Building computational pipelines

LECTURE 7

SEP 27, 2010

Programming: Scripting vs Software Engineering

- There are several kinds of programming.

- *Scripting* (from "writing little scripts") is centered on fairly small tasks.

- *Software Engineering* (from, um, "engineering") is centered on building bigger programs (like Avida).

- The skill sets do not actually overlap much.

Scripting

Building small scripts is useful for many things:

- Automating things you do again and again
  - Configuring machines
  - Running programs repeatedly

- Doing simple kinds of data analysis:
  - Combining data sets
  - Various kinds of statistical analyses
  - Graphing

Several kinds of languages used for scripting

- Shell scripts
  - Useful for task automation

- Python (and Perl, and Ruby, and …)
  - Particularly useful for text data manipulation.
  - More general languages, too.
  - Can be used for graphing and statistics.

- R (and Matlab)
  - Statistics, statistics, statistics!
  - Graphing

The real challenge in programming

- Computers are dumb.

- You need to match **what computers can do** with **what problems you need to solve**.

- This is the "computational thinking" part of things, basically.

The *annoying* challenge in programming

- I know what I need to do.

- How can I make the computer understand me??

- Analogous to speaking a foreign language in another country.

- This is "just details" as opposed to culture-shock aspect ("Ground floor" in Europe… floor 1 in US!)

Introduction to scripting

1) Shell script to install an EC2 machine

2) Shell script to compile Avida

3) Shell script to run Avida 10 times

4) Python script to "parse" summary data from running Avida 10 times

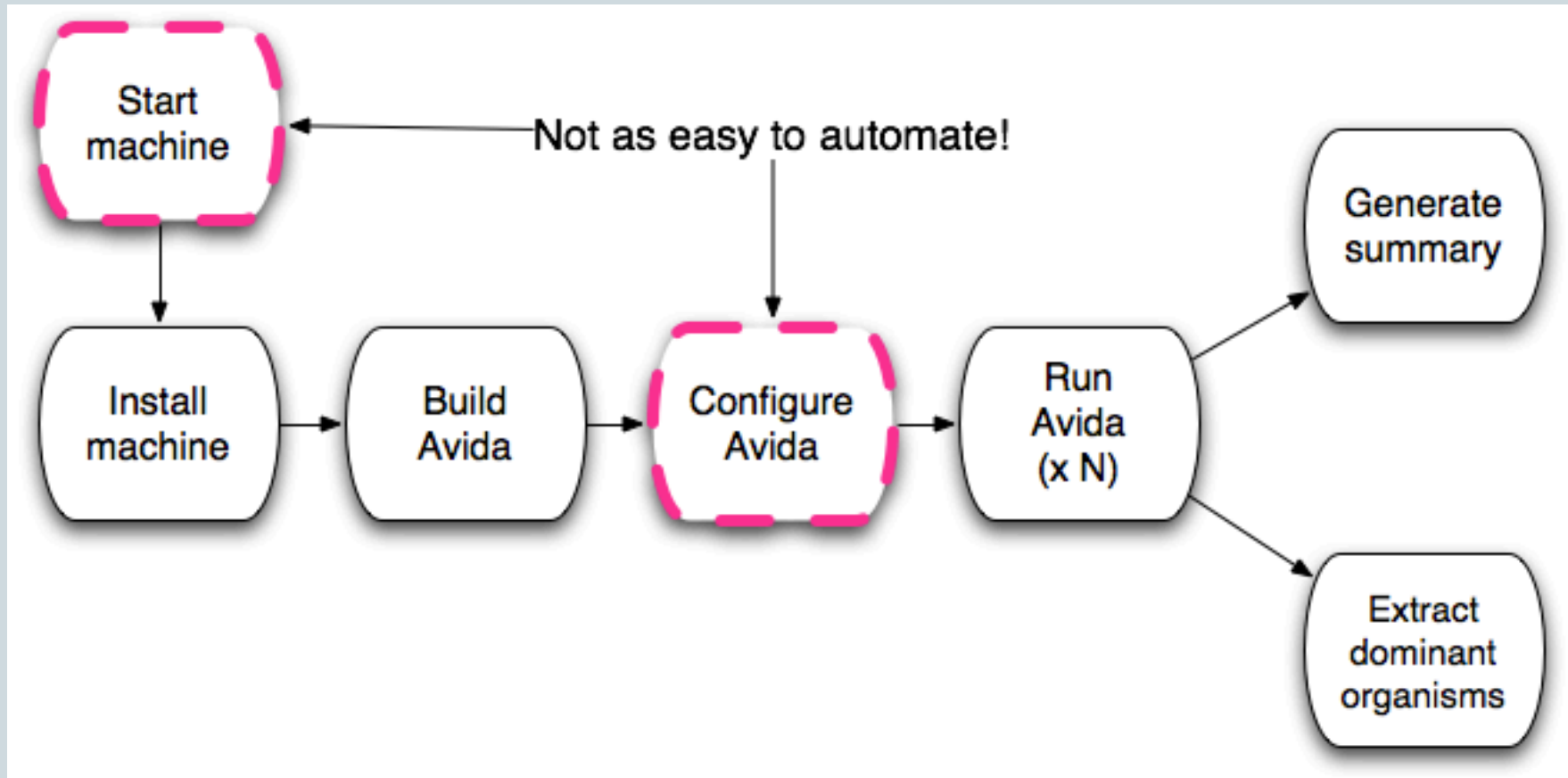5) Python script to extract dominant organism

Here's the real secret

- Don't ever write your own scripts from scratch!

- Take mine (or someone else's) and modify them!

- Iterative process:
  - Run script, verify that it "works"
  - Modify slightly
  - Repeat
  - SAVE EACH WORKING VERSION!

What we did, in diagram form

The Trap

- Once you know how to automate programs and analyses, it becomes *really, really easy* to generate extremely large amounts of data.

- However,
  - Keeping track of this data (and your scripts) gets difficult.
  - Keep your eye on the ball: "easy" doesn't necessarily mean worthwhile!
  - "Easy" also doesn't mean "correct."