VAMDC technology

Guy Rixon

Innsbruck, February 2013

VAMDC infrastructure

Data

Application

The VAMDC infrastructure is an intermediary: a device for getting data from databases to applications in an interoperable way. VAMDC provides neither the databases or the applications. The infrastructure imposes a number of standards on the data flow.

Plan A

Dump each database into a file and put on web.

Pro:

- "The simplest thing that could possibly work"
- Everything you can get has its own URL

Con:

- Data-sets too large (up to 10GB)
- No easy way to make data extracts

Plan B

Pre-compute all possible data extracts and dump on web

Pro:

- Selection now easy
- One URL for each possible extract

Con:

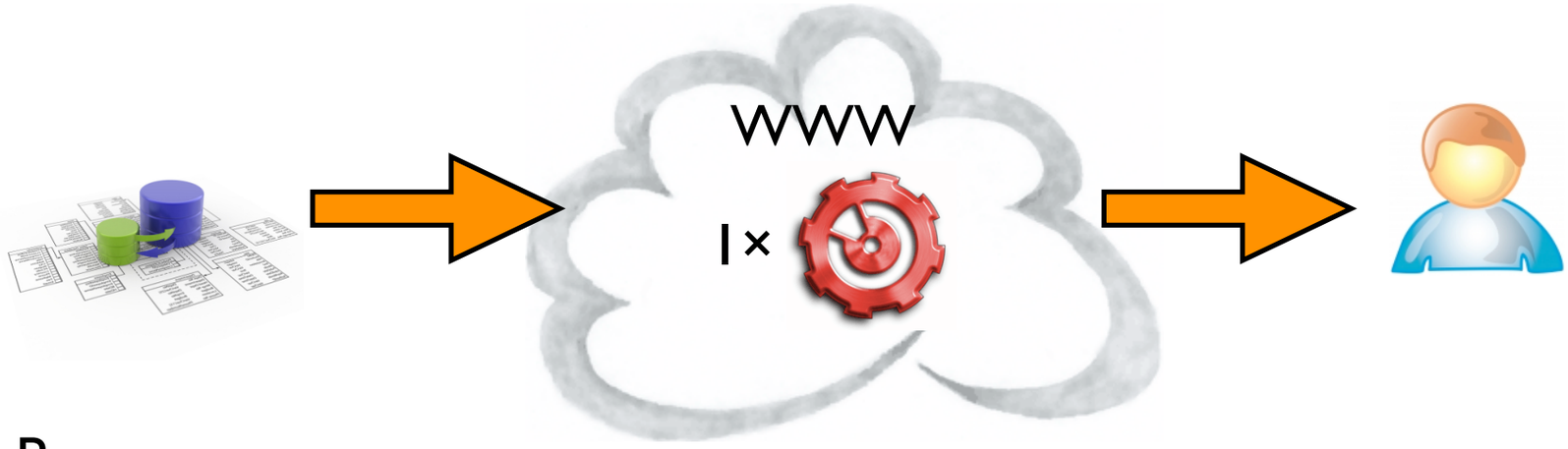
- Impossible to implement!

Plan C

Compute data extracts on demand

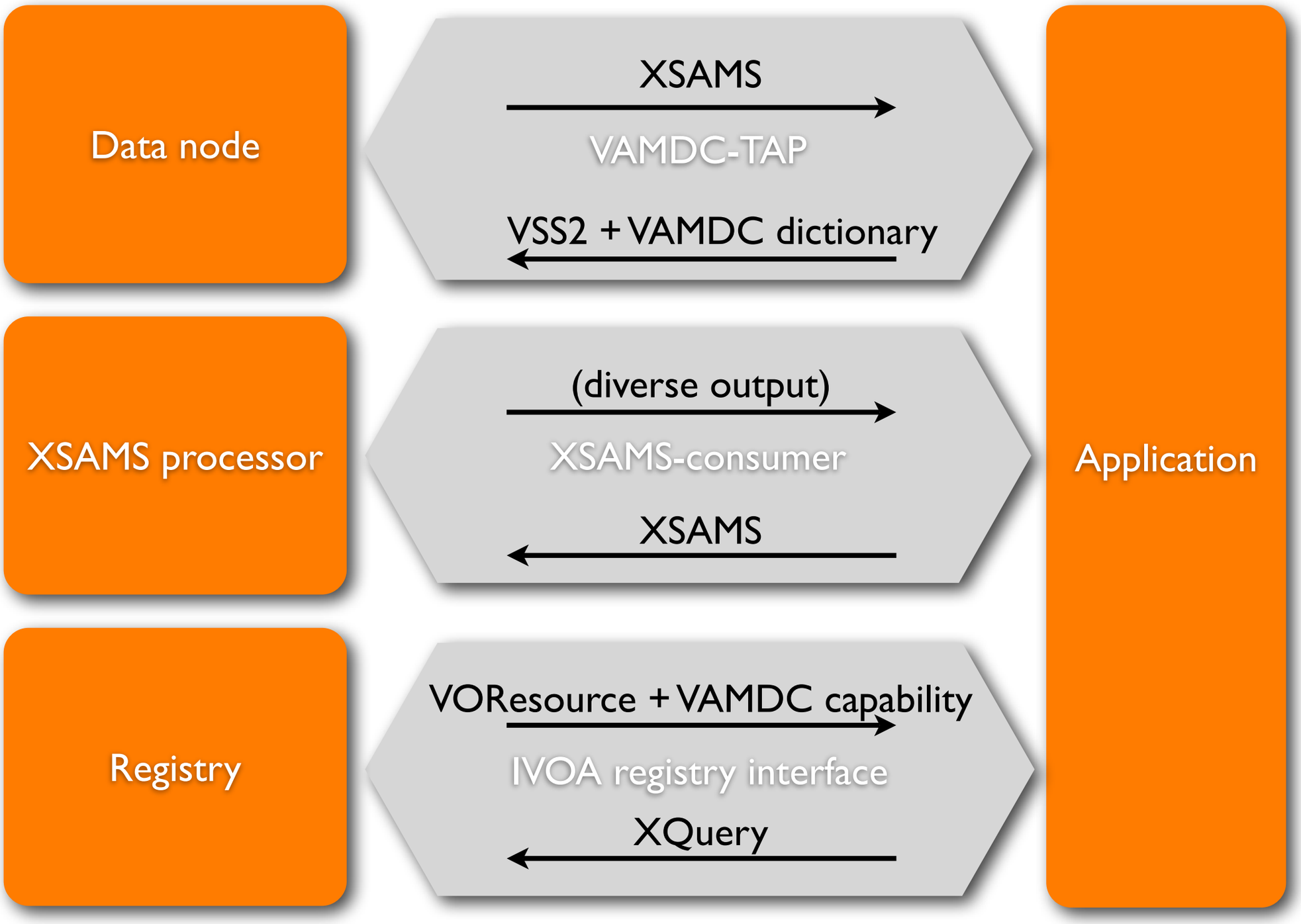
but

index them on the web as if pre-computed



- Pro:
  - Implementation now feasible
  - Still have a URL for every data-extract
- Con:
  - Some assembly required
  - Need to define standards for services, queries etc.

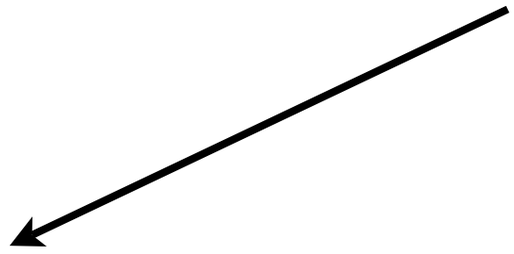
The core standards



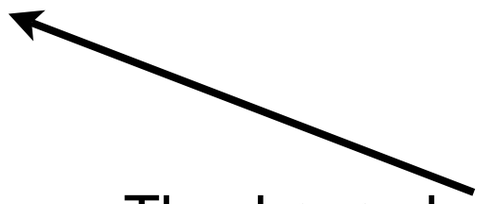
One group of standards for each type of web service. Each data node gives access to one database; VSS2 is the query language, the dictionary states the terms to use in the query, and XSAMS is the data format for the result. Note that VAMDC gives out data extracts resulting from queries, not entire or pre-prepared data-sets. The XSAMS processors transform XSAMS documents into other forms. The registry lists the locations and details of the other services.

Data URLs

The address of the service
(this one for CDMS)



http://cdms.ph1.uni-koeln.de:8090/DjCDMS/tap/sync?
REQUEST=doQuery&LANG=VSS2&FORMAT=XSAMS
&QUERY=SELECT+ALL+WHERE+MoleculeStoichiometricFormula+%3D+%27CO%27



The data to be extracted

Call these with HTTP GET

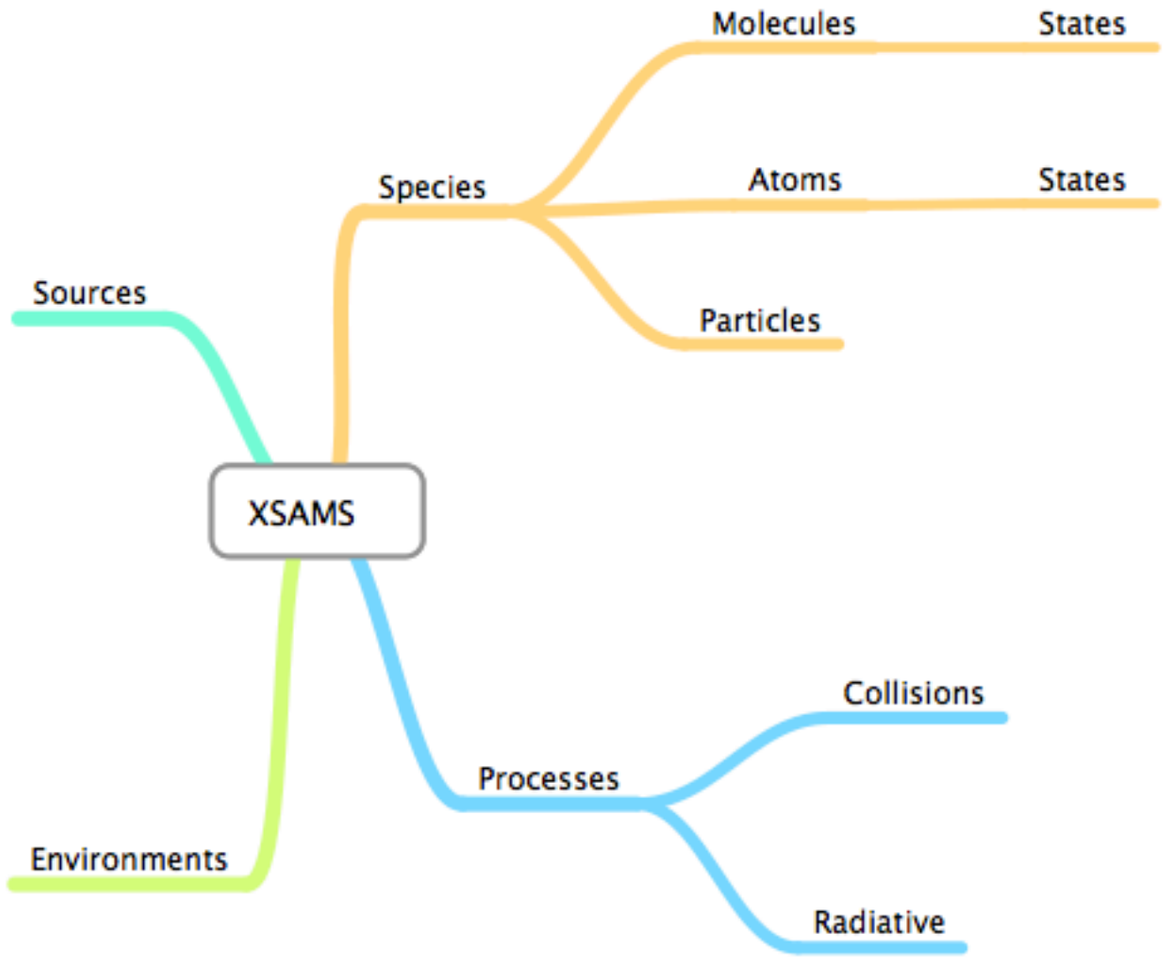
Our service protocol embeds the query language which in turn embeds terms from the dictionary. This lets us form a URL for an data extract. These URLs are easy to use – HTTP GET or POST – so we don't need libraries to implement the protocol.

XSAMS

- "XML schema for Atoms, Molecules & Solids"
- Developed by IAEA & VAMDC:
  - Proposed 2003, at IAEA DCN meeting
  - First versions by (IAEA, NIST, ORNL U. Pierre & Marie Curie, OPM, RFNC-VNIITF)
  - Subsequent development by VAMDC
- See http://www-amdis.iaea.org/xsams/
- See also http://www.vamdc.org/documents/standards/#data-model

XSAMS
XML Schema for Atoms, Molecules, and Solids

XSAMS structure: top



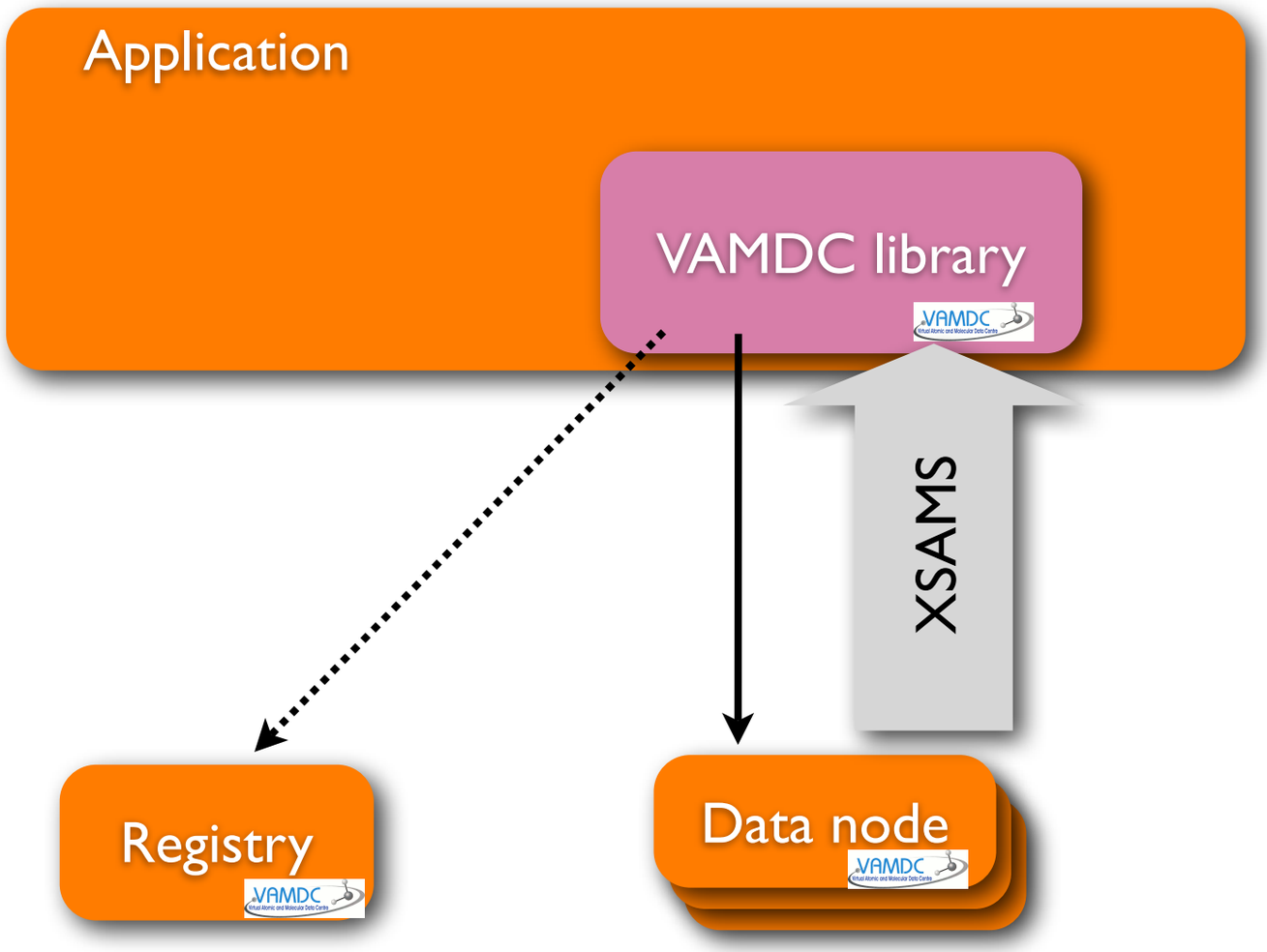
XSAMS structure: bottom

- All quantities have units
- All values can have associated uncertainties
- All values can have a source reference
- XML ⇒ no encoding issues for numbers

XSAMS for molecules

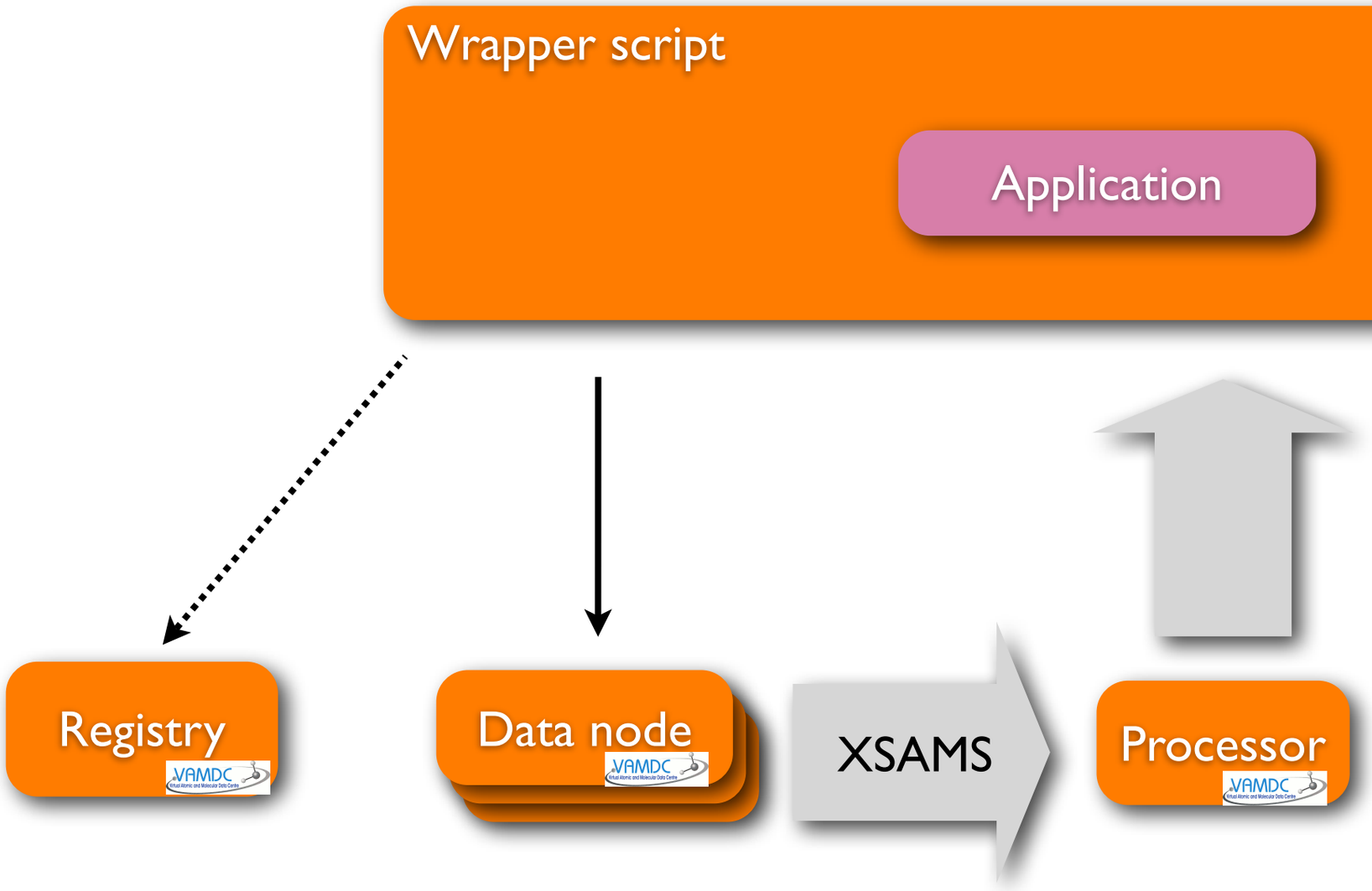
- "Case-by-case" XSAMS:
- Separate, additional schema for each class of molecule:
  - 1. Diatomic closed shell (`dcs`): CO, N2, NO+
  - 2. Hund's case (a) diatomics (`hunda`): NO, OH [for low J]
  - 3. Hund's case (b) diatomics (`hundb`): O2, OH [for high J]
  - 4. Closed-shell, linear triatomic molecules (`ltcs`): CO2, HCN
  - ...etc up to at least 12 cases

Adapted application



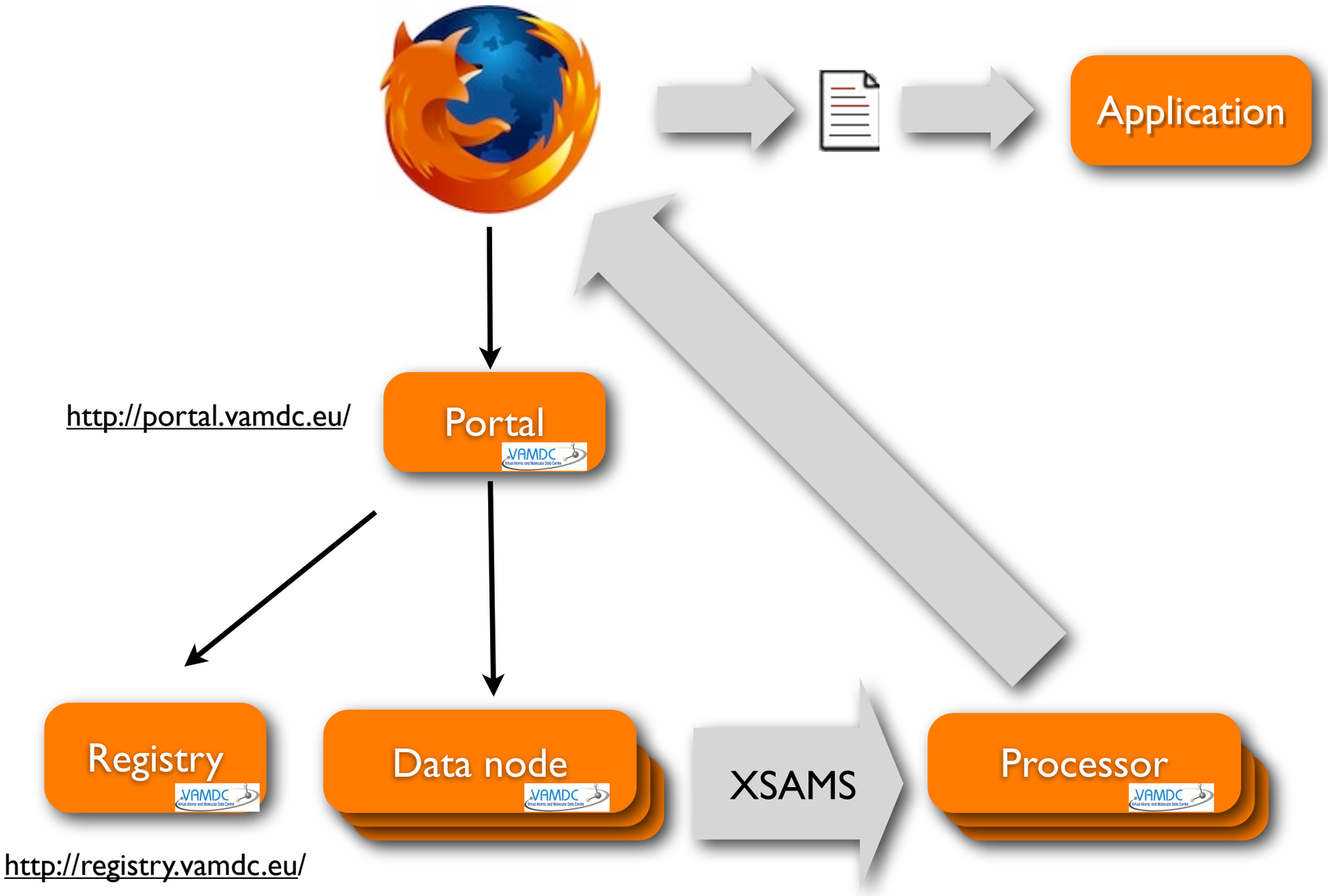
Normally, a science code is run on a user's own computer. It is possible to use VAMDC facilities with these codes by writing a wrapper script, which can be a simple shell script or a small program in, e.g., Python. All VAMDC services can be called easily from scripts. Narrow arrows represent commands, broad arrows represent data flows.

Wrapped application



Normally, a science code is run on a user's own computer. It is possible to use VAMDC facilities with these codes by writing a wrapper script, which can be a simple shell script or a small program in, e.g., Python. All VAMDC services can be called easily from scripts.

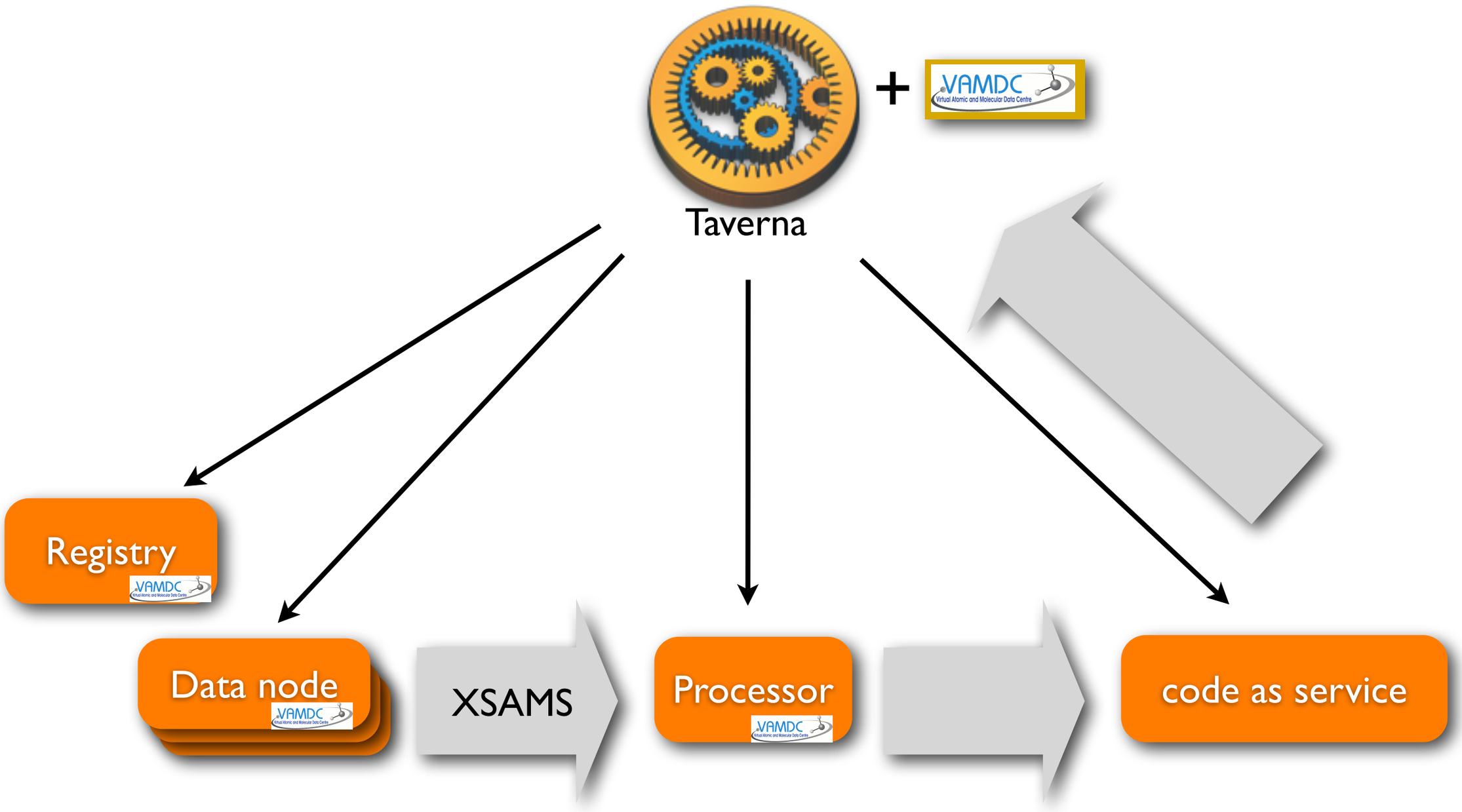
Portal, nodes & processors



VAMDC provides a web portal as a proxy for a connected application. The user drives the portal with a web browser and forms the queries interactively, typically routing the results to a processor for transformation into a desired format. The results are saved on disc and then offered to a local application.

In a system, there is one portal, many data nodes and many processors, all on different sites.

Taverna; code as service



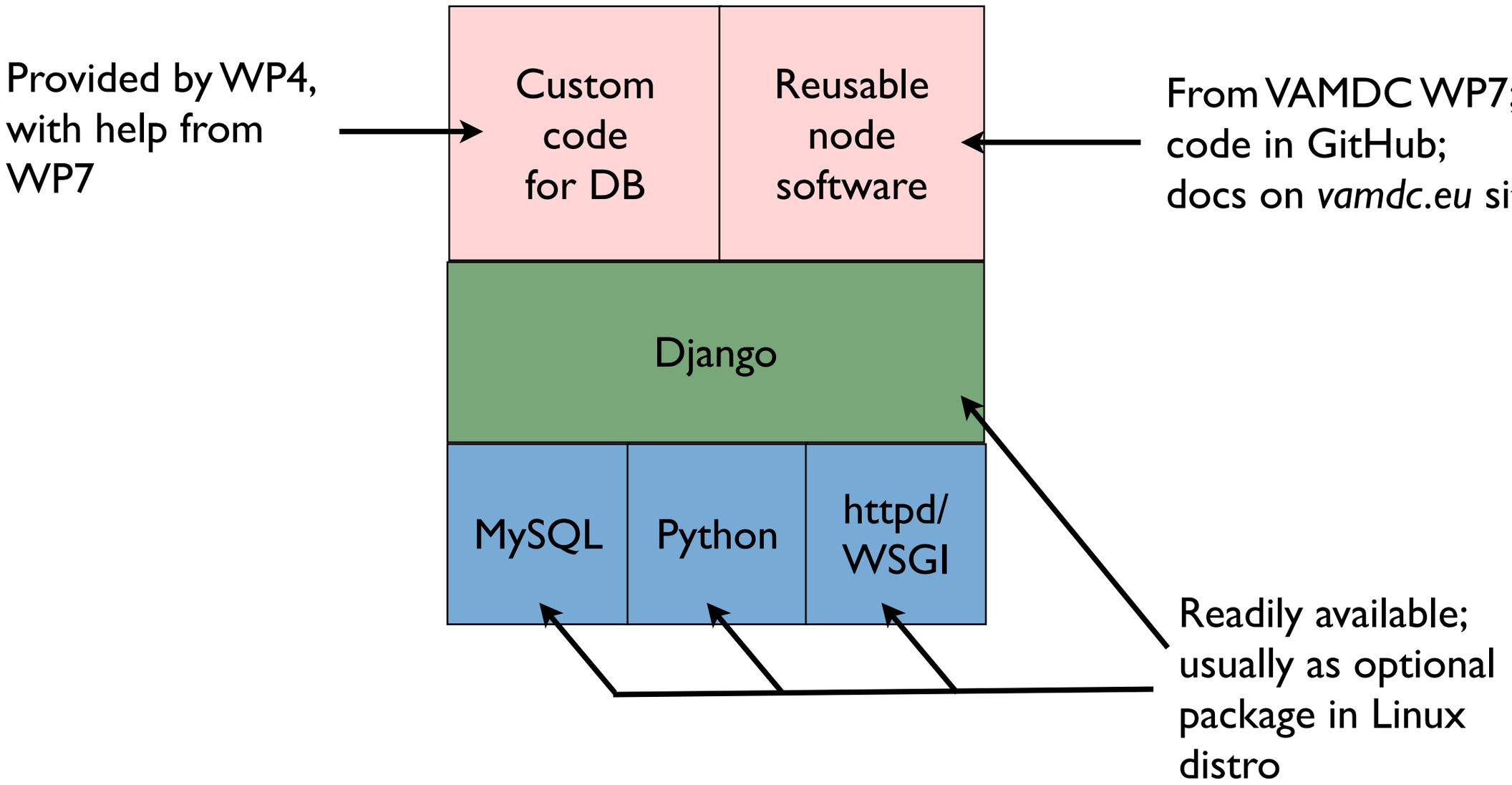
Science codes can be wrapped as web-services and run using the Taverna workflow-system. These services would not be part of the VAMDC system itself, but Taverna can compose calls to VAMDC services with call to the the code; it needs a VAMDC plug-in for this and that plug-in is a released, VAMDC product.

So how do I make a node?

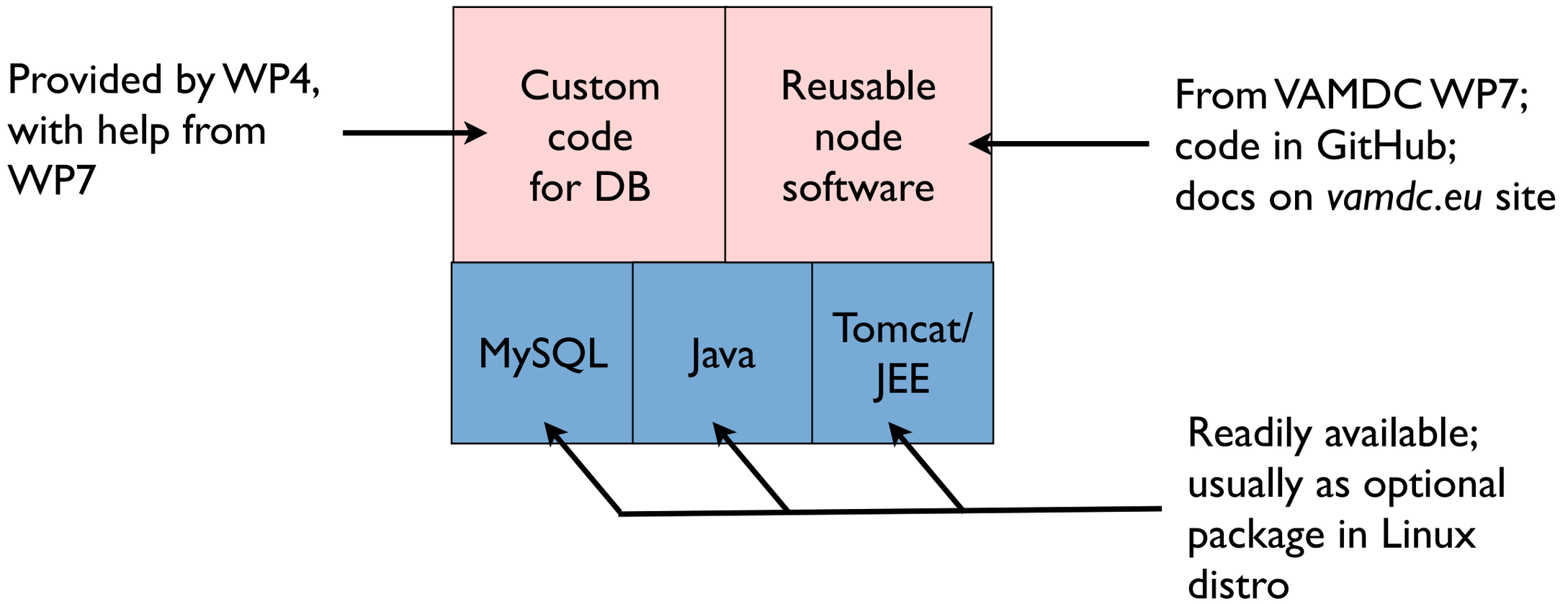
(And will it hurt?)

Node = database + web server + node software

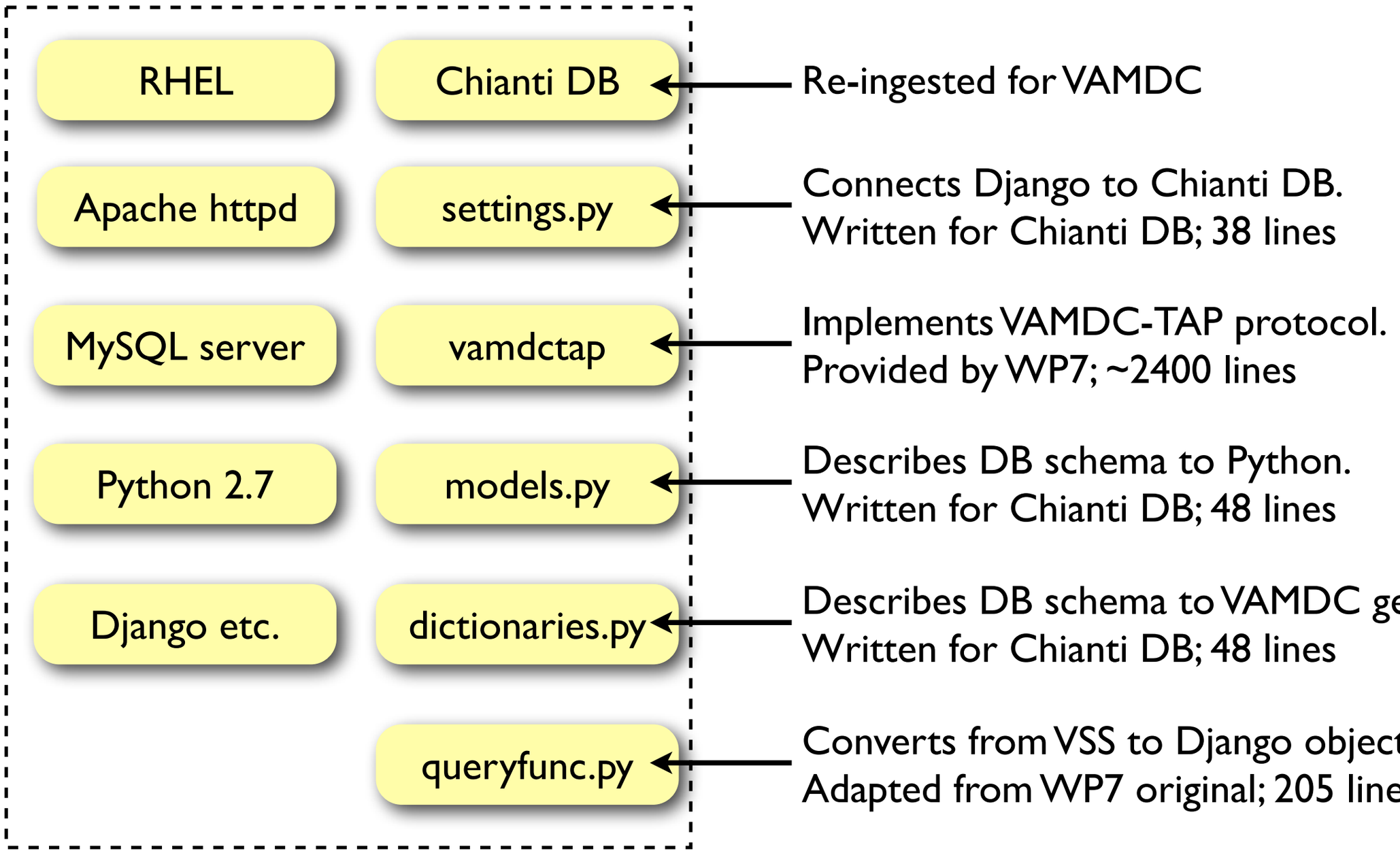
Node software in Python



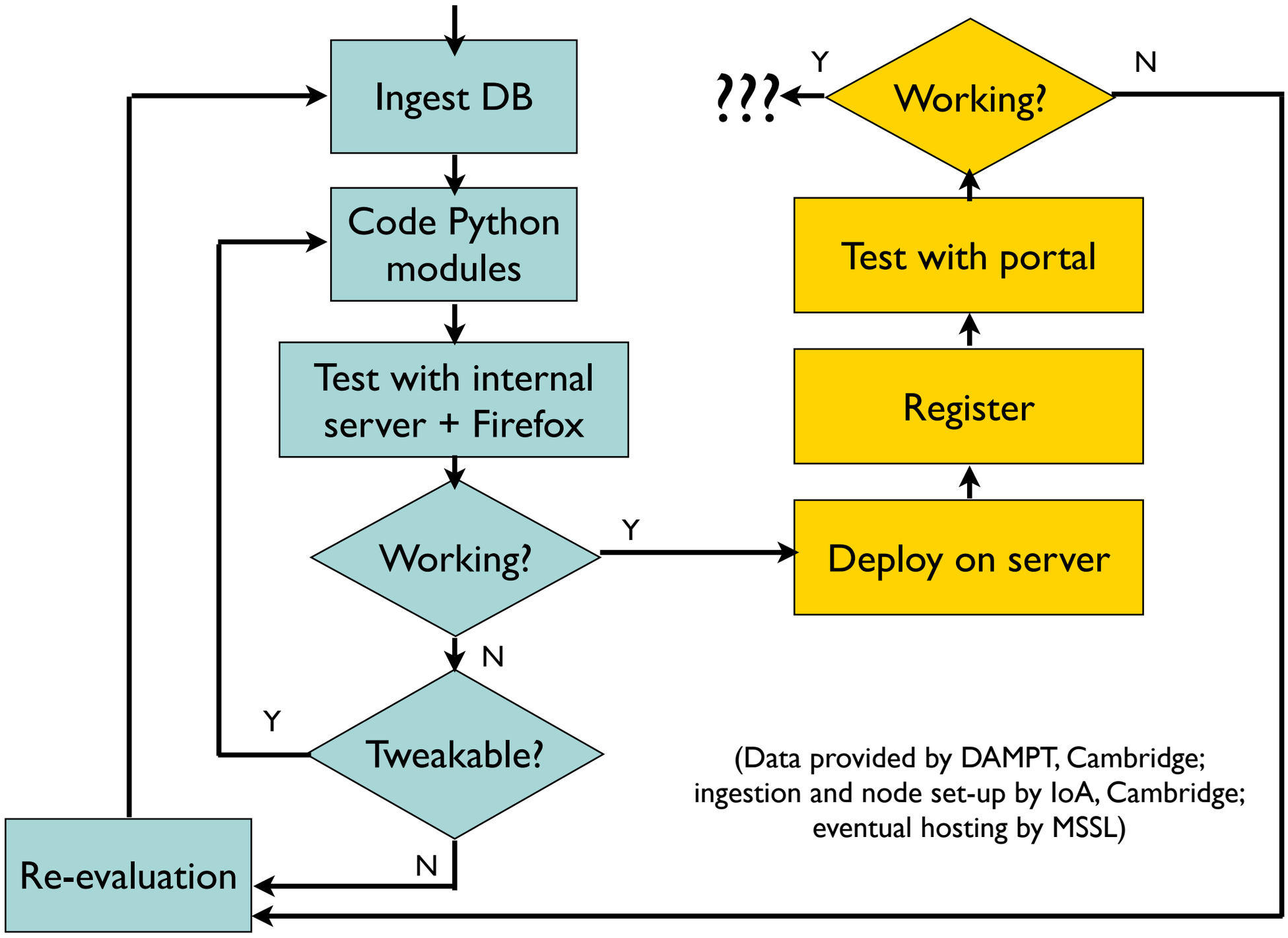
Node software in Java



Node example: Chianti



Chianti example (cont.)

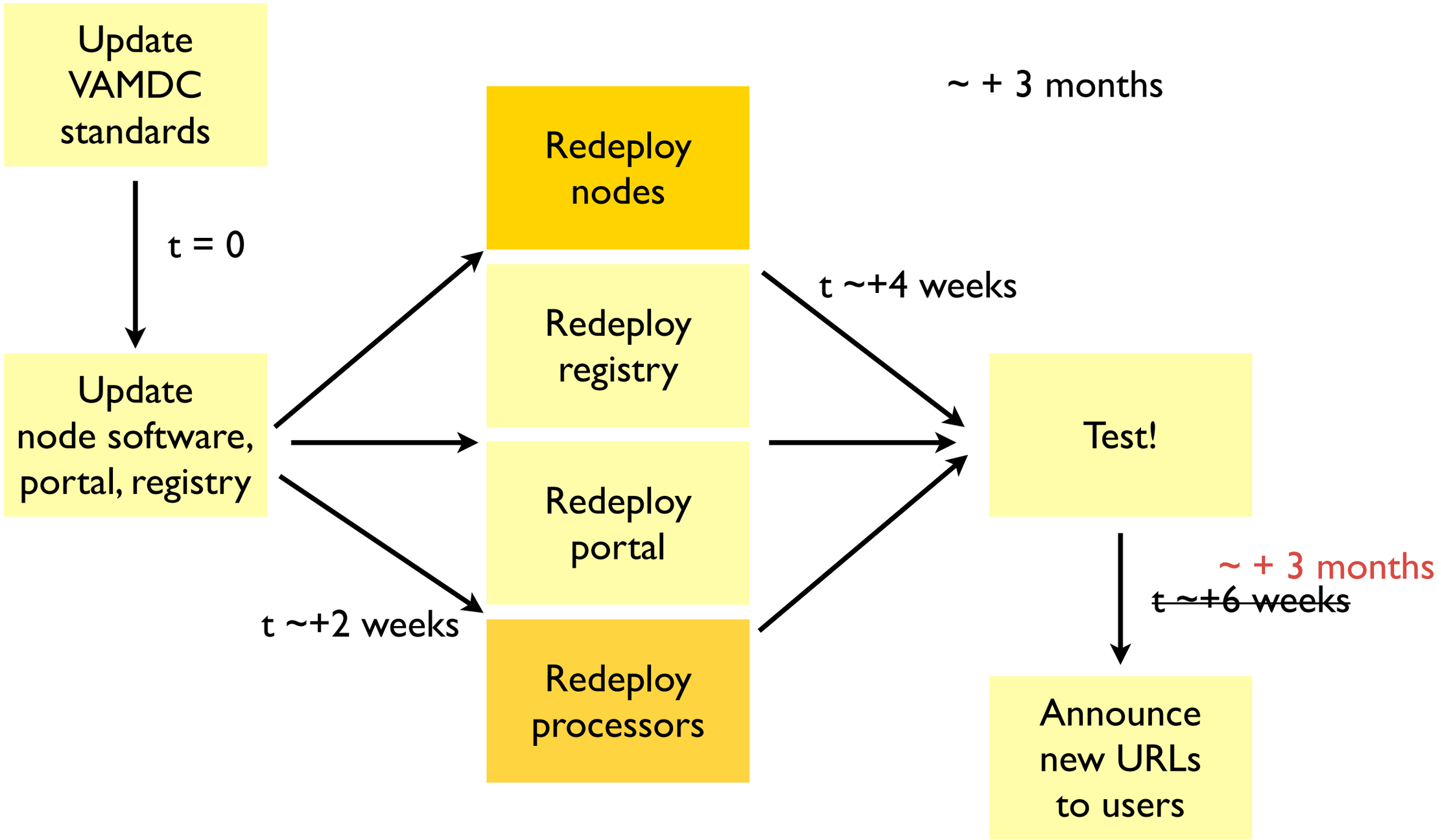


(Data provided by DAMPT, Cambridge;
ingestion and node set-up by IoA, Cambridge;
eventual hosting by MSSL)

System versions

- System version defined by standards version
- Three so far:
  - 11.05 (withdrawn)
  - 11.12 (current, released)
  - 12.07 (in preparation, to be released in 2013)
- Expect one new version per year from now on
- new standards ⇒ new deployments on new URLs

Annual updates of standards



The times for the transitions between stages are the approximate minimum times.
“Redeploy xxx” includes local testing, particularly by the owners of nodes.
The “test” box at the right is general integration-testing of the system.

Timeline of node registrations