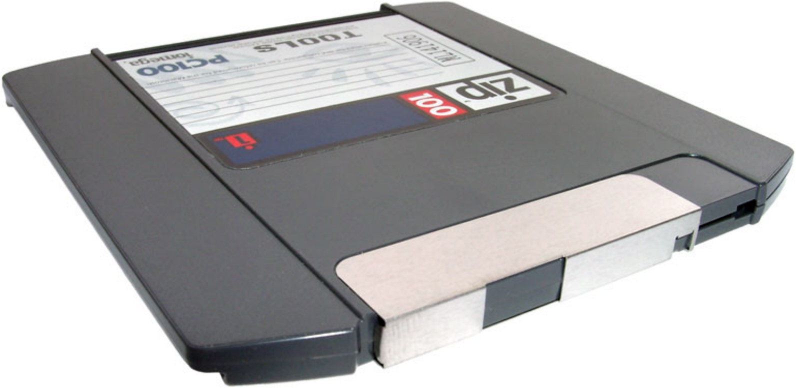
An introduction to data pipelines

Miklós Koren (@korenmiklos) Zoltán Tóth (@zoltanctoth)

CODEDTHINKING.
by Koren

Data science on the command line

Can you carbon date me?



My tools

```
economics,1994-
econometrics,1996-
stata,1997-
python,2003-
julia,2017-
```

Why data pipelines?

Many tools in our data science stacks.

Need to connect them properly.

1. reproducible
2. performant

Why command line?

Linux tools are built to do one thing well. They have had the need/solution to connect them for 40+ years.

For large datasets (often on cloud servers), you need tools to reduce size for analysis.

1. select columns
2. filter rows
3. aggregate by groups

Backstory

You are waiting for your flight at Paris CDG when you get a call from the Directeur du Trésor. He needs some analysis stat.

Your laptop battery just died. You only have access to a public internet terminal that runs Linux. You want to do an analysis that

1. only uses the tools at your disposal
2. can be fully reproduced later when you recharge your laptop
3. understandable by your colleagues

Data

1. Tenders Electronic Daily is a database compiled by the European Commission on public procurements in the EU. We use a 100k sample of the 2019 Contract Award Notices. The file in .csv format.
2. Resolve country codes.

A row in the TED dataset is a contract. Each contract has one or more buyers (there may be joint procurements of several agencies) and zero or more winners. Tenders may be invalidated (hence zero winner), but can also have multiple winners. We will focus on contracts with one buyer and one winner.