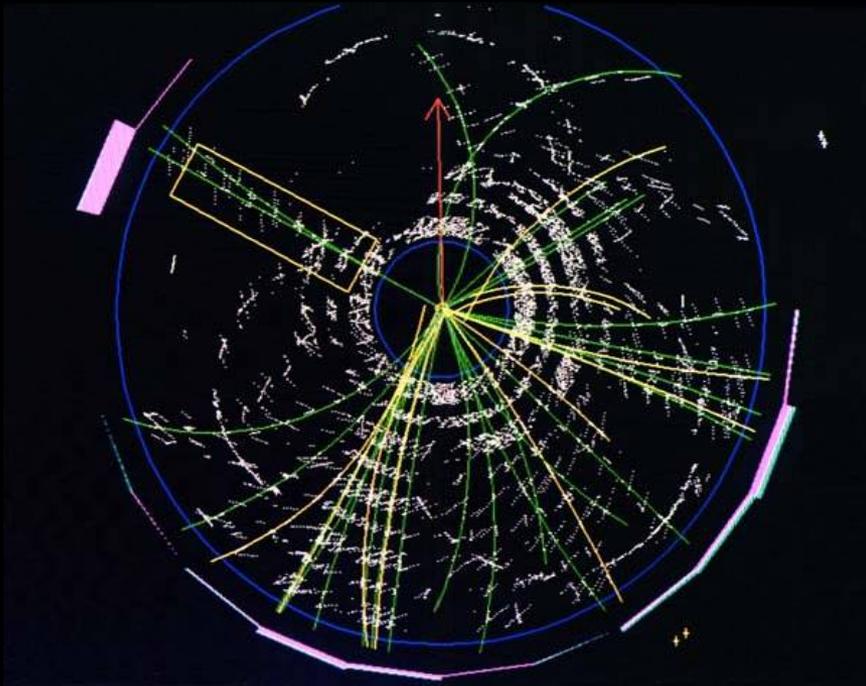


Data Preservation in Particle Physics

Robert Roser
Fermilab

SuperComputing 2012



Contact Info.

630-399-2609 (c)

roser@fnal.gov

Why is Fermilab so involved:

Fermilab is the only United States National Particle Physics Lab

- Our mission is to enable the US community to tackle the most fundamental physics questions of our era
- Decades long history of accomplishments in Big Data storage, management and access for Science.
- Experience Integrating and providing the “Glue” for the Community across the US Universities and other laboratories for both national and international programs.



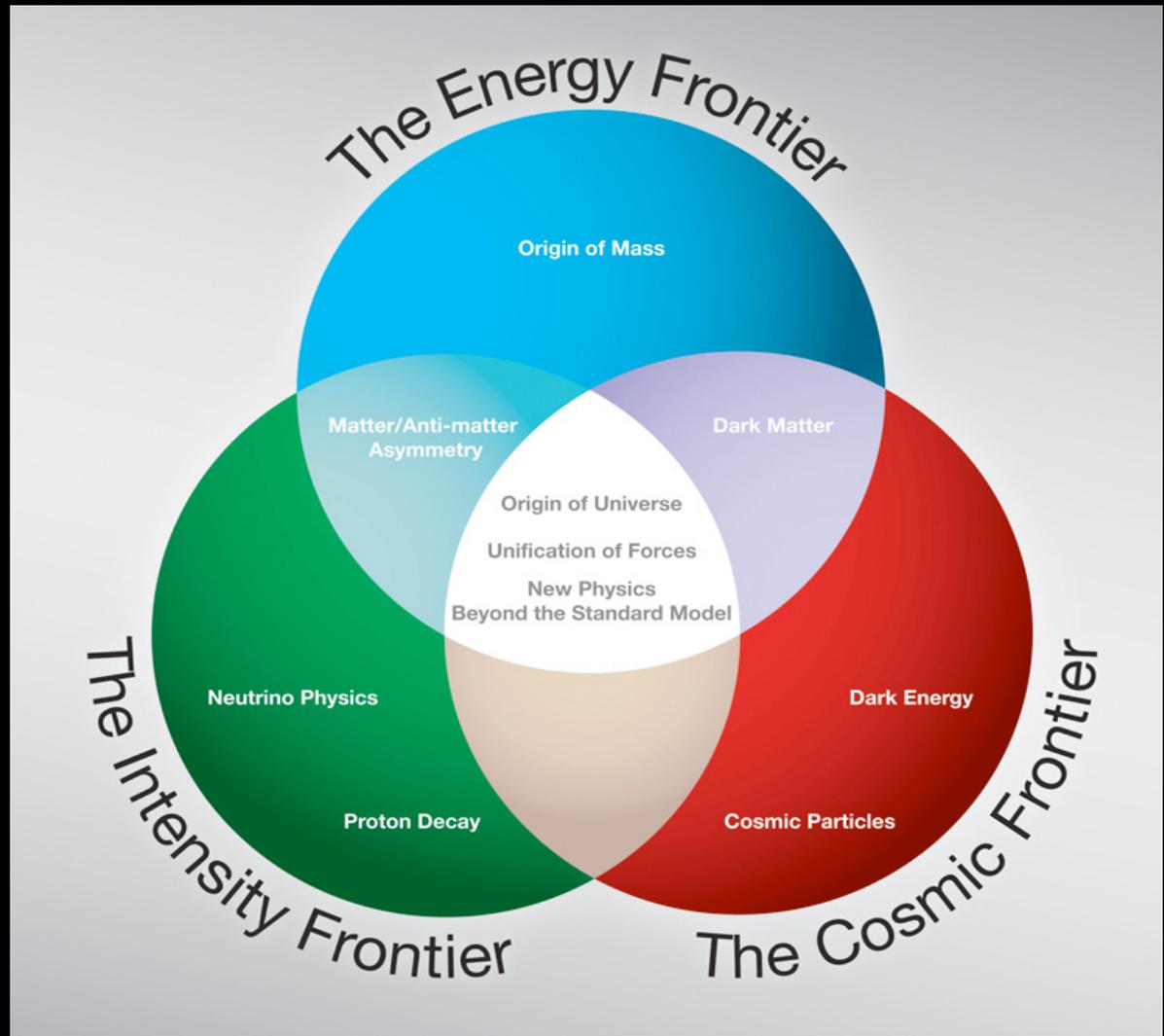
US High Energy Physics Program

Built on:

Accelerators

Detectors

Computing



Fermilab Characteristics

- ~1700 employees; ~\$400M DOE Funded with help from NSF
- 2300 users (visiting scientists)
- 6800 acres, park-like site
- Tevatron: Flagship machine for 25 years – now off and lab is in transition to its new priority



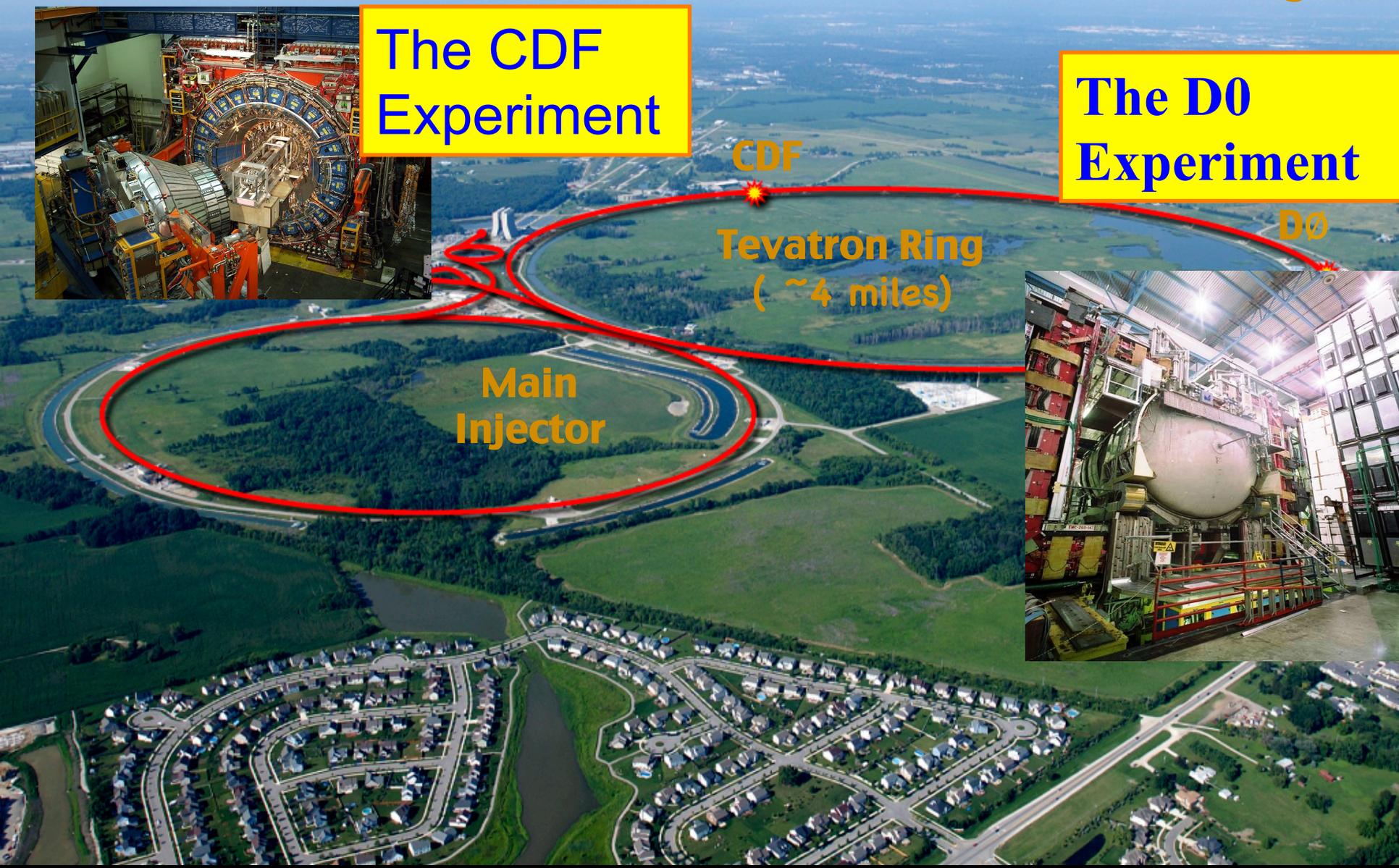
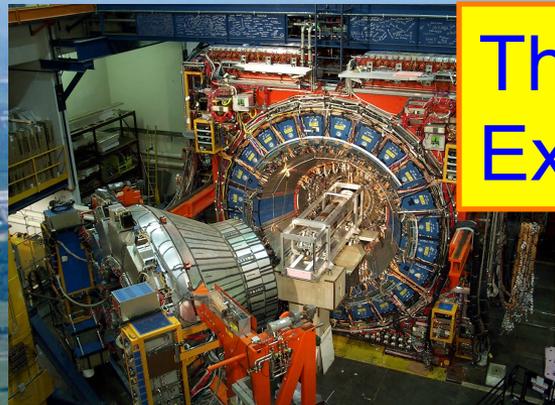
- Highest intensity neutrino beams (low and high energy)
- World class astrophysics , particle theory and computation programs
- Advanced detector and accelerator technology R&D

America's Most Powerful Accelerator: The Tevatron and its "Modern Marvels" the experiments

Chicago

The CDF
Experiment

The D0
Experiment



CDF

Tevatron Ring
(~4 miles)

D0

Main
Injector

Trying to Answer the Important Questions....

The sense of mystery has never been more acute

Where does mass come from?

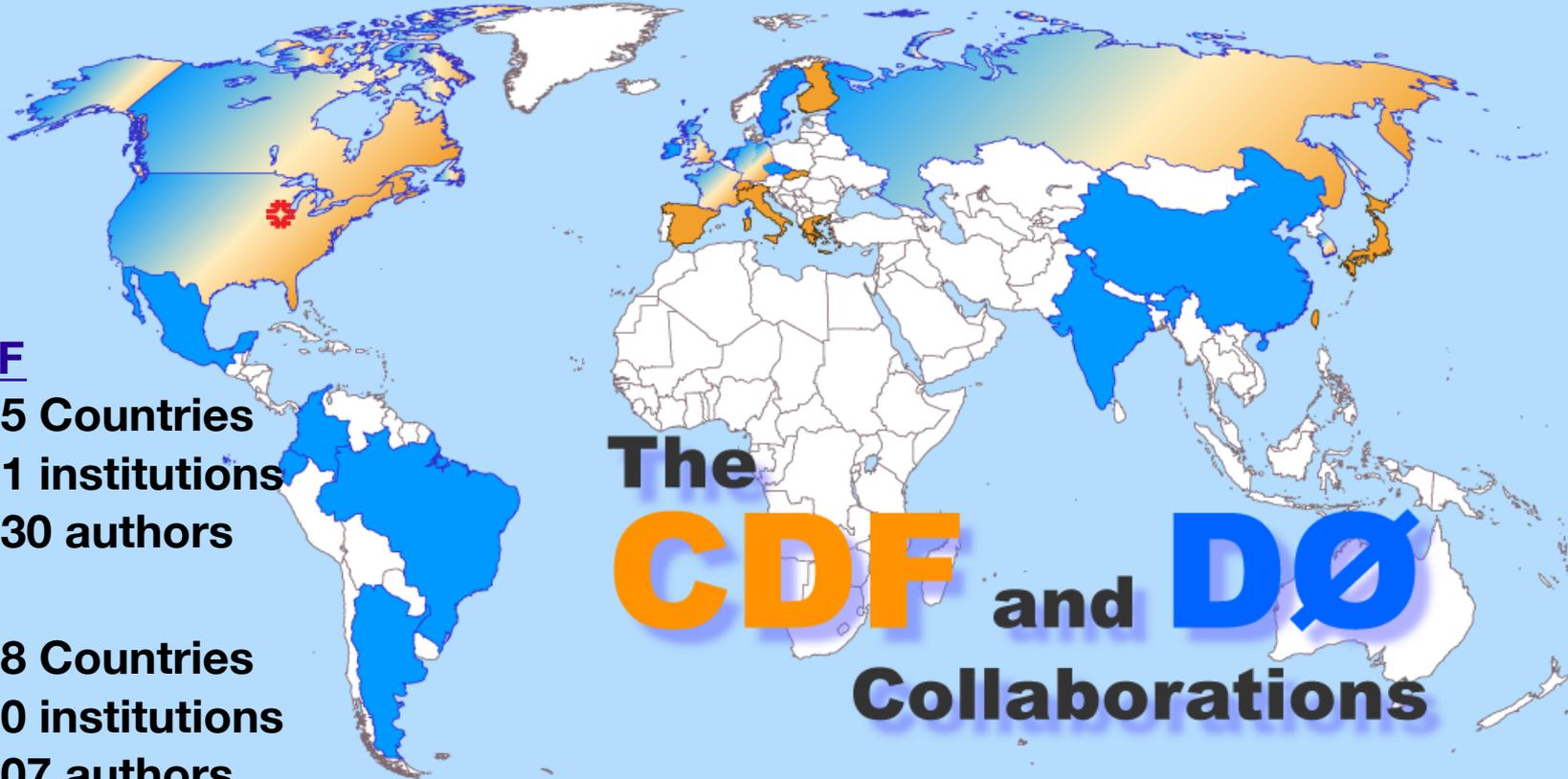
Are there extra dimensions of space?

Why is the Universe matter dominant?

What are the neutrino masses and what do they tell us?

What is dark matter?

What is dark energy?



The CDF and DØ Collaborations

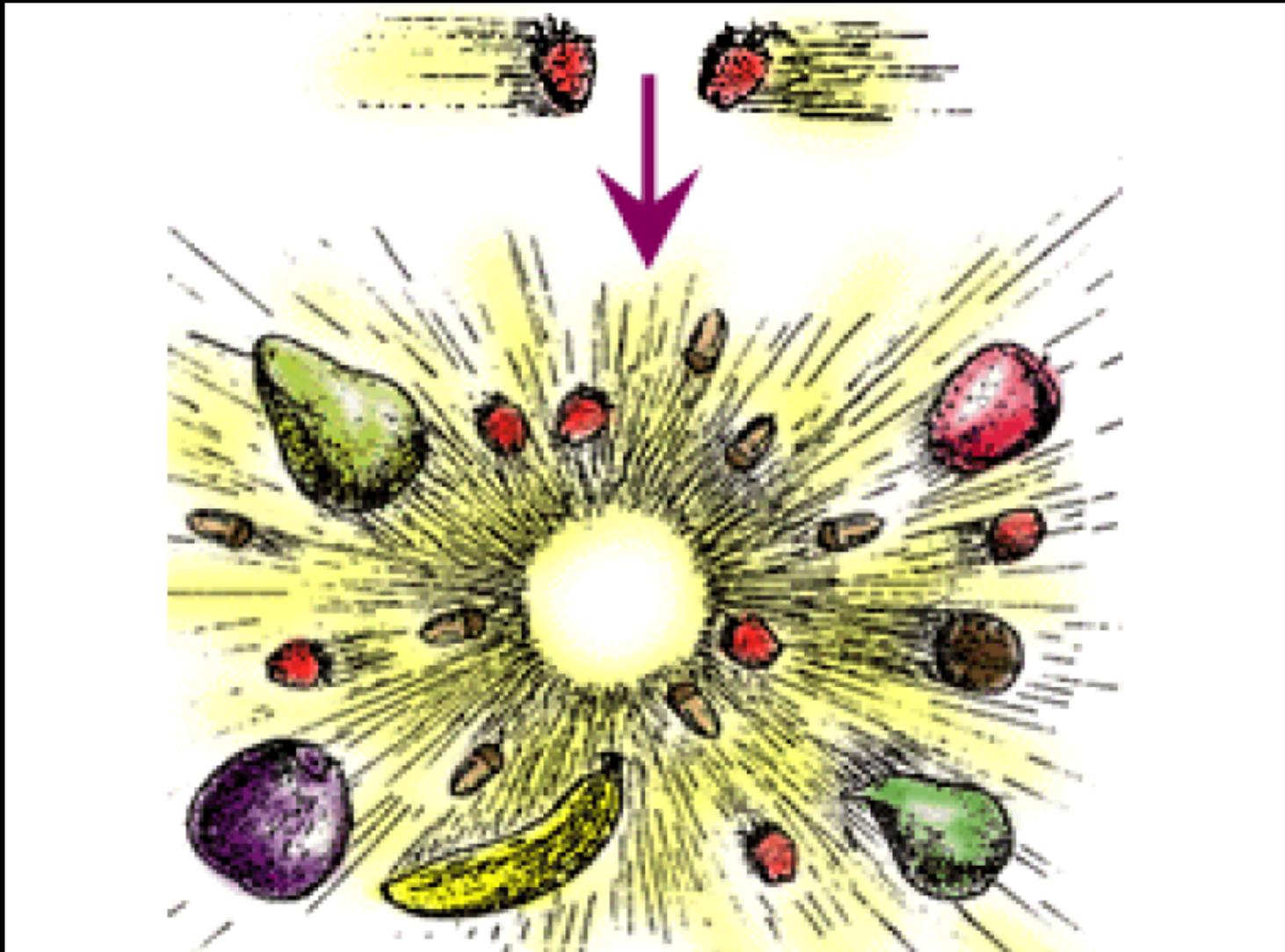
CDF

- ◆ 15 Countries
- ◆ 61 institutions
- ◆ 530 authors

DØ

- ◆ 18 Countries
- ◆ 90 institutions
- ◆ 507 authors

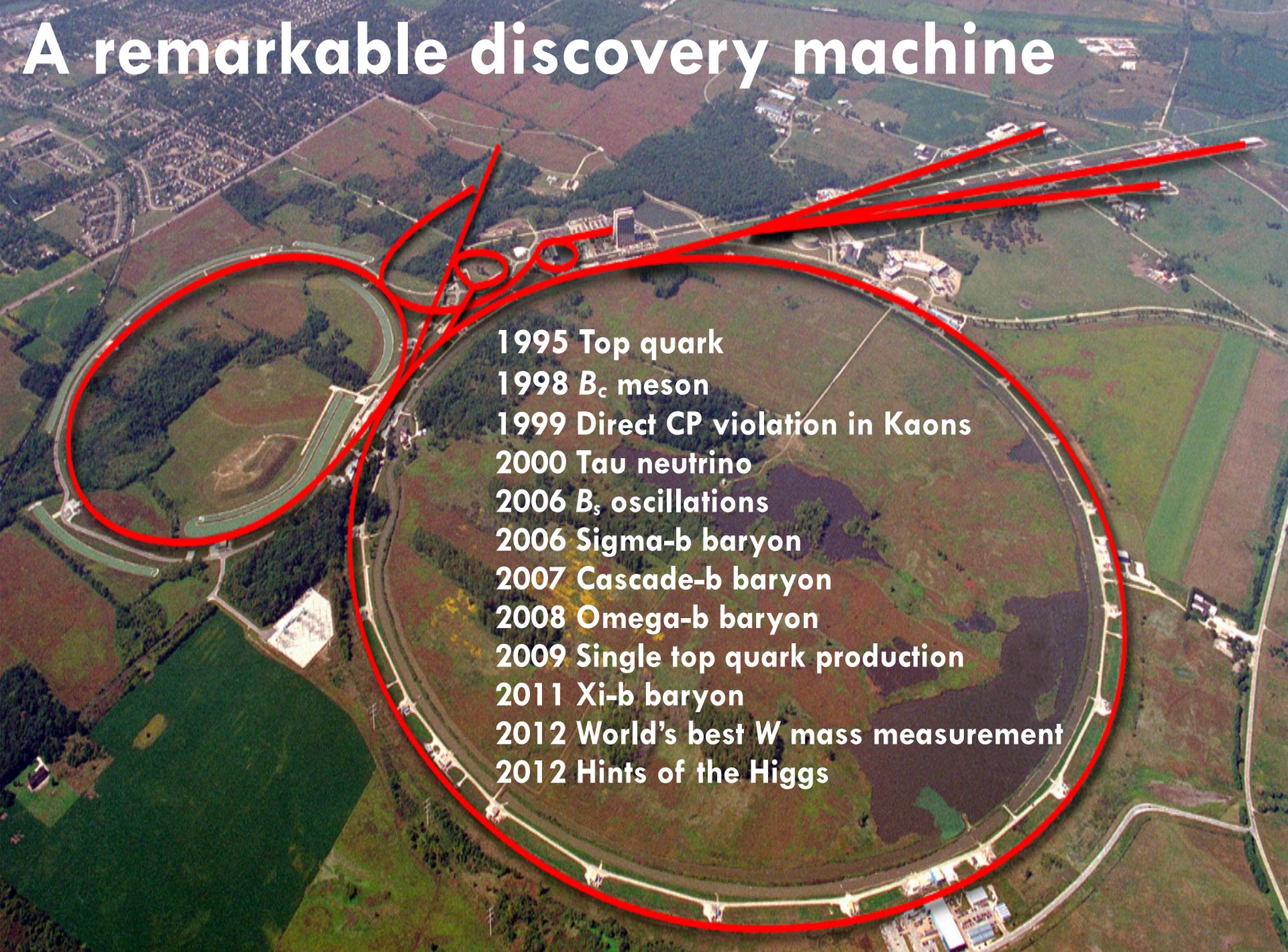




Can we make the worlds most delicious fruit?

Shamelessly stolen from Ian Shipsey

A remarkable discovery machine

An aerial photograph of the Fermilab particle accelerator complex. The image shows several large circular and oval-shaped tracks, with a prominent red circle highlighting the Tevatron ring. Red lines and smaller circles are drawn over the image, connecting various parts of the facility to a central list of scientific discoveries. The surrounding landscape is a mix of green fields and brown agricultural land, with some buildings and infrastructure visible.

1995 Top quark

1998 B_c meson

1999 Direct CP violation in Kaons

2000 Tau neutrino

2006 B_s oscillations

2006 Sigma-b baryon

2007 Cascade-b baryon

2008 Omega-b baryon

2009 Single top quark production

2011 Xi-b baryon

2012 World's best W mass measurement

2012 Hints of the Higgs

Tevatron Data Preservation

- The Two experiments ran for 10 years and acquired a unique and important data set that will most likely never be replicated
 - Graduate students on the experiments need support for analysis for the next ~6 years
 - Other experiments (e.g. LHC) with un-expected results may find value in checking within the Tevatron data.
 - Crucial to continue to have mechanisms to trust any results from these analyses.

Why All of a Sudden is this an Issue?

- Prior to the Tevatron Run II, particle physics experiments took data for a limited time and there was always a next experiment with an improved detector
- Run II is different – unlikely there will ever be another 2 TeV proton-anti-proton data set taken – thus this is unique
- There are physics advantages of proton – anti-proton collisions vs proton-proton that make this data set complimentary to what is now being collected at CERN
- Run II is no longer unique – other running experiments will also likely have important legacy data
- **It is important for the field to learn how to do this!**

Tevatron Data Preservation

- There is approximately 10 petabytes of data split about evenly between data and simulation (Monte Carlo)
 - The size of the data poses challenges in storage e.g. will it need to be migrated to new tape media – if so that is a big job in its own right? Need continuous curation activities/efforts.
 - Transport of data sets to local and/or international universities not likely to be able to be “immediate response”.
 - Re-analysis of the full data set will take significant length of time.

Tevatron Data Preservation

- Data Preservation is more than just curating the data
 - that is the easy part
- Must preserve ability to build and execute more than ten million lines of code for each experiment needed for re-processing and re-analysis of the datasets.
- Comparison of simulation to acquired events a critical dependency for trusting any results.
- Detector experts often called upon for help during current analyses.

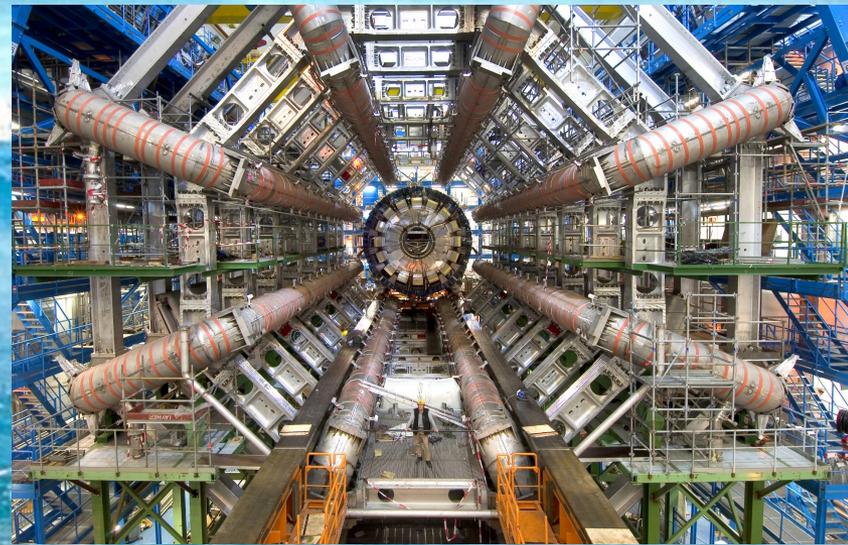
Tevatron Data Preservation

- Its about providing the tools, and documentation that can analyze the data in a proper fashion that will be useable for a decade or more
 - “virtualization” is not a silver-bullet. Need to carefully consider return on investments of different/new infrastructures.
 - Documentation, as well as tools, needs to be tested and curated to be trusted.
 - Physics culture relies on self-described data formats with lots of domain knowledge to interpret meaning from attribute definitions
 - Early preservation work will have direct access to existing experts. Preparing for the day when they are not available is a change in culture.

Tevatron Data Preservation
will help us understand how
to preserve this data set!

The LHC

The largest experiment... *EVER*



The LHC Experiment

- Science at an unprecedented scale
- Currently collecting data samples whose size in one year match the entire Tevatron output
- Ambitions to increase this by a factor of 10 or more in the coming years.
- This is “Big Data” personified
- Experiences in curating both the data and our tools to make sense of it for the Tevatron will guide us in preserving this first LHC data set.

Take Away Points..

- Particle Physics is only now coming to grips with long term data preservation
- Recently completed experiments (Like the Tevatron) have produced data that will not likely be re-made
- The job of data preservation goes well beyond data curation – requires the preservation of the analysis capabilities as well
- Lessons learned here will help with the preservation of other data sets
- Here to learn of others experiences and successes