DAWN: Infrastructure for Usable Machine Learning

Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia

dawn.cs.stanford.edu
It’s the Golden Age of Data*

Incredible advances in image recognition, natural language processing, planning, info retrieval

Society-scale impact: autonomous vehicles, personalized medicine, real-time translation

No end in sight for advances in ML

*for the best-funded, best-trained engineering teams
Building ML Products is Too Hard

Major successes (e.g., Siri, AlphaGo) require hundreds to thousands of engineers

Most effort in data acquisition, preparation, testing and productionizing: not just core ML!
“Only a fraction of real-world ML systems is composed of ML code”
The DAWN Question

What if anyone with domain expertise could build their own production-quality ML products?

• Without a PhD in machine learning
• Without being an expert in systems
• Without understanding the latest hardware

It’s happened before
It’s happened before: Search

**Before:** Decades of research on information retrieval, indexes, ranking, etc

**After:** any developer can add search to an app using a library (e.g. Solr); any user can use search

**Key idea:** end-to-end systems that tackle the barriers to access & production use
The DAWN Stack

Data Acquisition
- Snorkel
- DeepDive

Feature Engineering
- MacroBase (Streaming Data)
- Data Fusion

Model Training
- ModelSnap
- AutoRec, SimDex (Recommendation)
- Mulligan (SQL+graph+ML)

Productionizing
- ModelQA
- NoScope (Video)

Hardware
- CPU
- GPU
- FPGA
- Cluster
- Mobile

New Hardware: FuzzyBit, Plasticine CGRA

End-to-End Compilers: Weld, Delite
Example: MacroBase for Continuous Analytics

End-to-end system for anomaly identification

multi-dimensional data streams → MacroBase → anomalies & explanations
<table>
<thead>
<tr>
<th>record_id</th>
<th>user_id</th>
<th>state</th>
<th>hw_make</th>
<th>hw_model</th>
<th>firmware_version</th>
<th>app_version</th>
<th>avg_temp</th>
<th>battery_drain</th>
<th>trip_time</th>
</tr>
</thead>
<tbody>
<tr>
<td>131920</td>
<td>49e36c5b031141ddcf240f7</td>
<td>CO</td>
<td>Lenovo</td>
<td>Lenovo_K910L</td>
<td>4.4.2</td>
<td>v21</td>
<td>79.251224</td>
<td>0.205834</td>
<td>40.910145</td>
</tr>
<tr>
<td>131921</td>
<td>a670eab2b6d4e5991ea4269</td>
<td>WV</td>
<td>TCT (Alcatel)</td>
<td>4009A</td>
<td>7.1.1</td>
<td>v36</td>
<td>72.136380</td>
<td>0.184874</td>
<td>47.253076</td>
</tr>
<tr>
<td>131922</td>
<td>247c64e8a87438295ff1999</td>
<td>UT</td>
<td>TCT (Alcatel)</td>
<td>4009A</td>
<td>7.1.1</td>
<td>v31</td>
<td>77.300103</td>
<td>0.230015</td>
<td>25.342140</td>
</tr>
<tr>
<td>131924</td>
<td>6bd9af7242ca480a9675d90d</td>
<td>OH</td>
<td>HTC</td>
<td>HTC_M10u</td>
<td>6.0.1</td>
<td>v38</td>
<td>70.937014</td>
<td>0.454293</td>
<td>38.661611</td>
</tr>
<tr>
<td>131926</td>
<td>d449b12dcb3634d7af1021de</td>
<td>HI</td>
<td>HTC</td>
<td>HTC_Wildfire_S_A510b</td>
<td>6.0</td>
<td>v46</td>
<td>75.436764</td>
<td>0.151338</td>
<td>17.785555</td>
</tr>
<tr>
<td>131927</td>
<td>fff907a14e4a50ab76bd46</td>
<td>HI</td>
<td>bq</td>
<td>Aquaris_E4.5</td>
<td>4.4.1</td>
<td>v38</td>
<td>70.208187</td>
<td>0.286005</td>
<td>60.443799</td>
</tr>
<tr>
<td>131929</td>
<td>8226cd65bb1f4d61a66c4f55</td>
<td>MI</td>
<td>TCT (Alcatel)</td>
<td>ALCATEL_one_touch_97</td>
<td>6.0.1</td>
<td>v35</td>
<td>73.113370</td>
<td>0.249834</td>
<td>16.881133</td>
</tr>
<tr>
<td>131930</td>
<td>30e726fadc6744b2ace2d76b</td>
<td>LA</td>
<td>TCT (Alcatel)</td>
<td>ALCATEL_ONE_TOUCH_60</td>
<td>5.0</td>
<td>v40</td>
<td>77.918077</td>
<td>0.405417</td>
<td>51.163642</td>
</tr>
<tr>
<td>131931</td>
<td>569f35993da246f4af83c2e</td>
<td>FL</td>
<td>Lava</td>
<td>S1</td>
<td>6.0.1</td>
<td>v44</td>
<td>76.558080</td>
<td>0.416760</td>
<td>42.252460</td>
</tr>
<tr>
<td>131932</td>
<td>9d2db241316c433788b7ecc14</td>
<td>AL</td>
<td>LGE</td>
<td>LG-D724</td>
<td>7.0</td>
<td>v29</td>
<td>76.760340</td>
<td>0.334446</td>
<td>37.922632</td>
</tr>
<tr>
<td>131933</td>
<td>484a1ced0a6a4646874861c</td>
<td>LA</td>
<td>Hisense</td>
<td>LED42K680X3DU</td>
<td>4.4.4</td>
<td>v49</td>
<td>77.138769</td>
<td>0.409485</td>
<td>23.345804</td>
</tr>
<tr>
<td>131934</td>
<td>d375d5a0e10d4649e913e43</td>
<td>MI</td>
<td>Techno</td>
<td>TECNO_P5S</td>
<td>6.0</td>
<td>v31</td>
<td>70.115019</td>
<td>0.179464</td>
<td>45.051123</td>
</tr>
<tr>
<td>131936</td>
<td>e4385a64d96e4e89997ec07</td>
<td>WI</td>
<td>ZTE</td>
<td>Z828</td>
<td>6.0.1</td>
<td>v35</td>
<td>71.615570</td>
<td>0.396389</td>
<td>47.662474</td>
</tr>
<tr>
<td>131937</td>
<td>cf00ae2105bb4ec3cb43b64b2</td>
<td>FL</td>
<td>Spice</td>
<td>Spice_Mi-498H</td>
<td>5.0</td>
<td>v42</td>
<td>72.045184</td>
<td>0.327405</td>
<td>45.099422</td>
</tr>
<tr>
<td>131939</td>
<td>c94d264a8461490f85f1c2e</td>
<td>RI</td>
<td>Infocus</td>
<td>InFocus_M320u</td>
<td>4.4.1</td>
<td>v49</td>
<td>73.543359</td>
<td>0.224504</td>
<td>19.069803</td>
</tr>
<tr>
<td>131940</td>
<td>c3c82947ab5a4db9a2fe21</td>
<td>MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131943</td>
<td>4e4566143b14be1809ad4a9</td>
<td>RI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131944</td>
<td>0ee8ff3606b496392bedd49</td>
<td>NE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131946</td>
<td>0ee8ff3606b496392bedd49</td>
<td>OK</td>
<td>LGE</td>
<td>LS670</td>
<td>4.0.4</td>
<td>v30</td>
<td>78.186519</td>
<td>0.381604</td>
<td>27.601968</td>
</tr>
<tr>
<td>131947</td>
<td>7a4bc456aa54d20b1119d42</td>
<td>NV</td>
<td>Oppo</td>
<td>F1f</td>
<td>4.3.1</td>
<td>v38</td>
<td>77.434095</td>
<td>0.436364</td>
<td>40.869723</td>
</tr>
<tr>
<td>131948</td>
<td>c3c82947ab5a4db9a2fe21</td>
<td>GA</td>
<td>Huawei</td>
<td>HUAWEI_Y320-U151</td>
<td>4.4.3</td>
<td>v42</td>
<td>77.715329</td>
<td>0.281726</td>
<td>23.077248</td>
</tr>
<tr>
<td>131949</td>
<td>2b0acfc3a91f49b6aa70cbe5</td>
<td>MO</td>
<td>ZTE</td>
<td>KPN_Smart_300</td>
<td>4.4.4</td>
<td>v39</td>
<td>75.368614</td>
<td>0.371224</td>
<td>44.295975</td>
</tr>
<tr>
<td>131950</td>
<td>5aab0148ea7942deacfc9a27</td>
<td>NV</td>
<td>Ketab</td>
<td>TR10CS1</td>
<td>5.0</td>
<td>v49</td>
<td>79.459844</td>
<td>0.491424</td>
<td>37.653744</td>
</tr>
</tbody>
</table>

Too much data for manual inspection
Even harder when data is streaming
Early successful users: manufacturing, automotive, online video, mobile apps

github.com/stanford-futuredata/macrobase
The DAWN Stack

Data Acquisition
- Snorkel
- DeepDive

Feature Engineering
- MacroBase (Streaming Data)
- Data Fusion

Model Training
- ModelSnap
- AutoRec, SimDex (Recommendation)
- Mulligan (SQL+graph+ML)

Productionizing
- ModelQA
- NoScope (Video)

Hardware
- End-to-End Compilers: Weld, Delite
- New Hardware: FuzzyBit, Plasticine CGRA

Interfaces
- CPUs, GPUs, FPGAs, Clusters, Mobile

Algorithms
- Mulligan (SQL+graph+ML)

Systems
- Spark, Apache
Snorkel’s Approach: Weak Supervision

1) User writes *labeling functions*: short programs that may not always give right label
   - E.g. pattern to search in text

2) Snorkel simultaneously learns *noise* in LFs and a *noise-aware* target model (e.g. LSTM)

Result: 4 hours writing labeling functions matches months of hand-labeling 10,000+ documents
The DAWN Stack

Data Acquisition
- Snorkel
- DeepDive

Feature Engineering
- MacroBase (Streaming Data)
- Data Fusion

Model Training
- ModelSnap
- NoScope (Video)
- AutoRec, SimDex (Recommendation)
- Mulligan (SQL+graph+ML)

Productionizing
- ModelQA

Hardware
- End-to-End Compilers: Weld, Delite
- New Hardware: FuzzyBit, Plasticine CGRA

Interfaces
- CPU
- GPU
- FPGA
- Cluster
- Mobile

Algorithmic
- Systems
- Interfaces
- Algorithms
- Hardware
NoScope: Fast CNN-Based Video Queries

**Opportunity:** CNNs allow more accurate queries on visual data than ever

**Challenge:** processing 1 video in real time requires a $1000 GPU

**Result:** 100-3000x faster with <1% loss in accuracy via
- Model specialization
- Adaptive cascades

[github.com/stanford-futuredata/noscope](https://github.com/stanford-futuredata/noscope)
NoScope Results

**Elevator (best result)**
- 40x faster @ 99.9% accuracy
- 5858x faster @ 96% accuracy

**Taipei (worst result)**
- 36.5x faster @ 99.9% accuracy
- 206x faster @ 96% accuracy
DAWN: AI for everyone via new systems that tackle the barriers to real-world use

Whitepaper & more at dawn.stanford.edu