Streaming Data Explanation with MacroBase

Kai Sheng Tai

in collaboration with

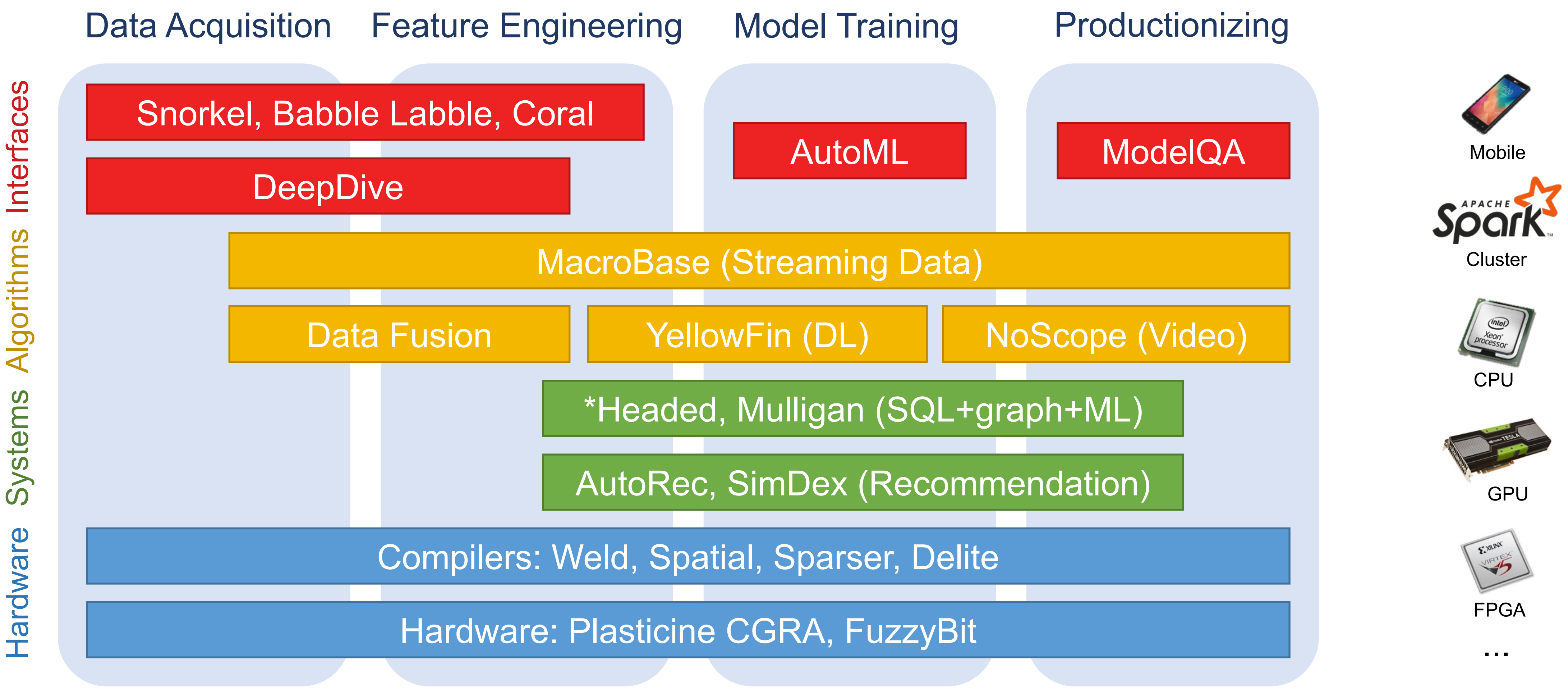
Peter Bailis, Edward Gan, Kexin Rong, Sahaana Suri,
Firas Abuzaid, Jialin Ding, Vatsal Sharan, Greg Valiant

Stanford DAWN Project

DAWN Project: Making ML More Accessible

PIs: Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia

dawn.cs.stanford.edu



DAWN Project: Making ML More Accessible

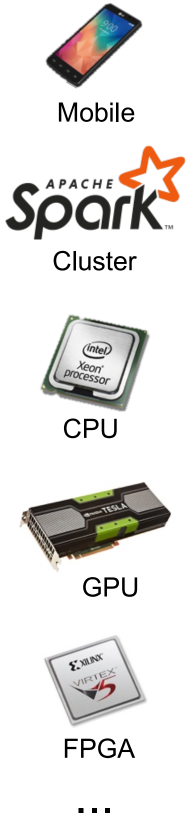
PIs: Peter Bailis, Kunle Olukotun, Chris Ré, Matei Zaharia

dawn.cs.stanford.edu

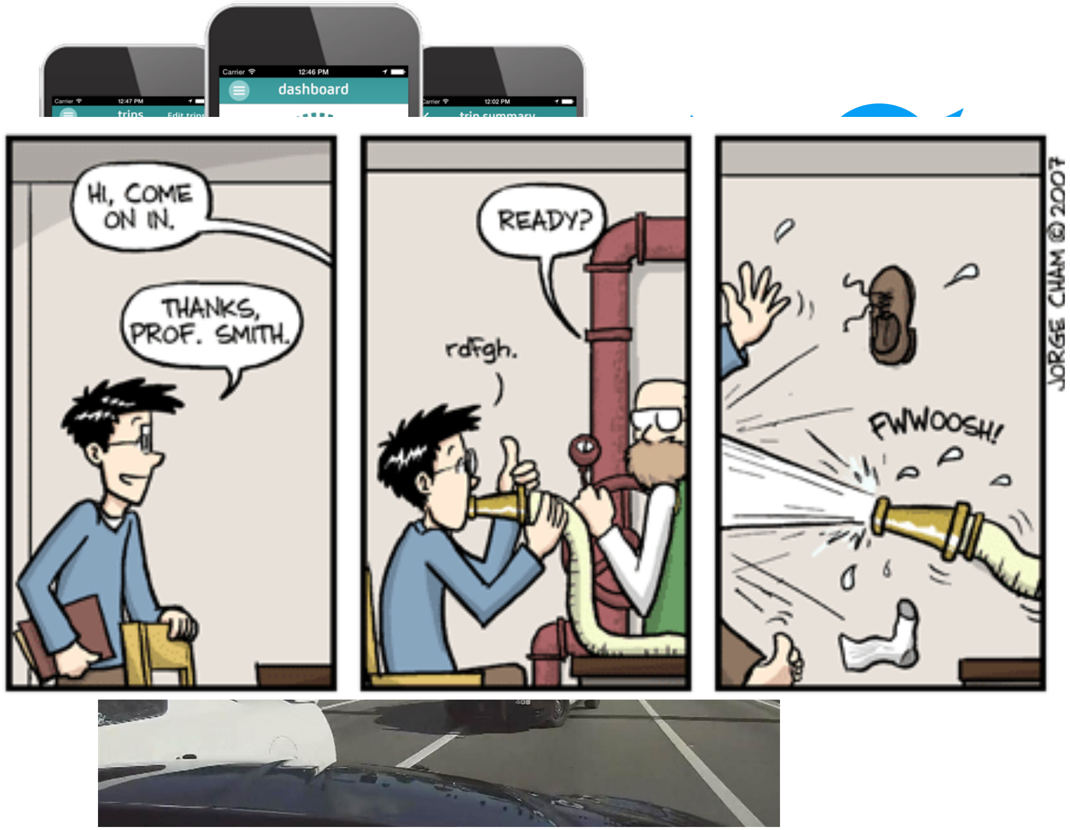
Data Acquisition | Feature Engineering | Model Training | Productionizing

Interfaces / Algorithms / Systems / Hardware

- Snorkel, Babble Labble, Coral
- DeepDive
- AutoML
- ModelQA
- MacroBase (Streaming Data)
- Data Fusion
- YellowFin (DL)
- NoScope (Video)
- *Headed, Mulligan (SQL+graph+ML)
- AutoRec, SimDex (Recommendation)
- Compilers: Weld, Spatial, Sparser, Delite
- Hardware: Plasticine CGRA, FuzzyBit



Continued Growth of Streaming Data Volumes



- Telemetry from mobile devices
  - >2B smartphones worldwide
- Application logs from web services
- Visual features from video streams
  - 1000s of dashcams, security cameras

MacroBase:
prioritizing *human attention*
via *feature selection*

MacroBase: Example Use Case

Input: stream of logs from mobile app (based on a real application)

Errors	Non-Errors
{iPhone7, USA}	{iPhone8, USA}
{iPhone7, **Canada**}	{iPhone7, USA}
{iPhone8, **Canada**}	{iPhoneX, USA}
{iPhone7, USA}	{iPhone7, USA}
{iPhone8, **Canada**}	{iPhone7, USA}
	{iPhone8, USA}
	{iPhone7, USA}
	{iPhone7, USA}

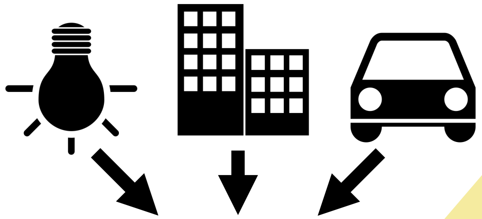
Explain error class to analyst with `[location = Canada]`

Challenges

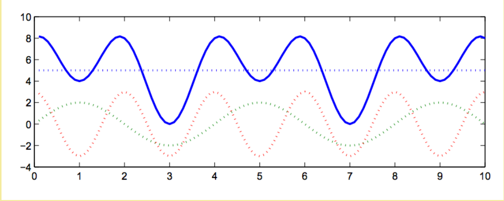
- **Throughput**:
  streams with millions of events/sec
- **Resource constraints**:
  limited computation and memory
- **Dimensionality**:
  high-order feature combinations
  (# phone models) x (# locations) x …

MacroBase Stream Analytics

macrobase.stanford.edu

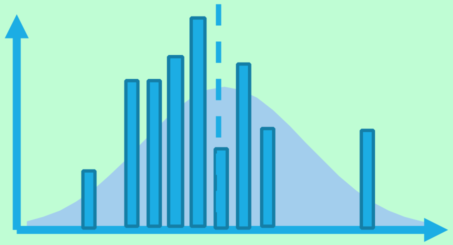


TRANSFORM



extract domain-specific signals

CLASSIFY



identify data in tails

EXPLAIN

find disproportionately correlated attributes

Outliers	**Inliers**
{iPhone6, Canada}	{iPhone6, USA}
{iPhone6, USA}	{iPhone6, USA}
{iPhone5, Canada}	{iPhone5, USA}

In production at:

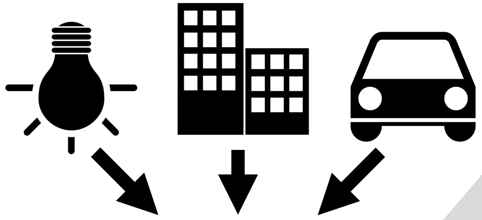
- major web service provider
- mobile app company
- video streaming service

Other projects:

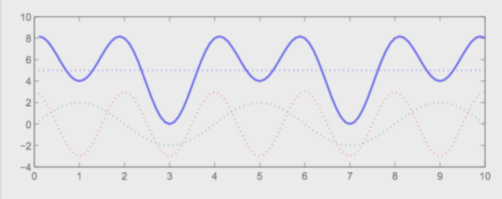
- Kernel density estimation
- Dimensionality reduction
- Faster CNN queries on video
- Method-of-moments for quantile estimation
- Time series visualization

MacroBase Stream Analytics

macrobase.stanford.edu

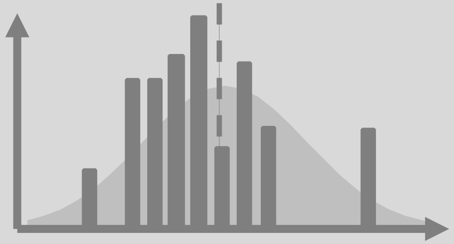


TRANSFORM



extract domain-specific signals

CLASSIFY



identify data in tails

EXPLAIN

find disproportionately correlated attributes

Outliers	**Inliers**
{iPhone6, Canada}	{iPhone6, USA}
{iPhone6, USA}	{iPhone6, USA}
{iPhone5, Canada}	{iPhone5, USA}

In production at:

- major web service provider
- mobile app company
- video streaming service

Other projects:

- Kernel density estimation
- Dimensionality reduction
- Faster CNN queries on video
- Method-of-moments for quantile estimation
- Time series visualization

This talk: Online feature selection on streams

MacroBase: Streaming Feature Selection

Setup: online learning of a linear classifier (e.g. logistic regression)
Goal: return top-*k* most discriminative features to the user

Track most *frequent* features?
Not necessarily the most *discriminative*

Sparsity-inducing regularization?
Hard to tune *a priori* to satisfy memory constraints

Weight-Median Sketch

[Tai, Sharan, Bailis, Valiant. arXiv 1711.02305]

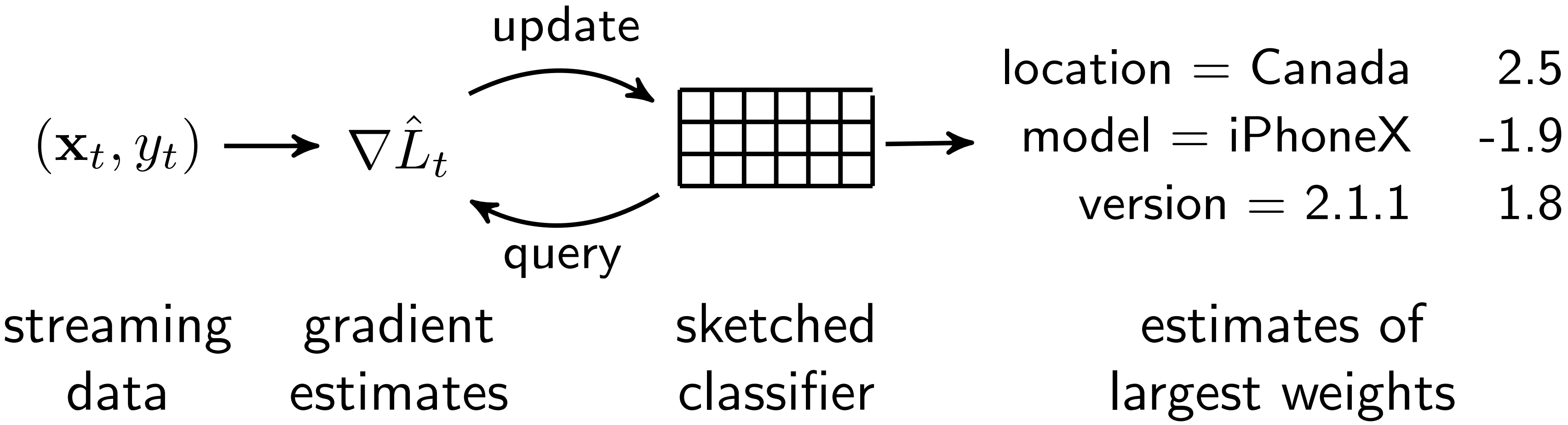
Maintain a *compressed* version (a **sketch**) of a linear classifier…

- … that supports fast updates
- … that supports queries for estimates of each weight
- … with (ϵ, δ)-approximation guarantee vs. uncompressed classifier

Track (approximation of) *k* most heavily-weighted features

Sketched Linear Classifiers

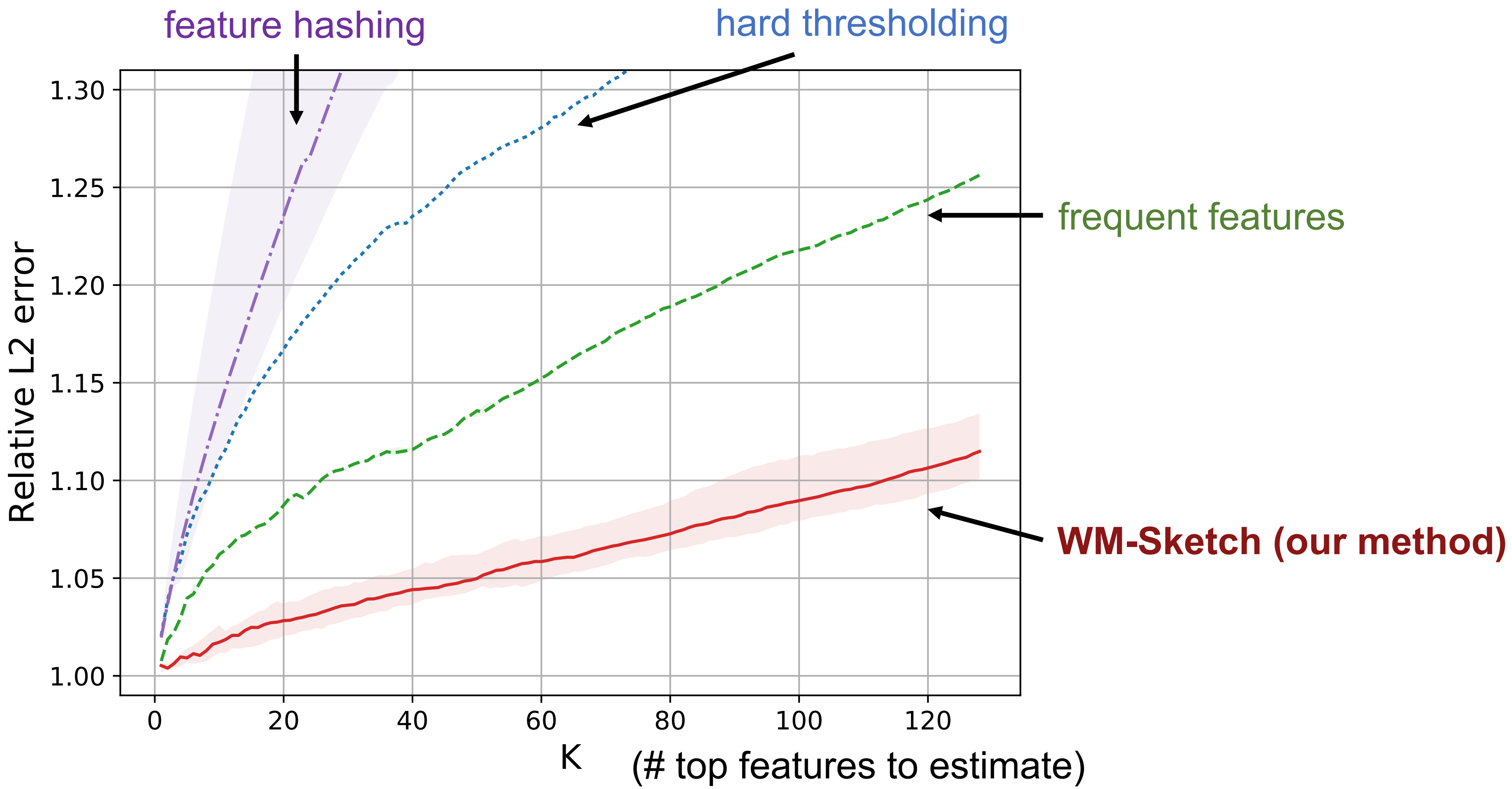
- Sketch of x: random projection of x to low dimension



Accurate weight recovery in practice

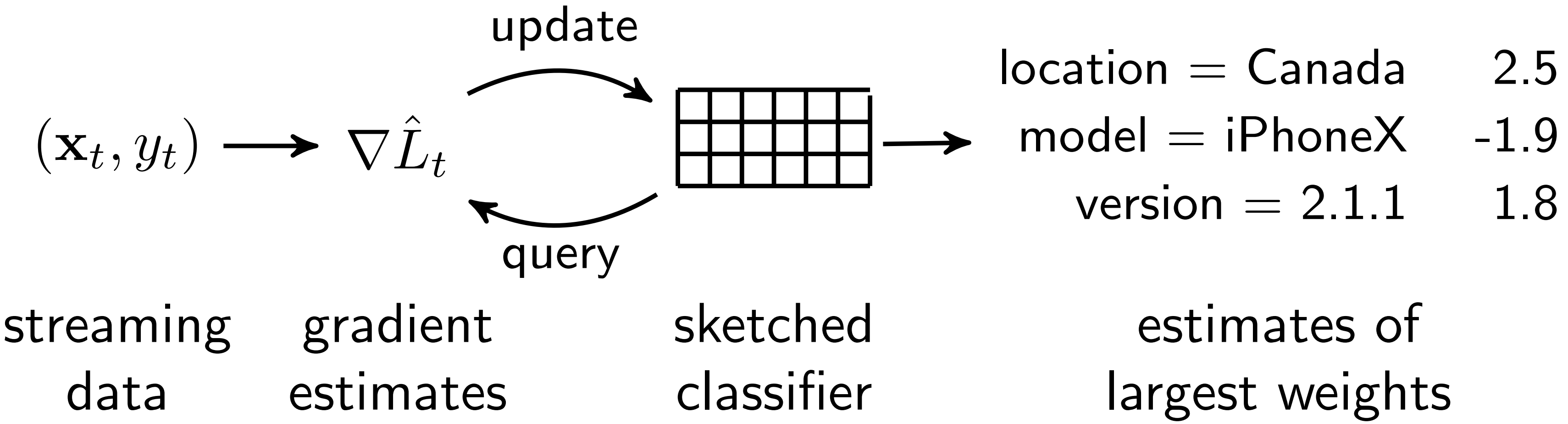
Online logistic regression on Reuters RCV1 with 4KB memory budget

(lower is better)



Sketched Linear Classifiers

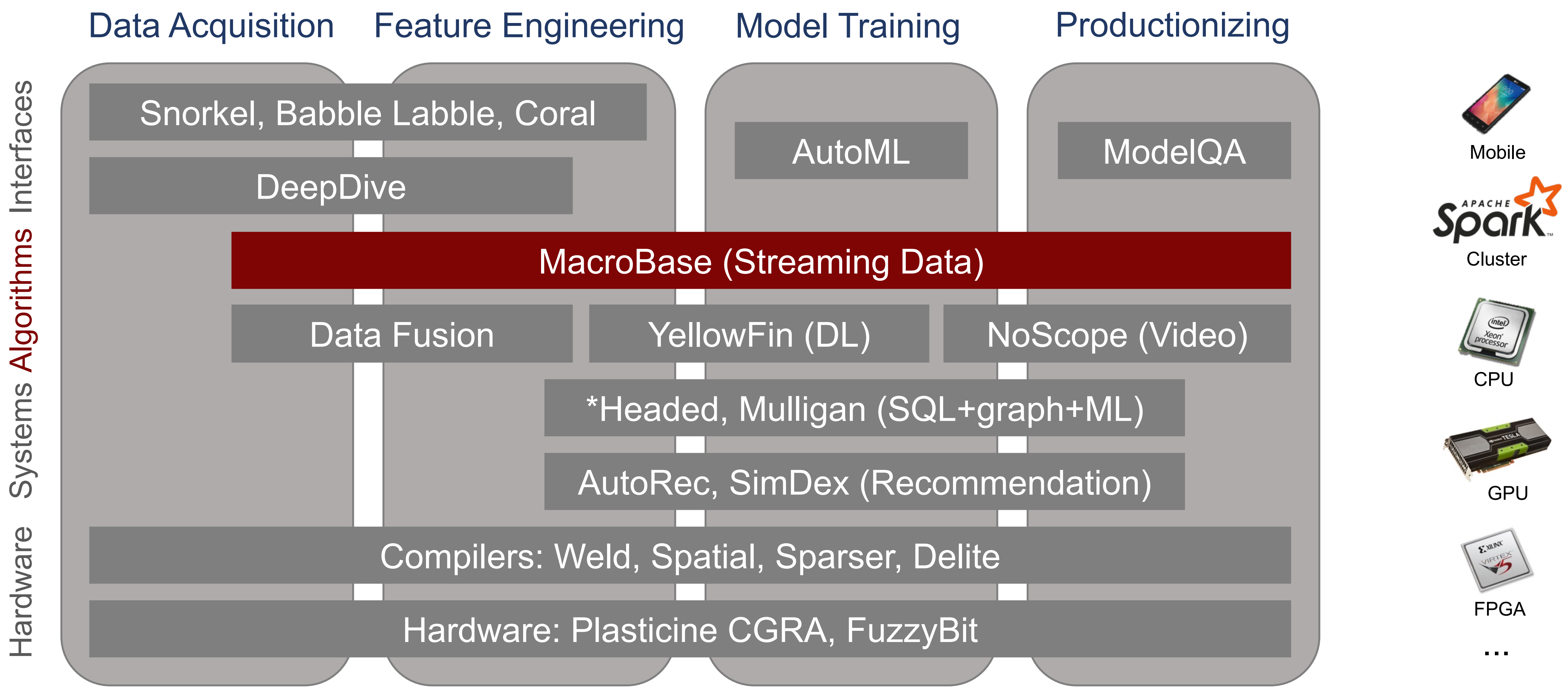
- Sketch of x: random projection of x to low dimension



Takeaways

- ***Count-Sketch*** data structure can be adapted to streaming feature selection
- Essentially ***feature hashing*** with ***highest-magnitude features in heap***
- Need only space ***logarithmic*** in original dimension

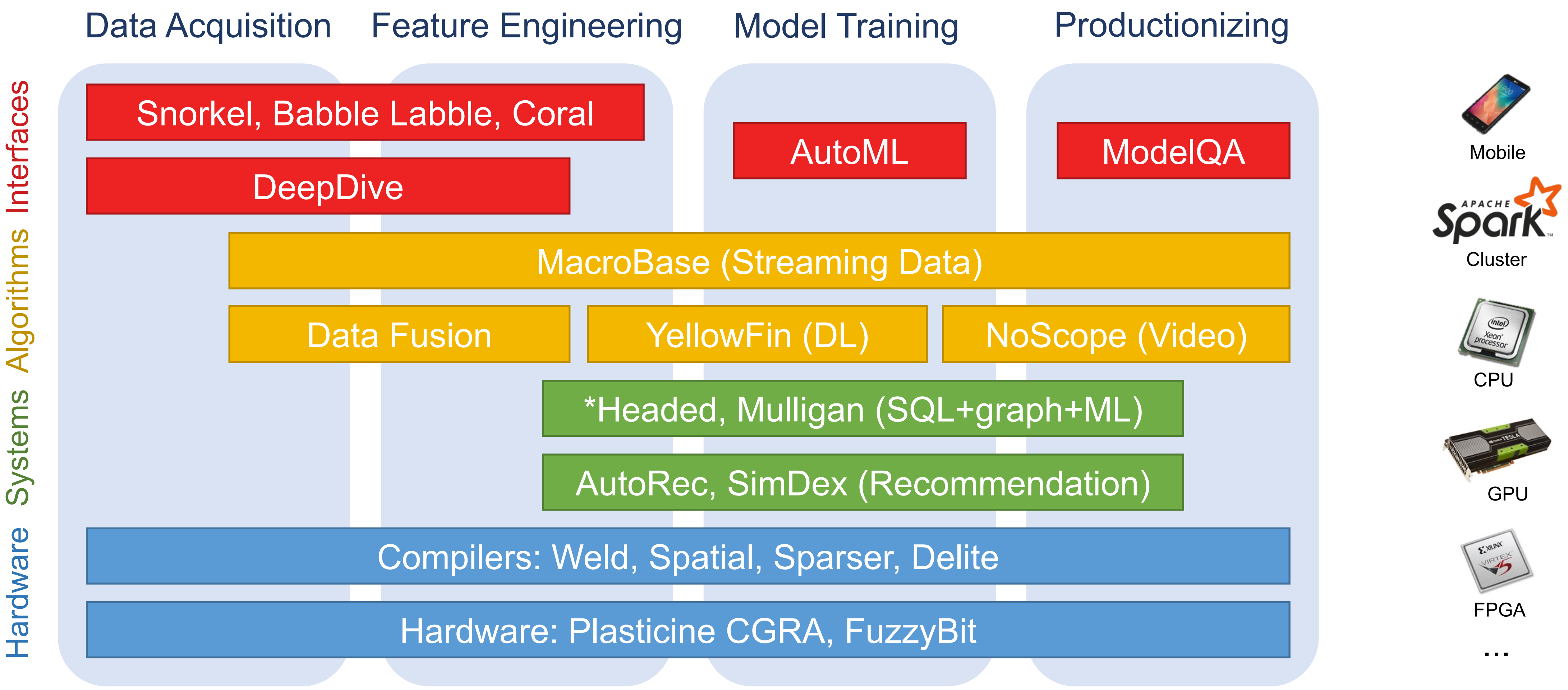
DAWN Stack



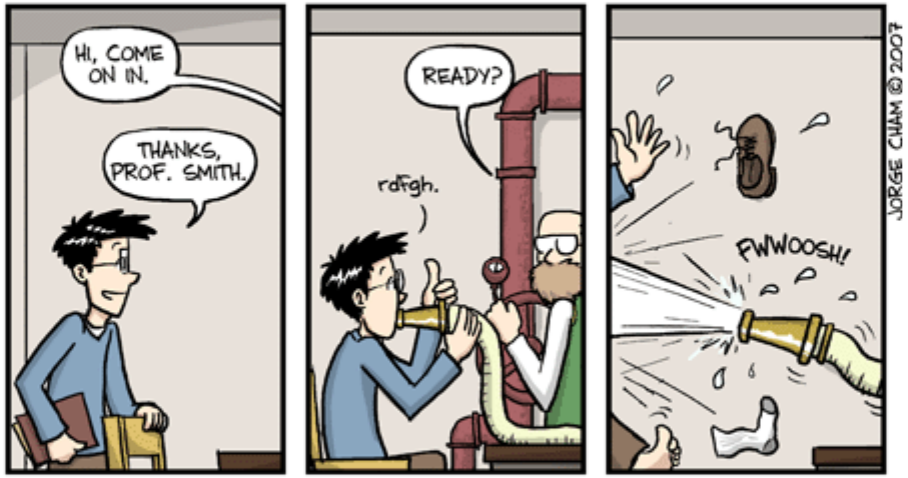
Data Acquisition
Feature Engineering
Model Training
Productionizing
Hardware
Systems
Algorithms
Interfaces
Snorkel, Babble Labble, Coral
DeepDive
AutoML
ModelQA
MacroBase (Streaming Data)
Data Fusion
YellowFin (DL)
NoScope (Video)
*Headed, Mulligan (SQL+graph+ML)
AutoRec, SimDex (Recommendation)
Compilers: Weld, Spatial, Sparser, Delite
Hardware: Plasticine CGRA, FuzzyBit
Mobile
Spark
Cluster
CPU
GPU
FPGA
...

DAWN Stack

Find out more @
dawn.cs.stanford.edu/blog



Recap



MacroBase: making sense of the firehose

This talk: Online feature selection by sketching linear classifiers

Check out other DAWN projects:
hardware + systems + ML

STANFORD DAWN

macrobase.stanford.edu

dawn.cs.stanford.edu

Kai Sheng Tai / kst@cs.stanford.edu