

fMRIPrep-next: Preprocessing as a fit-transform model

Christopher J Markiewicz¹, Mathias Goncalves¹, John Kruper², Ma Feilong³, Guarav H Patel⁴, Juan Sanchez-Pena⁴, Robert E Smith⁵, Lea Waller⁶, Joseph Wexler¹, Russell Poldrack¹, Oscar Esteban⁷

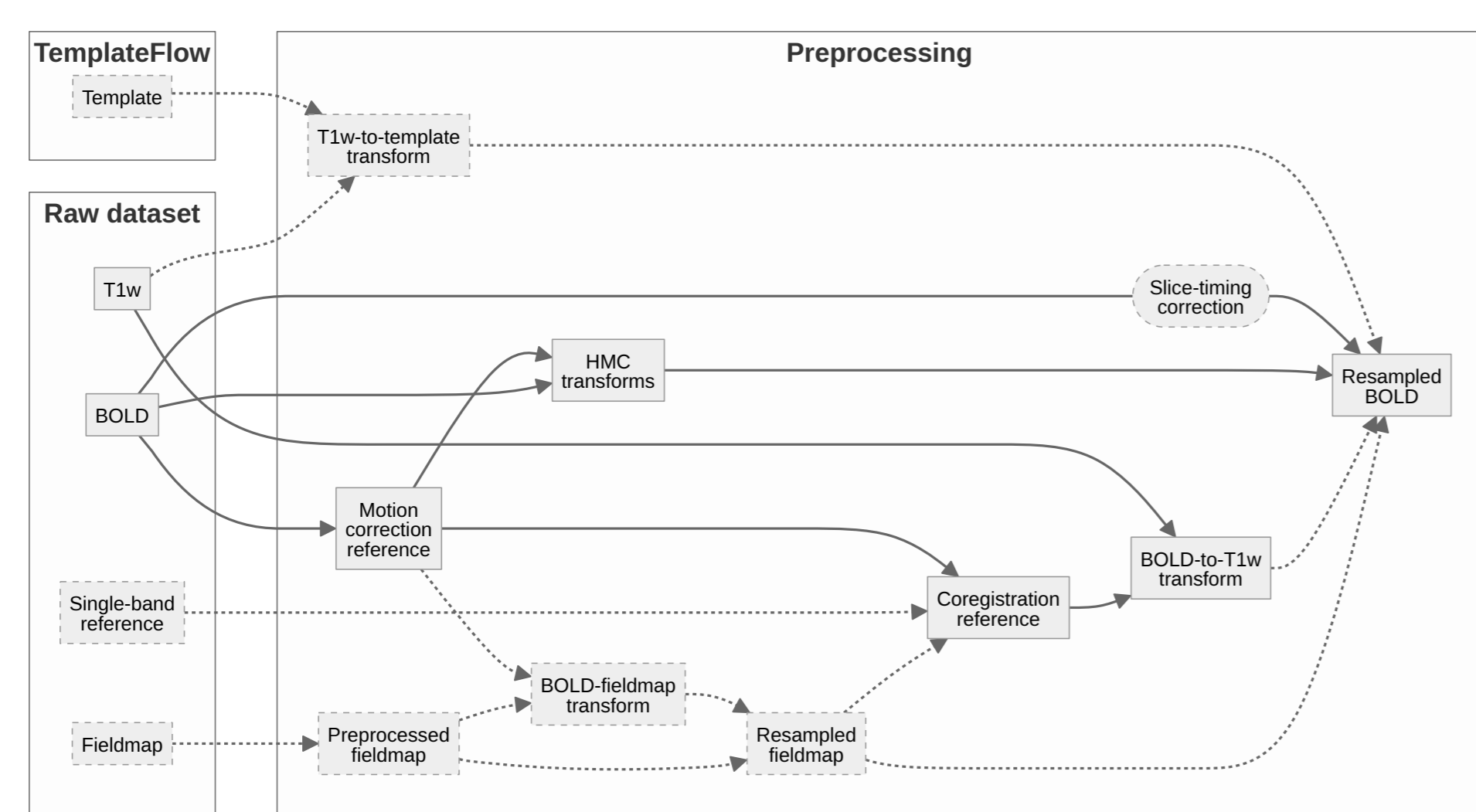
¹Stanford University, Stanford, CA, ²University of Washington, Seattle, WA, ³Dartmouth College, Hanover, NH, ⁴Columbia University, New York, NY, ⁵Florey Institute of Neuroscience and Mental Health, Melbourne, VIC, ⁶Charité Universitätsmedizin Berlin, Berlin, Berlin, ⁷Lausanne University Hospital and University of Lausanne, Lausanne, VD

fMRIPrep: <https://fmriprep.org>
NiPreps: <https://nipreps.org>

OHBM Poster #2257

End-to-end fMRI preprocessing pipelines are best suited to *small workloads*

- ▶ Data are preprocessed and analyzed on one system
- ▶ Subjects can be run all-at-once
- ▶ Little to no customization or modification is needed
- ▶ Derivative BOLD series can easily be stored together

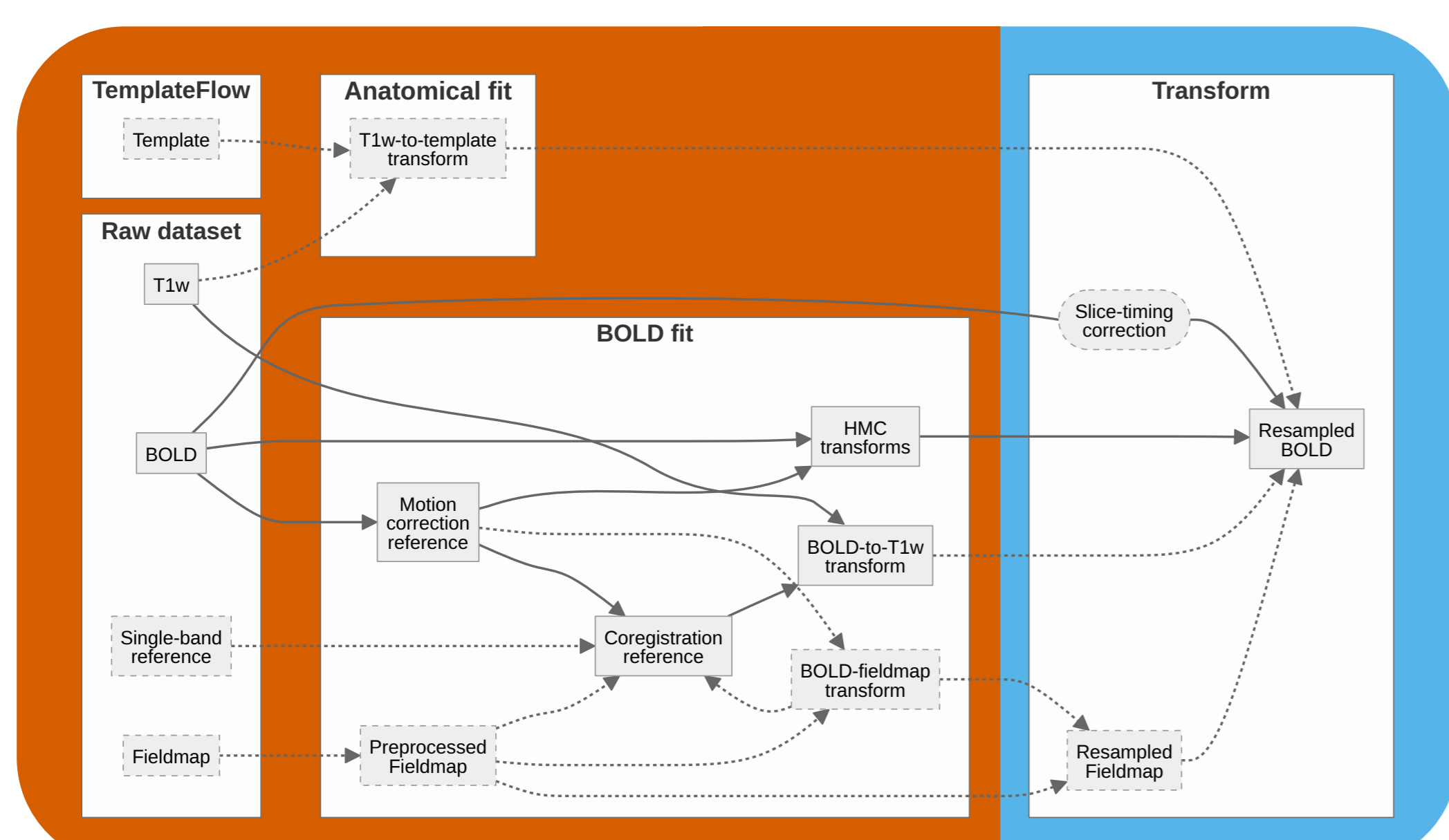


Schematic overview of the preprocessing workflow to generated a BOLD series resampled into a target space. Arrows indicate data dependencies, and dashes indicate optional inputs or steps.

When processing **large datasets**, or targeting **multiple output spaces**, generating derivative BOLD series demands **significant storage resources**.

Such workflows can be conceptually split into two parts (borrowing terms from scikit-learn):

- FIT** - Expensive estimates of required corrections
- TRANSFORM** - Cheap application of corrections



The **FIT** workflow contains most computationally intensive and all non-deterministic components of the preprocessing workflow. The **TRANSFORM** workflow generates most of the space-consuming derivatives.

We rewrote fMRIPrep to create a minimal set of derivatives, from which the output BOLD series can be generated

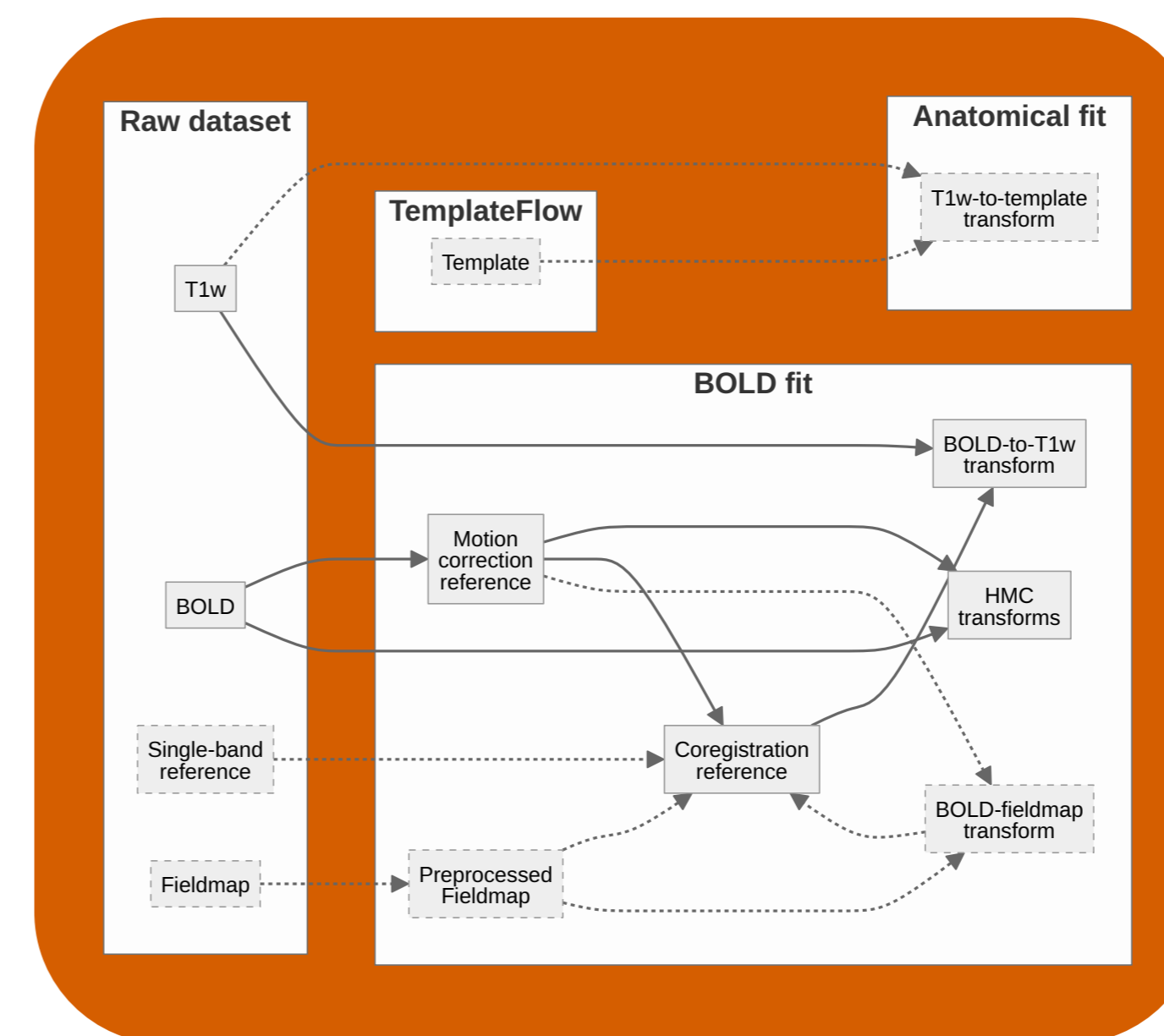
- ▶ The minimal derivatives for BOLD series include transforms, reference volumes, and fieldmaps. and the files needed to calculate them.
- ▶ Minimal derivatives are expensive to calculate, but small in storage size (e.g., head motion estimates).
- ▶ Fit-only workflows reduce runtime by >20% and storage requirements by >90% by size (>80% by file count).

Space/time comparisons:

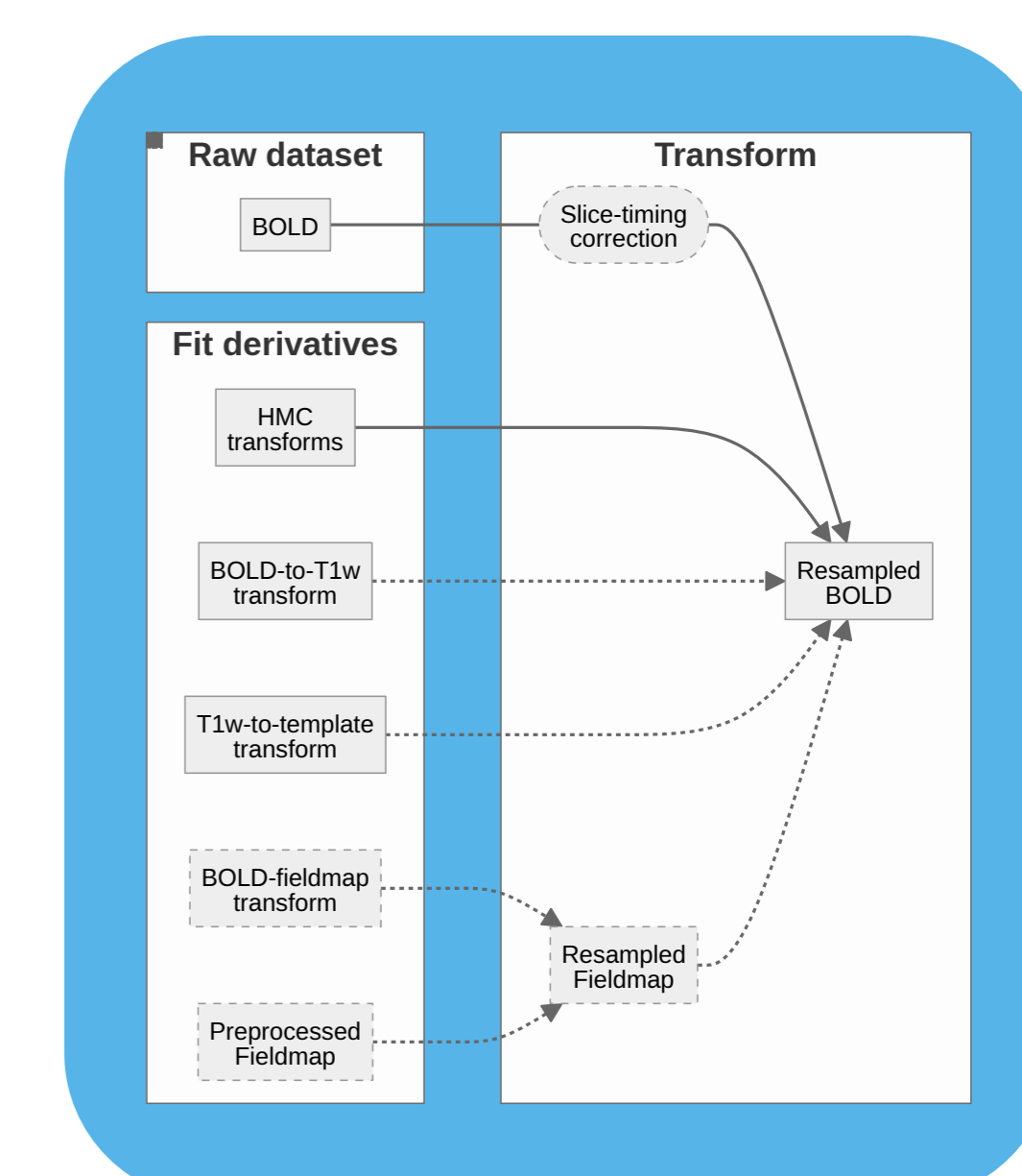
<https://fmriprep.org/en/latest/benchmarks.html>



Minimal derivatives enable new use cases



Running only the **FIT** workflow produces a compact set of derivatives. More target spaces can be selected with only a small impact on storage requirements.



Running only the **TRANSFORM** workflow avoids unnecessary recomputation. Resampling targets may be chosen for the specific analysis.

Use Case: Secondary research

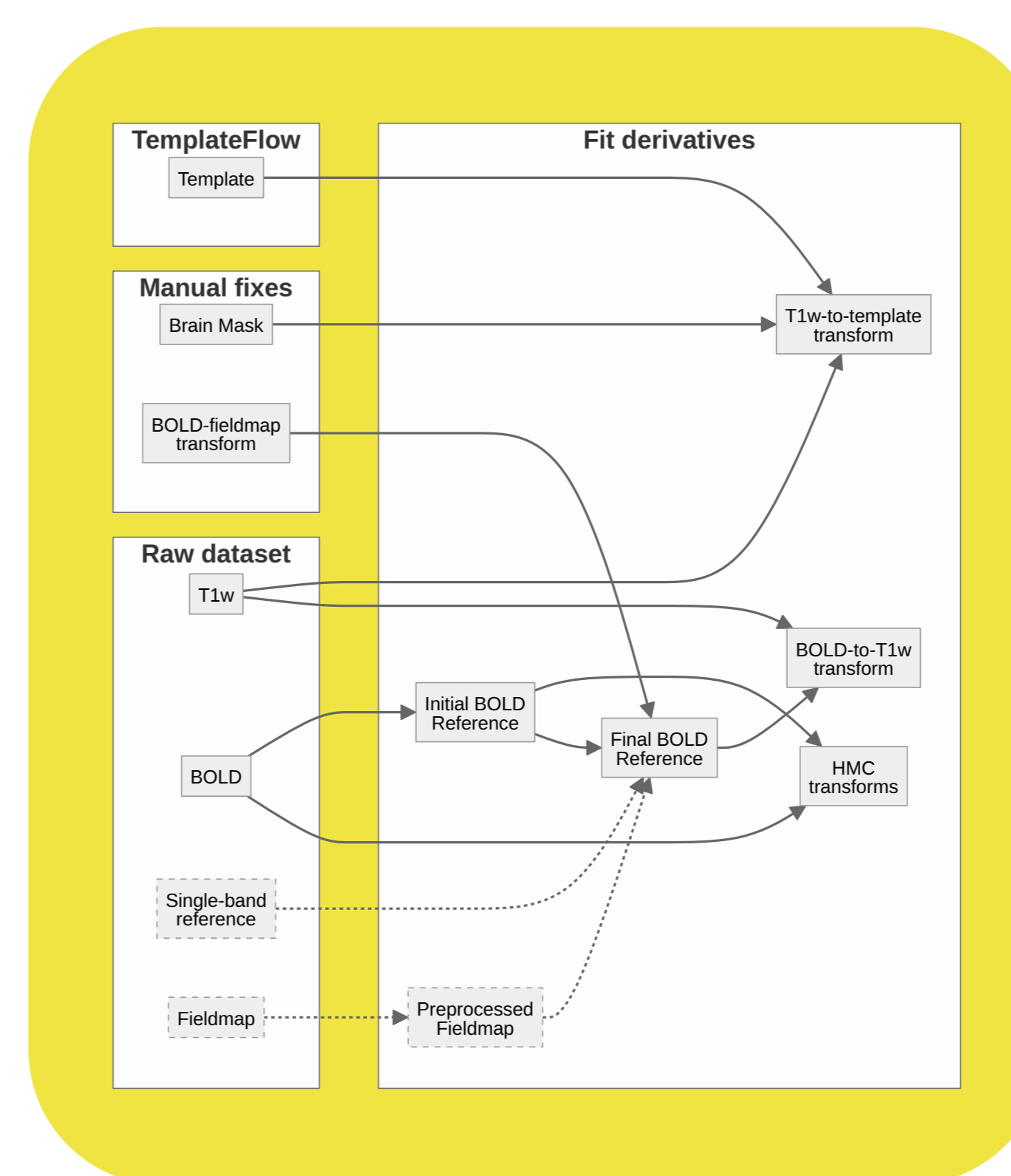
Alice publishes her raw data and the **FIT** workflow derivatives to a public archive. She provides transforms to multiple volumetric and surface templates. Bob and Carol each fetch the data and run the **TRANSFORM** workflow to reconstruct identical BOLD series in MNI space for analysis.

```
fmriprep raw/ out/ participant --level minimal \
--output-spaces ...
```

```
fmriprep raw/ out/ participant --level full \
--derivatives minimal/ --output-spaces ...
```

Use Case: Manual intervention

Dave uses an alternative brain mask and BOLD-to-fieldmap registration to improve the quality of his preprocessing. He reruns fMRIPrep with these derivatives as inputs.



```
fmriprep raw/ out/ \
participant \
--level minimal \
--derivatives fixes/ \
-output-spaces ...
```

Future directions

- ▶ Extend the fit-transform model to other nipreps tools. It has been adopted in ASLPrep already.
- ▶ fMRIPost library to generate secondary derivatives on-the-fly in downstream tools.
- ▶ Convert transform workflow into a standalone tool that is lighter than the fMRIPrep container.

Acknowledgements

NiPreps development, including fMRIPrep supported by the BRAIN Initiative award 1RF1MH121867-01A1. This research is part of the Frontera computing project at the Texas Advanced Computing Center. Frontera is made possible by National Science Foundation award OAC-1818253.