

AI: THE CATHEDRAL AND THE BAZAAR

SOFTWARE FREEDOM DAY, PANJAB UNIVERSITY, 2019

JAIDEV DESHPANDE, SENIOR DATA SCIENTIST



/@jaidevd

“The most important book about technology today,
with implications that go far beyond programming.”
—Guy Kawasaki

Revised & Expanded

THE CATHEDRAL & THE BAZAAR

MUSINGS ON LINUX AND OPEN SOURCE
BY AN ACCIDENTAL REVOLUTIONARY

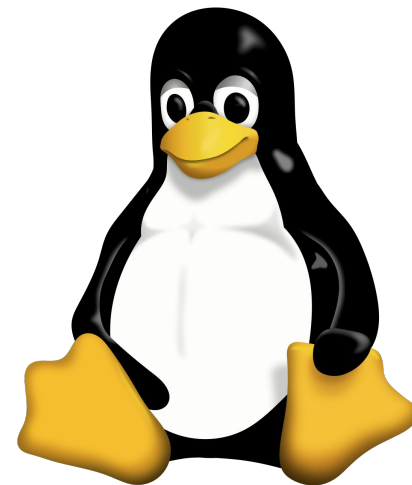


ERIC S. RAYMOND

WITH A FOREWORD BY BOB YOUNG, CHAIRMAN & CEO OF RED HAT, INC.



WIKIPEDIA
The Free Encyclopedia



NEW YORK TIMES BESTSELLER

“Provocative and fascinating.” —MALCOLM GLADWELL

Daniel H. Pink

author of *A Whole New Mind*

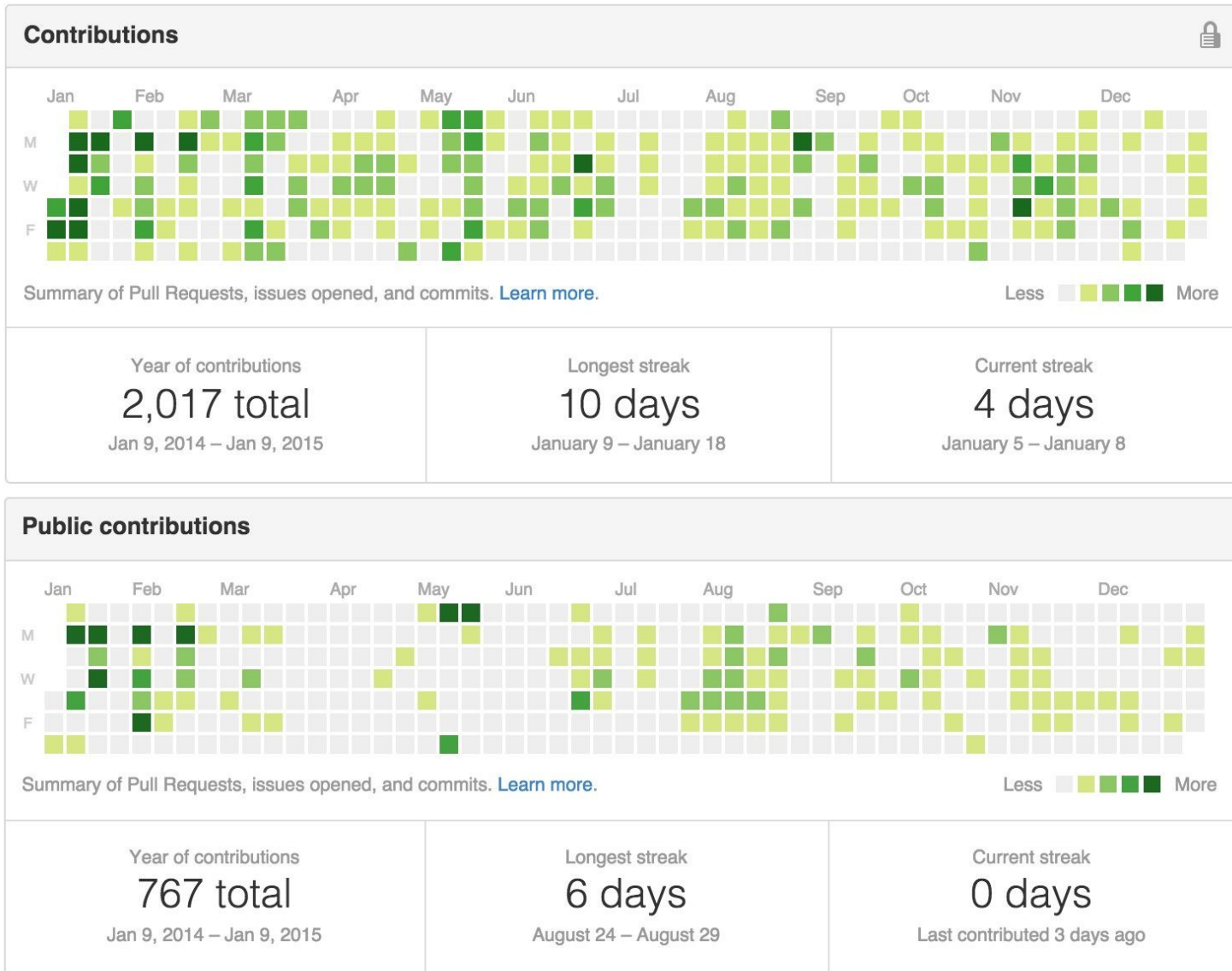
DRiVE

The Surprising Truth
About What Motivates Us

- **MASTERY** – If you’re a FOSS contributor, you already are above average.
- **AUTONOMY** – Responsibility & Ownership for the software you produce
- **PURPOSE** –
Every good work of software starts by scratching a developer’s personal itch.

– Eric Raymond

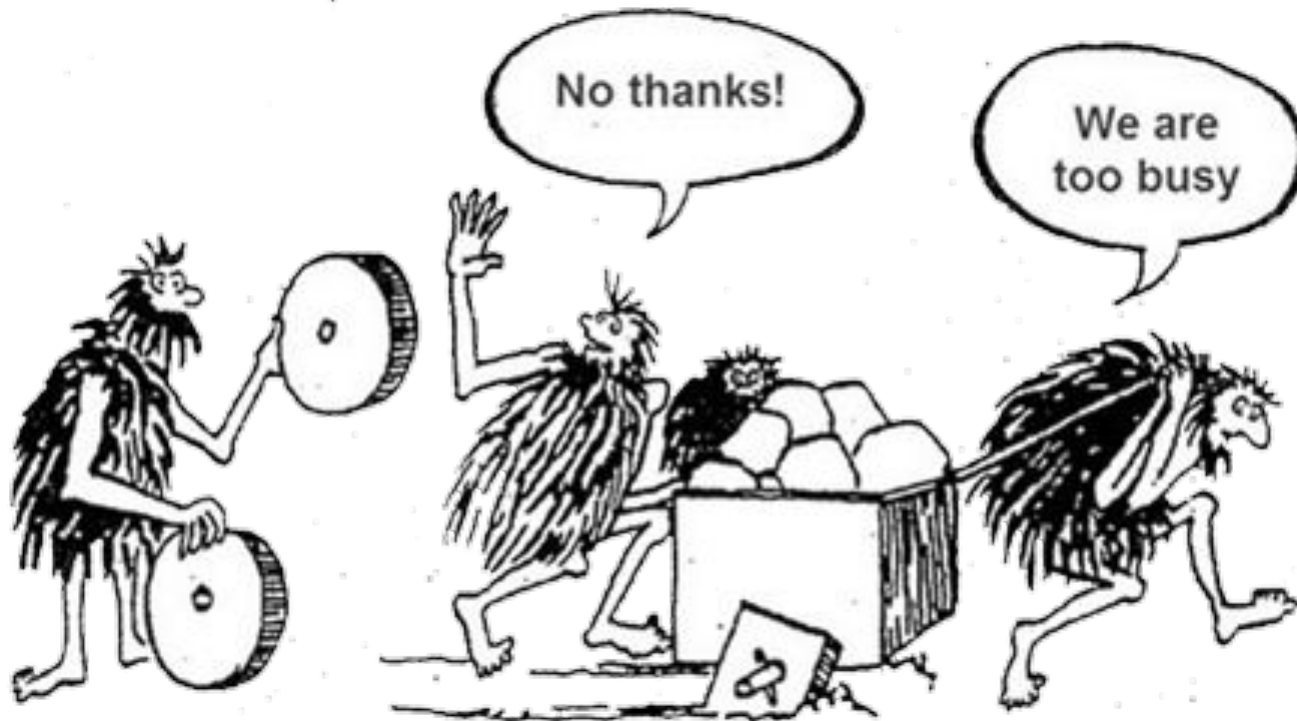
THE DARK SIDE OF OPEN SOURCE



THE DARK SIDE OF OPEN SOURCE

- Some of the best FOSS work is not on GitHub
- Contribution graphs can be misleading, (sometimes as misleading as LinkedIn endorsements)
- Continuous commit streaks are easy to create and game
- Standalone scripts vs modular code
- ***EVERY DEVELOPMENT PRODUCTIVITY METRIC IS WRONG!***

THE “NOT INVENTED HERE” SYNDROME

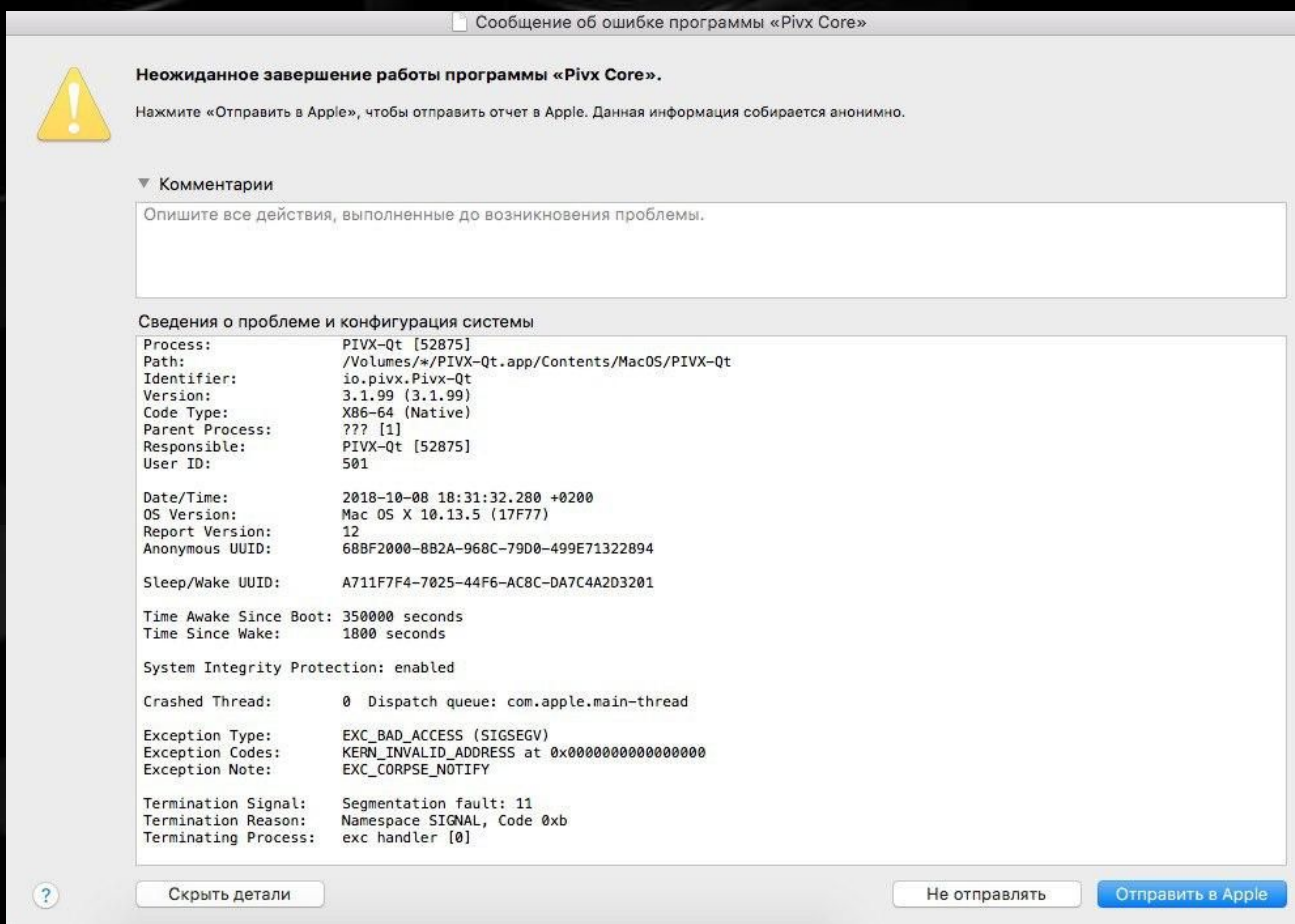


You better have very good reasons for not standing on the shoulders of giants.

THE YEAR 2011

- My first MOOCs:
 - Convex Optimization by Stephen Boyd
 - Machine Learning by Andrew Ng
- My First PyCon and SciPy!
- My First FOSS Projects
 - github.com/jaidevd/pyhht
 - github.com/scikit-signal/tftb

THE LESS KNOWN PYTORCH



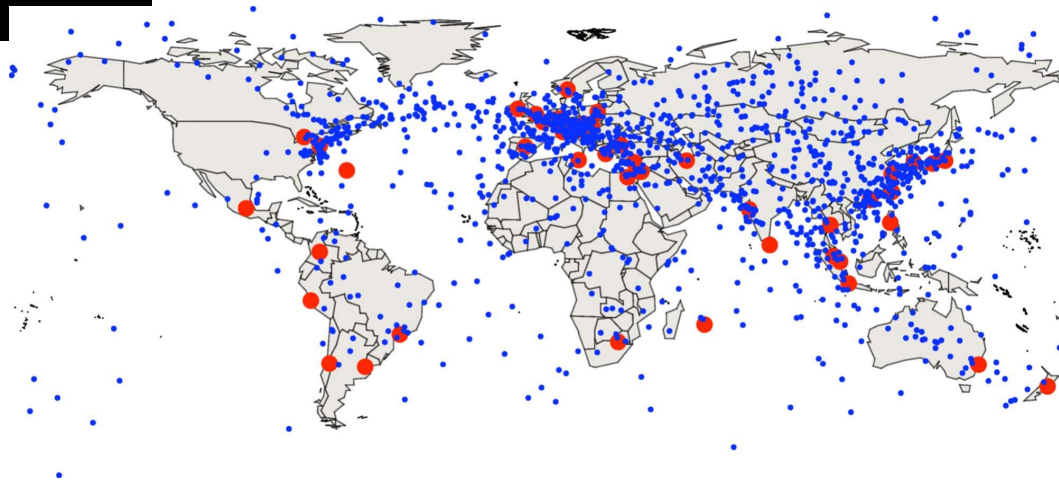
THE *HIGH ALTITUDE* SERIES BY MICHAEL NAJJAR



HOW ALGORITHMS SHAPE OUR WORLD - KEVIN SLAVIN



“We’re writing things we can no longer read. We’ve lost the sense of what’s happening in this world we’ve made.”



Wissner-Gross, A.D. and Freer, C.E., 2010.
Relativistic statistical arbitrage. *Physical Review E*, 82(5), p.056104.



WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

THE HONEST VIEW

1. A Redundant Misnomer
(@johndcook)
2. Applied Statistics
3. *A lot* of software engineering
4. Deliverables are end-to-end
software services or products

THE FASHIONABLE VIEW

1. The Sexist Job of the 21st Century
(HBR)
2. Artificial Intelligence / Deep Learning
3. Notebooks and Spreadsheets
4. Deliverables are cool visualizations
and deep networks



WHO IS A DATA SCIENTIST?

WHO IS A DATA SCIENTIST?

- *“Someone who knows more statistics than an average computer scientist and more computer science than the average statistician.” – DJ Patil*
- Someone who:
 - knows linear algebra, calculus, probability & stats
 - is happy to write application code and likes working with databases and filesystems
- Someone who knows:
 - Linear Algebra
 - Probability and Statistics
 - Calculus
 - Programming
- Almost every STEM student studies this!





Programmers Need To Learn Statistics Or I Will Kill Them All

I have a major pet peeve that I need to confess. I go insane when I hear programmers talking about statistics like they know shit when it's clearly obvious they do not. I've been studying it for years and years and **still** don't think I know anything. This article is my call for all programmers to finally learn enough about statistics to at least know **they don't know shit**. I have no idea why, but their confidence in their lacking knowledge is only surpassed by their lack of confidence in their personal appearance.

RECENT POSTS

[Bob Ross, Pigmented Lullabies, And WordPress Sucks](#)

[The Billionaires vs. BrandonM](#)

[Photographing Art](#)

[What If It Worked?](#)

[Protected: The PSF's Next Steps](#)

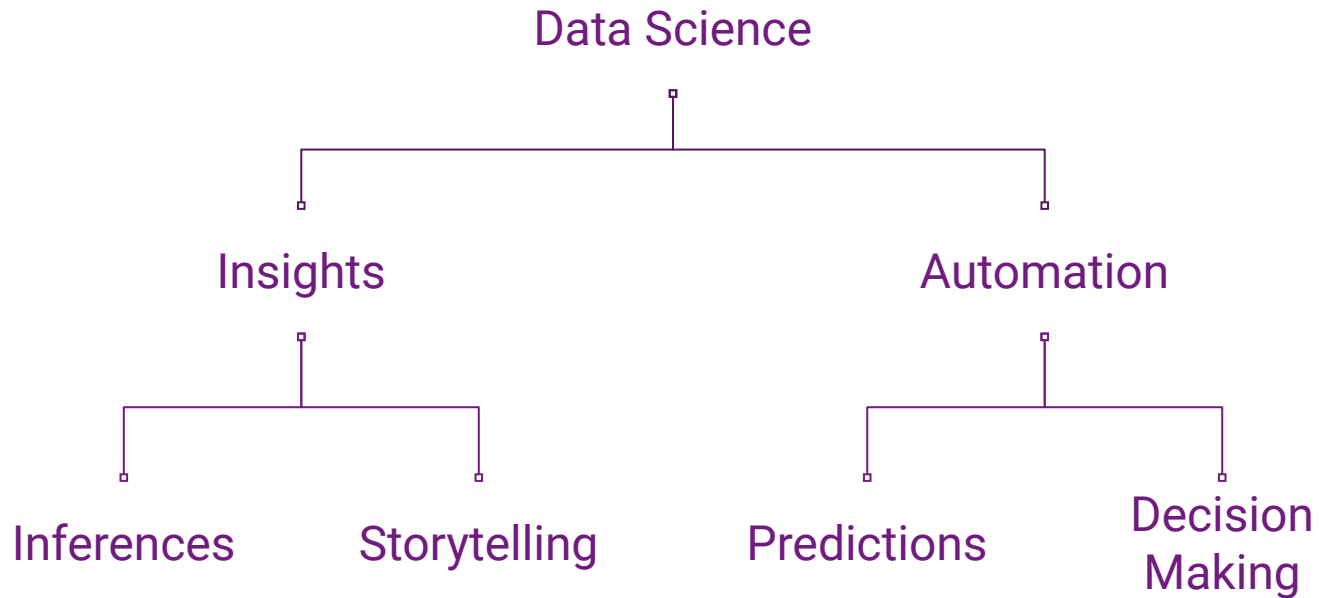
THE “HOW TO...” OF DATA SCIENCE

- All of the above can be learnt
- Programming is important because:
 - Obvious reasons
 - Programming ability correlates with analytical thinking
 - Test of how well you understand the material
- What's the best resource? No right answer, but:
 - Pick a problem and a learning resource - *any* problem and *any* resource
 - Keep learning and trying to solve the problem until:
 - The resource is no longer relevant or the problem is solved
 - ***DROP* the resource!**
 - Pick another resource & repeat
 - Write code every step of the way

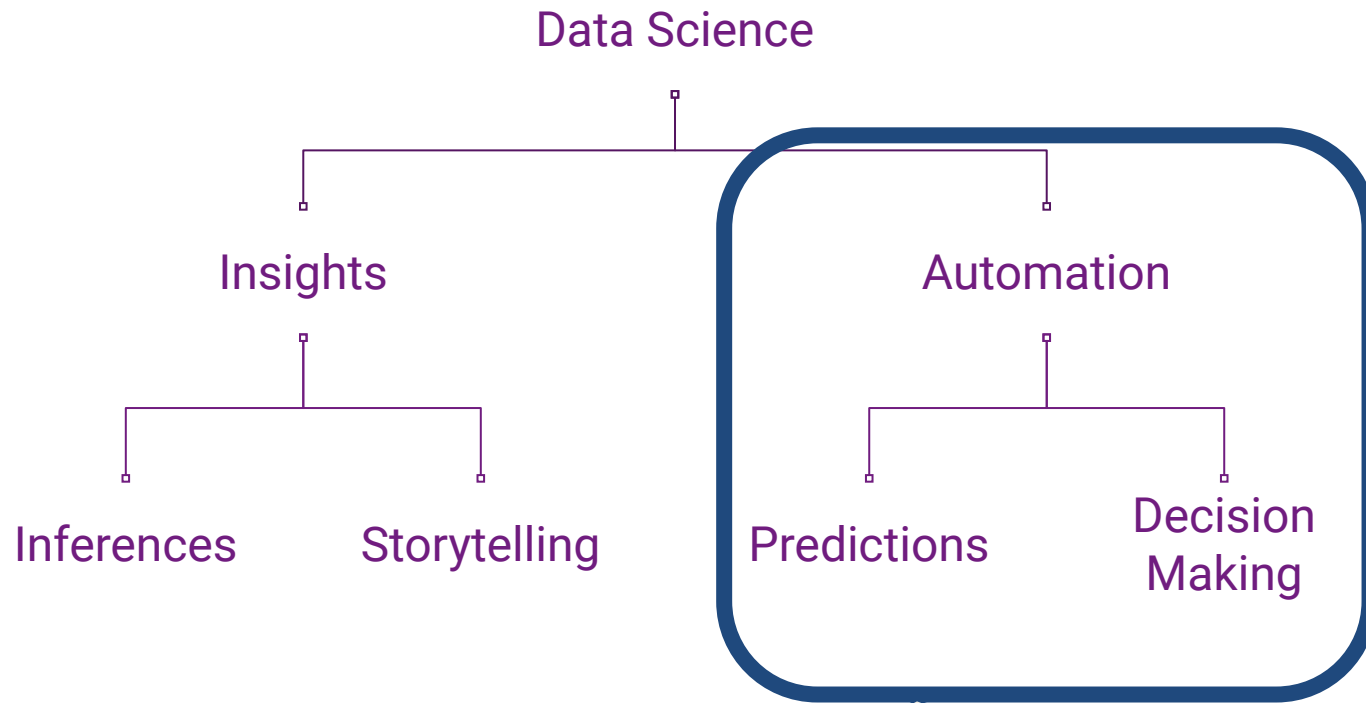


COMPONENTS OF AI / ML

INSIGHTS VS AUTOMATION

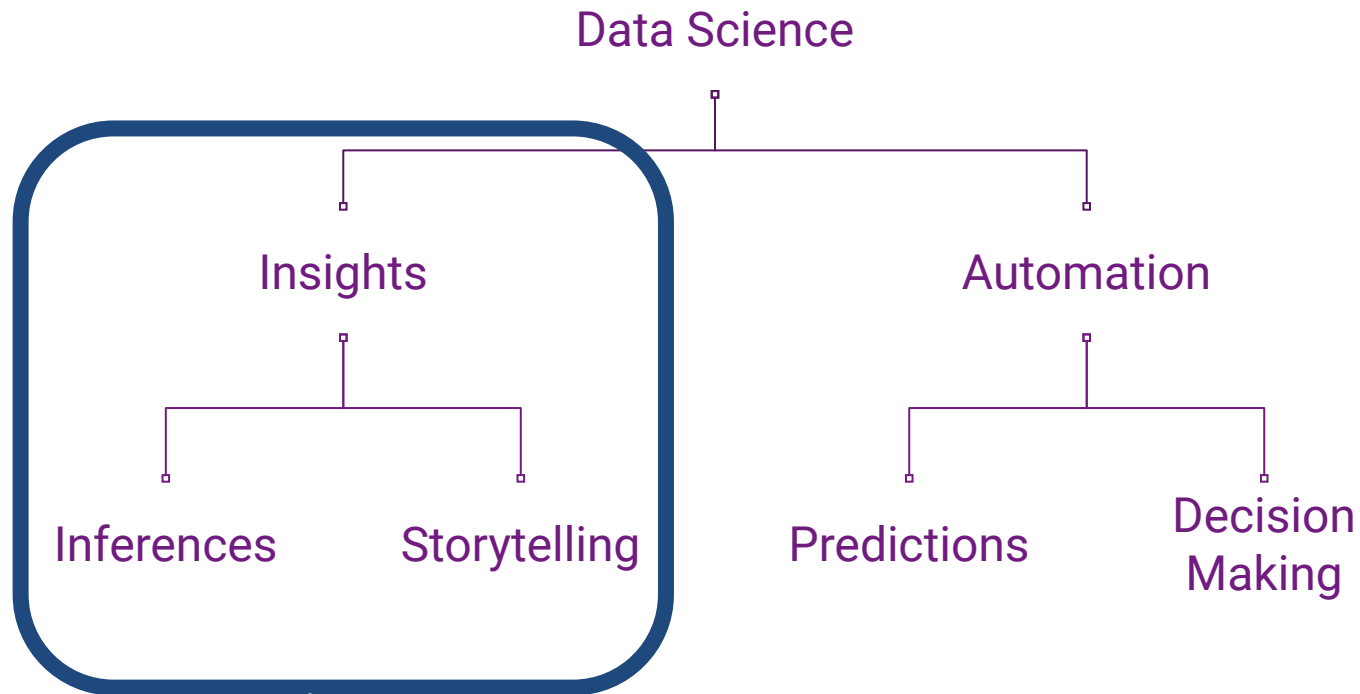


INSIGHTS VS AUTOMATION



- Science > Art
- Lots of machine learning, engineering and infrastructure
- Fairly structured problems – lots of scope for automation

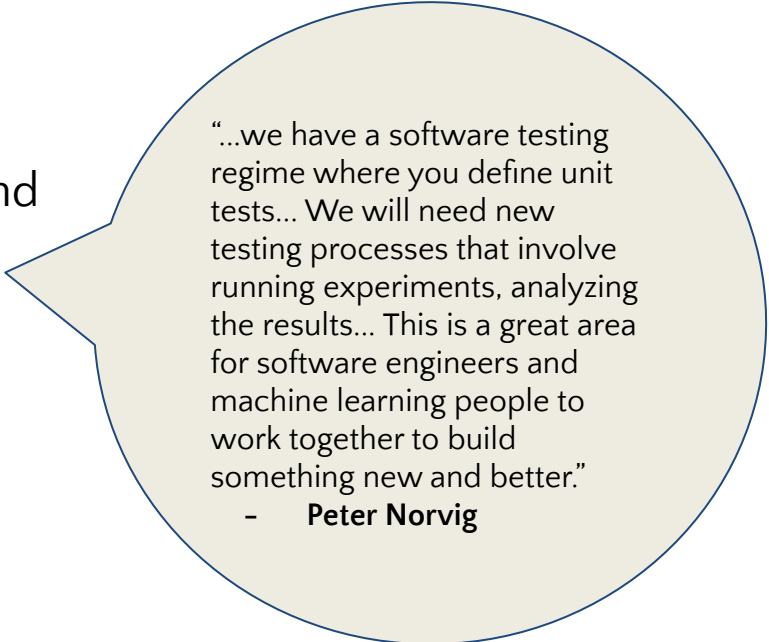
INSIGHTS VS AUTOMATION



- Art > Science
- Lots of visualization and statistics
- Mostly open-ended - lot of room for creativity
- **Distinguishing Factor*

TOP PROBLEMS IN DATA SCIENCE

- Explainability of machine learning:
 - Should it be a black box? Should it be explainable?
 - How do you trust your model?
- Better tooling:
 - bridging the gap between research and practice
 - how the “programming tax” can be reduced
- Awareness:
 - Does the problem need machine learning?
 - Does it need deep learning?
 - Does it need big data solutions?
 - Hype vs Reality: “Worrying about the AI apocalypse is like worrying about overpopulating Mars.” – **Andrew Ng**



“...we have a software testing regime where you define unit tests... We will need new testing processes that involve running experiments, analyzing the results... This is a great area for software engineers and machine learning people to work together to build something new and better.”

– **Peter Norvig**

THANK YOU!



/@jaidevd



deshpande.jaidev@gmail.com