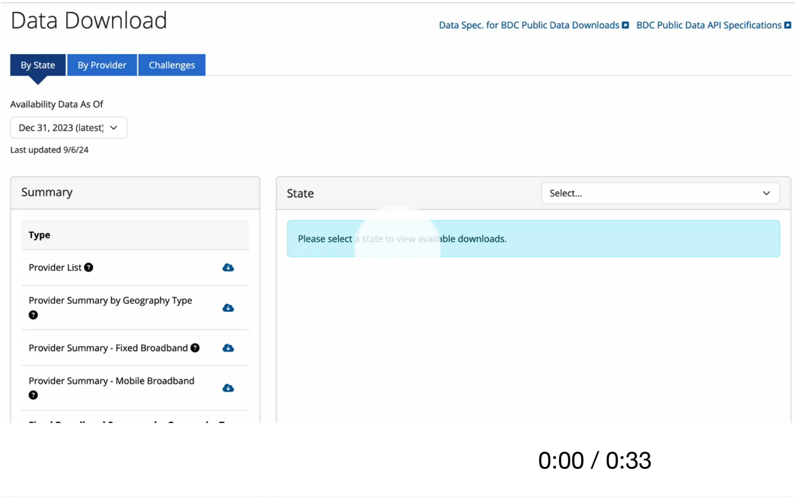
BROADBAND ACCESS SINCE 2014 BUT:

- 2 **voluminous** datasets, hundreds of files
- No "**road safety barriers**": multiple encodings, erronous values (and more..)

Center on Rural Innovation

BEHIND THE SCENE:

HOW IT STARTED: MANUEL EXTRACTION


Data Download
Data Spec. for BDC Public Data Downloads
BDC Public Data API Specifications
By State
By Provider
Challenges
Summary
Type
Provider List
Provider Summary by Geography Type
Provider Summary - Fixed Broadband
Provider Summary - Mobile Broadband
State
0:00 / 0:33

Availability Data As Of: Dec 31, 2023 (latest)

Last updated 9/6/24

State: Select...

Please select a state to view available downloads.

- 🔁Repeat for every State (56)
- 🔁Repeat for every version (??)
- 🔴Error prone (500 hundred clicks)

… AUTOMATION WITH CODE …

```
# [...]
# Download all FCC data needed
all : dirs dl_file list_file

include config.mk

## dirs:    Create specific dirs to download data
dirs : config.mk
	@pwd
	@mkdir -p $(DATA_PATH)
	echo "create $(DATA_PATH)"

## dl_file: Create a list of files availiable, download them
dl_file : $(CSV_SRC) get_csv_dl.R
	Rscript get_csv_dl.R $(CSV_SRC)
	Rscript list_file_dl.R $(DATA_PATH) $(CSV_SRC)
.PHONY : dl_file

```

- New projects -> New specs

CHALLENGE: COPY/PASTING AND MAINTAINING X SIMILAR CODE SOURCES

- Spoiler: Not everyone like SQL

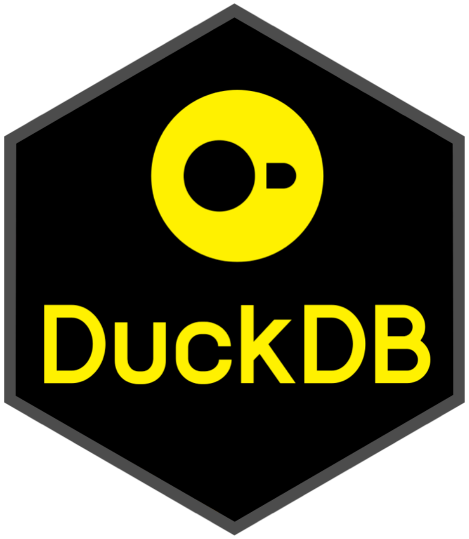
CHALLENGE: MAKE IT MORE ACCESSIBLE

- Out of memory -> Process in DB

CHALLENGE: SLOW, LACK OF

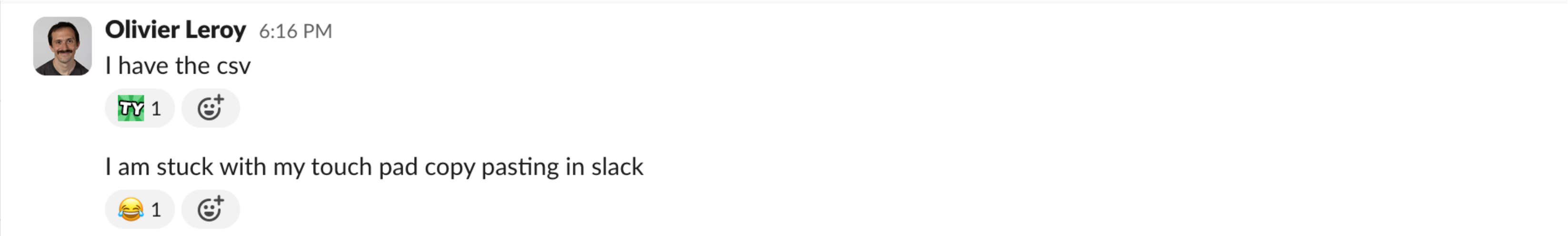
💡 LIGHT FROM POSIT::CONF

Workshop on DuckDB (Parquet / Arrow / OLAP)



SLACK IS BUZZING!

> **Warning**
>
> Contains real slack messages with typos!



Olivier Leroy 6:16 PM

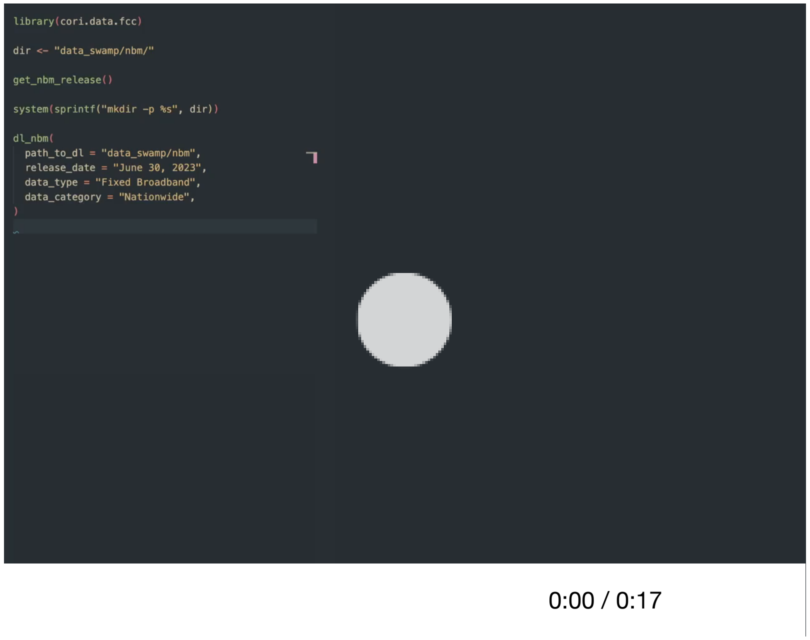
I have the csv

I am stuck with my touch pad copy pasting in slack

DUCKDB SOLVED THE LACK OF REACTIVITY!

… CODE PACKAGED

Visual Code



```
library(cori.data.fcc)

dir <- "data_swamp/nbm/"

get_nbm_release()

system(sprintf("mkdir -p %s", dir))

dl_nbm(
  path_to_dl = "data_swamp/nbm",
  release_date = "June 30, 2023",
  data_type = "Fixed Broadband",
  data_category = "Nationwide",
)
```

- DuckDB: in-process
- Reduce errors
- Complexity is in **one** place and **abstracted** so we can focus on what bring **value**!

CORI.DATA.FCC

Center
on Rural
Innovation

WHAT IS PACKAGED CODE?

ONE WHITE LIE: CODE THAT WORK OUTSIDE OF "MY" COMPUTER

BENEFITS FROM THE LANGUAGE ECOSYSTEM

- Tested against GitHub runner: it does not just work on my computer!
- Website to share it/distribute it

OPEN SOURCE: WHY SHARING IT?

- Way to **publish**: we get more value that way
- Way to **share**: code is used!
- A place were we can **exchange** in a constructive way

HOW TO GET IT:

The source code is hosted on GitHub (Version control)

You need the R package {remotes}.

```
install.packages("remotes")
remotes::install_github("ruralinnovation/cori.data.fcc")
```

🚧Check the version you have and see what new versions are offering 🚧

```
packageVersion("cori.data.fcc")
```

```
[1] '0.0.1'
```

WHAT CAN IT DO: CHOOSE YOUR OWN ADVENTURE!

- I need to go back to the source: Download data from FCC
- I need primary data but working with hundreds of CSV is not for me: Download raw data from CORI (NBM / Form 477)
- Census block is perfect for me: Download tranformed data for NBM from CORI (ISP / County)
- Guide you in those steps!

PACKAGING CODE = EXTRA TEAM MEMBER 👷

> we’ve talked about how packages can act like team members such as the **IT Guy**, **Analyst**, Tech Lead, or Project Manager.[1]
>
> > […] developed packages are […] extra expert team members.[2]

1. https://www.emilyriederer.com/post/team-of-packages/#collaboration
2. https://milesmcbain.xyz/posts/data-analysis-reuse/

CORI Center on Rural Innovation

PACKAGING CODE = INCREASE SPEED TO INOVATION / INSIGHT

- Functions designed to get data and load it in your environment
- **Capitalize on it**!
  - New project that is requiring FCC staff block estimates -> add it!

Center on Rural Innovation

EXAMPLES OF USES CASES

FIRST EXAMPLE: BLOCK COVERED BEAD LIKE

Place holder for Camdem blog post

WHAT ARE THE ISP IN OHIO?

Visual Code

Show 5 entries

Search:

Provider Name	Cnt Block	Cnt Rel.
Cox Communications, Inc	2163	9414
Ottoville Mutual Telephone Company	270	992
Telephone Service Company	570	2516

Provider Name	Cnt Block	Cnt Rel.
SYCAMORE TELEPHONE COMPANY	404	1437
The Champaign Telephone Company	531	2287

Showing 1 to 5 of 121 entries

Previous 1 2 3 4 5 … 25 Next

SUMMARY

- Primary data can give more power (especialy in new field)!
- (Primary) data is only useful if you can **effectively** use it.
- Use code that are abstracting the pain
- Easier to **capitalize/build** on it: generating less tech debt
- we need those extra “team members”

CONTACTS

Website: https://ruralinnovation.us

LinkedIn | Twitter | Facebook | Instagram | YouTube