

The Machine Translation Toolpack for LoonyBin

Machine Translation and HyperWorkflows

Jonathan Clark, Jonathan Weese,
Byung Gyu Ahn, Andreas Zollmann, Qin Gao,
Kenneth Heafield, Alon Lavie

4th Machine Translation Marathon
Monday, January 25, 2009



Outline

- A (Brief) Guilt Trip
- LoonyBin
 - What goes in
 - What goes on
 - What comes out
- The MT Toolpack for LoonyBin
- A Few Recommendations

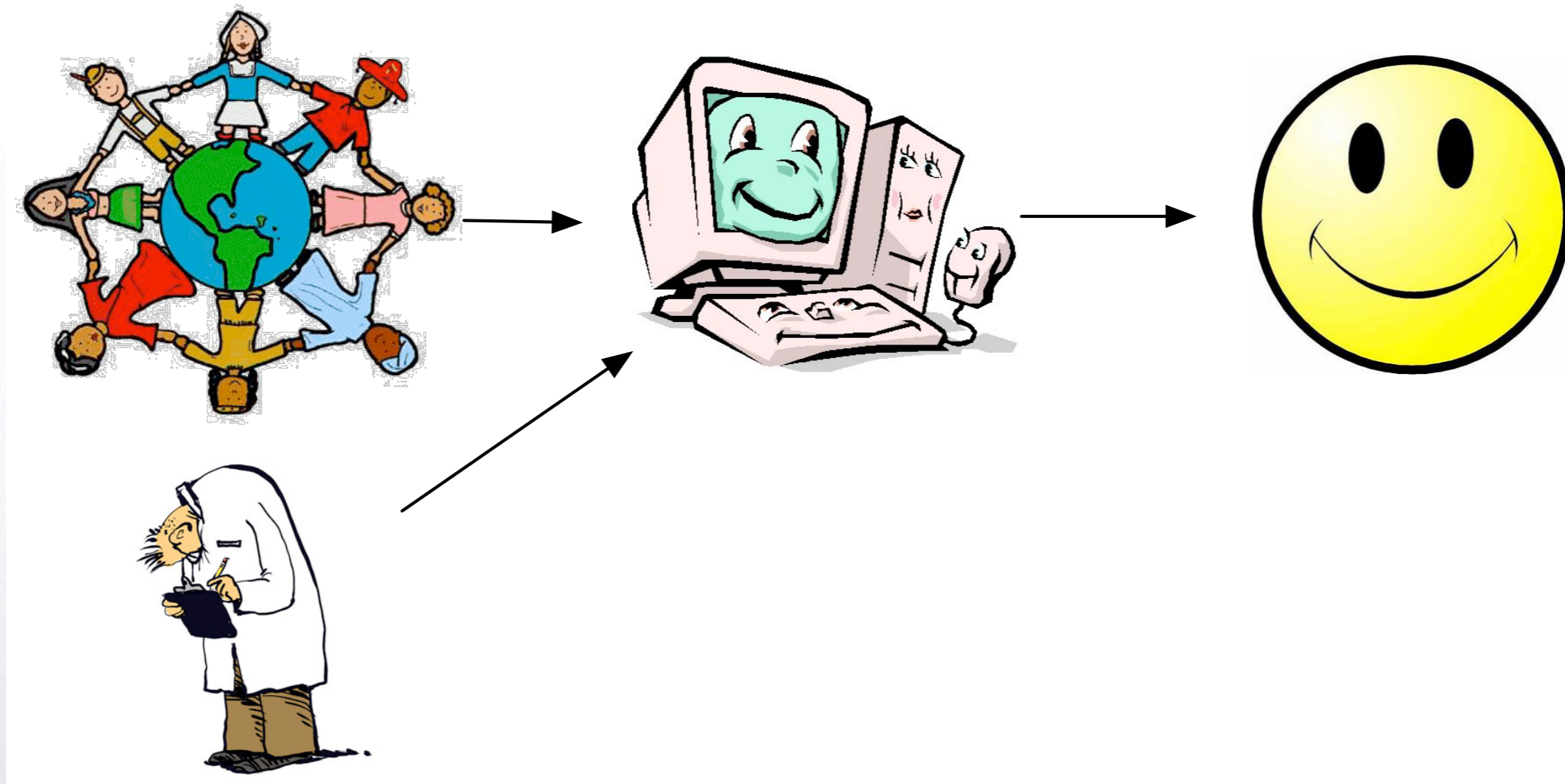


The Guilt Trip

- What does an ideal experiment look like?
 - Small Δ 's, Reproducible, Detailed Analysis/Logs...
- What do most MT experiments look like?
 - Convoluted scripts, lazy evaluation of data analysis...



MT Workflows in Papers





Actual MT Workflows

```
if ($_HELP) {
    print "Train Phrase Model

Steps: (--first-step to --last-step)
(1) prepare corpus
(2) run GIZA
(3) align words
(4) learn lexical translation
(5) extract phrases
(6) score phrases
(7) learn reordering model
(8) learn generation model
(9) create decoder config file

For more, please check manual or contact koehn\@inf.ed.ac.uk\n";
    exit(1);
}

my $___FACTOR_DELIMITER = $_FACTOR_DELIMITER;
$_FACTOR_DELIMITER = '|' unless ($_FACTOR_DELIMITER);

print STDERR "Using SCRIPTS_ROOTDIR: $SCRIPTS_ROOTDIR\n";

# supporting binaries from other packages
my $GIZA = "$BINDIR/GIZA++";
my $SNT2COOC = "$BINDIR/snt2cooc.out";
```



Issues

- Automation
- Reproducibility
- Variability
- Scripting Bugs
- Multiple machines, clusters, and schedulers
- Hard to see Big Picture



What goes into LoonyBin



Going in

- Knowledge from self 6 months ago
- Knowledge from predecessor 8 years ago about removing the 300-character underscore out of the corpus
- Visual representation of input/output files and parameters as a DAG

Obligatory Screenshot

The screenshot displays the LoonyBin HyperDAG Designer V0.4.0 interface. The main window shows a workflow diagram titled "Untitled Workflow" with the file name "gale-p4-audio-eval.pipe". The workflow consists of numerous nodes connected by arrows, representing a sequence of processing steps. A callout bubble labeled "Available Tools" points to a sidebar on the left containing a tree view of tool categories such as "MANUAL FILESYSTEM", "MANUAL HDFS", "OR", "PARAMETER BOX", and "Machine Translation". Another callout bubble labeled "Drag and Drop" points to a node in the workflow. A third callout bubble labeled "Tooltips" points to the right-hand panel, which displays configuration details for a selected tool, including "Tool Name: GALE Packager", "Step Name: 7340-packag", "sysName: CMU-StatXfer-201", "occassion: P4 Audio Evaluation", and "Machine Config: barrow". A fourth callout bubble labeled "Machine/Scheduler" points to a dropdown menu at the bottom of the workflow diagram.

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))

7330-unstitch-sgml ((default))

7320-extract-top-best ((default))

7320-format-nbest ((default))

unstitch-nbest ((default))

7230-score-shadow ((default))

7220-shadow-topbest ((default))

7210-unstitch-shadow ((default))

7200-decode ((default))

7190-filter-lm ((default))

7128-format-for-joshua ((default))

7122-add-lex-probs ((default))

7121-add-blank-leaf ((default))

7120-add-phrase-penalty ((default))

7110-prune-pt ((default))

FS-lm ((default))

FS-lexicons ((default))

7180-get-target-vocab ((default))

Tool Name: GALE Packager

Step Name: 7340-packag

sysName: CMU-StatXfer-201

occassion: P4 Audio Evaluation

Machine Config: barrow

Available Tools

Drag and Drop

Tooltips

Machine/Scheduler

Obligatory Screenshot

The screenshot displays the LoonyBin HyperDAG Designer V0.4.0 interface. At the top, the title bar reads "LoonyBin HyperDAG Designer V0.4.0". Below it, a menu bar includes "Pipeline" and "Options". A toolbar shows "Mouse Mode" with sub-modes: "Scrolling", "Selecting", "Editing", and "Transforming Mode".

On the left, a "Tools" sidebar lists various components: "MANUAL FILESYSTEM", "MANUAL HDFS", "OR", "PARAMETER BOX", and a "Machine Translation" folder containing "Decoders", "Grammars and Tables", "Language Modeling", "Mono Corpus", "Output", "Parallel Corpus", "Parsing", "Scoring", "Tuning", and "Word Alignment".

The main workspace shows a workflow diagram with nodes: "topbestOut" (blue circle), "refs" (blue circle), and "hyps" (red circle). Below this, a modal dialog box is open, containing three buttons: "OK", "Cancel", and "Auto".

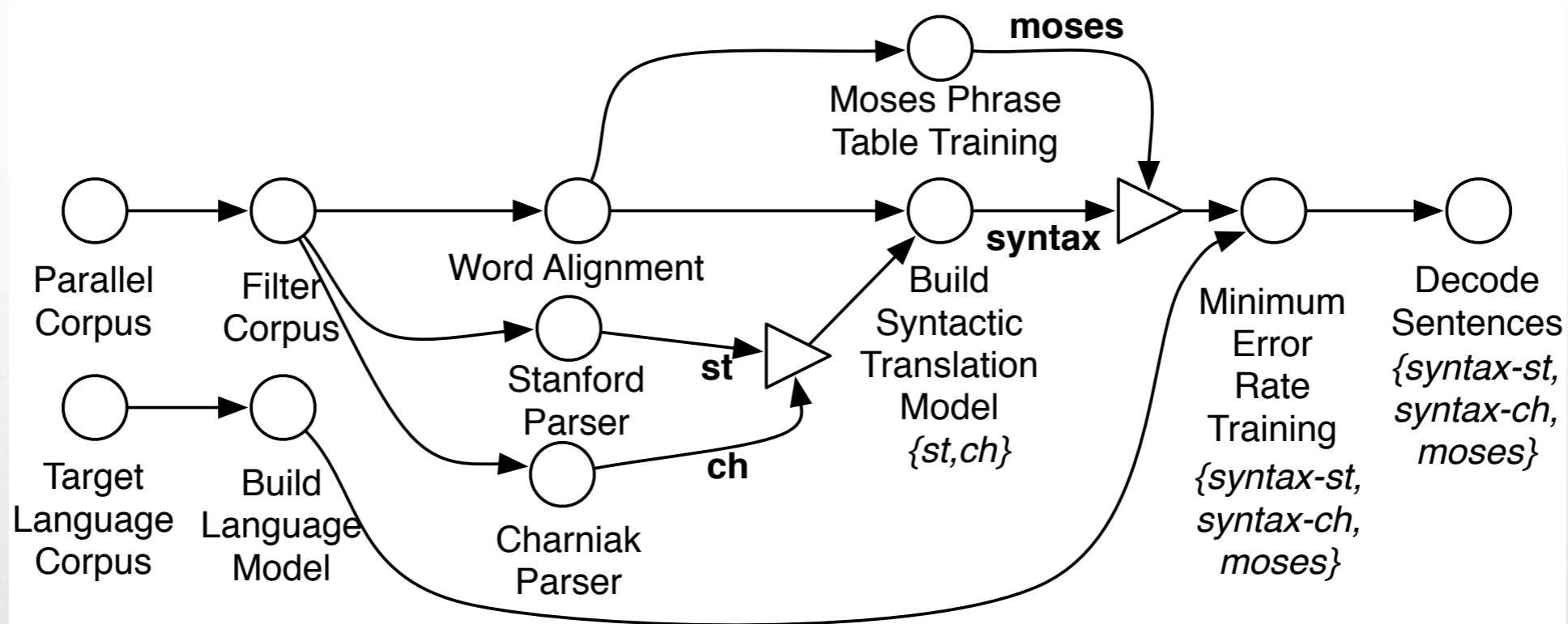
On the right, a "Machine Config" panel shows a list of configurations: "GALE Packager", "7340-packag", "CMU-StatXfer-201", "P4 Audio Evaluation", and "barrow".

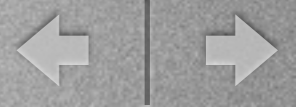
Three blue speech bubble annotations are present: "Available Tools" points to the Tools sidebar; "tips" points to the modal dialog; and "Machine/Scheduler" points to the Machine Config panel.



HyperWorkflows

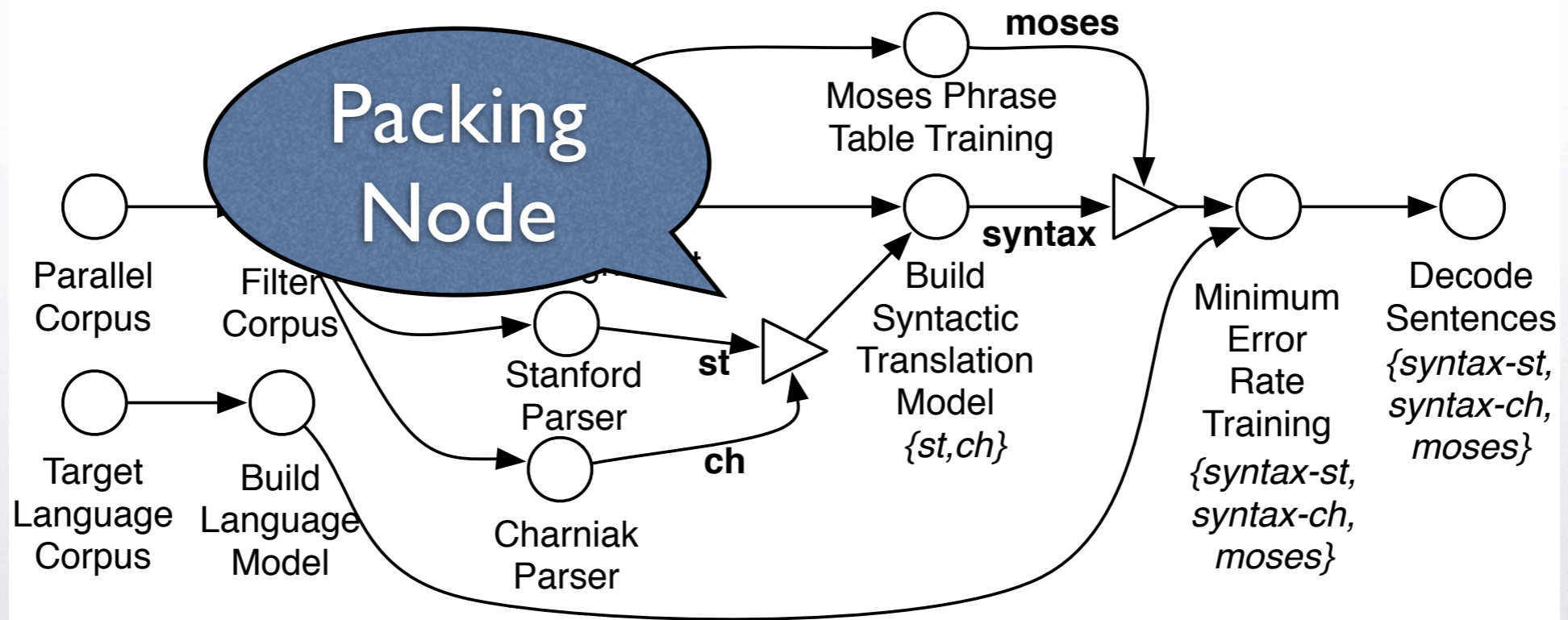
- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG





HyperWorkflows

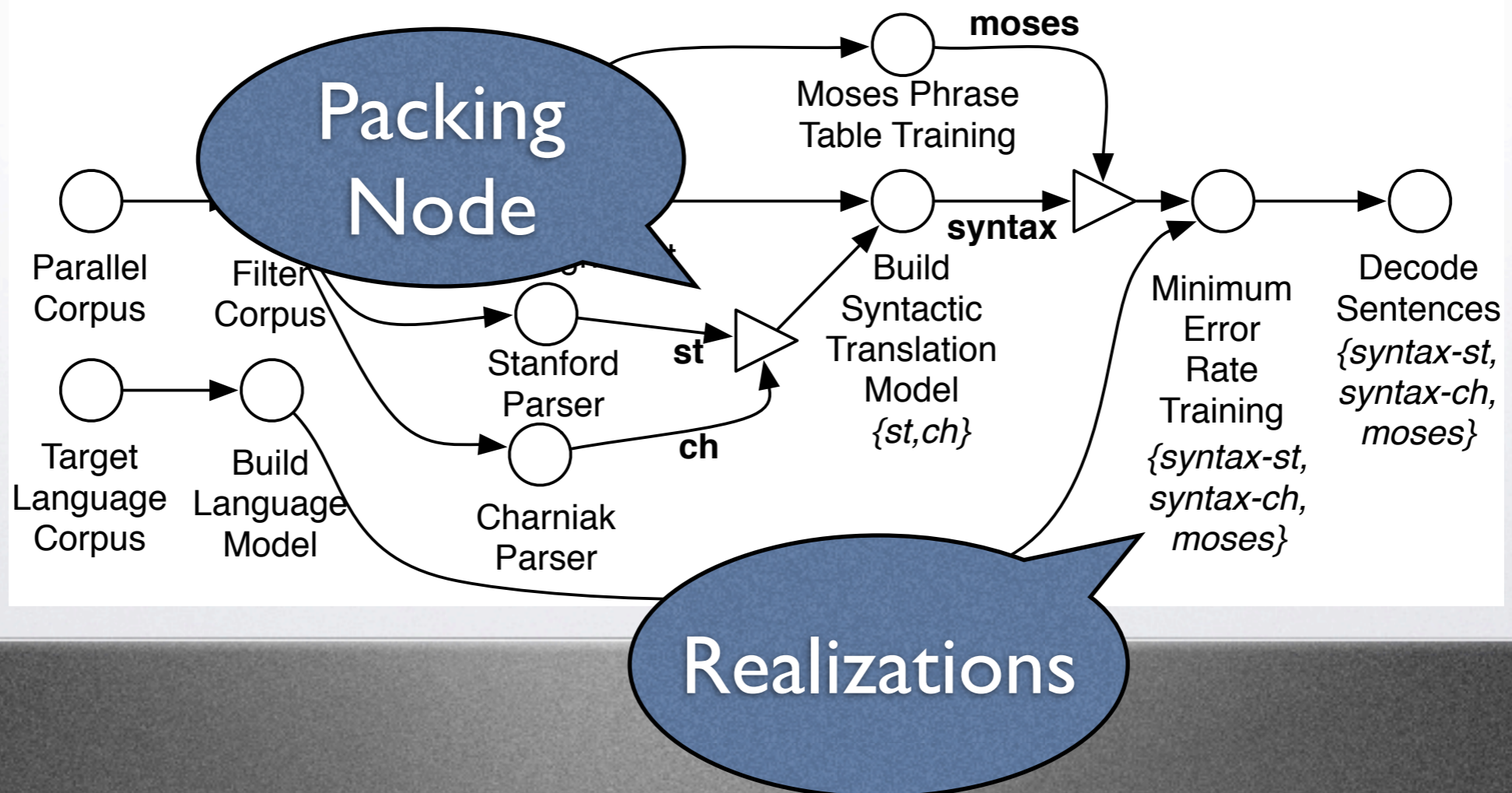
- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG





HyperWorkflows

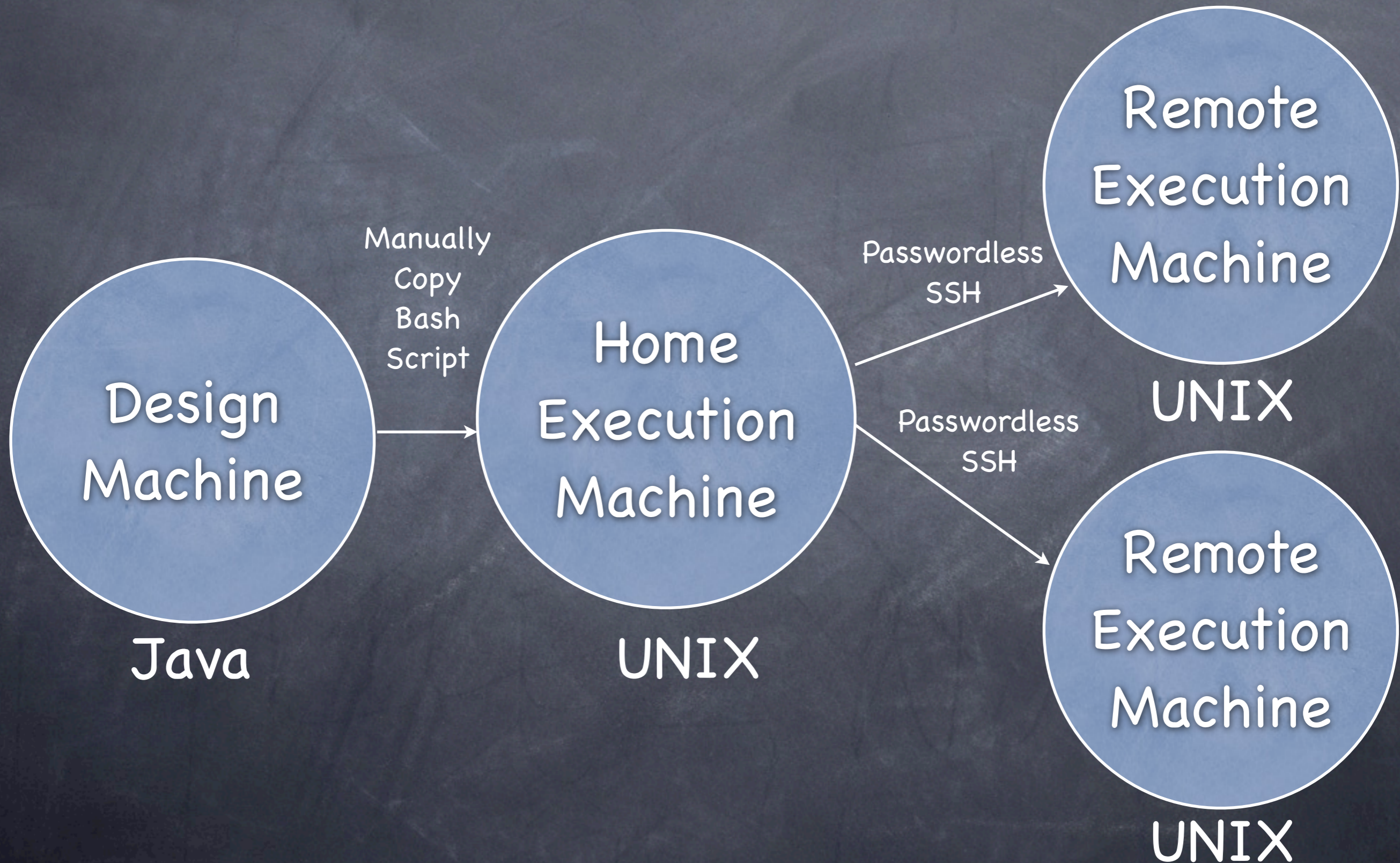
- **HyperWorkflows:** Shared substructure in experiments
- Encode small variations in a HyperDAG





What goes on
and
What comes out

What happens where?





What goes on

- Check if files and tools are present *first*
- Sanity checking at each step
- Copying of files (including to HDFS)
- Automatic login to remote machines (via passwordless SSH)
- Scheduler wrappers (e.g. Torque/Condor/SGE)

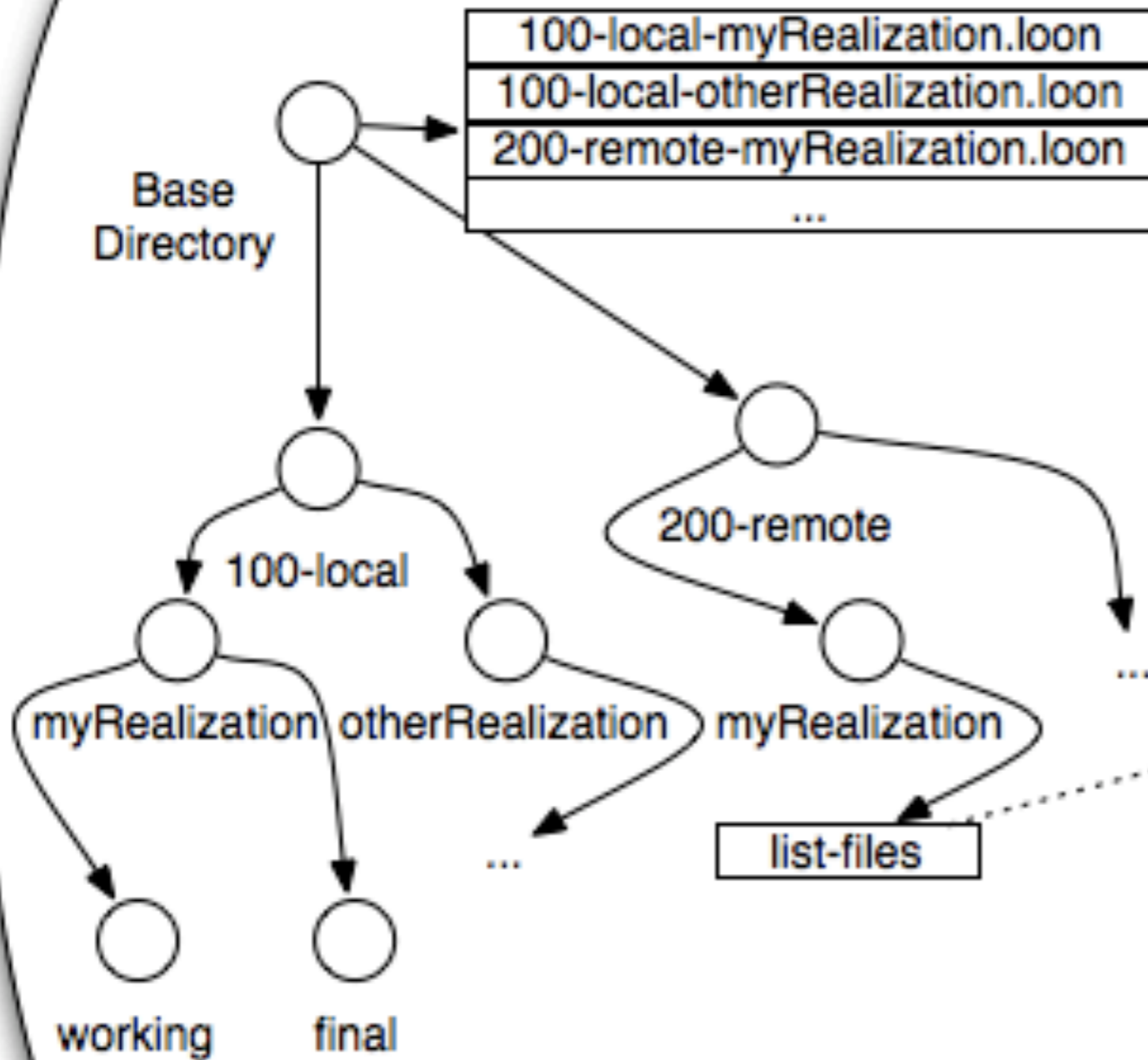


What comes out

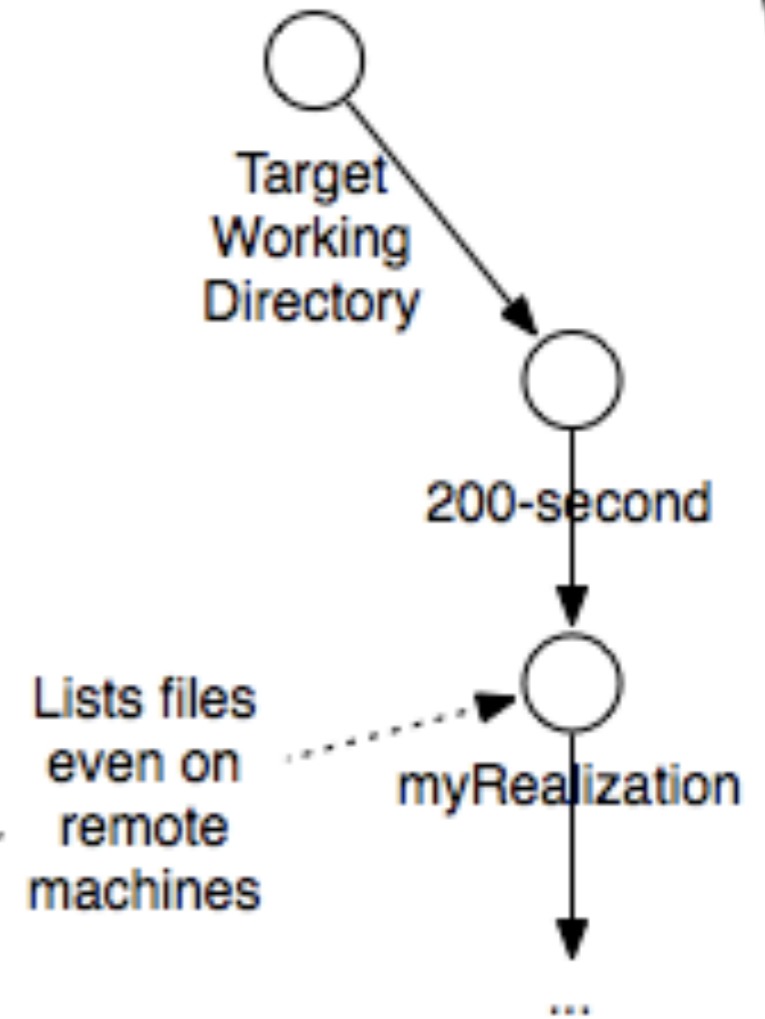
- Artifacts of the workflow in an organized directory structure
- Log with detailed information about data (corpus, alignments, parses, etc.) after pipeline step
 - Simple text format
 - Complete history in each file
- Email/SMS notifications of completion/failure

Directory Structure

Home Machine



Remote Machine 1



Lists files even on remote machines

Example Log Output

```
5000-tune.it1.AvgWeight.pt_wordcount      -1.76
5000-tune.it1.ExampleTopbestHyp.1 oslo 6-2 -lrb- afp -rrb- - terje roed..
5000-tune.it2.hypotheses.Total            712902
5000-tune.it2.hypotheses.PerSentence      396
5000-tune.it2.hypotheses.AddedTotal       272703
5000-tune.it2.hypotheses.AddedPerSentence 151
5000-tune.it2.Weight.lm                   1.55
4250-prune-pt-default.MachineName         gritgw1005.yahooresearchcluster.com
4250-prune-pt-default.Datestamp Tue Oct 6 22:38:35 UTC 2009
4250-prune-pt-default.TimeElapsed         0:17:17
4250-prune-pt-default.COUNT.Phrase_Records_Read 14561086
4250-prune-pt-default.COUNT.Source_Sides_After_Pruning 176529
4250-prune-pt-default.FileSystemCounters.FILE_BYTES_READ 308358509
```



MT Toolpack for LoonyBin

- Includes
 - Joshua training pipeline including Berkeley aligner and recasing (Jonny and Byung @ JHU)
 - Moses training pipeline
 - MGIZA/Chaksi (Qin)
 - SAMT (Andreas)
 - Multi-Metrics -- BLEU/NIST/Meteor/TER (Kenneth)
 - LM training via SRILM
 - MEMT (Kenneth)



Adding You Own Tools (Please)

- Just implement a python interface
 - Inputs
 - Outputs
 - Parameters
 - How to form UNIX command
 - Analyzers (optional, but recommended)



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



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



LoonyBin Best Practices

- Lots of steps. Why?
 - Continue on failures
 - Interchange components easily
 - Record effect of each component on data
- Workflows can have many granularities!



Conclusion

- **Make your life easier**
- **Make our lives easier**
- **Be a more responsible scientist**

Questions?

loonybin.sourceforge.net