Jupyter Trends in 2018

Paco Nathan @pacoid
Jupyter provides a rich set of extensible, re-usable building blocks, expressed through various open protocols, APIs, and standards.

These get combine for a wide variety of use cases, as extensible software architecture for interactive computing with data.

Over the past year since JupyterCon 2017, we’ve noted three distinct trends emerging ➔
We’ve seen large organizations adopt Jupyter for their analytics infrastructure, in a “leap frog” effect over commercial offerings.

Many people hired out of universities already know how to write ML apps in Jupyter – and those without coding backgrounds can learn rapidly via Jupyter.

Why spend money re-training your staff to use proprietary frameworks when there are more effective means available?
An **emerging trend** disrupts the past 15-20 years of software engineering practice:

**hardware > software > process**

Hardware is now evolving more rapidly than software, which is evolving more rapidly than effective process. Jupyter helps “future proof” efforts during this period of chaos / rapid evolution.

BTW, that dovetails quite nicely with cloud services.
A recent interview with Andrew Feldman, founder/CEO of Cerebras Systems, gives a good overview of the blossoming area of **specialized hardware** for machine learning, edge computing, decentralization, etc.:

https://www.oreilly.com/ideas/specialized-hardware-for-deep-learning-will-unleash-innovation
As we see enterprise, government, universities, etc., roll out interactive computing at scale, the *organizational challenges* arise next:

Practices regarding collaboration, **data privacy**, ethics, security, compliance, etc.

Jupyter addresses critical needs – which **Silicon Valley** hadn’t previously focused on enough.

Watch within the **highly regulated environments**, where that rapid evolution in open source is happening.
O’Reilly did a recent study about ML adoption in enterprise, with 8000+ respondents worldwide, which provides relevant insights: [https://www.oreilly.com/ideas/5-findings-from-oreilly-machine-learning-adoption-survey-companies-should-know](https://www.oreilly.com/ideas/5-findings-from-oreilly-machine-learning-adoption-survey-companies-should-know)
an even larger **challenge** looms:

We’re here now, **29 years** after Tim Berners-Lee created WWW – **55 years** after Ted Nelson invented hypertext – **73+ years** after Vannevar Bush (and Jorge Luis Borges) first described it.

Online media expands, while the business of *print media* has all but tanked.

Science, given its “publish or perish” onus, has become a vast and scattered library of “**digital paper**” –

*all neatly indexed by keyword search and wiki entries…*
an even larger challenge looms:

We’re here now, **29 years** after **Tim Berners-Lee** created WWW – **55 years** after **Ted Nelson** invented hypertext – **73+ years** after **Vannevar Bush** (and **Jorge Luis Borges**) first described it.

Online media expands, while the business of print media has all but tanked. Science, given its “publish or perish” onus, has become a vast and scattered library of “digital paper” – all neatly indexed by keyword search and wiki entries…

**except when it isn’t**

---

**WIRED**

**Using Artificial Intelligence to Fix Wikipedia’s Gender Problem**

Adelson was among thousands of names flagged by Quicksilver, a software tool by San Francisco startup Primer designed to help Wikipedia editors fill in blind spots in the crowdsourced encyclopedia. Its underrepresentation of women in science is a particular target. The world’s fifth-most-visited website has a long-running problem with gender bias: Only 18 percent of its biographies are of women. Surveys estimate that between 84 and 90 percent of Wikipedia editors are male.
Those pioneers dreamt of entirely new ways for us to collaborate, to extend our shared understanding.

However, they hadn’t dreamt of trolling and harassment … **Russian bot swarms** … **climate science attacked** due to lack of reproducible papers … ML leveraged to **polarize public animosity** … cyberthreats holding **hospital IT for ransom** …

Plus other ways of befouling scientific advances, online media, etc. While we’re talking about **open source**, these are exploits – as attempts to undermine **open society**.
Karl Popper, however, warned about precisely that:

“non-reproducible single occurrences are of no significance to science”

as explored in *The Logic of Scientific Discovery* (1934) and later in *The Open Society and Its Enemies* (1945)
Karl Popper, however, warned about precisely that: “non-reproducible single occurrences are of no significance to science” as explored in *The Logic of Scientific Discovery* (1934) and later in *The Open Society and Its Enemies* (1945). If you have not studied the latter in detail, you should.
Check out **astrophysics research** applied to analyze and detect cyberthreats in media, e.g., work by Steve Kramer, et al.:

Eight decades later, we inherit a blend of what both Bush and Popper had scried from the rubble and ashes of WWII.

Reproducibility in science – and, importantly, the closely related aspect of falsifiability – become foremost concerns.

To wit, unmitigated power craves universal statements for its own whims; however, universal statements can be disproven by singular events.
Reproducible science has close analogues in other fields on which, as we find, an *open society* depends:

- **data science** – vital for any organization that depends on analytics, as the key to shared, accountable judgement
- **machine learning** – interpretation, verification, transparency, ethics
- **software engineering** – continuous integration (CI/CD), testability, security audits, reliability for critical infrastructure
- **teaching** – to help instructors manage the scaffolding needed to make course materials more engaging, immediately hands-on; to give learners confidence and direct experience
- **journalism** – how we demonstrate tangible, quantifiable evidence about what might otherwise be dismissed as ephemeral reports
Reproducible science has close analogues in other fields on which, as we find, an open society depends:

- **data science** – vital for any organization that depends on analytics, as the key to shared, accountable judgement
- **machine learning** – interpretation, verification, transparency, ethics
- **software engineering** – continuous integration (CI/CD), testability, security audits, reliability for critical infrastructure
- **teaching** – to help instructors manage the scaffolding needed to make course materials more engaging, immediately hands-on; to give learners confidence and direct experience
- **journalism** – how we demonstrate tangible, quantifiable evidence about what might otherwise be dismissed as ephemeral reports

Q: where else?
BTW, reproducible workflows in machine learning are notoriously difficult, due to a variety of reasons: e.g., the stochastic nature of training models, non-deterministic floating-point math on GPUs, etc.

A new category of tooling approaches reproducible ML workflows in innovative ways, including:

- **Biome** by Recognai
- **PEDL** by Determined AI
Meanwhile, there’s a compelling dynamic in which both reproducible science and open source are necessary for collaboration at scale.

Both disciplines have much to learn from each other.

Let’s work together to discover and articulate that part about “where else?”
Meanwhile, there’s a compelling dynamic in which both reproducible science and open source are necessary for collaboration at scale. Both disciplines have much to learn from each other. Let’s work together to discover and articulate that part about “what else?”

Ultimately, much of our program at JupyterCon 2018 is about what these disciplines collected here now must learn from each other.
Thank you.
publications, interviews, conference summaries...

https://derwen.ai/paco
@pacoid