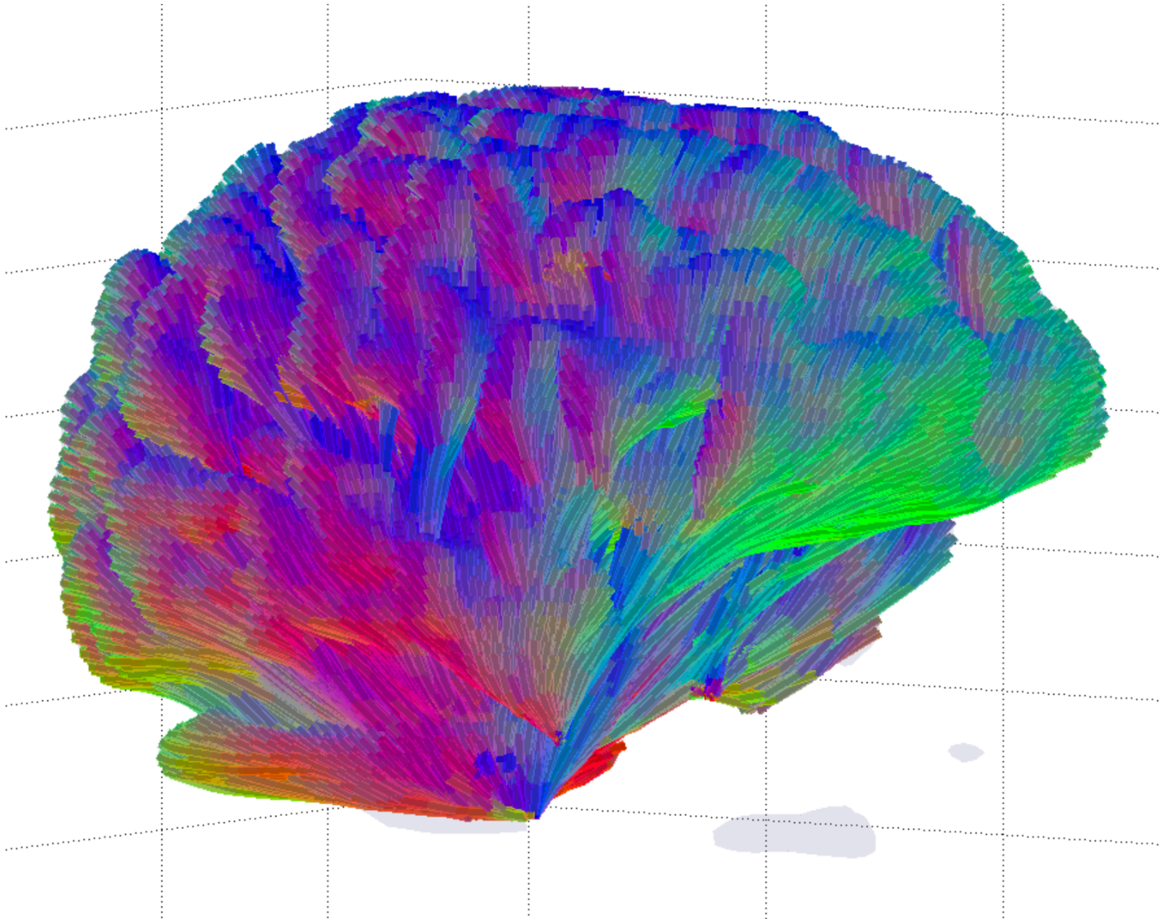
SESSION SUMMARY

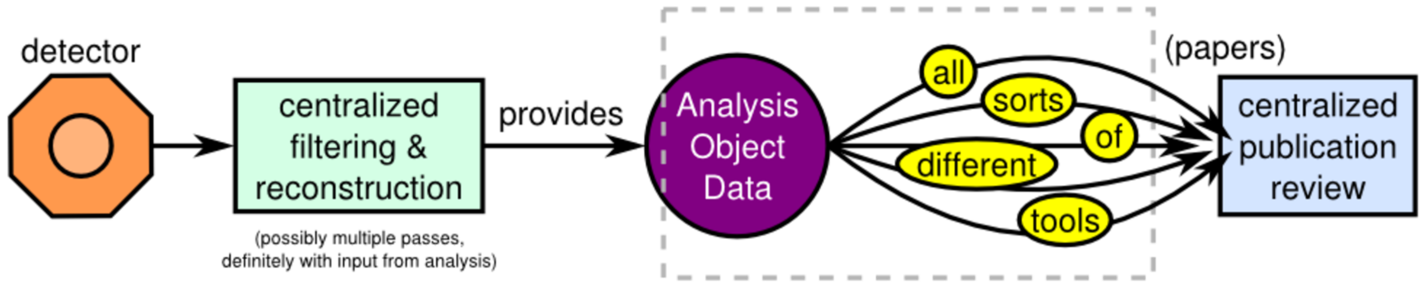
Data Intensive Analysis Tools, Visualization



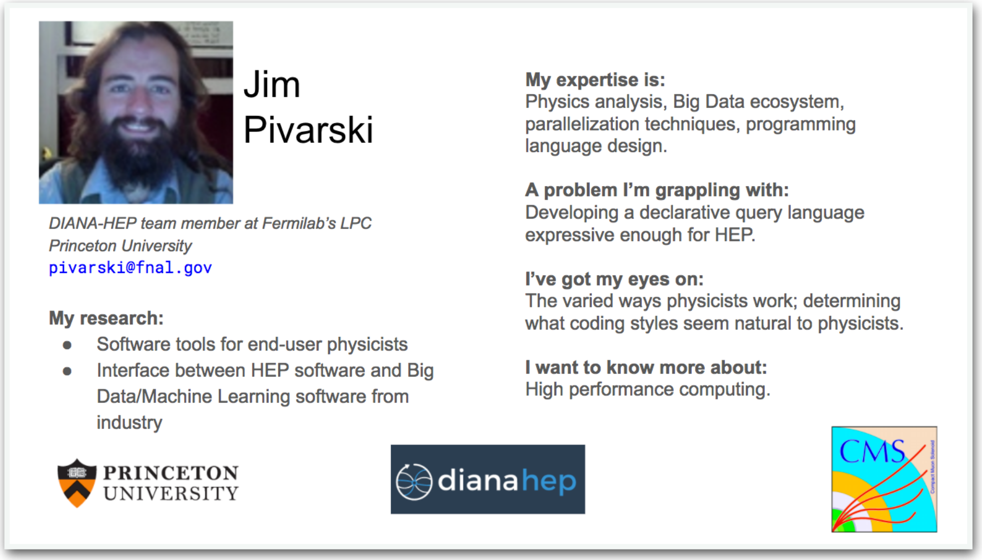
Fernanda Psihas

Indiana University

Session Intro / Summary


detector
centralized filtering & reconstruction
provides
Analysis Object Data
all
sorts
of
different
tools
(papers)
centralized publication review

(possibly multiple passes, definitely with input from analysis)


Jim Pivarski

DIANA-HEP team member at Fermilab's LPC
Princeton University
pivarski@fnal.gov

My research:

- Software tools for end-user physicists
- Interface between HEP software and Big Data/Machine Learning software from industry

My expertise is:
Physics analysis, Big Data ecosystem, parallelization techniques, programming language design.

A problem I'm grappling with:
Developing a declarative query language expressive enough for HEP.

I've got my eyes on:
The varied ways physicists work; determining what coding styles seem natural to physicists.

I want to know more about:
High performance computing.

PRINCETON UNIVERSITY

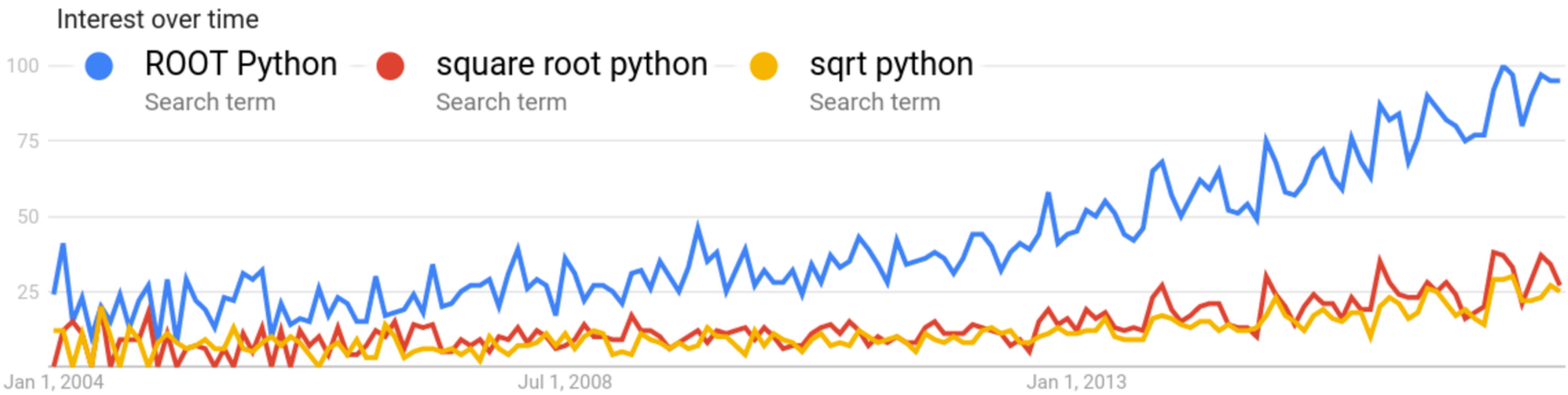
dianahep

CMS

What are the big or new ideas?

What are indicated R&D paths?

What would you like to see in an Software Institute?

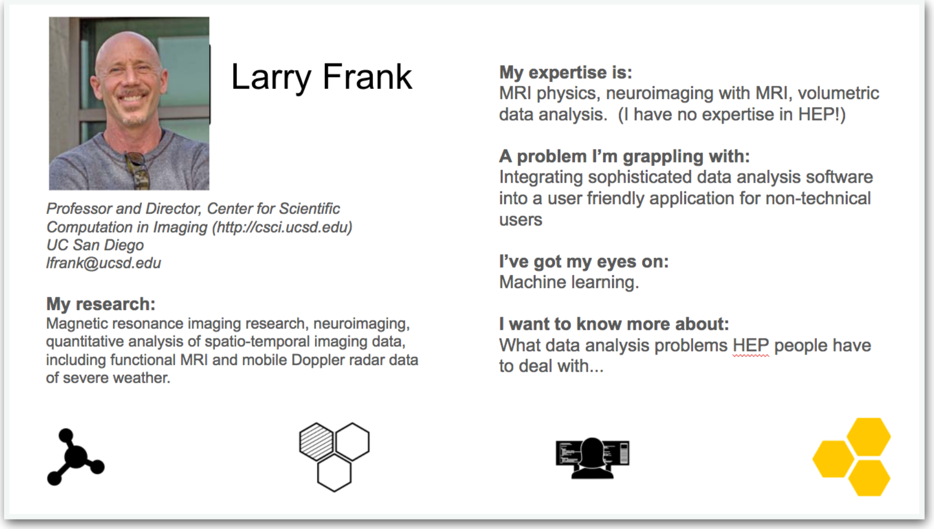


Challenges in Neuroimaging

Looking for spatial-temporal patterns in neural structure.

Many common challenges with HEP data intensive analysis.

Discussion followed regarding collaboration opportunities, data



Larry Frank

Professor and Director, Center for Scientific Computation in Imaging (http://csci.ucsd.edu)
UC San Diego
lfrank@ucsd.edu

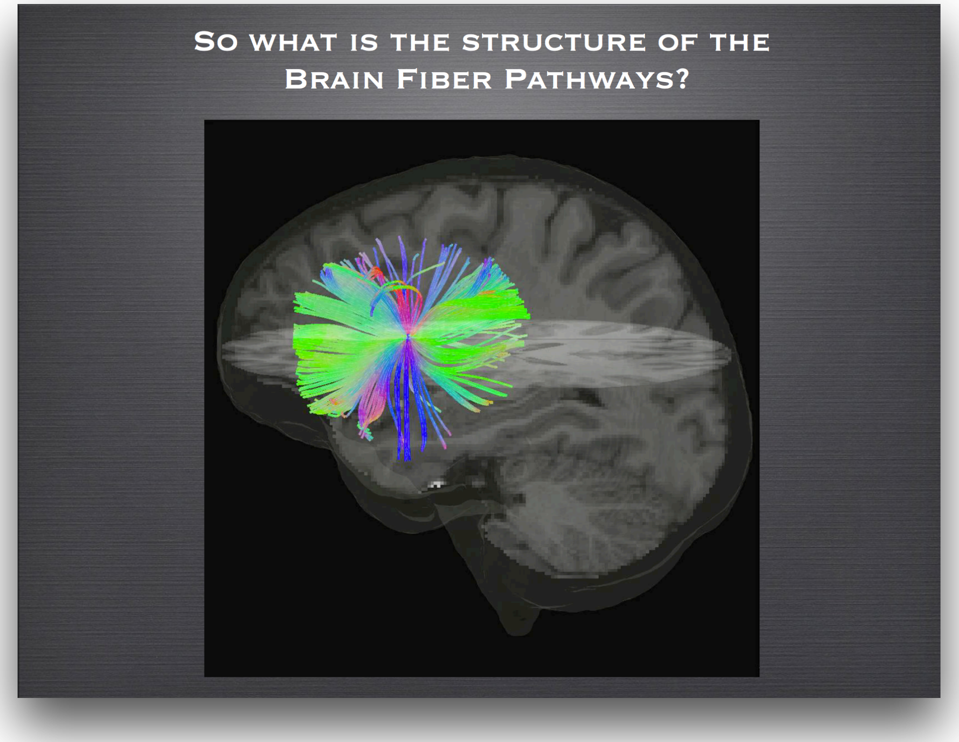
My research:
Magnetic resonance imaging research, neuroimaging, quantitative analysis of spatio-temporal imaging data, including functional MRI and mobile Doppler radar data of severe weather.

My expertise is:
MRI physics, neuroimaging with MRI, volumetric data analysis. (I have no expertise in HEP!)

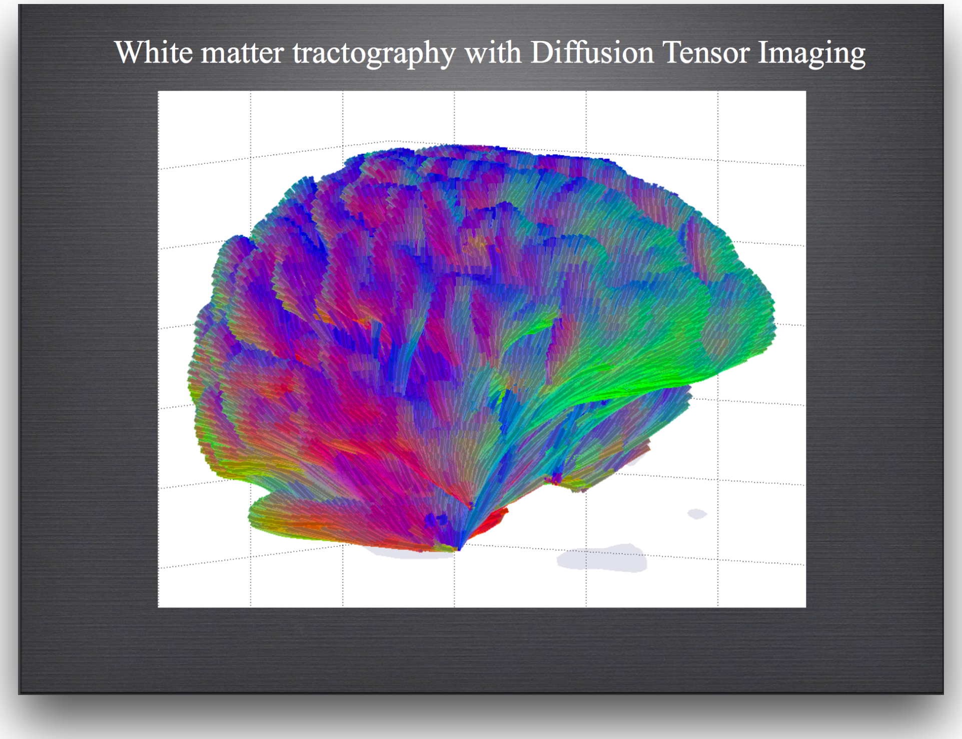
A problem I'm grappling with:
Integrating sophisticated data analysis software into a user friendly application for non-technical users

I've got my eyes on:
Machine learning.

I want to know more about:
What data analysis problems HEP people have to deal with...

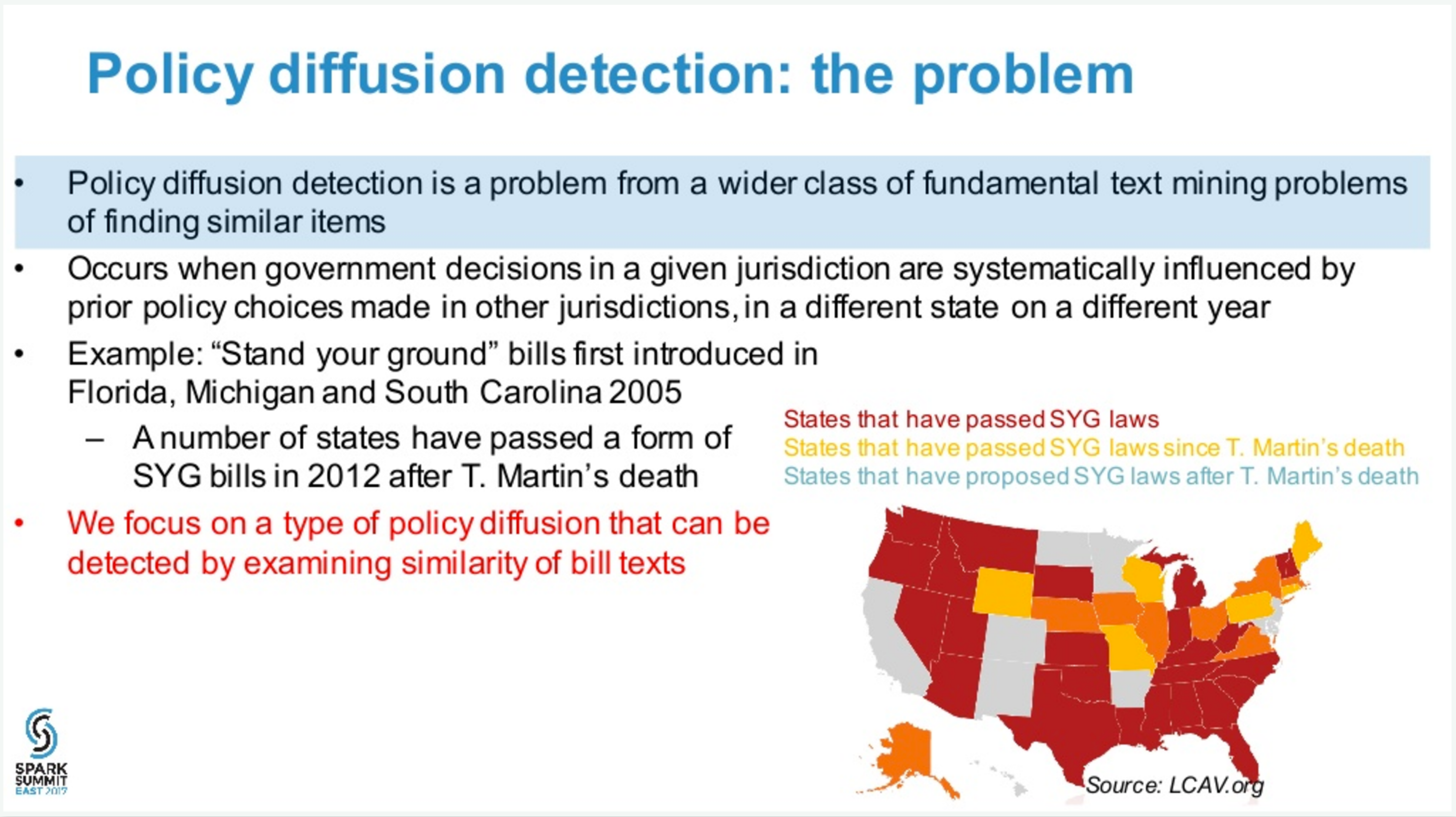


SO WHAT IS THE STRUCTURE OF THE BRAIN FIBER PATHWAYS?



White matter tractography with Diffusion Tensor Imaging

ML pipelines with Spark

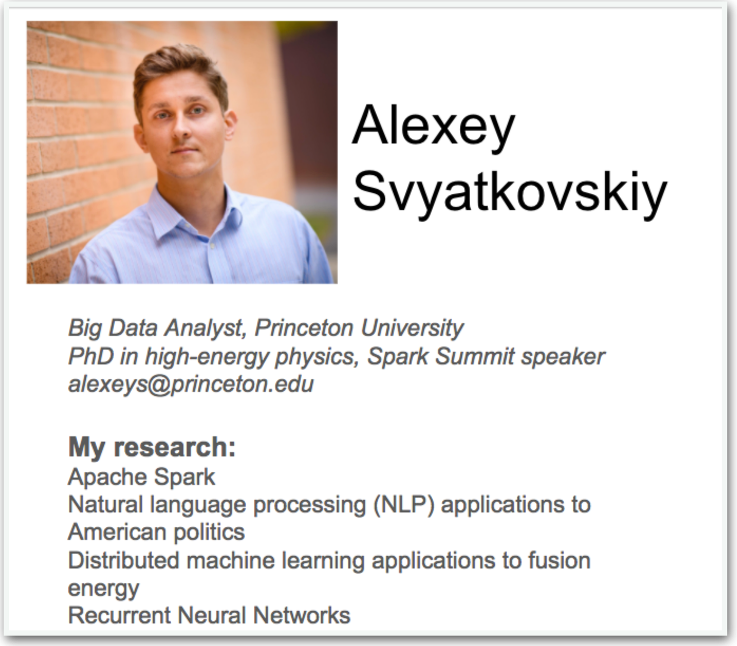


Policy diffusion detection: the problem

- Policy diffusion detection is a problem from a wider class of fundamental text mining problems of finding similar items
- Occurs when government decisions in a given jurisdiction are systematically influenced by prior policy choices made in other jurisdictions, in a different state on a different year
- Example: “Stand your ground” bills first introduced in Florida, Michigan and South Carolina 2005
  - A number of states have passed a form of SYG bills in 2012 after T. Martin’s death
- We focus on a type of policy diffusion that can be detected by examining similarity of bill texts

States that have passed SYG laws
States that have passed SYG laws since T. Martin’s death
States that have proposed SYG laws after T. Martin’s death

Source: LCAV.org



Alexey Svyatkovskiy

Big Data Analyst, Princeton University
PhD in high-energy physics, Spark Summit speaker
alexeys@princeton.edu

My research:
Apache Spark
Natural language processing (NLP) applications to American politics
Distributed machine learning applications to fusion energy
Recurrent Neural Networks

Large-scale text processing pipeline with Spark ML and GraphFrames

Showed evaluation of Apache Spark to Study Policy Diffusion (when government decisions are influenced by prior policy choices in nearby jurisdictions)

XENON1T,Open Source & python

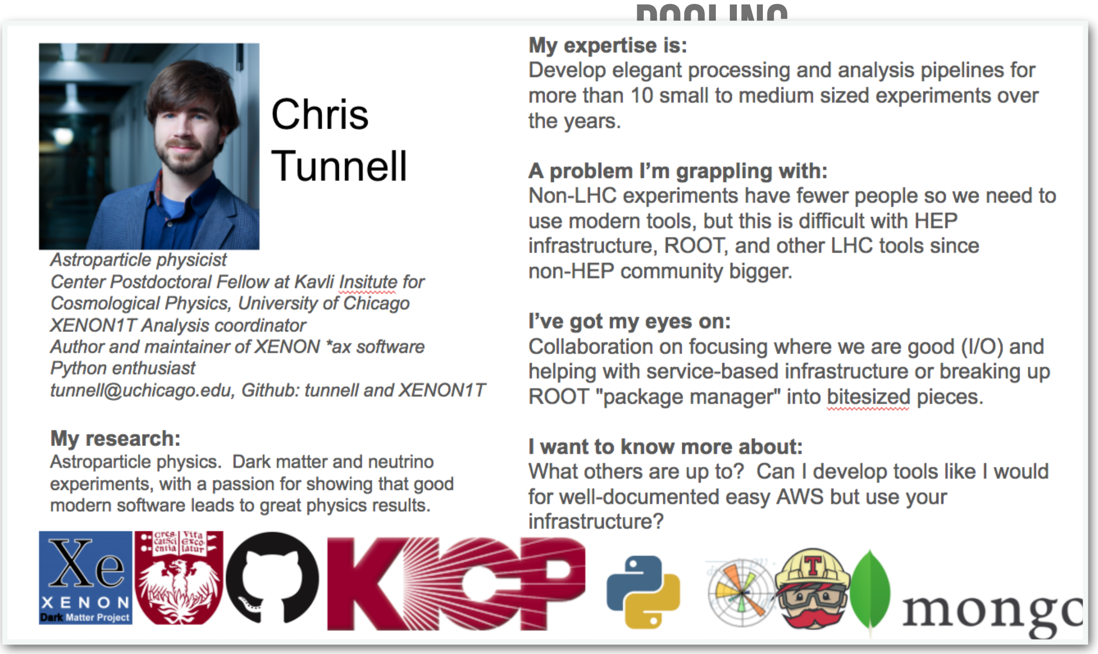
Focus on python and HDF5

“Wrote XENON DAQ/processing/analysis software to get a “real” job...but it worked!”

Effort to move away from domain-specific tools in order to get transferable skills (expertise with specific tools)

Focus on PyData but lots of discussion of other efforts emerged.

LRN
POOLING


Chris Tunnell

Astroparticle physicist
Center Postdoctoral Fellow at Kavli Insitute for Cosmological Physics, University of Chicago
XENON1T Analysis coordinator
*Author and maintainer of XENON *ax software*
Python enthusiast
tunnell@uchicago.edu, Github: tunnell and XENON1T

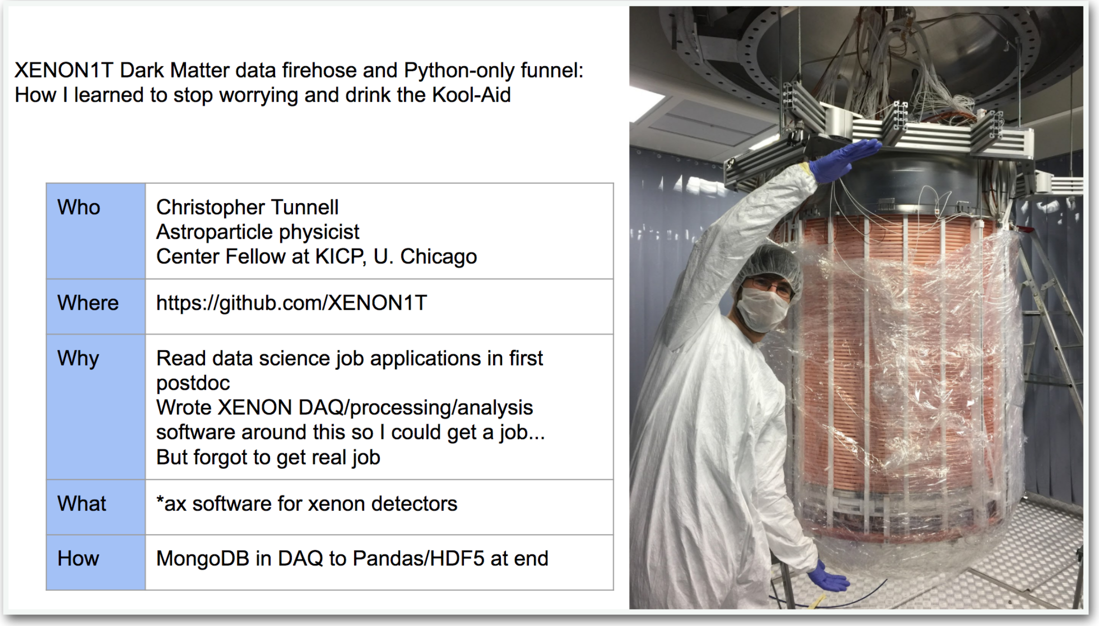
My research:
Astroparticle physics. Dark matter and neutrino experiments, with a passion for showing that good modern software leads to great physics results.

My expertise is:
Develop elegant processing and analysis pipelines for more than 10 small to medium sized experiments over the years.

A problem I’m grappling with:
Non-LHC experiments have fewer people so we need to use modern tools, but this is difficult with HEP infrastructure, ROOT, and other LHC tools since non-HEP community bigger.

I’ve got my eyes on:
Collaboration on focusing where we are good (I/O) and helping with service-based infrastructure or breaking up ROOT "package manager" into bitesized pieces.

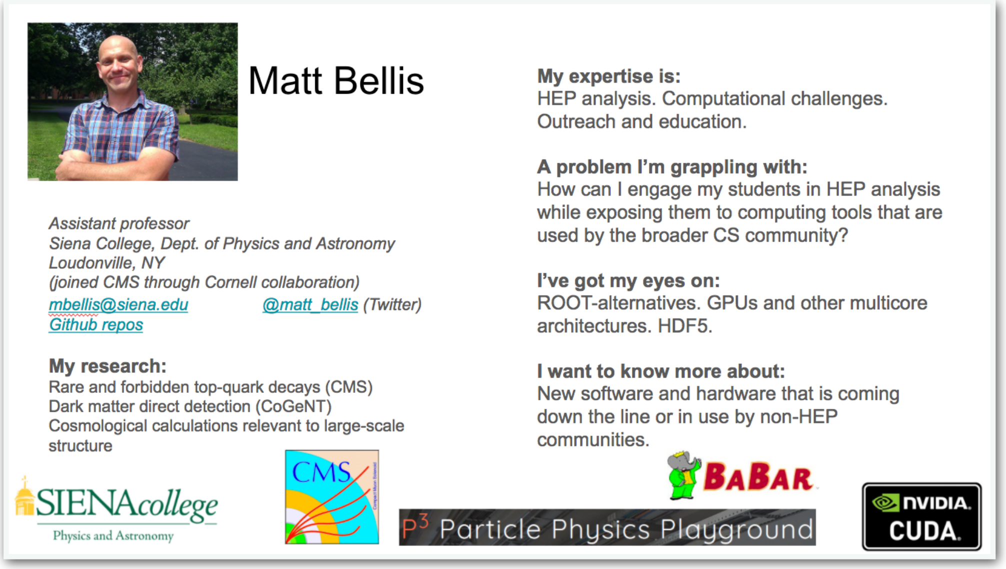
I want to know more about:
What others are up to? Can I develop tools like I would for well-documented easy AWS but use your infrastructure?


XENON1T Dark Matter data firehose and Python-only funnel:
How I learned to stop worrying and drink the Kool-Aid

Who	Christopher Tunnell Astroparticle physicist Center Fellow at KICP, U. Chicago
Where	https://github.com/XENON1T
Why	Read data science job applications in first postdoc Wrote XENON DAQ/processing/analysis software around this so I could get a job... But forgot to get real job
What	*ax software for xenon detectors
How	MongoDB in DAQ to Pandas/HDF5 at end

Fernanda Psihas

ROOT-less workflow?



Matt Bellis

Assistant professor
Siena College, Dept. of Physics and Astronomy
Loudonville, NY
(joined CMS through Cornell collaboration)
mbellis@siena.edu *@matt_bellis (Twitter)*
Github repos

My research:
Rare and forbidden top-quark decays (CMS)
Dark matter direct detection (CoGeNT)
Cosmological calculations relevant to large-scale structure

My expertise is:
HEP analysis. Computational challenges. Outreach and education.

A problem I'm grappling with:
How can I engage my students in HEP analysis while exposing them to computing tools that are used by the broader CS community?

I've got my eyes on:
ROOT-alternatives. GPUs and other multicore architectures. HDF5.

I want to know more about:
New software and hardware that is coming down the line or in use by non-HEP communities.

Will ROOT be used for HL-LHC?

If not, what will we use for fileld? Development environment? Language/ libraries?

Need test cases now to see what works and what doesn't. Maybe ROOT is the right answer!

Should we write code to harness maximum benefits of language, rather than writing C-like Python or Python-like C?

Should we minimize inheritance to maximize sustainability?

ROOT related discussion

ROOT I/O format

Can we retain the benefits of ROOT I/O for dealing with HEP data?

Lots of discussion about HDF5 and data formats. Should we be investing on a replacement to the ROOT I/O?

ROOT analysis framework

How does a change of input formats affect the existing analysis workflows?

ROOT is a package manager for shipping physics code in pre-boost library era. Can it be modularized?

Arising questions to feed into goals of the institute

Overarching question is about how to (and to what extent) implement analysis tools currently in use in industry and how to handle the interplay with our tools/data.

Can S2I2 recommend to ROOT/experiment/DIANA/etc teams to provide various adapters, e.g. ROOT->CSV, ROOT->NumPy, ROOT->Avro, etc.

Can the institute invest on development of middleware which translate ROOT into another (bring/new) data format natively.

How do we justify teaching tools which are domain-specific when 9/10 people leave physics and the world past us in the last decade?

Arising questions to feed into goals of the institute

How do we deal with the incompatibility with new analysis tools, but benefit from the subset of tools that HEP does well (I/O)?

Can the institute engage in training developers on new analysis tools and data formats? (Efforts exist from individuals but there is no cohesive effort in place)

In the case of new data formats/analysis tools arising, can the institute invest on developing tools which allow backwards compatibility and reproducibility?

On collaboration. How can the institute engage the community in collaborations with the CS community?