Data Science in the Wild

Ryan G. McClarren, PhD.
University of Notre Dame

My Background

- I have been working on using data analysis techniques to understand how we can use computer simulation to make predictions including reasonable estimates of uncertainty.

- I also helped to found a business that does data science/predictive analytics for retailers such as
  - J. Crew
  - Dick's Sporting Goods
  - White Castle
  - Kroger

- For marketing purposes, we also would do things like forecast the Oscars or opine on the NFL draft.

- All of these ventures brought me into contact with the world of Big Data, Data Science, Machine Learning, and many other buzzwords.

Some Press

Big Data Goes To The Oscars And Other Friday Stories

FEB 28, 2014 @ 11:26 AM 492 VIEWS

Howard Baldwin
CONTRIBUTOR
FOLLOW ON FORBES (72)

Opinions expressed by Forbes Contributors are their own.

FULL BIO

Back in the last century, there was nothing I loved more than the Oscar ceremony. My first job as a journalist was reviewing movies (my first was *Magnum Force*, so that tells you how long ago it was). Today, if it's not on a DVD with subtitles and a pause button for bio-breaks, I'm just not that interested.

Yet I was piqued by this *InformationWeek* column in which a company called Farsite claims to be able to pick the winners using big data analytics. Using historical data, previous wins from other groups, and even gossip, the company claims to have picked five of the six big winners last year, so it'll be interesting to see if they can do it again. (By the way, the film critic for the *San Francisco Chronicle* does the same kind of analysis each year, using criteria such as the actor's age, whether the part required a disability, and other parameters; I don't think he uses a computer.)

THE WALL STREET JOURNAL.
WSJ.com

April 26, 2013, 10:00 AM ET

For Rookies, Draft Position Isn't Destiny

ByCarl Bialik

Getty Images

Even if he's picked in the third round, Eddie Lacy could get lots of carries in his first year in the NFL.

Everyone knows that where a player gets picked in the NFL draft that started Thursday determines how much money he makes and how much playing time he'll get as a rookie. Less obvious is that there's one high-profile position for which being picked later may mean more first-year opportunities.

by Farsite chief science officer Ryan McClarren. The typical rookie drafted in the first half of the third round got 7.5 more carries than did his far more touted peer drafted about 50 spots earlier, in the second half of the first round. The effect goes only so far, though: The median number of carries for a rookie running back drafted in the fourth round or later is just 14, compared to 122 for the first half of the third round and 62 for the second half of the third round. That's good news for all the running backs, including Alabama's Eddie Lacy, waiting to hear their names after none were chosen in the first round Thursday.

A Working Definition

- “Big Data” refers to the bleeding edge of data analysis for a particular field
- Contains all of the data collected in the field
- Is of a size that cannot be handled by standard tools
- Has not typically been analyzed in totality before.
- Has some sort of magic behind it.
  - Why else would someone use a term so vague?

Other Buzzwords: Data Science, Analytics, Predictive Analytics

- Another term that is used commonly is Data Science:
  - This is a blend of computer science, statistics, signal processing, etc.
  - Wide but not deep – the term statistics should usually suffice.

- Analytics, Predictive Analytics
  - Analytics comes from the business world and historically referred to reporting, dashboards, etc.
    - How many widgets did we make in each factory today
  - Predictive Analytics combines statistical models with analytics to get more information than raw numbers and summaries.
    - Basically statistics on top of a database.

Domain Dependence

- What is Big Data to a farmer is not Big Data to an analyst at Facebook
  - Gigabytes of weather data v. Every interaction each of the billions of users of Facebook has *ever* had with the site

- In Science and Engineering there are also differences in domain
  - The Large Hadron Collider 3 GB/s
  - Simulation run for engineering design ~GBs/day

- Also, the means and abilities to process data will vary based on the domain.
  - Facebook, Google, etc. use large computing clusters to analyze data

Velocity - Variety - Volume

- Big Data typically refers to a combination of these being large, and potentially all three being large.
- Velocity refers to the speed at which the data is created.
- Variety refers to the kinds of data that are created
  - Structured/Unstructured
  - Text, images, audio, video, sensors
  - Social Media
- Volume refers to the amount of data.
- All of these can cause issues in data analysis.
- The definition is also evolving because what was Big Data five years ago is not necessarily big today.

“Data does not need to be Big to be Powerful”

It is not the size of the data, but how you use it.

- Despite the number of people talking about using data to improve outcomes, there is much more talk than use of data.

- Most businesses/researchers are collecting large amounts of data, but are lost when it comes to using it.

- Building more and more haystacks, but no one is looking for the needles.

- It is possible to make an impact by extracting small data sets from the vast lake of data in an organization.

How to Apply Data Science

Data science is an iceberg.

- The results that the outside world sees represents a small fraction of the total work.
- Often one spends more time collecting, cleaning, organizing, and exploring the data than building sophisticated models.
- A data scientist is skilled in the areas of data collection, processing, statistical modeling, visualization, and implementation.
- Few people have all of these skills; almost no one is an expert in all of them.

What is the role of data in an organization

- This is the key question to ask and to answer
- Are data-driven decisions valued?
- Is data used to justify decisions that have already been made?

- For any important decision, we should ask
  - What data supports the decision?
  - What data do we wish we had to make a scientific decision?
  - Are we going to be able to assess this decision later?

Use data to make humans more effective

- Find the easy decisions and make them automatic
  - Don’t spend time on the easy decisions.
  - If all the data and models agree, the decision should be easy.
- Deliver the data and models to support the difficult decisions
  - Data and models point to different decisions
  - Here is where the human intervention is necessary
- Processes should be set up to make it clear which decisions are which

Data scientists are most effective when they can find the important problems.

- The most valuable data scientists are given direction according what the important metrics in a firm are.

- The data is available in an easy to deal with format
  - Not usually the case.

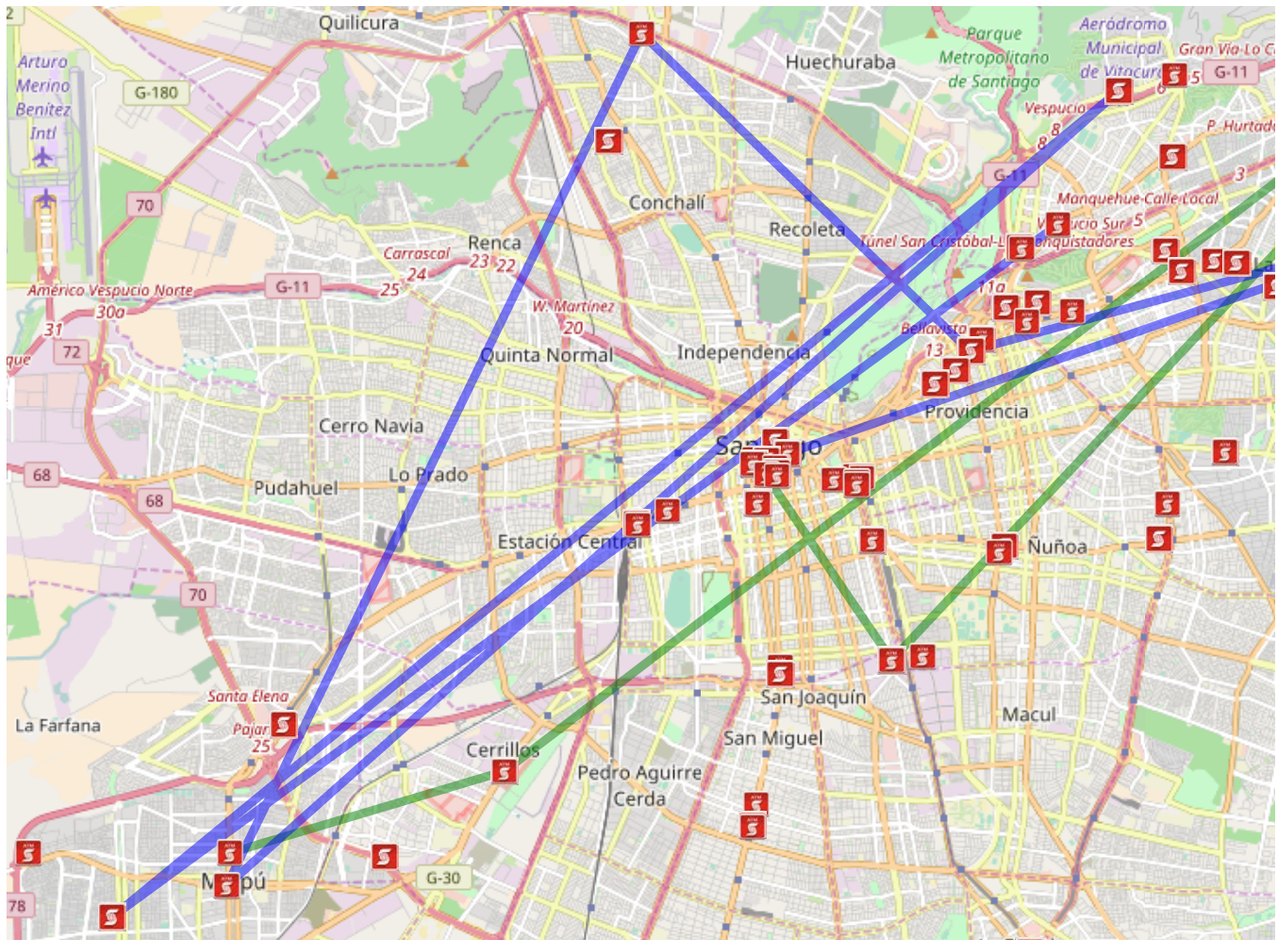
- Explore the data for relationships that can be exploited to improve the performance of a business.

- The organization must also value the insights from data.

Geospatial Data Science
Location Matters

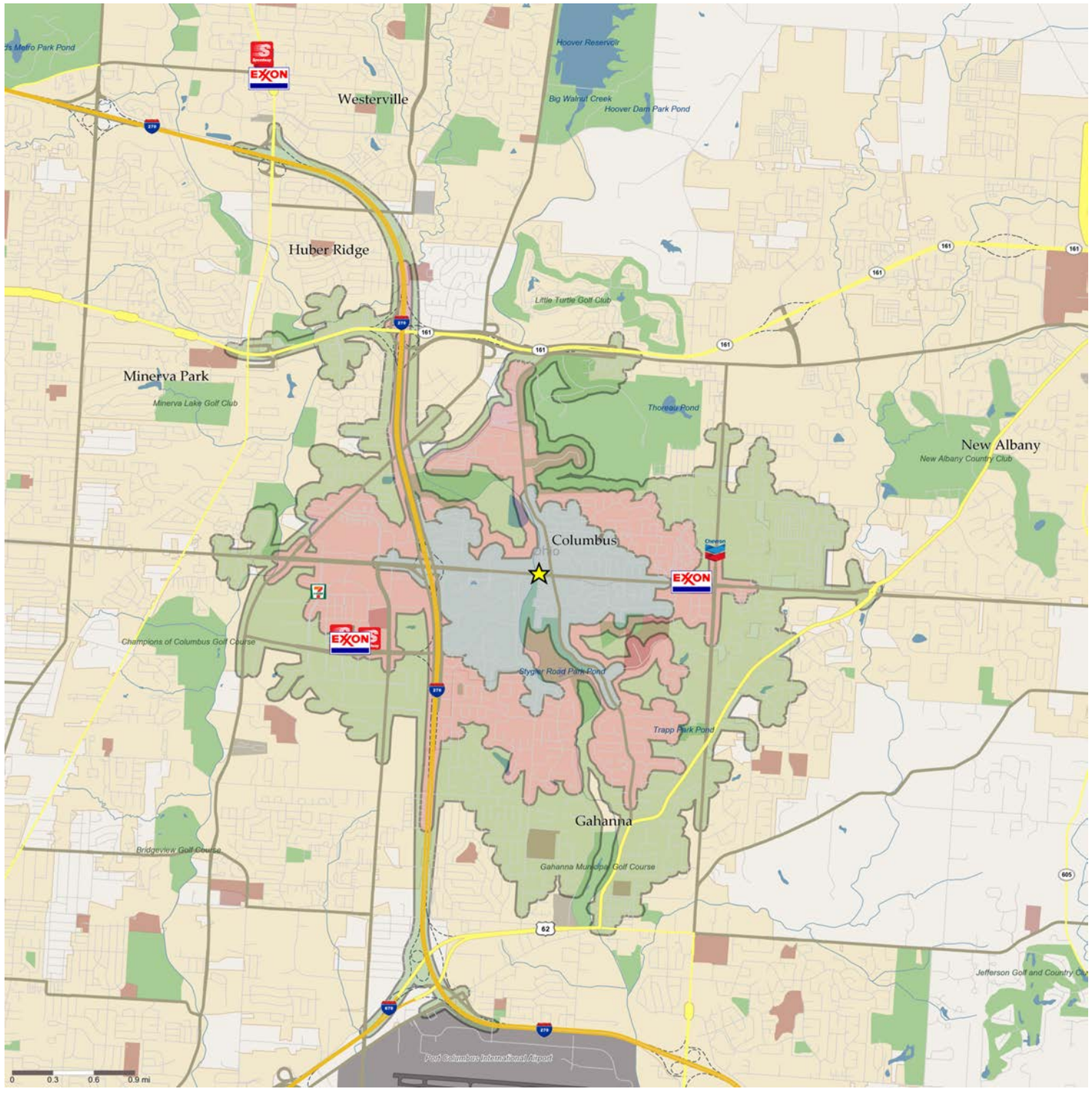
Location Data Presents a Unique Opportunity

- Location can be very powerful
  - The stores/restaurants/ banks you go to are likely strongly influenced by location
    - Even more than other factors (price, quality, etc.)



- One can learn a lot about a process or behavior by combining location with time series data.

Forecasting Performance by Location



Drivetime (Minutes)	Population	Daytime Pop.
0-3	4,679	2,603
3-5	9,756	5,436
5-7	15,056	8,157

Drivetime (Minutes)	Med HH Inc	Commuters
0-3	65,583	2,419
3-5	67,728	5,014
5-7	68,205	7,582

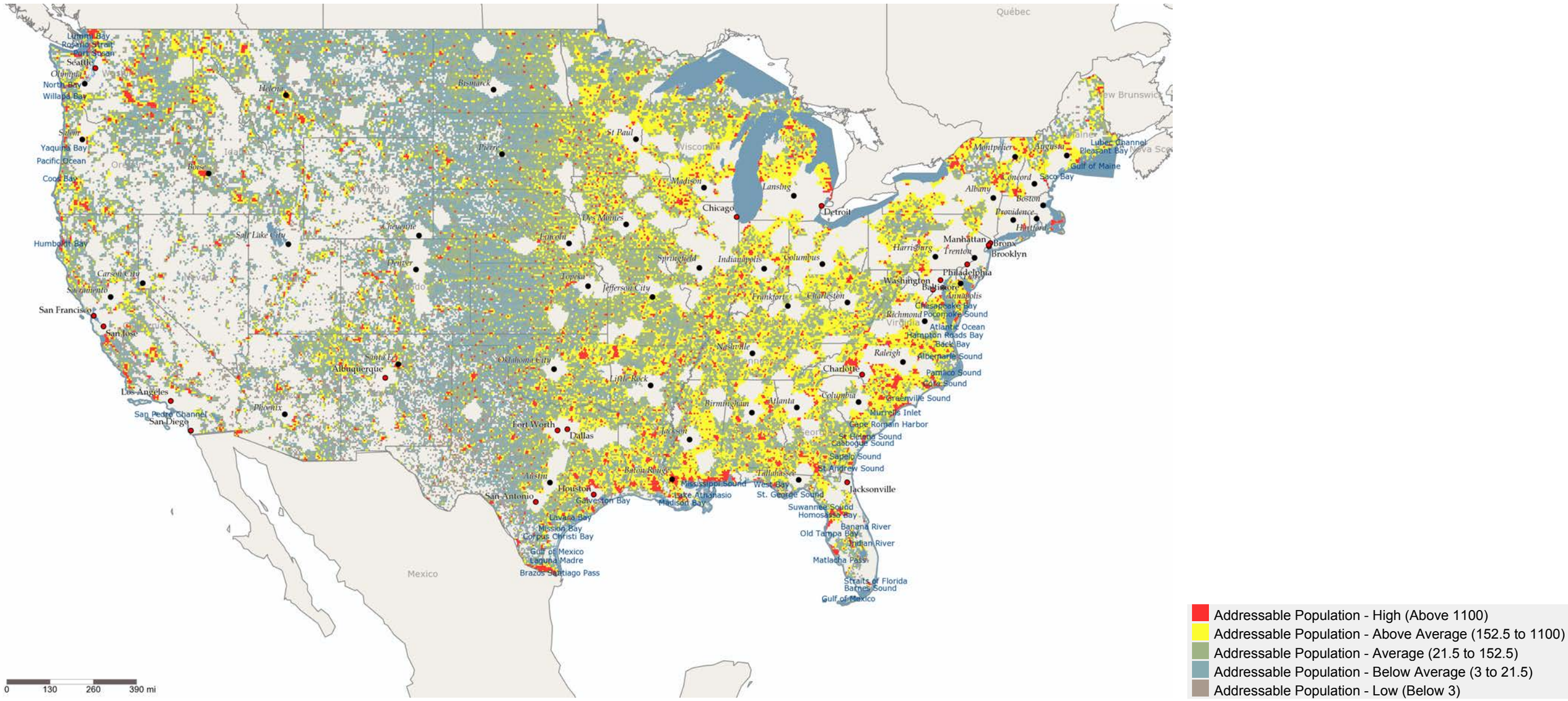
Drivetime (Minutes)	Median Age	Pop Growth
0-3	34.0	-2%
3-5	36.0	3%
5-7	35.0	5%

Competitor	Count	Closest Drivetime
7-11	1	4.50
Chevron	4	3.80
Exxon	3	4.30
Speedway	3	5.10

- Prospective Site
- Exxon
- Speedway
- Chevron
- 7-11
- Trade Areas - 3 Minute Drive Time
- Trade Areas - 3-5 Minute Drive Time
- Trade Areas - 5-7 Minute Drive Time

Growth Forecasting for Private Equity/Investment Banking

- Here we estimate market size for regions of the entire USA where there are no competitors within 100km.
- We have divided the nation into 30 km x 30 km tiles and analyzed all of them to produce this map.



Data Science to Predict the Oscars

Bringing Data to the Academy Awards

The academy awards (Oscars) are presented every year to the recognize excellence in movies.

- The Oscars are voted on by a panel of movie industry professionals known as the Academy of Motion Picture Arts and Sciences (e.g. directors, actors, producers, writers, editors, etc.).
- There are 24 awards with 6 being the major awards:
  - Best Picture
  - Best Director
  - Best Actor
  - Best Actress
  - Best Supporting Actor
  - Best Supporting Actress
- Each award has 5 nominees (Best Picture has 8).

The Oscars are voted via secret ballot by the 7,000 Academy Members

- We as movie lovers wanted to predict the outcome of the awards using our data science and machine learning tools.

- Problem: No polls of Oscar voters exist. How can we tell what the voters will do?

- One way would be to look at the film awards from other organizations such as the Golden Globes or the Critics' Choice Awards.
  - This could correlate with the Oscar voters but the voters for these awards are different people.

The Oscar voters do tell us what they like, in a way.

- Each part of the movie industry has “guild” awards given by voters made up of every worker in that role:
  - Screen Actors Guild Awards (every actor/actress votes)
  - Producers Guild Awards (every producer)
  - Writers Guild of America Award (every writer)

- Each of these awards has voters that also will be Academy Awards voters.
  - Therefore winners of these awards give us a signal for how voters from each category will vote.

- Also, in the past global “prediction markets” where traders bet real money on the outcome of events. These could include insider betting from Academy voters.

The hard part was collecting the data and putting it in a useful form.

- I started doing this in 2012 so I had to do scrape a lot of this data myself.

- I needed the winners and nominees in each of the 6 categories for the past 30 years from
  - The Oscars
  - The guild awards
  - The other awards (Golden Globes, etc.)

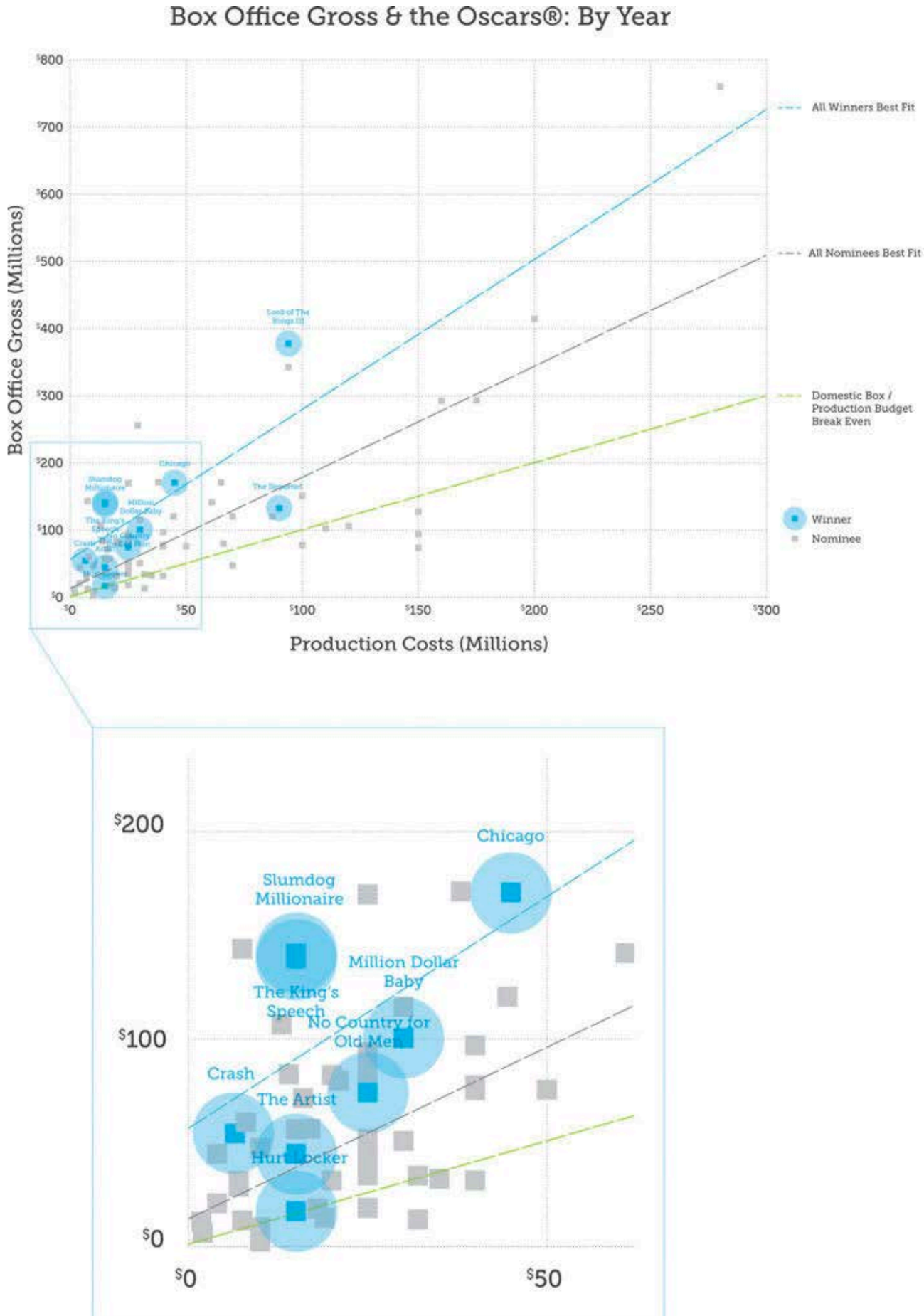
- Most of the awards data was scraped using Python and Wikipedia.

- I also needed information about each film: the running length, the stars, type of film (comedy, drama, …)
  - The websites Rotten Tomatoes and IMDB were helpful here due to their APIs

- We wanted all this data so we could look for important indicators of winning an Oscar.
  - Does a film being nominated in multiple categories increase the chance of winning a specific award?
  - Is it better to have multiple actors/actresses nominated from a film for the lead to win a Best Actor/Actress award (or is it the opposite)?
  - Does being nominated or winning the award previously matter for individual awards.

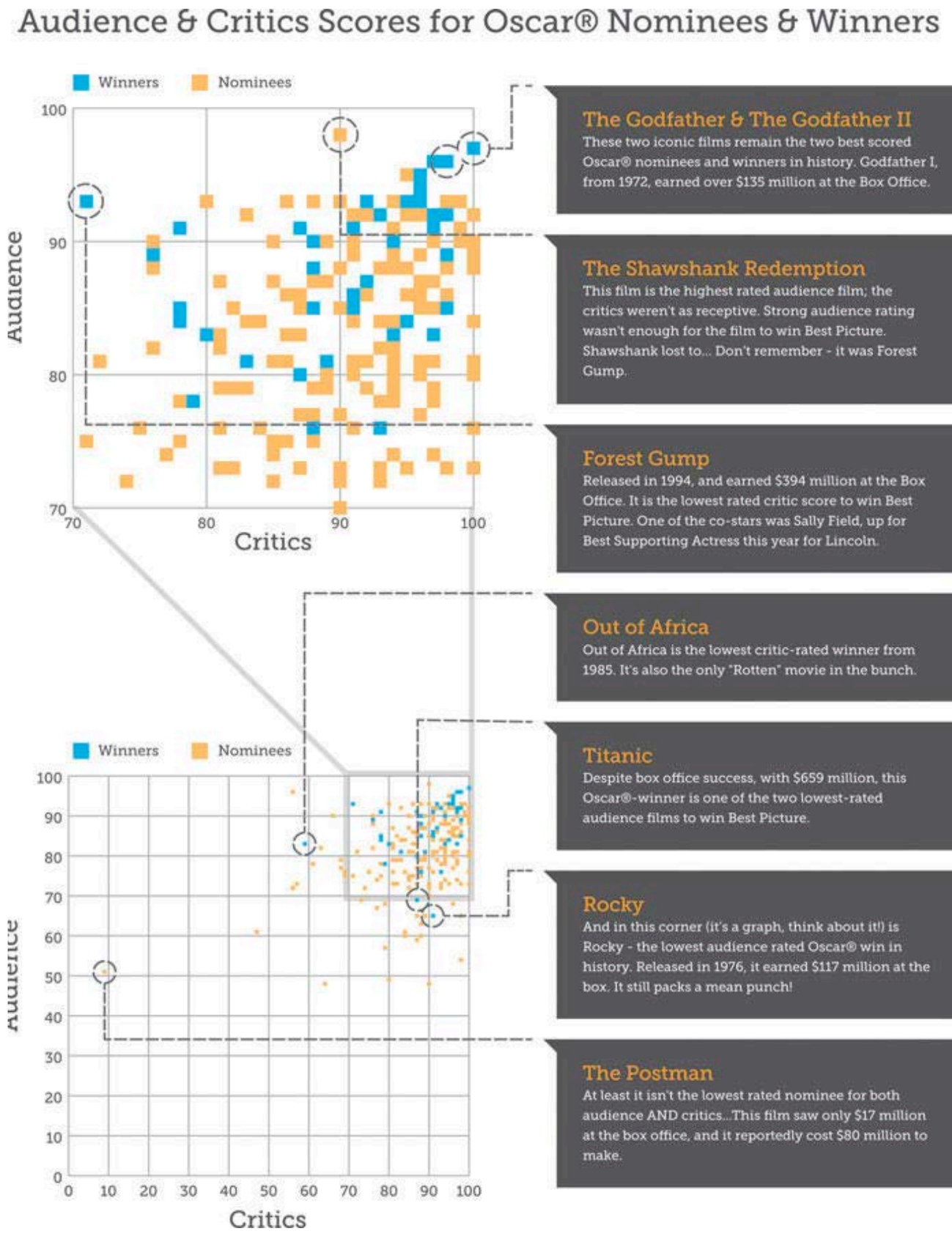
One variable we looked at was Movie Earnings

- In particular, do winners of the Best Picture Oscar have greater earnings per cost?
- It appears that winners of the Best Picture Oscar do better in terms of profit, than other nominees.
- However, this is not predictive because some of the ticket sales occur after the film wins.



Rotten Tomatoes and the Oscars

- Rotten Tomatoes is a website that aggregates the critics reviews for a movie, as well as user submitted reviews.
- A movie gets a Critics Score (1-100) and Audience Score (1-100) indicating the percentage of favorable reviews by critics or the general public.
- Once again, we cannot separate reviews from before the awards with after.



We built a multinomial logistic regression model to predict the winners.

- For each category the key variables were chosen by identifying which had the strongest predictive power.

- Then we created models that were the best predictions given which other awards have been given.
  - Right after the nomination, we had a model that was based on only characteristics of the movie and the other nominations.
  - After, for example, the Producers Guild awards some models would be updated to include that variable.

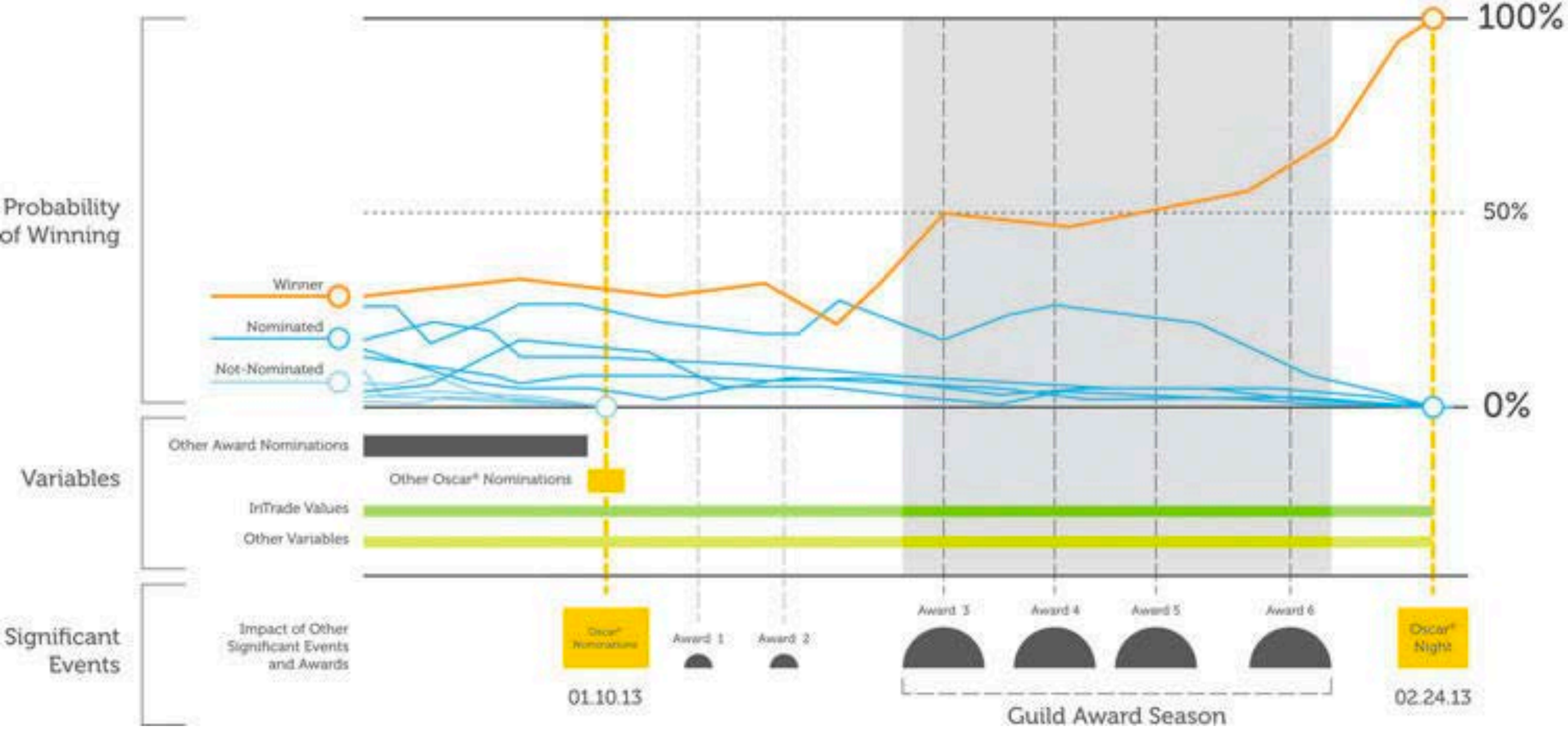
- The models have the form
  prob. of winning $= f\left(c + \beta_1(\text{won award A}) + \beta_2(\text{\# of nominees}) + \ldots\right)$

- The coefficients and the variables in the model changed as the

- Therefore, our prediction varied over time as more data was available.

How The Oscar® Is Won

The likely winner in a given category is predicted based on a series of EVENTS and other VARIABLES. These SIGNALS, accumulated throughout awards season, inform and evolve the probability of winning.



In 2013 Our Model Predicted 5/6 Categories Correctly.

- The categories of Best Actor, Best Actress, and Best Supporting Actress were widely predicted in other sources.
- *Argo* as best picture and Christoph Waltz as best supporting actor were correct picks that most other pundits
- We (and almost everyone else) incorrectly picked Steven Spielberg over Ang Lee.
- Other predictions: *Atlantic* magazine (4/6), *New York Times* (5/6), Nate Silver, 538 (4/6).

Best Picture Nominees	Farsite
Argo	45.6%
Lincoln	10.3%
Silver Linings Playbook	21.5%
Les Miserables	0.5%
Life of Pi	4.5%
Amour	4.0%
Zero Dark Thirty	1.2%
Django Unchained	0.2%
Beasts of the Southern Wild	12.1%

Best Director Nominees	Farsite
Steven Spielberg	78.6%
Ang Lee	12.0%
David O Russell	5.4%
Michael Haneke	3.0%
Benh Zeitlin	1.0%

Best Actor Nominees	Farsite
Daniel Day-Lewis	57.5%
Hugh Jackman	18.6%
Bradley Cooper	17.2%
Joaquin Phoenix	6.2%
Denzel Washington	0.5%

Best Actress Nominees	Farsite
Jennifer Lawrence	65.6%
Emmanuelle Riva	19.3%
Jessica Chastain	9.0%
Naomi Watts	1.5%
Quvenzhane Wallis	4.7%

Best Supporting Actor Nominees	Farsite
Christoph Waltz	44.0%
Tommy Lee Jones	35.4%
Robert De Niro	11.7%
Philip Seymour Hoffman	4.9%
Alan Arkin	4.0%

Best Supporting Actress Nominees	Farsite
Anne Hathaway	94.3%
Helen Hunt	4.4%
Sally Field	0.9%
Jacki Weaver	0.2%
Amy Adams	0.2%

	Best Picture	Best Director	Best Actor	Best Actress	Best Supporting Actor	Best Supporting Actress
Actual Winners	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Farsite Forecast	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
PredictWise (David Rothschild)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Ben Zauzmer	Gravity	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Betfair	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Thelma Adams (Yahoo)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Matt Atchity (Rotten Tomatoes)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	Oprah Winfrey
Kyle Buchanan (New York Vulture)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Mike Cidoni (Associated Press)	**12 Years a Slave**	**Alfonso Cuaron**	Bruce Dern	**Cate Blanchett**	**Jared Leto**	June Squibb
Edward Douglas (Coming Soon)	American Hustle	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Scott Feinberg (Hollywood Reporter)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Thom Geler (Ent. Weekly)	**12 Years a Slave**	**Alfonso Cuaron**	Bruce Dern	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Pete Hammond (Deadline Hollywood)	American Hustle	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Mark Harris (Grantland)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Time Hayne (Moviefone)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Michael Hogan (Vanity Fair)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Dave Karger (Fandango)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Tariq Khan (Fox News)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	Michael Fassbender	Oprah Winfrey
Guy Lodge (In Contention HitFix)	American Hustle	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Scott Manty (Access Hollywood)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Mary Milliken (Reuters)	**12 Years a Slave**	Steve McQueen	Chitewel Ejiofer	Sandra Bullock	Michael Fassbender	**Lupita Nyong'o**
Michael Musto (Village Voice)	**12 Years a Slave**	Steve McQueen	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Tom O'Neil (Gold Derby)	American Hustle	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Kevin Polowy (Yahoo)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Steve Pond (The Wrap)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Claudia Puig (USA Today)	**12 Years a Slave**	**Alfonso Cuaron**	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	Oprah Winfrey
Christopher Rosen (Huffington Post)	**12 Years a Slave**	**Alfonso Cuaron**	Bruce Dern	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Paul Sheehan (Gold Derby)	American Hustle	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Keith Simanton (IMDB)	**12 Years a Slave**	**Alfonso Cuaron**	Robert Redford	**Cate Blanchett**	**Jared Leto**	Jennifer Lawrence
Sasha Stone (Awards Daily)	**12 Years a Slave**	Steve McQueen	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Jeff Wells (Hollywood Elsewhere)	American Hustle	Steve McQueen	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Glenn Whipp (LA Times)	**12 Years a Slave**	**Alfonso Cuaron**	**Mathew McConaughey**	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**
Susan Wloszczyna (RogerEbert.com)	**12 Years a Slave**	Steve McQueen	Chitewel Ejiofer	**Cate Blanchett**	**Jared Leto**	**Lupita Nyong'o**

We went 6/6 in 2014

It was an "easy" year however with no upsets.

And now some Science: Improving Fusion experiments with Machine Learning

Inertial Confinement fusion attempts to create a sustained, controlled nuclear fusion reaction in the laboratory.

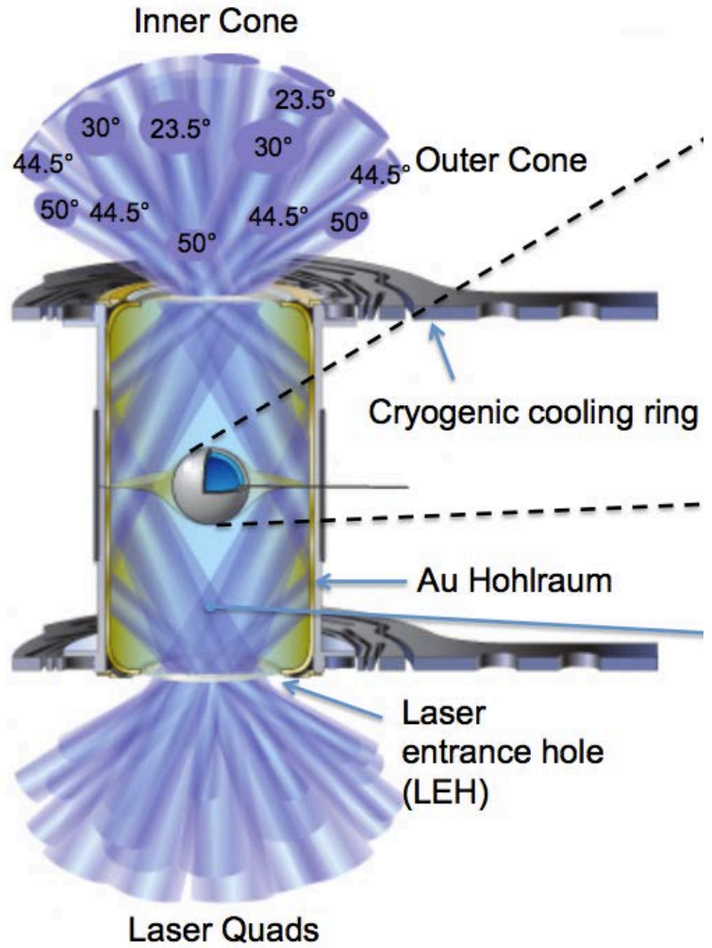
- Or asks the question: what can we do if we take the energy from a laser this size

And focus it here in 20 ns.

Image credits: atomicinsights.com and wired.com

National Ignition Facility fires 192 laser beams at a target to compress a pellet of cryogenic hydrogen to high-density

- The process involves lasers heating the gold to plasma temperatures.
- The gold plasma radiates x-rays that heat the edge of the fuel pellet.
- Ablation of the pellet drives a shock wave into the fuel.
- The converging shocks and compression of the fuel lead to ignition of the fusion reaction.
- The fusion products are helium ions and neutrons.



Magnitude of implosion is a remarkable feat.

- High implosion velocity
- 300 – 450 km/s
- Compression ratio
  - 20-40x in radius
  - Volume reduced by 30,000x
- Stagnation pressure ~100 Gbar
- Each NIF shot begins at 16 K and ends at 50 million K

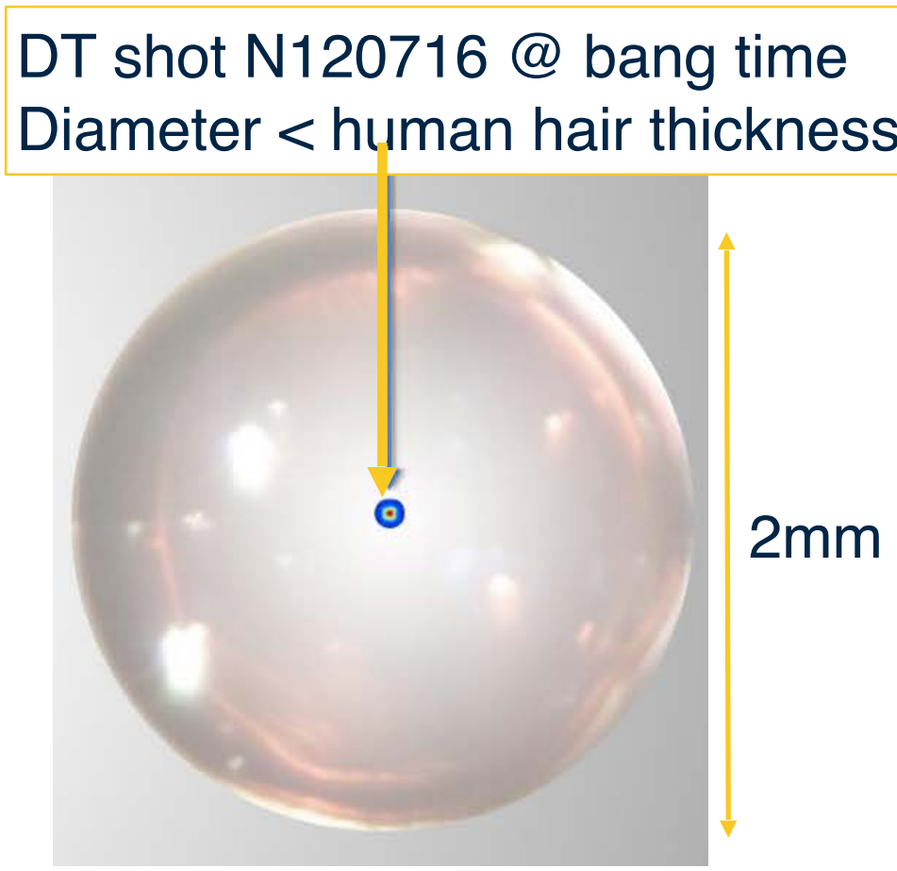
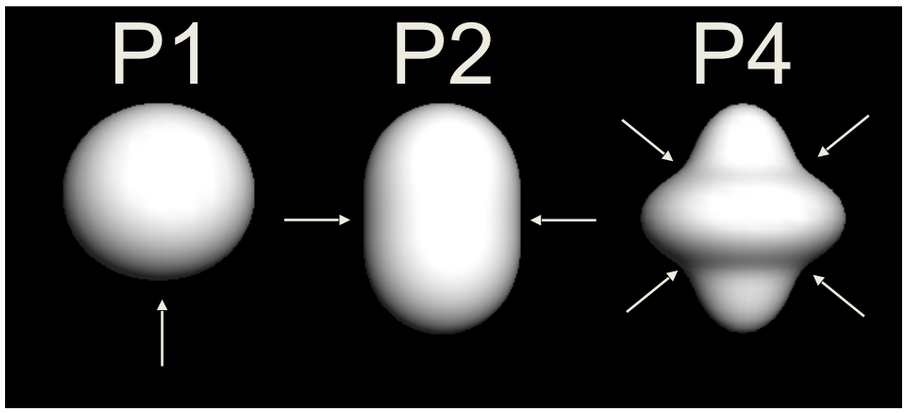


Image: Steve Obenschain

The Trinity Open Science I Database was the largest ever created for ICF simulations

- 3-shock HDC Ignition Design
  - HYDRA 2D Capsule (D. Ho)

- Varied 9 parameters
  - Time-varying drive asymmetry
    - Legendre Modes P1, P2, P4
  - Time-varying drive amplitude
  - Capsule gas fill density



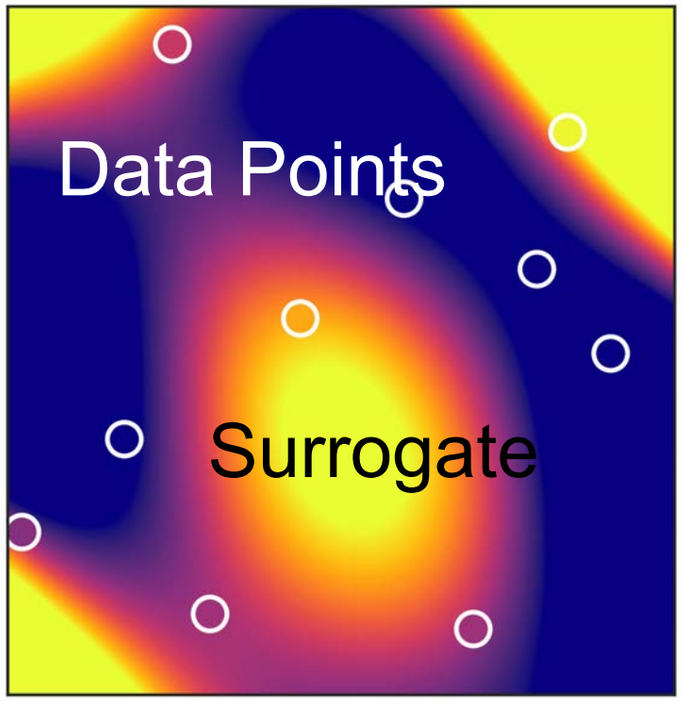
- Successfully completed over 60k simulations, 39 Million CPU Hours (Trinity @ LANL),
  5 PB Raw Data, 100 TB Processed & Zipped

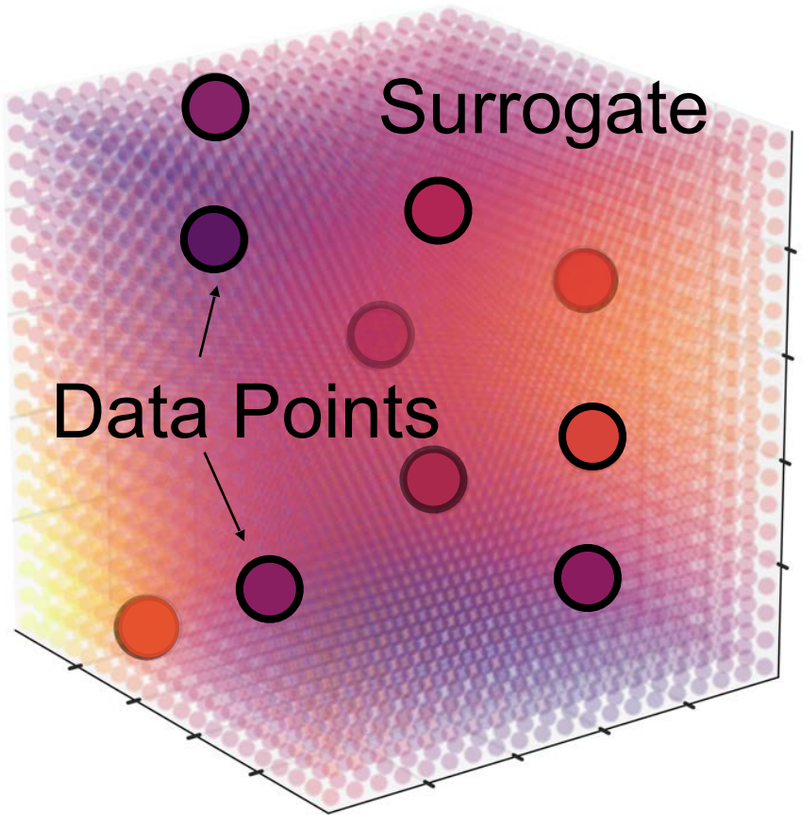
Goal: build surrogate model and search 9D design space for a high performance implosion

Surrogate models are fits to simulation data

$$y \simeq f(x_1, x_2, x_3, ...)$$

- Trained on data of arbitrary dimensionality
- Quickly estimate result where no data exist
- Examples:
  - Polynomial curve fitting
  - Power laws
  - Random forest regressors
  - Neural Networks

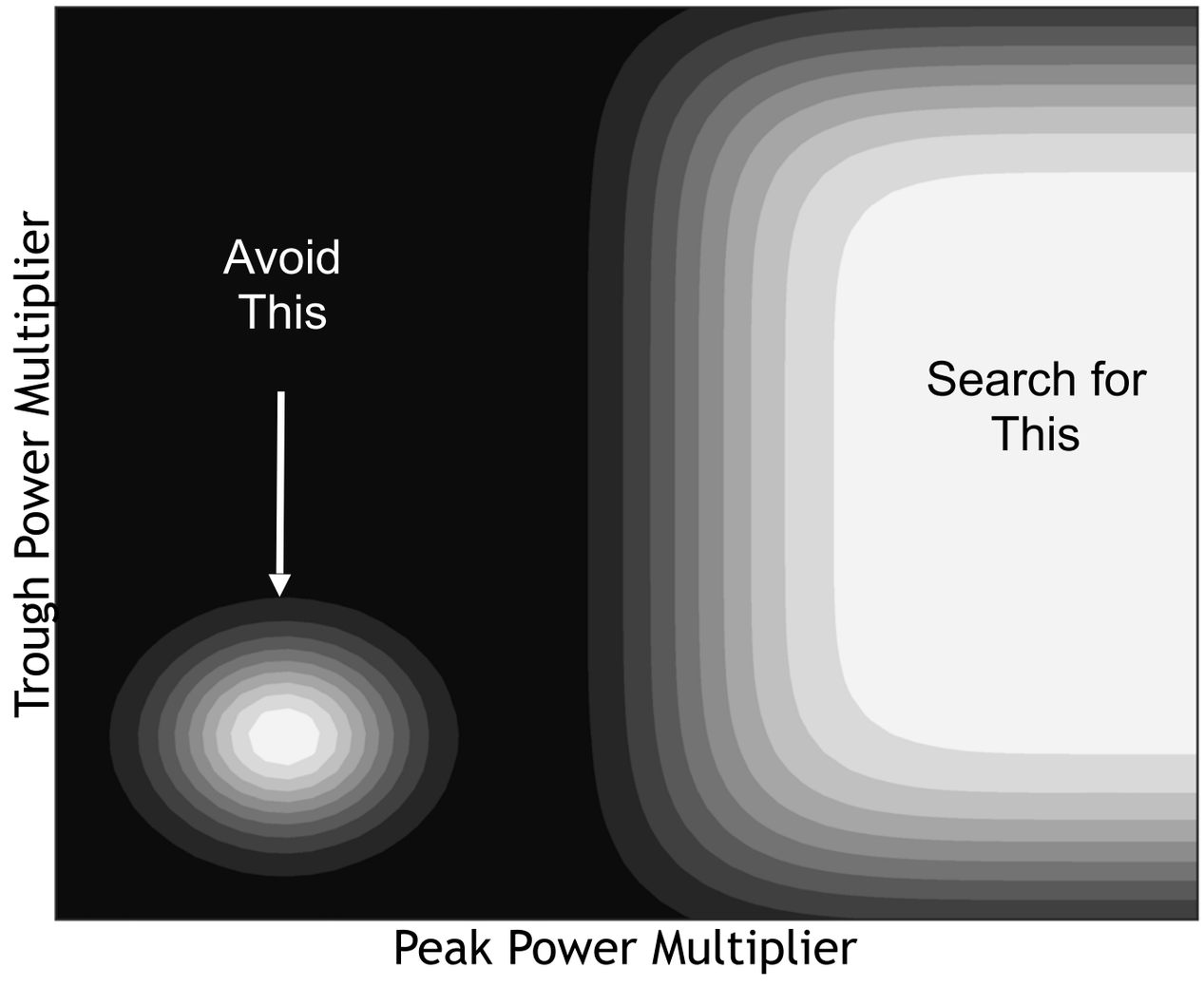




Surrogates are fast approximations to expensive simulations

Surrogate model can help quantify implosion robustness

- Goal: search for a high yield “plateau” in design space (if one exists)

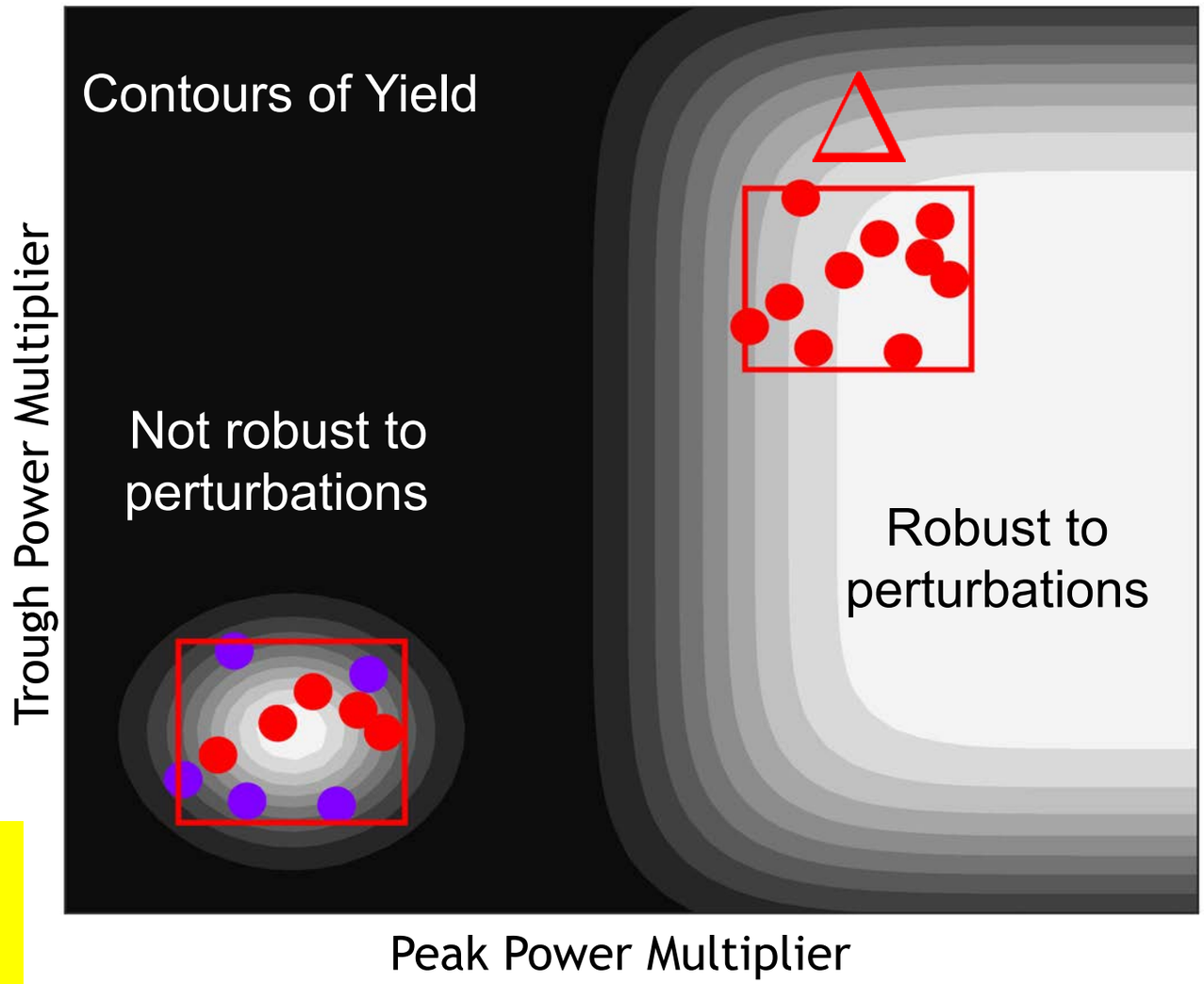


Surrogate model can help quantify implosion robustness

- Goal: search for a high yield “plateau” in design space (if one exists)

- Use surrogate to quantify robustness of yield to perturbations
  - Pick a point
  - Randomly perturb many times (1000)
  - Count fraction of random samples that surrogate says achieve high yield:

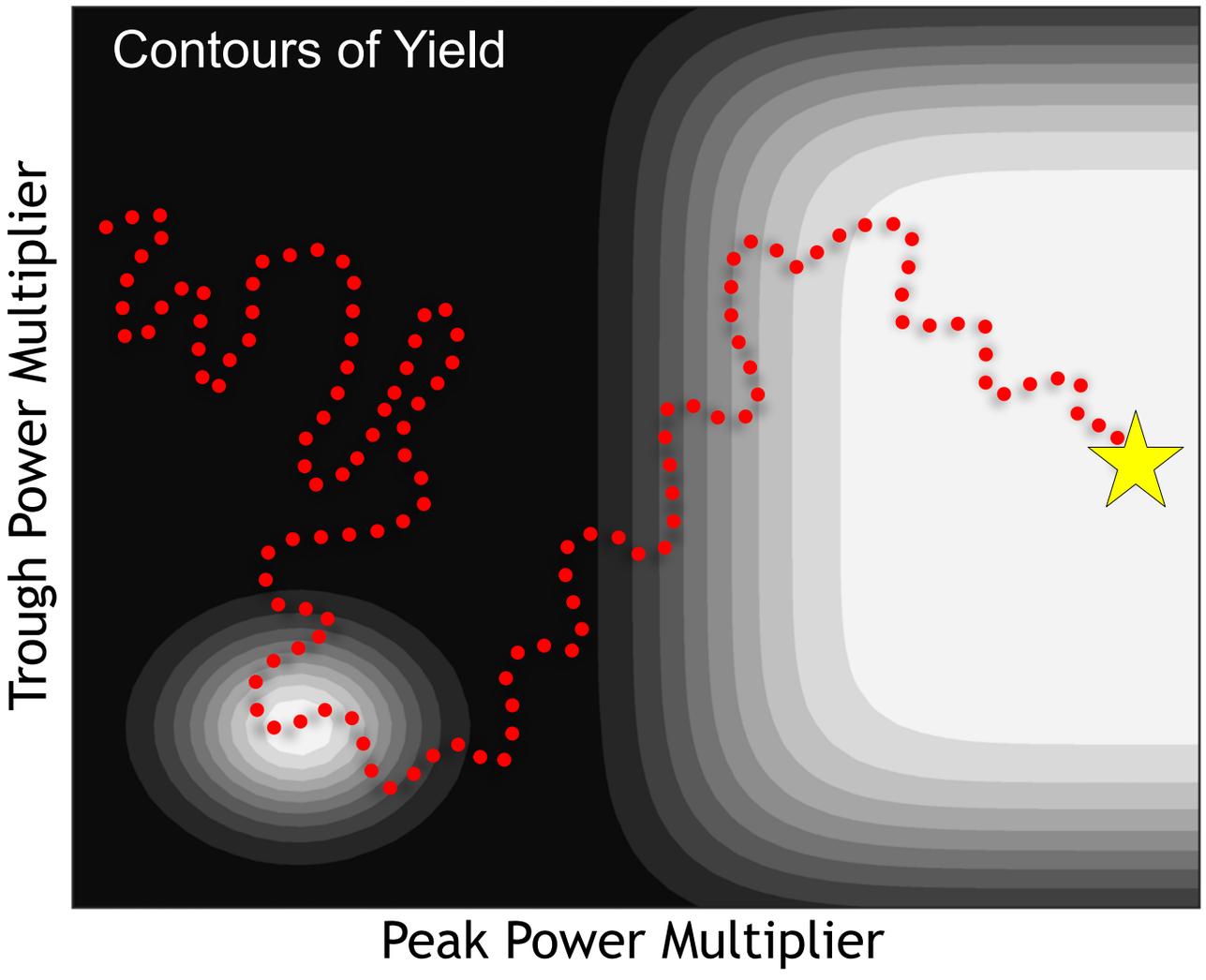
A robust design ignites when perturbed



Surrogates enable rapid search through design space for a robust, high yield implosion

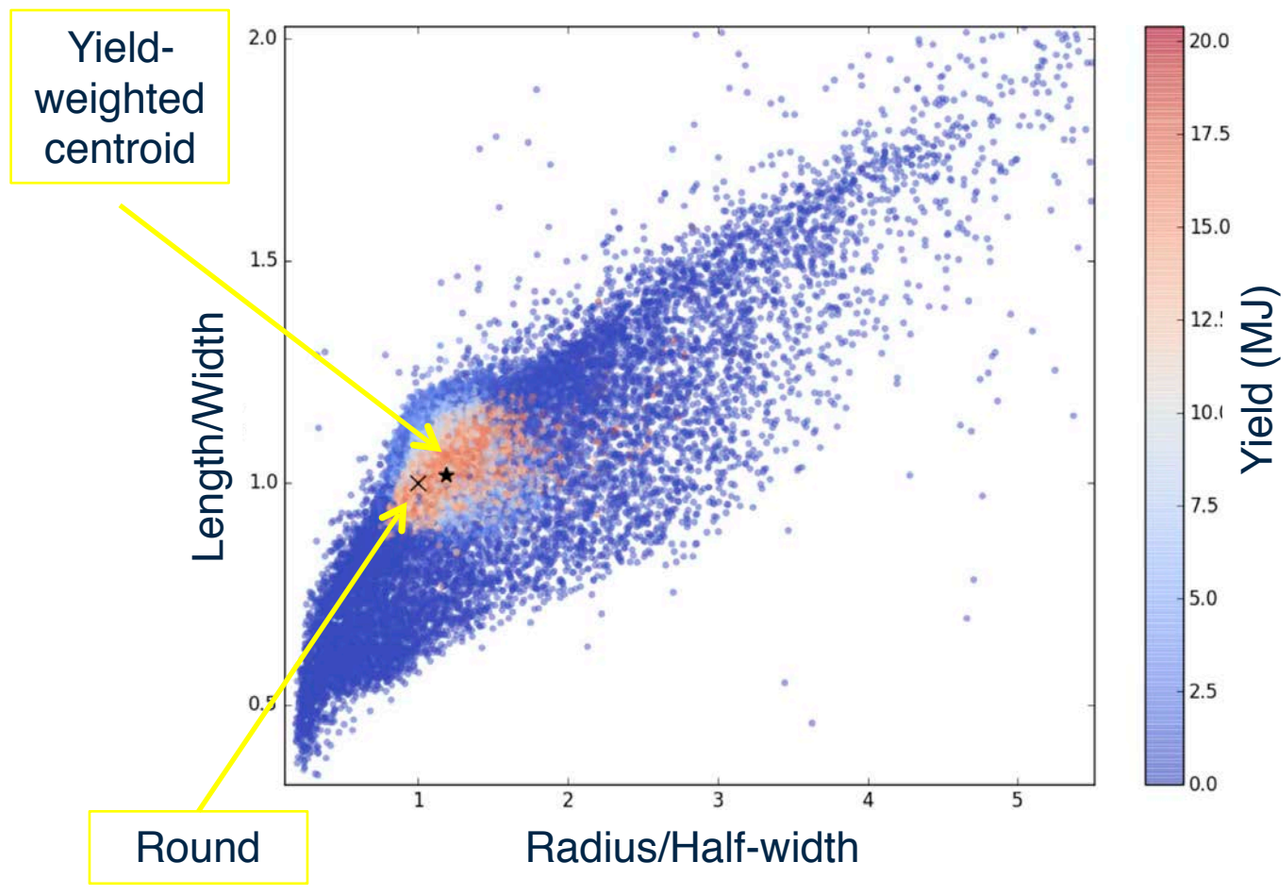
Search requires 5 million surrogate evaluations:

~many million CPU hours on Trinity for HYDRA
OR
~an hour with the surrogate



High yield implosions characterized by shapes that are typically not round

- A central tenet of inertial confinement fusion is that spherical implosions are best.
- Using the random forest regression model, determine where in input space are the largest, most stable yields.
- Using machine learning we are questioning long held beliefs of the community.
- *Artificial intelligence discovered how to do fusion.*



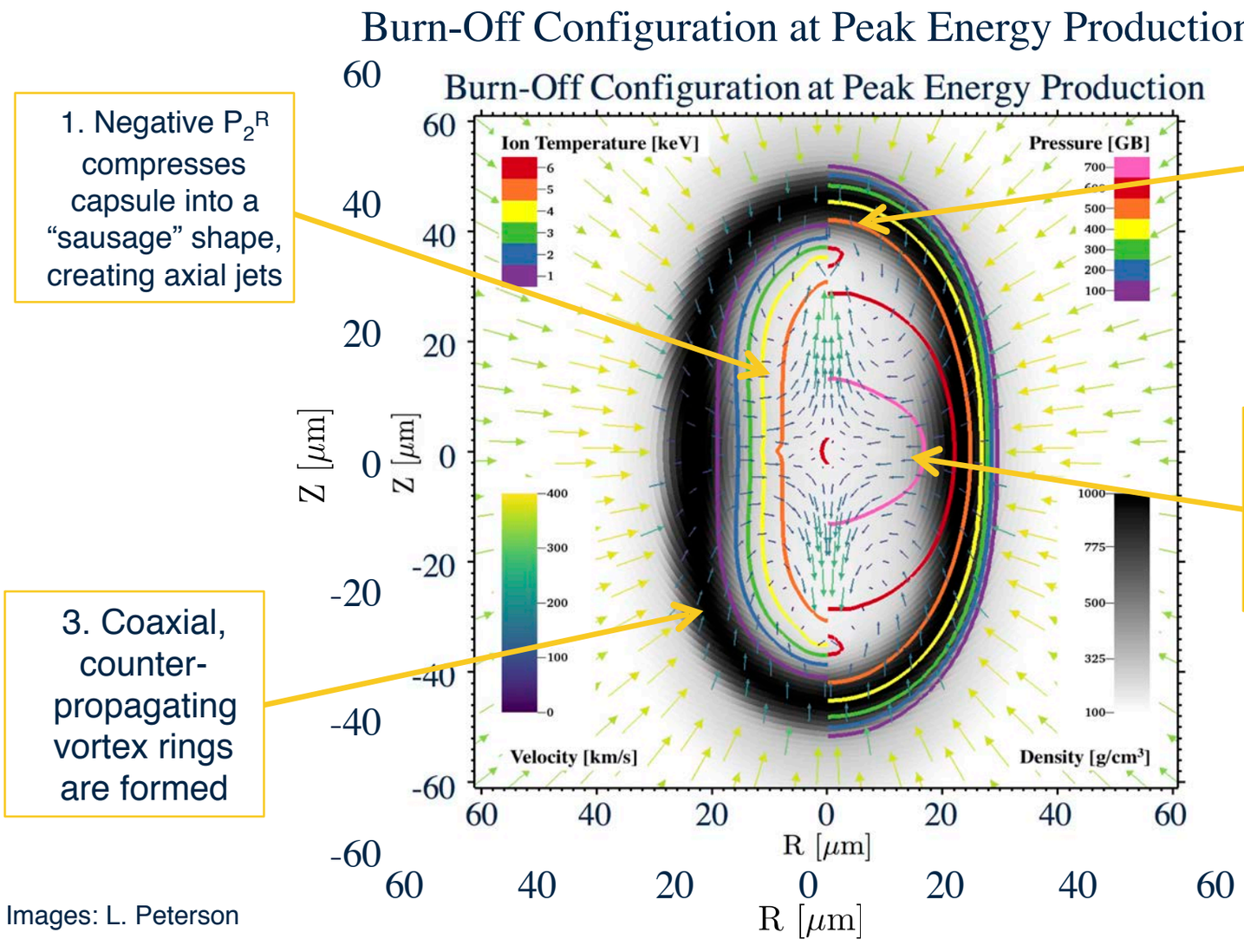
Squash the Pellet from the side and then from the top to get higher fusion yield.

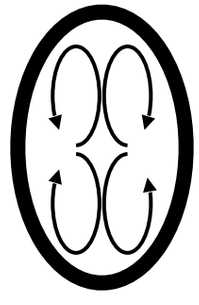
1. Negative P_2^R compresses capsule into a “sausage” shape, creating axial jets

2. Positive P_2^P prevents jets from escaping

3. Coaxial, counter-propagating vortex rings are formed

Flows circulate into hotspot at equator, out at the poles.

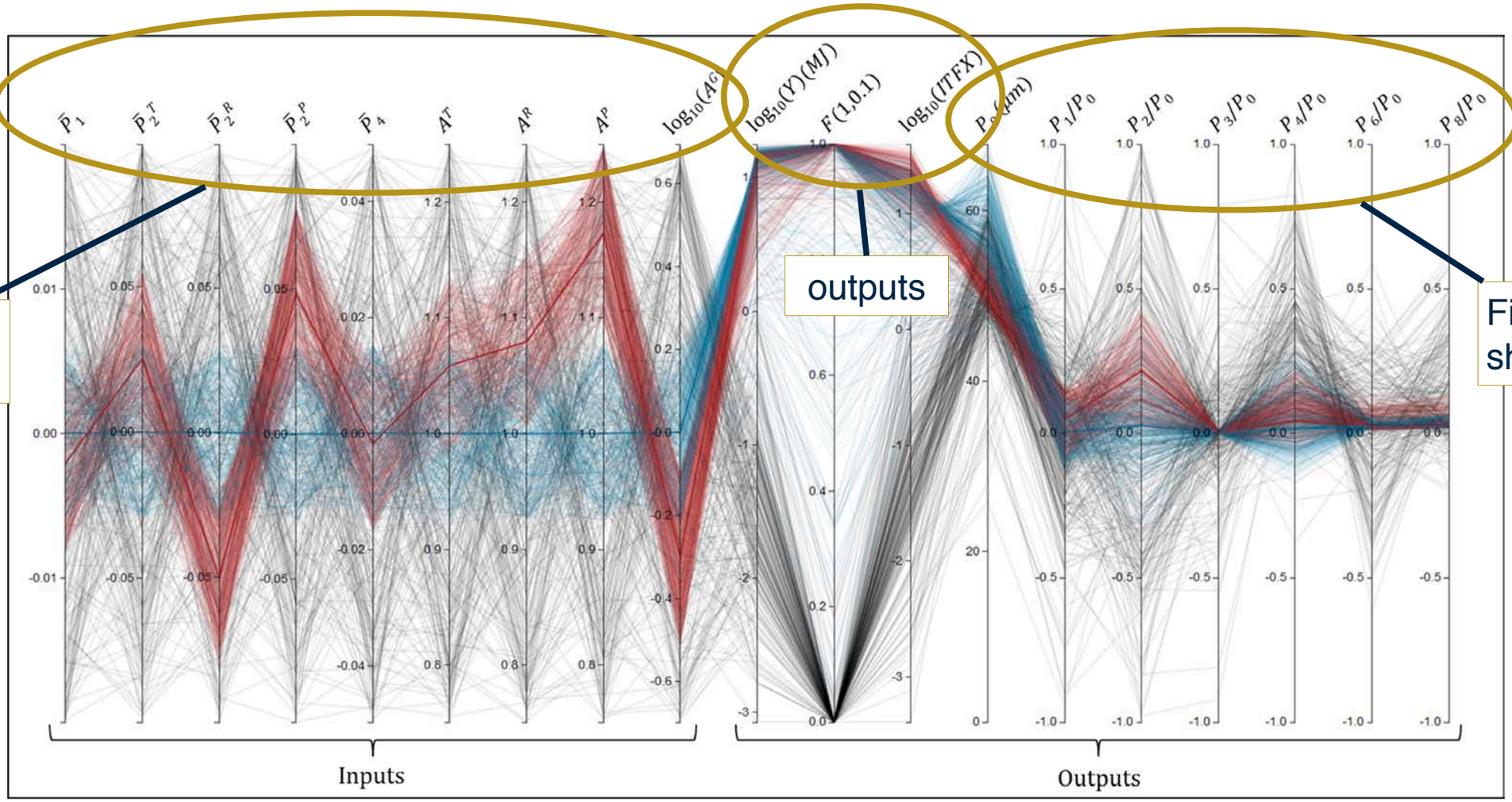




Images: L. Peterson

Petersen, Humbird, et al., Physics of Plasmas > Volume 24, Issue 3 > 10.1063/1.4977912

Two families of high performing implosions: round (blue), time-varying P_2 ovoid (red).



Inputs: P_1, $P_2^{(T,R,P)}$, P_4 : asymmetry perturbations. $A^{(T,R,P)}$: drive magnitude; A^G: gas fill density. (**T**rough-end of first shock, **R**ise to peak laser power, **P**eak drive.)

Data Science is truly an Interdisciplinary Endeavor

- It is interdisciplinary in the underpinning tools:
  - Statistics
  - Computer Science
  - Mathematics
  - Engineering

- Data Science and Machine Learning can be applied to a variety of fields (interdisciplinary in application).

- This is a great field to work in for those that love new challenges, and being generalists (in the best sense of the word).
  - Can always apply your skills to new problems.
  - Need to know a lot about a range of topics.

- These skills are also clearly in demand worldwide.

DATA SCIENTIST JOB DEMAND AT A GLANCE

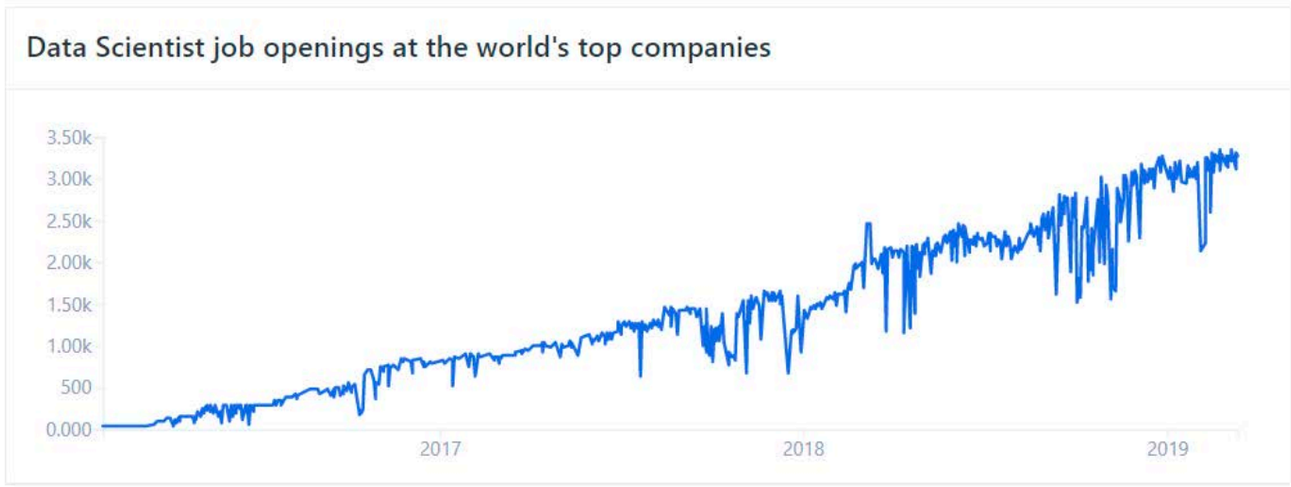
28%
Projected increase in demand for data scientists through 2020
Forbes.com

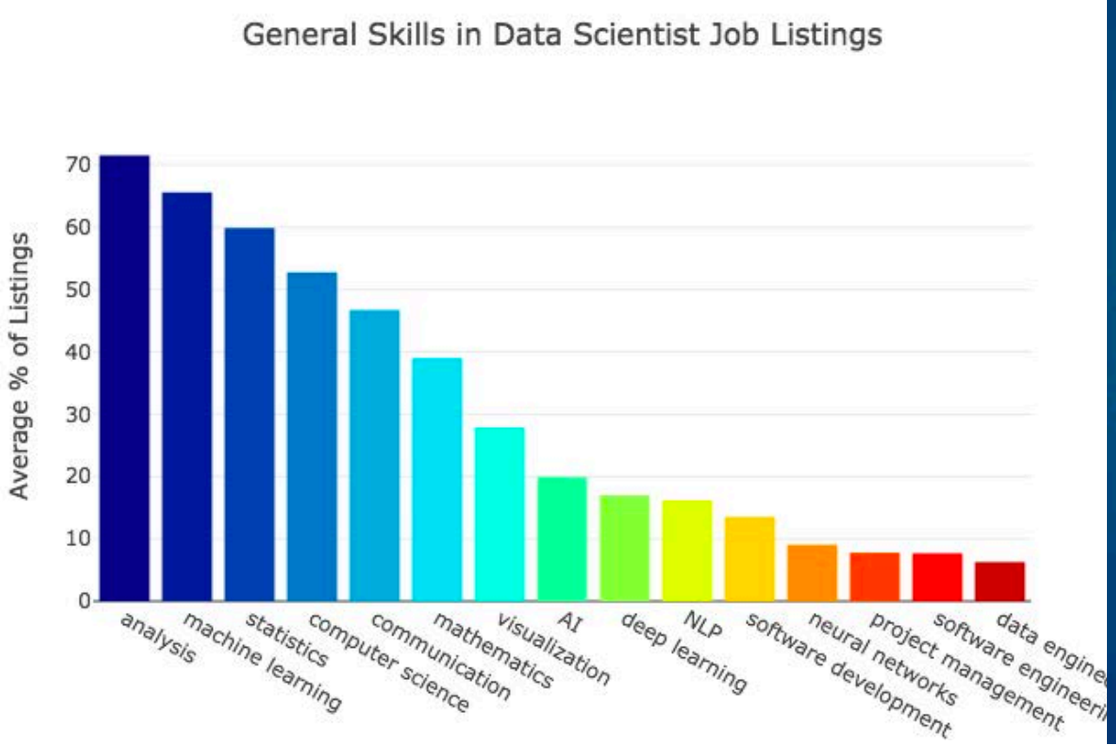
Demand for data scientists could outpace supply by **250,000** jobs by 2024
UsNews.com

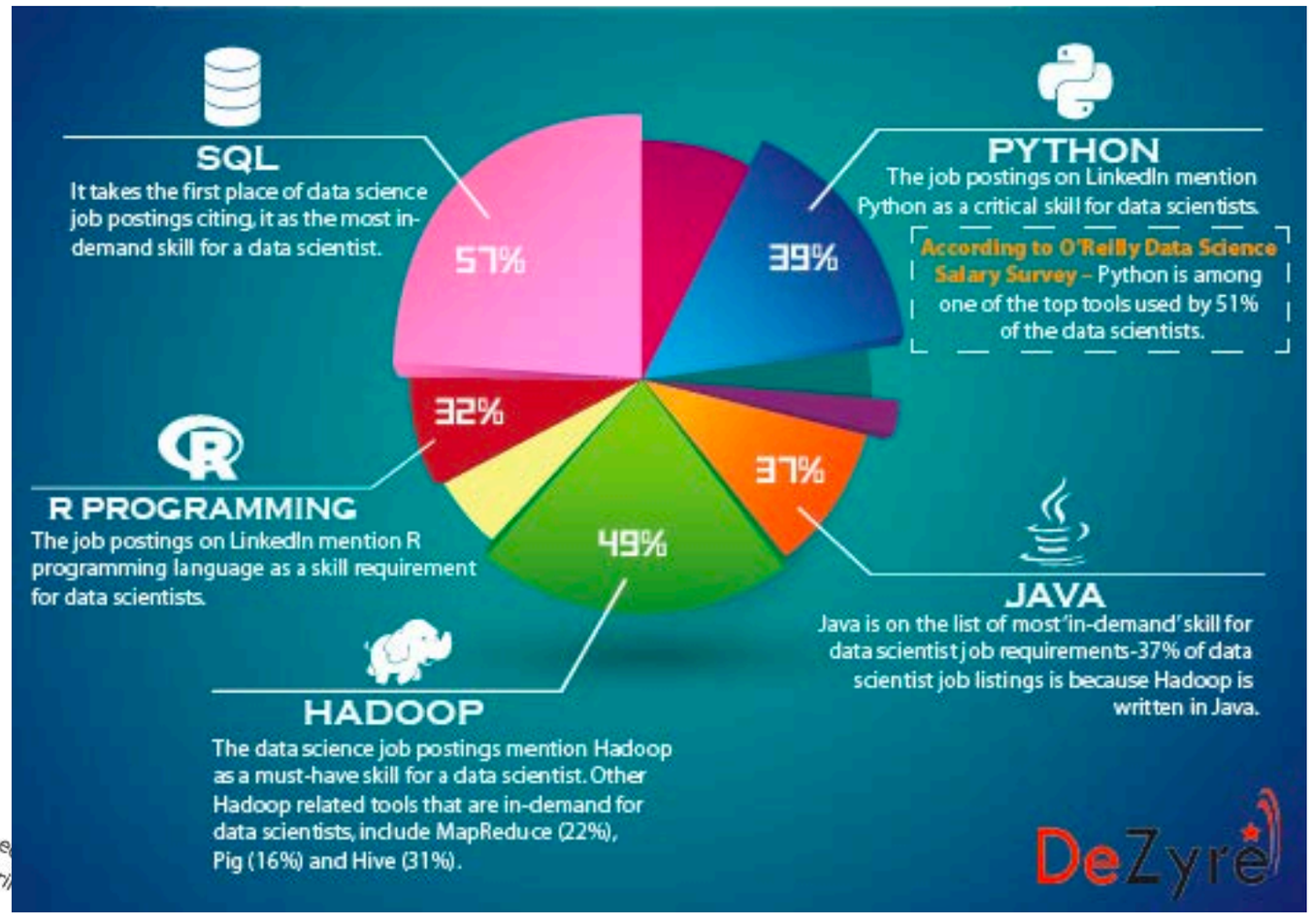
About **3 in 5** data science and analytics jobs are in three fields:
Forbes.com

- Finance & Insurance
- IT
- Professional Services

$ 110,000
Median income for data scientist job listings
Forbes.com

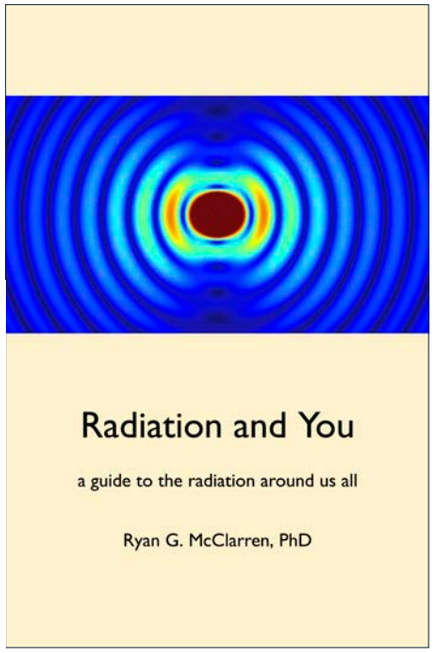






My books

I have three books available:
Uncertainty Quantification and Predictive Computational Science from Springer https://www.springer.com/gp/book/9783319995243



Computational Nuclear Engineering and Radiological Science Using Python from Academic Press http://a.co/2HdisVb