Banff’s next step: an open-source data editing system for advanced tools and collaboration

Delivering insight through data for a better Canada

UNECE Expert Meeting on
Statistical Data Editing

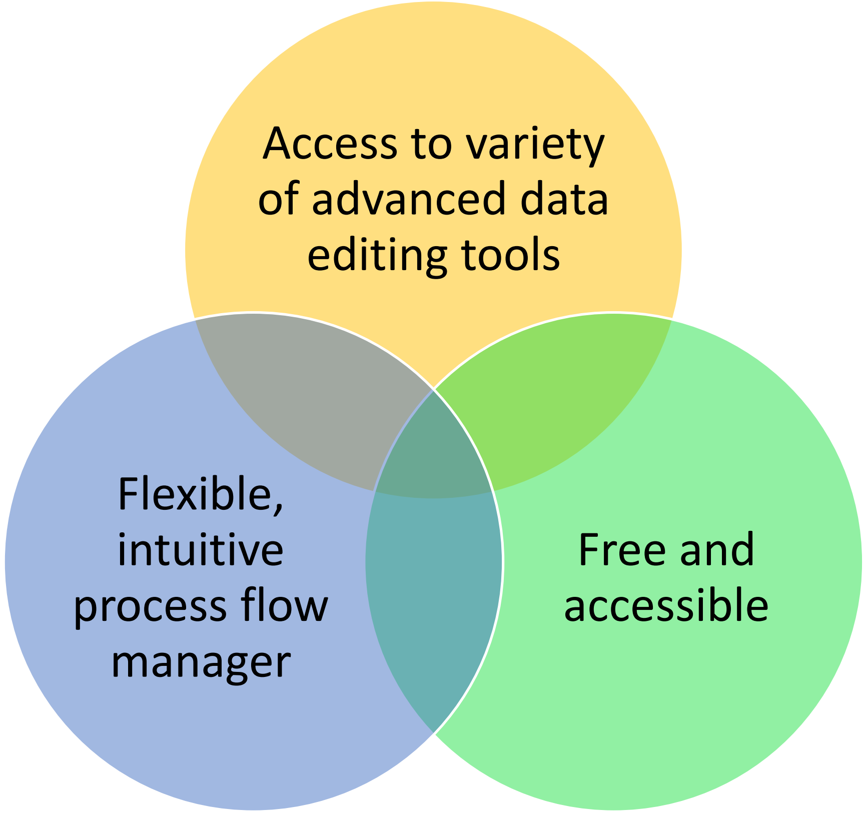
October 3-7, 2022

Darren Gray

Statistics Canada

What’s our goal?

A system to design and process production-scale data editing, with an expanding catalogue of expert-vetted, community-supported tools, accessible to everyone.


Access to variety of advanced data editing tools

Flexible, intuitive process flow manager

Free and accessible

What is Banff?

- Generalized edit and imputation system developed and maintained by Statistics Canada
- Features nine data editing procedures performing various tasks, including outlier detection, error localization, and donor imputation
- Includes Banff Processor: metadata-driven process flow manager
- Currently runs on SAS architecture

What are the plans?

- Major changes:
  - Free, standalone software package (remove SAS dependency)
  - New criterion for standardized Banff modules
  - Completely redesigned Banff processor

- Release: Spring 2024
  - Banff team to support users as they migrate to new system
  - Begin integration of new data editing tools

Banff Modules

Objectives

1. Remove SAS dependency while maintaining all current functionality
2. Eliminate as much user data management as possible
3. Facilitate the integration of external tools

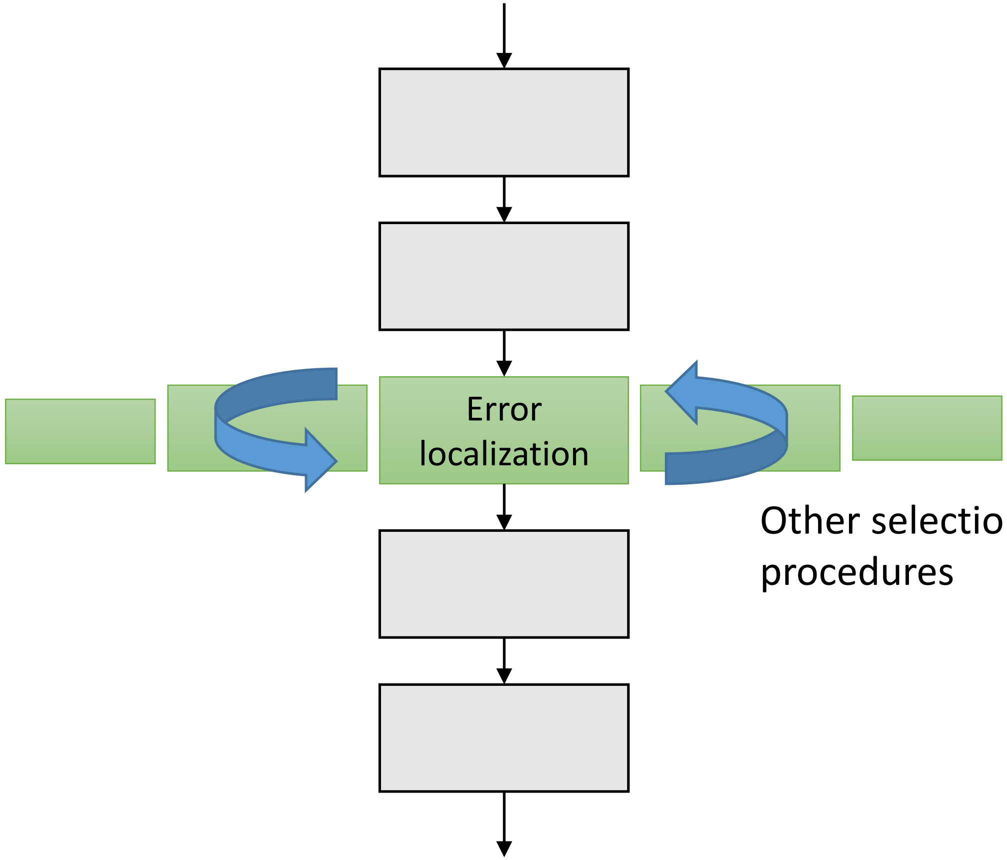
Module standardization

What do we need?

- Modules run in sequence shouldn’t require intermediate data management
- Modules performing similar tasks should be interchangeable

Integration of new modules requires clear instructions on how the modules should interact!

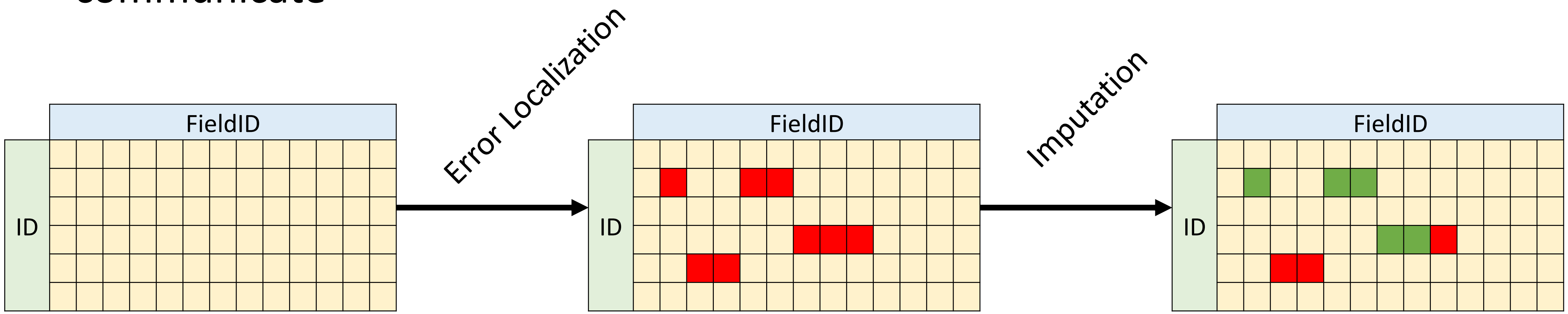
Statistical data editing process flow



Other selection procedures

Modular interaction

- Outputs from one module -> inputs for subsequent ones
- Two key inputs/outputs: statistical data and the Banff status file
- Banff status file contains key metadata (status flags) that modules use to communicate



Banff status flags

- Status flags serve two important roles in process flow:
  - Used in subsequent process steps
  - As an audit trail

- Encode relevant metadata at three levels:
  - Field level
  - Record level
  - Process level

New!

Banff status flags: examples

- Selection flags: records or fields that require specific treatment
- Exclusion flags: records or fields that should be excluded from certain steps
- Imputation flags: records or fields successfully imputed
- Warning flags: process steps that fail to run successfully

Goal: standardize as much metadata as possible

Can I create my own modules?

Yes!

This is a key objective of the standardization project – to make it as easy as possible to adapt / modify / wrap external tools into the Banff system. The Banff team will provide guidance and tools for this purpose.

Banff Processor

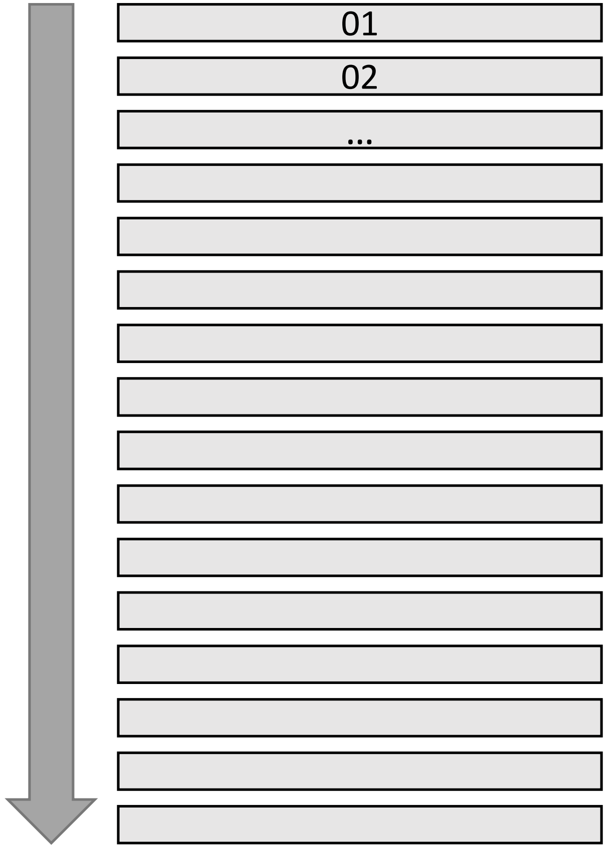
Objectives

1. Remove SAS dependency
2. Simplify the design stage, while improving functionality
3. Performance efficiency

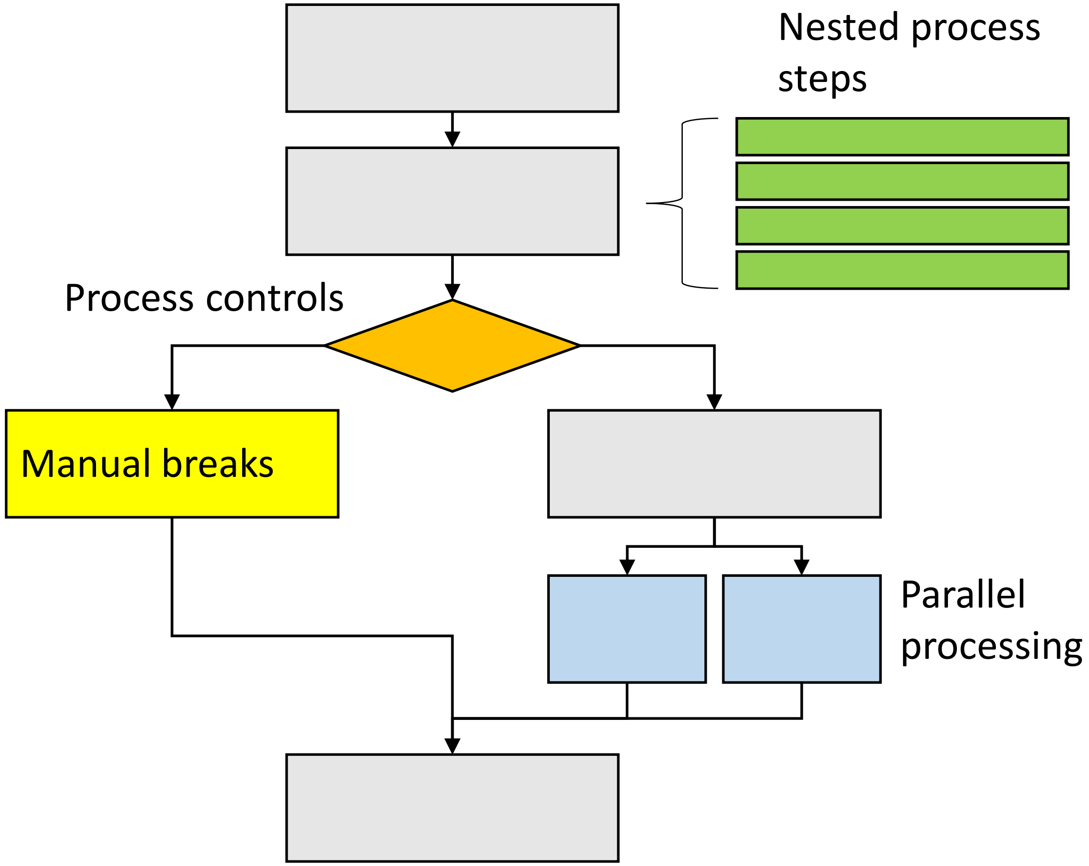
Highlights of redesign

- Point-and-click interface – likely a web application
- Access to built-in Banff modules, plus custom user modules
- New features focused on process flow design, convenience and efficiency

Linear process flow
(Current processor)

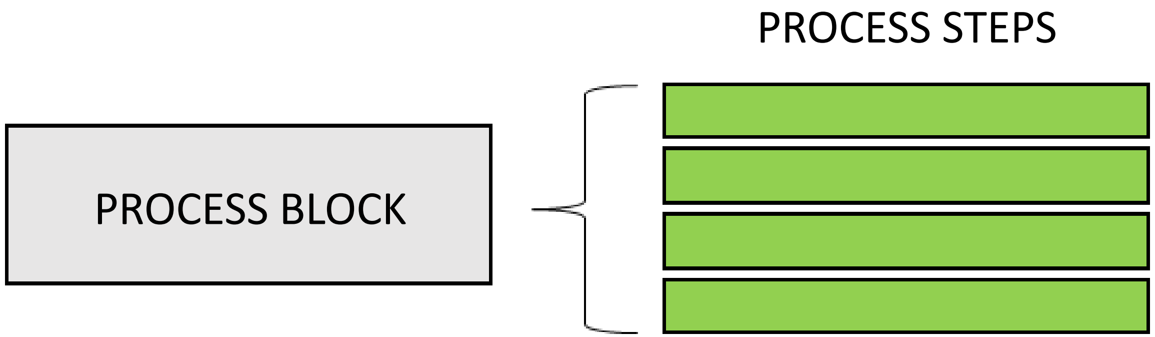


Complex process flow
(Redesign)



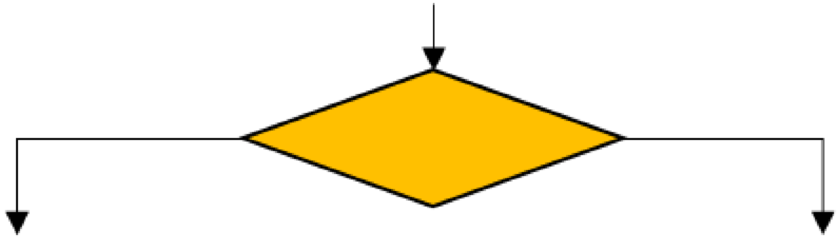
Nested process steps

- Allow users to group multiple steps into a single process block
- Multiple levels permitted
- Any specifications applied to the process block affects all elements
- Simplifies design and reduces repetition



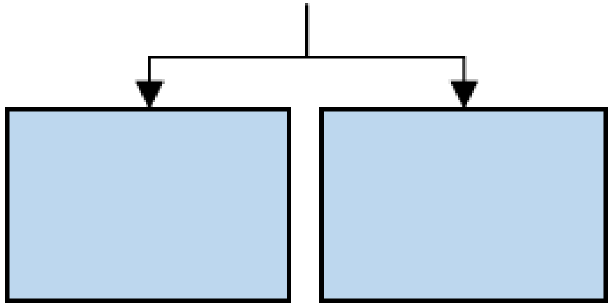
Process controls

- Introduce process controls to determine:
  - When a process step is executed
  - Which statistical data should be processed
  - Which metadata should be processed
- Gives users high-level control over process flow



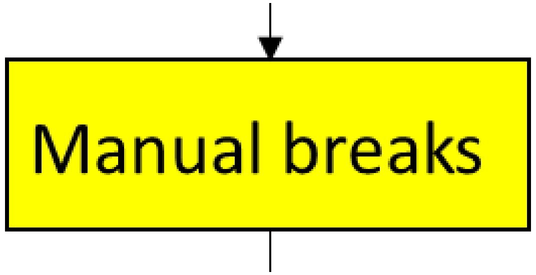
Parallel processing

- Take advantage of any available parallel processing infrastructure to improve performance efficiency



Manual breaks

- Trigger a pause in processing until outputs of manual process step are loaded into system



Conclusion

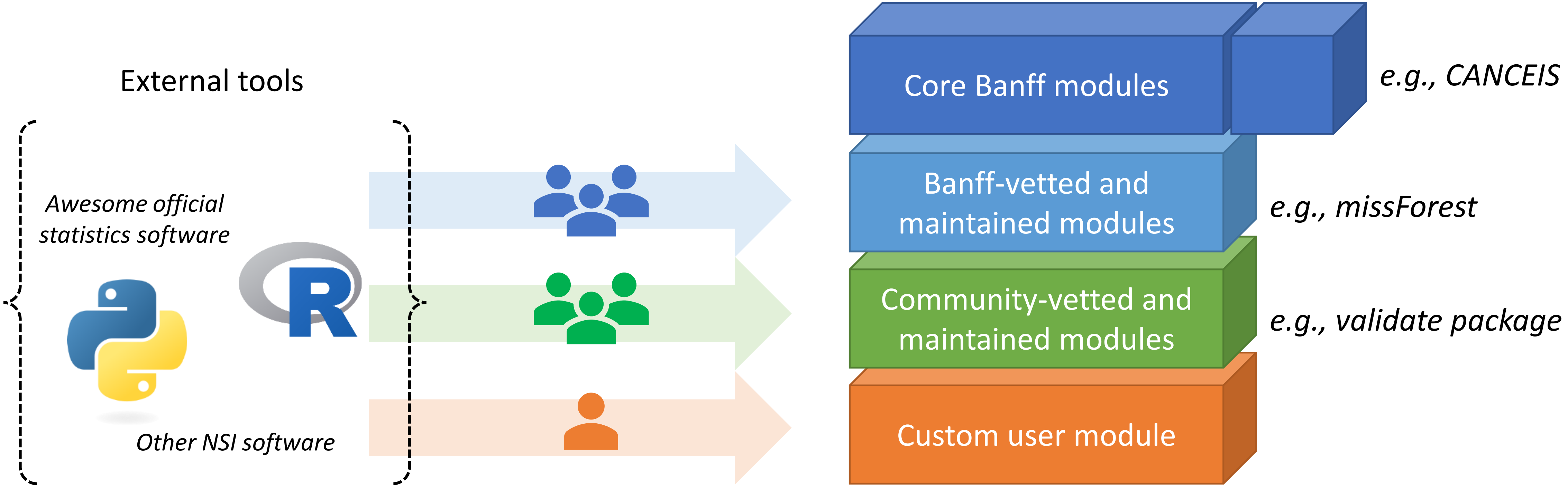
Summary of changes

- Target release: Spring 2024
- Built on open-source technologies
- Updated Banff modules:
  - Streamlined and modular
  - Expanded scope of status flags
- Completely redesigned Banff Processor

Less time spent coding, more time spent on design and testing

- Lay foundation for new opportunities in research, innovation, and collaboration, through the integration of external methods and tools

Building a catalogue of Banff modules



External tools

Awesome official statistics software

Other NSI software

Core Banff modules

e.g., CANCEIS

Banff-vetted and maintained modules

e.g., missForest

Community-vetted and maintained modules

e.g., validate package

Custom user module

THANK YOU!

darren.gray@statcan.gc.ca

banff@statcan.gc.ca