Specialized Computing in the Agencies: Some Examples, Scenarios, and Lessons

Peter B. Meyer
Research Economist
BLS Office of Productivity and Technology
November 2016

Views expressed in this work do not represent the U.S. Bureau of Labor Statistics.

Goal: draw from Federal experience in specialized statistical computing

■ I asked Federal developers of novel statistical estimation systems about challenges they experienced relating to:

▶ Procurement

▶ Security

▶ Open source software

▶ Software development

Interview sources

These developers had made specialized, unique, admirable statistical estimation programs.

Interviews were informal. I drew also from past teams and work groups and advice.

Jonathan Lisic (NASS)	Ben Klemens (Treasury, Census)
Nathan Cruze (NASS)	Mike Messner (EPA)
Ted To (BLS)	Zhengwei Yang (NASS)
Ann Vega (EPA)	Jurgen Kropf (BLS)

3

Their projects (I)

Lisic (NASS) and Cruze (NASS) – Developed research statistical systems in R, C, SAS, and Stan, and transitioned them to production. One project: Bayesian estimates of crop yield based on many data sources, e.g. weather.

To (BLS) – runs focused cost-efficient Linux systems for economic research computation

Vega (EPA) – manager bringing on R/Shiny for both research and production

Kropf (BLS) – manager of systems bringing in data for monthly Employment Situation report; Java and SAS

Their projects (II)

Klemens (Treasury, Census) – developed Apophenia library; Tea for imputation; tax revenue estimation

Messner (EPA) – makes hierarchical Bayesian model estimates of water quality based on samples

Yang (NASS) – co-developer of CropScape and VegScape
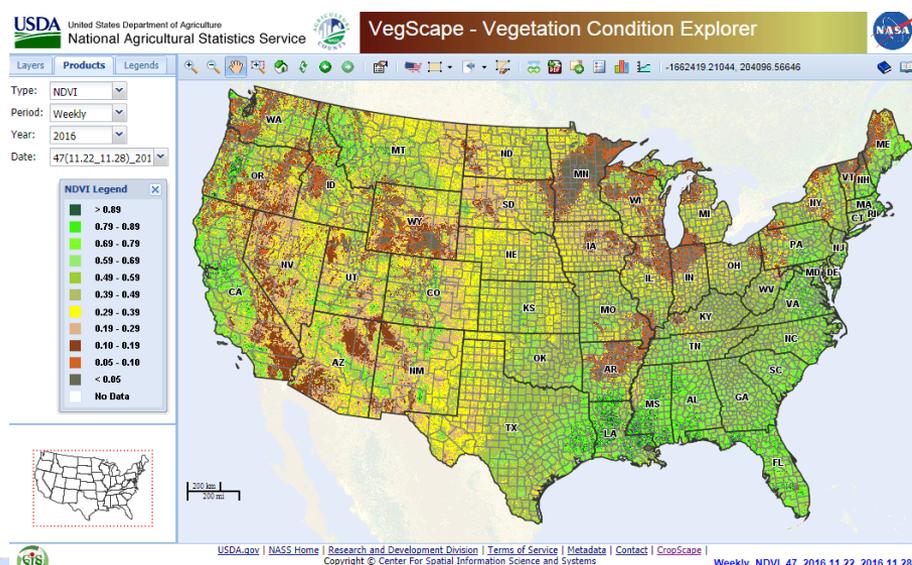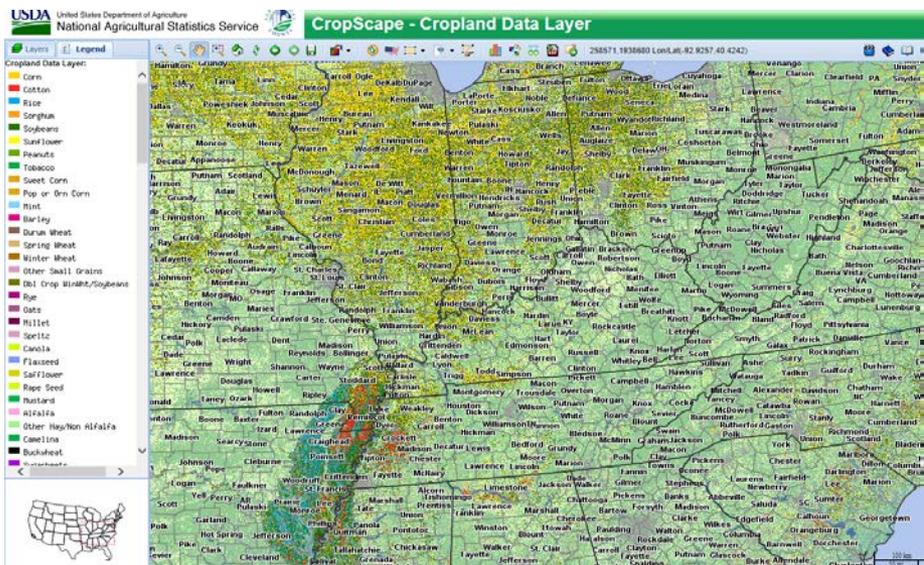
Theme: Constraints on procurement, installation, and licensing

- Developers often not authorized to install software themselves
- Must be approved/certified for use in the agency
  - Has been a severe problem for open-source software
  - Who supports it?
  - Where's Section 508 accessibility VPAT?
- "Terms of service" not always agreed between suppliers and agencies
  - Open source software can help
- Outside partners can help

Theme: Research vs. Production (I)

- A "two cultures" kind of problem
  - Production is prioritized generally
  - NY Fed has separate networks for research and production

- "Why do you need THAT?"   (e.g. for research)
  - ▶ "Need" – in research, uncertainty and ambiguity
    - – We try new tools to learn, improve, and be able to hire
  - ▶ Asking for new software creates creates work
    - – Delays, costs, frictions, research schedule is uncertain

Contrast: Production software run schedule

		DEEPS Manual Links	1st Closing Production Schedule				
1			**1st Closing Production Schedule**				
2	DIDD Calendar	**DEEPS Manual Links**	**Year: 2016**		**Operator(s):**		
3/4	**AEPP DWI**						
5			**Month:**				
6	**AEPP Manual**	Usual Start Time			**Processing Date and Time--**		
7	**Weekday**		**Job Description**	**Started:**	**Completed:**		**COMMENTS**
9	**Friday**	15:00	**ALL/** Production Registry Update				
10		15:30	**ALL/** Registry Synchronization				
11		16:00	**ALL/** Microdata Synchronization (Usable, TABNAC10, Post Processing Changes)				
12		18:00	**ALL/** Microdata Update & Screening Tests				
13		20:30	**SPECIAL/** Transpose Microdata & Create HRRDB/CMI Reports				
14		20:40	**ALL/** Prepare Datasets for Screening Review Applet: (Special Circumstances)				
15		21:00	**ALL/** Copy SAS datasets from Production to COOP Environment (MODIFIED WITHIN SPECIFIC DAYS)				
17	Monday	16:30	**ALL/** Reintegrate Microdata Post Analyst Screening Review				
18		16:45	**ALL/** Registry Synchronization				
19		17:15	**ALL/** Microdata Synchronization (Usable, TABNAC10, Post Processing Changes)				
20		18:00	**ALL/** Microdata Update & Screening Tests				
21		20:30	**SPECIAL/** Transpose Microdata & Create HRRDB/CMI Reports				
22		20:45	**ALL/** Generate Preliminary AEPP Estimates				
23		21:05	**ALL/** Preliminary Monthly Estimate DB Updates for Non GME datatypes				
24		21:25	**ALL/** Create HEDB extract and transfer				
		21:25	**FIRST/** Generate Seasonally Adjusted Published at 1st Closing Comparison Table (Run X12, Produce Table) {No Database Update- V2.0}				

Sheet tabs: **1st Closing** | 2nd Closing | 3rd Closing | Real Earnings | Real Earnings - JAN | Partial | 1st Closing Holiday-Monday | ...

Theme: Research vs. Production (II)

- Turbulent library/packages add challenges in production use

- Lisic is documenting guidance for transitioning research projects in R to production

Theme: Developers of advanced systems are isolated or unsupported

- Often innovators seem alone, when org is committed to an existing production system
- To address this we have
  - FCSM research and policy conferences
  - SCOPE community of practice across stats agencies
  - Statipedia – shared reference/workspace across agencies
  - ICSP and OIRA – connecting offices/institutions
- They sometimes need support from the top
  - Aka "top cover" (Burke and Dennehy 2015)
  - Procurement depends on this

10

Theme: Software engineering (I)

- Statistical agencies don't always treat this as a core program job
  - ▶ Instead, thinking statistics or economics are core; software tasks handed off to IT dept or contractors
  - ▶ Addressing engineering issues is "a necessary evil" (?!)
    - – "We're not software engineers." (different from NASA)

- Issues often involve "systems thinking"
- **Error handling -** identifying sources of a problem
- Simple focused **interface**, not a statistical package
- Installation, upgrades of our own sw

- **Interfaces** between languages, software layers
- **Publishing** libraries/packages
- **Cross-platform** transitions, interop
- Data from many sources & formats
- **Modularity**

Theme: Software engineering (II)

- After hiring economists, agencies may re-discover periodically that they are not software engineers

- Adaption: treat it as a core function: train, plan, hire, reward, get tools for it
  - ▶ Much of this happens; maybe more is needed

- New Federal Open Source Code policy may help
  - ▶ Helps staff team up across agencies, and beyond

BLS

Themes again

- Procurement, licensing, installation constraints
- Research and production differences
- Statistical software developers can be isolated
- Our agencies do software engineering

Contact Information

Peter B. Meyer

Research Economist

Office of Productivity and Technology

www.bls.gov/dpr

202-691-5678

meyer.peter@bls.gov

BLS