Microsoft

Microsoft Research Faculty Summit 2013

Microsoft

Interactive Machine Learning Challenges

Patrice Simard
Distinguished Engineer
Microsoft Research

Microsoft Research
Faculty Summit

Large unstructured data = untapped value

Lack of structure keeps the value locked in the data

Data examples: Web pages, queries, tweets, classified, eBay, email, ads, documents, etc.
Value tasks: Matching, querying, pivoting, clustering, analytics, etc.
Not covered: structured data. Ex: click prediction, graph analysis, Chess, etc.

Humans can extract value, inefficiently.

Task: Find all the job posting and resumes among 100M web pages. At 10 seconds per page: a lifetime.

Machines can do this at scale, but…

Building classifiers and field extractors to replace humans is difficult and not always accurate.
Machine learning is the method of choice but it is expensive, slow to build, and not widely accessible.

Bottleneck: ML experts and infrastructure

With time and energy, we can build a few classifiers

ML experts can build a few classifiers and form fillers for the head. Typical procedures:

For each new classifier or form filler, repeat:

Repeat until good enough (cycle time is weeks or months)

Collect small (biased) data set (e.g. 50K examples)

Label data (outsourced)

Feature engineering and ML tuning (highly paid ML expert)

Deploy on real data (and pray)

The standard approach does not scale to the tail

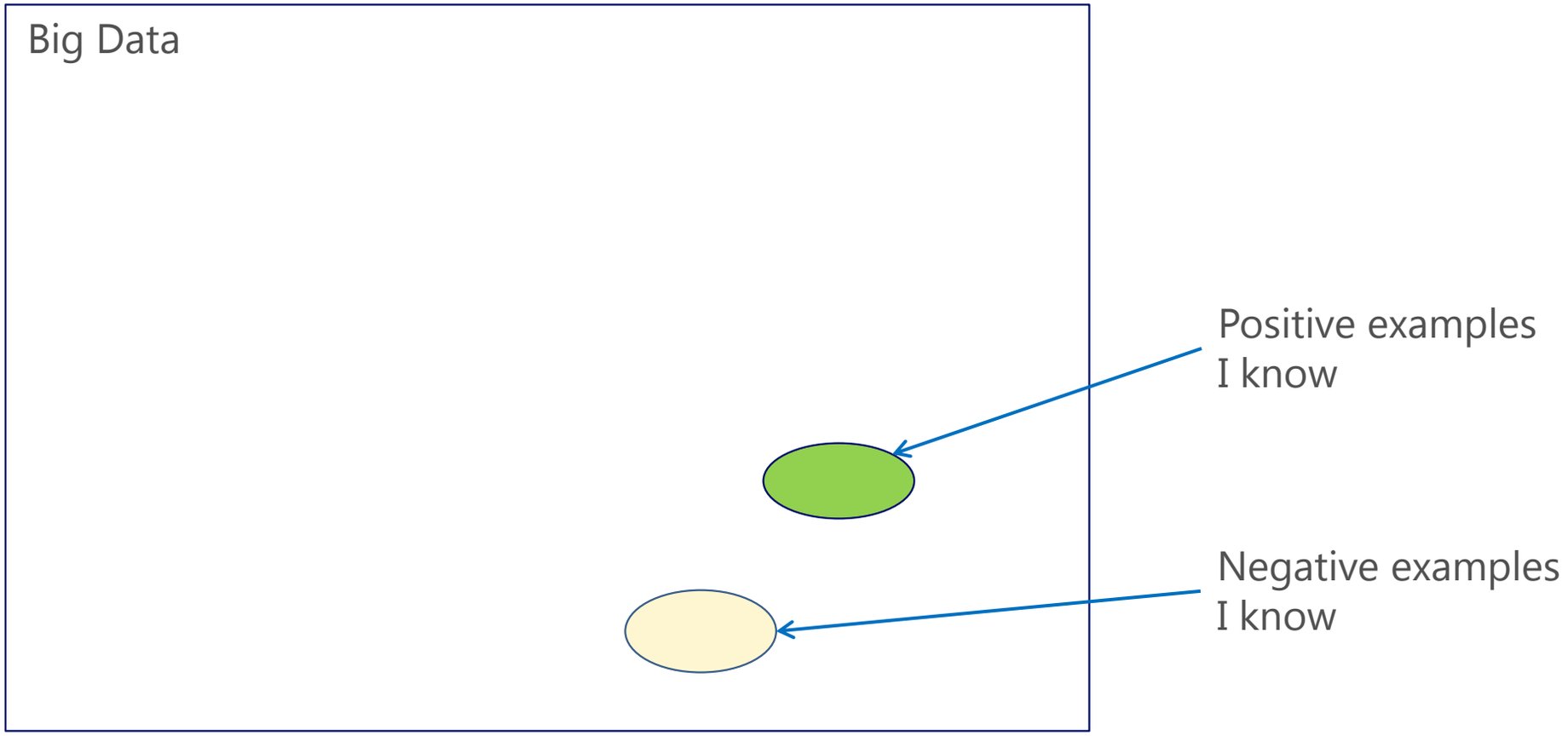
ML experts are expensive, hard to find, and hard to evaluate.

Large startup cost (data collection, labeling, ML testing, deployment, etc).

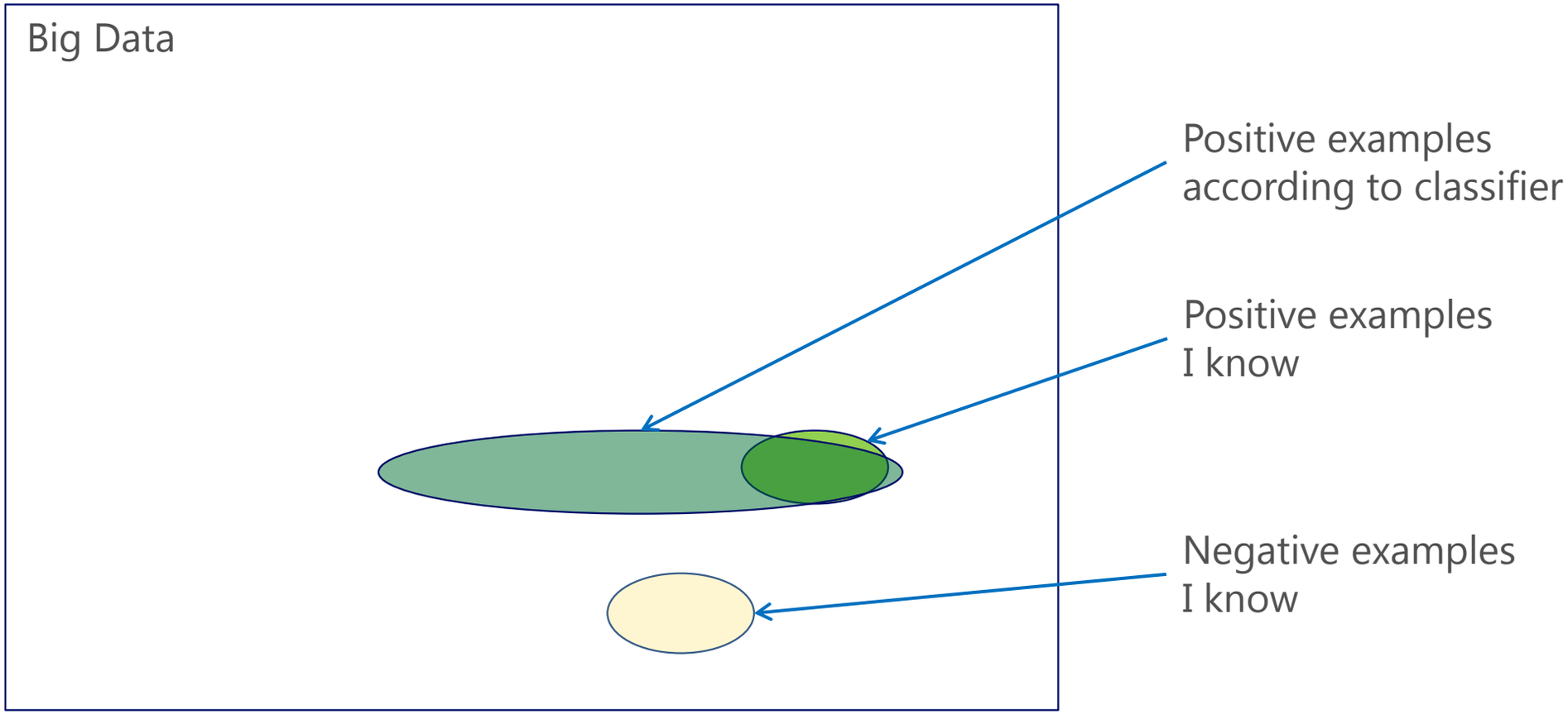
Each classifier can take weeks to reach adequate accuracy.

Errors discovered during deployment are unpredictable and costly.

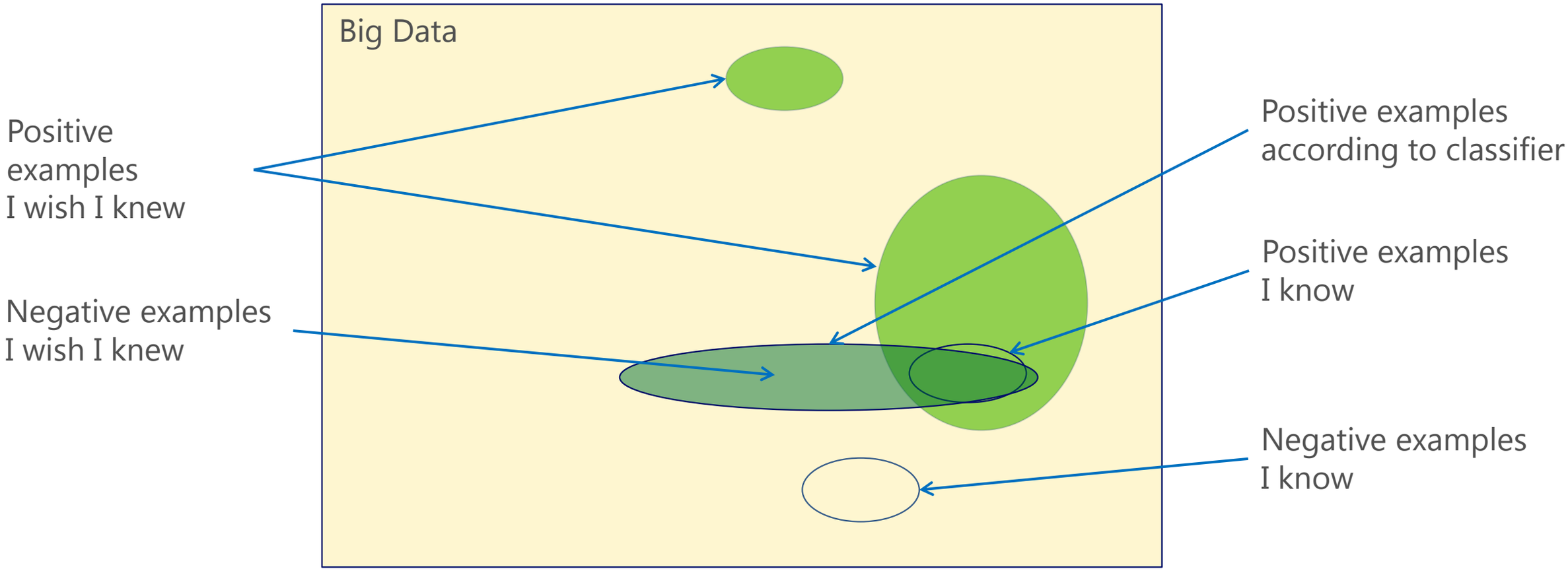
Big Data Picture



Big Data Picture



Big Data Picture



Solution: Interactive ML

Leverage ordinary humans for semantic information

We can design a system that leverages human contributions.

- Collect large sample of data (e.g. 100 million)
- For each classifier or entity extractor:
  - Repeat until good enough (cycle time is seconds or minutes)
    - Sample large data set using classifier scores
    - User provides a few labels or feature edits
    - Classifier is retrained on labeled data with new features
    - Rescore whole set with new classifier
  - Deploy

Take ML expert out of the loop.

No prior ML or engineering expertise is required to build classifiers or entity extractors.

With instant feedback loop, user quickly becomes as good or better than ML expert.

Solution: Interactive ML (Continued)

Shared infrastructure.

The data is collected once (an engineer is required for this step).
Features and classifiers are created on demand and shared across a community.
Labels and features are the asset.
Infrastructure could be offered as a service.

No surprises

The user continually interacts with a large data set that has the same distribution as the real world.
Deployment is outside the development loop. It is predictable and happens once.

ICE (Interactive Classification & Extraction)

Infrastructure

In-memory column store (7.7 TB of RAM)

60 machines (score 50M web pages in 2 sec)

UX

Labeling: User can sample (see figure) or query

Featuring: User provides list of words (dictionaries) representing concepts.

ML

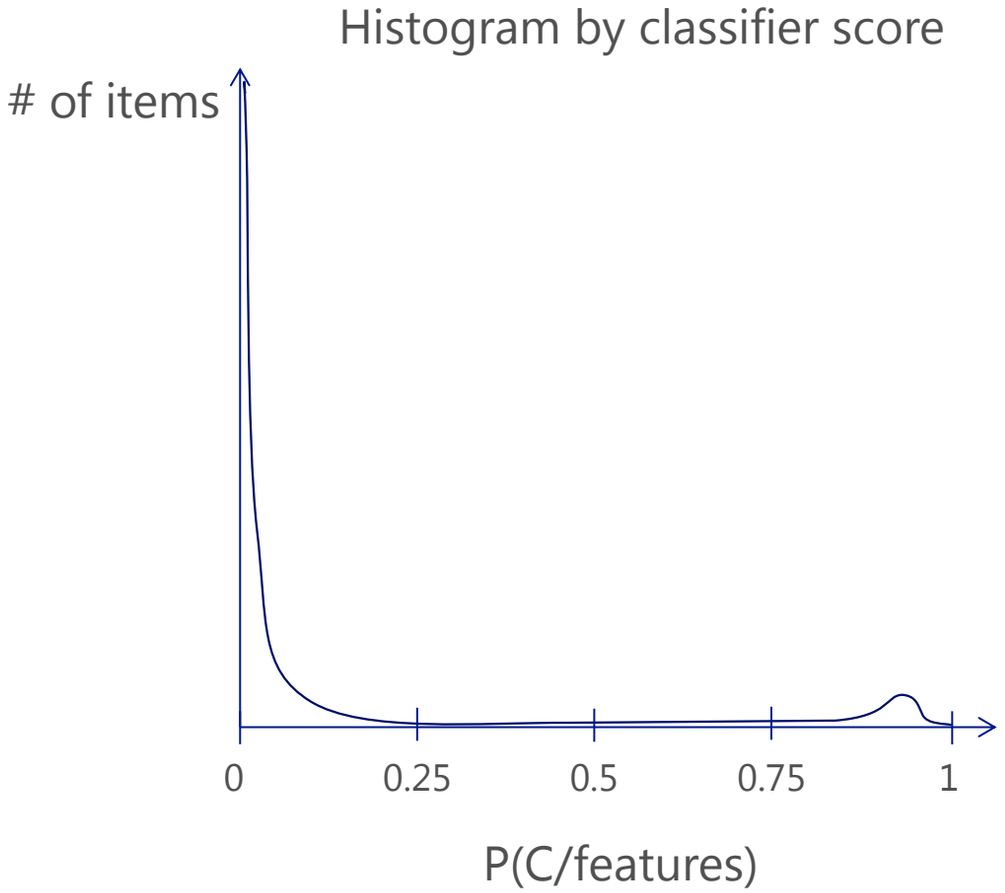
Simple classifiers (logistic regression)

Regularization (architecture and features)

Progress is measured using reachability

Demo

How to be robust against what we do not know?
Answer: Make exploration part of the design.



Challenges for interactive ML on big data

Sampling Lopsided data (UX & ML)

Exploration and cold start
Measuring progress (we cannot compute recall).

Featuring & Debugging (ML & UX)

Simple user defined features and cold start (limits accuracy).
Regularization (must be automatic, must complement feature simplicity).
Generalization (errors must be predictable and actionable).
Modularity and schemas (for entity extraction).

Infrastructure (System & UX)

Fast response time (seconds) for re-training labeled data and re-scoring 100s millions items
Sharing of features, classifiers, and entity extractors (quality, staleness, discoverability, volume)

© 2013 Microsoft Corporation. All rights reserved. Microsoft, Windows and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on