Automating the Generation of Hardware Component Knowledge Bases

Luke Hsiao, Sen Wu, Nicholas Chiang, Christopher Ré, and Philip Levis
lwhsiao@stanford.edu | LCTES 2019

Introduction

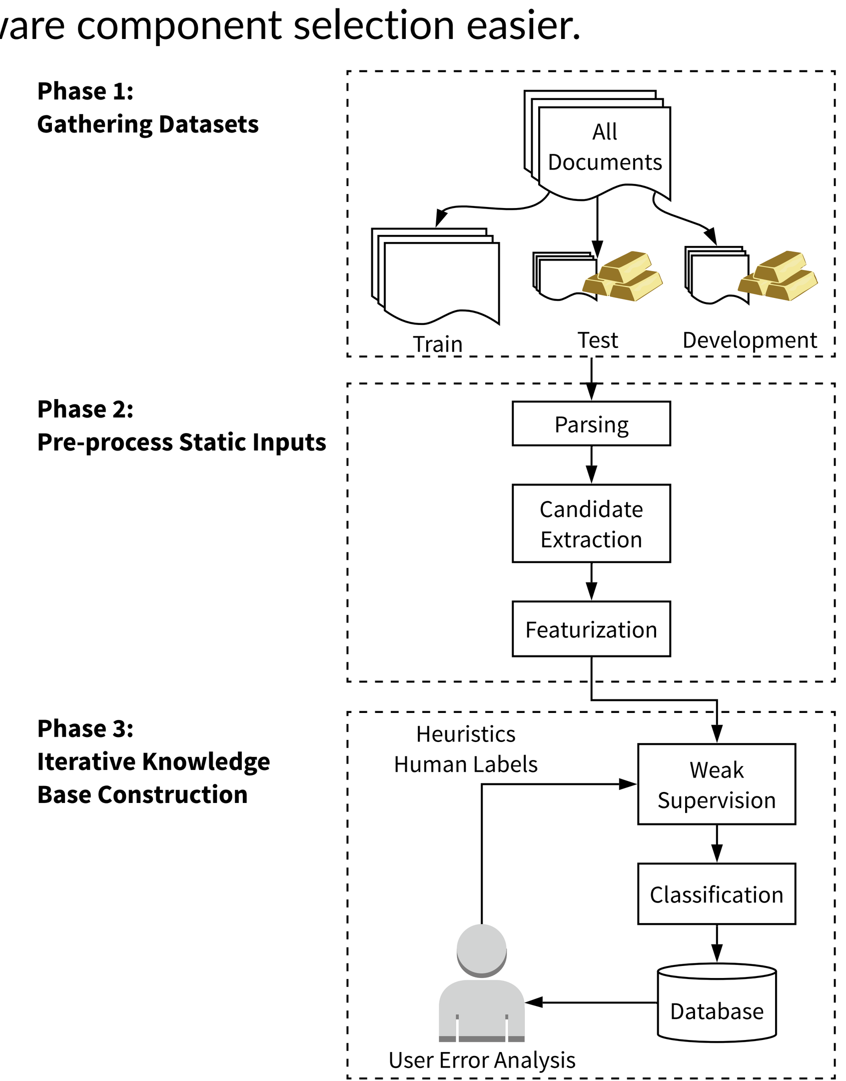
Embedded system design productivity benefits from hardware component information that is both available and accessible. Ultimately, troves of hardware component data is inaccessibly locked away in datasheets. Challenges of building databases from PDFs include:

- **Relational Data**: Users query quantitative values, where traditional text-search is insufficient.
- **Technical Jargon**: Datasheets assume a technical reader, which precludes untrained crowdsourcing.
- **Input Format Variety**: Data is presented in many formats and styles using multiple data modalities.

Contributions

We make the following contributions:

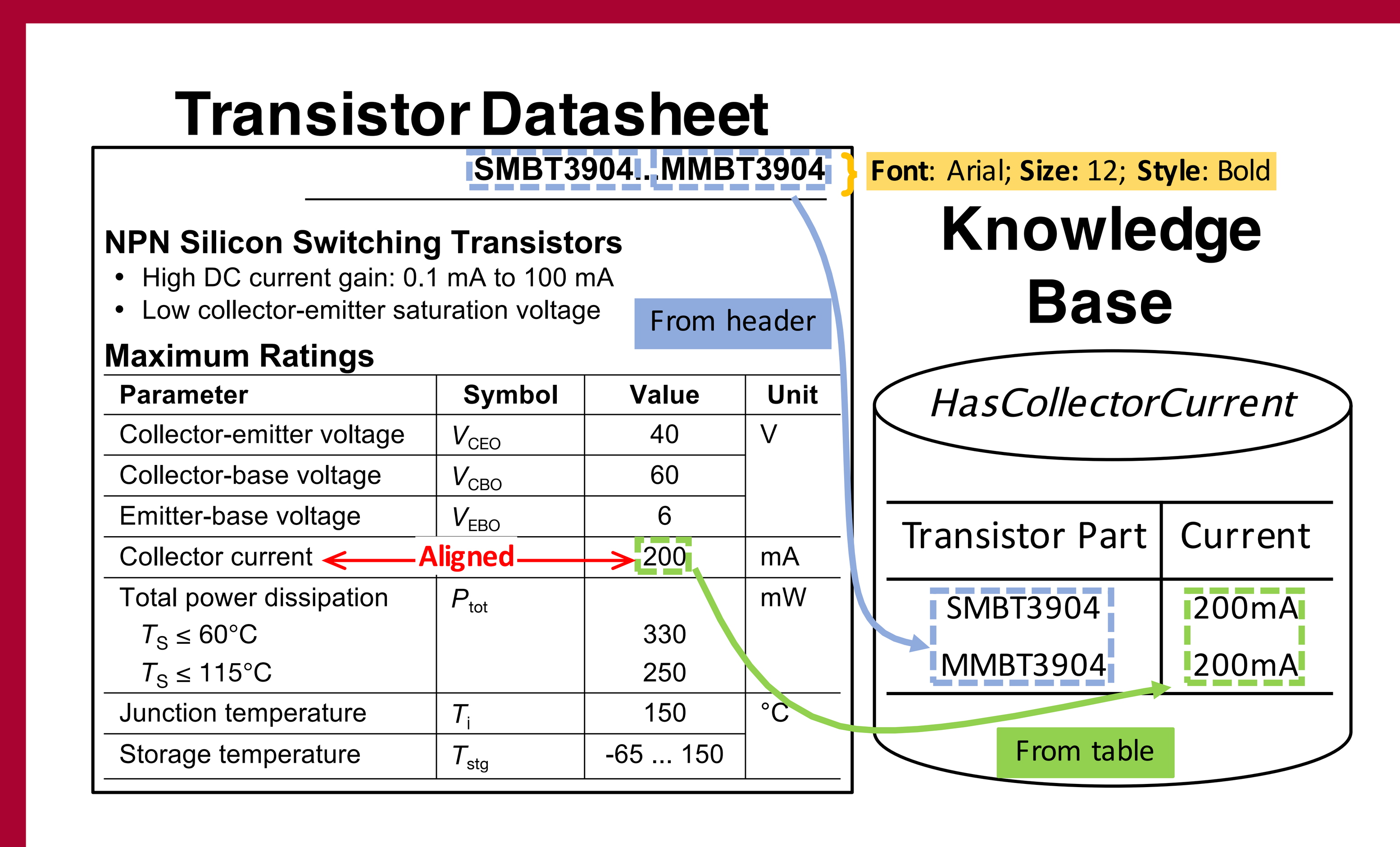
1. A **general methodology** for building hardware component knowledge bases with machine learning.
2. The evaluation on **multiple hardware components**, extracting **both textual and non-textual** information.
3. **Application studies** that show how this makes hardware component selection easier.



Summary of Results

- We leverage expertise from both heuristics and human labels to achieve an average of **75 F1 points**.
- Using a dataset of 15 000+ datasheets, we extract **multiple relations** from **multiple data modalities** such as numerical values and product thumbnails.
- On average, we **surpass Digi-Key by 12 F1 points**—improving recall 24 % at a cost of 9 % in precision.
- Our errors are often **systematic errors**, not random human errors, and more readily identified.

Enhance design tools by generating hardware component databases **at scale** with **human-like quality** using machine learning.



Take a picture
to download the full paper or visit
https://doi.org/10.1145/3316482.3326344

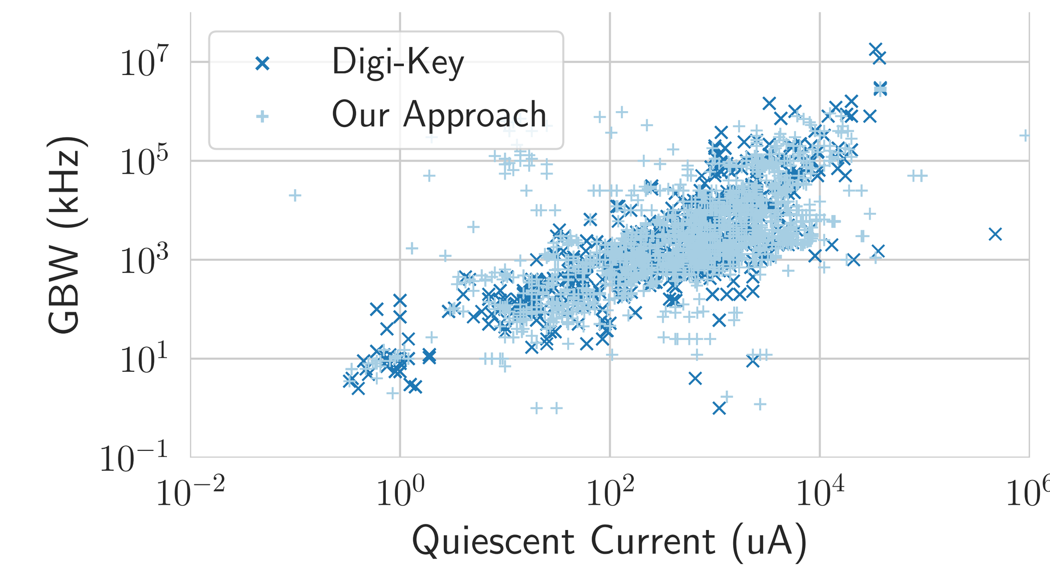
Summary of our Datasets

Dataset	Size	#Docs	#Pgs/Doc	#Rels
Bipolar Junction Transistors	3 GB	6.9 k	5.5	4
Circular Connectors	3 GB	5.1 k	3.2	1
Operational Amplifiers	5 GB	3.3 k	23.3	2

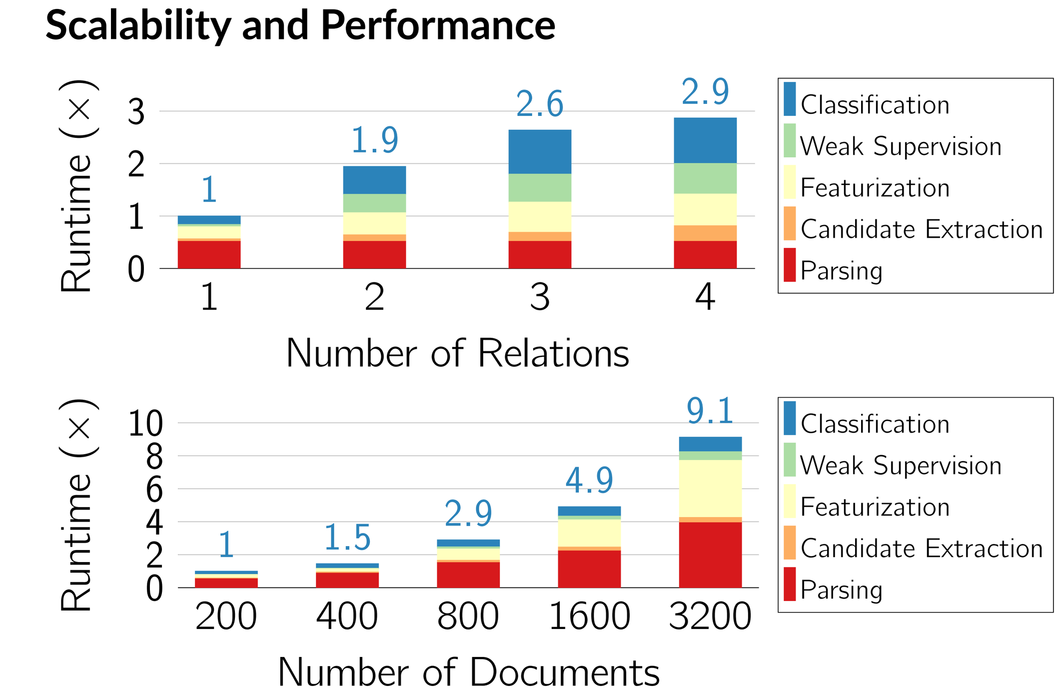
End-to-end Quality vs. Digi-Key

Relation	Source	Prec.	Rec.	F1
Polarity	Digi-Key	**1.00**	0.67	0.80
	Our Approach	0.94	**0.94**	**0.94**
Max Collector-Emitter Volt.	Digi-Key	**0.97**	0.67	**0.79**
	Our Approach	0.75	**0.77**	0.76
Gain Bandwidth Product	Digi-Key	**0.91**	0.62	0.74
	Our Approach	0.88	**0.84**	**0.86**
Quiescent Current	Digi-Key	**0.93**	0.45	0.61
	Our Approach	0.89	**0.80**	**0.84**

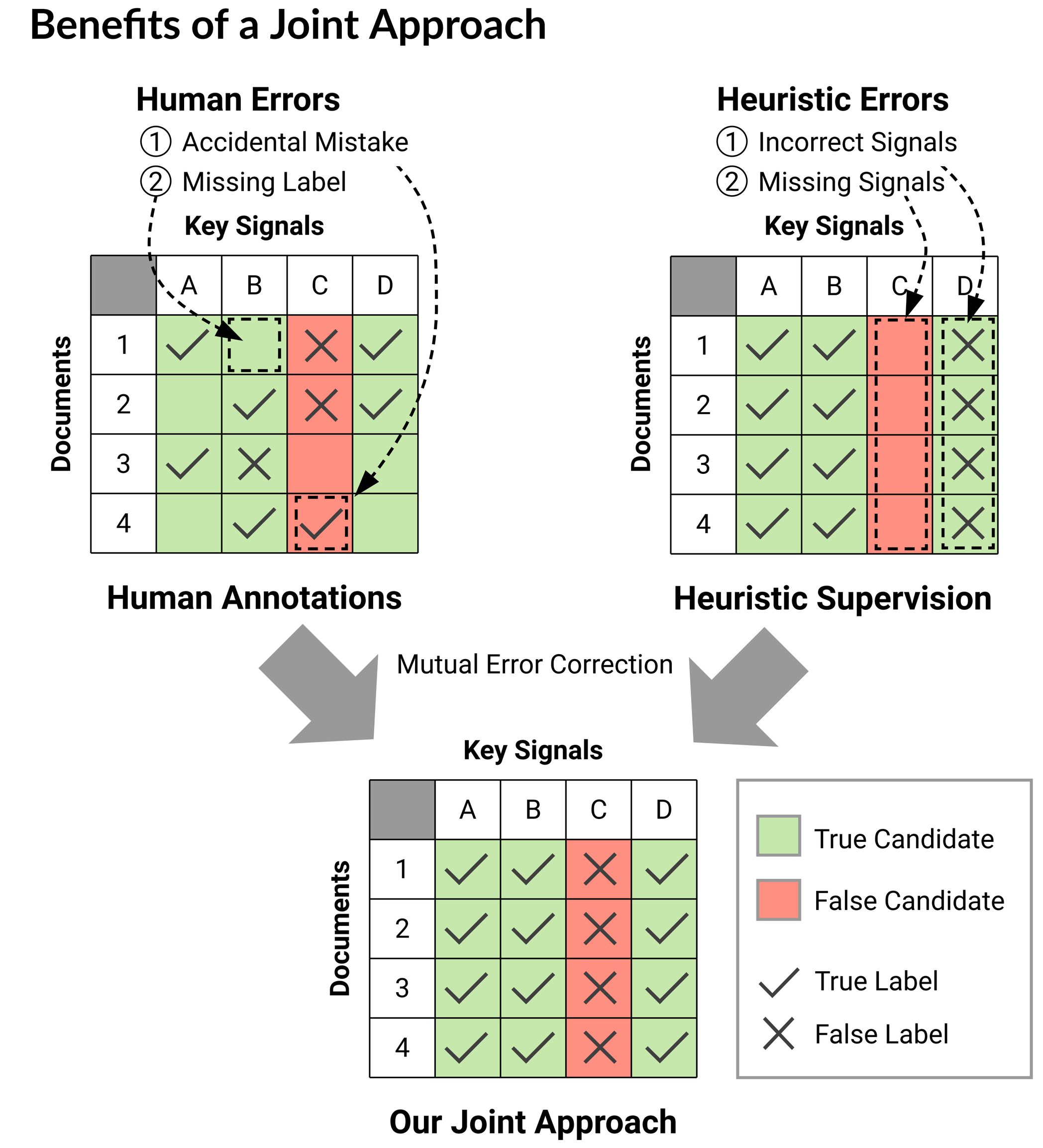
Enabling Design Applications



Scalability and Performance



Benefits of a Joint Approach



Stanford