LoonyBin:

Keeping Language Technologists Sane through Automated Management of (Hyper)Workflows

Jonathan Clark and Alon Lavie
Carnegie-Mellon University

LREC 2010
Thursday, May 20, 2010

Outline

- Empirical NLP Research
  - Day-to-day issues
  - Current problems
    - LoonyBin's solutions
- Workflows
- HyperWorkflows

Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers

Empirical NLP

- Plumbing: Gluing (Linux) tools together

- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers

Empirical NLP

- Plumbing: Gluing (Linux) tools together

- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers

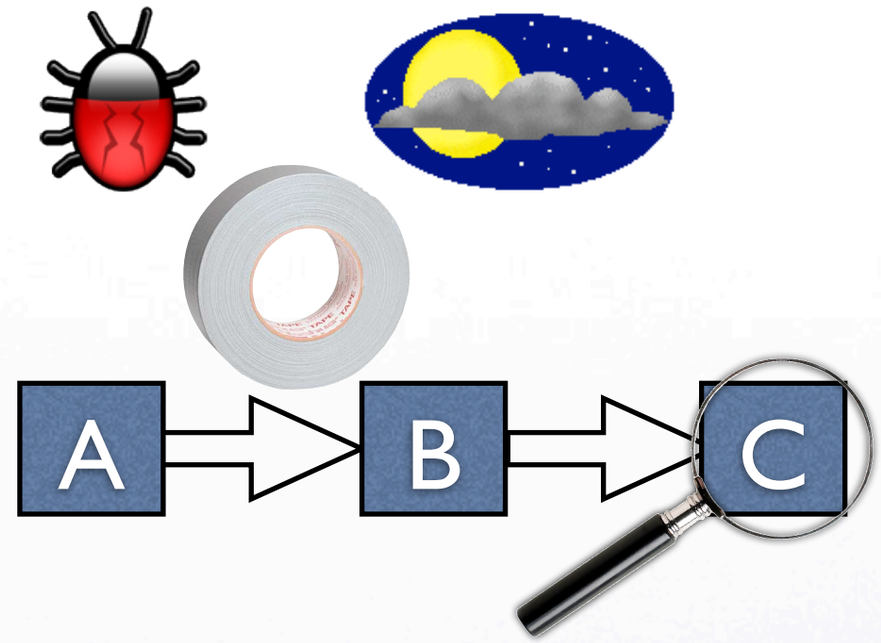
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results

- Sanity checking
- Running variations
- Moving between clusters & schedulers

Empirical NLP

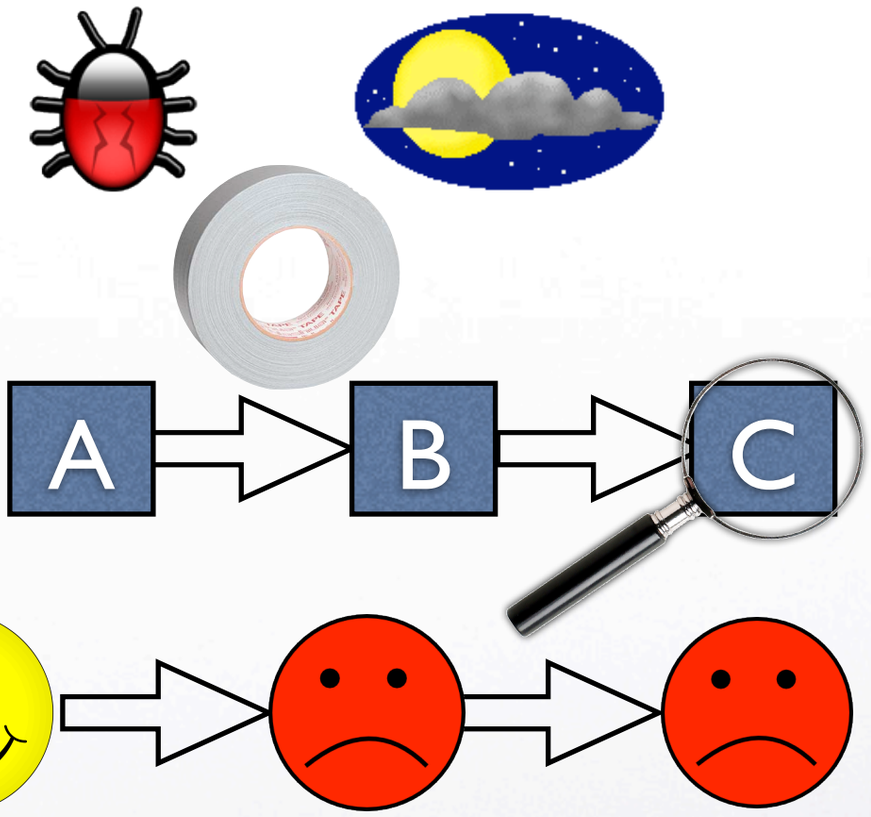
- Plumbing: Gluing (Linux) tools together
- Recording results



- Sanity checking
- Running variations
- Moving between clusters & schedulers

Empirical NLP

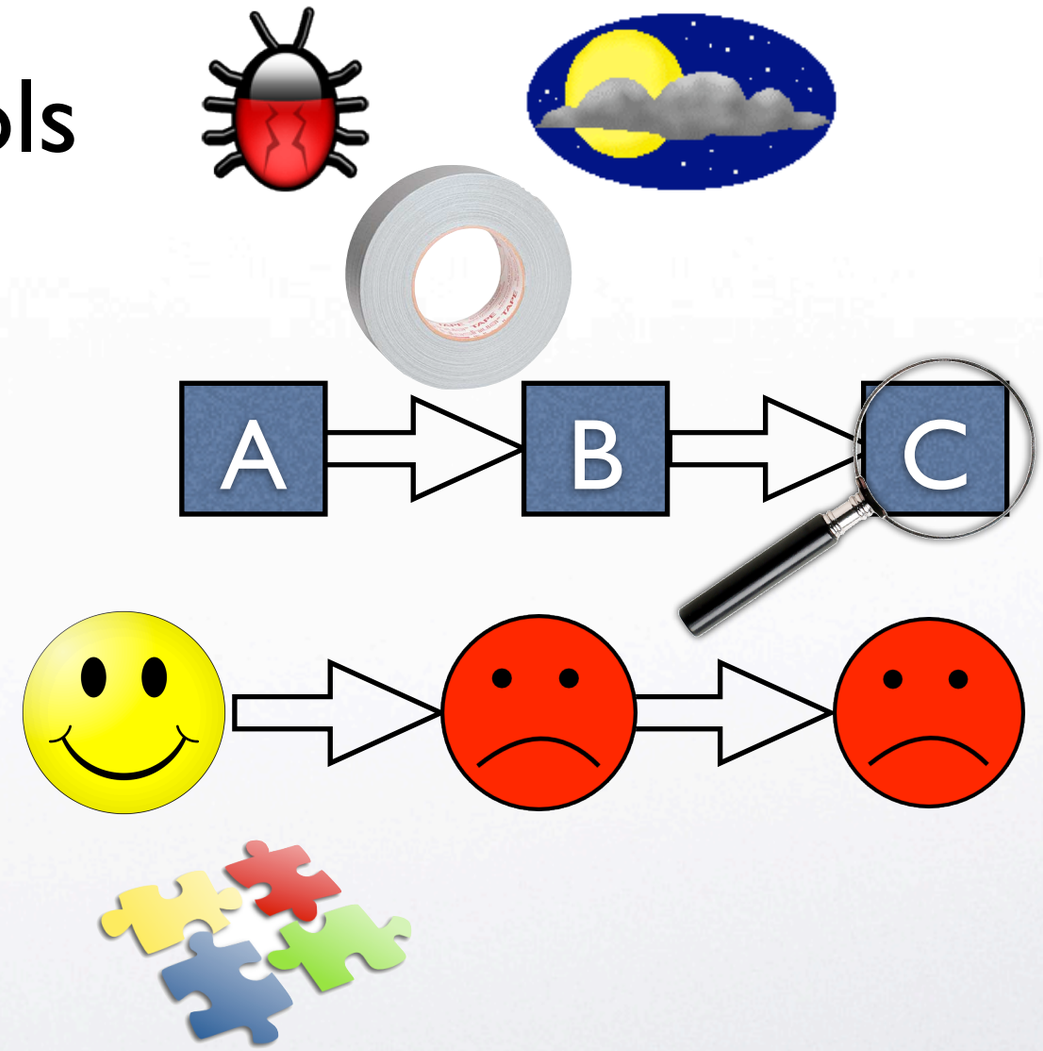
- Plumbing: Gluing (Linux) tools together
- Recording results



- Sanity checking
- Running variations
- Moving between clusters & schedulers

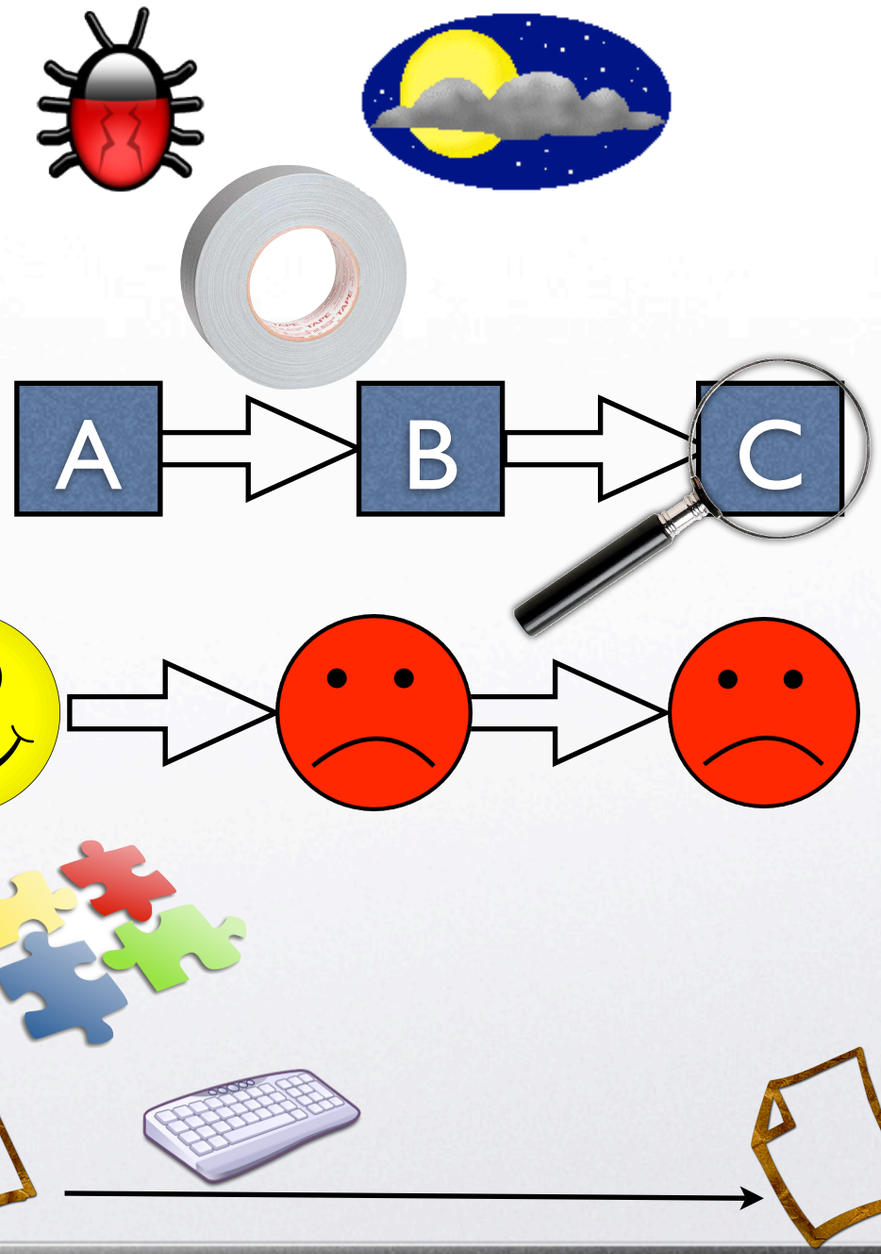
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



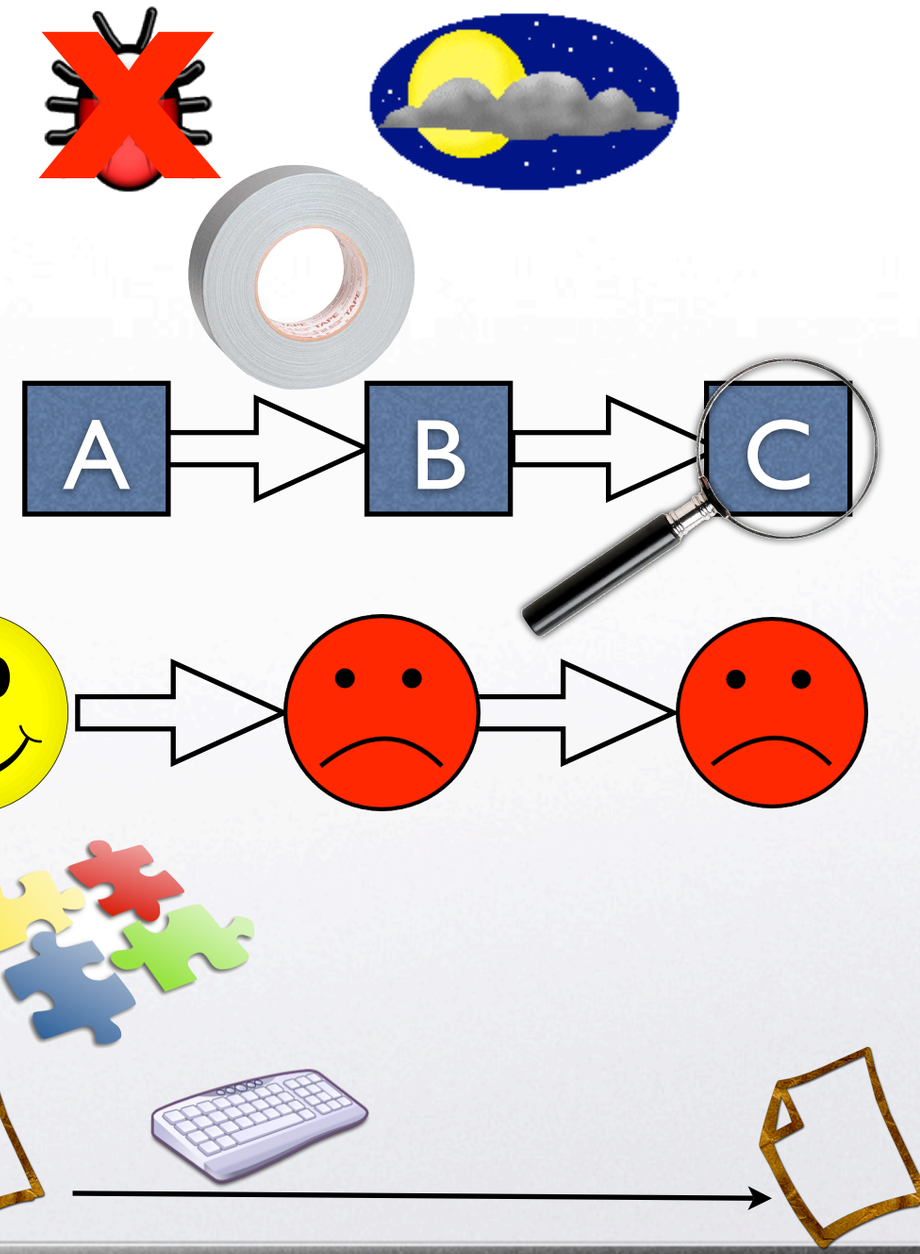
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



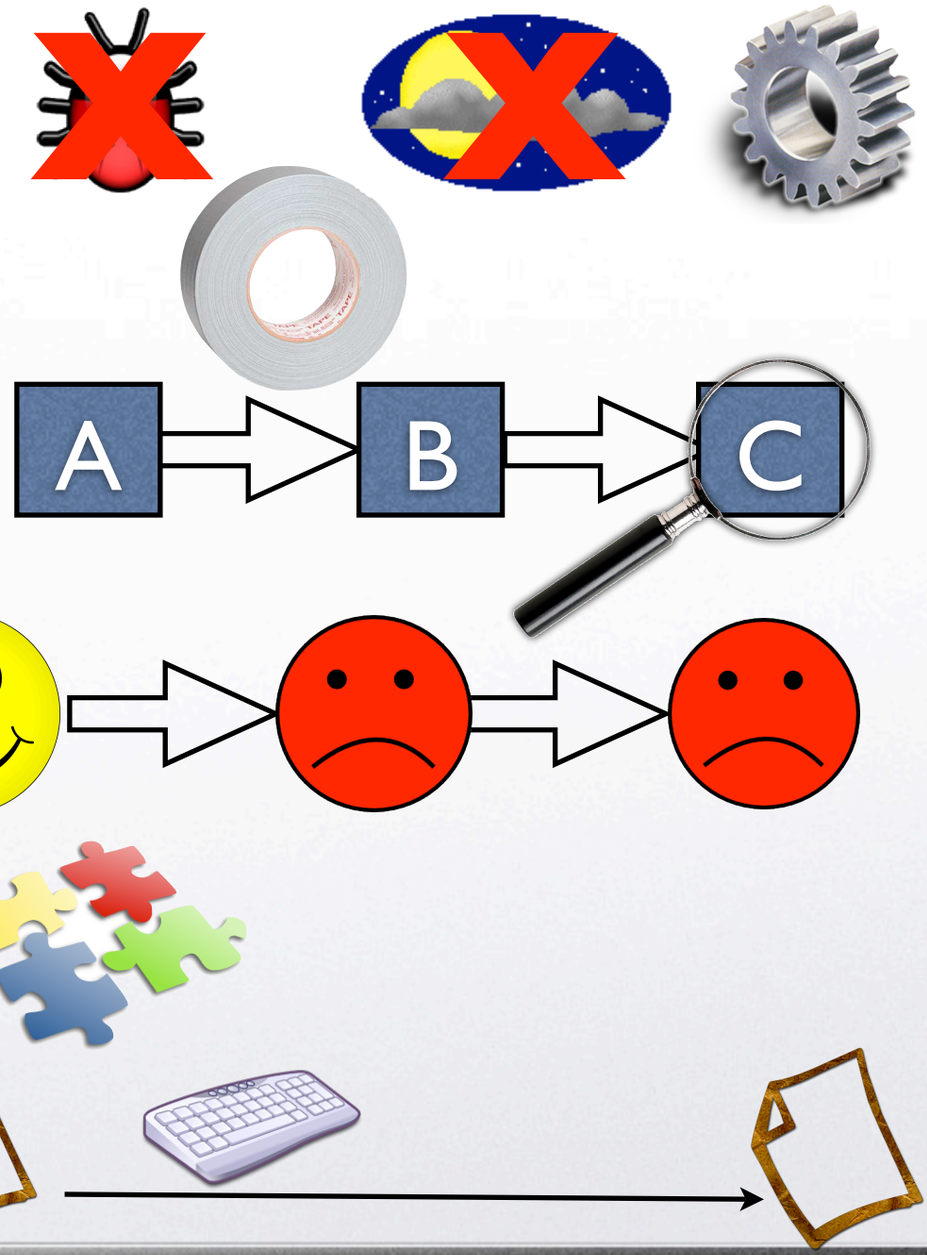
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



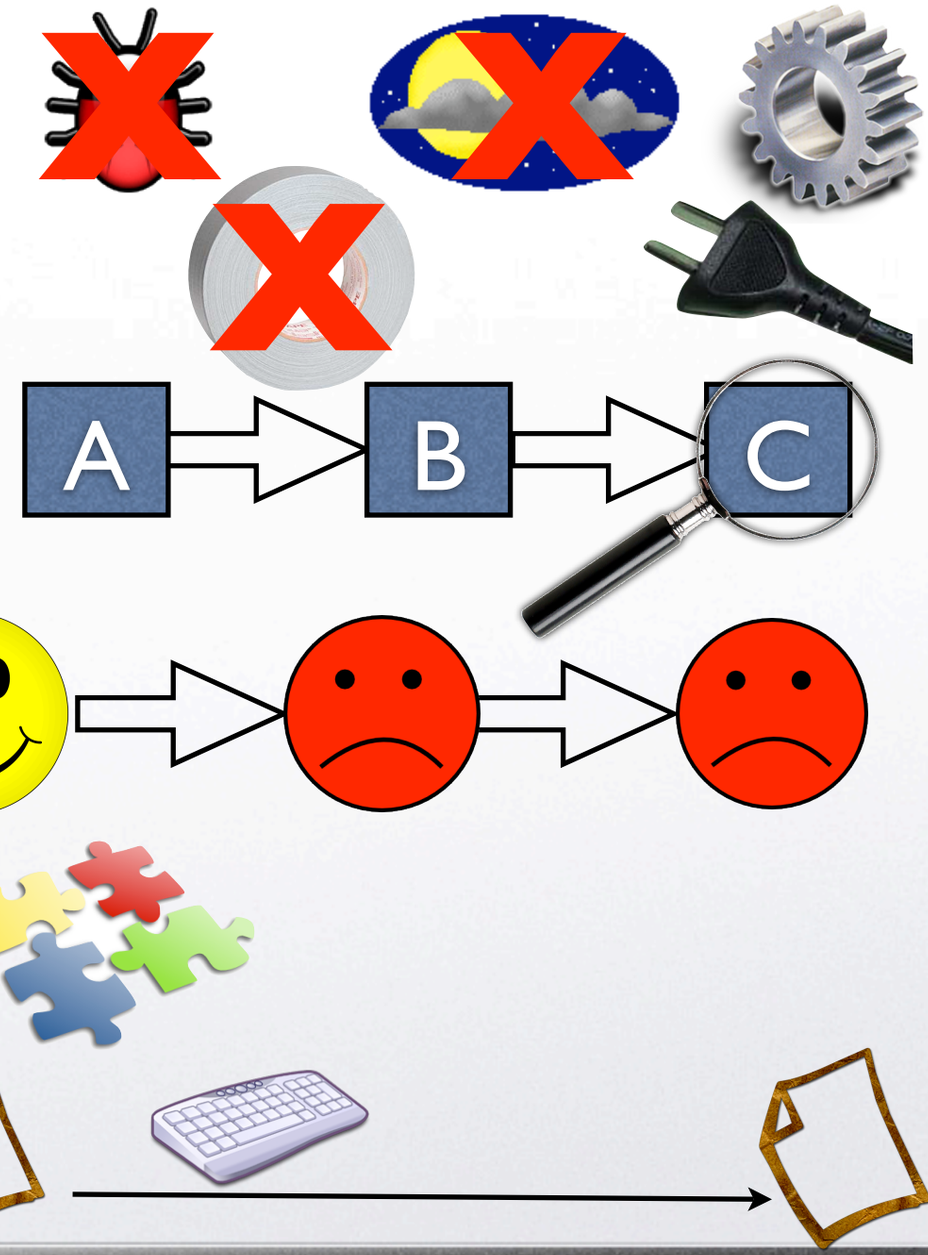
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



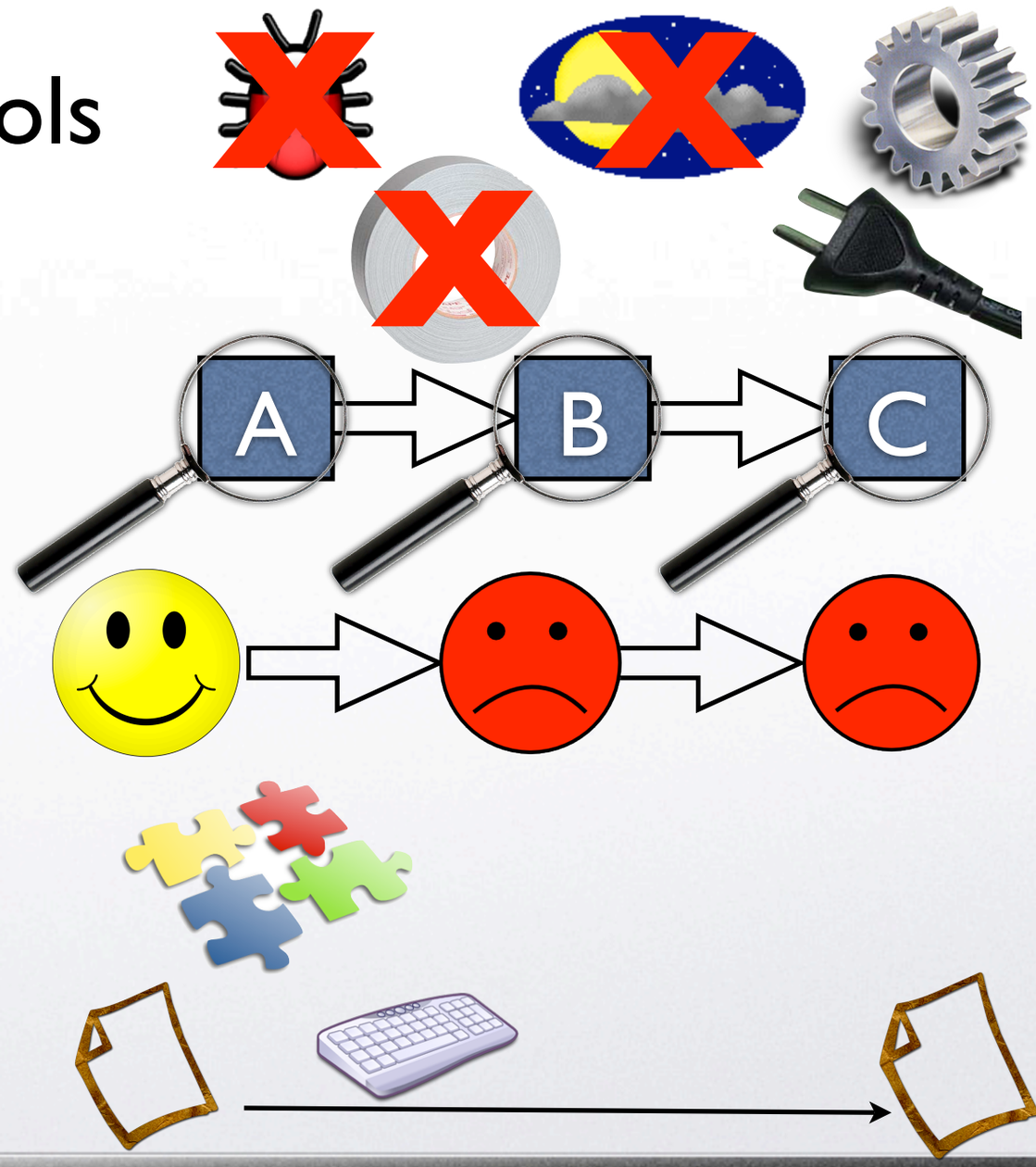
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



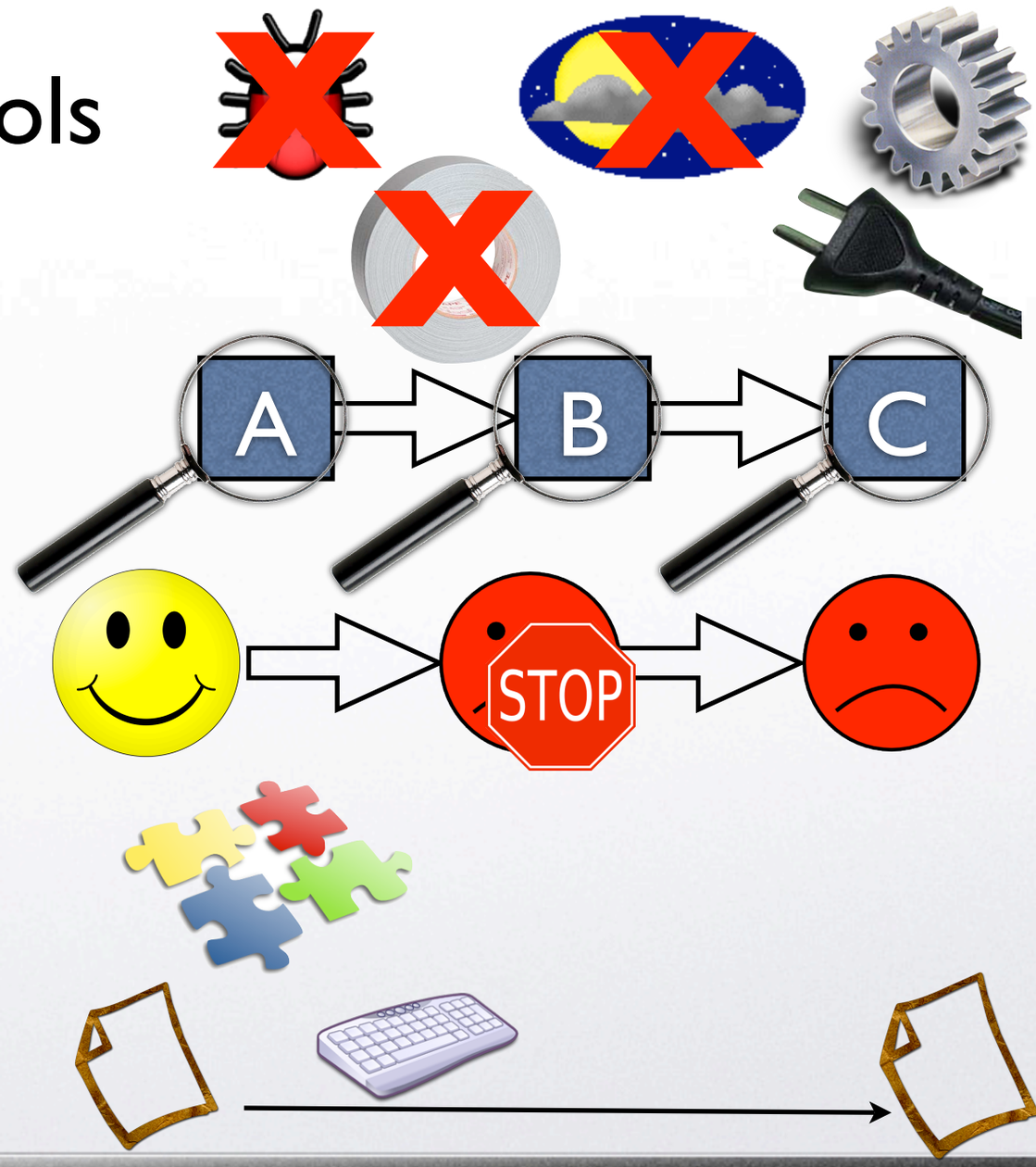
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



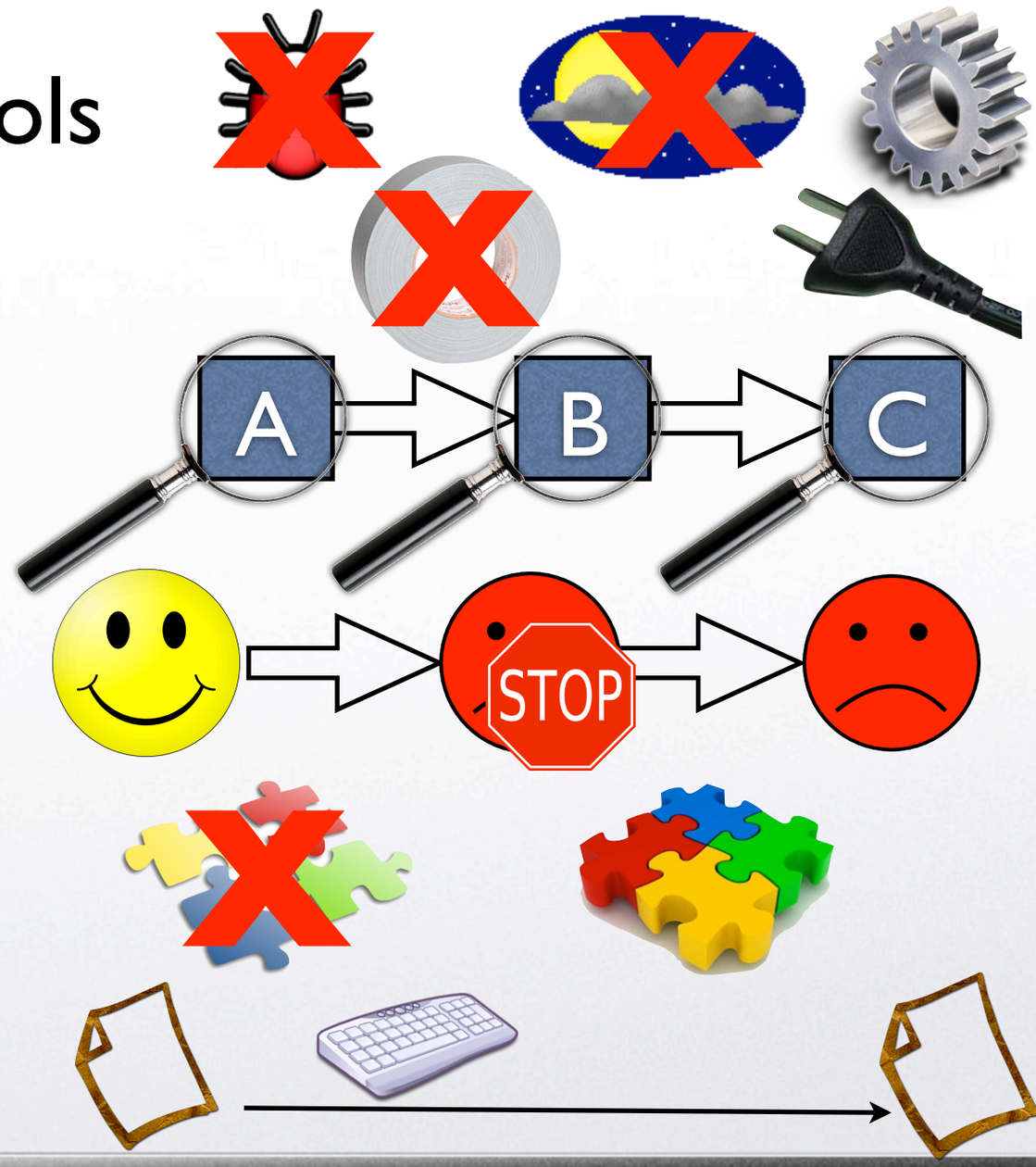
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



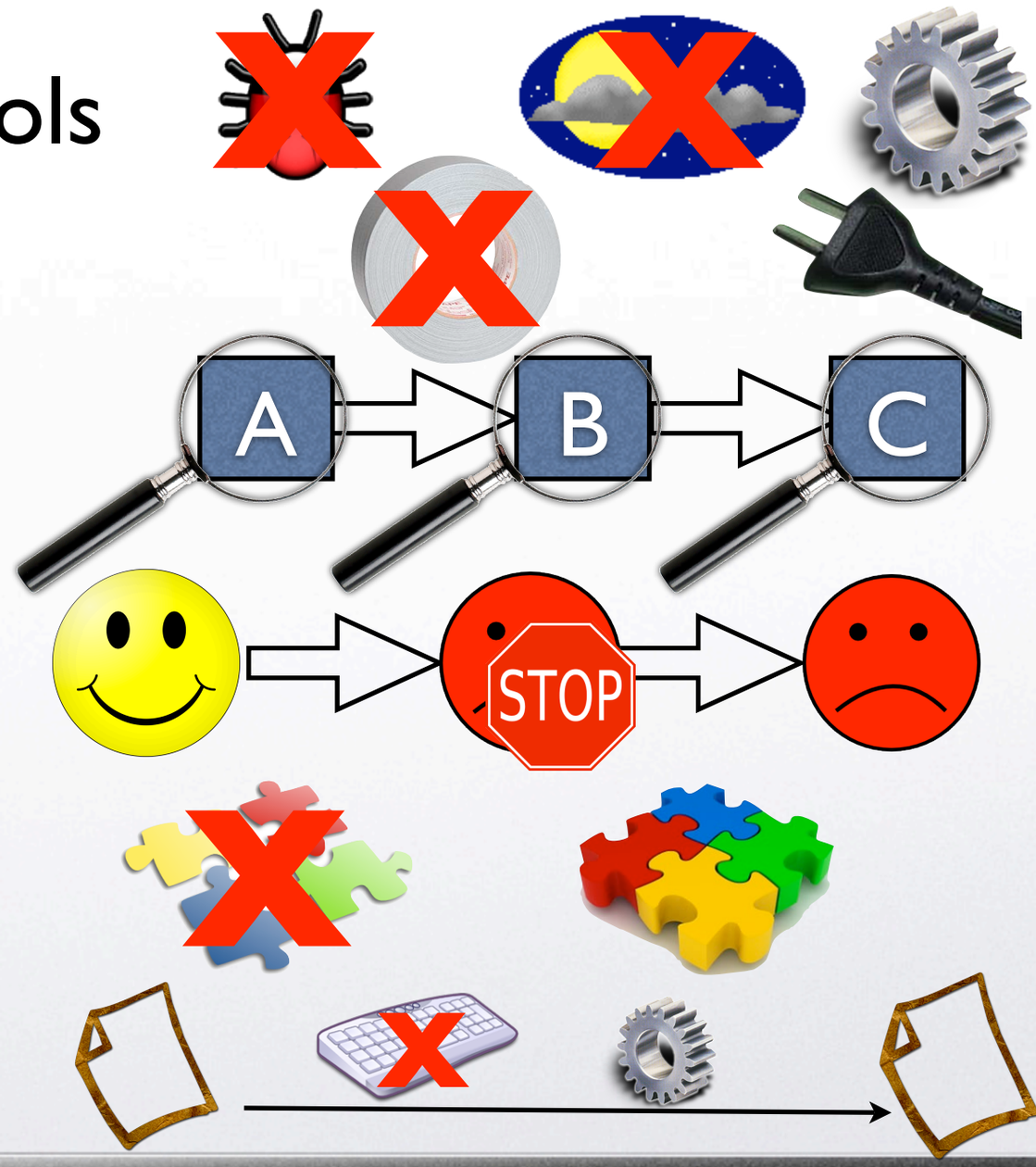
Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



Empirical NLP

- Plumbing: Gluing (Linux) tools together
- Recording results
- Sanity checking
- Running variations
- Moving between clusters & schedulers



Proposed Solution:

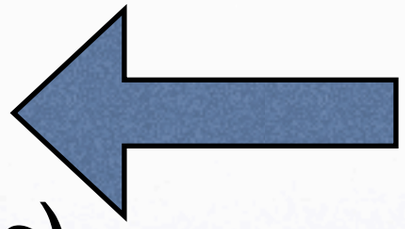
HyperWorkflow Management

LoonyBin

- Define the tools
  (inputs/outputs/parameters → shell commands)
- Define the workflow
  (DAG of steps and dependencies)
- Generate & run a shell script

LoonyBin

- Define the tools
  (inputs/outputs/parameters → shell commands)
- Define the workflow
  (DAG of steps and dependencies)



- Generate & run a shell script

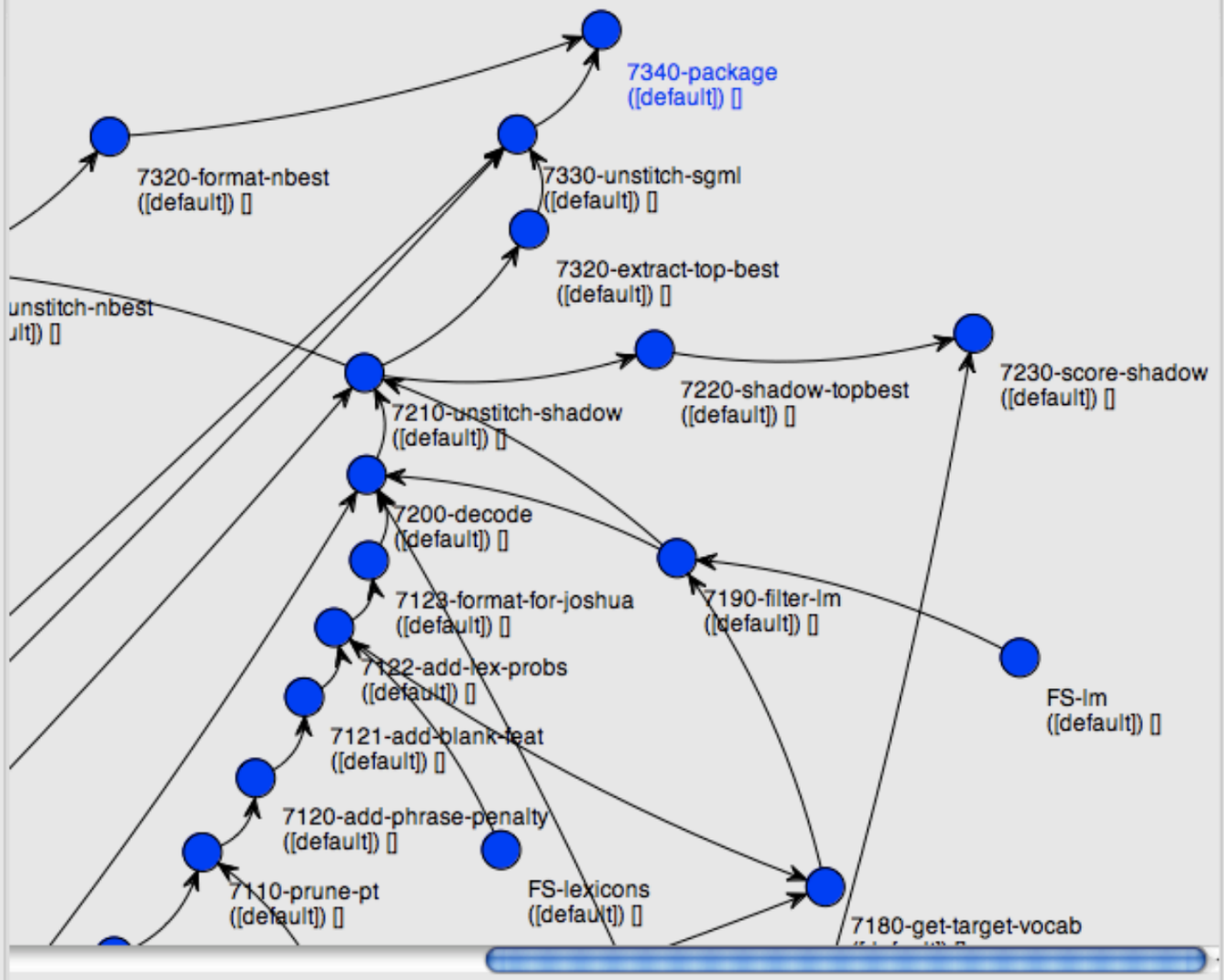
LoonyBin HyperDAG Designer V0.4.0

Pipeline Options

Mouse Mode | Scrolling | Selecting | Editing | Transforming Mode

- Tools
  - MANUAL FILESYSTEM
  - MANUAL HDFS
  - OR
  - PARAMETER BOX
  - Machine Translation
    - Decoders
    - Grammars and Tables
    - Language Modeling
    - Mono Corpus
    - Output
    - Parallel Corpus
    - Parsing
    - Scoring
    - Tuning
    - Word Alignment
  - Trashed Tools
    - Errors

Untitled Workflow | gale-p4-audio-eval.pipe



Tool Name: GALE Packager

Step Name: 7340-packag

sysName: CMU-StatXfer-201

occassion: P4 Audio Evaluatio

Machine Config: barrow

LoonyBin HyperDAG Designer V0.4.0

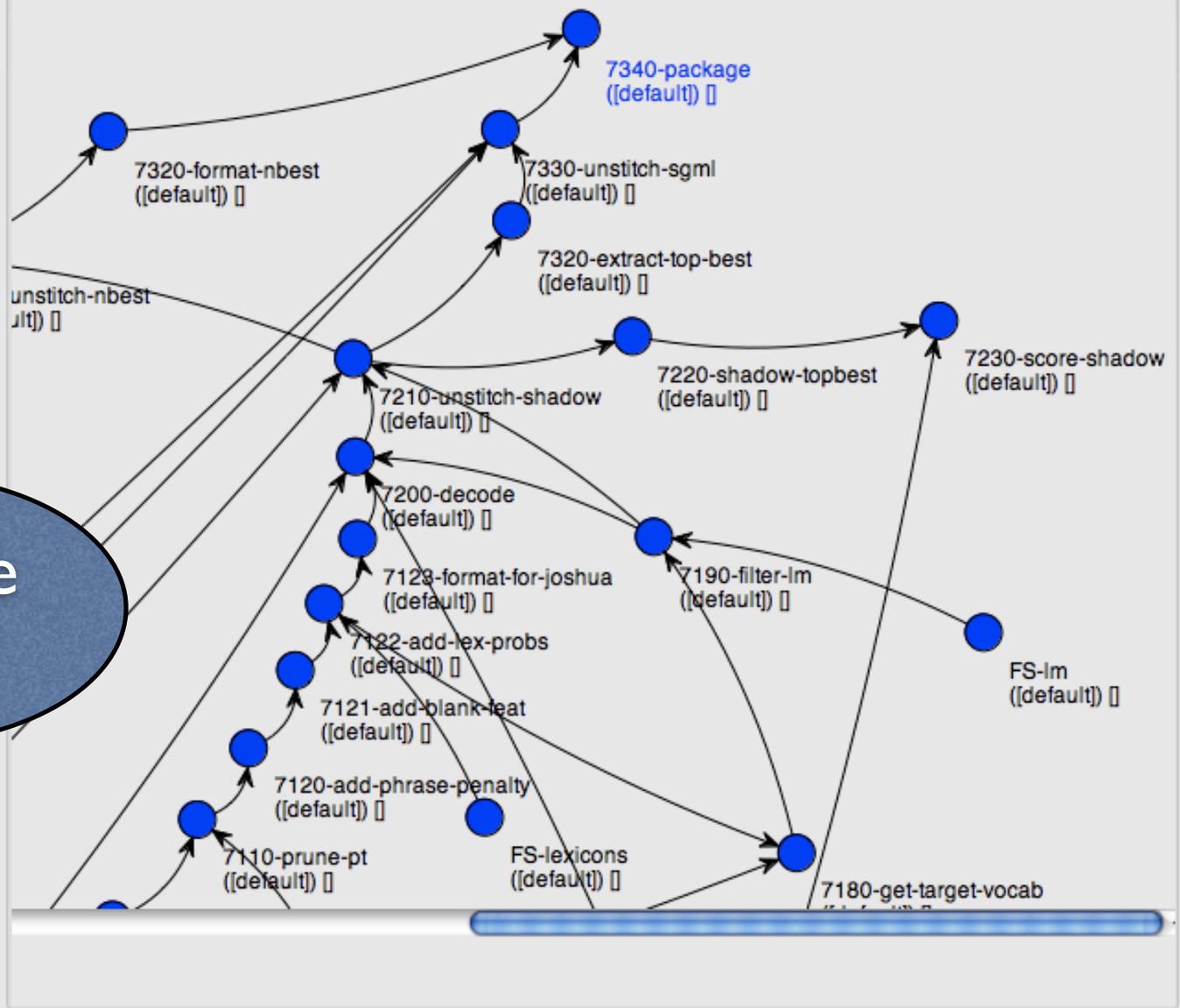
Pipeline Options

Mouse Mode | Scrolling | Selecting | Editing | Transforming Mode

- Tools
  - MANUAL FILESYSTEM
  - MANUAL HDFS
  - OR
  - PARAMETER BOX
- Machine Translation
  - Decoders
  - Grammars and Tables
  - Language Modeling
  - Mono Corpus
  - Output
  - Parallel Corpus
  - Parsing
  - Scoring
  - Tuning
  - Word Alignment

Available Tools

Untitled Workflow | gale-p4-audio-eval.pipe



Tool Name: GALE Packager

Step Name: 7340-packag

sysName: CMU-StatXfer-201

occassion: P4 Audio Evaluatio

Machine Config: barrow

LoonyBin HyperDAG Designer V0.4.0

Pipeline Options

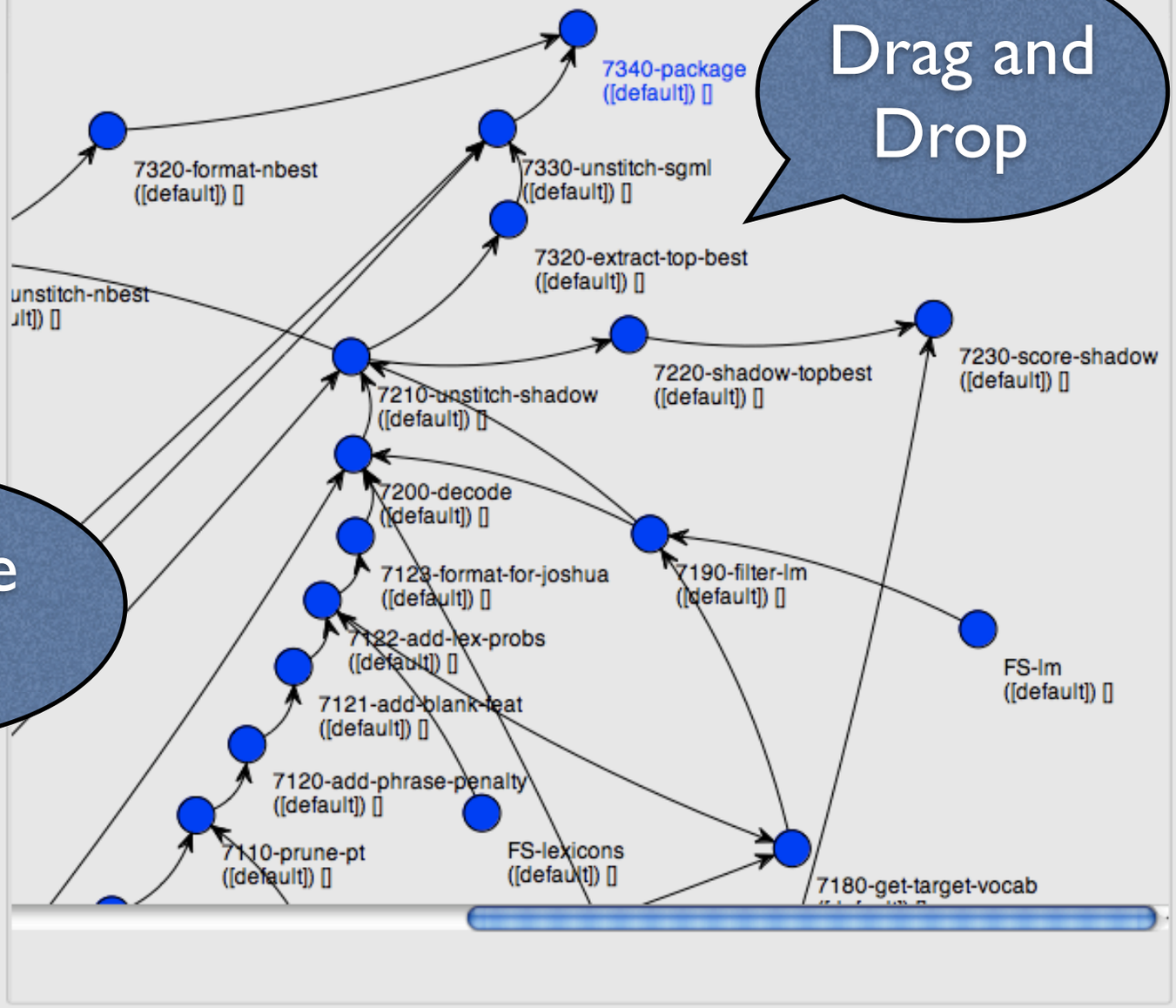
Mouse Mode | Scrolling | Selecting | Editing | Transforming Mode

- Tools
  - MANUAL FILESYSTEM
  - MANUAL HDFS
  - OR
  - PARAMETER BOX
- Machine Translation
  - Decoders
  - Grammars and Tables
  - Language Modeling
  - Mono Corpus
  - Output
  - Parallel Corpus
  - Parsing
  - Scoring
  - Tuning
  - Word Alignment

Available Tools

Untitled Workflow | gale-p4-audio-eval.pipe

Drag and Drop



Tool Name: GALE Packager

Step Name: 7340-packag

sysName: CMU-StatXfer-201

occassion: P4 Audio Evaluatio

Machine Config: barrow

LoonyBin HyperDAG Designer V0.4.0

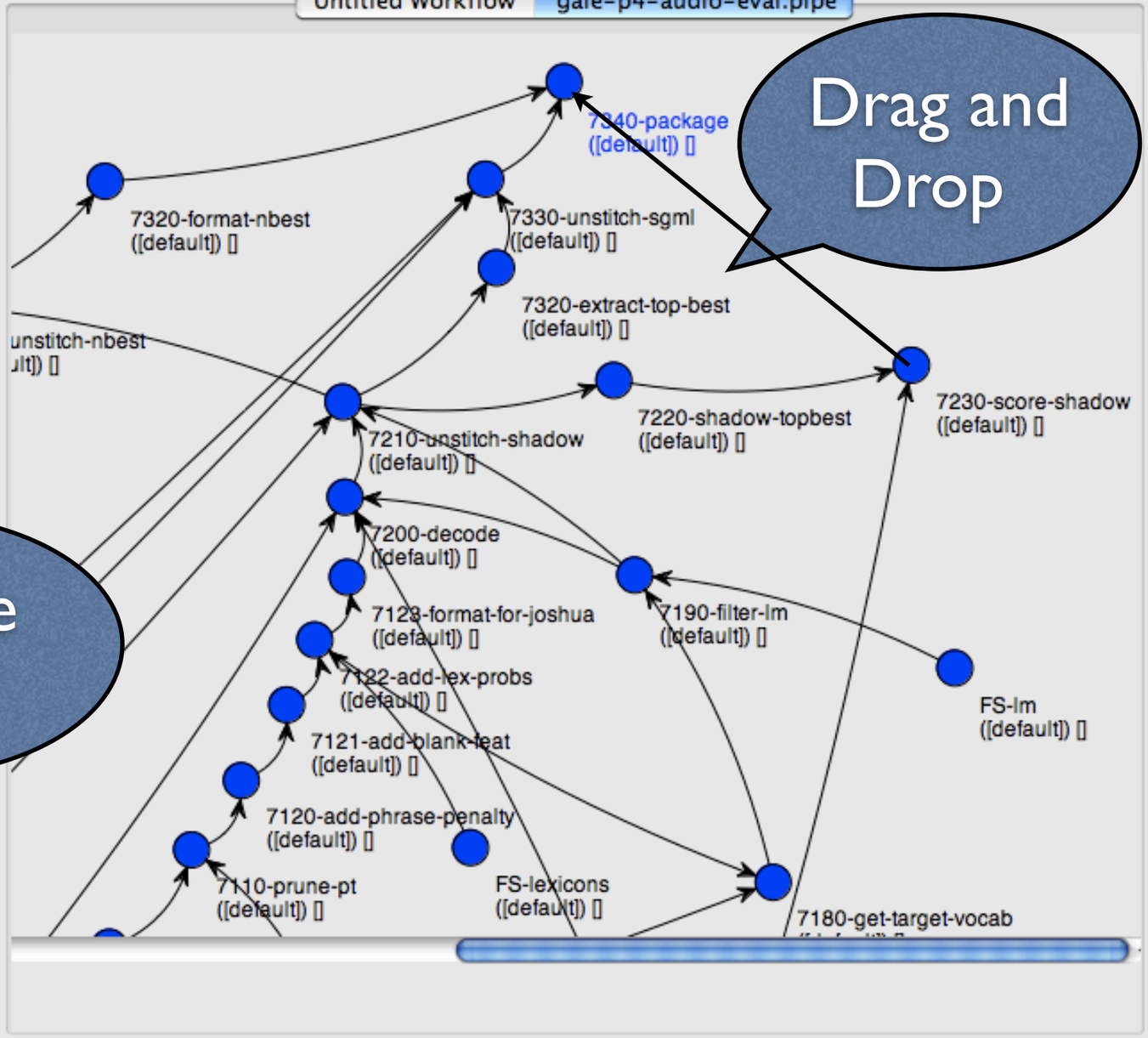
Pipeline Options

Mouse Mode | Scrolling | Selecting | Editing | Transforming Mode

- Tools
  - MANUAL FILESYSTEM
  - MANUAL HDFS
  - OR
  - PARAMETER BOX
- Machine Translation
  - Decoders
  - Grammars and Tables
  - Language Modeling
  - Mono Corpus
  - Output
  - Parallel Corpus
  - Parsing
  - Scoring
  - Tuning
  - Word Alignment

Available Tools

Untitled Workflow | gale-p4-audio-eval.pipe



Drag and Drop

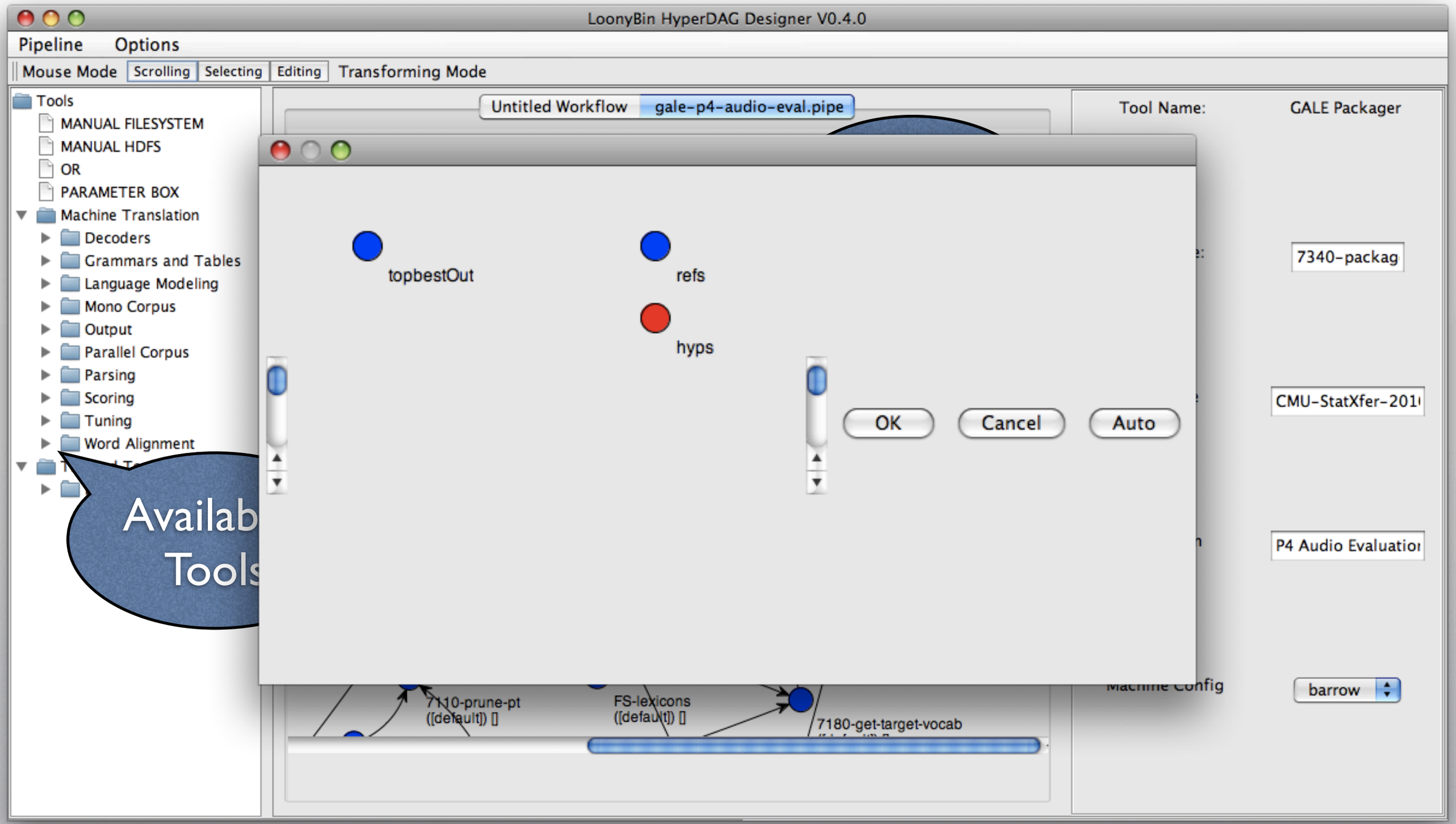
Tool Name: GALE Packager

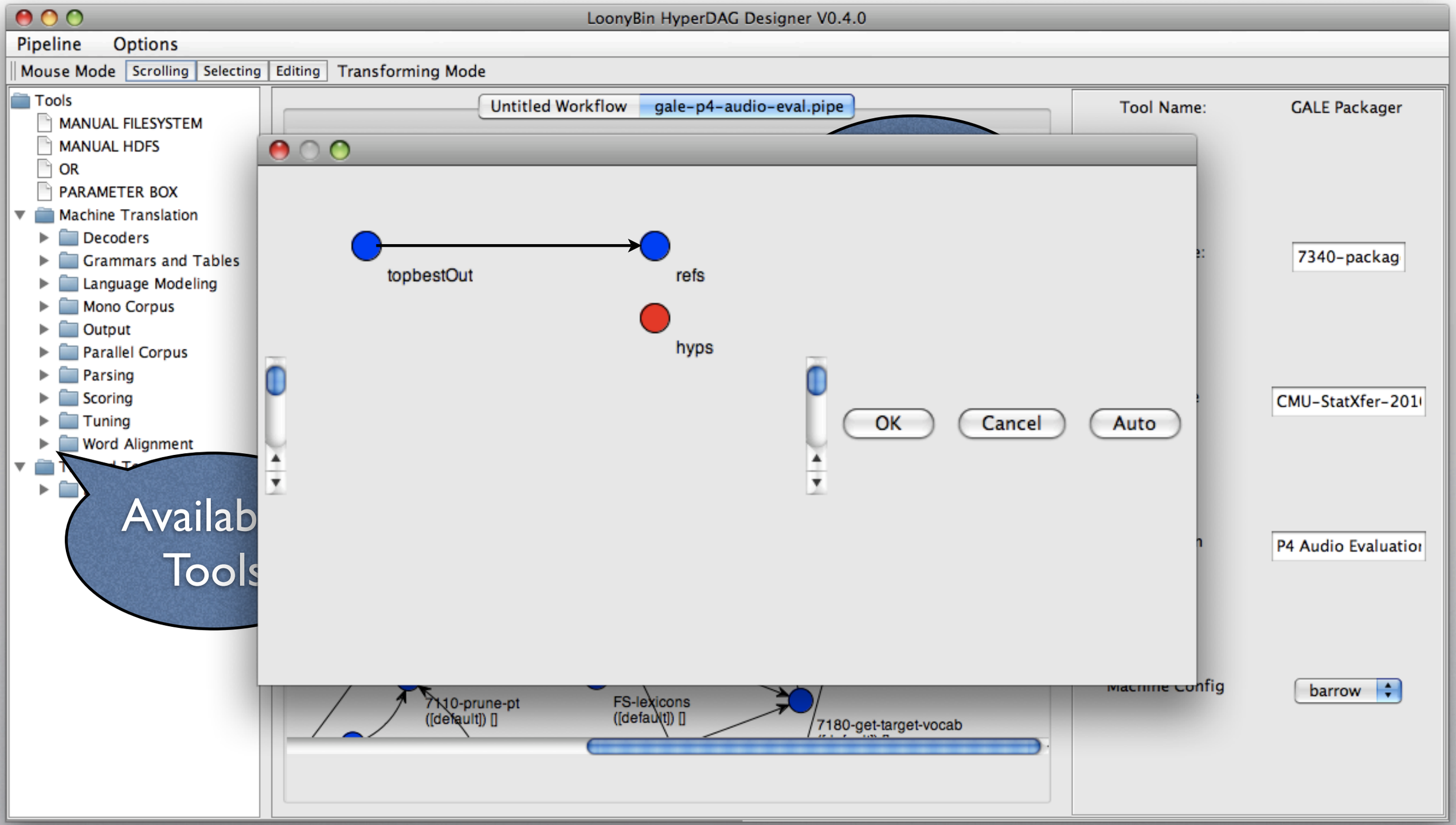
Step Name: 7340-packag

sysName: CMU-StatXfer-201

occassion: P4 Audio Evaluatio

Machine Config: barrow


LoonyBin HyperDAG Designer V0.4.0
Pipeline
Options
Mouse Mode
Scrolling
Selecting
Editing
Transforming Mode
Tools
MANUAL FILESYSTEM
MANUAL HDFS
OR
PARAMETER BOX
Machine Translation
Decoders
Grammars and Tables
Language Modeling
Mono Corpus
Output
Parallel Corpus
Parsing
Scoring
Tuning
Word Alignment
Availab
Tools
Untitled Workflow
gale-p4-audio-eval.pipe
topbestOut
refs
hyps
OK
Cancel
Auto
7110-prune-pt
([default]) []
FS-lexicons
([default]) []
7180-get-target-vocab
Tool Name:
GALE Packager
7340-packag
CMU-StatXfer-201
P4 Audio Evaluatio
barrow


LoonyBin HyperDAG Designer V0.4.0
Pipeline
Options
Mouse Mode
Scrolling
Selecting
Editing
Transforming Mode
Tools
MANUAL FILESYSTEM
MANUAL HDFS
OR
PARAMETER BOX
Machine Translation
Decoders
Grammars and Tables
Language Modeling
Mono Corpus
Output
Parallel Corpus
Parsing
Scoring
Tuning
Word Alignment
Availab
Tools
Untitled Workflow
gale-p4-audio-eval.pipe
topbestOut
refs
hyps
OK
Cancel
Auto
Tool Name:
GALE Packager
7340-packag
CMU-StatXfer-201
P4 Audio Evaluatio
barrow
7110-prune-pt
([default]) []
FS-lexicons
([default]) []
7180-get-target-vocab

LoonyBin HyperDAG Designer V0.4.0

Pipeline Options

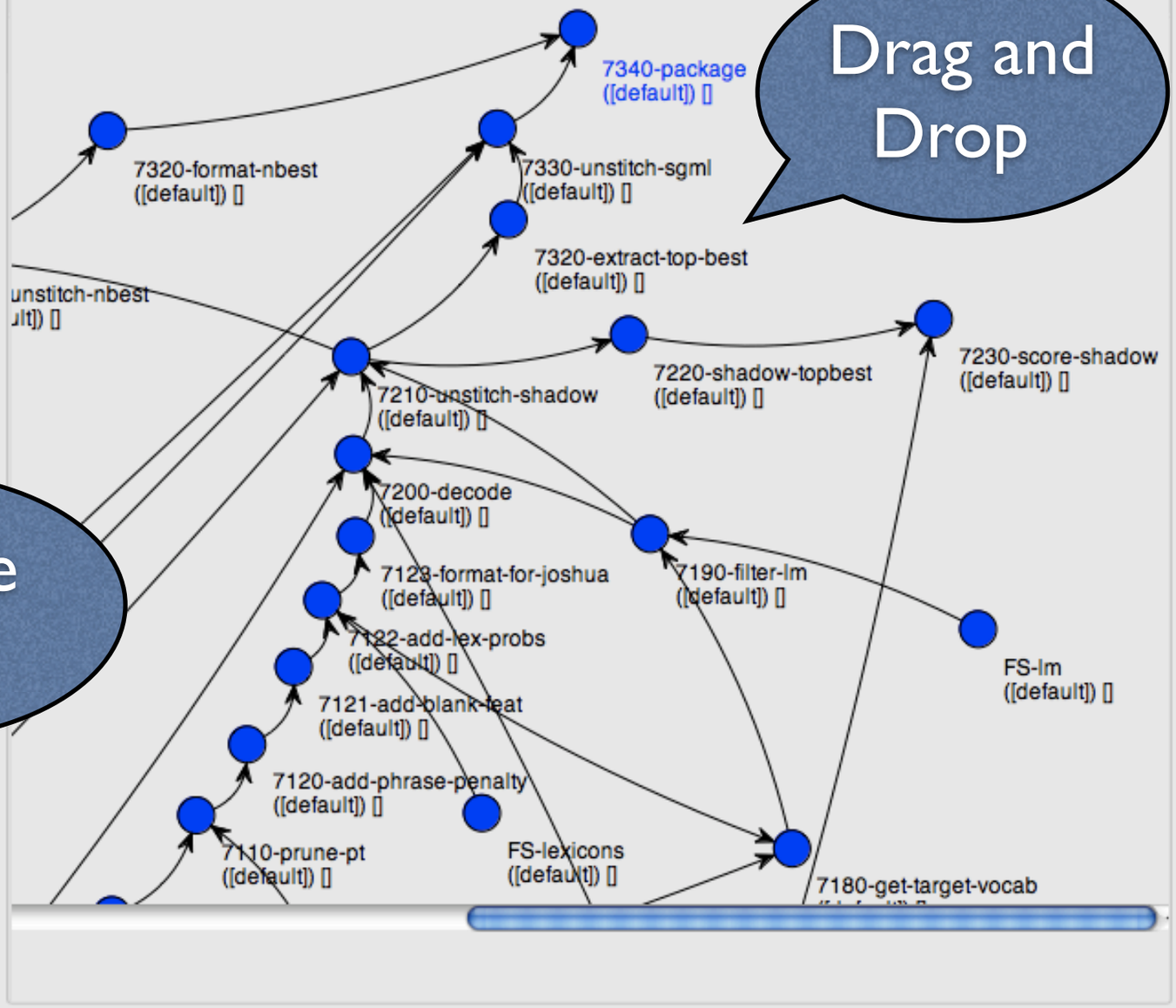
Mouse Mode | Scrolling | Selecting | Editing | Transforming Mode

- Tools
  - MANUAL FILESYSTEM
  - MANUAL HDFS
  - OR
  - PARAMETER BOX
- Machine Translation
  - Decoders
  - Grammars and Tables
  - Language Modeling
  - Mono Corpus
  - Output
  - Parallel Corpus
  - Parsing
  - Scoring
  - Tuning
  - Word Alignment

Available Tools

Untitled Workflow | gale-p4-audio-eval.pipe

Drag and Drop



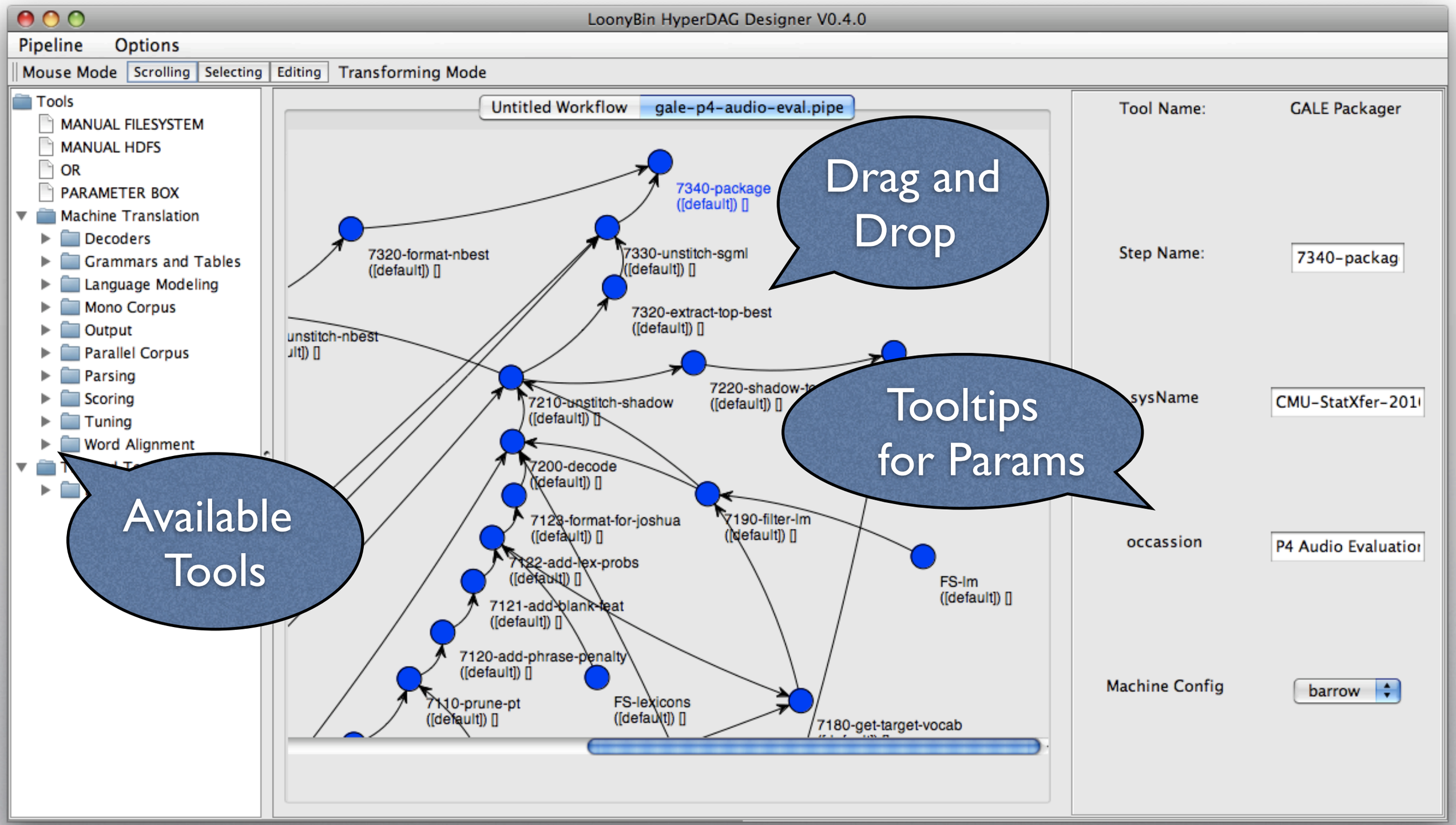
Tool Name: GALE Packager

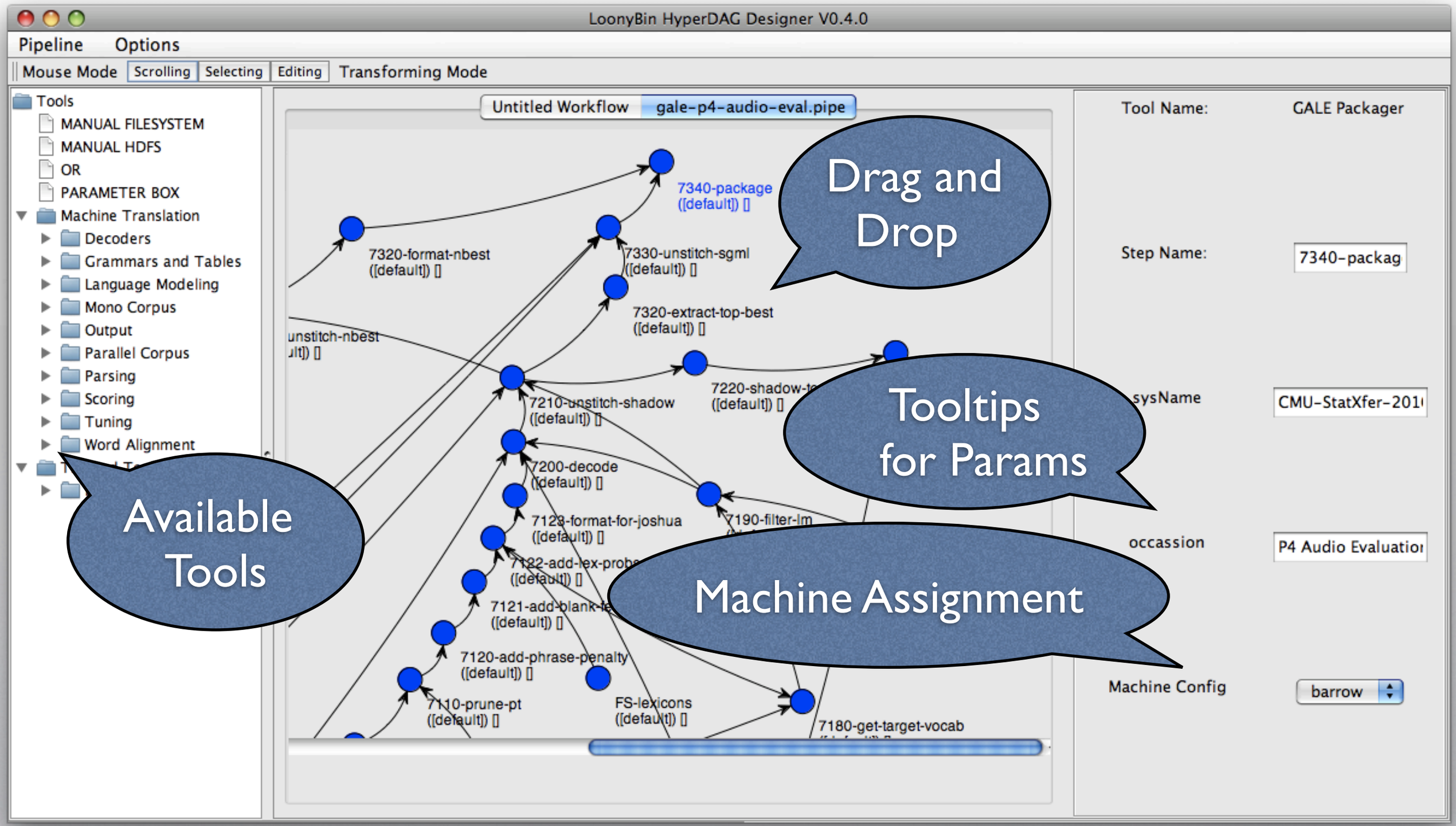
Step Name: 7340-packag

sysName: CMU-StatXfer-201

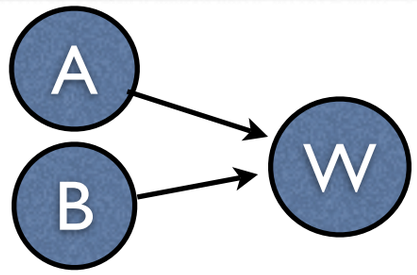
occassion: P4 Audio Evaluatio

Machine Config: barrow


LoonyBin HyperDAG Designer V0.4.0
Pipeline
Options
Mouse Mode
Scrolling
Selecting
Editing
Transforming Mode
Tools
MANUAL FILESYSTEM
MANUAL HDFS
OR
PARAMETER BOX
Machine Translation
Decoders
Grammars and Tables
Language Modeling
Mono Corpus
Output
Parallel Corpus
Parsing
Scoring
Tuning
Word Alignment
Untitled Workflow
gale-p4-audio-eval.pipe
7340-package
([default]) []
7320-format-nbest
([default]) []
7330-unstitch-sgml
([default]) []
7320-extract-top-best
([default]) []
unstitch-nbest
ult]) []
7210-unstitch-shadow
([default]) []
7220-shadow-t
([default]) []
7200-decode
([default]) []
7123-format-for-joshua
([default]) []
7190-filter-lm
([default]) []
7122-add-lex-probs
([default]) []
FS-lm
([default]) []
7121-add-blank-feat
([default]) []
7120-add-phrase-penalty
([default]) []
7110-prune-pt
([default]) []
FS-lexicons
([default]) []
7180-get-target-vocab
Drag and Drop
Tooltips for Params
Available Tools
Tool Name:
GALE Packager
Step Name:
7340-packag
sysName
CMU-StatXfer-201
occassion
P4 Audio Evaluatio
Machine Config
barrow


LoonyBin HyperDAG Designer V0.4.0
Pipeline
Options
Mouse Mode
Scrolling
Selecting
Editing
Transforming Mode
Tools
MANUAL FILESYSTEM
MANUAL HDFS
OR
PARAMETER BOX
Machine Translation
Decoders
Grammars and Tables
Language Modeling
Mono Corpus
Output
Parallel Corpus
Parsing
Scoring
Tuning
Word Alignment
Untitled Workflow
gale-p4-audio-eval.pipe
7340-package
([default]) []
7320-format-nbest
([default]) []
7330-unstitch-sgml
([default]) []
7320-extract-top-best
([default]) []
unstitch-nbest
ult]) []
7210-unstitch-shadow
([default]) []
7220-shadow-t
([default]) []
7200-decode
([default]) []
7123-format-for-joshua
([default]) []
7190-filter-lm
7122-add-lex-prob
([default]) []
7121-add-blank-le
([default]) []
7120-add-phrase-penalty
([default]) []
7110-prune-pt
([default]) []
FS-lexicons
([default]) []
7180-get-target-vocab
Drag and Drop
Tooltips for Params
Available Tools
Machine Assignment
Tool Name:
GALE Packager
Step Name:
7340-packag
sysName
CMU-StatXfer-201
occassion
P4 Audio Evaluatio
Machine Config
barrow

Generating a Script for



INPUTS

foreignCorpus

nativeCorpus

PARAMETERS

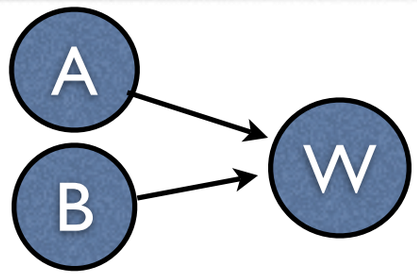
fertility

OUTPUTS

alignments

Python Tool Descriptor

Generating a Script for



INPUTS

foreignCorpus	A's output "x"
nativeCorpus	B's output "y"

PARAMETERS

fertility	0.01

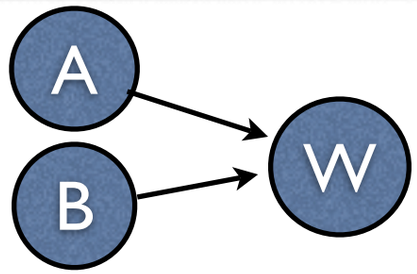
OUTPUTS

alignments

Parameters & dependencies from workflow

Python Tool Descriptor

Generating a Script for



INPUTS

foreignCorpus	A's output "x"	…/inputs/f
nativeCorpus	B's output "y"	…/inputs/n

PARAMETERS

fertility	0.01

OUTPUTS

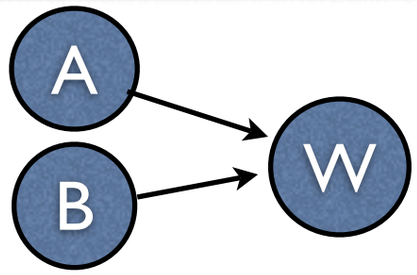
alignments	…/outputs/wa

LoonyBin assigns paths

Parameters & dependencies from workflow

Python Tool Descriptor

Generating a Script for



INPUTS

foreignCorpus	A's output "x"	…/inputs/f
nativeCorpus	B's output "y"	…/inputs/n

PARAMETERS

fertility	0.01

OUTPUTS

alignments	…/outputs/wa

LoonyBin assigns paths

Parameters & dependencies from workflow

Python Tool Descriptor

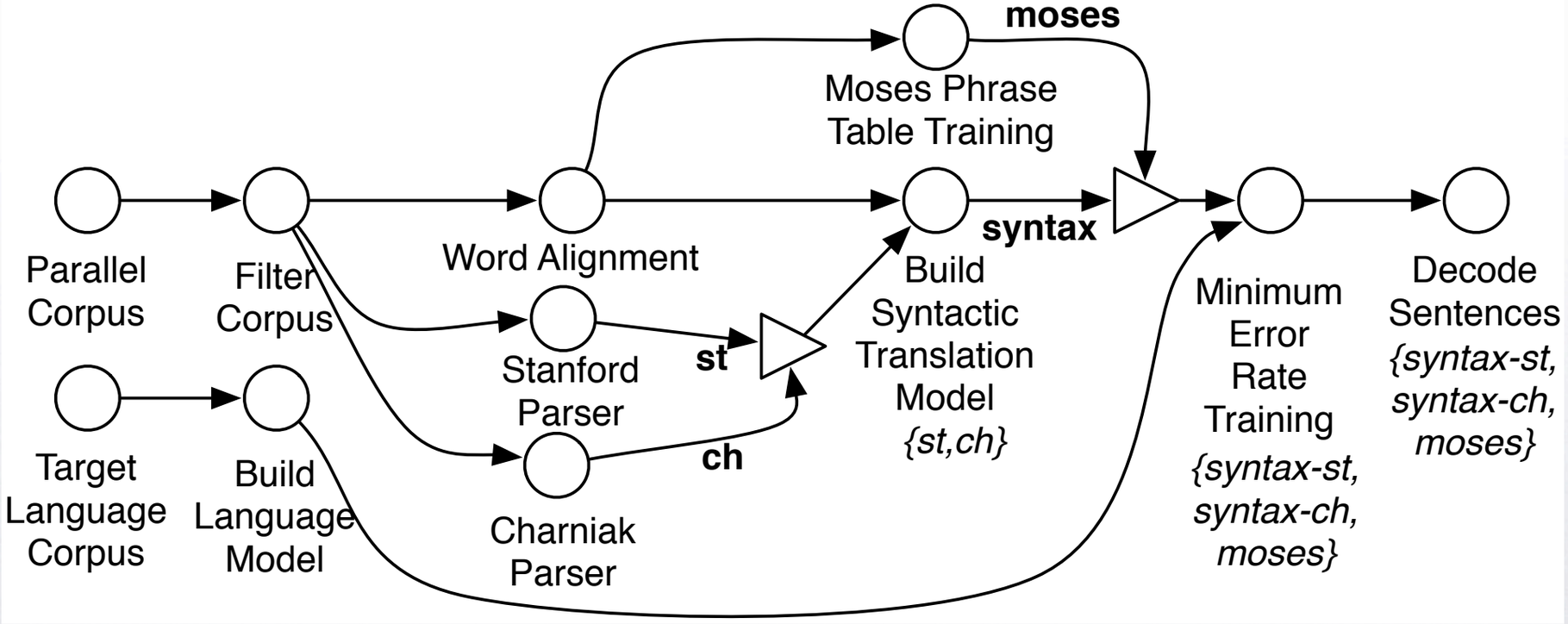
java edu.cmu.Tokenizer ../inputs/f ../inputs/n > ../outputs/wa

So far...

- Complaints about current implementation of empirical NLP experiments
- Define the tools
  (inputs/outputs/parameters)
- Define the workflow
  (DAG of steps and dependencies)
- Generate & run a shell script

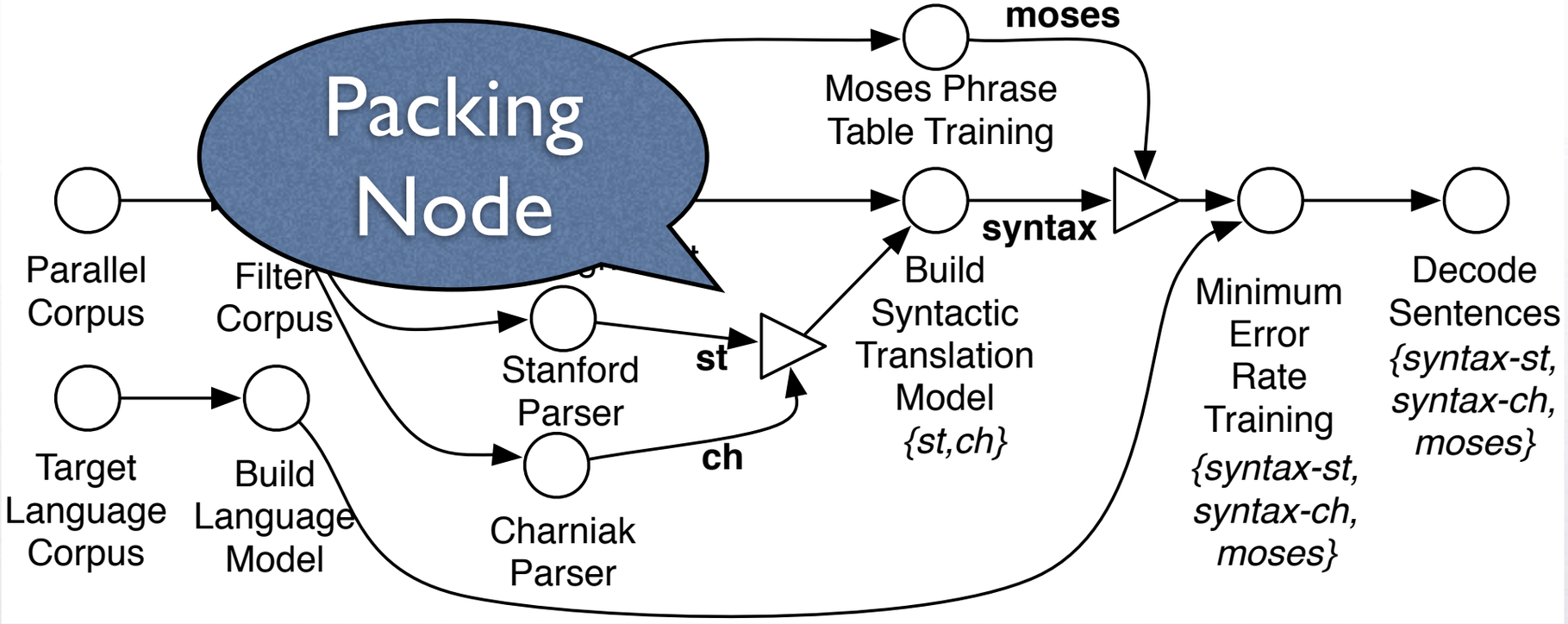
HyperWorkflows

- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG



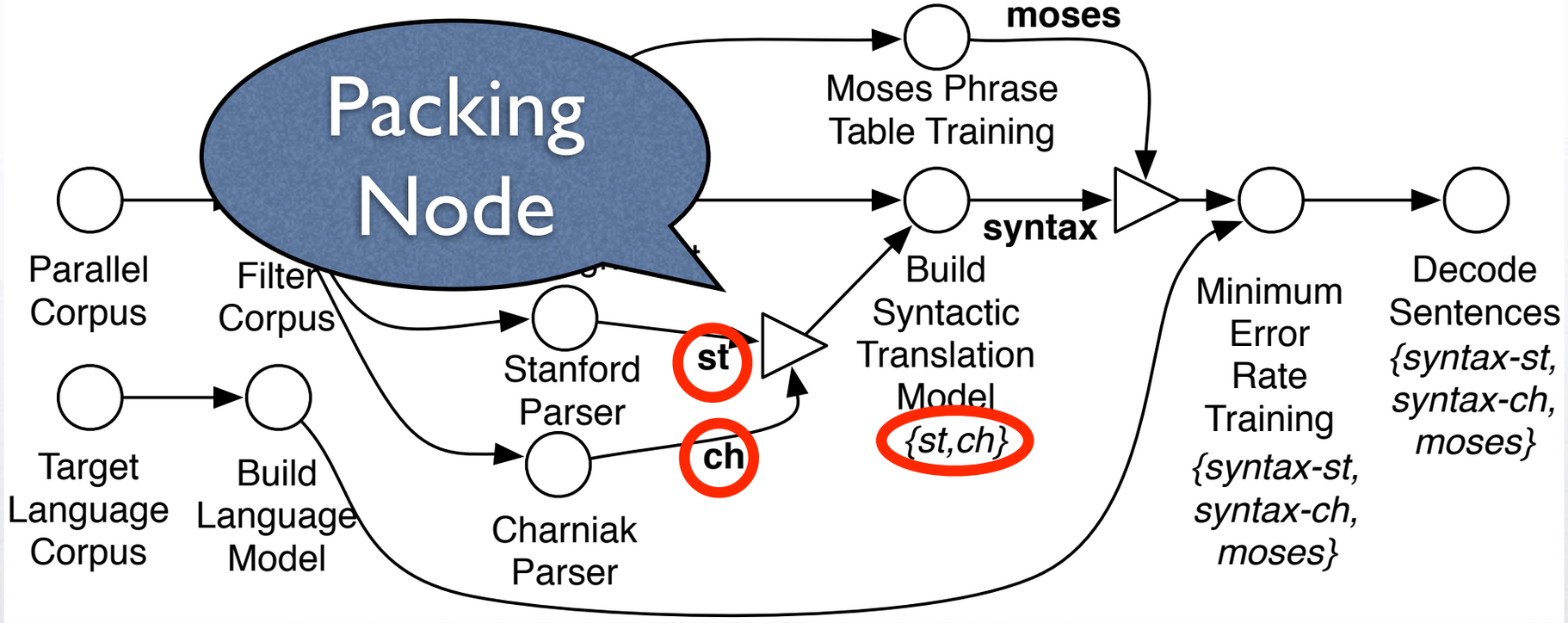
HyperWorkflows

- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG



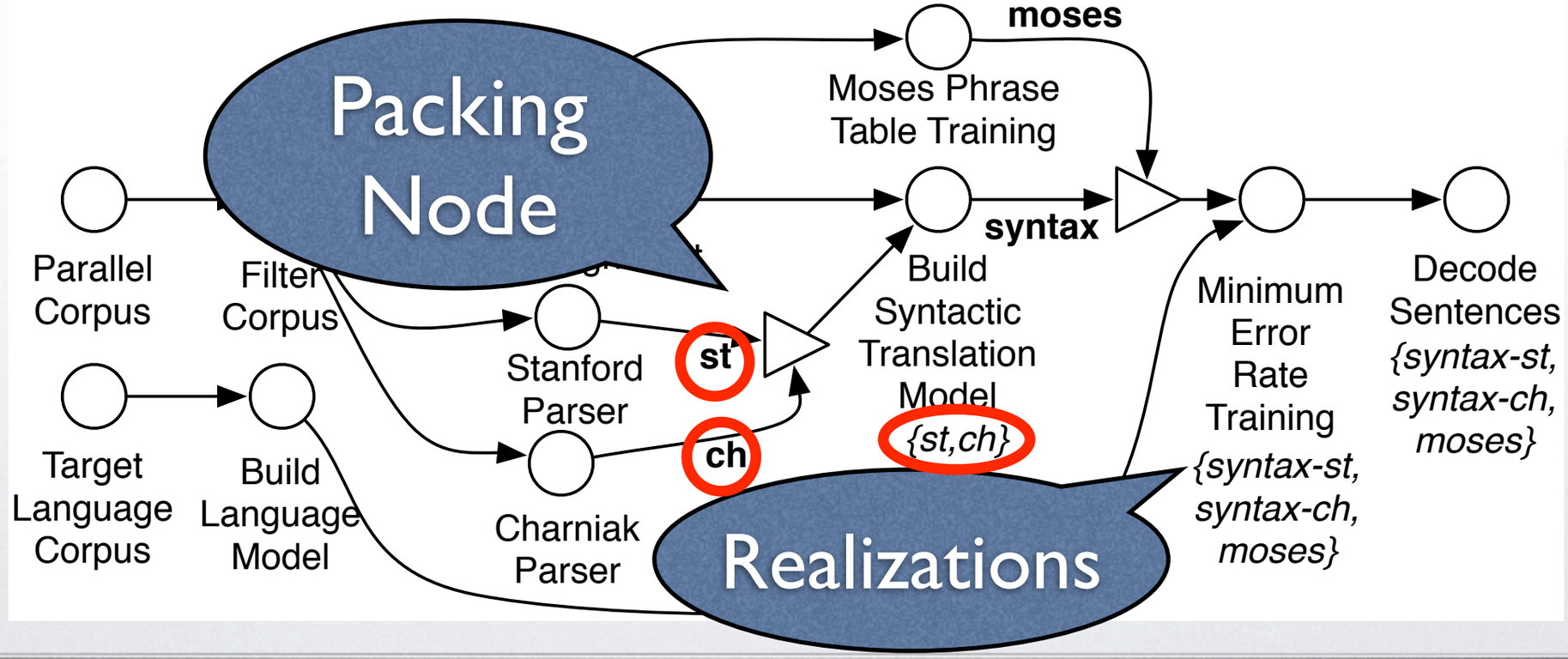
HyperWorkflows

- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG



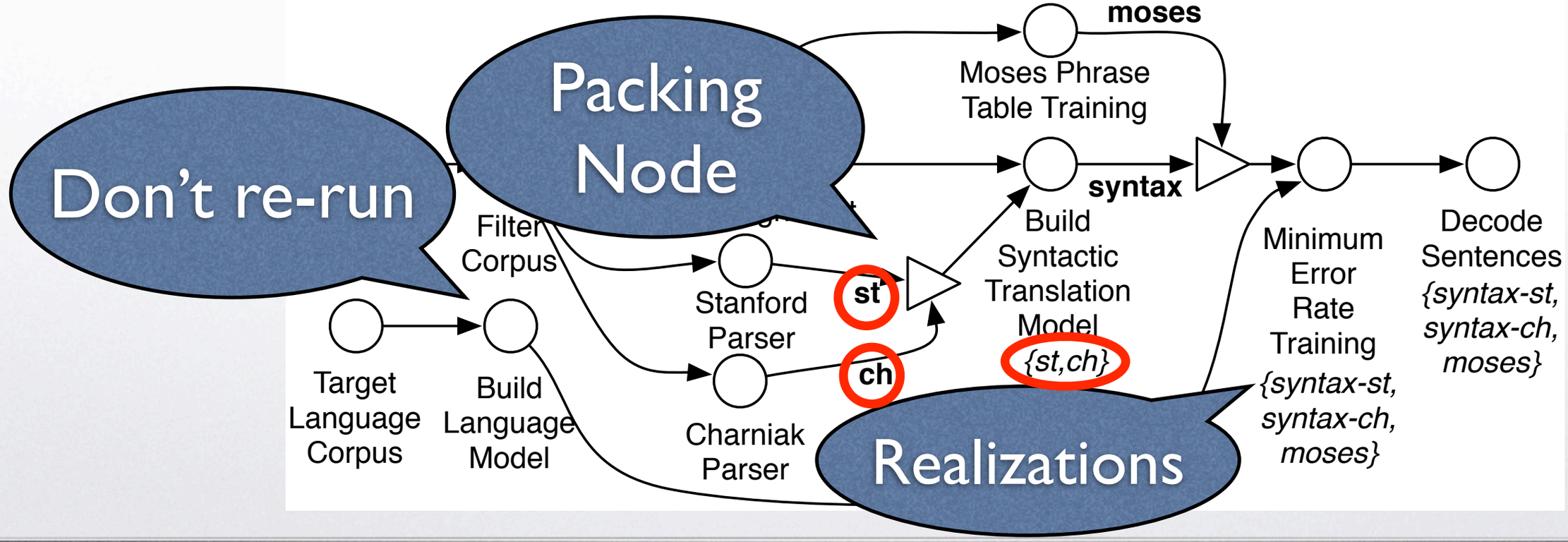
HyperWorkflows

- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG



HyperWorkflows

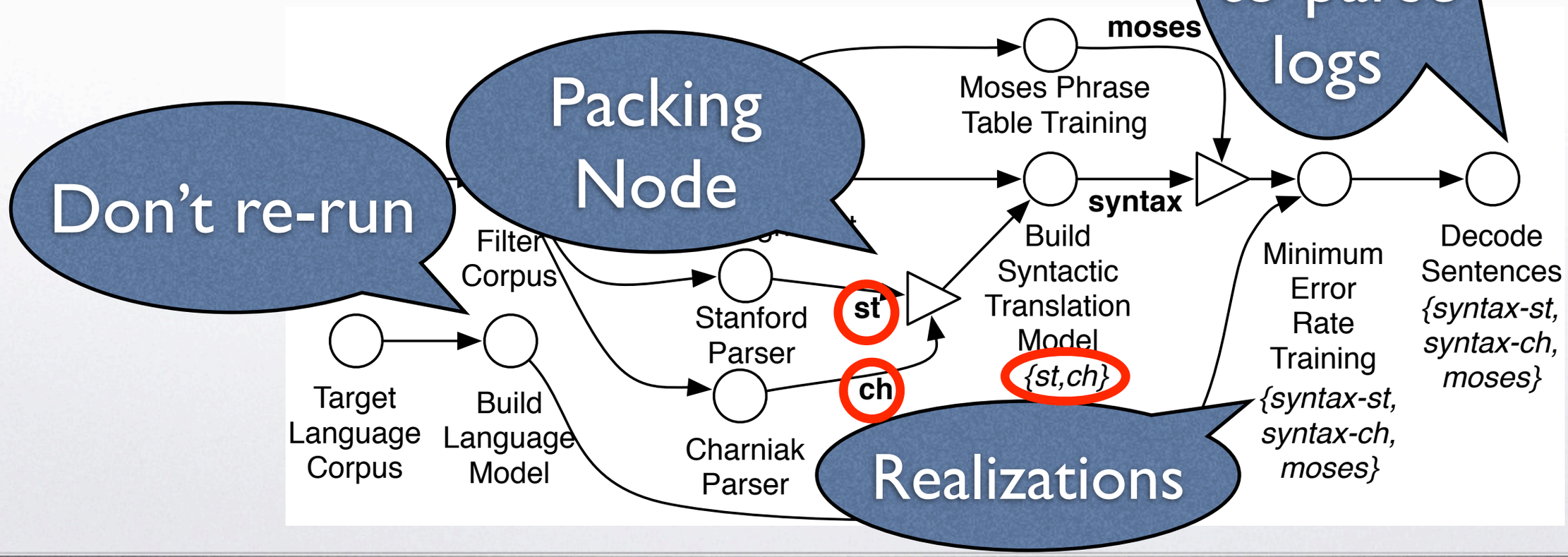
- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG



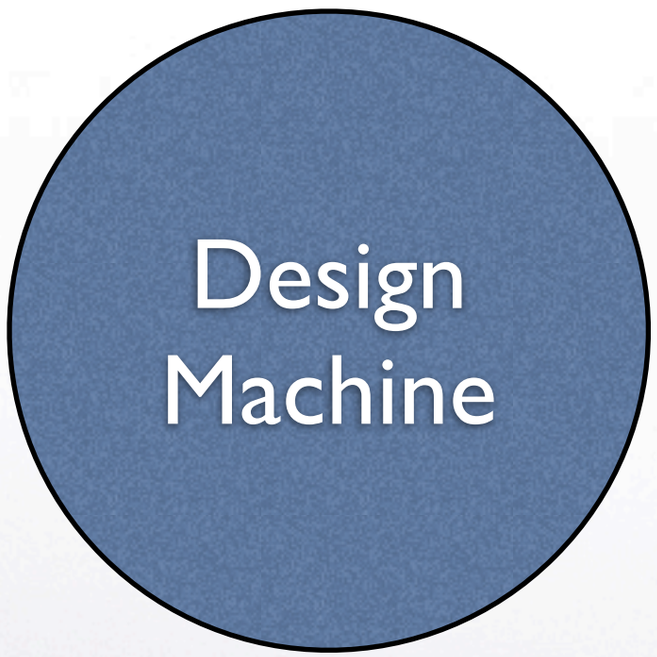
HyperWorkflo

- **HyperWorkflows:** Shared s in experiments
- Encode small variations in a Hyp

Organized directory structure & easy-to-parse logs

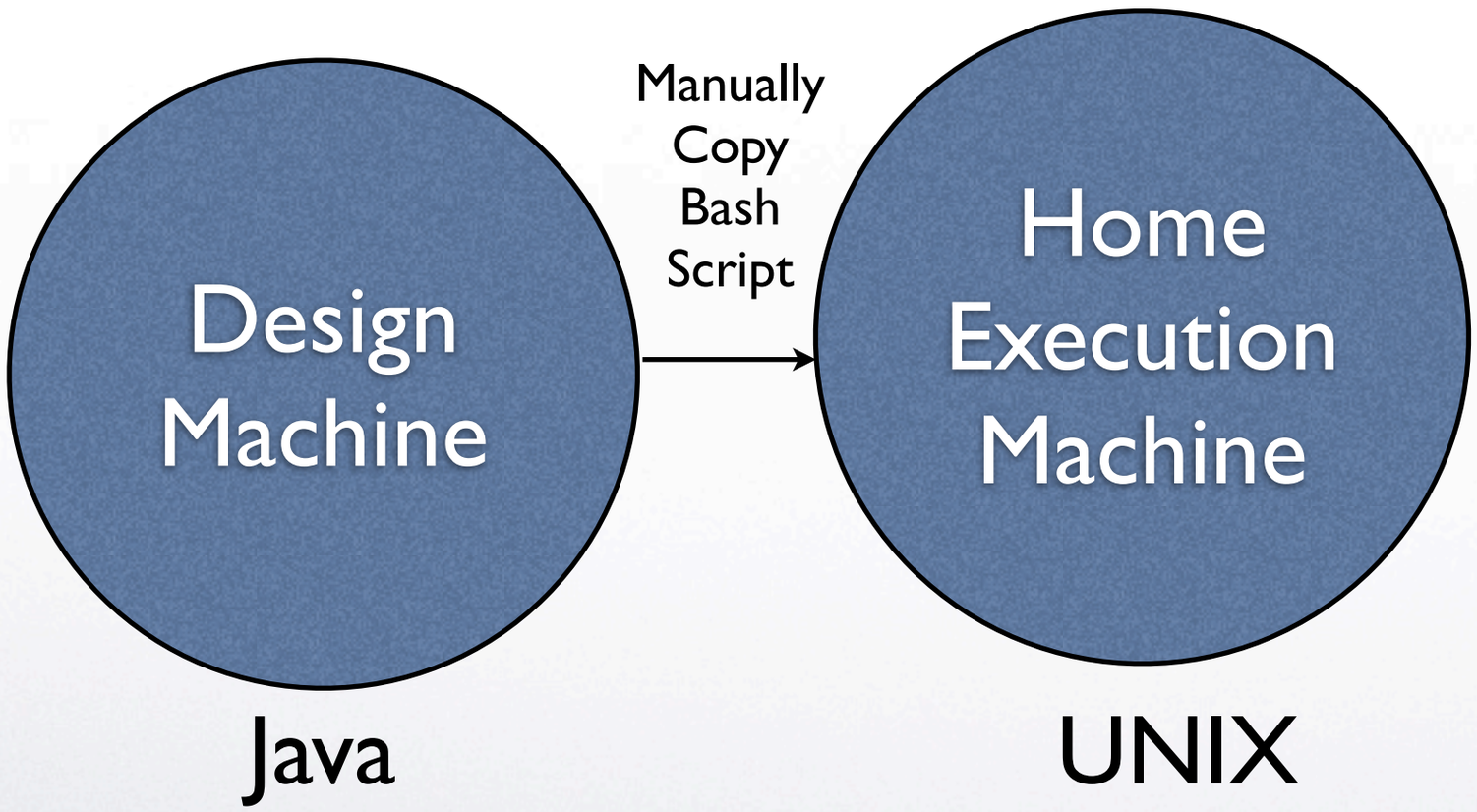


Multiple Machines and Schedulers

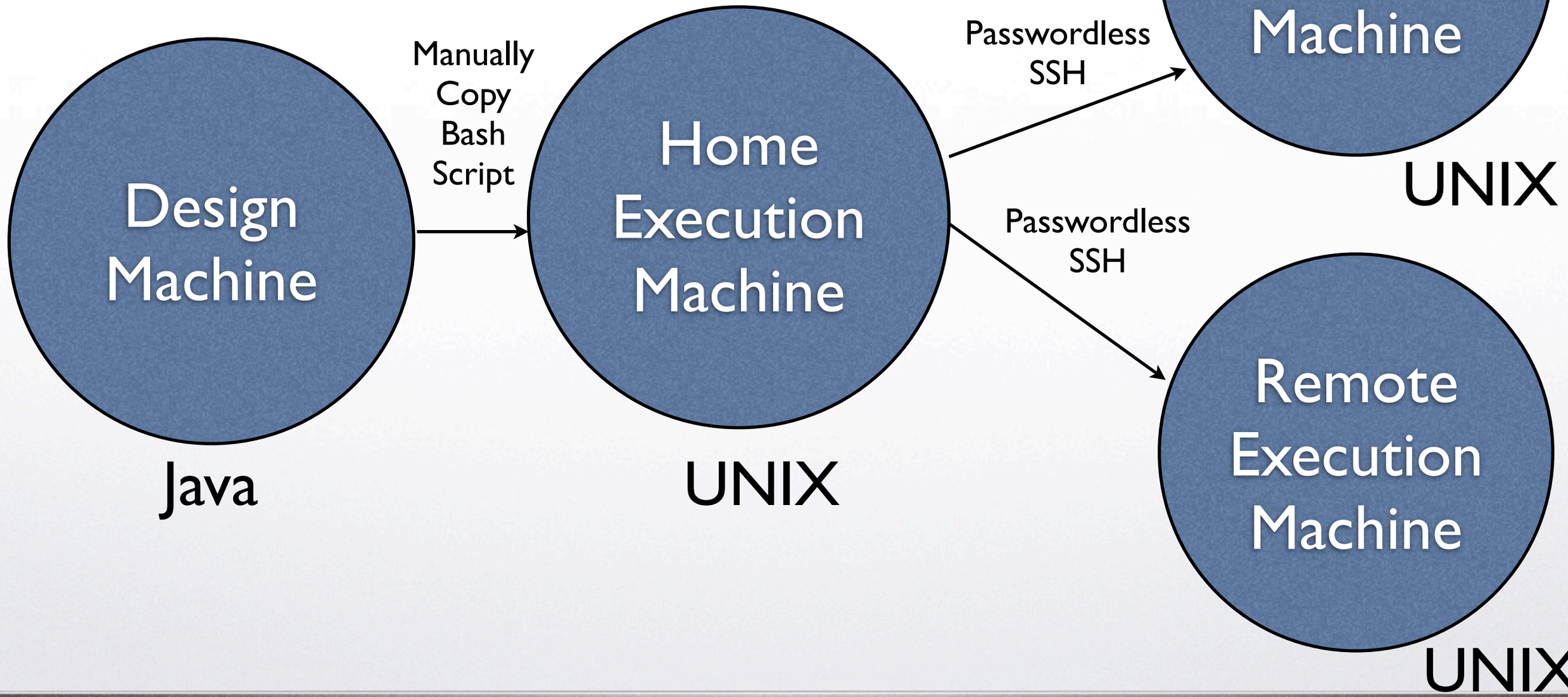


Java

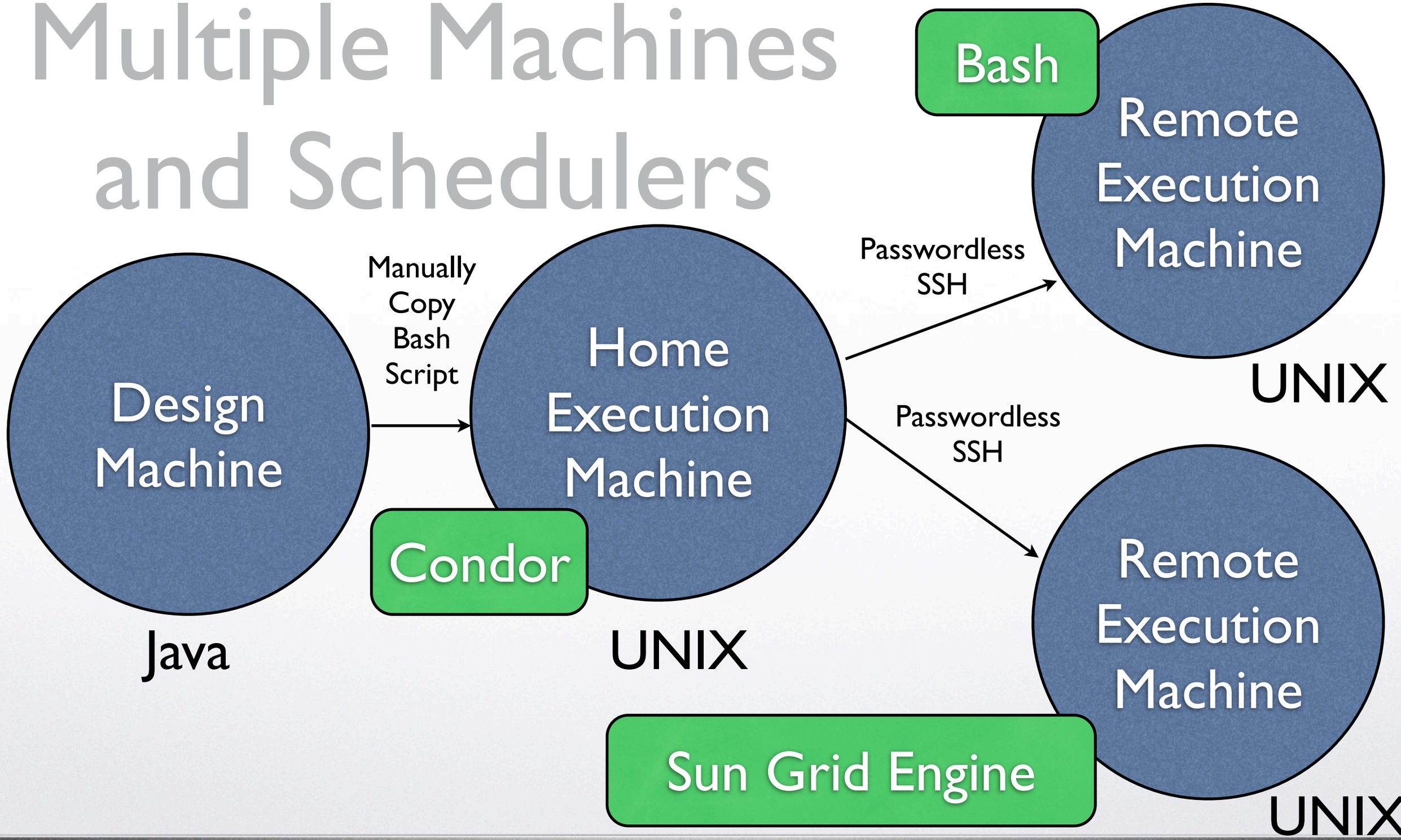
Multiple Machines and Schedulers



Multiple Machines and Schedulers



Multiple Machines and Schedulers



Other Things to Make Life Easier

- Sanity checking at each step (embedded in Tool Descriptors)
- Copying of files (including to HDFS)
- Text-based workflow definition (in SVN)
- Open-source LGPL License

WANTED

- Users & Contributors
- Machine Translation Toolpack (released)
- Corpus Processing Toolpack?
- Parsing Toolpack?
- Question Answering Toolpack?
- Resource Directory Toolpack?
- Speech Recognition Toolpack?

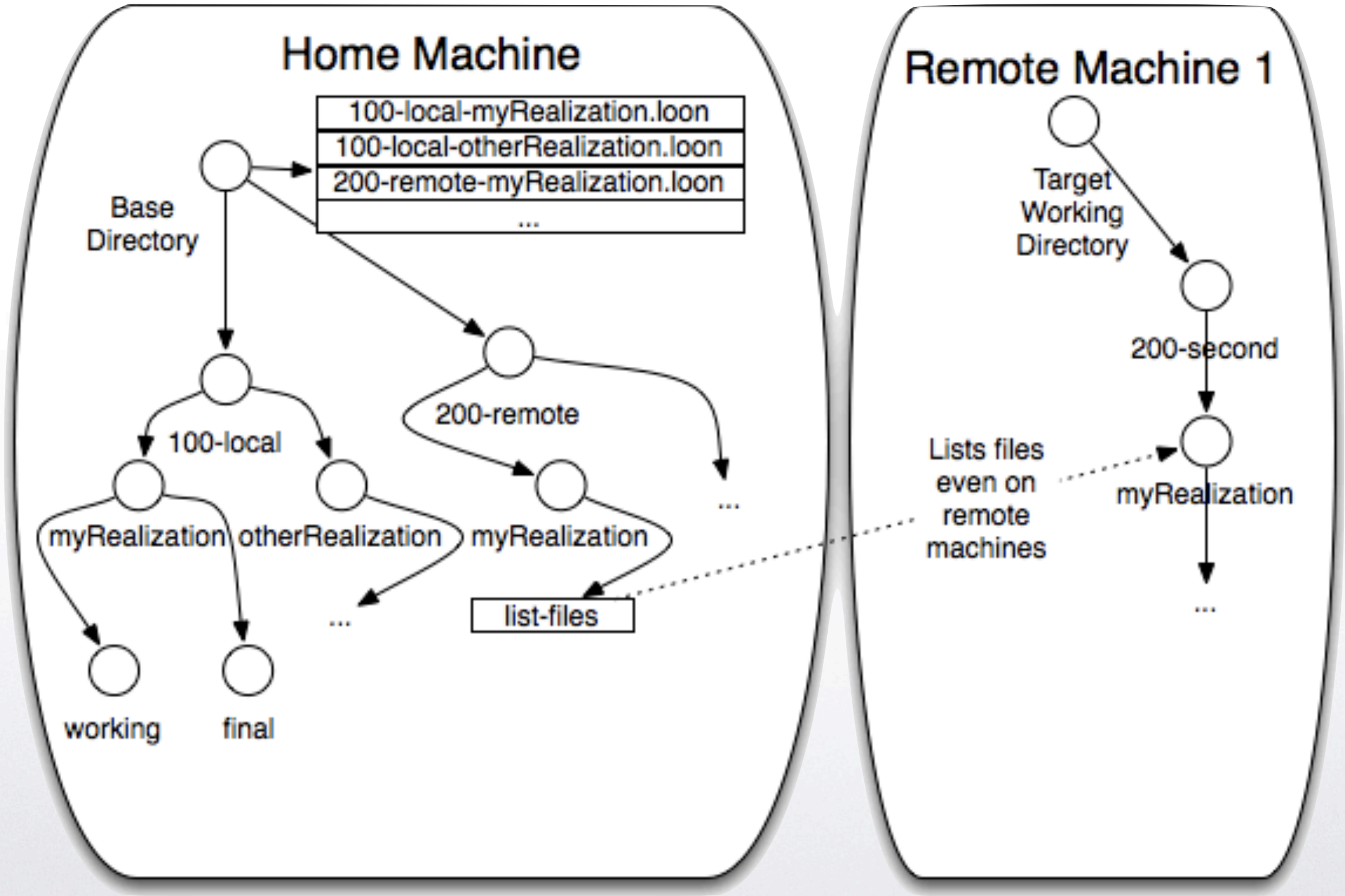
Conclusion

- Make your life easier
  - Automation
  - Sanity Checking
  - Logging
- Make your colleagues' lives easier
  - Reproducibility
  - Modularity

Questions?

Tutorial & Software at

http://loonybin.sourceforge.net


Home Machine
100-local-myRealization.loon
100-local-otherRealization.loon
200-remote-myRealization.loon
...
Base Directory
100-local
200-remote
...
myRealization
otherRealization
myRealization
...
list-files
working
final
Remote Machine 1
Target Working Directory
200-second
Lists files even on remote machines
myRealization
...

Practical Issues

- How does LoonyBin know when to run/ rerun a step?
  - Each vertex x realization has a loon log file. If the file does not exist, the step is (re)run
- What if I don't want that many steps?
  - Workflows have many granularities!

Recommendations

- Store your workflow files in SVN
- Store your log files in SVN -- experimental data is useful long after we get annoyed with size of data files!
- Log the SVN revision of frequently changing tools in your Loon logs -- Build them from SVN every time to ensure you're executing that version

Future Work

- Default parameters -- Short-term
- Asynchronous DAG execution (currently all steps are run in serial) -- Mid-Term
- Workflow monitoring and reprioritization during execution -- Long-term
- Encapsulation of workflows as “tools” (hierarchical tools) -- Long-term
- Automatic file compression -- Long-term