# Advanced Data Management (CSCI 640/490)

## Databases & Data Wrangling

Dr. David Koop

# pandas

- Contains high-level data structures and manipulation tools designed to make data analysis fast and easy in Python
- Originally built on top of NumPy
- Built with the following requirements:
  - Data structures with labeled axes (aligning data)
  - Support time series data
  - Do arithmetic operations that include metadata (labels)
  - Handle missing data
  - Add merge and relational operations

# polars

- Contains high-level data structures and manipulation tools designed to make data analysis **"lightning"** fast and easy in Python
  - Built using Apache Arrow
  - Written from scratch using Rust but with a Python API
  - Parallelized (uses multiple cores)
  - Intuitive API: "I came for the speed, but stayed for the syntax"

# Assignment 2

- Assignment 1 Questions with polars, DuckDB, and pandas
- CS 640 students do all, CS 490 do polars & DuckDB (pandas is EC)
- Can work by framework or by query
- Most questions can be answered with a single statement… but that statement can take a while to write
  - Read documentation
  - Check hints

# DuckDB

- A fast analytical, portable, in-process, open-source database system
  - Zero dependencies
  - Runs on all popular OSes, has client APIs for many programming languages
  - Supports common file formats (JSON, CSV, Parquet)
  - Fast
  - Extensible (extensions for spatial data, etc.
- "A modern embedded analytics database that runs on your machine and lets you efficiently process and query gigabytes of data from different sources" [M. Needham et al.]
- Minimal support for transactions: OLAP not OLTP!

# DuckDB: Bringing analytical SQL directly to your Python shell

[P. Holanda, 2023]

# DuckDB: Crunching data anywhere, from laptops to servers

[G. Szárnyas, 2024]

# Chicago Food Inspections Exploration

- Using Polars
- Using Pandas
- Using DuckDB