A Platform for Large-Scale Machine Learning on Web Design

**What is Machine Learning?**
An algorithmic method that “learns” the characteristics of a domain from a “training set” of examples and makes predictions about these characteristics on new data.

**Why Machine Learning?**
With several billion pages, the Web provides a large and diverse corpus of examples. Only an algorithmic approach to example-based Web design tools is capable of accounting for this scale. However, formulaic would also not account for the diversity of Web design patterns, nor could they simulate non-rule-based dimensions – creativity, for example.

**Training Set Requirements:**
1. A page renders the same way every time it is requested.
2. All pages within the corpus must be complete copies of their online counterparts: all content they request must be stored in our corpus, and source code cannot be modified.
3. A representation of a page that captures the structural relationships in its visual layout.
4. The corpus must be scalable and preserving the graph structure of the Web (links between pages and their resources) is key.

**Web Crawler + Proxy Server**
Multithreaded web crawler contains an embedded browser configured to use a proxy server.
Proxy server sees every resource request made by a page and saves this to the corpus.
Custom HTTP Headers used to preserve page-resource links (graph structure).
Corpus-only mode serves content directly from corpus ensuring consistent rendering every time.

**Bento: Page Segmenter**
With Javascript and CSS3, it is possible to arbitrarily reposition content such that it no longer corresponds to the tree structure of the Document Object Model (DOM).
Bento “re-DOMs” a page such that nodes and the hierarchy better correspond to the visual layout of a page.
Try it online at http://bento.stanford.edu

**Crowdsourced Page Labelling**
Labels corresponding to design-oriented features would allow for learning of less concrete characteristics – e.g. “minimalist” or “elegant.”
Built an interface to crowdsource application of labels to a page. labels to a page.

In a preliminary trial, hired designers to label 300 pages.
However, this is not scalable. Designers must have an incentive to label pages without financial compensation.