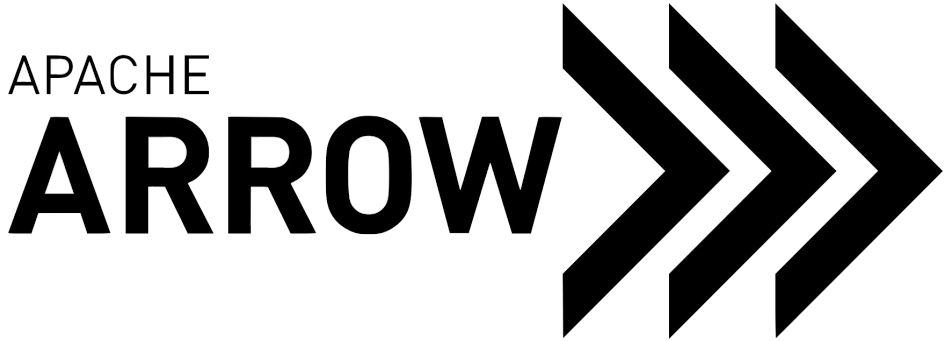


Introduction to Arrow

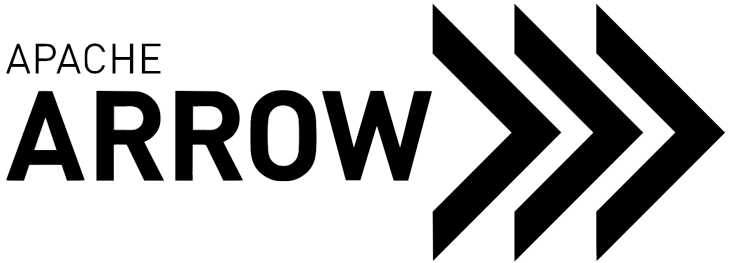
RLadies Johannesburg

November 25th, 2021

Nic Crane

What is Arrow?

- Standardised format for in-memory analytics
- Fast & scales well
- Implementations in multiple languages

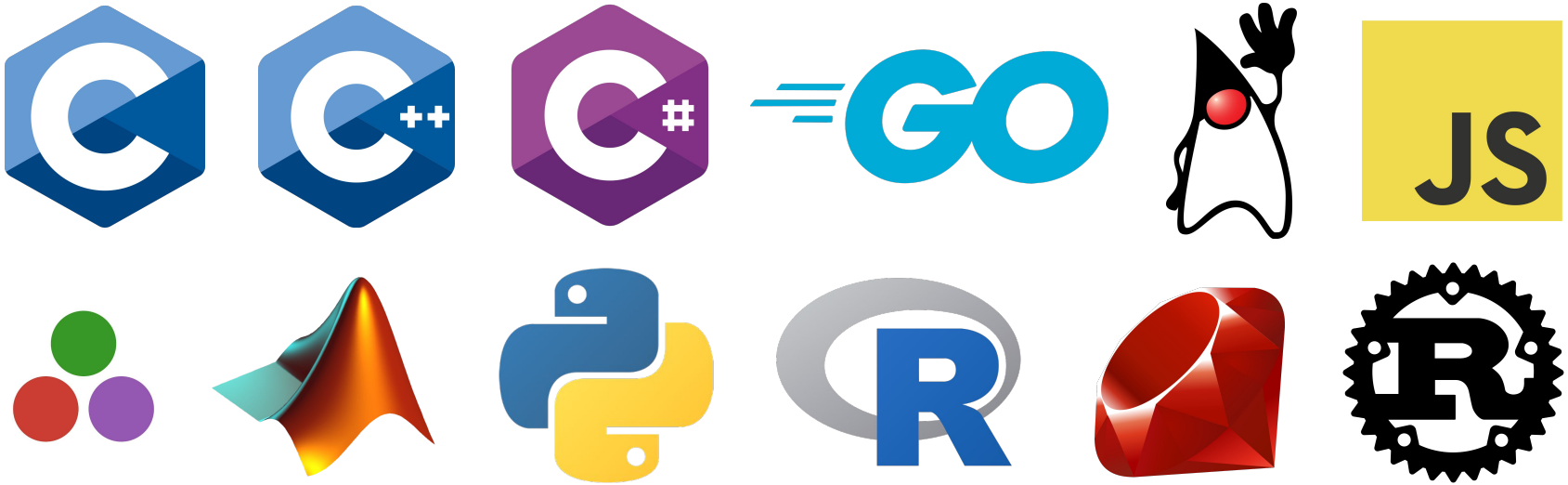


For a more detailed overview, see: https://arrow.apache.org/overview/

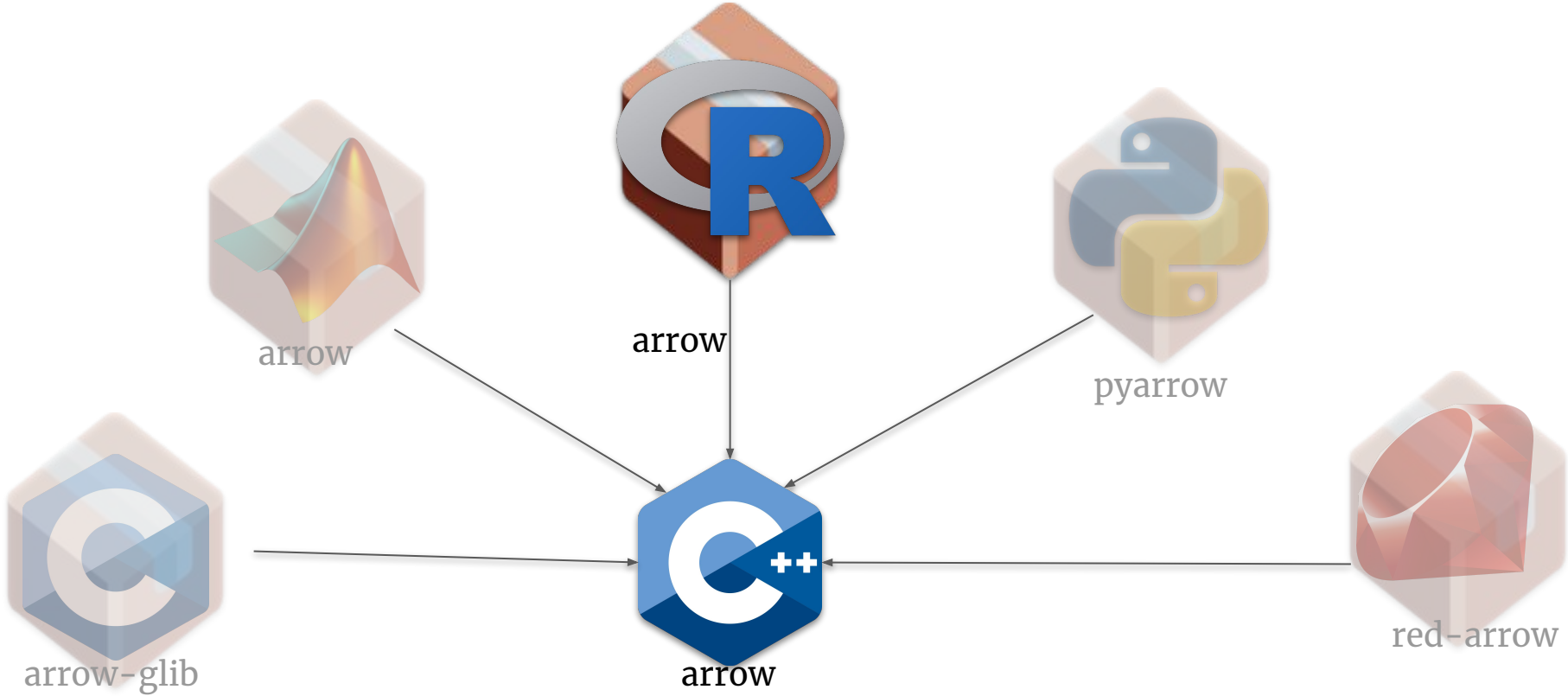
What can I do in R with Arrow?

- **analyse larger-than-memory datasets without the need to maintain a database**
- **query datasets using dplyr syntax and functions from base R and the tidyverse via Arrow's compute functions**
- work with data stored in S3 buckets without having to download it
- work with multi-file datasets without having to first import and combine them in-memory
- have super-efficient interoperability with other Arrow-based things
- have tighter control over column types
- more

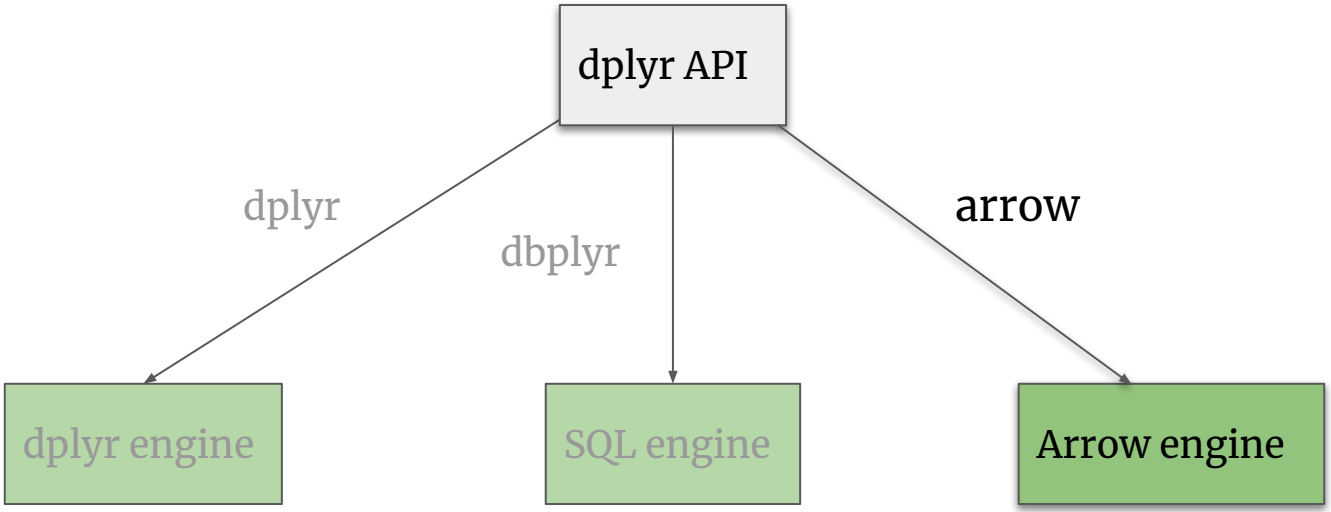
Apache Arrow



How does this work for the R package?



How does this work for the R package?



Arrow compute functions

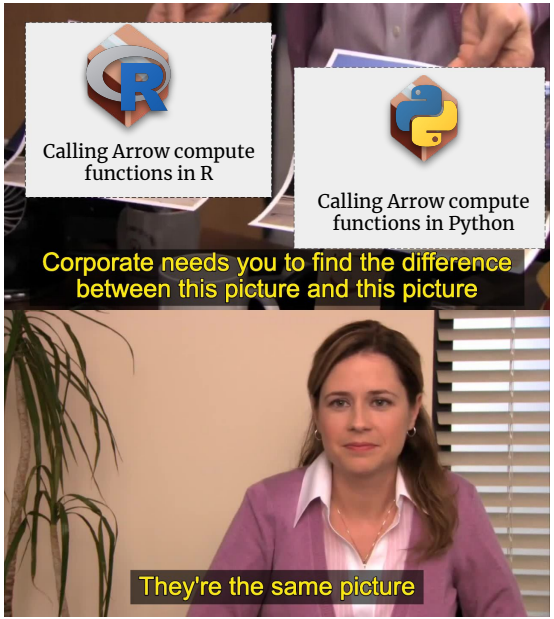
- Written in C++
- Bindings between many Arrow compute functions and their base R or tidyverse equivalent
- Same arguments, expect same outputs
- More being added all the time but open an issue if there's a specific one you would find useful

Demo 1: Work with huge datasets in-memory

https://github.com/thisisnic/intro_to_arrow/blob/main/01_arrow_scales.R

Demo 2: Use familiar tidyverse and base R syntax

https://github.com/thisisnic/intro_to_arrow/blob/main/02_arrow_dplyr.R



What's new?

In v 6.0.0:

- group_by + summarise
- joins
- lots of new string parsing functions
- date extraction functions
- DuckDB integration

What's next?

Currently looking at:

- supporting more function bindings
- expanding and updating R documentation and tutorials
- more resources for new contributors
- performance and efficiency improvements
- window functions
- more!

Resources

- {pkgdown} site - https://arrow.apache.org/docs/r/
- Apache Arrow R Cookbook - https://arrow.apache.org/cookbook/r/
- JIRA issue tracker - https://issues.apache.org/jira/projects/ARROW/

Want to know more?

For a higher-level overview of the project, technical components, history of the project, and its future directions:

Wes McKinney - New Directions for Apache Arrow

https://www.youtube.com/watch?v=u7DecbDw3QE

For more on why Arrow is so fast:

Neal Richardson - Bigger Data With Ease Using Apache Arrow

https://www.youtube.com/watch?v=zND-Wj2XPvc

For the data engineering perspective:

Ian Cook - Apache Arrow Enabling Data Engineering Tasks in R

https://www.youtube.com/watch?v=SXbq4OYtsFA

Q & A

Nic Crane

@nic_crane