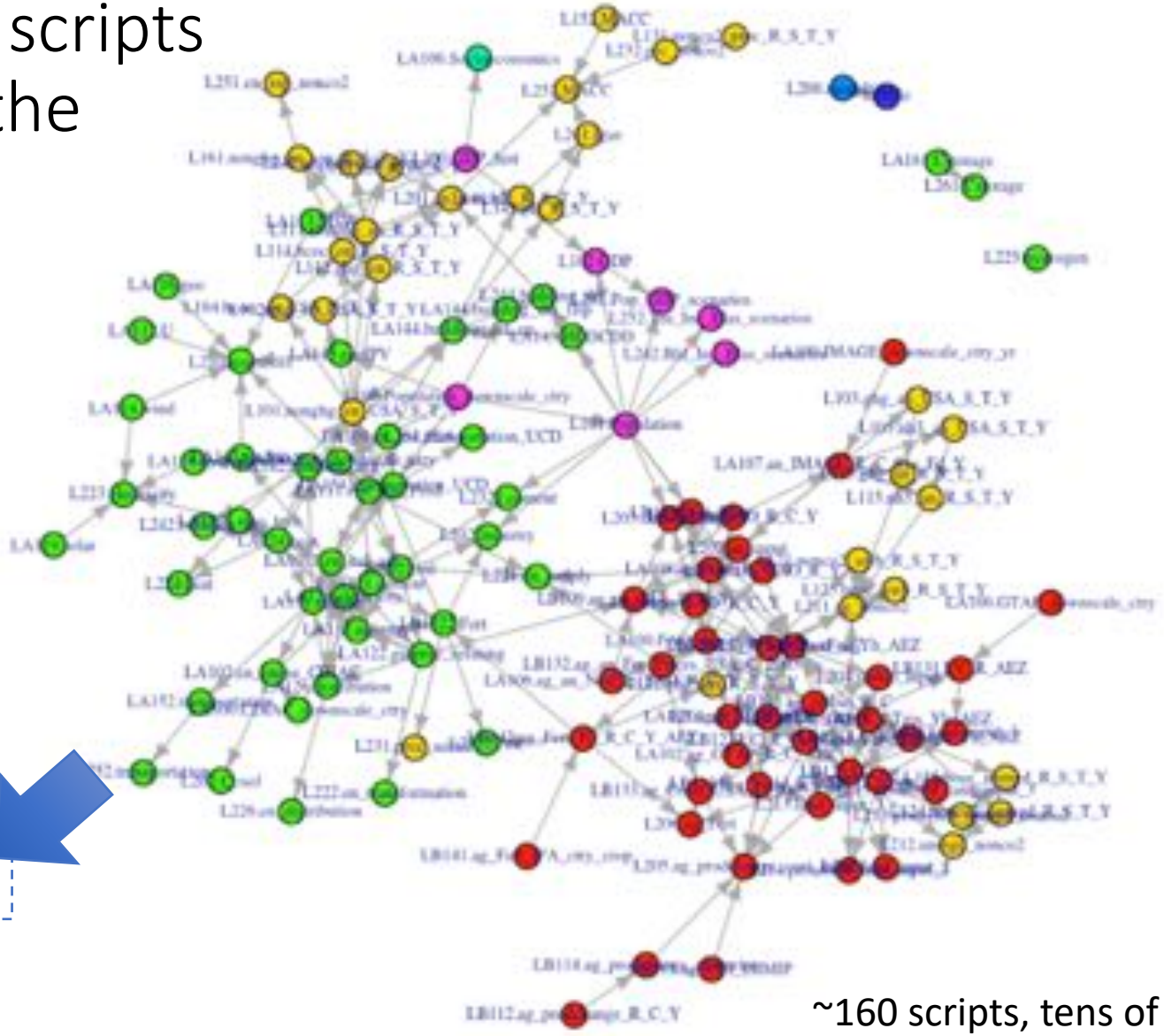
The GCAM data system

- The new version of GCAM has a completely rewritten data system that assembles, checks, and calculates its inputs
- This sophistication is necessary because GCAM is a *computational engine*, not a static model
- The entire structure of the model can be defined and dynamically changed by its inputs

Before, a complex system of R scripts comprised the GCAM *data system*



xml

GCAM

~160 scripts, tens of thousands of lines of code

Goals for the new data system

- Spread the knowledge and find problems
- Better documentation throughout
- Flexibility (change assumptions)
- Robustness
  - quickly know when things go wrong
- Code clarity
- Tools to diagnose, explore, modify, test
- Easy to use
- Speed

Principles

- Clear and clean code, documentation, abstracted common code, discrete functions
- Lots and lots of “unit testing”
- R *package*: lots of good things for free, and very easy to install and run

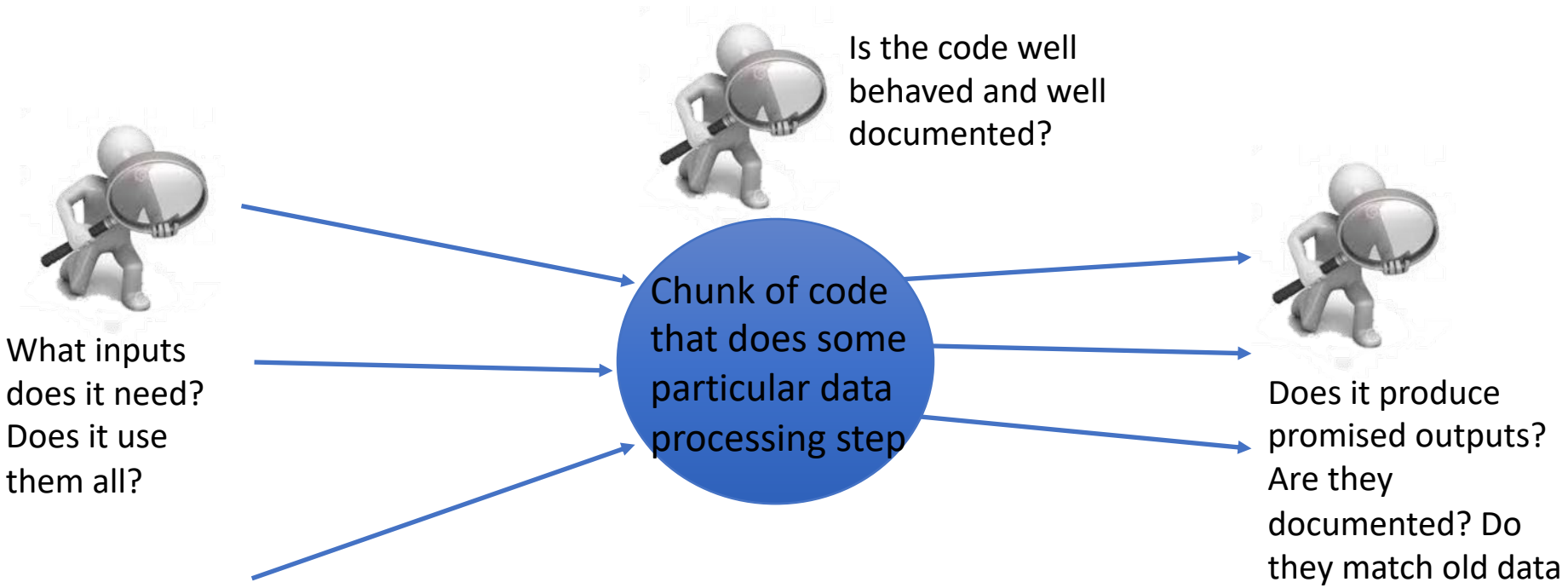
```
> library(gcamdata)
> driver()
GCAM Data System v0.4
Found 190 chunks
Found 1345 chunk data requirement
Found 833 chunk data products
[1]"module_aglu_L2242.land_input”
[1]"- make 0.10”
[1]"module_aglu_LA100.0_LDS_prepr
”
[1]"- make 2.69”
[1]"module_aglu_LA100.FAO_downsca
[1]"- make 3.81"
```

(etc.)

Testing and its benefits

- With what's called "continuous integration", i.e. continuous automated testing, we can enforce *lots* of good things
  - Correct outputs
  - Robust behavior
  - Code reviews
  - Documentation – system *will not build* with incomplete documentation

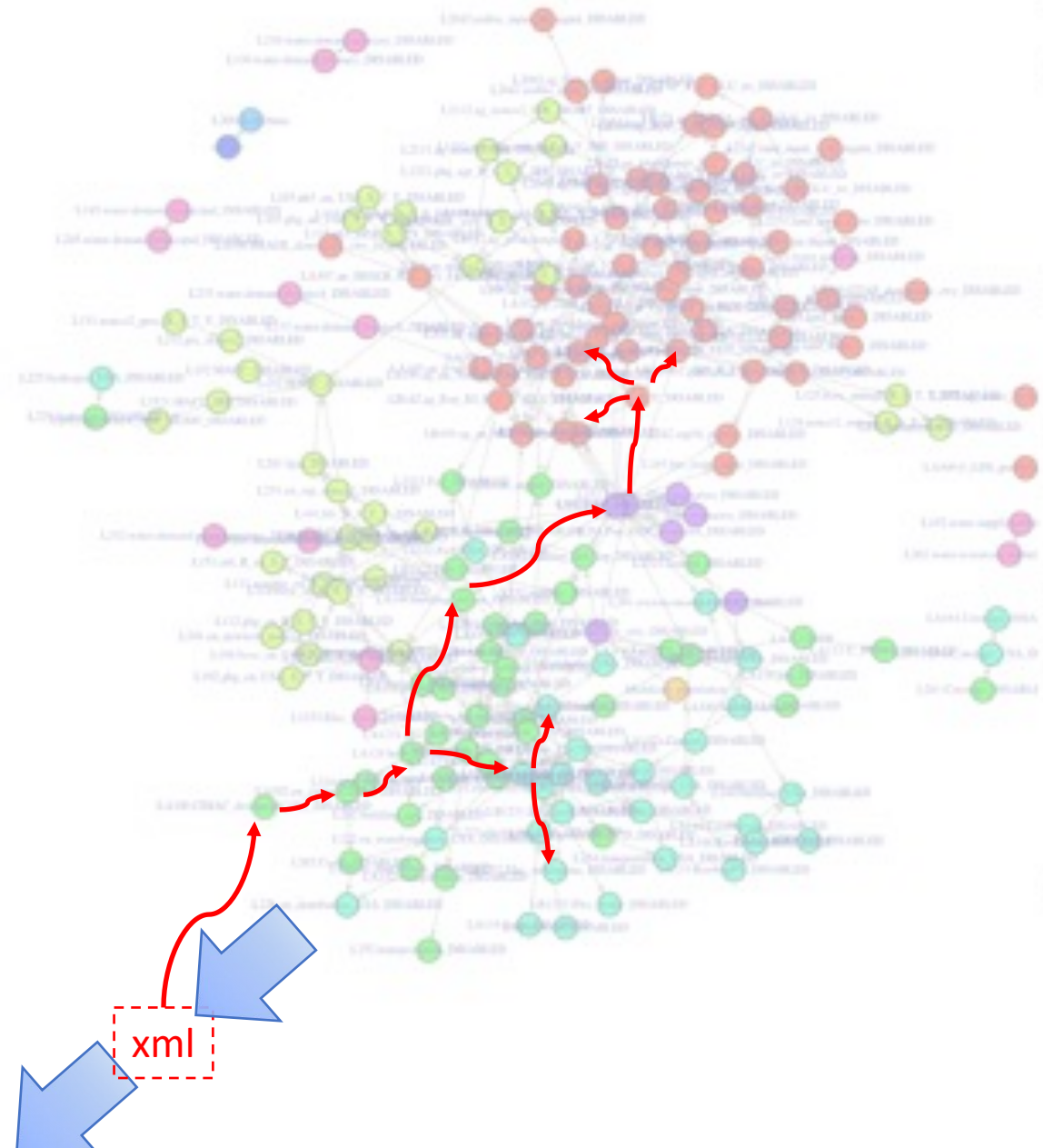
Testing and its benefits



New capabilities enable better and faster research

- Because we define and enforce chunk behavior, we can do useful things:
  - Graph data flow
  - Get the provenance of any piece of data
  - Trace data and identify problems
  - ‘Shim’ into the data system to examine/change/test
- I.e., this can be a useful tool not just cumbersome scripts
- Data provenance: we can query the new data system for all steps and data sources that produced a particular GCAM input


```
> trace("bld_agg_gSSP1.xml")
1 - bld_agg_gSSP1.xml
        produced by module_socio_batch_bld_agg_
        Precursor: L242.IncomeElasticity_bld_gS
2 - L242.IncomeElasticity_bld_gSSP1
        produced by module_socioeconomics_L242.
        Building Income Elasticity: gSSP1 (Unit
        Uses previously calculated per-capita G
        Building income elasticity for each GCA
        Precursor: common/GCAM_region_names (#3
        Precursor: socioeconomics/A42.inc_elas
        Precursor: L102.pcgdp_thous90USD_Scen_R
3 - common/GCAM_region_names - read from file
        GCAM 32-region names (NA)
        Maps GCAM region IDs to region names
        Read from extdata/common/GCAM_region_na
        No precursors
4 - socioeconomics/A42.inc_elas - read from fi
        Building sector income elasticity, (pcg
inc_elas:unitless (% change in service demand ,
        aggregate buildings sector income elast
        Read from extdata/socioeconomics/A42.in
        No precursors
5 - L102.pcgdp_thous90USD_Scen_R_Y - produced
        Gross Domestic Product (GDP) per capita
of 1990 USD (MER))
        Computed as GDP/population.  Values pri
        historical; values subsequent are from
        Precursor: common/iso_GCAM_regID (#6 be
        Precursor: socioeconomics/SSP_database_
        Precursor: socioeconomics/IMF_GDP_growt
        Precursor: L100.gdp_mil90usd_ctry_Yh (#
        Precursor: L101.Pop_thous_R_Yh (#10 bel
        Precursor: L101.Pop_thous_Scen_R_Yfut (
6 - common/iso_GCAM_regID - read from file
        ISO to GCAM region mapping (NA)
        Maps iso codes to GCAM regions (includi
        ----------,,Former GCAM regions,
        Read from extdata/common/iso_GCAM_regID
        No precursors
```




```
> trace("bld_agg_gSSP1.xml")
1 - bld_agg_gSSP1.xml
        produced by module_socio_batch_bld_agg_
        Precursor: L242.IncomeElasticity_bld_gS
2 - L242.IncomeElasticity_bld_gSSP1
        produced by module_socioeconomics_L242.
        Building Income Elasticity: gSSP1 (Unit
        Uses previously calculated per-capita G
        Building income elasticity for each GCA
        Precursor: common/GCAM_region_names (#3
        Precursor: socioeconomics/A42.inc_elas
        Precursor: L102.pcgdp_thous90USD_Scen_R
3 - common/GCAM_region_names - read from file
        GCAM 32-region names (NA)
        Maps GCAM region IDs to region names
        Read from extdata/common/GCAM_region_na
        No precursors
4 - socioeconomics/A42.inc_elas - read from fi
        Building sector income elasticity. (pcg
inc_elas:unitless (% change in service demand
        aggregate buildings sector income elast
        Read from extdata/socioeconomics/A42.in
        No precursors
5 - L102.pcgdp_thous90USD_Scen_R_Y - produced
        Gross Domestic Product (GDP) per capita
of 1990 USD (MER))
        Computed as GDP/population.  Values pri
        historical; values subsequent are from
        Precursor: common/iso_GCAM_regID (#6 be
        Precursor: socioeconomics/SSP_database_
        Precursor: socioeconomics/IMF_GDP_growt
        Precursor: L100.gdp_mil90usd_ctry_Yh (#
        Precursor: L101.Pop_thous_R_Yh (#10 bel
        Precursor: L101.Pop_thous_Scen_R_Yfut (
6 - common/iso_GCAM_regID - read from file
        ISO to GCAM region mapping (NA)
        Maps iso codes to GCAM regions (includi
        ----------,,Former GCAM regions,
        Read from extdata/common/iso_GCAM_regID
        No precursors
```

More capabilities

- Not tie ourselves specific data sources
  - E.g. scale to arbitrary input
- Easy to incorporate upgrade/different data assumptions
- Track units associated with data
- Produce smooth, arbitrary time series for backcasting experiments
- Easy-to-install and run R package