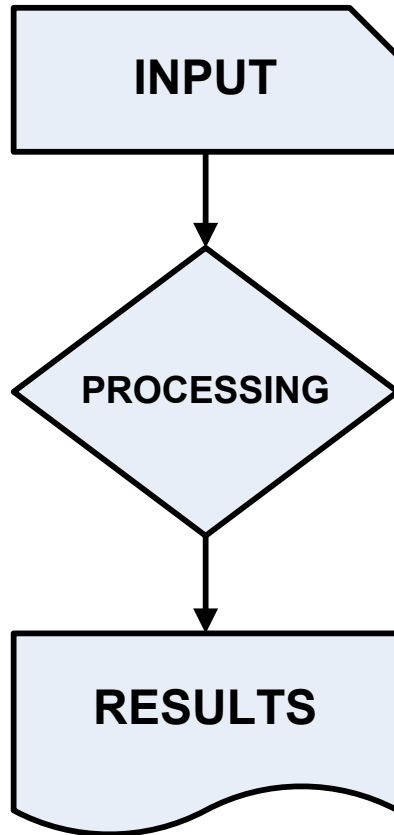# SCIENCE = DATA

Why Optimizating Your Workflow and Data Management on SciNet Enables More and Better Science, With a Bird's-Eye View of How To Achieve This Based on Several Successful Use-Cases, and Including The Possible Use of HPSS for Big Data.

SNUG

Sep/12/2012

NEED TO ANSWER ALL THESE QUESTIONS

INPUT

$\left\{\begin{array}{l}\text{LOCATION (PATH) ?} \\ \text{FORMAT (ASCII, binary, tarball, gzip'ed) ?} \\ \text{AMOUNT (TB, # of Files, # of directories) ?} \\ \text{RATE (MB/s, IOPs) ?}\end{array}\right.$

PROCESSING

$\left\{\begin{array}{l}\text{RESOURCES (# of CPUs/Nodes, Modules) ?} \\ \text{SCHEDULE (walltime, checkpoints) ?}\end{array}\right.$

RESULTS

LOCATION, FORMAT, AMOUNT, RATE?

SCRATCH
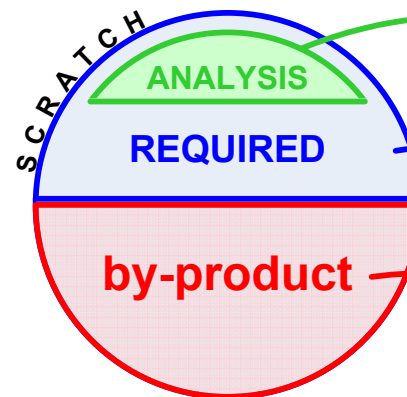
ANALYSIS → **HPSS and/or project (tarball)**

REQUIRED → **HPSS (tarball)**

by-product → **Immediate deletion**

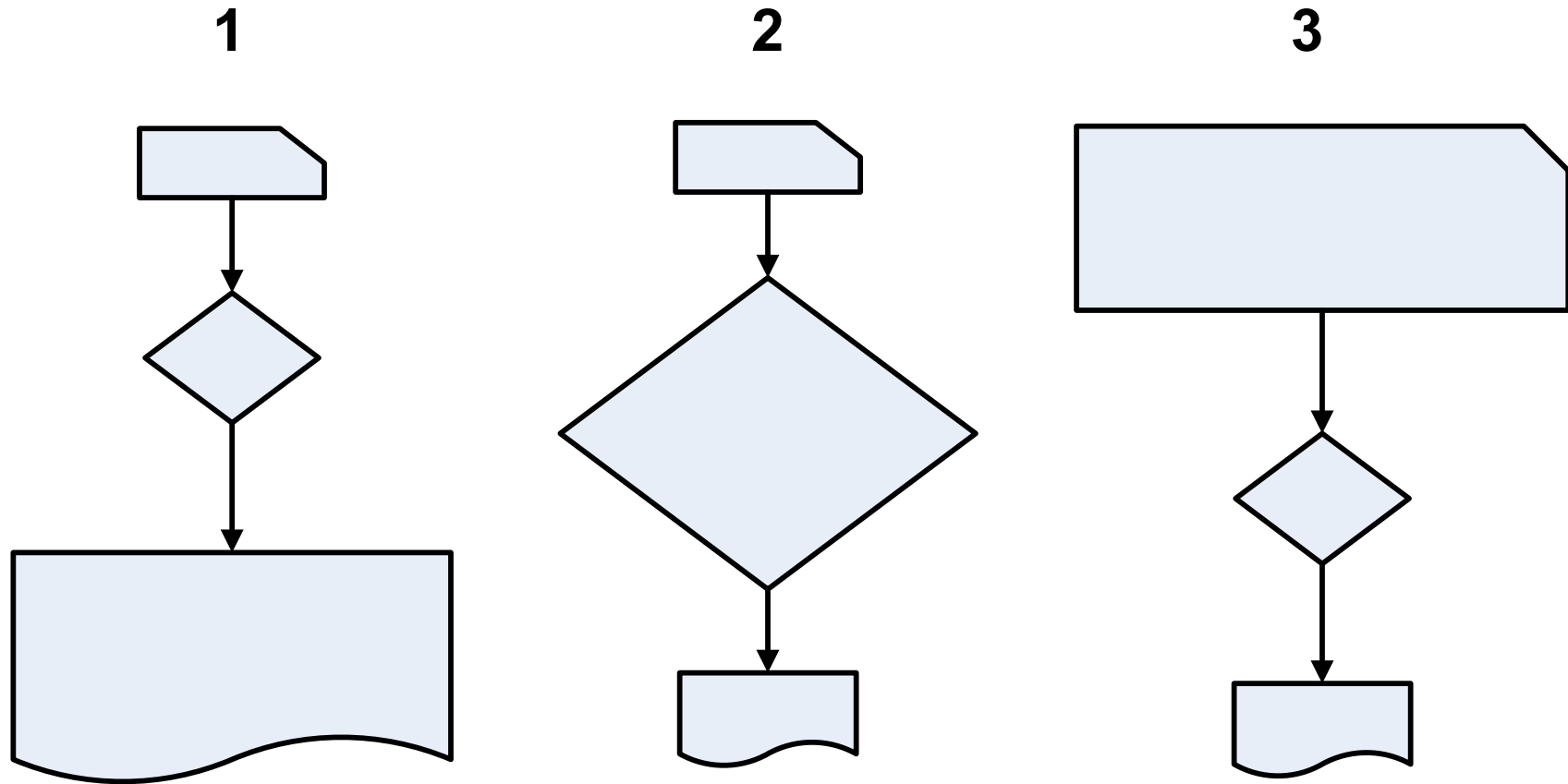**UNIT OF COMPUTATION OR SIMULATION?**

Month?
Cell?
Tree node?
Sample?
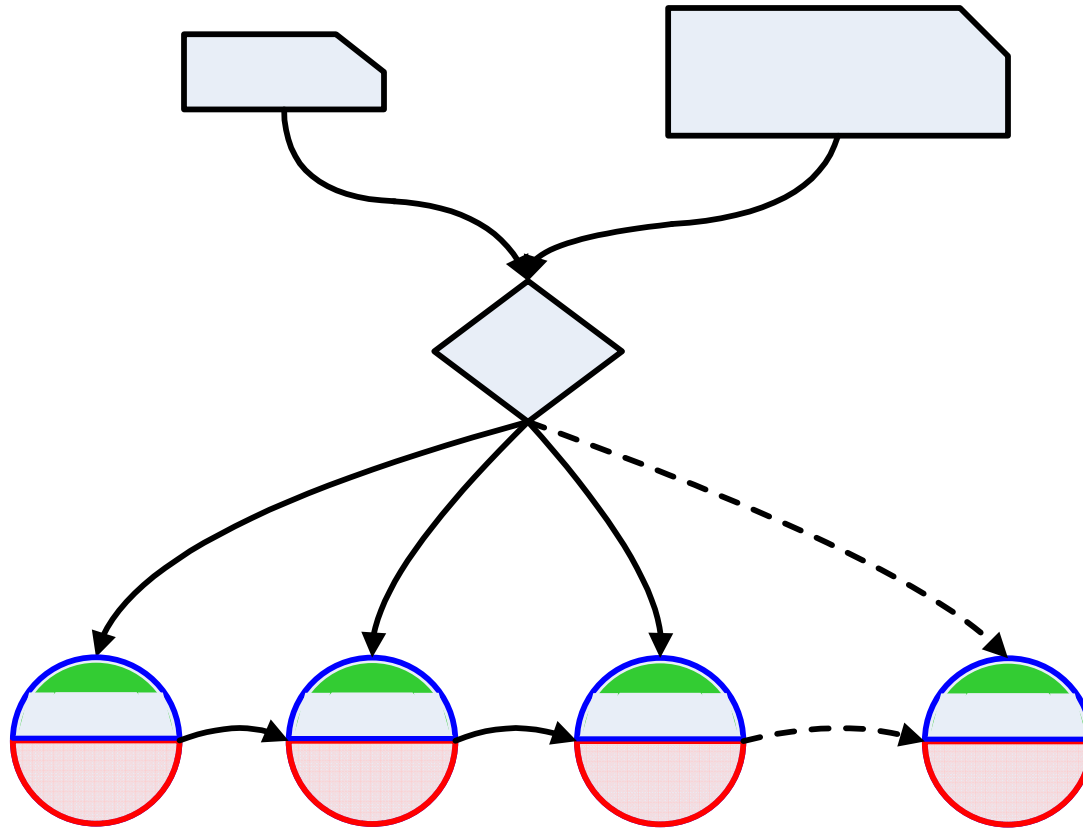
AWARENESS OF THE WEIGTH OF EACH COMPONENT

SCENARIOS

TIME-tax vs. STORAGE-tax ?
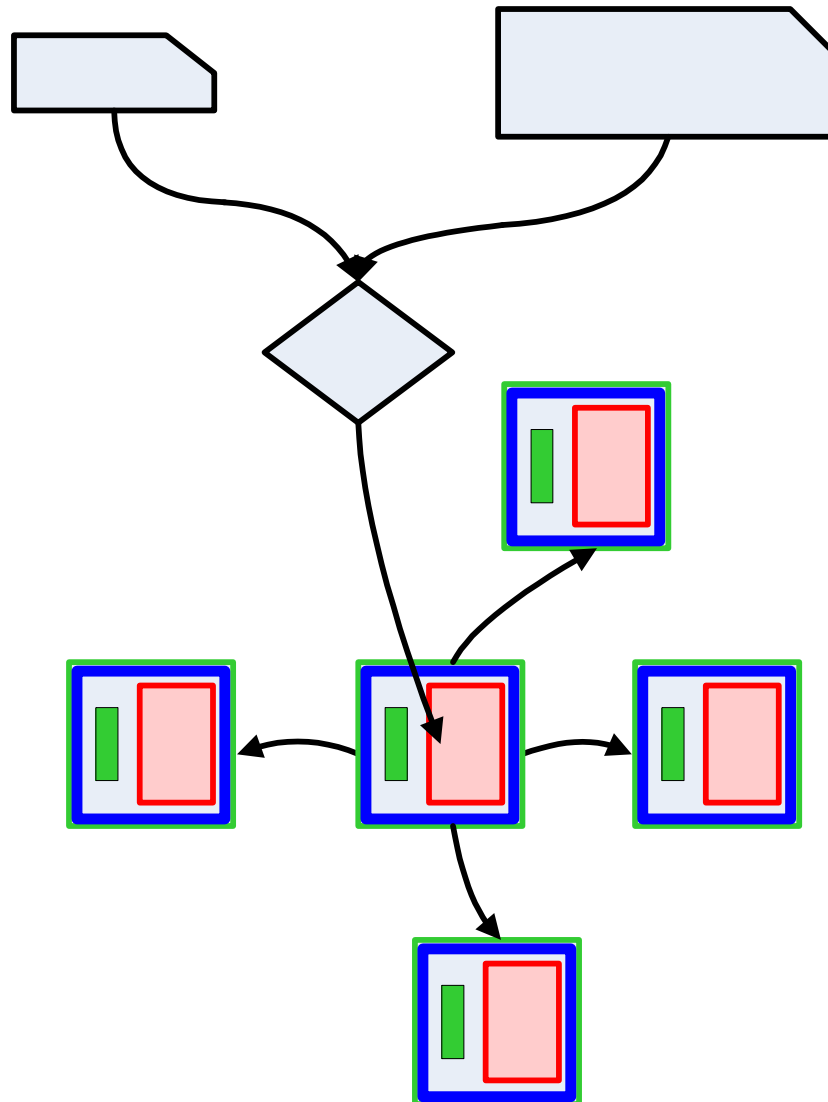**Redundancies and inefficiencies within the group ?**

CASE STUDY 1

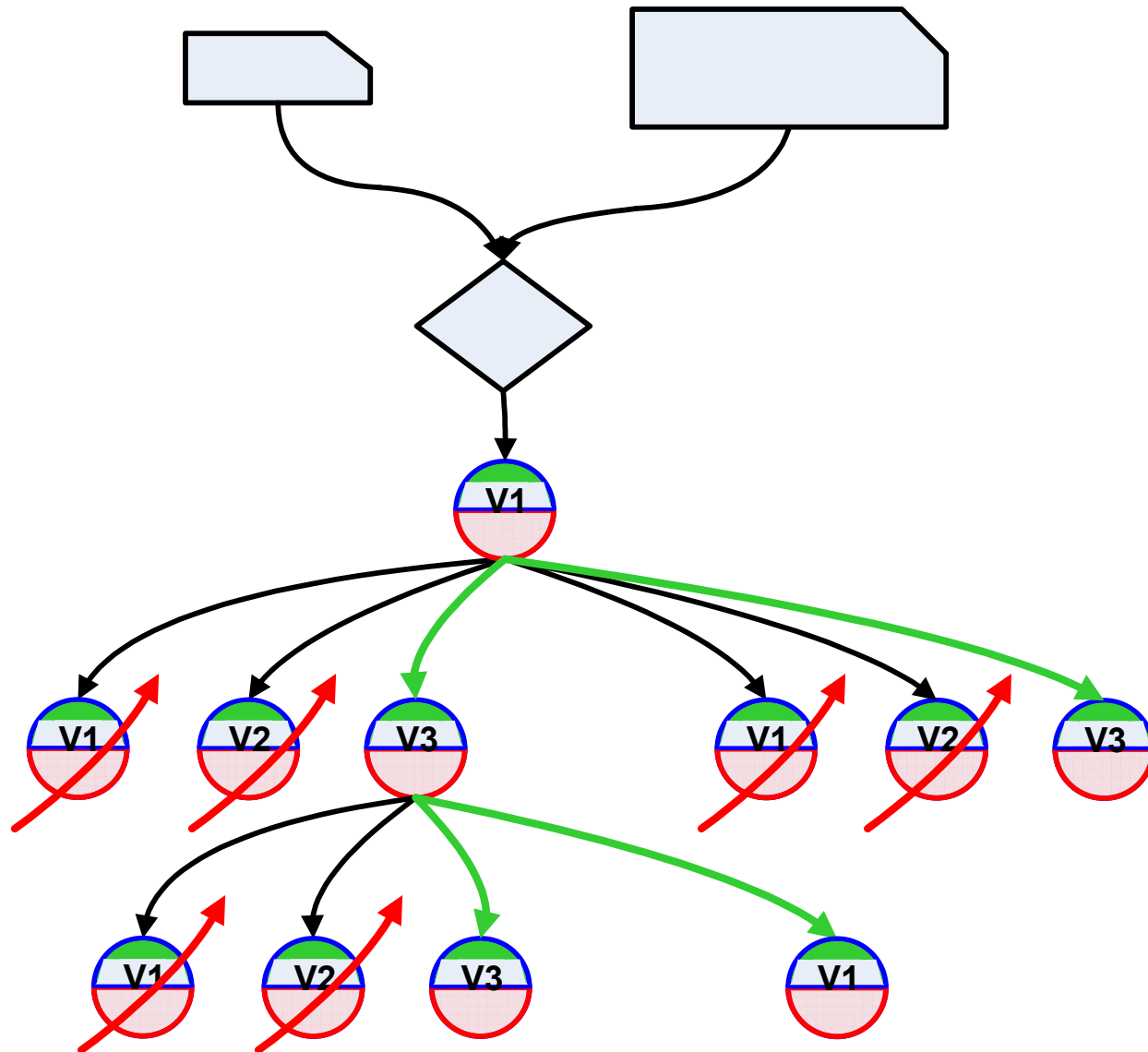## Sequential dependency

CASE STUDY 2

Boundary dependency

CASE STUDY 3

Unit refinement

CASE STUDY 4

Sampling