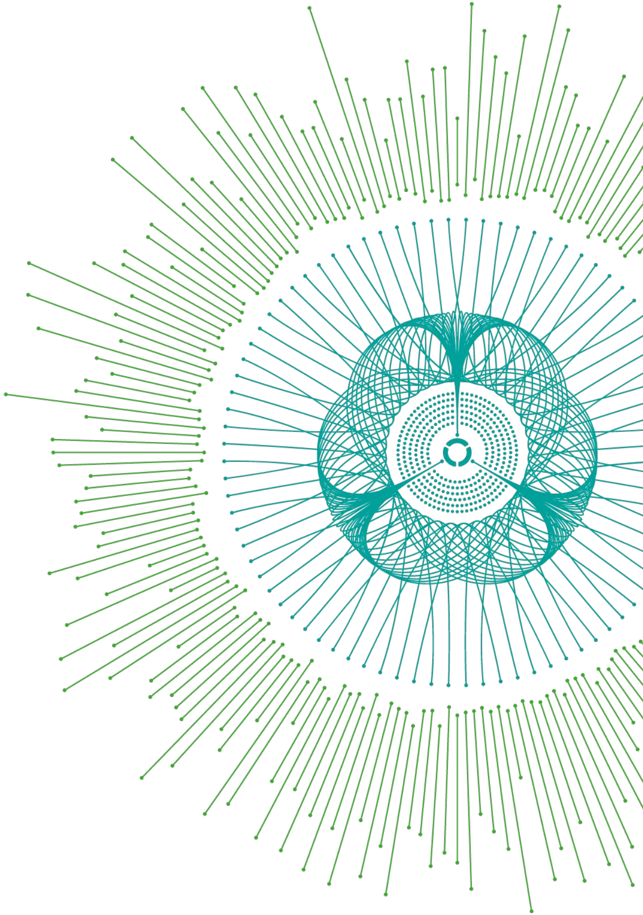
ANACONDA®

Enterprise Data Science Comes of Age

Peter Wang
CEO & Co-founder, Anaconda Inc.

March 23, 2022

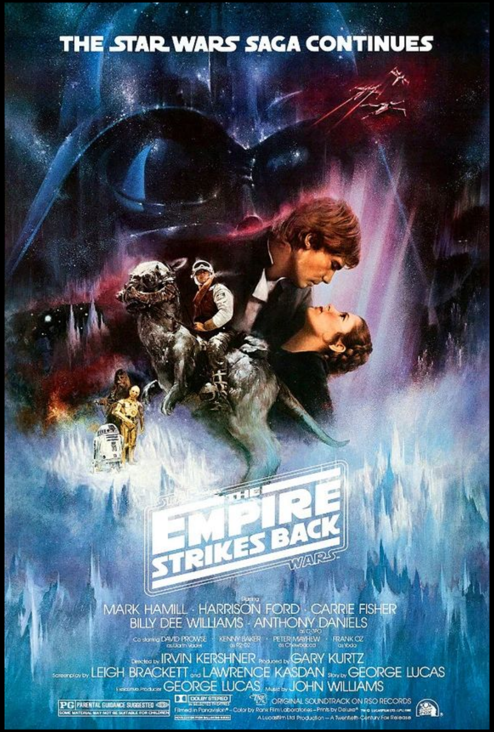


Topics

- Python adoption and growth
- Enterprise adoption of data science & ML
- MLOps beyond the data warehouse
- Open source innovation and commercialization
- Democratizing the AI future

TWENTIETH CENTURY-FOX Presents A LUCASFILM LTD PRODUCTION STAR WARS
Starring MARK HAMILL HARRISON FORD CARRIE FISHER
PETER CUSHING
and
ALEC GUINNESS
Written and Directed by GEORGE LUCAS
Produced by GARY KURTZ
Music by JOHN WILLIAMS
PANAVISION® PRINTS BY DE LUXE® TECHNICOLOR®
DOLBY SYSTEM
STAR WARS

"A New Language"


THE STAR WARS SAGA CONTINUES
STAR WARS
THE EMPIRE STRIKES BACK
MARK HAMILL · HARRISON FORD · CARRIE FISHER
BILLY DEE WILLIAMS · ANTHONY DANIELS
Directed by IRVIN KERSHNER
Produced by GARY KURTZ
Screenplay by LEIGH BRACKETT and LAWRENCE KASDAN
Story by GEORGE LUCAS
Executive Producer GEORGE LUCAS
Music by JOHN WILLIAMS
ORIGINAL SOUNDTRACK ON RSO RECORDS
A Lucasfilm Ltd Production – A Twentieth Century Fox Release

"The Enterprise Strikes Back"

STAR WARS
RETURN OF THE JEDI
MARK HAMILL · HARRISON FORD · CARRIE FISHER
BILLY DEE WILLIAMS · ANTHONY DANIELS
Directed by RICHARD MARQUAND
Produced by HOWARD KAZANJIAN
Story by GEORGE LUCAS
Screenplay by LAWRENCE KASDAN and GEORGE LUCAS
Executive Producer GEORGE LUCAS
Music by JOHN WILLIAMS

"The Return of Cybernetics"

Episode IV

A NEW LANGUAGE

It is a period of disruption.

Big Data, arising from adoption of cloud and mobile, has shaken the stranglehold of SQL and Java.

Rebel scientists have managed to create a new language, powerful and easy to learn....

How Did Python Take Over?

My Thesis in 2012

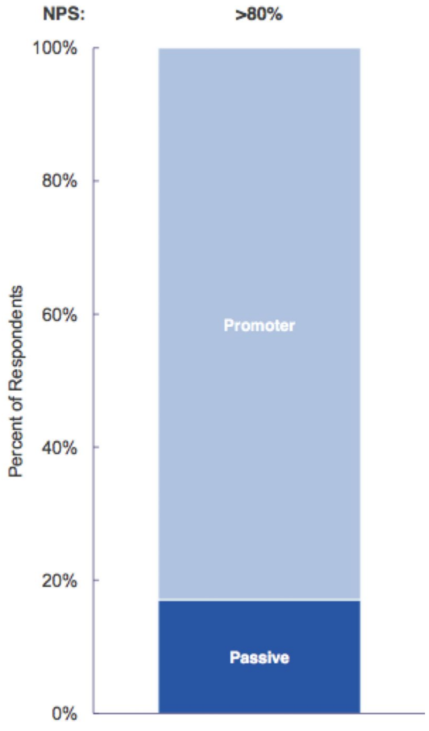
“There is a unique double-disruption in business and technical computing.”

Big Data is disrupting traditional data warehousing and business intelligence;
Predictive analytics and machine learning are green-field areas for everyone.

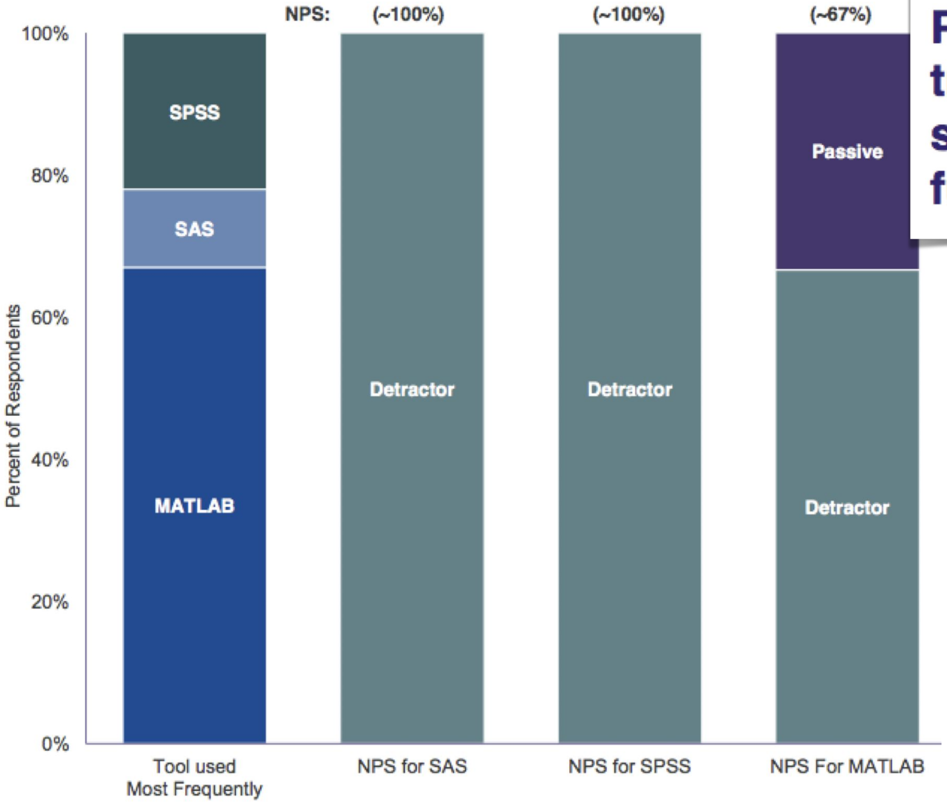
Cloud Computing and new vector hardware (GPUs) give businesses access to 100x or 1000x their previous maximum analytical capability, on an hourly rental basis, with zero capital investment.

Incumbent Tools Were... Not Loved

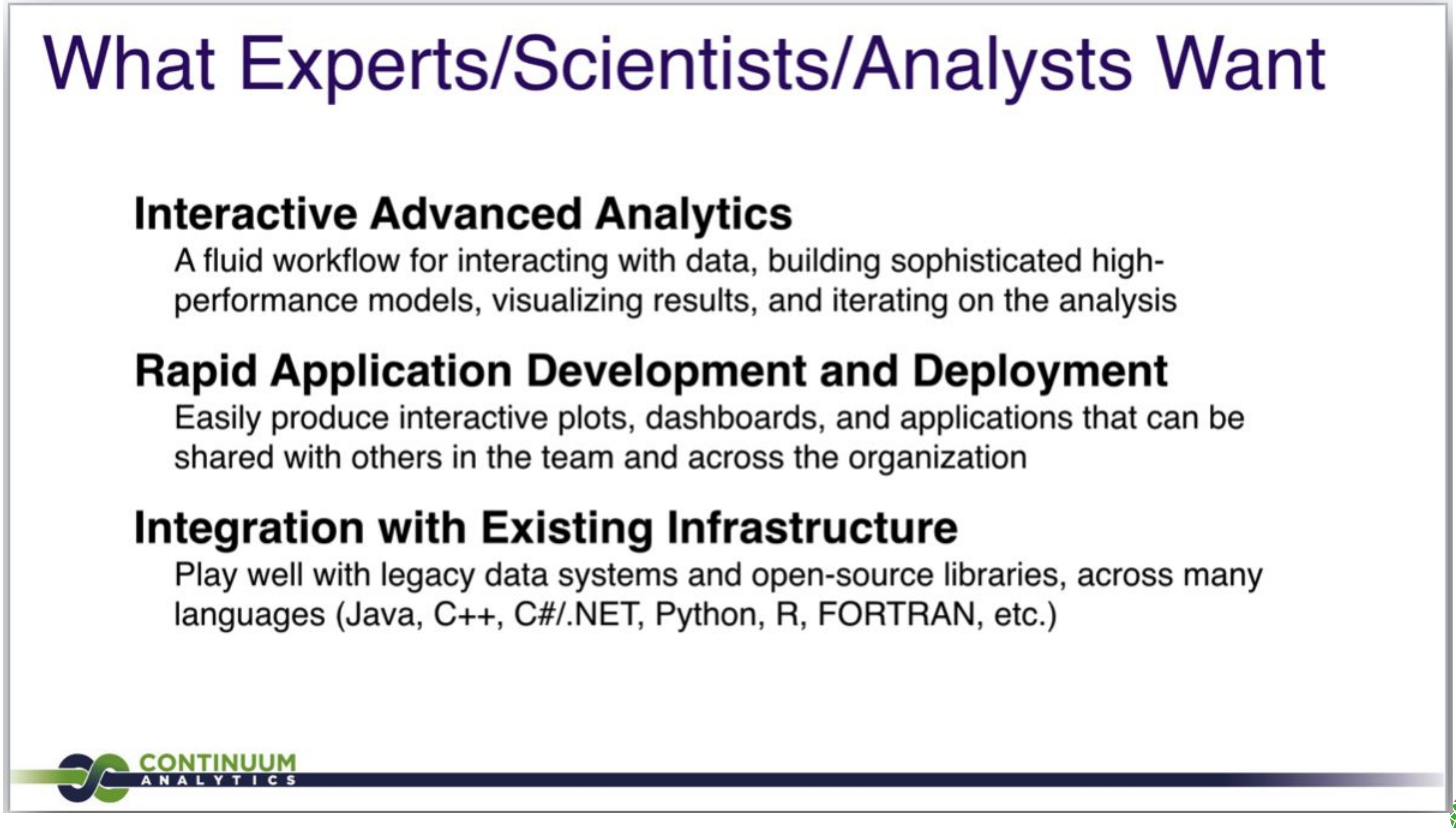
Python NPS Rating



Use of Incumbent Solutions



End users report that Python-based tools trump incumbent solutions across all functionality criteria.



What Experts/Scientists/Analysts Want

Interactive Advanced Analytics

A fluid workflow for interacting with data, building sophisticated high-performance models, visualizing results, and iterating on the analysis

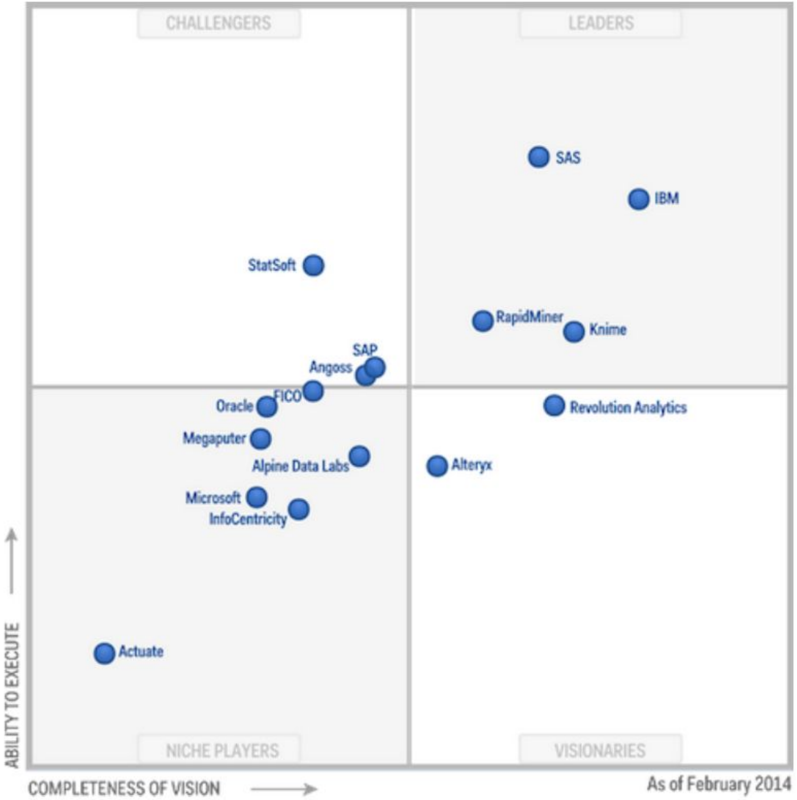
Rapid Application Development and Deployment

Easily produce interactive plots, dashboards, and applications that can be shared with others in the team and across the organization

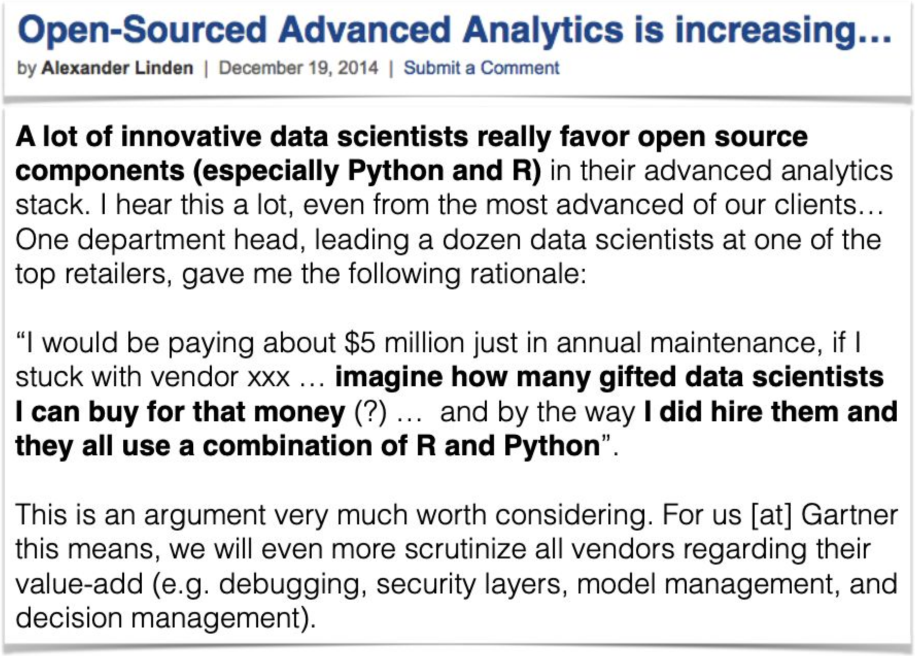
Integration with Existing Infrastructure

Play well with legacy data systems and open-source libraries, across many languages (Java, C++, C#/.NET, Python, R, FORTRAN, etc.)

Gartner Recognized This in 2014…



As of February 2014



Open-Sourced Advanced Analytics is increasing…

by **Alexander Linden** | December 19, 2014 | Submit a Comment

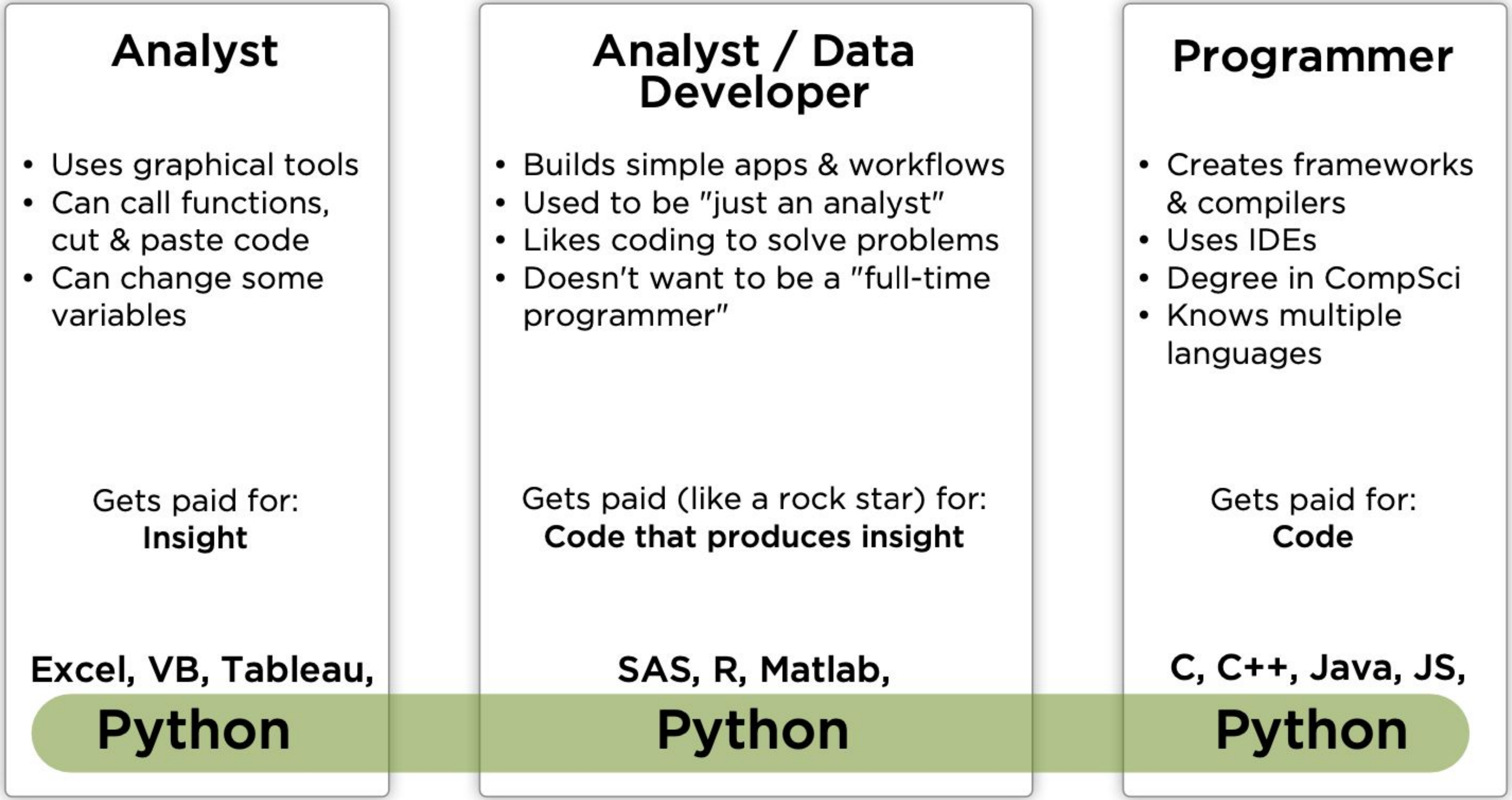
A lot of innovative data scientists really favor open source components (especially Python and R) in their advanced analytics stack. I hear this a lot, even from the most advanced of our clients… One department head, leading a dozen data scientists at one of the top retailers, gave me the following rationale:

"I would be paying about $5 million just in annual maintenance, if I stuck with vendor xxx … **imagine how many gifted data scientists I can buy for that money** (?) … and by the way **I did hire them and they all use a combination of R and Python**".

This is an argument very much worth considering. For us [at] Gartner this means, we will even more scrutinize all vendors regarding their value-add (e.g. debugging, security layers, model management, and decision management).

Alexander Linden
Research Director
2 years at Gartner
29 years IT Industry

Python’s Unique Position



Analyst

- Uses graphical tools
- Can call functions, cut & paste code
- Can change some variables

Gets paid for:
Insight

Excel, VB, Tableau,

Python

Analyst / Data Developer

- Builds simple apps & workflows
- Used to be "just an analyst"
- Likes coding to solve problems
- Doesn't want to be a "full-time programmer"

Gets paid (like a rock star) for:
Code that produces insight

SAS, R, Matlab,

Python

Programmer

- Creates frameworks & compilers
- Uses IDEs
- Degree in CompSci
- Knows multiple languages

Gets paid for:
Code

C, C++, Java, JS,

Python

Why Python?

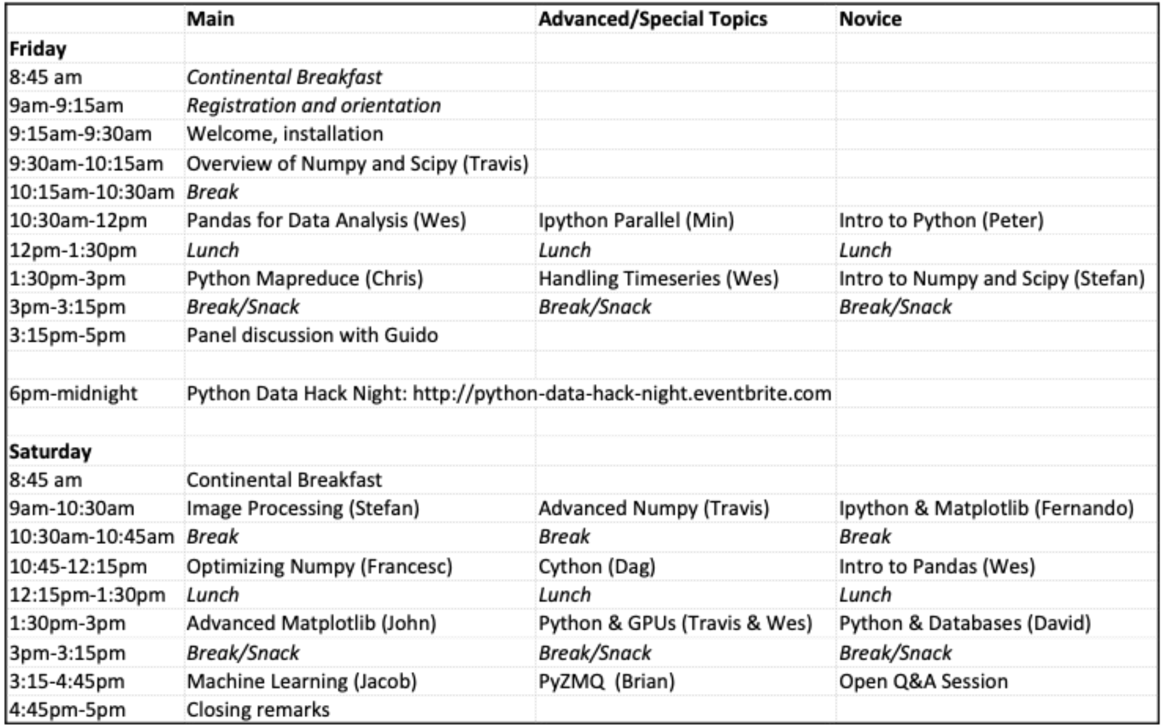
- Easy to learn ("Time to first plot")
- Rich library ecosystem ("Batteries included")
- Designed for readability ("Fits in your head")
- Communities are generally fairly pleasant

Why Python?

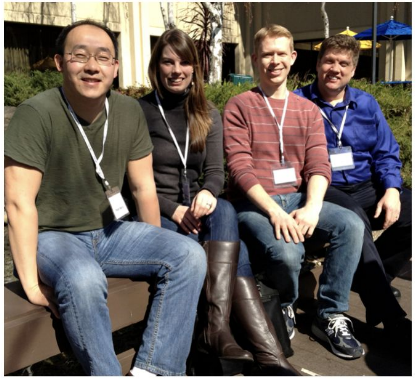
- Easy to learn ("Time to first plot")
- Rich library ecosystem ("Batteries included")
- Designed for readability ("Fits in your head")
- Communities are generally fairly pleasant
- CPython VM has a low-level C API that allowed scientific programmers to write a huge number of high-performance bindings to C++ and FORTRAN.

  ("Honda Civic with mounting bolts for warp drive.")

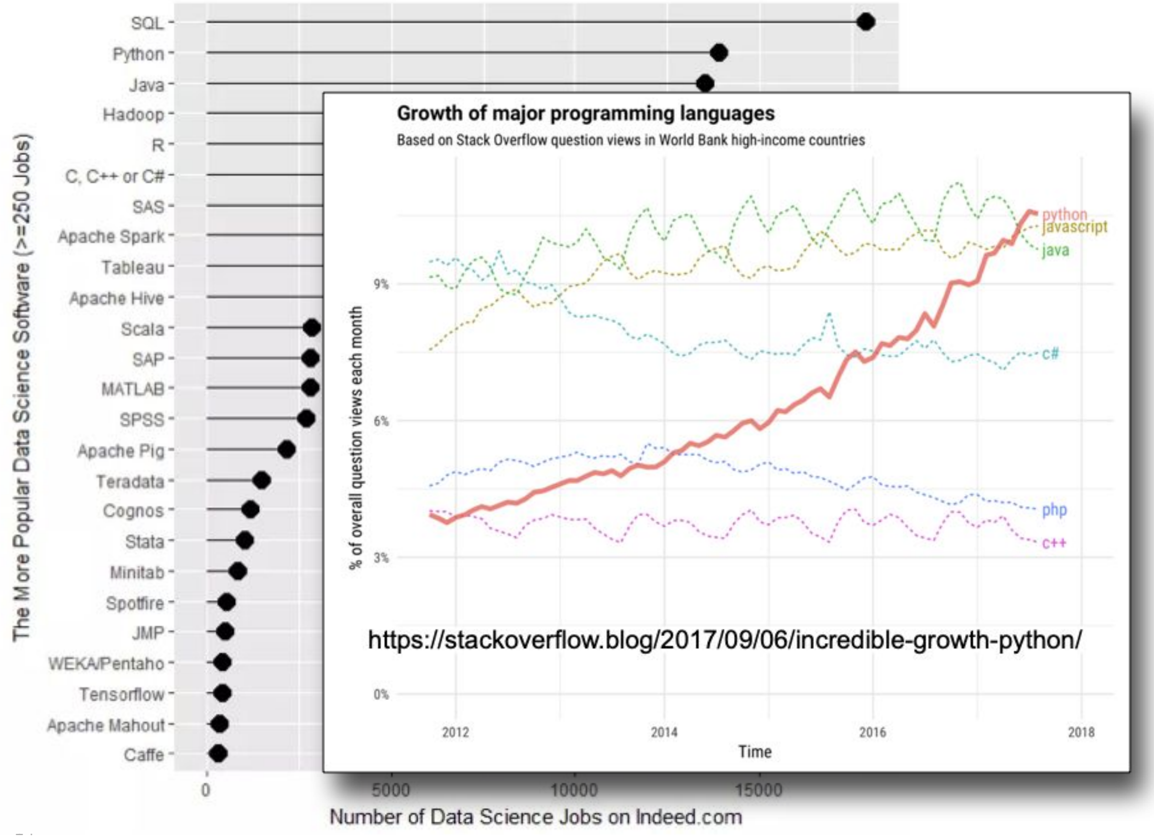
10 Years Ago… First PyData Meetup, March 2012



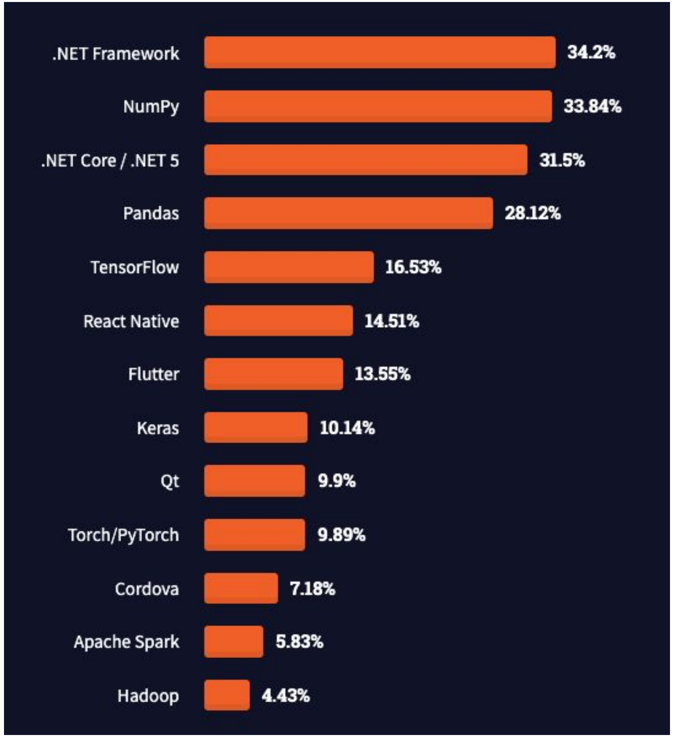
	Main	Advanced/Special Topics	Novice
Friday			
8:45 am	*Continental Breakfast*		
9am-9:15am	*Registration and orientation*		
9:15am-9:30am	Welcome, installation		
9:30am-10:15am	Overview of Numpy and Scipy (Travis)		
10:15am-10:30am	*Break*		
10:30am-12pm	Pandas for Data Analysis (Wes)	Ipython Parallel (Min)	Intro to Python (Peter)
12pm-1:30pm	*Lunch*	*Lunch*	*Lunch*
1:30pm-3pm	Python Mapreduce (Chris)	Handling Timeseries (Wes)	Intro to Numpy and Scipy (Stefan)
3pm-3:15pm	*Break/Snack*	*Break/Snack*	*Break/Snack*
3:15pm-5pm	Panel discussion with Guido		
6pm-midnight	Python Data Hack Night: http://python-data-hack-night.eventbrite.com		
Saturday			
8:45 am	Continental Breakfast		
9am-10:30am	Image Processing (Stefan)	Advanced Numpy (Travis)	Ipython & Matplotlib (Fernando)
10:30am-10:45am	*Break*	*Break*	*Break*
10:45-12:15pm	Optimizing Numpy (Francesc)	Cython (Dag)	Intro to Pandas (Wes)
12:15pm-1:30pm	*Lunch*	*Lunch*	*Lunch*
1:30pm-3pm	Advanced Matplotlib (John)	Python & GPUs (Travis & Wes)	Python & Databases (David)
3pm-3:15pm	*Break/Snack*	*Break/Snack*	*Break/Snack*
3:15-4:45pm	Machine Learning (Jacob)	PyZMQ (Brian)	Open Q&A Session
4:45pm-5pm	Closing remarks		

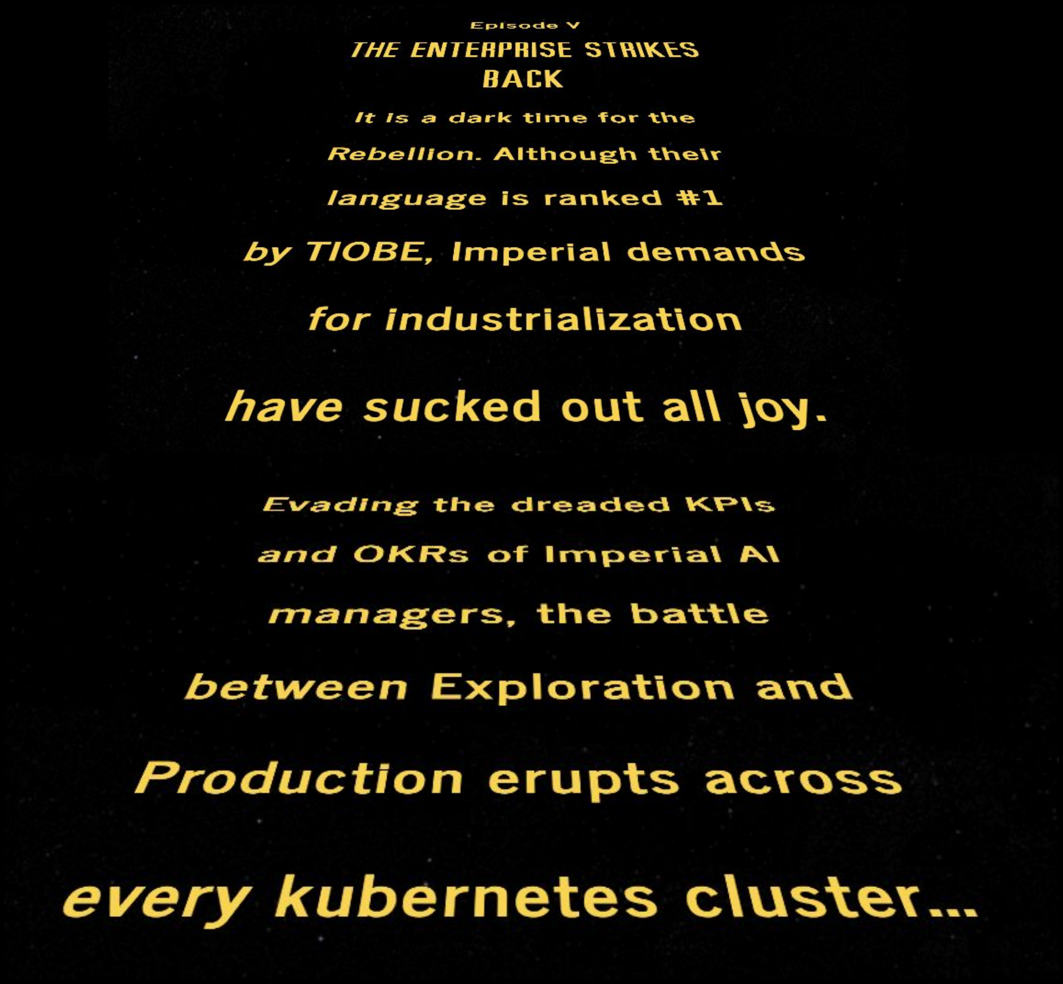


Python Led the Way for the Growth of Data Science



https://stackoverflow.blog/2017/09/06/incredible-growth-python/





Episode V

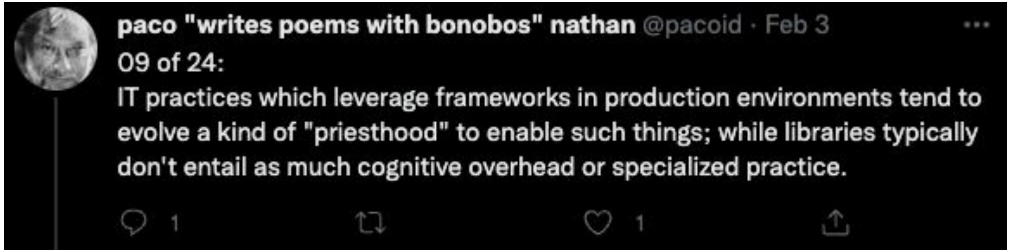
THE ENTERPRISE STRIKES BACK

It is a dark time for the Rebellion. Although their language is ranked #1 by TIOBE, Imperial demands for industrialization have sucked out all joy.

Evading the dreaded KPIs and OKRs of Imperial AI managers, the battle between Exploration and Production erupts across every kubernetes cluster...

“Whither Python in the Era of MLOps?”

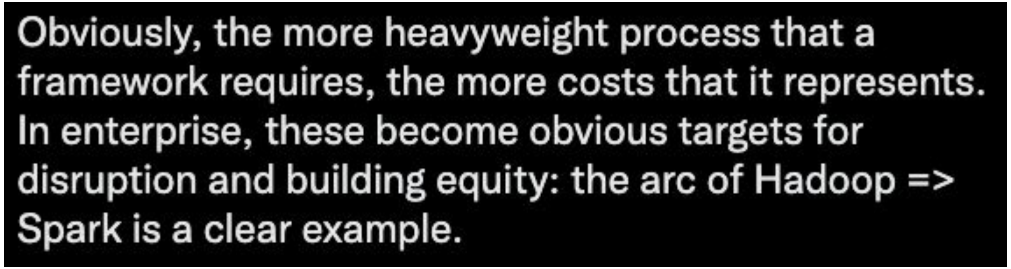
Paco Nathan on Kubernetes…



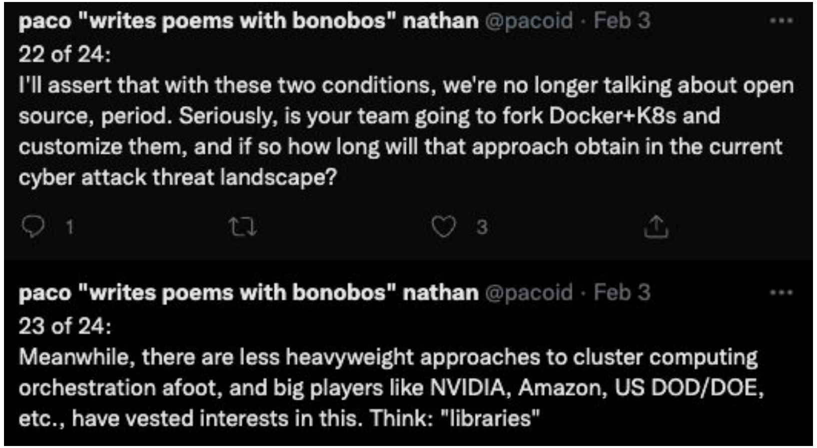
paco "writes poems with bonobos" nathan @pacoid · Feb 3

09 of 24:
IT practices which leverage frameworks in production environments tend to evolve a kind of "priesthood" to enable such things; while libraries typically don't entail as much cognitive overhead or specialized practice.

1 1



Obviously, the more heavyweight process that a framework requires, the more costs that it represents. In enterprise, these become obvious targets for disruption and building equity: the arc of Hadoop => Spark is a clear example.



paco "writes poems with bonobos" nathan @pacoid · Feb 3

22 of 24:
I'll assert that with these two conditions, we're no longer talking about open source, period. Seriously, is your team going to fork Docker+K8s and customize them, and if so how long will that approach obtain in the current cyber attack threat landscape?

1 3

paco "writes poems with bonobos" nathan @pacoid · Feb 3

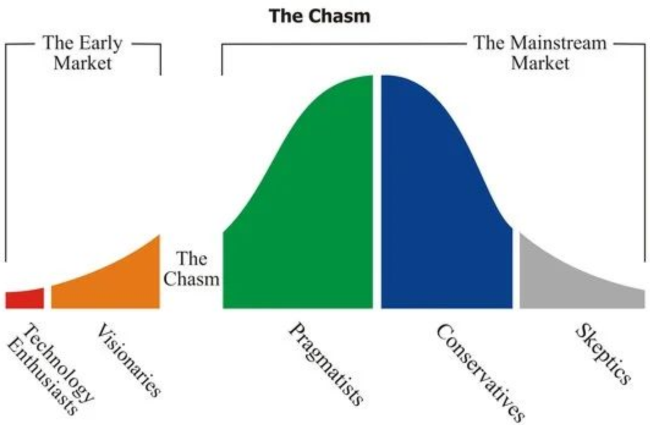
23 of 24:
Meanwhile, there are less heavyweight approaches to cluster computing orchestration afoot, and big players like NVIDIA, Amazon, US DOD/DOE, etc., have vested interests in this. Think: "libraries"

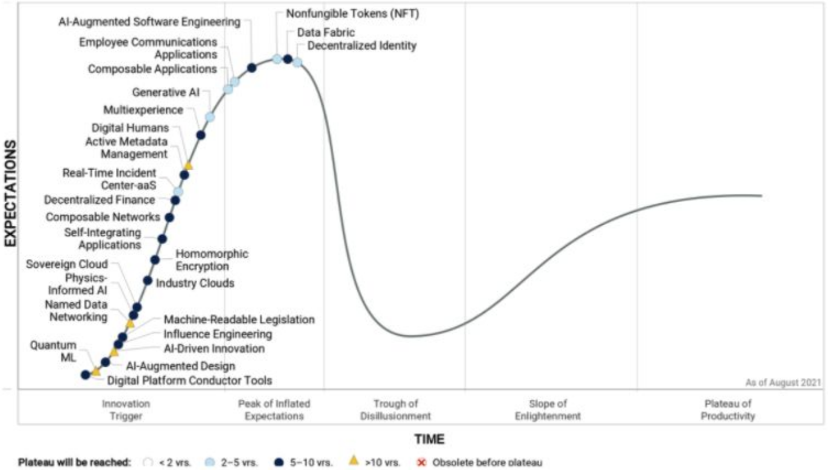
How Do Enterprises Adopt Tech?

- Less about the tech, and more about the structure of Enterprise IT orgs!
- Classic "crossing the chasm" dynamics in tech adoption curve
- Many IT orgs *can afford to be perpetually ineffective*.
  - There are always myriad risks and scapegoats for why innovation doesn't happen
  - It's OK to encumber the business if everyone else in your sector also sucks

“Every morning in the savanna, a herd of antelopes wakes up, knowing that no matter how slow they run, if the company doesn’t miss its quarterly earnings, the VP lions won’t reduce their budget.”

The Chasm and the Hype Cycle



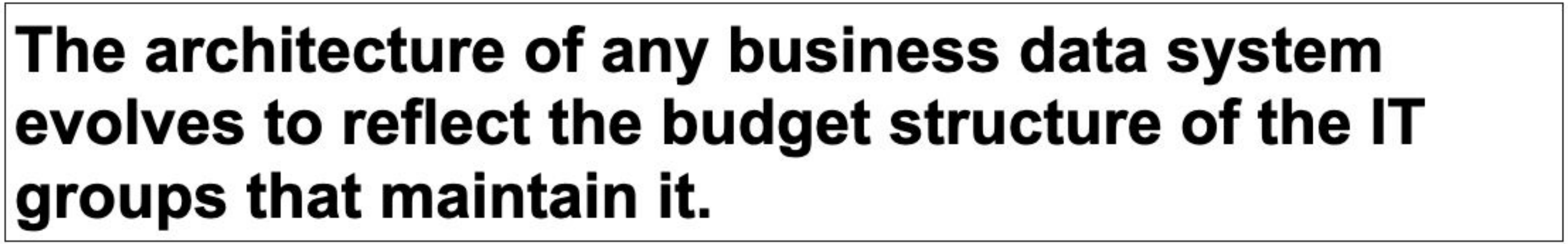


If the growth rate of ignorance/confusion exceeds the growth of confident information, then the sense-making gap will fill with BS.

The data/ML/AI space will continue to outpace the sensemaking of adopters.

This is a semi-permanent problem.

Peter’s Corollary to Conway’s Law



The architecture of any business data system evolves to reflect the budget structure of the IT groups that maintain it.

… not strategic or operational needs

… not ensuring future analytical agility

… not optimizing for rapid insights

	Exploration	Production
Data	● Fast, unfettered access ● Ease of introducing new, varied, messy datasets ● Reproducibility	● Strict, governed access ● Well-defined schema ● Provenance & auditability
Compute Infrastructure	● High performance ● Low latency, interactive ● Individualized & specialized	● Scalable, high-availability ● Manageable at scale ● Cost amortization over many machines and users
Organization	● Individual high-achievers with lots of context & capability ● Agile, able to quickly learn new skills and approaches	● Sustain operations at lowest possible cost ● Robustness against unintended change

Better Framing: Pioneer vs Settler

	Settler	Pioneer
Objectives	Minimize risk given known threats and failure modes, while maintaining operations within set bounds.	Maximize opportunity, learning, innovation for given limits of time/resources
Scaling	Operational efficiency of repeatable processes	Individual Capability: speed, breadth of context, flow, creativity
Governing Philosophy	Improve baseline (bias) by managing “known knowns”	Exploring high-variance “unknown unknowns”

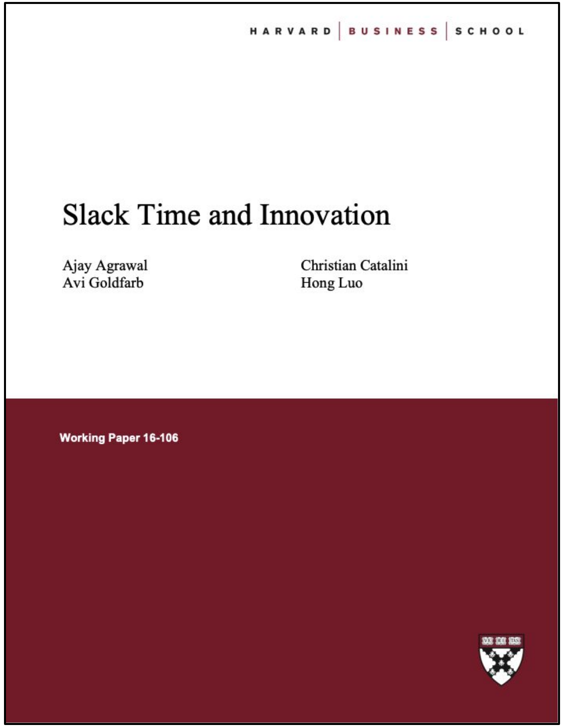
The Heart of the Issue: “Scale” of what, exactly?

Original sin of Enterprise Software Development: Programmers are fungible labor. Scale via headcount.

Alternative Thought Experiment: How far can one *single* person/pioneer get?

- How fast to answer a query?
- How sophisticated of a business question?
- How accurately can they make a prediction?

Against Efficiency



HARVARD | BUSINESS | SCHOOL

Slack Time and Innovation

Ajay Agrawal
Avi Goldfarb

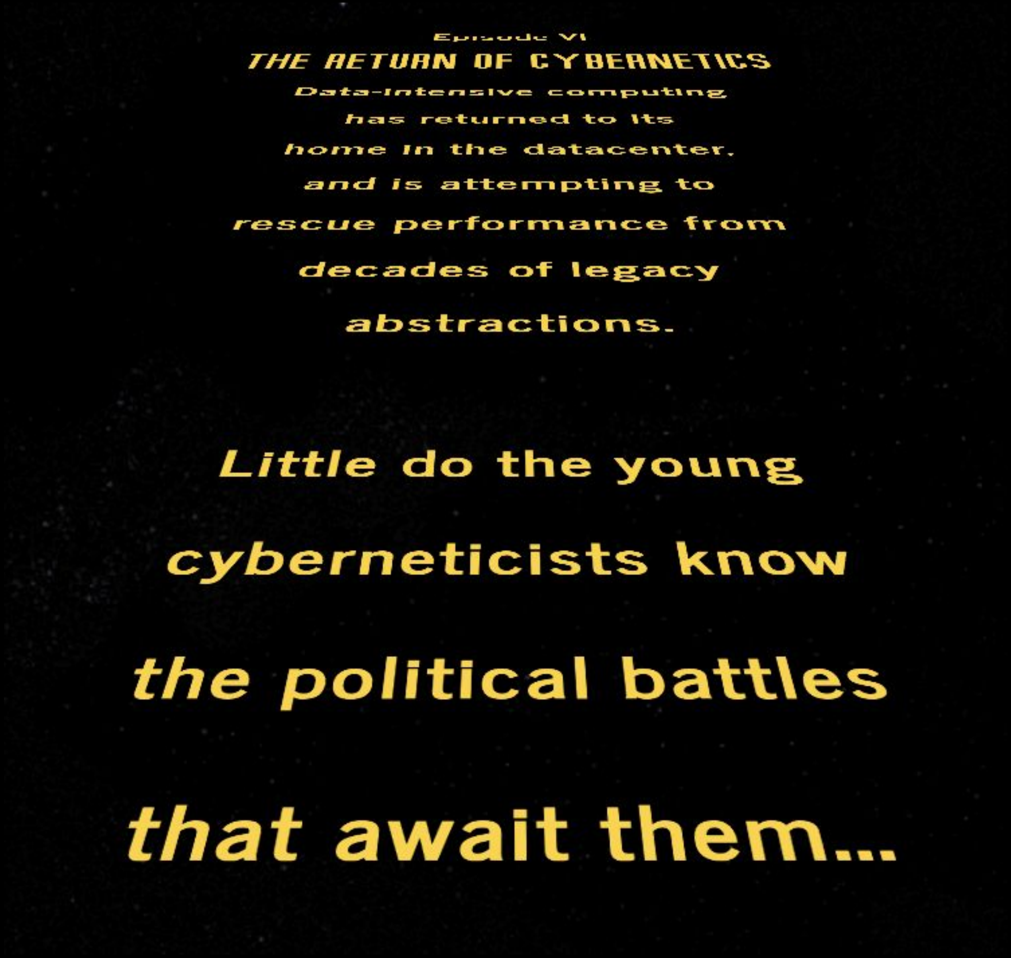
Christian Catalini
Hong Luo

Working Paper 16-106

Operational scaling in mature orgs always minimizes slack *in situ*. (There may be budget elsewhere for “innovation”.)

All creative growth requires “slack” to explore the adjacent possible.

Principle of Subsidiarity should be respected: intelligence lives in the leaf nodes.

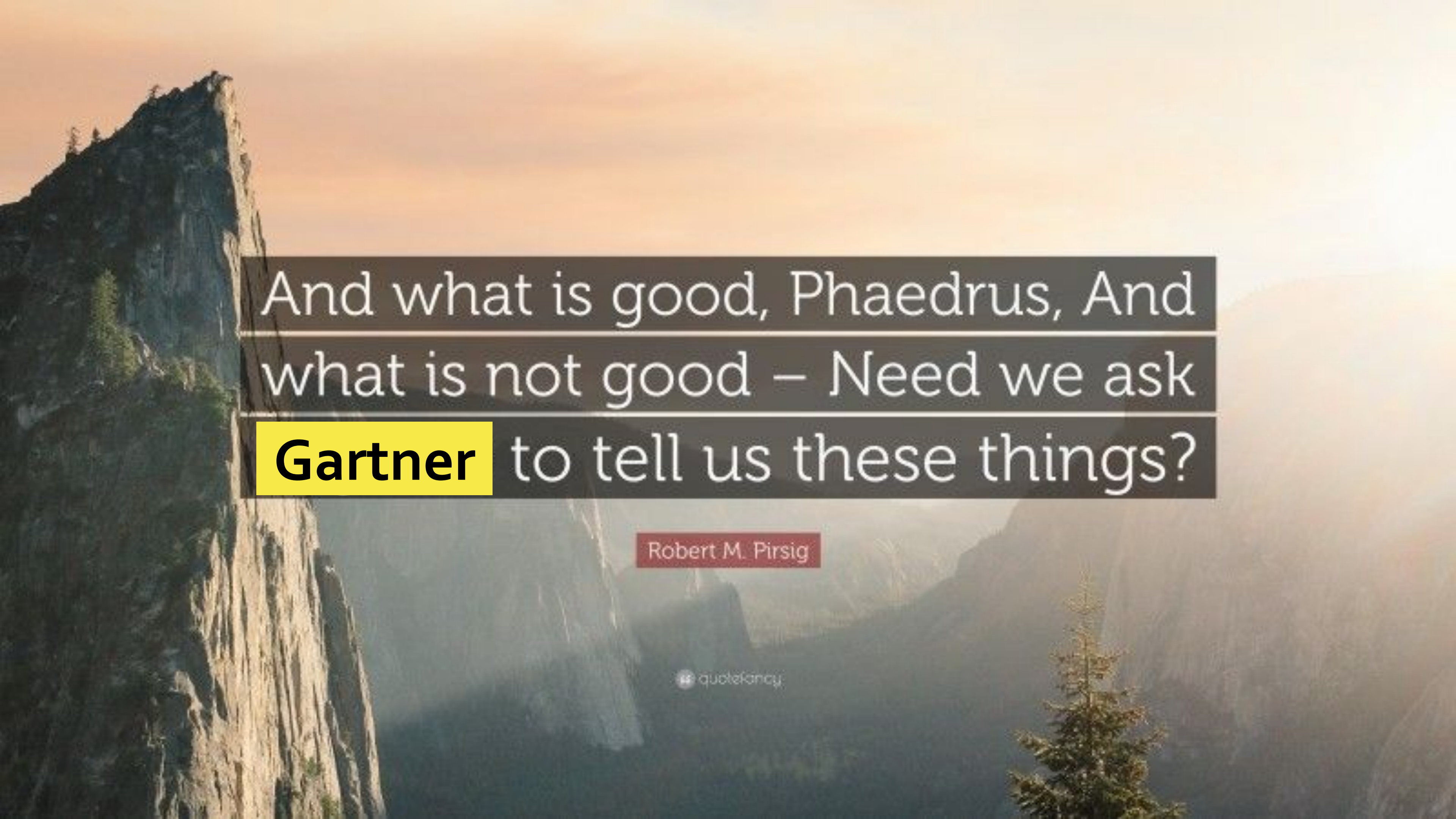


Episode VI

THE RETURN OF CYBERNETICS

Data-intensive computing has returned to its home in the datacenter, and is attempting to rescue performance from decades of legacy abstractions.

Little do the young cyberneticists know the political battles that await them...



And what is good, Phaedrus, And what is not good – Need we ask **Gartner** to tell us these things?

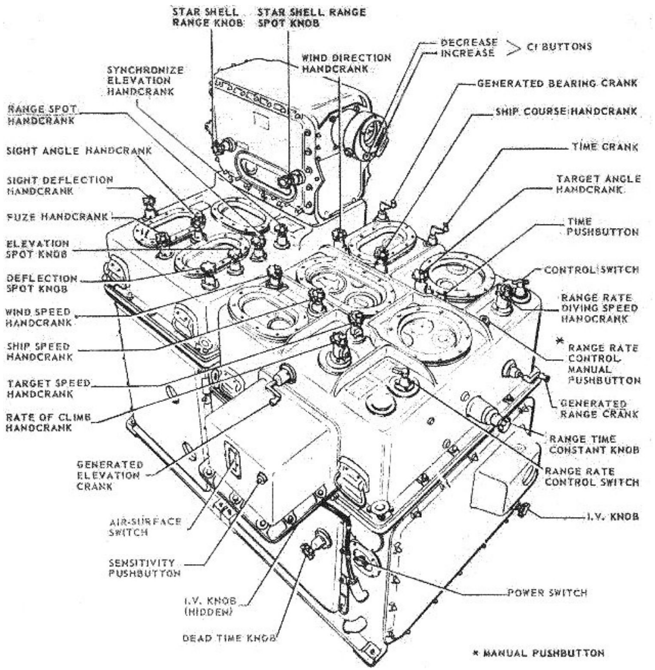
Robert M. Pirsig

The Perpetual Chasm

- Data scientists still continue to struggle to get IT to meet their needs
- C-suite continues to be bombarded with hype about AI & ML
- Data, software development, infrastructure are all confused about the frothy tech landscape. ("Influx of VC cash from quantitative easing has led to qualitative strife.")

We must educate the practitioner community, and facilitate peer learning amongst all stakeholders!

Back to the Future



Computing started with Prediction.

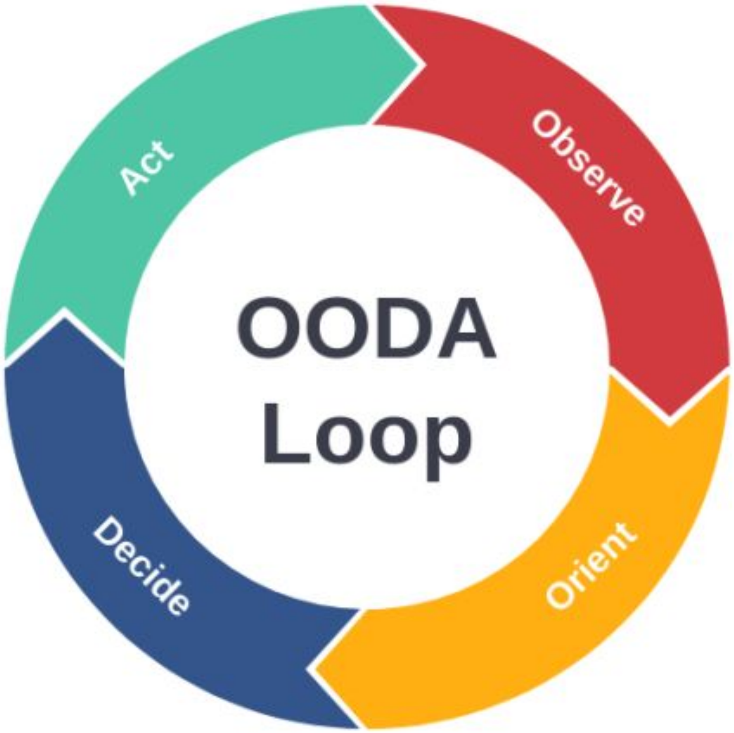
Post-WW2, “just counting things” came to dominate.

In PC era, transactional systems dominated. Prediction became niche (except in certain industries)

With explosion of social/mobile data, and 100X processors, we are now back to prediction.

In a world of connected sensors & actuators, this is now Cybernetics.

“Computational Decisioning” Is the Future



Observe

What is the current situation? What is the reason you want to change? how bad do you want to change?

Orient

Where are you currently at relative to where you want to go? How far is it to your destination?

Decide

What is the exact path you are going to take? How are you going to handle challenges and set backs?

Act

What's the approach and method you will take to implement the decisions? What is your action plan?

Cybernetic Systems Require a Holistic Approach

- Correctness is multi-dimensional and transcends any individual component:
    - Performance (latency and throughput) is an integral part of correctness
    - Agility is an integral part of correctness, to maximize data scientist productivity
- Cannot alienate data management, from code development, from infrastructure provisioning and planning.
- Change management has to track Code, Data, and Semantics. 😳

Zen of Cybernetics

Know Thy Data

There is no data, only frozen models.

Move code to data; move context into code.

Information is a verb. It is the dance, not the dancer.

Zen of Cybernetics

Know Thy Data

There is no data, only frozen models.

Move code to data; move context into code.

Information is a verb. It is the dance, not the dancer.

Know Thy Computer

Understand the nominal performance of your system.

Get a bigger computer. there’s no extra credit for using a distributed system.

Focus on bandwidth, not software abstractions.

Zen of Cybernetics

Know Thy Data

There is no data, only frozen models.

Move code to data; move context into code.

Information is a verb. It is the dance, not the dancer.

Know Thy Computer

Understand the nominal performance of your system.

Get a bigger computer. there’s no extra credit for using a distributed system.

Focus on bandwidth, not software abstractions.

Know Thy Org

There is one veritable problem: human relations.

Smart people don’t scale. Some things require smart people.

Creativity and innovation require slack in the system.

JOIN US!

Wednesday, March 23
7:30 - 9:30 PM CDT
KUNG FU SALOON,
716 W 6th St, Austin

BREWS & BBQ

snowflake®

ANACONDA®