AI within O’Reilly Media: How does Media make use?

Paco Nathan @pacoid
Director, Learning Group @ O’Reilly Media
2017-06-29
Personal route: +35 years, so far

Design process after Tim Brennan (~1990)

At an off-site for Apple Computer's Creative Services department, Tim Brennan began a presentation of his group's work by showing this model. "Here's how we work," he said. "Somebody calls up with a project; we do some stuff; and the money follows."

Brennan captures important aspects of the process:
- the potential for play
- its similarity to a "random walk"
- the importance of iteration
- its irreducible "black-box" nature

kudos: Dubberly Design Office
Personal route: +35 years, so far

- AI (then) @ Stanford, NASA, Bell Labs (1982-88)
- later: ML for custom search, behavioral targeting, recommender systems, network security, etc.
- led Data teams in industry:
  - one of the early “100% Cloud” architectures
  - first large-scale Hadoop use case on AWS
  - circa 2009 something called “Data Science”
  - analytics @ Eric Ries’ firm where “Lean Startup” began
- just prior: community evangelist for Apache Spark
Q: How does a Media company use AI?
We started with that question

More obvious answers:
writing articles and case studies about other firms adopting AI?
perhaps building some chat bots?

Less obvious answers:
surfacing the structure of content, in detail, as it evolves
enhancing next-gen search / recommendations
guiding editors through content gap analysis, formative assessment, curriculum pathing, etc.
AI NY 2016:
O’Reilly staff came away from this conference last year recognizing several AI applications now possible in Media

We spent 3 months evaluating approaches, then put together a cross-team effort to build new apps

Currently we’re moving into live customer experiments, then production use on Platform
Media in a world before...
Changes underway in Media

- people don’t buy stacks of printed books anymore
- learning curves accelerate as traditional “career paths” vanish
- editors are overwhelmed, authors even more so
- people consume media on mobile devices
- video is in high demand, generationally, although it’s “expensive” for editorial
- conversational interfaces are becoming quite interesting
- also take into account a shift toward computable content
- engineers may not be heads-stuck in IDEs much longer
Media in a world after... ?!

Sunspring

http://benjamin.wtf/

It’s No Game
Learning experiences
Peer Teaching through a range of Media

- books, videos
- conferences
- live online courses
- computable content
- case studies
- learning paths
- articles
- podcast interviews
- chat forums
We’re a learning company. Conferences are live learning events, closely linked into the depth of learning experiences offered on our Platform. Both are key to helping companies get started and stay ahead with AI, Big Data, SRE, Design Thinking, Leadership, etc.
White paper: "How do you learn?"
White paper: "Why self assessments improve learning"
White paper: "How great companies make change happen"
White paper: "Learn alongside innovators, thought-by-thought"
O’Reilly Media – Platform

Our **AI @ O’Reilly** team comes from a background of working to implement *personalized learning*. Internal *data products* at O’Reilly create foundations for building that

**Peer teaching + AI + Platform → Personalized learning**
Key insight for Media + AI:
Any content that can represented as text can be parsed by NLP, then manipulated by available tooling
More observations about Media...

- labeled images get *really* interesting
- text or images within a context has an inherent structure
- representation of that structure is all but nonexistent in the Media vertical
- O’Reilly leverages AI to fill that void
Problem space
Disambiguation in content discovery

Consider searching for the term `react` on Google, where the first page of results includes:

- acting coaches
- video games
- student engagement
- children’s charities
- *UI web components (what our audience wants?)*
- surveys
Options...

- using previous browser history can help provide context for disambiguation
- customers *might* scroll through many varied results
- if neither of the above work well, then do something better
- see also [Query Understanding](#)
We have lots of popular content about the JavaScript library React.

We also have content about Psychology, Chemistry, Physics, etc., all of which are valid searches for the term `react`.

O’Reilly Media – Platform
People are great at catching subtle nuances about “good” versus “poor” search results.

Deep learning would be interesting; however, without good labels first, DL isn’t particularly useful

Too expensive for editors to annotate books/videos from 200+ publishers

A good case for active learning, weak supervision, etc., to augment our labeling and help extend our ontology
You need an ontology
AI > Deep Learning

*Perception, prediction, memory* – are necessary; however, they do not address *understanding*

We must distinguish between what humans and machines do well: “human-in-the-loop”

- *cognitive load, speed, scale, repeatability:* 
  - **computers** > humans

- *curation* (captchas, as an example): 
  - computers < **humans**
Ontology

- what Deep Learning doesn’t provide directly
- aka, “knowledge graph” – a computable thesaurus
- maps the semantics of business relationships
- the hard part, a relatively expensive investment

- S/V/O: “nouns”, some “verbs”, a few “adjectives”
- a complement to DL which brings ROI to AI, FTW
- conversational interfaces (e.g., Google Assistant) begin to improve UX by importing ontologies through APIs
Ontology

- we built atop OWL/RDF, with SKOS as the core
- Library of Congress and DBpedia as upper ontology
- using rdflib and NetworkX to manage the graph
- mid-level, across domains for 200+ publishers
- vector embedding used to expand the graph
- reified with other data sets, converted into JSON for the React UI to use
- where we believe the L&D market is heading, to manage learning materials more effectively
Ontology

Deep learning (deep structured learning) is a subfield of machine learning based on a set of algorithms that build on ideas using multiple processing layers, with emphasis on data transformations. Deep learning has become an umbrella term for neural networks.

- Deep learning (deep structured learning) is a subfield of machine learning based on a set of algorithms that build on ideas using multiple processing layers, with emphasis on data transformations. Deep learning has become an umbrella term for neural networks.
Which parts do people or machines do best?

- **Human scale**
- **Primary structure**

**Team goal:** *solve et coagula* – maintain structural correspondence

**Big win:** inferences across the graph
Which parts do people or machines do best?
Which parts do people or machines do best?

Python

- [ipython, 0.7184, 0.7184]
- [ruby, 0.7088, 0.7088]
- [perl, 0.6965, 0.6965]
- ['standard library', 0.6856, 0.6856]
- [python code, 0.6851, 0.6851]
- [numpy, 0.6797, 0.6797]
- [matplotlib, 0.6794, 0.6794]

NumFOCUS

PSF

functional programming

scripting

web development

data science

OSCON

Strata

continuum analytics

open source framework

software library

fluent

oscon
Example use: “heavy-lifting” in our UX
Apps and tooling
Relative scale

Using the same data size, same algorithms which required a Hadoop cluster a decade ago ... these now run within a few minutes in Python on my laptop.

Pause to think about that before staffing a team, making technology selections, etc.
Components

- **rdflib + NetworkX**: ontology graph represented as N3 “turtle”
- **PyTextRank**: NLP parsing, features, summarization
- **Jupyter + nbtransom**: human-in-the-loop ML pipelines
- **Apache Spark**: sort, partitioning, task management
- **scikit-learn**: machine learning models
- **gensim**: vector embedding / deep learning
- **datasketch**: approximation algorithms
- **Flask, React, Node.js**: microservices, UI web components
- **Redis**: in-memory indexing, full-text search
PyTextRank

TextRank (R Mihalcea, P Tarau, 2004) a graph algorithm that extracts key phrases and summarizes texts – for improved NLP over use of keywords, n-grams, etc.

- construct a graph from a paragraph of text
- run PageRank on that graph
- extract the highly ranked phrases

Python implementation atop spaCy, NetworkX, datasketch:

- https://pypi.python.org/pypi/pytextrank/
Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.
Video transcription

- AI services in the cloud transcribe audio
- ontology provides domain-specific “hints”
- comparable results with human transcription services, at least for use in NLP
- speeds development of formative assessment, and other text-based use of video content
- costs pennies on the dollar, fast turnaround
- otherwise, video is too time-intensive for editorial staff
- to-do: synthesized outlines
Transcript: let's take a look at a few examples often when people are first learning about Docker they try and put it in one of a few existing categories sometimes people think it's a virtualization tool like VMware or virtualbox also known as a hypervisor these are tools which are emulating hardware for virtual software

Confidence: 0.973419129848
Semantic similarity

Semantic similarity between two texts can be measured using a **Jaccard** measure, and approximated using **MinHash** and **LSH**

Used to construct a graph of the overall content, i.e., showing structure
Management strategy
Design pattern: Human-in-the-loop

*Real-World Active Learning: Applications and Strategies for Human-in-the-Loop Machine Learning*


Ted Cuzzillo

*O’Reilly Media, 2015-02-05*

Develop a policy for how human experts select exemplars:

- bias toward labels most likely to influence the classifier
- bias toward ensemble disagreement
- bias toward denser regions of training data
Design pattern: Human-in-the-loop

Building a business that combines human experts and data science

Eric Colson, StitchFix
O’Reilly Data Show, 2016-01-28

“what machines can’t do are things around cognition, things that have to do with ambient information, or appreciation of aesthetics, or even the ability to relate to another human”
Design pattern: Human-in-the-loop

*Creating large training data sets quickly*
oreilly.com/ideas/creating-large-training-data-sets-quickly

**Alex Ratner**, Stanford
*O’Reilly Data Show*, 2017-06-08

Snorkel: “weak supervision” and “data programming” as another instance of human-in-the-loop
github.com/HazyResearch/snorkel

conferences.oreilly.com/strata/strata-ny/public/schedule/detail/61849
Active learning

- special case of semi-supervised machine learning
- send difficult tasks or edge cases to human experts, let algorithms handle more routine examples
- works best in cases which have lots of cheap, unlabeled data
- for example: an abundance of content to be classified, where the cost of labeling is expensive

https://en.wikipedia.org/wiki/Active_learning_(machine_learning)
Active learning

*Data preparation in the age of deep learning*

**Luke Biewald**, CrowdFlower

*O’Reilly Data Show, 2017-05-04*

send human workers cases where machine learning algorithms signal uncertainty (low probability scores)

or when your ensemble of machine learning algorithms signals disagreement
Buy vs. build

Recognizing commercial services such as CrowdFlower and WorkFusion, we needed to initiate this work with a relatively small team, working closely to define our methods, processes, and policies.

Another way to look at this: as a bootstrap process, preparing our corpus for subsequent use of deep learning, etc., after we have sufficient labels.
domain driven design  a pattern language for software development, focused on domain logic

detailed design document  a written description of a software product

debugger tools  a computer program used to test and debug other programs

Sams Teach Yourself Java™ in 21 Days (Covering Java 8)
by Rogers Cadnehead
Published by Sams

In just 21 days, you can acquire the knowledge and skills necessary to develop applications on your/vcomputer and apps that run on Android phones and tablets. With this incomplete tutorial you'll ...

Java® Performance Companion
by Bengt Rutisson, Charlie Hunt, Monica Beckwith, Poonam Parhar
Published by Addison-Wesley

"Java® Performance® Companion shows how to systematically and proactively improve Java performance with today's advanced multicore/hardware and complex

Core Java® Volume I - Fundamentals, Tenth Edition
by Cay S. Horstmann
Published by Prentice Hall

* Core Java® has long been recognized as the leading, no-nonsense tutorial and reference for new/experienced programmers who want to write robust Java code for real-world applications. Now, Core ...

Expand
Collaboration through Jupyter

Notebooks used to manage ML pipelines, where machines and people collaborate

- “Human-in-the-loop design pattern” talk @ JupyterCon NY 2017
- experts adjust hyperparameters in ML pipelines
- machines write “logs” of ML modeling and evaluation
- experts run `jupyter notebook` via SSH tunnel for remote access – in lieu of dedicated UIs
- this work anticipates upcoming collaborative document features in Jupyter
- https://pypi.python.org/pypi/nbtransom
Collaboration through Jupyter

Notebooks used to manage ML pipelines machines and people collaborate

- “Human-in-the-loop design pattern” talk @ ...
- experts
- machines
- evaluate
- experts
- for remote
- this work
- document
- https://pypi.python.org/pypi/nbtransom

“Human-in-the-loop design pattern”

session @ JupyterCon NY 2017
Collaboration through Jupyter
Collaboration through Jupyter

- running a notebook via SSH tunnel removes a need for dedicated UIs
- this work anticipates upcoming collaborative document features in Jupyter:

### Real-time collaboration in Jupyter notebooks

Ian Rose

JupyterCon, Aug 24
Expert review

- ML pipeline reports the results – labels
- human-in-the-loop expert reviews content, especially for the edge cases

```
| orm:React | 9781783558568 | ch03s03.html | 0.03 0.97 | Mastering React |
| orm:React | 9780134546513 | ch02.html | 0.07 0.93 | Learning React |
| orm:React | 9781786462558 | ch09s08.html | 0.10 0.90 | React Native Cookbook |
| orm:React | 9781939902436 | ch04.html | 0.13 0.87 | Developing a Redux Edge |
| orm:React | 9781783551620 | ch02.html | 0.17 0.83 | React.js Essentials |
| orm:React | 9781783986668 | ch07s02.html | 0.29 0.71 | Clojure Reactive Programming |
| orm:React | 9781783550258 | ch02.html | 0.40 0.60 | Mastering CryENGINE |
| orm:React | 9780981531656 | h3_id_19.html | 0.42 0.58 | Actors in Scala |
| orm:React | 05965227225 | ch22s04.html | 0.47 0.53 | Cisco IOS Cookbook, 2nd Edition |
| orm:React | 9780132542913 | ch06.html | 0.49 0.51 | The Clean Coder |
| null | 9780071625098 | ch09.html | 0.91 0.09 | JavaServer Faces 2.0 |
| null | 9780133846904 | ch01.html | 0.81 0.19 | Reactive Messaging Patterns with the Actor Model |
| null | 978932558205 | Chapter01.xhtml | 0.73 0.27 | Engineering Chemistry |
| null | 0131428240 | ch04.html | 0.65 0.35 | Walking the Tightrope |
| null | 9780470490068 | ch03.html | 0.61 0.39 | The Reactor Factor |
```
Nuances for active learning

Our ML pipeline is based on good/poor exemplars from search results – active learning features are being integrated as an “expert mode” in the UI

**No Free Lunch theorem**: it’s better for us to err on the side of less false positives / more false negatives

We bias toward exemplars most likely to influence the classifier
Human-in-the-loop as management strategy

*Personal opinion:* the “game” isn’t to replace people – instead, it’s about leveraging AI to augment people, so companies can retain people with valuable domain expertise, making their work even more vital.
Using AI to create new jobs
“We won’t run out of work until we run out of problems...

Our main advances have come when we invested in other people’s children – massive investment in EU following WWII, built from something that resembles Syria today.”

– Tim O’Reilly
Business outcomes
But how is this AI?

- more work, quicker, than could be performed by editors – who are already super busy people
- exceeding human parity as a benchmark
- relieves pressure on the organization, as learning curves accelerate
- augments some of our most valuable personnel, so they can do more
- helps evolve our management strategy, for net creation of jobs
Data products

- search < recommendations
- notifications for emerging trends
- conversational interfaces (e.g., Groupbot)
- content acquisition gap analysis
- better content complexity estimated
- scoping needs for assessment
- curriculum pathing suggestions
- auto-summarization (editorial drafts?)
- how content themes evolve over time
Hindsight 20/20

- A former exec urged against full-text NLP, concerned that it’d be too difficult/expensive
- Entire corpus was parsed on an old desktop server, < 400 lines of Python for NLP
- < 50K titles in corpus
- < 5K terms in working vocabulary
- < 3 Gb index, which fits in memory
- ~80% of search results can be pre-computed and become web cache hits
- Yet, this conference will generate ~100 hours of video, and we produce 20 conferences per year
Supply and demand

We need much much **less** of:

*brotopian scrum teams... building more tools... hellbent on more yet APIs and web apps*

We need much much **more** of:

*people who leverage extended dialog with machines*
“The future belongs to those who understand at a very deep level how to combine their unique expertise with what algorithms do best.”

– Pedro Domingos, *The Master Algorithm*
Miscellaneous caveats

- know your numbers, understand scaling reality (not myths)
- **InnerSource**, to break silos – avoid wasting lots of time
- infra + data eng at scale is **not** a good first step
- have only 2-3 people edit your ontology
- anyone who blurted “But, web apps” or “But, databases” or “But, Solr” got asked to leave the room ASAP
- AI projects have a low tolerance threshold for people who bank on tech debt
A plug for InnerSource...

We thought the introduction of data science had run headlong into enterprise silos and lingering tech debt. As if!

The introduction of AI exacerbates that problem even more so. Suggested responses:

- **be prepared for internal battles due to AI disruptions**
- **InnerSourceCommons.org** open source practices *within* enterprise
- design patterns for working across silos
- think: “Good house rules for guests” as other teams submit PRs
“In terms of industry adoption, I think data readiness is a pretty good predictor of how quickly a particular business, and by extension an industry, will adopt machine learning. You can buy all the software that you want; if you still are struggling to pull data from disparate silos, clean it up, make sense of it, you’re not going to get anywhere.”

– David Beyer, Machine Intelligence in the wild
O’Reilly Strata
CN, Jul 12-15
NY, Sep 25-28
SG, Dec 4-7
SJ, Mar 5-8
UK, May 21-24

O’Reilly Artificial Intelligence
SF, Sep 17-20

JupyterCon
NY, Aug 22-25
“Jupyter is the new front end for data science and AI.”

- Andrew Odewahn, CTO, O'Reilly Media

At JupyterCon, discover how to scale analytics to create business value from data—and transform your organization.
updates, reviews, conference summaries...

liber118.com/pxn/
@pacoid