DAWN: Data Analytics for What’s Next

A Stanford University Industrial Affiliates Program

DAWN Vision

DAWN is a five-year Stanford research project being launched in 2017 to design systems and tools for usable machine learning, enabling non-experts to build and run production-ready ML apps.

- Four faculty members that have created industry-changing technology such as Apache Spark, Apache Mesos, DeepDive, and Niagara.
- Close working collaboration with industrial partners: joint workshops, retreats, collaboration on real apps and data.
- Open source, permissive license release of all software.
- An approach validated by early results with dozens of partners.

DAWN Team

The DAWN team consists of four Stanford professors and their research groups:

Chris Ré: Expert on data and ML systems whose work on deep learning and dark data is used at hundreds of organizations (FDA, VA, major web companies).

Matei Zaharia: Expert on parallel computing systems and creator of Apache Spark, Apache Mesos, Tachyon, and other widely used open source software.

Peter Bailis: Expert on databases and stream processing with DAWN infrastructure already in use at many companies.

Kunle Olukotun: Hardware and programming model expert whose work led to today’s multicore CPUs, GPUs, and DSL-based programming models.

Research Goals

Although machine learning has tremendous potential, ML applications remain too hard and expensive to build. We believe that ML can be democratized in the same way as visualization, planning, SQL, and search have in the past, allowing hundreds of times more ML applications. We have three goals:

1) Democratize data preparation by creating new tools that let domain experts (e.g., IT or engineers) clean and label data without major human effort.
2) Democratize training by automatically creating robust models that can easily be explained and updated based on new requirements, input data, and target environments.
3) Democratize production operation by designing ML systems that run in real time across a variety of platforms and can be monitored and debugged.

As a moonshot result, we aim that with DAWN technology, a non-ML domain expert (e.g., DevOps specialist) will be able to build a production machine learning app (e.g., anomaly detection for a new service) in hours using a natural language interface.

Research Themes

The DAWN research agenda currently comprises the following impactful research themes.

Research Theme	Industry Impact
Data programming (Snorkel, DeepDive)	Let domain experts easily train models from domain-specific data (e.g., log entries, time series)
Real-time analytics (MacroBase)	Provide robust and scalable monitoring and anomaly detection across large-scale telemetry
Unsupervised and transfer learning	Leverage unlabeled data and patterns across customers automatically when building applications
Automatic training	Simplify the creation of effective models
High-level interfaces	Let engineers or customers easily build ML apps using domain-specific interfaces similar to search
Optimizing runtimes (Weld, Delite, Spark)	Deploy models on current and emerging hardware platforms or at edge devices without porting
Debugging/observability	Allow engineers to debug and tune models
Combining inference and actuation	Build robust actuation-based products with explainable models and guarantees
Personalized learning	Robustly combine customer-specific and industry data

Impact

Our existing projects including DeepDive (SIGMOD ’16), Snorkel (NIPS/ICML ’16), MacroBase (CIDR/SIGMOD ’17), and Weld (CIDR ’17) are already in use at industry and scientific labs for applications such as electric vehicle monitoring, personalized medicine, datacenter optimization, text mining, and big data analytics.

Engagement

Corporate members are an integral part of DAWN. They provide insights on real-world problems, opportunities, and constraints that inform and inspire our research. They engage with faculty and students. They provide a path to testing and applying our innovations, thus leading to large-scale impact.

Corporate engagement includes the following elements:

- Invitations to technical retreats every 6 months

- Invitations to on-campus workshops on DAWN technologies (e.g., Snorkel hackathon, MacroBase hackathon, etc.)—great opportunities to engage with students
- An opportunity to serve on the DAWN advisory board, thus providing input on industry considerations
- Early access to software prototypes that we develop, under a permissive open source license
- Access to faculty members and students to discuss research and pursue collaboration

Funding

DAWN is a five-year program commencing in 2017 that is supported by unrestricted funding from a small number of partner companies that each contribute $500k per year plus additional support from government funding agencies like NSF and DARPA. DAWN is a Stanford University industrial affiliates program and is subject to university policies for such programs including openness in research, publication and broad sharing of results, and faculty freedom to pursue research topics and methodology of their choice. See https://industrialaffiliates.stanford.edu/. DAWN is affiliated with the Stanford Data Science Initiative, a university-wide program for massive data, analytics, and algorithms.

IP

DAWN researchers will use and develop open-source software, and it is the intention of all DAWN researchers that any software released will be released under an open source model, such as BSD.

Information

For further information please contact any member of the DAWN team or Steve Eglash, Executive Director, seglash@stanford.edu.