Incremental workflow improvement through analysis of its data provenance

Paolo Missier
School of Computing,
Newcastle University, UK

In collaboration with the eScience Central group at Newcastle:
Prof. Paul Watson, Dr. Simon Woodman, Dr Hugo Hiden, Dr. Jacek Cala

TAPP'11 workshop
Heraklion, Crete
June 20-21, 2011

Motivation

Despite the growing momentum around provenance as a premiere form of metadata,
success exploitation stories trail models and technology advances

Motivation

Despite the growing momentum around provenance as a premiere form of metadata,
success exploitation stories trail models and technology advances

- We are getting very good at recording provenance of data:
  - multiple data models (OPM, Provenir, Janus, Karma, PML,...)
  - provenance-aware system / service architectures ...
    - PASS, Karma
  - ... and workflows
    - Kepler, Taverna, Galaxy, VisTrails,...

- But, what are systems/applications really doing with it?
  - deliver value to users? i.e., in e-science, in the Web
    - scientific reproducibility, quality, trust
  - optimize system analysis, performance?
    - enable partial re-run of resource-intensive processes

Broad goal and specific research context

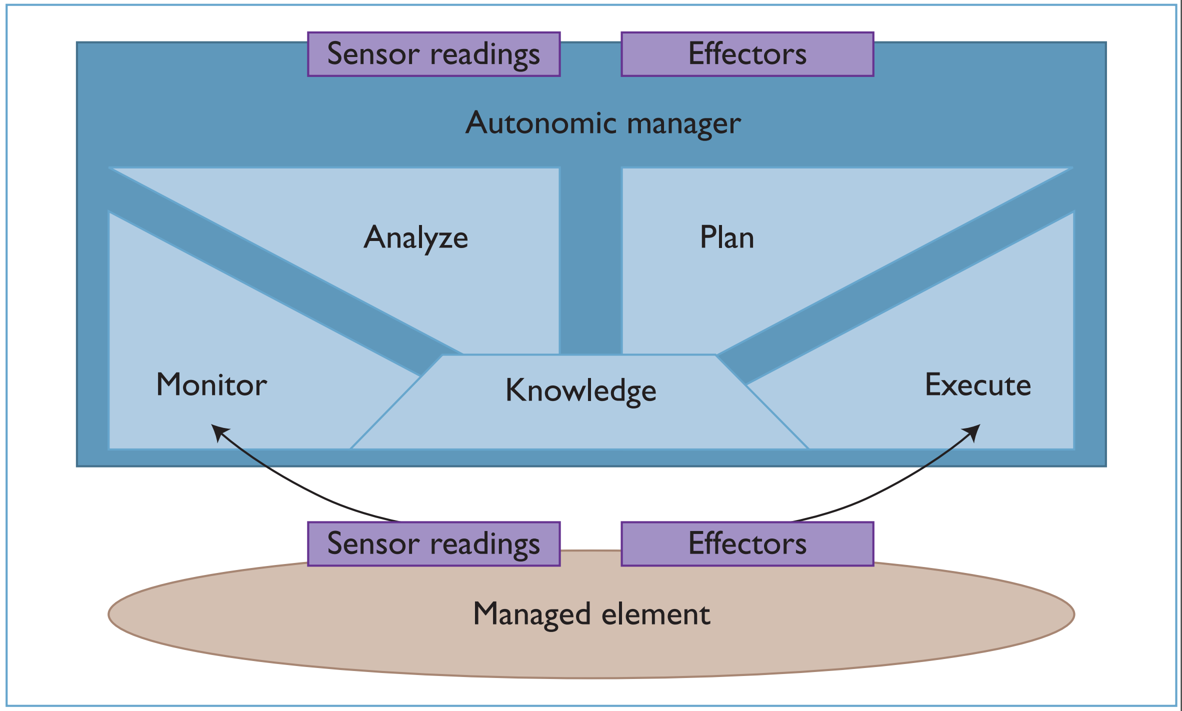
A systematic study on methods and applications of mining / learning techniques applied to large corpora of provenance metadata

- An opportunistic starting point:

  optimization of resource-intensive, repetitive, provenance-aware e-science workflows

  Cloud-based workflows come with a clear cost model

  - valuable testbed readily available:
    - real e-science applications
    - large provenance graphs
  - dynamic optimization requires many runs



- Reference framework:

  adaptive, self-managing software systems

  - systems that can dynamically adapt their behaviour in response to changing conditions in the inputs or in their environment [1,2]

[1] IBM. An architectural blueprint for autonomic computing. Tech. rep., IBM, 2011
[2] Huebscher, M. C., and McCann, J. A. A survey of autonomic computing - degrees, models, and applications. ACM Computing Surveys CSUR 40, 3 (2008), 1–28.

Research

Hypothesis: Provenance traces recorded for past runs of a workflow can be used to make future runs more efficient

Approach: Add adaptive control to an existing workflow, with provenance analysis at its core → new *recommender* task

Research

Hypothesis: Provenance traces recorded for past runs of a workflow can be used to make future runs more efficient

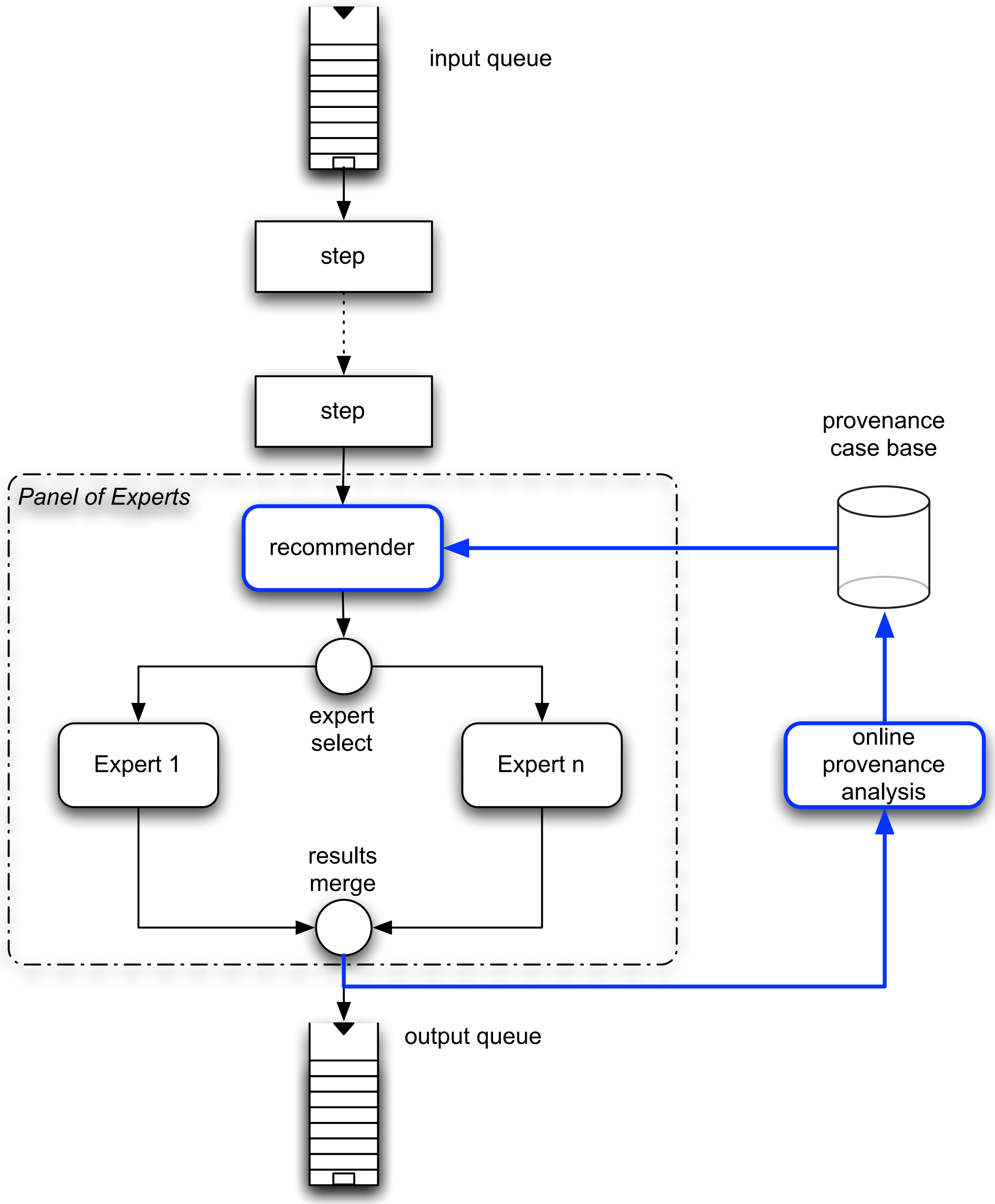
Approach: Add adaptive control to an existing workflow, with provenance analysis at its core → new *recommender* task

Applicable for instance to a "Panel of Experts" pattern:

– N experts are activated on same inputs, best outputs ar selected

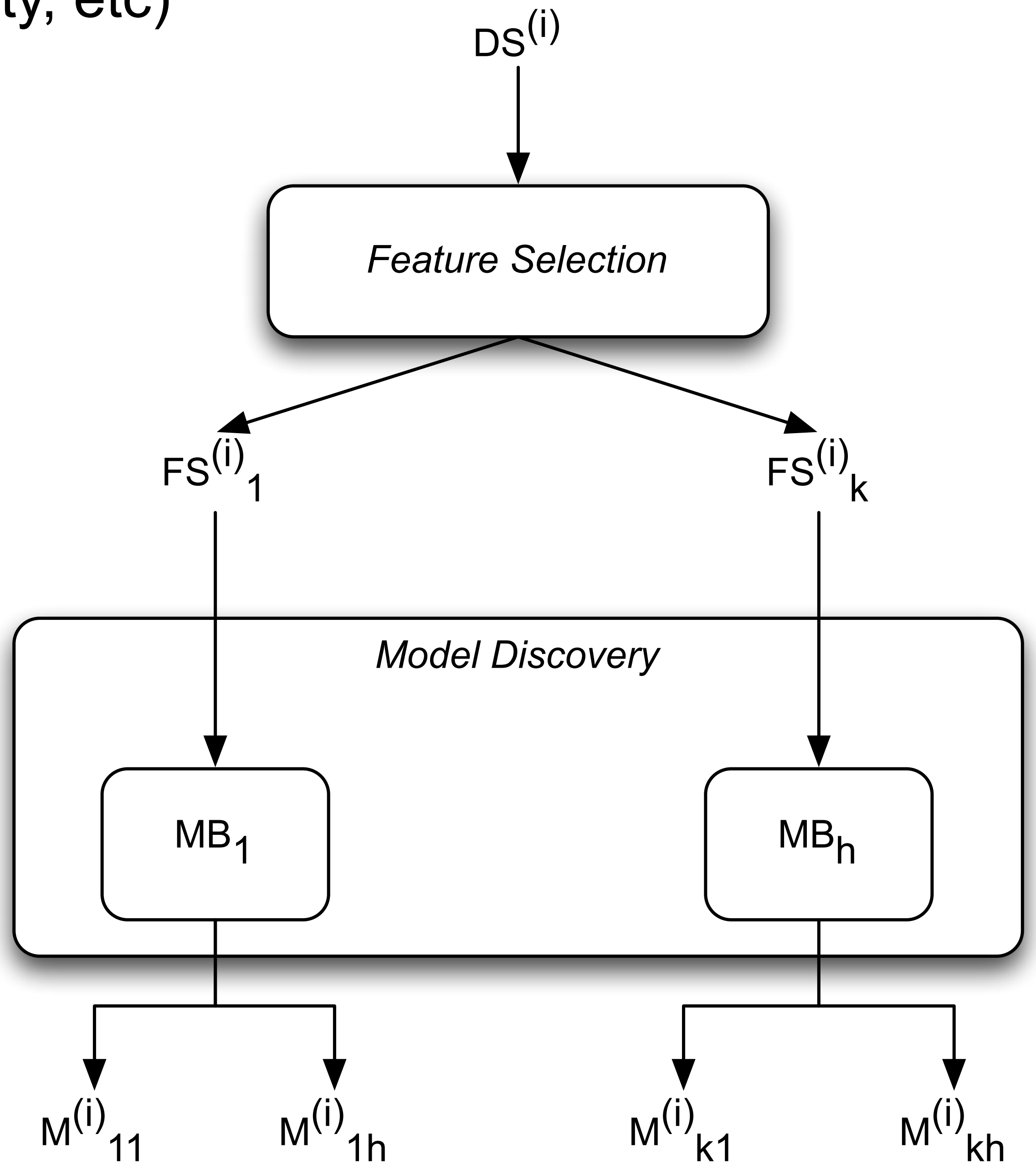
Provenance used for incremental correlation of the inputs to the experts' success rate

- Provenance of run *i* indicates which experts performed well on their input
- **Adaptively pruning the process space**: on run *i*+1, use provenance of output computed by runs 1..n to select/prioritize invocation of experts



Case study: DiscoveryBus workflow

- QSAR: Quantitative structure-activity relationships
  - at forefront of Chemical Engineering research
- OpenQsar project (http://www.openqsar.com):
  - Establish correlations between the structure of molecular compounds and some of their associated activities (toxicity, solubility, etc)
- DiscoveryBus: a workflow implementation of OpenQSAR
  - eScience Central cloud-based workflow system
  - datasets $DS^{(i)}$ are a family of structurally homogeneous molecules
  - Feature Selection extracts few relevant features from $DS^{(i)}$
  - Each learning scheme $MB_1 \ldots MB_h$ generates a different predictive models for molecular activity

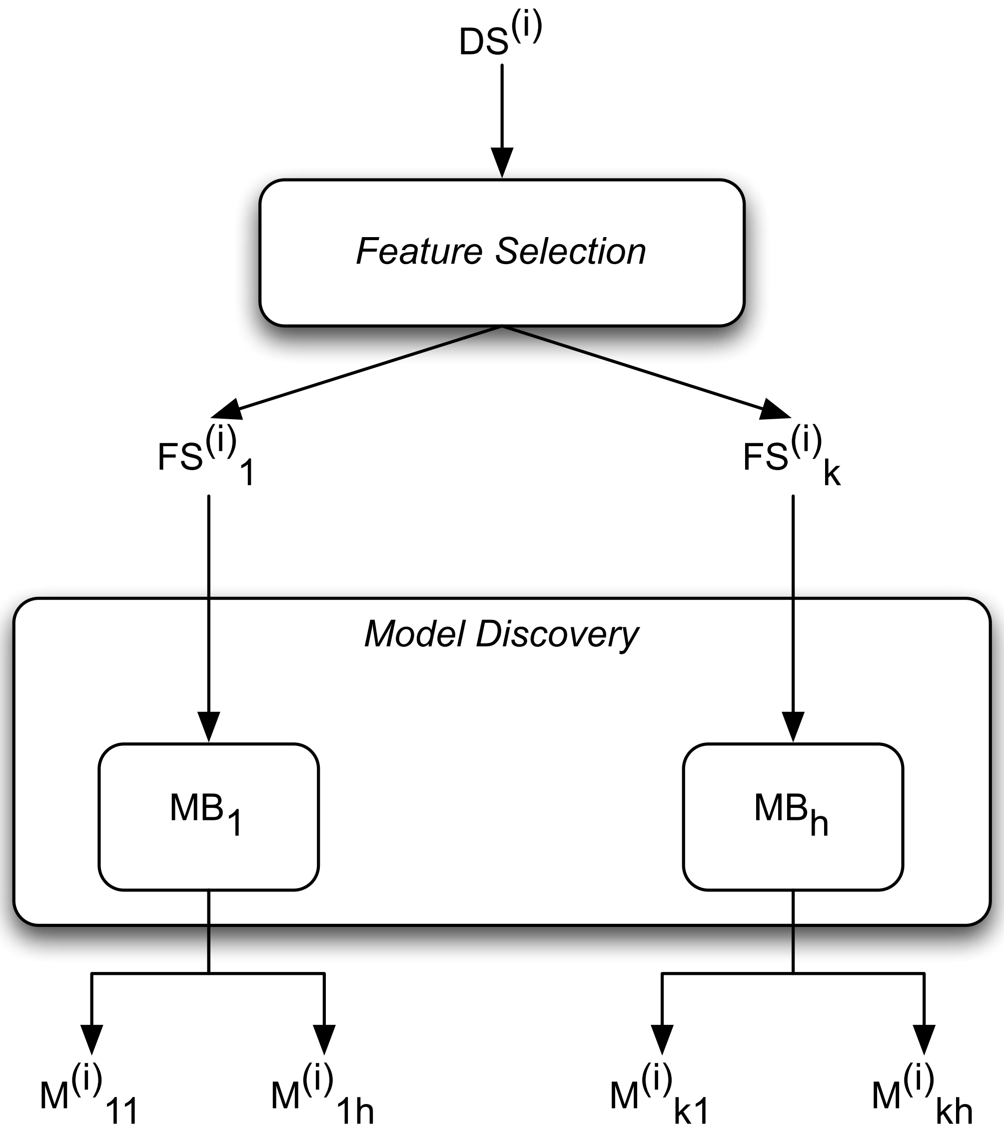


Repetitive and resource-intensive:

Workflow execution repeated over about 10K different input datasets

Role of provenance in DiscoveryBus

- Connection to Panel of Experts:
  - Experts ⇒ model builders MB_i
  - Experts outcome ⇒ quality of generated model (predictive power, stability)
- Optimization goal:
  - to prioritize invocation of the MB_i based on their past performance on inputs similar to $FS^{(i)}_j$



Provenance correlates quality of output models $M^{(i)}_{jk}$ to intermediate feature sets $FS^{(i)}_j$:

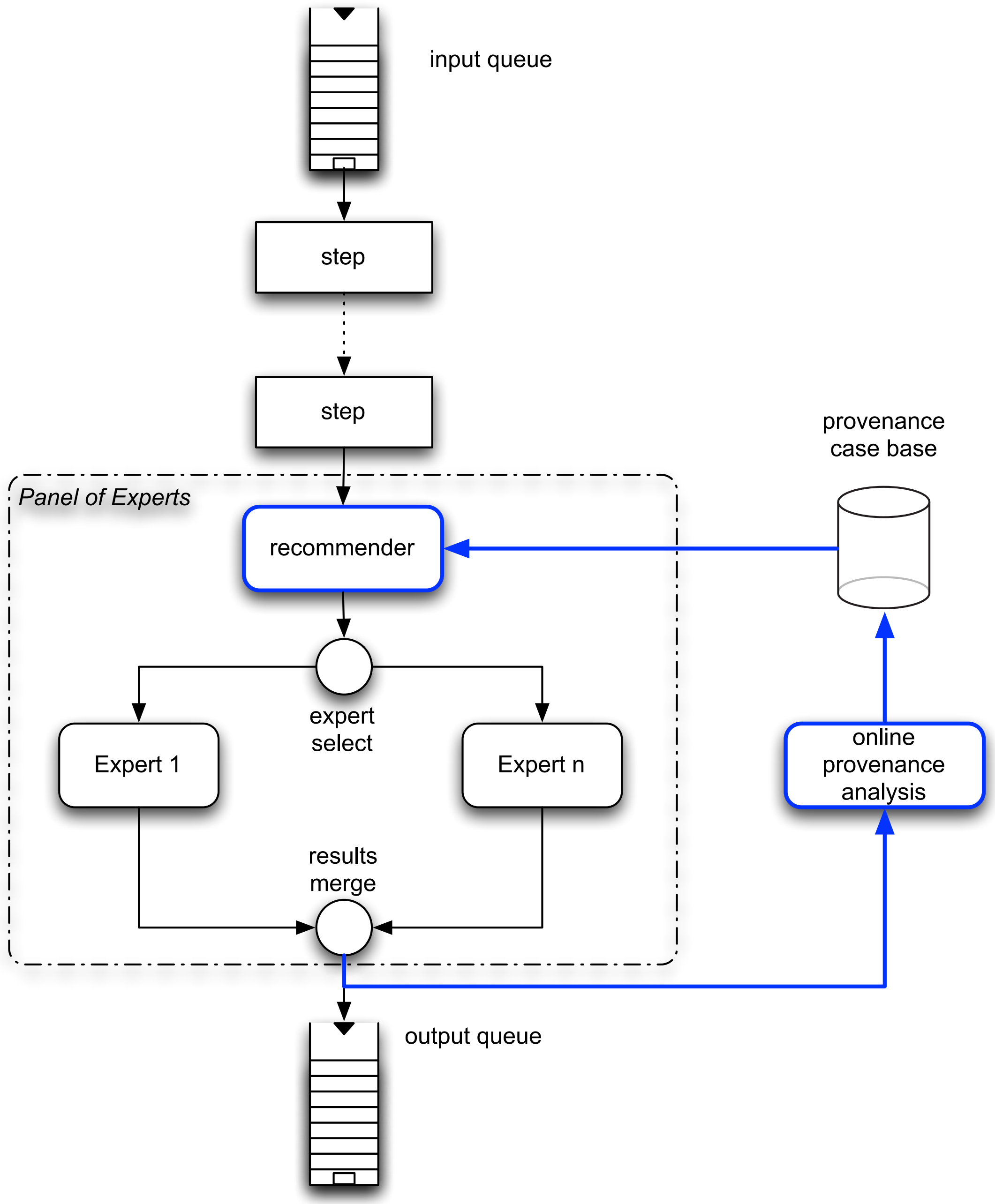
$$M^{(i)}_{jh} \stackrel{WasGeneratedBy}{\longrightarrow} MB_h \stackrel{used}{\longrightarrow} FS^{(i)}_j \stackrel{WasDerivedFrom}{\longrightarrow} DS^{(i)}$$

The recommender

- One Quality Matrix QM_{FS} is associated to each Feature Set FS
- $QM_{FS}[MB_i]$ encodes the success history of model builder MB_i in the workflow every time FS is used as input:

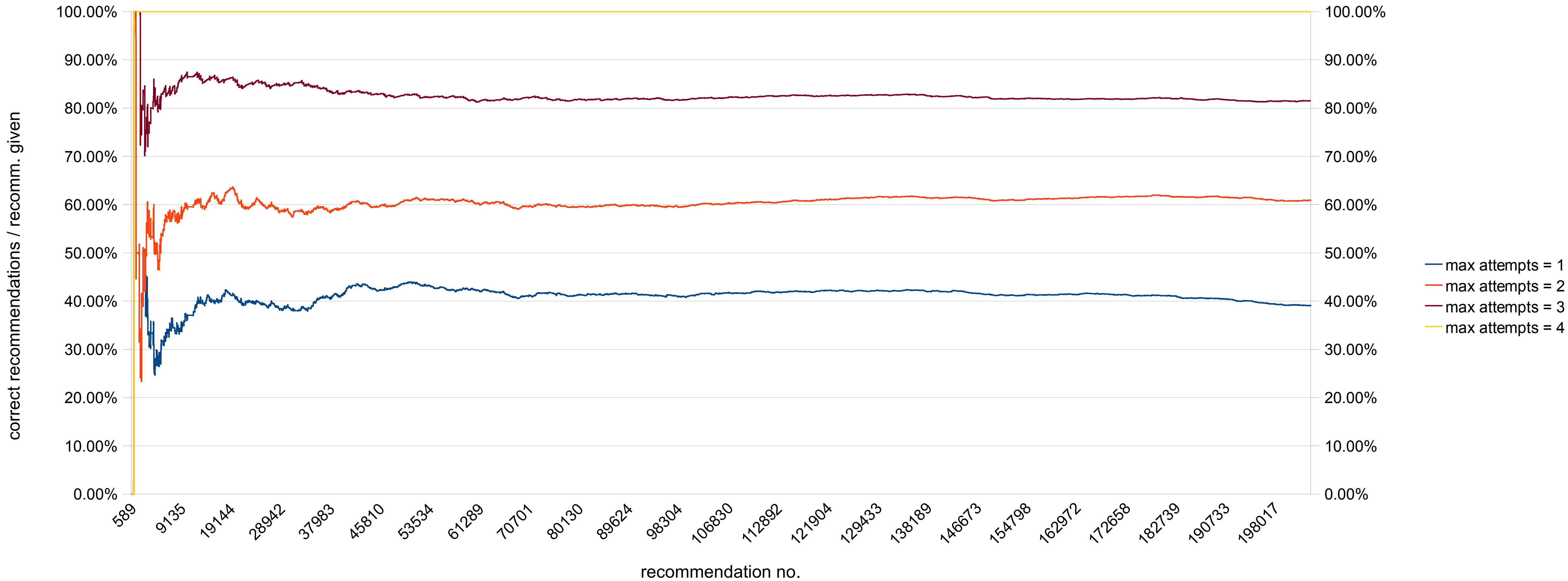
$$QM_{FS}[MB_h] = \langle G, B \rangle$$

- G (resp B): number of times MB_h has been observed to produce a good (resp. bad) model when applied to input FS
- QM_{FS} is updated every time FS appears in the provenance graph
- The builders' historical success rate G induces a dynamic partial order on the MB_i



- For each run, the Recommender:
  - intercepts FS in the flow
  - returns partial order from QM_{FS}

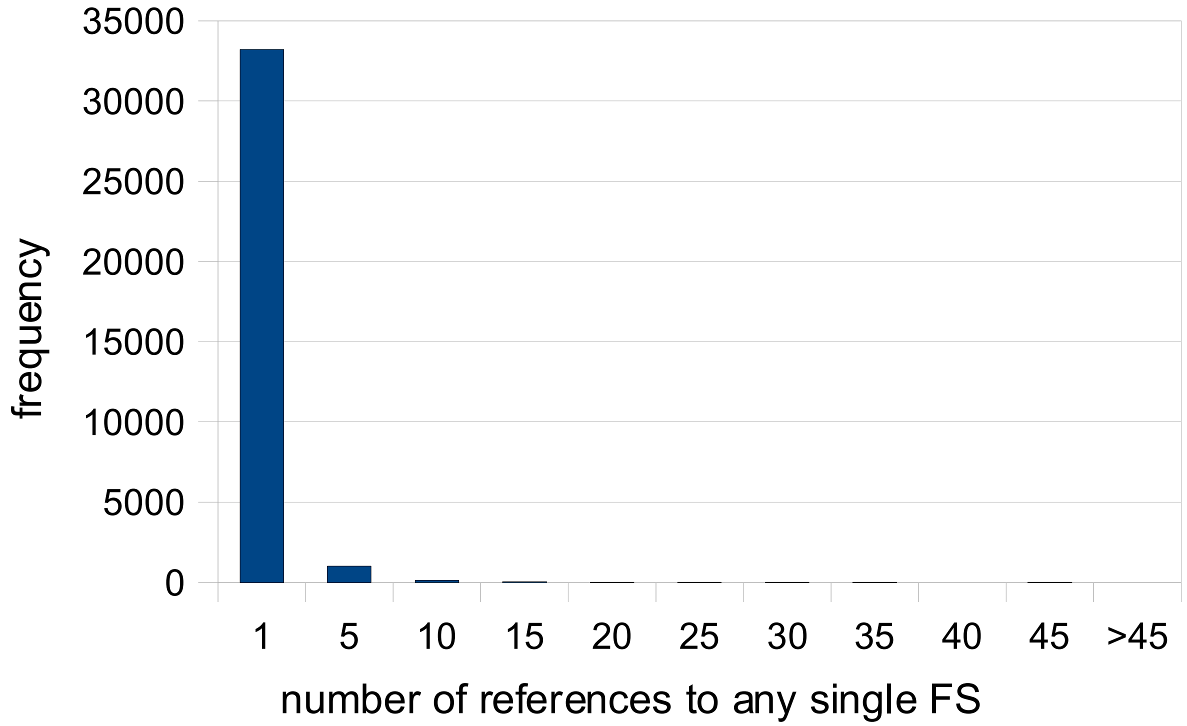
Early experimental results



- max_attempts is the accuracy/resources trade-off parameter
  - max_attempts = n: only first n out of H model builders are invoked
- Chart shows net accuracy over the entire available history of runs
  - success rate / number of recommendations ***given***

Limitations and ongoing work

- Approach suffers when FS space is sparse
  - most FS seen only once
- Recommender *abstains* when QM_{FS} not sufficiently populated
- This is the main hurdle to successful optimization



- Strategy: increase space density by clustering the FS
  - needs a distance metric over the set of FS
  - hierarchical clustering should provide a way to experiment with accuracy/efficiency trade-offs

Short summary

- What happens when you try to apply mining/learning techniques to large corpora of provenance metadata?
  - any interesting applications / use cases?
  - which techniques apply?
  - are there significant research challenges, or an orchard of low-hanging fruits?

- privacy in provenance mining
- what provenance models lend themselves well to mining
- ...