Deploying KNIME to the Enterprise
Reshaping Data & Architecture for Healthcare
We were founded in 1876 by Colonel Eli Lilly, a man committed to creating high-quality medicines that met real needs in an era of unreliable elixirs peddled by questionable characters. His charge to the generations of employees who have followed was this: “Take what you find here and make it better and better.”

More than 140 years later, we remain committed to his vision through every aspect of our business and the people we serve starting with those who take our medicines, and extending to health care professionals, employees and the communities in which we live.

**OUR PROMISE**
Lilly unites caring with discovery to make life better for people around the world.

**OUR MISSION**
We make medicines that help people live longer, healthier, more active lives.

**OUR VISION**
We will make a significant contribution to humanity by improving global health in the 21st century.

**OUR VALUES**
Integrity, excellence and respect for people.
Stable
Reliable, Reproducible, Transparent for development work

Secure
Providing the right data at the right time to the right people

Specific
The right analytics for my specific business problem

Scalable
Can it handle the volume and variety of our data
Reshaping Data: Example Production Capabilities

Stable
Reliable, Reproducible, Transparent for development work

Secure
Providing the right data at the right time to the right people
• Rapid / ‘Agile’ development
• 1 Production KNIME version per year (2.6 / 2.9 / 2.12 / 3.3)
• >130 internal nodes
• 340 Features/Bugs in 2016 (heavily focused on data and data security – NTLM/OAuth)
• 6 internal updates
KNIME Open Source install

Add Version Fixed Vendor and Community Plugins

Add Lilly internal Plugins

Create new Lilly installer

Add Lilly specific config file specific to Lilly laptop defining mem usage etc.

Mirror Lilly installer to US/UK/Spain

Packaging
• 100’s of nodes for data manipulation
• Data augmentation using calls to ‘best of breed’ cluster side algorithmic search methods
• Encapsulation of best analytics practice for problem x, towards automated reasoning
“Here be dragons”

- Download e.g.: Medline, JAPIC-Q, JMed
- Clean, De-duplicate, Augment, Join
- Text Mine (Cogito for taxonomies and ontologies)
- Select, Alert (Email)

...simple so far
“Here be dragons”
GxP!
Specific

The right analytics for my specific business problem
Automated Structure Verification:
Lilly Screening Collection representing all compounds made internally or purchased for projects. Routinely used to screen new disease targets for possible starting points.

1. Extract data from multiple DB’s to prepare instrument run files
2. Acquire data from instrument
3. Analyse using call to external tool server MestreLabs
4. Check results against criteria to score Pass/Fail/Review
5. Publish results to the Corporate DB

Multi year project to retrospectively check quality of all compounds run in months using KNIME
Complex and specific scientific needs only available from one vendor integrated through dedicated server build, serving predictions to KNIME via webserver API (SOA)

KNIME now analyses all new compound registrations daily (scheduled server task) to check (a) Analytical data is available and (b) is consistent with the registered structure and (c) emails submitter on Fail
Reshaping the Architecture: >140 RESTful Webservice API endpoints deployed in 2-3 years primarily for data retrieval and augmentation by KNIME
Many legacy, new and vendor applications deployed internally with webservice endpoints for KNIME integration. Team also built many API’s (& nodes) for many data sources.
Multiple other internal systems now utilising KNIME, but also independently accessing the SOA layer for programmatic data access.
Integrating via SOA:

- New KNIME versions
- Internal feature requests (tools/data)
- Data security
- Data-load/Volume
- Streaming
Timeouts and retries

* Thanks: Jason Ochoada
URL Over-ride and data truncation:

URL from node configuration is overriding the default value from the preferences.

data truncated
Trade offs: Downstream configure…

- Internal Mirror of many external data sources for secure querying
- Between version API changes (schema upgrades) break nodes if return columns are hard coded
- Alternate flexible column specification on return breaks downstream configure
Bottlenecks and frequent https calls:
How to break streaming:

Heavy compute calls to custom internal algorithmic code push a lot of data back and forth between compute layer and node.

Even with a high number of HPC calls, frequent breaks for data processing/filtering prevents application of streaming.
Scalable
Can it handle the volume and variety of our data
Video feed Meta data for KNIME Analytics

KNIME processes and integrates additional data streams alongside video feed

User front end (INSTANT) provides interactive analysis of meta data and ad-hoc analytics via KNIME automation
Scaling out Orchestration with Parallel Chunk

KNIME orchestrates file movement from secure file store, runs Lilly proprietary signal processing code then writes meta data (result) back to disk or DB:
- Custom node to correctly handle protected file shares
- Call external Lilly code
- Loops & schedule to automate
Real-world evidence helps us discern all the ways a medicine can be used. elil.ly/534g #pharma
How is an existing medicine X and treatment Y working in the health care system?

Querying Netezza (IBM Pure Data) Store: Metrics, Sentiment, Outcome…

Stuart Morton / Jeramy Brewster / Malika Mahoui
Research:
Dave Evans
Lewis Vidler
David Thorner
Jason Ochoada
Christos Nicolau
Beth Wright
Gary Sharman
Sylvain Demanze
Natalie Franklin
Steve Green

Research IT:
Luke Bullard
Thomas Wilkin
Nathan Roberts
Stuart Morton
Malika Mahouï
Jeramy Brewster

Other API & System Owners:
Hongzhou Zhang
Jim Hughes
Christos Nicolaou
James Rimell

… and the many other users and developers who broke things, fixed things, and did neat things with KNIME