MANCHESTER 1824
The University of Manchester

Automatic vs Manual Provenance Abstractions: Mind the Gap

Pınar Alper
TAPP
9 June 2016

Pınar ALPER
University of Manchester

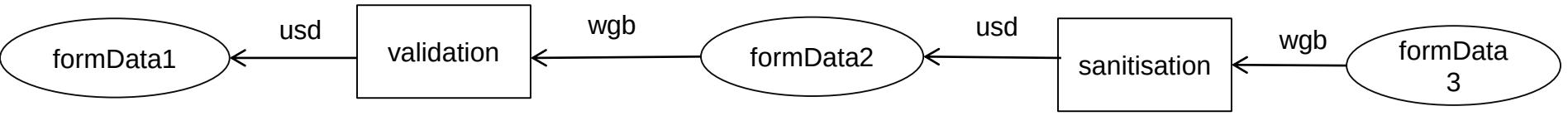
Khalid BELHAJJAME
Université Paris Dauphine

Carole A. GOBLE
University of Manchester

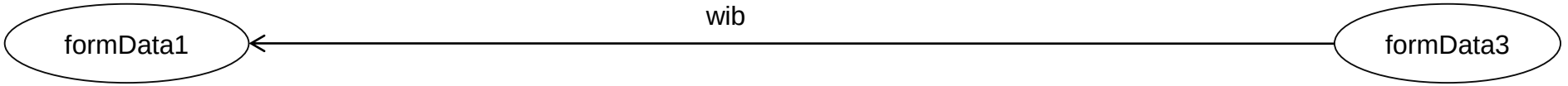
MANCHESTER 1824
The University of Manchester

Transparency brought by Provenance can be a double edged sword

Too revealing



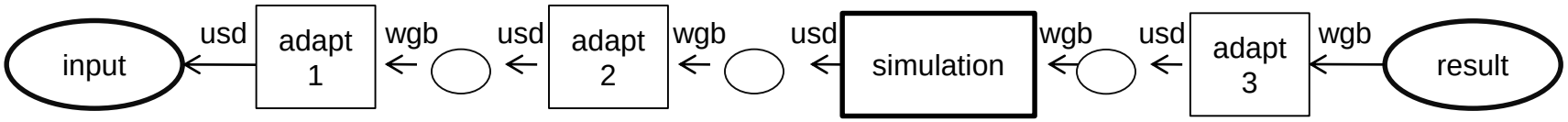
Secured Provenance



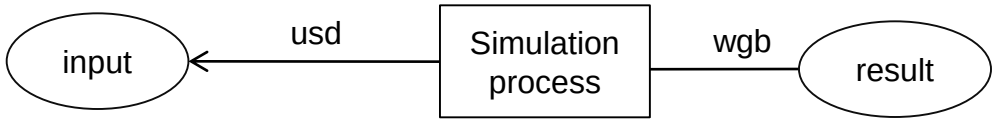
MANCHESTER 1824
The University of Manchester

Transparency brought by Provenance can be a double edged sword

Too complex



Simplified Provenance

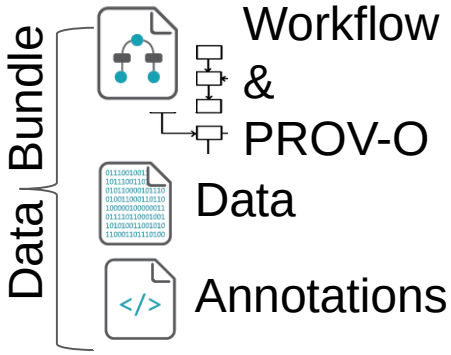


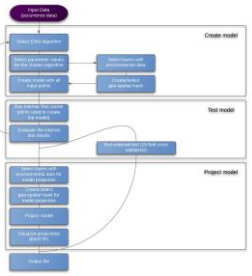
Our Goal & Context

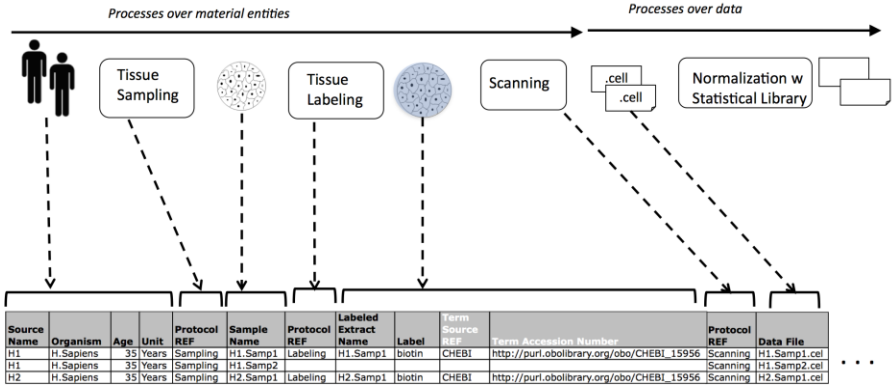
Compare Manual and Semi-automated abstractions that simplify workflows in the context of reporting data-oriented experiments.

→ – Reports are experimental metadata on:

- Method
- Data







Source Name	Organism	Age	Unit	Protocol REF	Sample Name	Protocol REF	Labeled Extract Name	Label	Term Source REF	Term Accession Number	Protocol REF	Data File
H1	H.Sapiens	35	Years	Sampling	H1.Samp1	Labeling	H1.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H1.Samp1.cel
H1	H.Sapiens	35	Years	Sampling	H1.Samp2						Scanning	H1.Samp2.cel
H2	H.Sapiens	35	Years	Sampling	H2.Samp1	Labeling	H2.Samp1	biotin	CHEBI	http://purl.obolibrary.org/obo/CHEBI_15956	Scanning	H2.Samp1.cel

"Best Practices for Workflow Design: How to Prevent Workflow Decay." In Proceedings of SWAT4LS, November 2012.

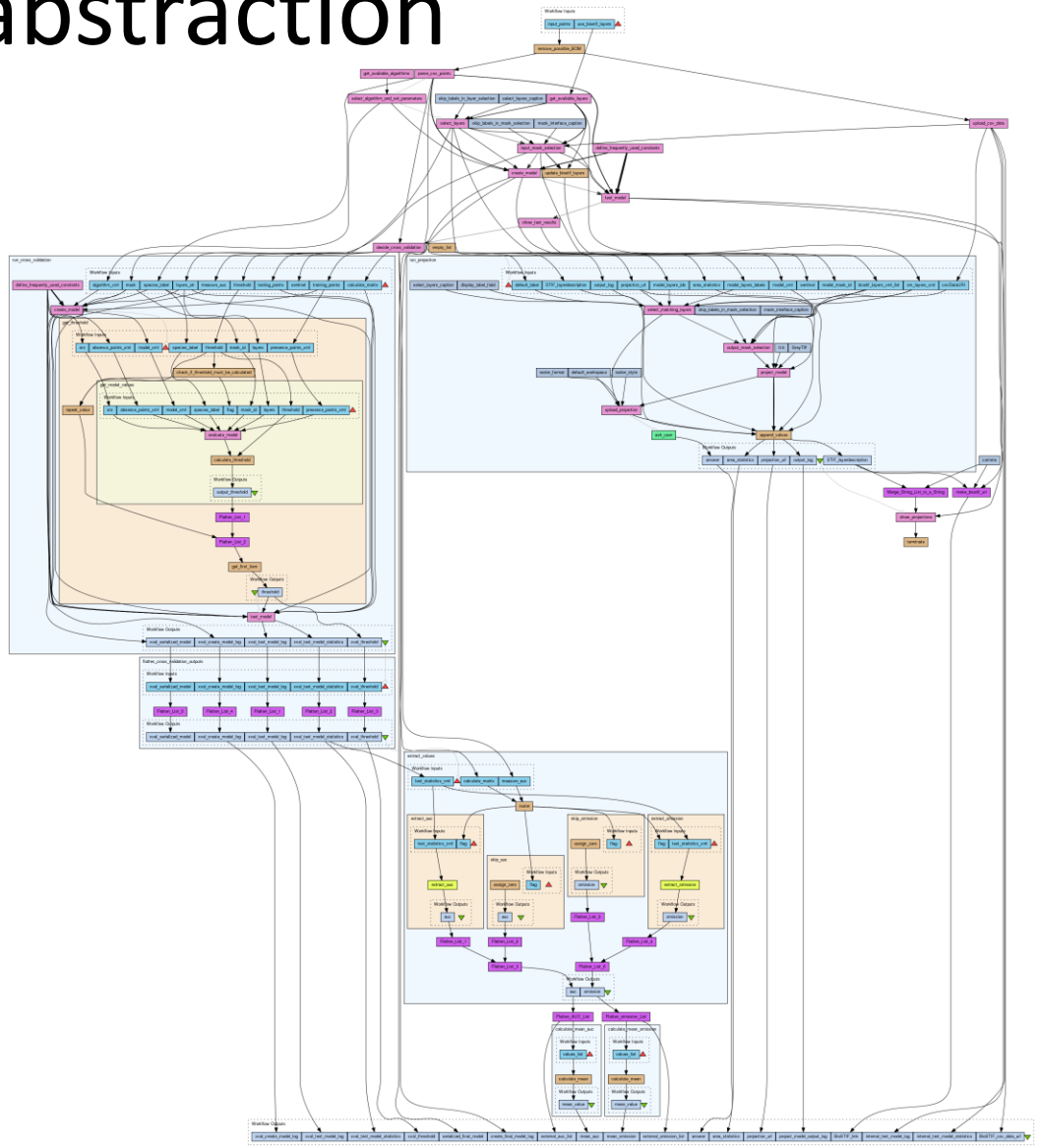
"Toward interoperable bioscience data." Nature Genetics, 44(2):121–126, February 2012.

"Scientific Data", Open-Access Journal. Nature Publishing Group, 2015, http://scientificdata.isa-explorer.org

MANCHESTER 1824
The University of Manchester

Workflow complexity necessitates abstraction

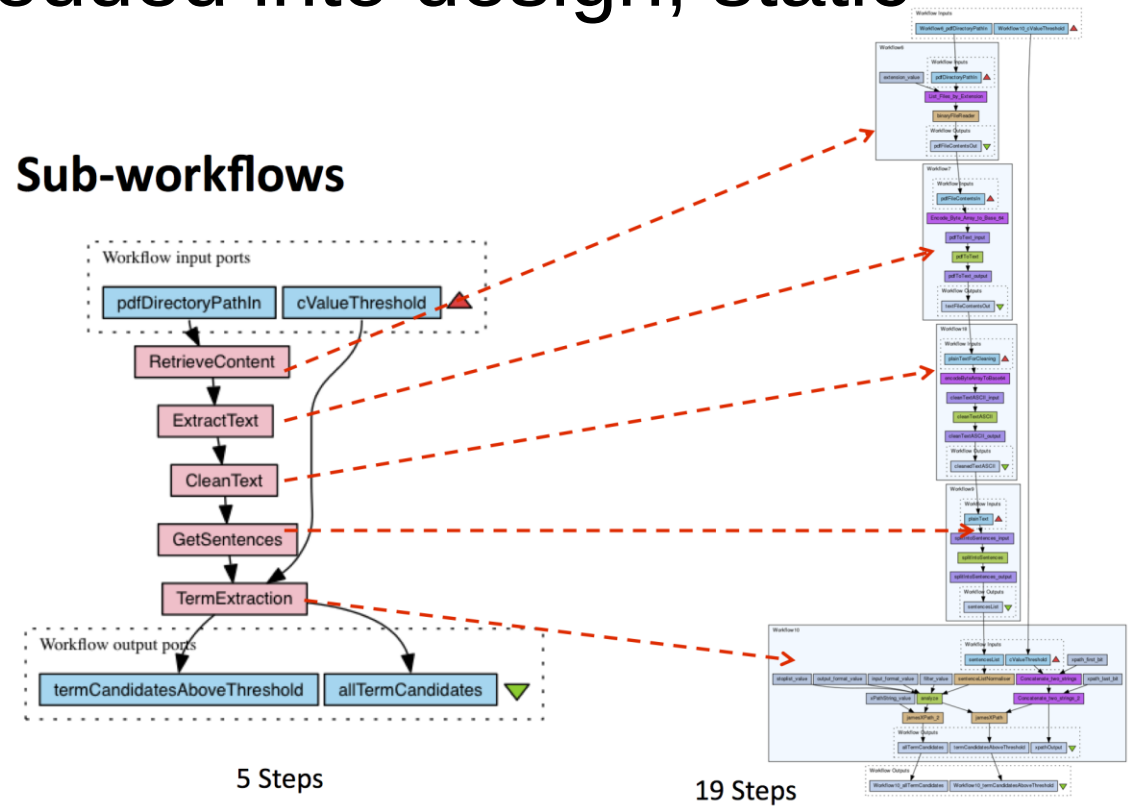
- Up to 50+ data processing tasks.
- Majority (70 %) dedicated to data adaptation
- Leads to complex provenance.



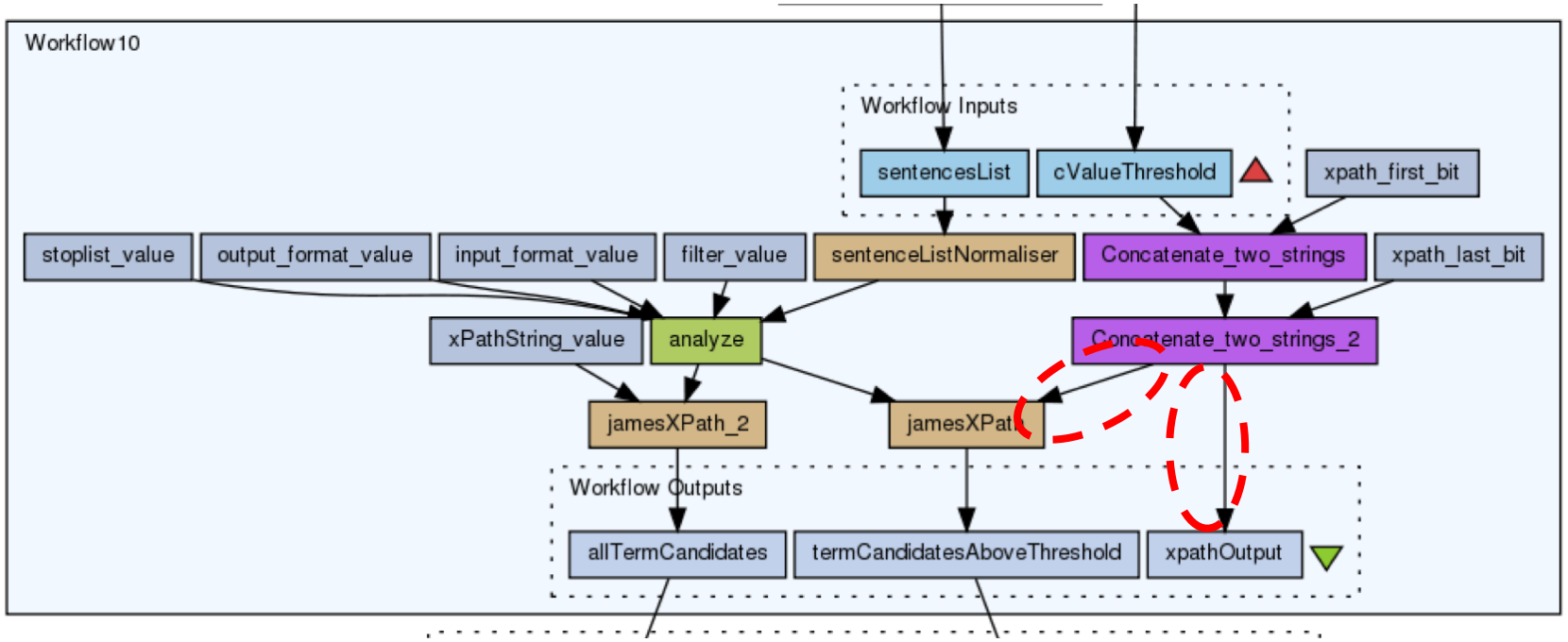
"Common Motifs in Scientific Workflows: An empirical analysis." Future Generation Comp. Sys. 36: 338-351 (2014).

Current Approach

- Manual, Design Abstractions observable in existing workflows myexperiment
- Embedded into design, static



Design Abstractions

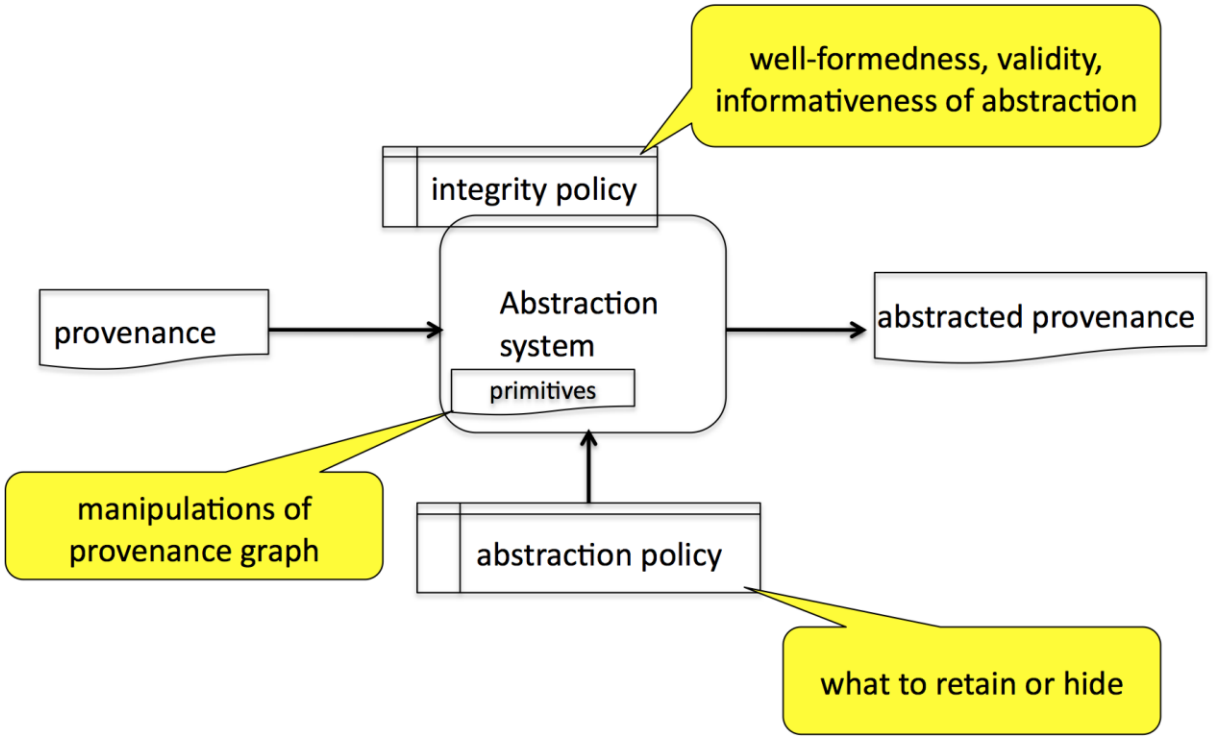


Bookmarked Intermediaries

Alternative Approach

Semi-automated abstraction systems:

- dynamic
- Workflow Summaries & ZOOM UserViews



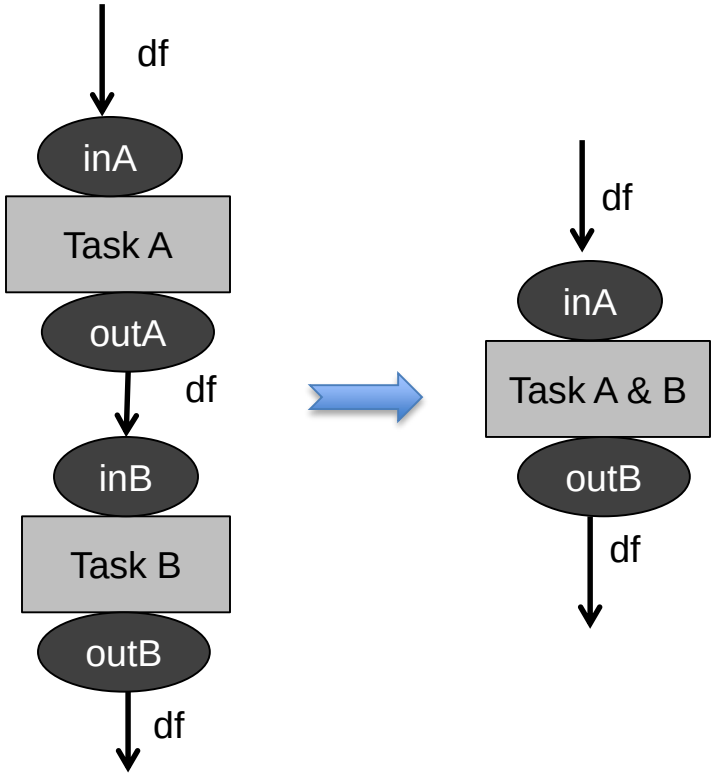
- ProvAbs, IPAW 2014
- Propub, SSDBM 2011
- TACLP, FGCS 2015
- SecProv, WAIM 2008
- Provenance Redaction, SACMAT 2011
- Surrogates, VLDB 2011

"Small is beautiful: Summarizing scientific workflows using semantic annotations." BigData 2013, pages 318-325.
"Querying and Managing Provenance through User Views in Scientific Workflows." ICDE 2008, pages 1072–1081.

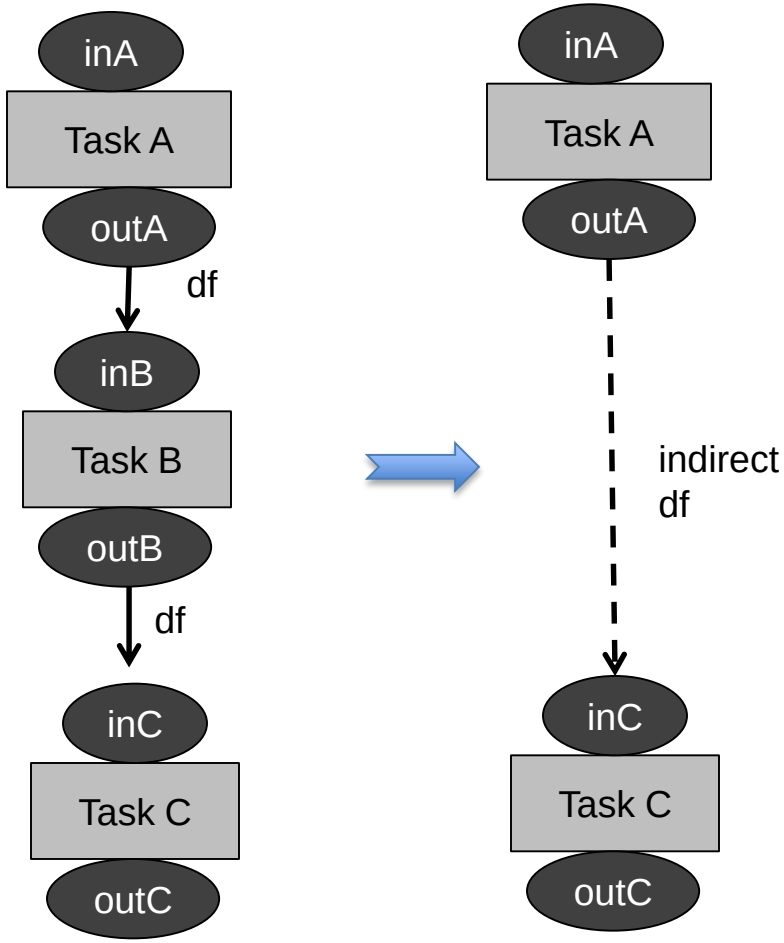
MANCHESTER 1824
The University of Manchester

Workflow Abstraction Primitives

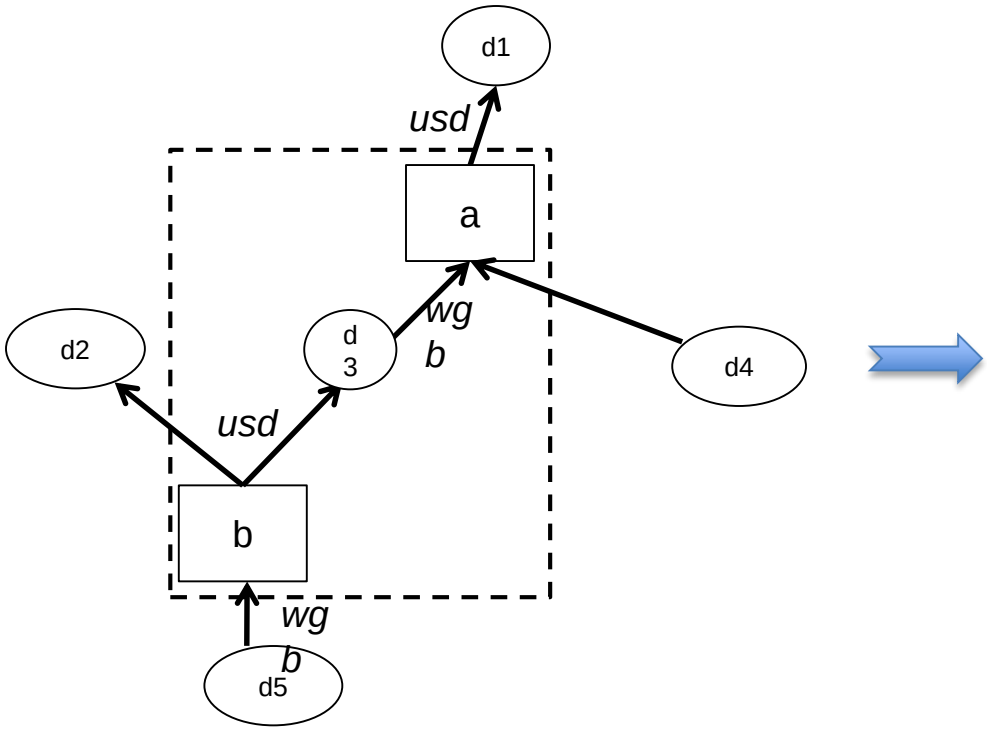
Task grouping

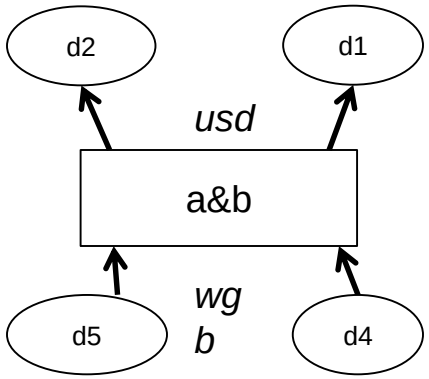


Task elimination



Integrity Policy: Soundness



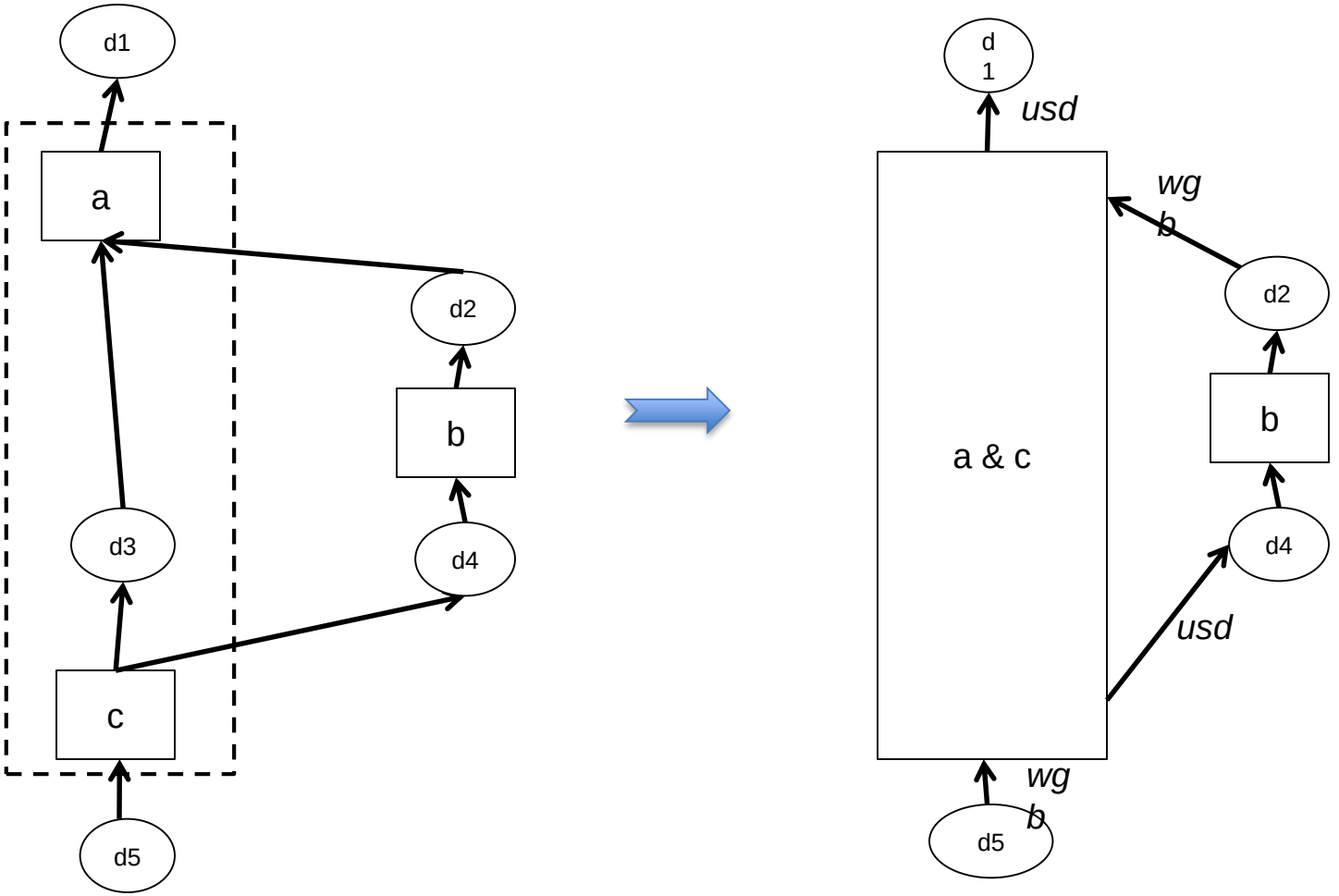


In the context of reporting soundness can be compromised.

Sub-workflow based design abstractions do not necessarily preserve soundness.

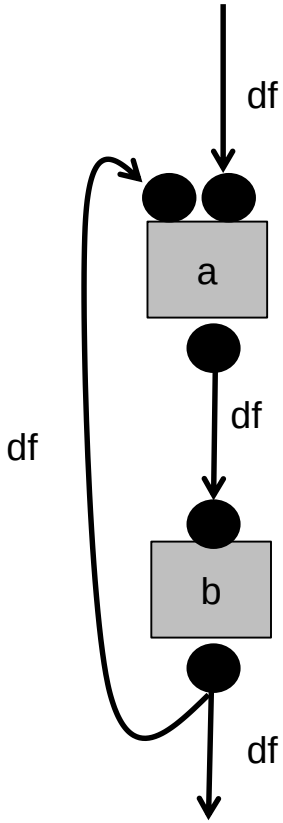


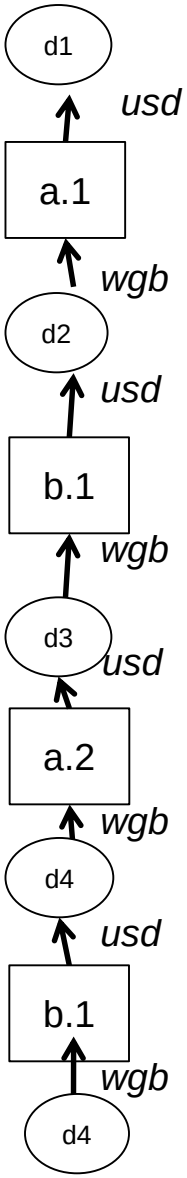
Integrity Policy: Acyclicity



From a modelling perspective cycles are allowed in provenance.

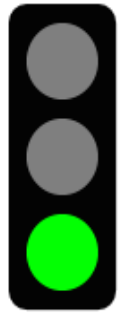
Cyclic dataflows unfold into Acyclic Lineage



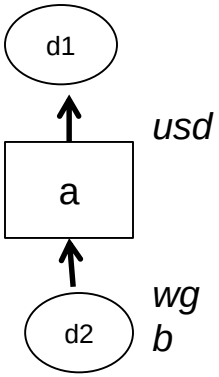


Raw workflow provenance is acyclic.

In the context of reporting (e.g. data tables, design abstractions) cycles are not observed.



Integrity Policy: Bipartiteness



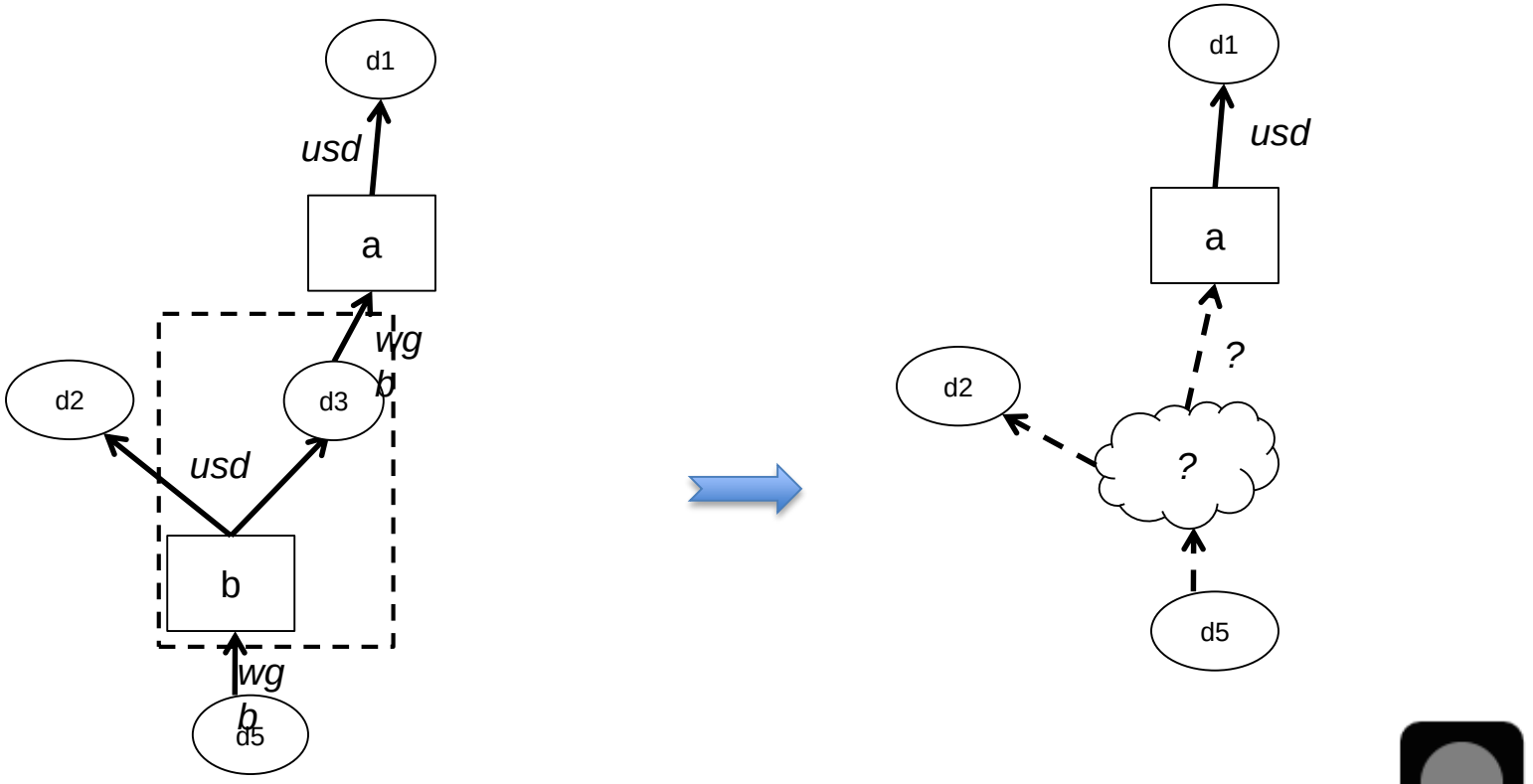
Workflow provenance is bi-partite.

Design abstractions preserve bipartiteness.

In the context of reporting bi-partiteness is advised but not always achieved.

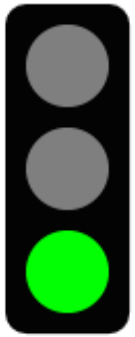


Integrity Policy: Validity

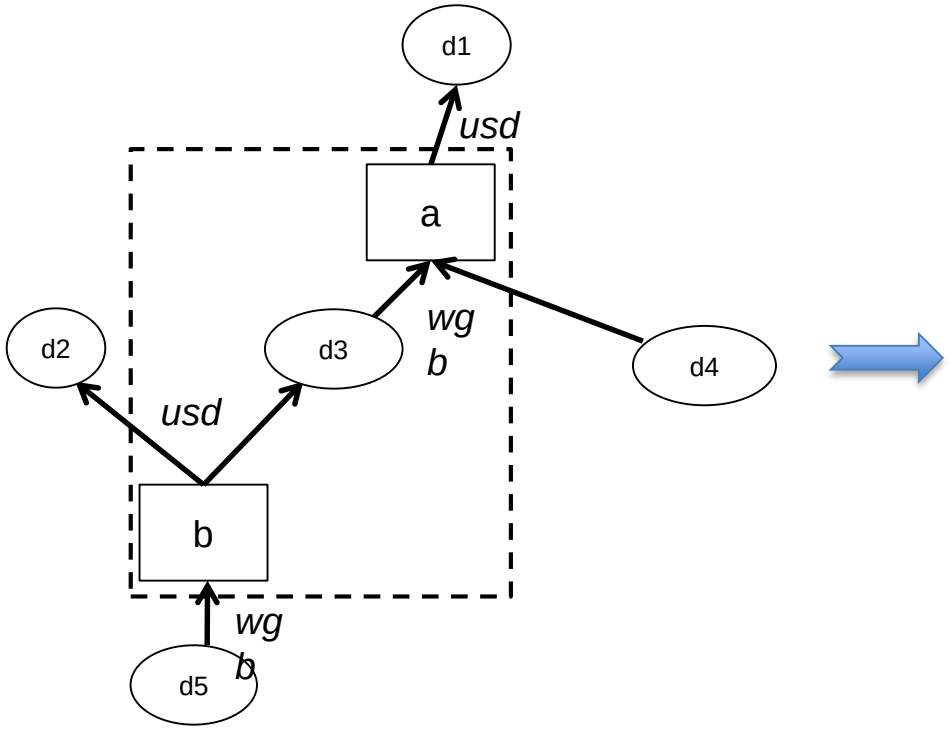


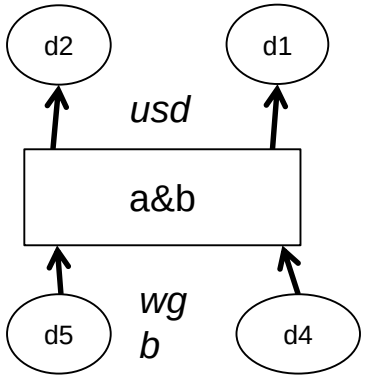
Design abstractions preserve validity.

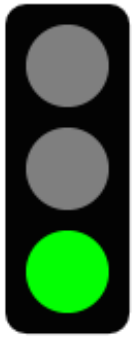
In the context of reporting validity is a necessary property.



Integrity Policy: Completeness







Determined by preservation of lineage relations

Design abstractions preserve completeness

In the context of reporting completeness is a necessary property.

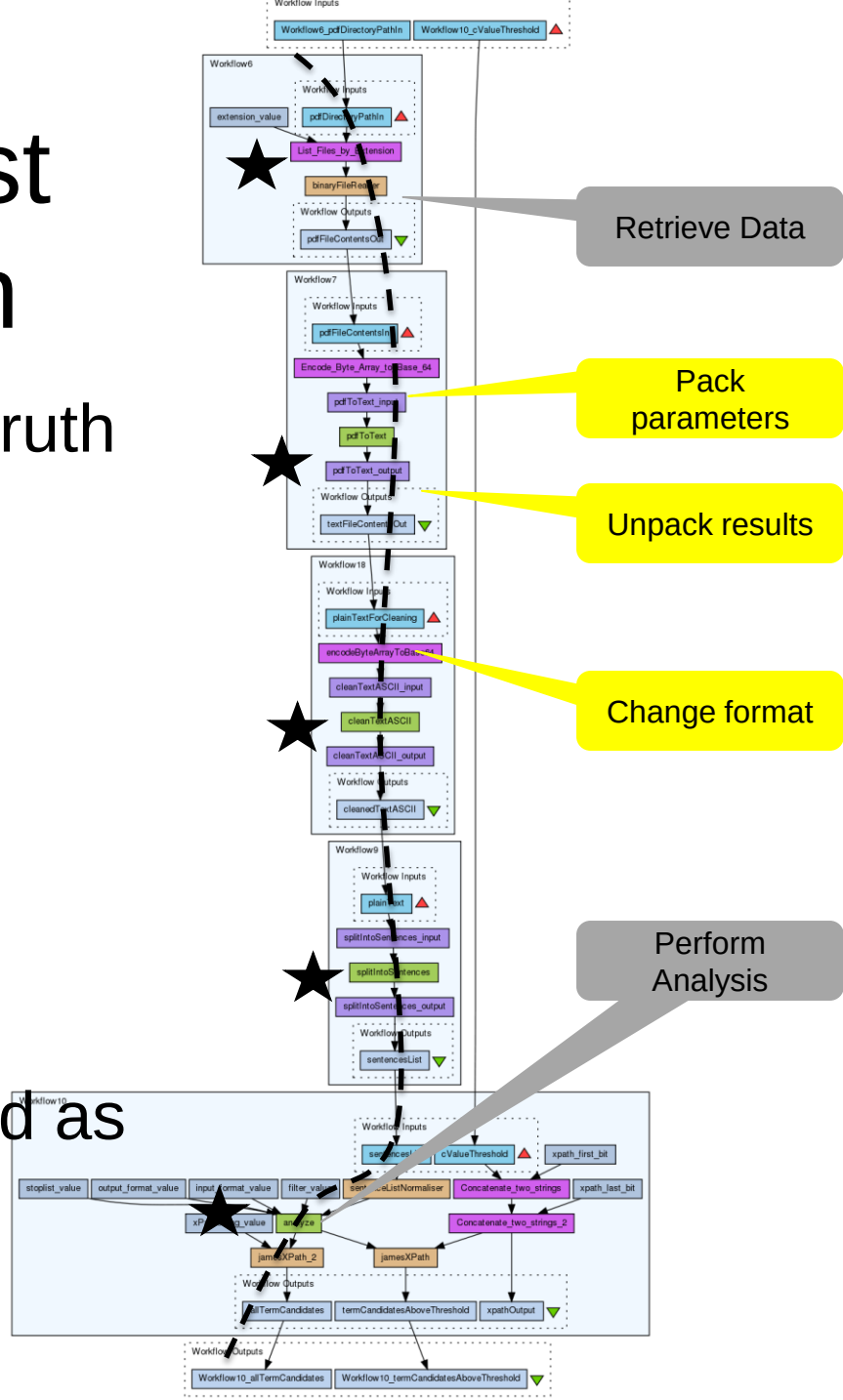
Comparison of Systems

	Workflow Summaries		ZOOM	Design Abstractions
Abstraction Policy	Annotation-Primitive pairs		Important Task Ids	--
Primitive	Grouping	Elimination	Grouping	Grouping
Integrity Policy				
Validity	✓	✓	✓	✓
Soundness		✓	✓	
Bipartiteness	✓		✓	✓
Completeness	✓	✓	✓	✓
Acyclicity	✓	✓		✓

MANCHESTER 1824
The University of Manchester

Comparison Against Design Abstraction

- Design Abstractions as groundtruth
  - Ports on main data derivation path
  - Sub-workflow tasks.
- Workflow Summaries
  - Eliminate All Adapters
  - Collapse All Adapters
  - Collapse
- ZOOM
  - Non-Adapter tasks designated as significant



Task Elimination

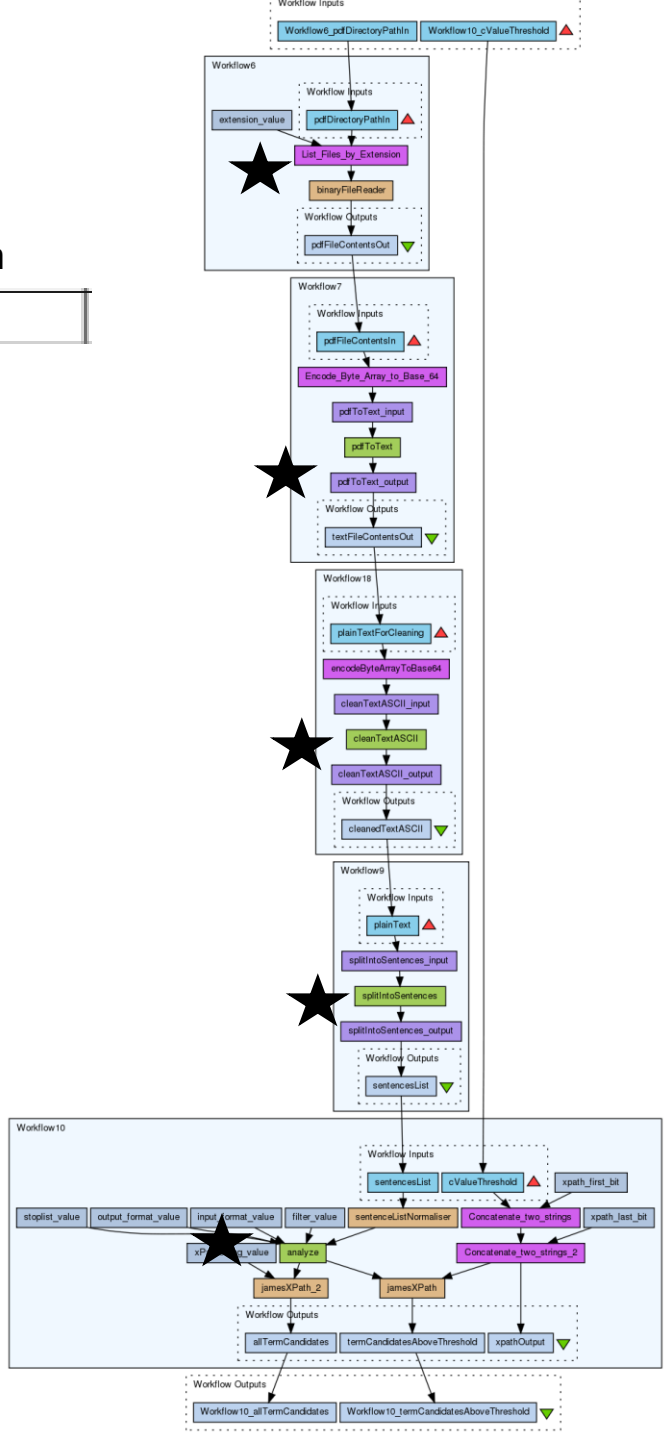
	Process Precision	Data Precision
Wf Summaries -Eliminate All	5/5	0/9

# of processes in the abstracted account overlapping w user's abstraction

total # of processes in the abstracted account

Hopping over traces does not simplify the account data –wise as as it does process-wise

A significant task and its report worthy output are not necessarily co-located



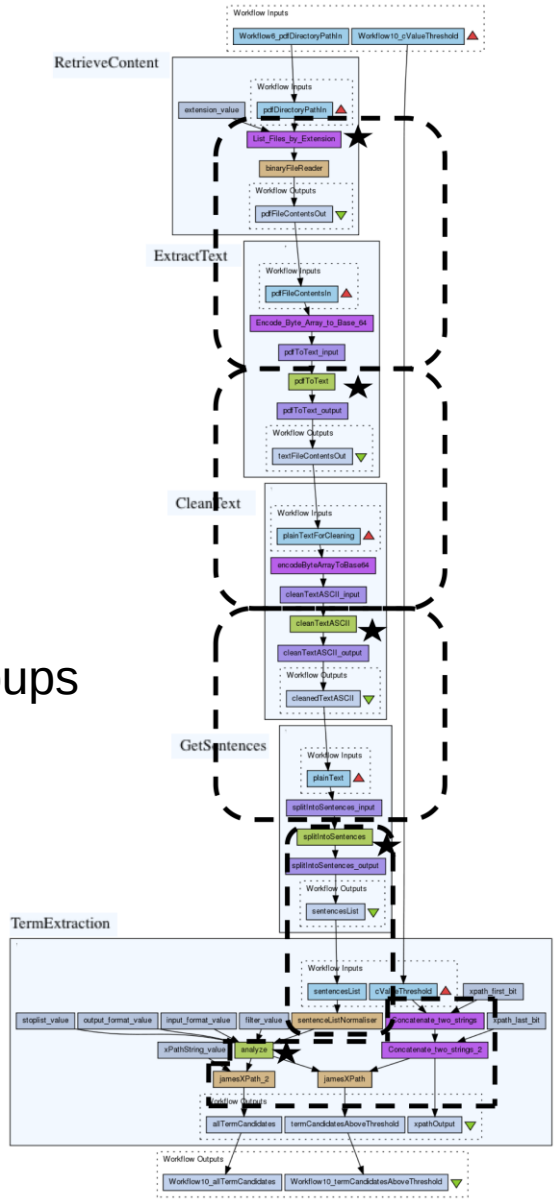
MANCHESTER 1824
The University of Manchester

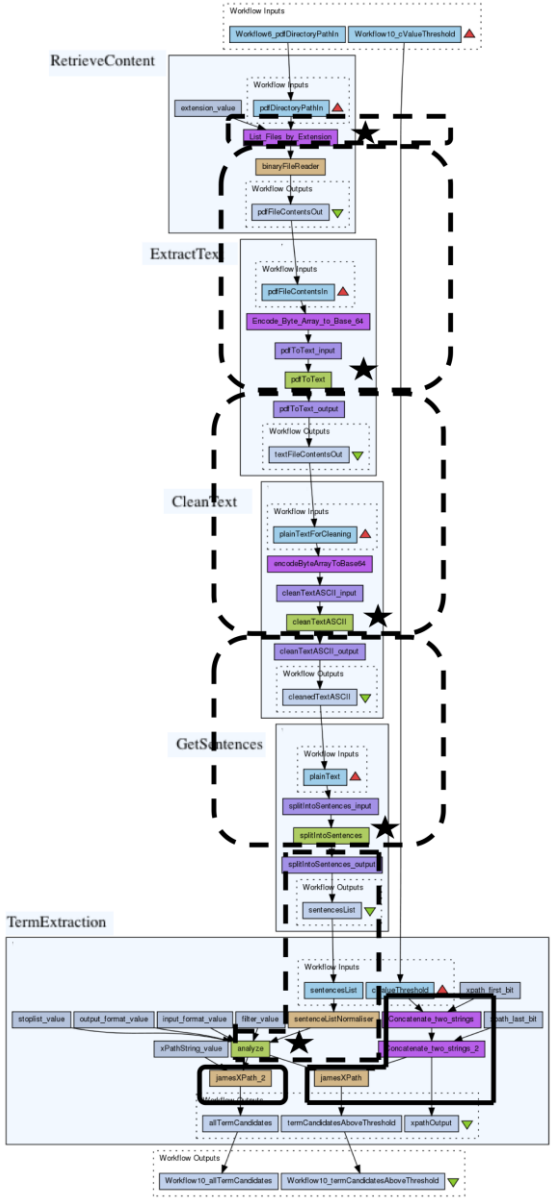
Task Grouping

	Process	Data
Wf Summaries -Collapse All	5/5	0/4
ZOOM	5/7	0/6

ZOOM's soundness policy creates two extra groups

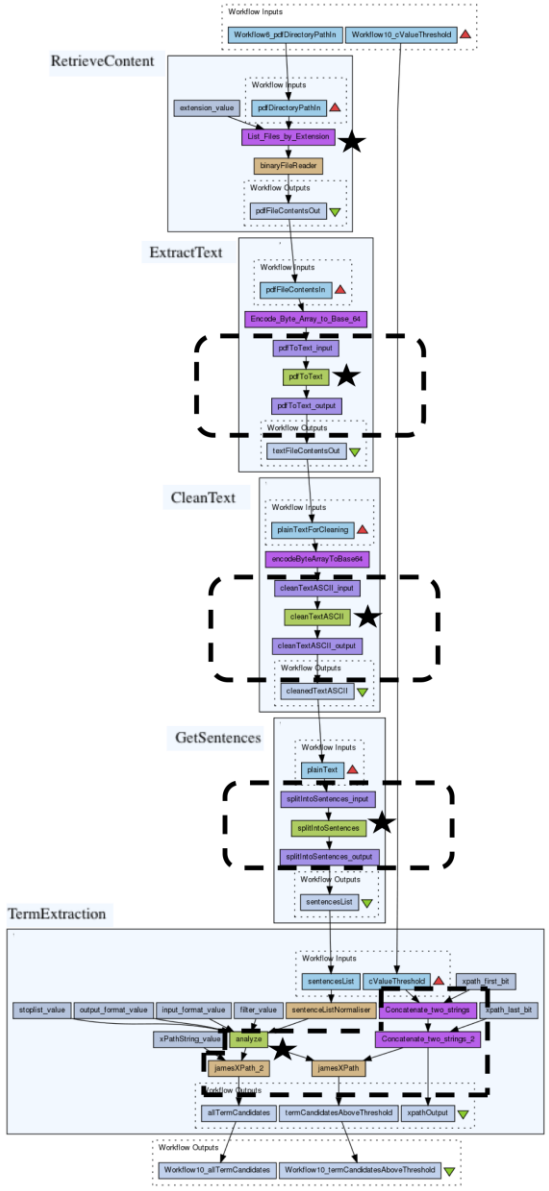
Where you put the boundary to groups matters for data abstraction!





Task Grouping

Abstracting selectively (less aggressively) by taking activity function (hence I/O characteristics) into account.



MANCHESTER 1824
The University of Manchester

Conclusions

- Abstraction systems focus on the process and do not directly cater for data significance.
- Are these observations generalisable?
  - Elimination unsuited for WFs with pack/unpack steps
  - Sweeping style grouping isn't helpful for data abstraction
  - Selective grouping relies on domain-specific policies.
- End-use informs suitable integrity policies.
- Scientists are accountable for reports. They are likely to favor having final say on the abstraction.
- Rethink abstraction as a prehoc process supporting workflow design.

"TOWARDS HARNESSING COMPUTATIONAL WORKFLOW PROVENANCE FOR EXPERIMENT REPORTING", PhD Dissertation, University of Manchester E-Scholar Repository. https://www.escholar.manchester.ac.uk/uk-ac-man-scw:300560

MANCHESTER 1824
The University of Manchester

Thank You!