Python for Data Science II

Amir Farbin

Lecture 1

A Data Science BS Degree

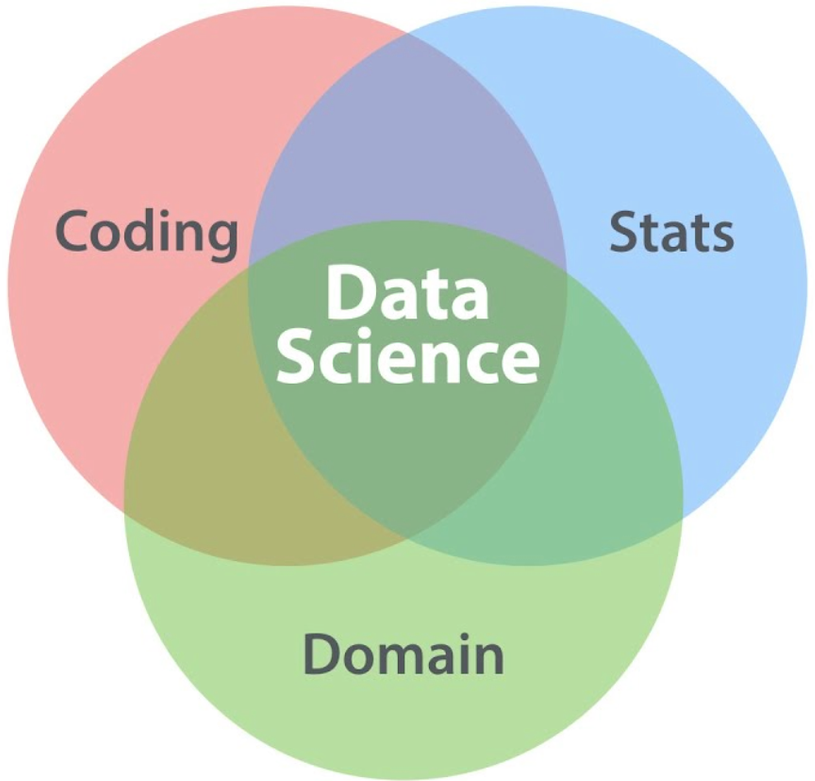
- The Degree Proposal was in progress for 3 years.
- **Approved** by *Texas Higher Education Coordinating Board* (October 22, 2020)
    - Full program launch Fall 2021.
- Courses available since Fall 2018.
- *Minor* defined Fall 2020.
- Unique Degree
    - Undergraduate
        - Most programs are professional masters or PhD.
    - Within *College of Science*
        - Most are in Computer Science or Business
        - Requires concentration and Capstone Project
- Aim to prepare students:
    - Entry-level Data Science jobs
    - Better Science Research
        - Undergraduate
        - Better positioned for Graduate School
            - Stronger Application
            - Start on research quicker

The UTA Data Science faculty

https://www.uta.edu/science/data-science/

The Challenge

- Practicing Data Science requires
    - Coding
    - Math (e.g. Statistics)
    - Domain expertise (e.g. physics, biology, …)
- Each of these areas can be a degree onto itself…
    - Usually people come into data science from one of these areas.
- *Challenge*: Start with a University Freshman with no preparation.
- This course will tackle all three areas…



The Degree

- Choose a concentration: Biology, Physics, Math, Psychology, Chemistry, Earth and Environmental Science, Geology
- Note courses are being renumbered/renamed for Fall 2021. Pending UCC approval:
    - ***DATA 1301*** – Introduction to Data Science
    - ***DATA 3311*** — Mathematics for Data Science
    - ***DATA 3401*** -- Python for Data Science 1
    - ***DATA 3402*** -- Python for Data Science 2
    - ***DATA 3421*** -- Data Mining, Management, and Curation
    - ***DATA 3441*** -- Statistical Methods for Data Science 1
    - ***DATA 3442*** -- Statistical Methods for Data Science 2
    - ***DATA 3461*** -- Machine Learning
    - ***DATA 4380*** -- Data Problems
    - ***DATA 4381*** -- Data Capstone Project 1
    - ***DATA 4382*** -- Data Capstone Project 2
- Major: core + these courses + 2 math courses (and calculus I)
- Minor: ~ 5 courses

The DS BS Degree requirements are as follows:

- UTA Core curriculum (46 hours)
- Data Science Foundations (18 hours)
    - Student Success
    - Intro Data Science
    - Python 1 and 2
    - Statistical Inference
    - Linear Algebra/Probability
- Core Data Science courses (16 hours)
    - Statistical Methods 1 and 2
    - Data management
    - Machine Learning
- Data Science Capstone courses (9 hours)
    - Data Problems
    - Capstone 1 and 2
- Domain specific courses (23 hours)
    - Domain Concentration Specific Requirements
- Science elective with lab (8 hours)
- Total = 120 hours

Syllabus

Introduce Yourself

- Your UTA Degree
    - What is your major?
    - What year?
    - When will you graduate?
- Interests
    - Is there a specific scientific or professional field?
    - Have you done any research?
    - Any hobbies, etc, that you can apply DS to …?
- Optional:
    - Do you work? Where? Can you apply DS there?
    - What else competes with your time to work on your degree?
- Your goals
    - What’s next (job, grad school)?
    - How can this course help?
- Your setup
    - What kind of computer? (Windows, Mac, Linux)
- Anything else you like to share…

Large Language Models (LLMs)

- LLMs are part of the reality of modern data science and research.
    - I use LLMs to:
    - Help develop and refine this course
    - Assist with my own research and software development
    - Explore new ideas, workflows, and explanations
- Many of you are already experimenting with LLMs:
    - To write code
    - To debug
    - To explain concepts
    - Sometimes to try to shortcut learning
- Ignoring LLMs is not realistic.
- Blindly relying on LLMs is also not effective.
- Our goal is not to ban or worship LLMs —our goal is to understand how to use them well.

A Shared Experiment: Learning How to Use LLMs Effectively

This course will treat LLM use as a shared, evolving experiment.

- I will try different ways of incorporating LLMs into:
  - Learning
  - Coding
  - Problem-solving
  - Interpretation
- You will:
  - Try using LLMs in different ways
  - Reflect on what helps and what hurts your understanding
  - Provide real data through your work and experiences
- Together, we will:
  - Analyze outcomes
  - Identify failure modes (hallucinations, shallow understanding, inefficiency)
  - Identify effective workflows

The objective:
To help you become the *best possible future version of yourself —*
a data scientist who can think critically, verify results, and use powerful tools responsibly.

Optional closing line (if you want it on Slide 2, smaller font)

The point is not whether AI can do the work —
the point is whether you understand the work well enough to know when it’s wrong.

If you want next, I can:

- Add a one-slide “rules of engagement” for LLM use on assignments
- Write a short syllabus policy that matches this framing
- Design a reflection prompt students submit about their LLM use
- Help you turn this into a mini research study on learning with LLMs

Just tell me how far you want to push it.

Why This Course (and Why It Matters with LLMs)

This course gives you **the core understanding needed to do a large fraction of real data science**.

The goal is to **learn the building blocks** —so you can **assemble solutions step by step**,**which is exactly how effective LLM use works today.**

Analogy:

- Most people can *drive* a car.
- **Race car drivers** know what's under the hood.
- They can push systems to the limit.
- They can work with mechanics and designers and as a team.
- They know when something is wrong — and how to fix it.

This course trains you to be a race car driver.

LLMs are powerful — but understanding is what lets you use them well.

What do *I* do ?

Was the Universe an Accident?

Artificial Intelligence may find the answer in data from the Large Hadron Collider

Amir Farbin

UNIVERSITY OF TEXAS ARLINGTON

ATLAS EXPERIMENT

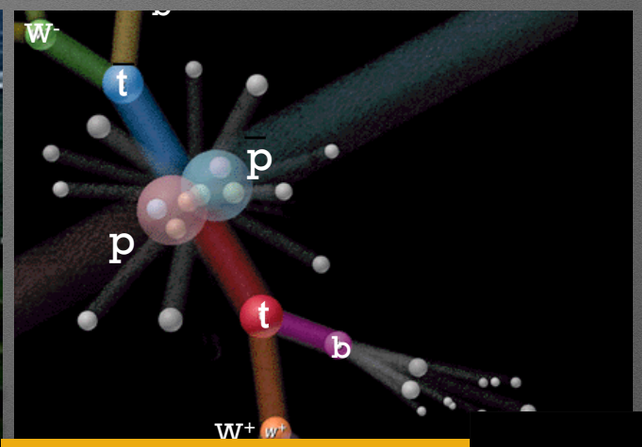
DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

What is HEP ?

Large Hadron Collider (LHC)

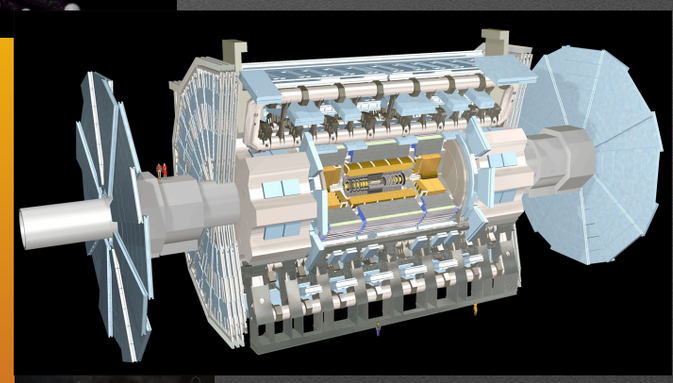
Largest Machine Ever Built

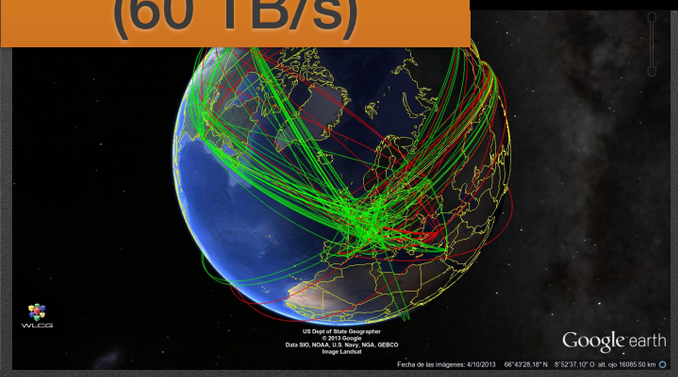
10^{11} Protons Collide 40 Million Times per Second



Create New Particles
Only 1 in 10^{11} are "interesting"

Record with 5 Story 100M Channel "Camera" (60 TB/s)





Processed by 300k Cores Around the World

Higgs Discovery - Nobel Prize Physics 2013

Physics Letters B 716 (2012) 1–29

Contents lists available at SciVerse ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

ATLAS Collaboration *

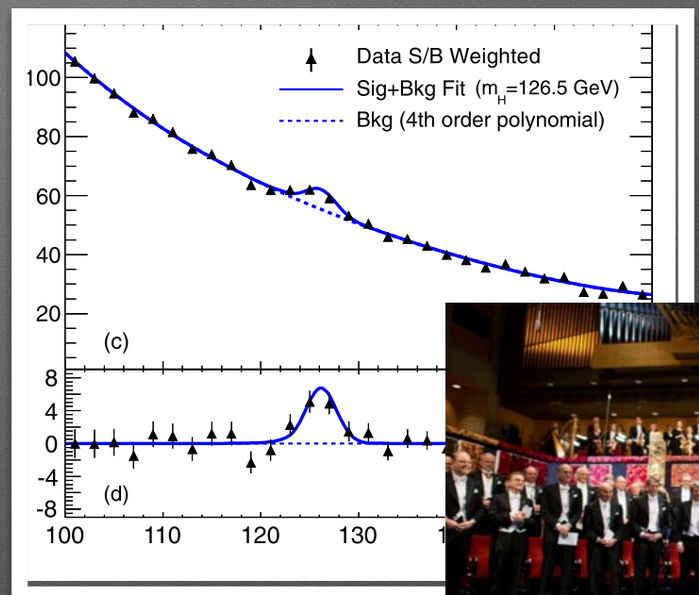
This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

ARTICLE INFO

Article history:
Received 31 July 2012
Received in revised form 8 August 2012
Accepted 11 August 2012
Available online 14 August 2012
Editor: W.-D. Schlatter

ABSTRACT

A search for the Standard Model Higgs boson in proton–proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of 126.0 ± 0.4 (stat) ± 0.4 (sys) GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.



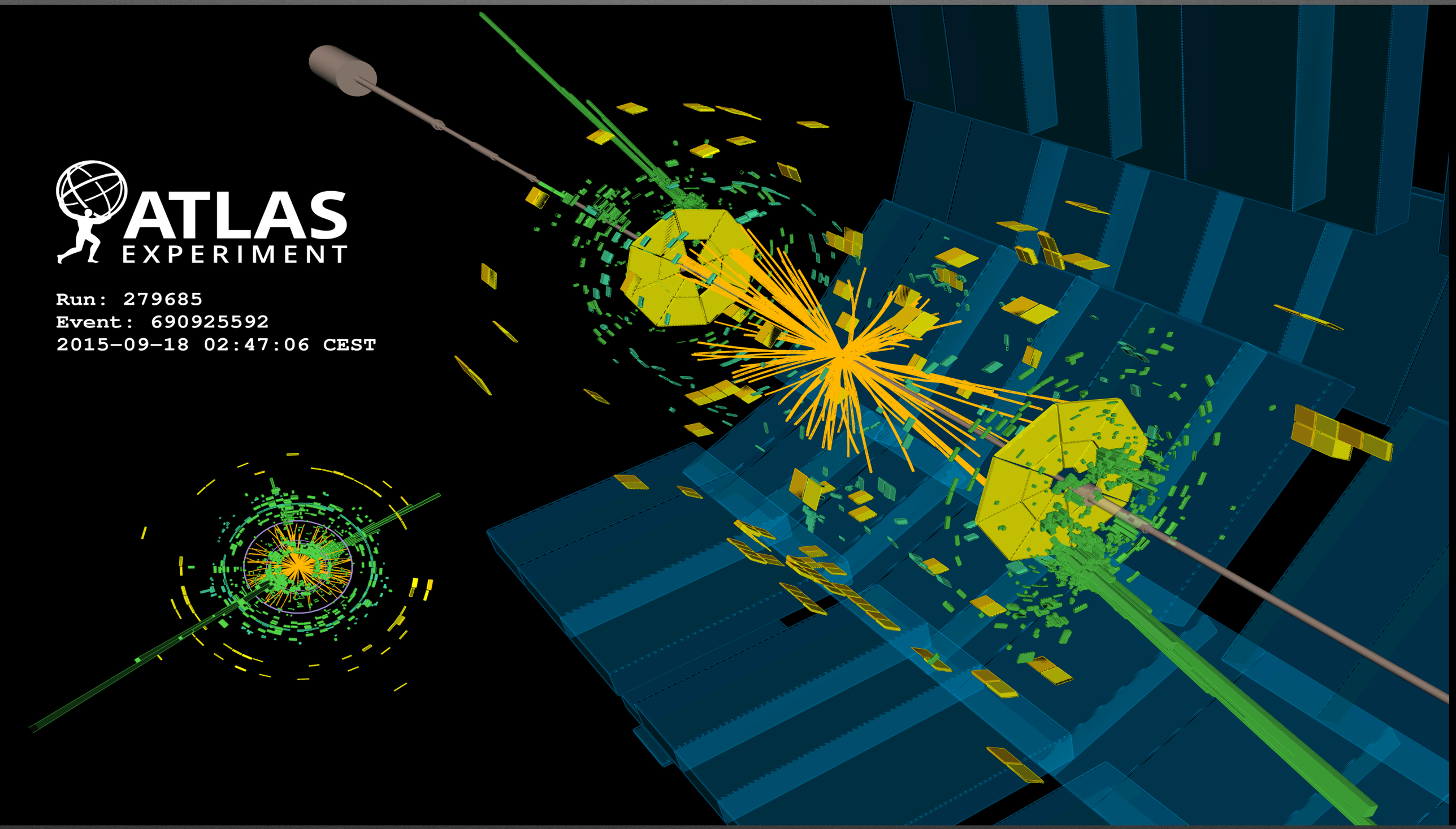
Last Piece of the Standard Model

Best Tested Theory… Ever.

Not Done… Higgs is Light! Possibilities:

- *Fine-tuned Theory*: Accident or Multiverse + Anthropic Principle
- *Mechanism*: Supersymmetry, Extra-Dimensions, Sub-structure
  - Focus of LHC
- *Design?*

Deep Learning in High Energy Physics



- Requires lots of computing
- Upgrade to LHC will give us 100x the data.
  - We won’t have 100x the computing power or storage.
- Use Artificial Intelligence and newest processors…

Animal Brains

- The brain takes in sensory data… ***builds hierarchical models of the world***.
- So effectively, a ***representation*** of the input is assembled in the brain.
  - Eyes see a limited window… but…
  - Location Cells
  - Imagining locations

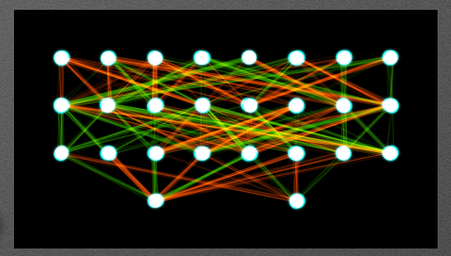
Brief History of AI

Artificial Intelligence

- Goal: Systems that reason and act as well as or better than humans
- Heuristic AI (1990's)
- Machine Learning AI
  - Knowledge learned from data
- Neural Networks ~ Brain inspired computing (1943)
  - Universal Computation Theorem (1989)
  - Multi-layer hidden networks (a.k.a. Deep) (1965)
  - Vanishing Gradient Problem (1991)

Deep Learning Renaissance (> 2007 - now)

- Driven by:
  - New NN Innovation
  - Big Data
  - Graphical Processing Units

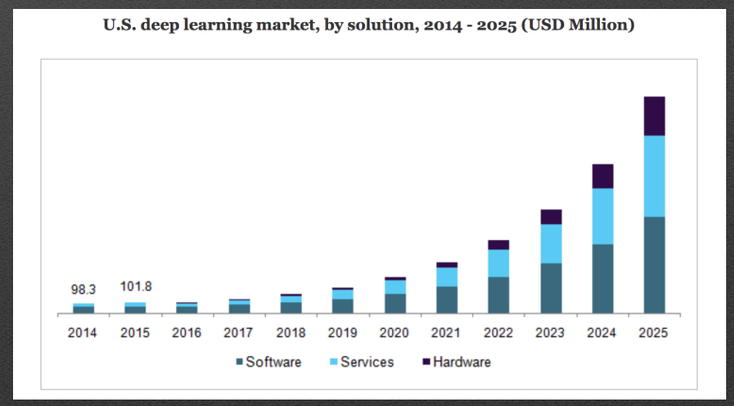


- Amazing Feats

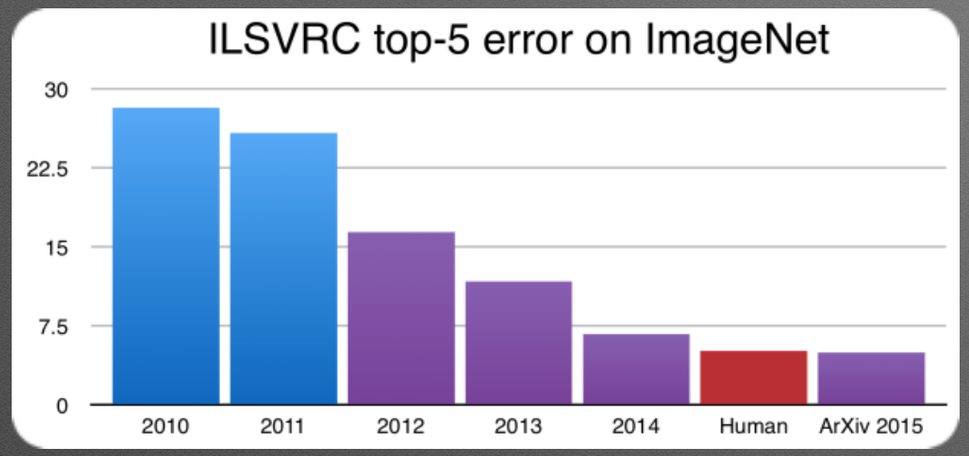
WIRED STAFF SCIENCE 06.26.12 11:15 AM

GOOGLE'S ARTIFICIAL BRAIN LEARNS TO FIND CAT VIDEOS

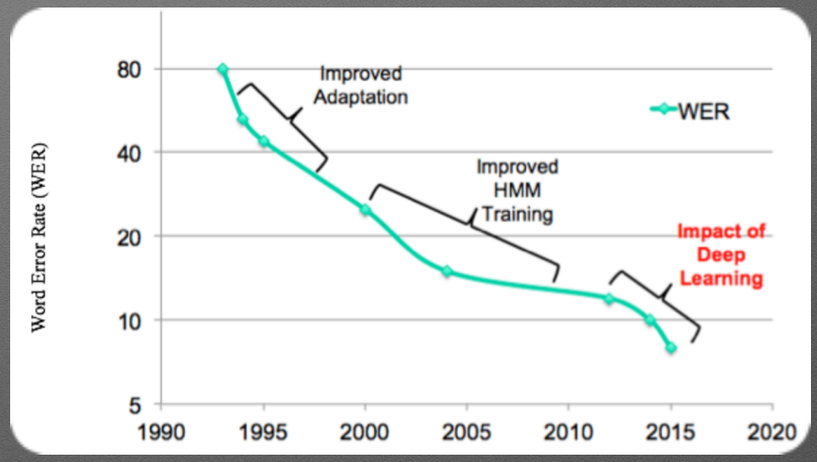
- Market Growth
- Industry Adoption



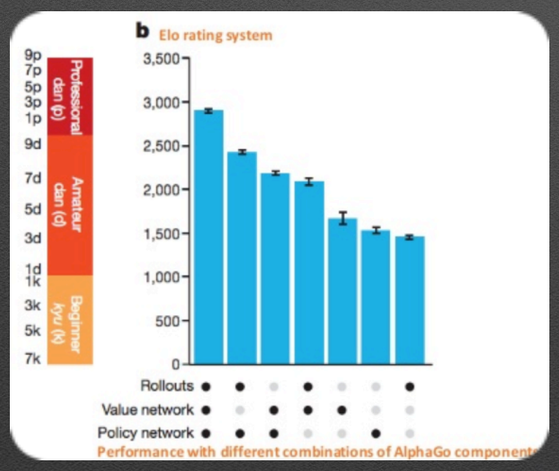
Amazing Feats : Some Examples



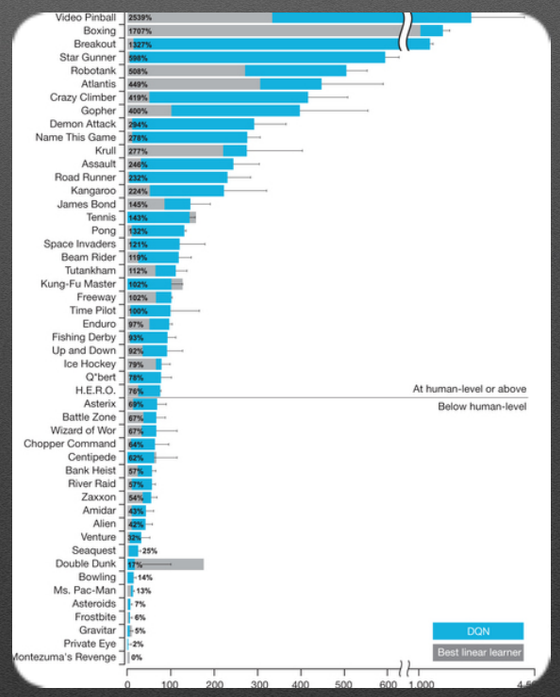
ImageNet Outperforms humans



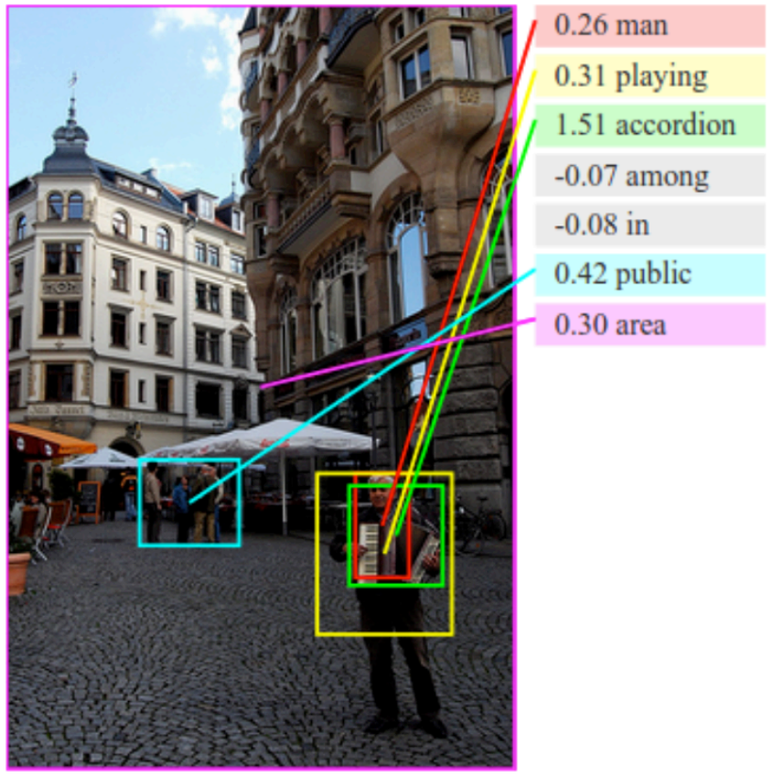
Rapid Advances in Speech Recognition

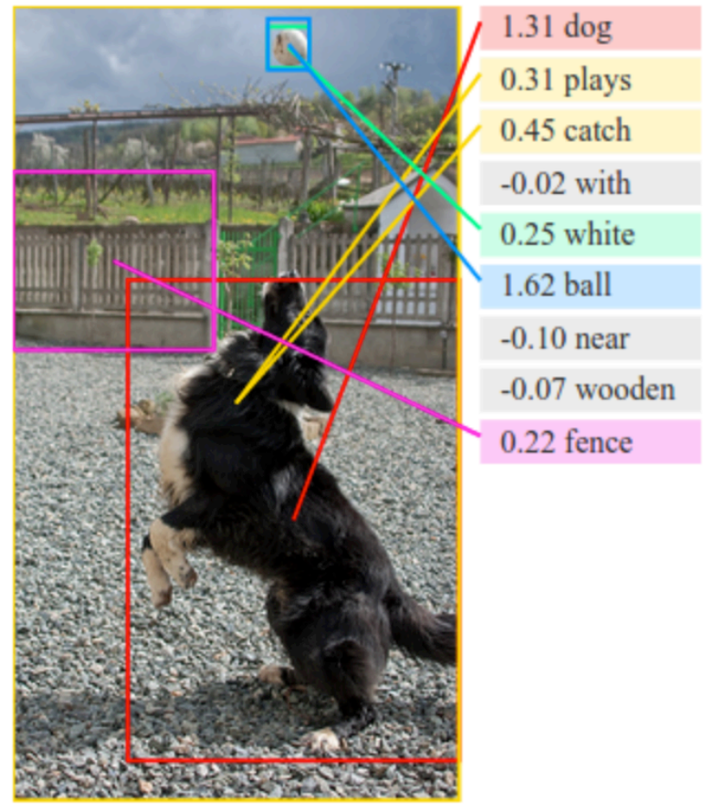


AlphaGo beats Lee Sedol



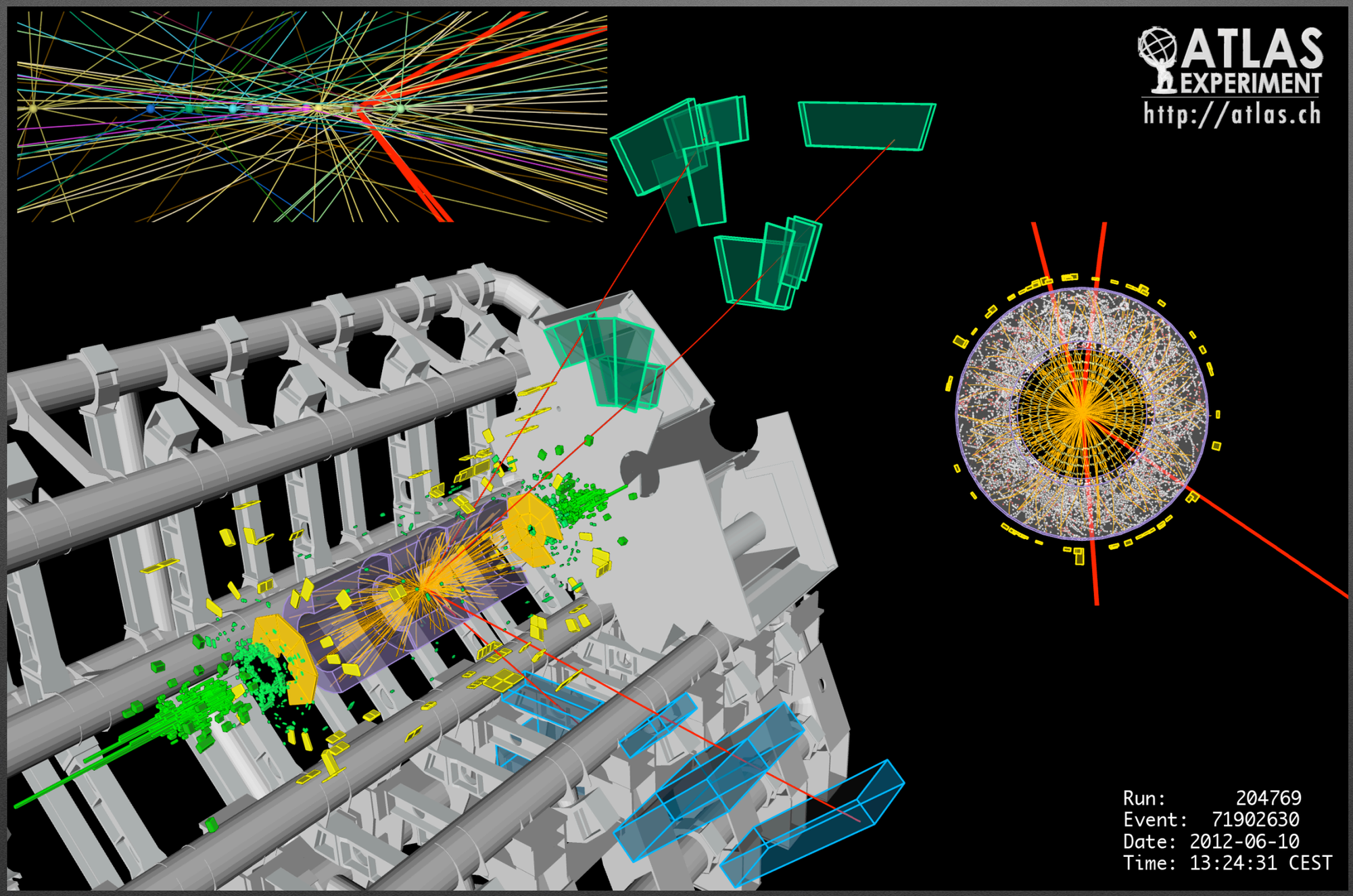
Human Level control in playing Atari games





Deep Learning can be used to tell a story with context through data

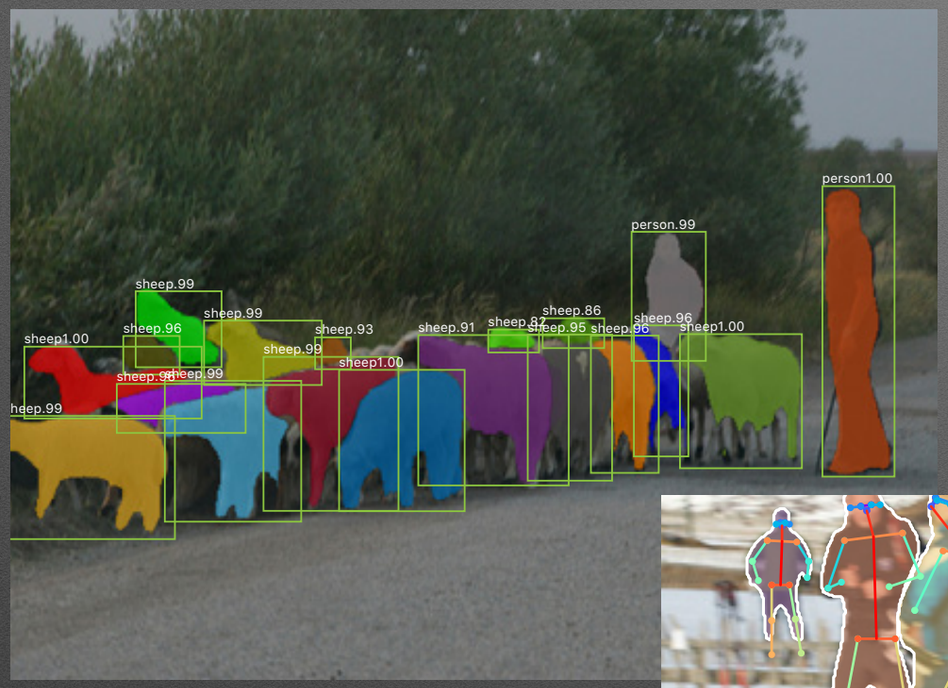
Used in HEP to understand what physics is happening

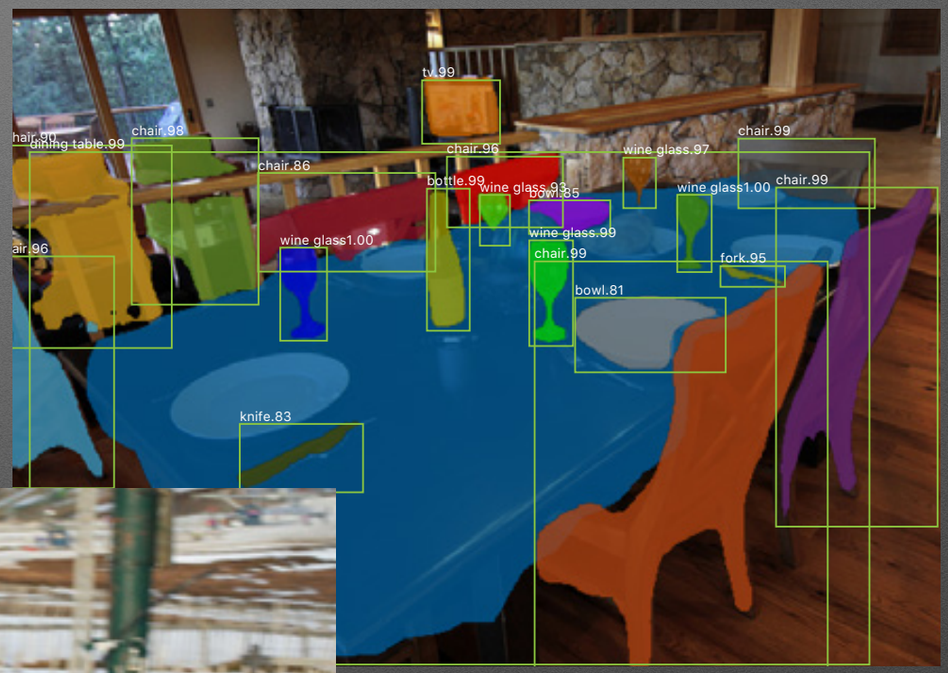


$H \rightarrow ZZ \rightarrow 4\ell$

We can use deep learning object tagging techniques in the data to find the decaying particle

Amazing Feats : Some More Examples





https://arxiv.org/pdf/1703.06870.pdf

Why go Deep?

Better Algorithms

- Better results
- Solution where there is none
- Make sense of complicated data

Easier Development

- Feature Learning, not Feature Engineering
- Save time and cost

Faster Algorithms

- DNNs Faster than traditional Algs
- Neuromorphic processors

Why Science + Data Science ?

Take my field, High Energy Physicists (HEP). It's ideally suited:

- HEP Systems and Machine Learning and Deep Learning Systems confront similar challenges
- Decades of Experience at the Data Frontier
- Bridge between science and industry
- HEP scientists are also engineers by training

MOVE OVER, CODERS— PHYSICISTS WILL SOON RULE SILICON VALLEY

… it's happening across Silicon Valley. …, ***the things that just about every internet company needs to do are more and more suited to the skill set of a physicist***.

new wave of data science and AI is something that suits physicists right down to their socks.

"There is something very natural about a physicist going into machine learning … more natural than a computer scientist."

Physicists know how to handle data … building these enormously complex systems requires its own breed of abstract thought.