Fonduer: Knowledge Base Construction from Richly Formatted Data

Sen Wu

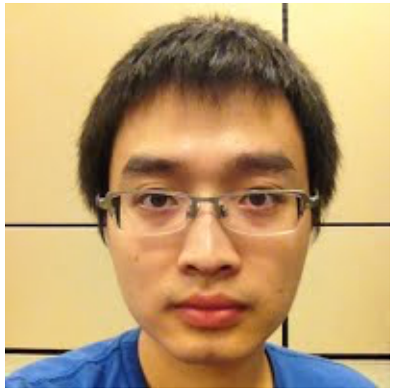
Stanford University

Thank you!

FONDUER

My Amazing Team:

Luke Hsiao



Xiao Cheng

Braden Hancock

Theodoros Rekatsinas

Philip Levis

Christopher Ré

STANFORD DAWN

STANFORD INFOLAB

My Awesome Collaborators:

DISTRICT ATTORNEY NEW YORK COUNTY

DARPA MEMEX

accenture

DS3 Sciences, Systems, Services

HITACHI Inspire the Next

Stanford MEDICINE

SING Stanford Information Networks Group

Macrostrat Lab

Alibaba Group

Knowledge bases are everywhere…

Unstructured Information

Knowledge Base Construction

Structured Knowledge Base

KNOWLEDGE VAULT

yago
select knowledge

ProbKB

IBM Watson

Product Graph

TextRunner/
ReVerb

NELL

DeepDive

And many more…

But, troves of "richly formatted" information remains untapped

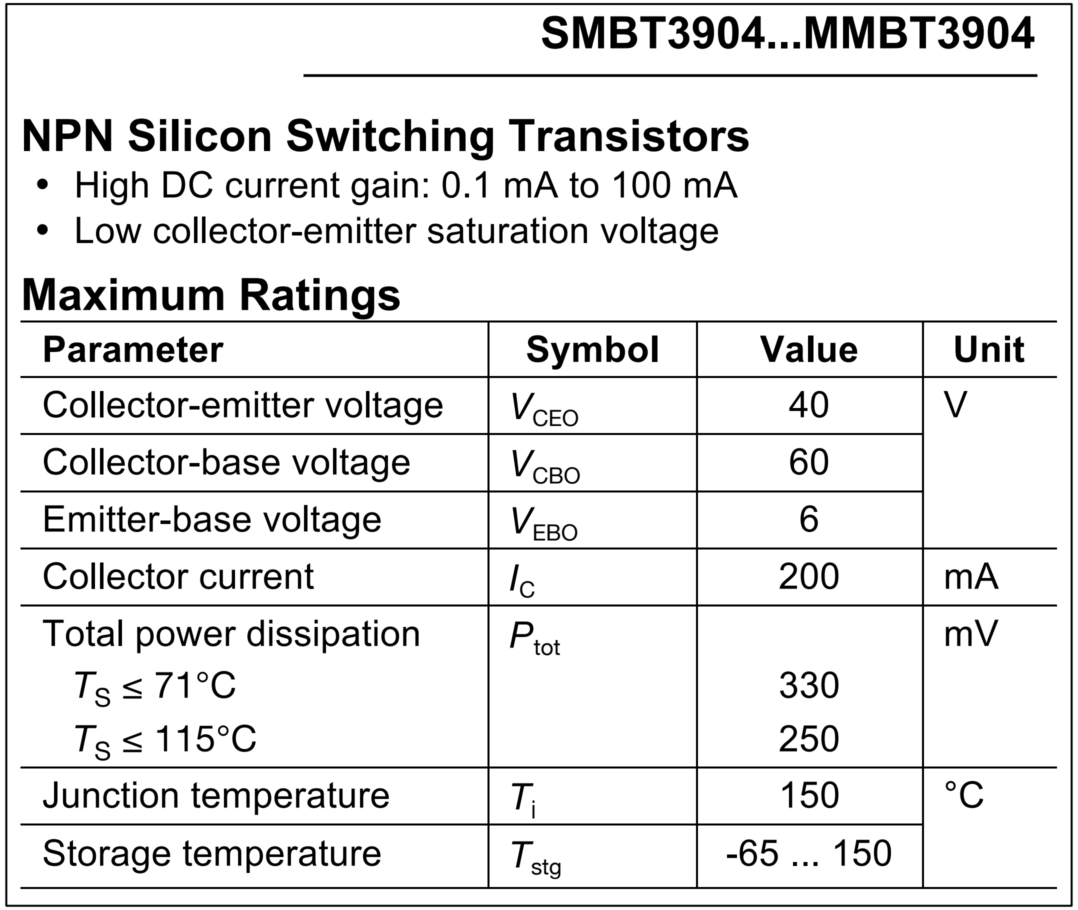
Richly formatted data

FONDUER

Richly formatted data: information is expressed via textual, structural, tabular, and visual cues.

HTML XML PDF DOC

Transistor Datasheet (PDF)



SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

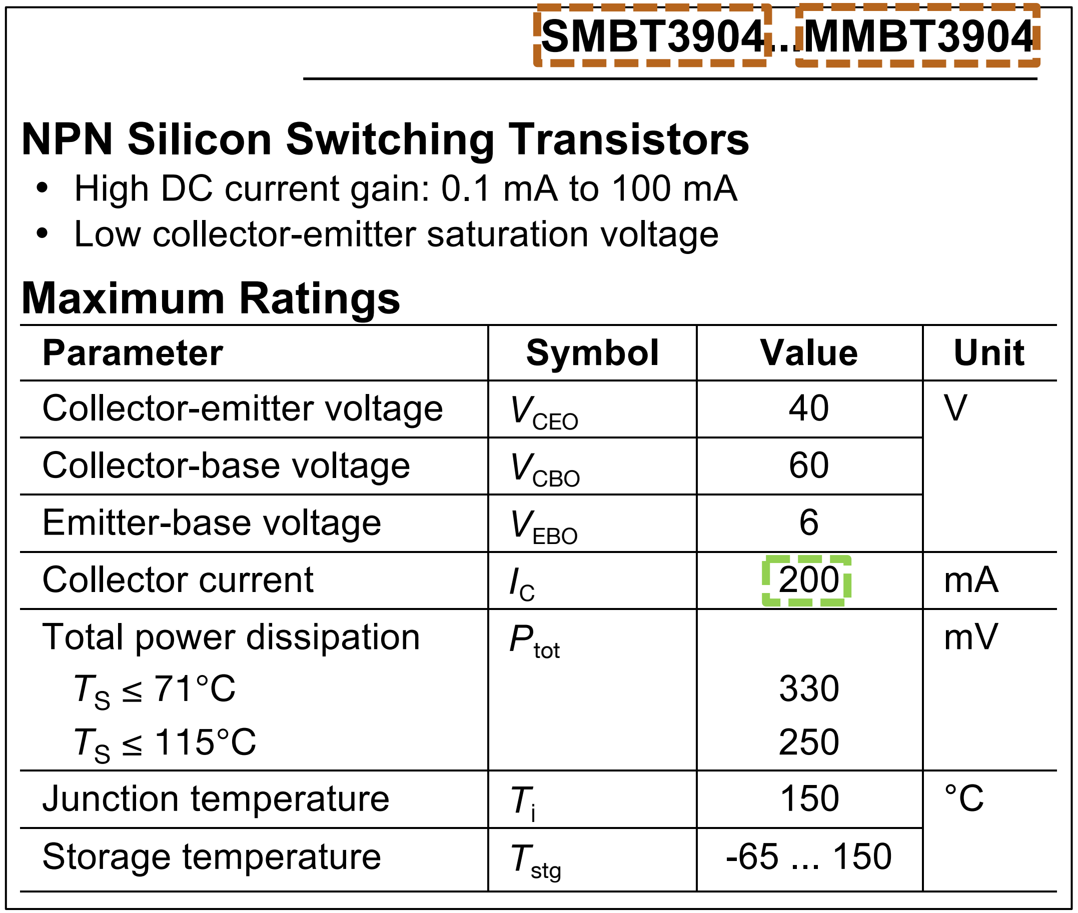
Maximum Ratings

Parameter	**Symbol**	**Value**	**Unit**
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_i	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Knowledge base construction from richly formatted data

Goal: extract maximum collector current from transistor datasheets

Transistor Datasheet



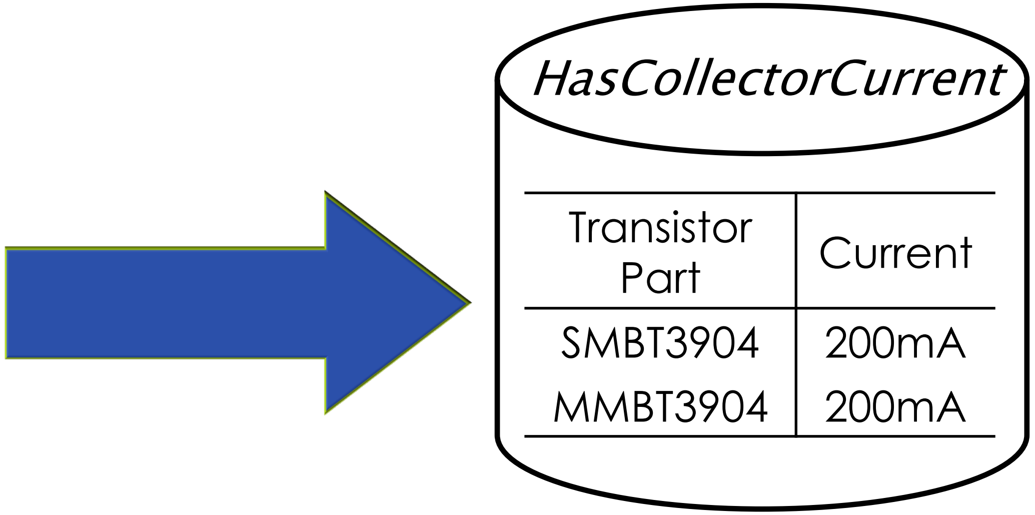
SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \le 71°C$		330	
$T_S \le 115°C$		250	
Junction temperature	T_i	150	°C
Storage temperature	T_{stg}	-65 ... 150	

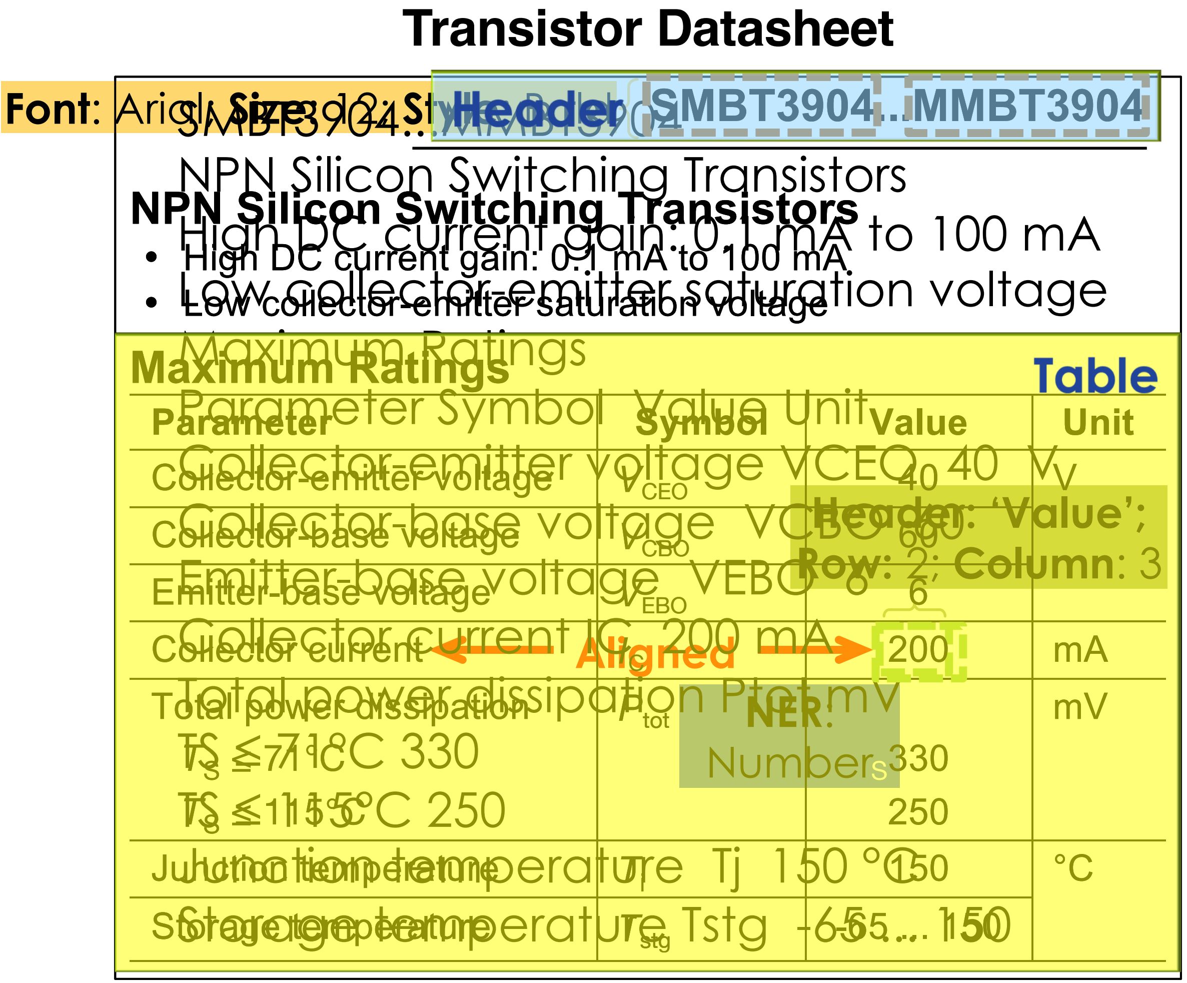


HasCollectorCurrent

Transistor Part	Current
SMBT3904	200mA
MMBT3904	200mA

Knowledge Base

Knowledge base construction from richly formatted data



Transistor Datasheet

SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \le 71$ °C		330	
$T_S \le 115$ °C		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

In richly formatted data, semantics are expressed in textual, structural, tabular, and visual modalities throughout a document

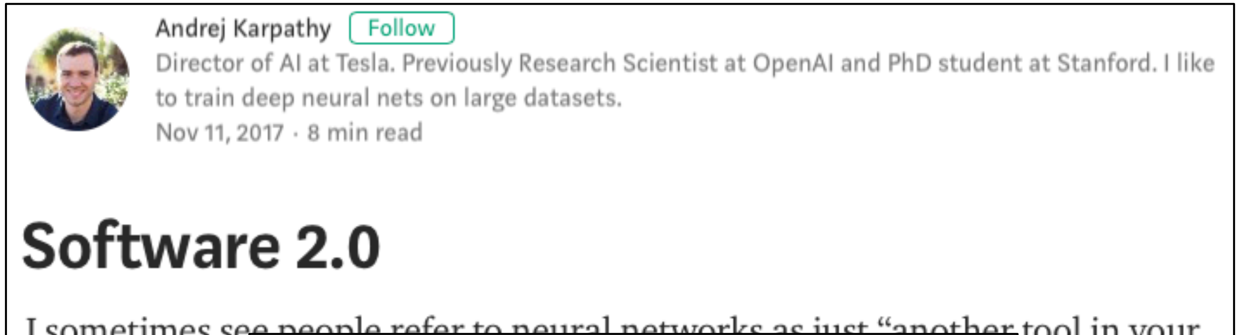
Conventional approach 1: Filter out other modalities besides unstructured text

Conventional approach 2: Limit the context scope to sentences or tables.

Problem: Misses important relations if you neglect multimodal information

Up to 97%
missed relations!

Deep learning is very successful in many domains



Andrej Karpathy Follow

Director of AI at Tesla. Previously Research Scientist at OpenAI and PhD student at Stanford. I like to train deep neural nets on large datasets.

Nov 11, 2017 · 8 min read

Software 2.0

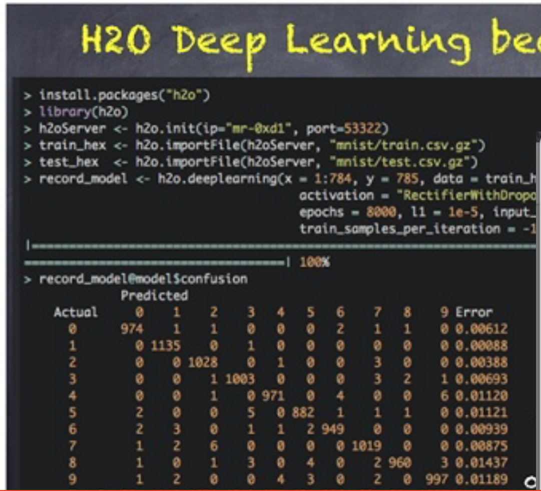
I sometimes se ... tool in your
machine learn ... work here or
there, and som ... ons.
Unfortunately,
Neural networ
of a fundamen

Alibaba's artificial intelligence bot beats humans at reading in a first for machines

A deep neural network model developed by Alibaba has scored higher than humans in a reading comprehension test, paving the way for bots to replace people in customer service jobs

PUBLISHED : Monday, 15 January, 2018, 11:33am
UPDATED : Monday, 15 January, 2018, 12:17pm

SQuAD
The Stanford Question Answering Dataset



H2O Deep Learning beats MNIST

```
> install.packages("h2o")
> library(h2o)
> h2oServer <- h2o.init(ip="mr-0xd1", port=53322)
> train_hex <- h2o.importFile(h2oServer, "mnist/train.csv.gz")
> test_hex  <- h2o.importFile(h2oServer, "mnist/test.csv.gz")
> record_model <- h2o.deeplearning(x = 1:784, y = 785, data = train_h
                                   activation = "RectifierWithDropo
                                   epochs = 8000, l1 = 1e-5, input_
                                   train_samples_per_iteration = -1
|==========================================================================
=====================================| 100%
> record_model@model$confusion
          Predicted
Actual     0    1    2    3   4   5   6    7   8   9 Error
   0     974    1    1    0   0   0   2    1   1   0 0.00612
   1       0 1135    0    1   0   0   0    0   0   0 0.00088
   2       0    0 1028    0   1   0   0    3   0   0 0.00388
   3       0    0    1 1003   0   0   0    3   2   1 0.00693
   4       0    0    1    0 971   0   4    0   0   6 0.01120
   5       2    0    0    5   0 882   1    1   1   0 0.01121
   6       2    3    0    1   1   2 949    0   0   0 0.00939
   7       1    2    6    0   0   0   0 1019   0   0 0.00875
   8       1    0    1    3   0   4   0    2 960   3 0.01437
   9       1    2    0    0   4   3   0    2   0 997 0.01189
```

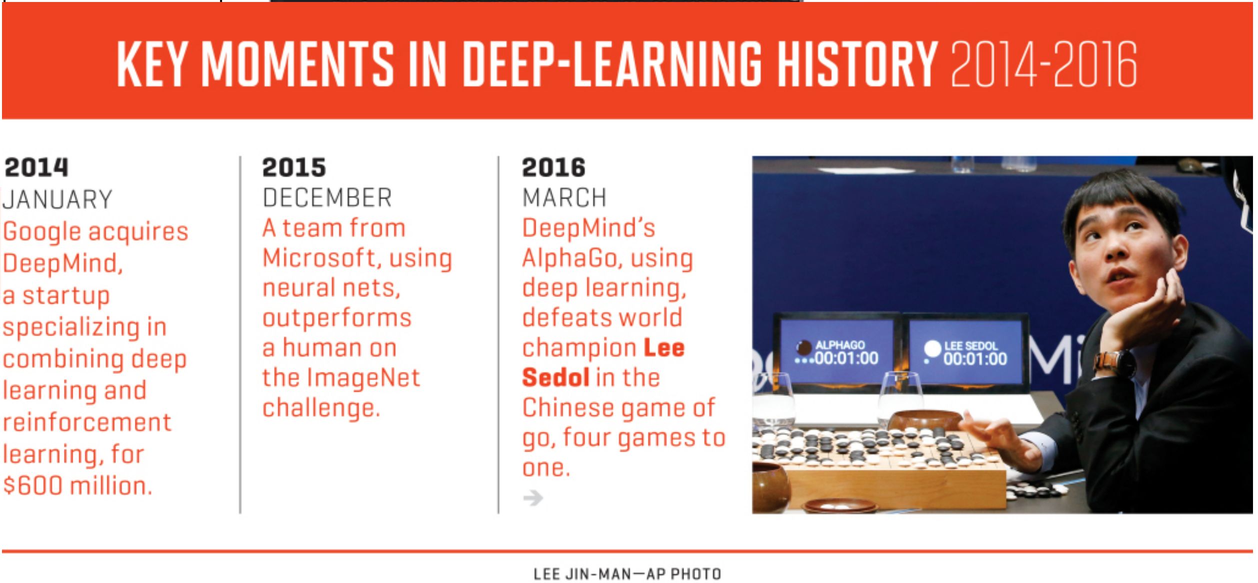
François Chollet
@fchollet
Following

It is my impression that the world of deep learning *research* is starting to plateau. What's booming: deploying DL to real-world problems.

11:19 AM - 9 Sep 2017

186 Retweets 484 Likes

26 186 484



KEY MOMENTS IN DEEP-LEARNING HISTORY 2014-2016

2014
JANUARY
Google acquires DeepMind, a startup specializing in combining deep learning and reinforcement learning, for $600 million.

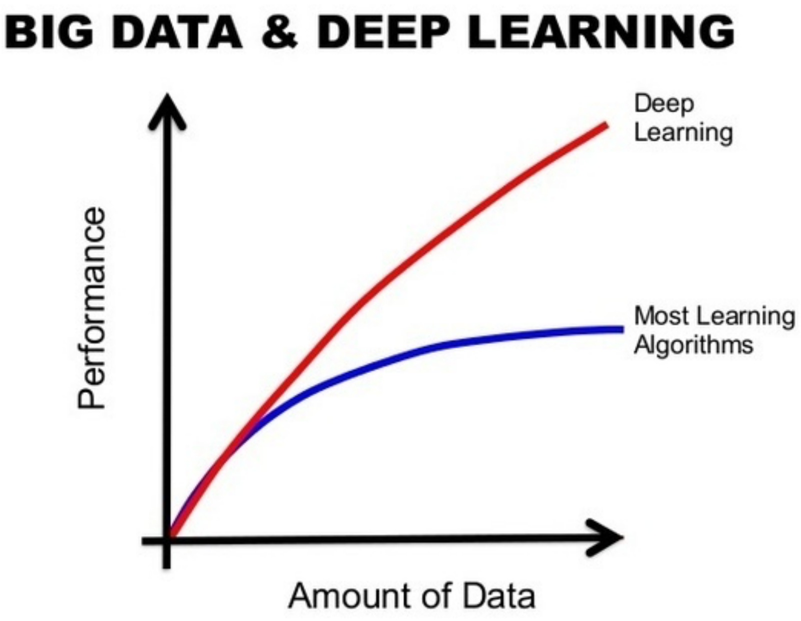
2015
DECEMBER
A team from Microsoft, using neural nets, outperforms a human on the ImageNet challenge.

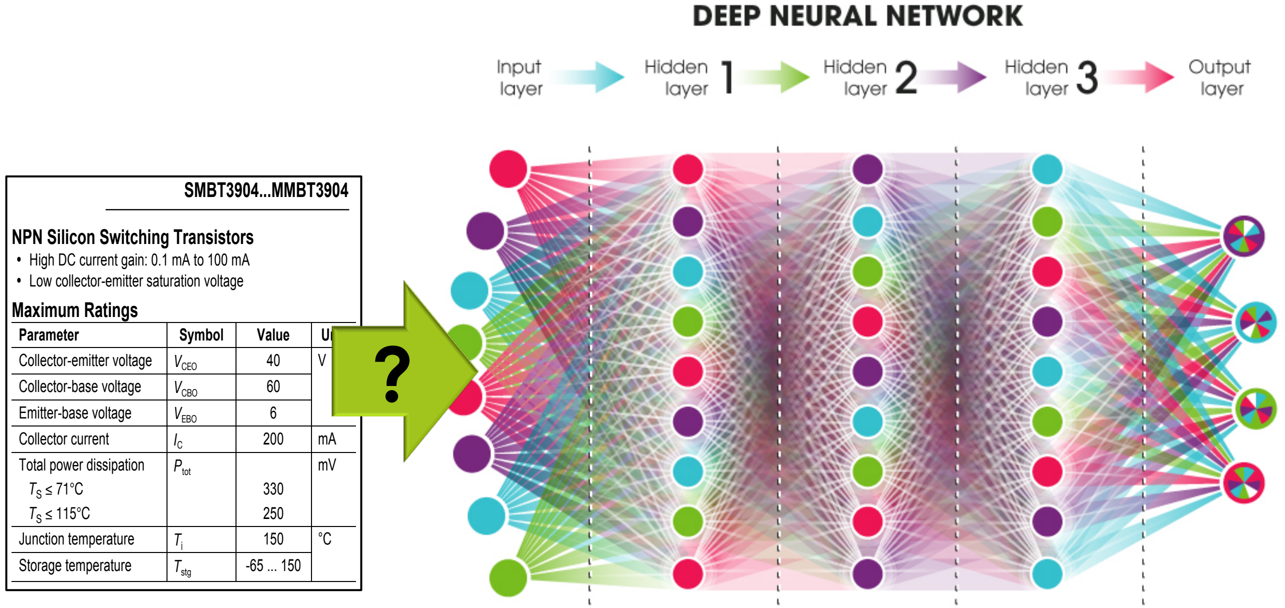
2016
MARCH
DeepMind's AlphaGo, using deep learning, defeats world champion **Lee Sedol** in the Chinese game of go, four games to one.
→

LEE JIN-MAN—AP PHOTO

Can we take advantage of this powerful tool and apply it to our problem?

Keys to utilizing deep learning





SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	U
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation $T_S \leq 71°C$ $T_S \leq 115°C$	P_{tot}	 330 250	mV
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

How do we gather enough labeled, richly formatted data?

How do we model the characteristics of richly formatted data in DL?

FONDUER

Fonduer

A *weakly supervised* deep learning framework for knowledge base construction from richly formatted data

Fonduer in practice!

FONDUER

DARPA MEMEX

accenture

Alibaba Group

Stanford MEDICINE

SING
Stanford Information Networks Group

Macrostrat Lab

Anti-Human Trafficking

司法拍卖 公开·公平·公正 sf.taobao.com

【第一次拍卖】浙A72R67小型普通轿车一辆

34,000

00 03 24 33.9

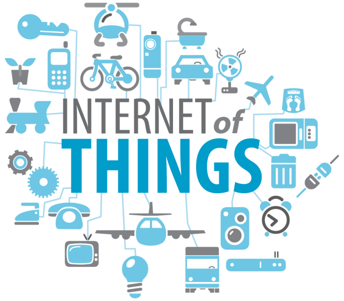
34,500

15 | 176 | 3923

竞买公告 | 竞买须知 | 标的物介绍 | 竞买记录 (17) | 大家都在问

Search Engine

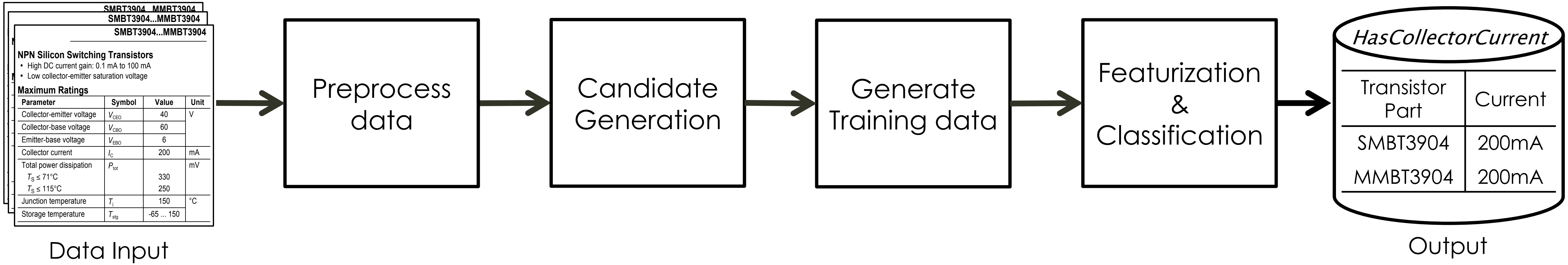
Genome-wide
Association Studies



Internet
of Things

Paleontology

Fonduer pipeline



SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \le 71°C$		330	
$T_S \le 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

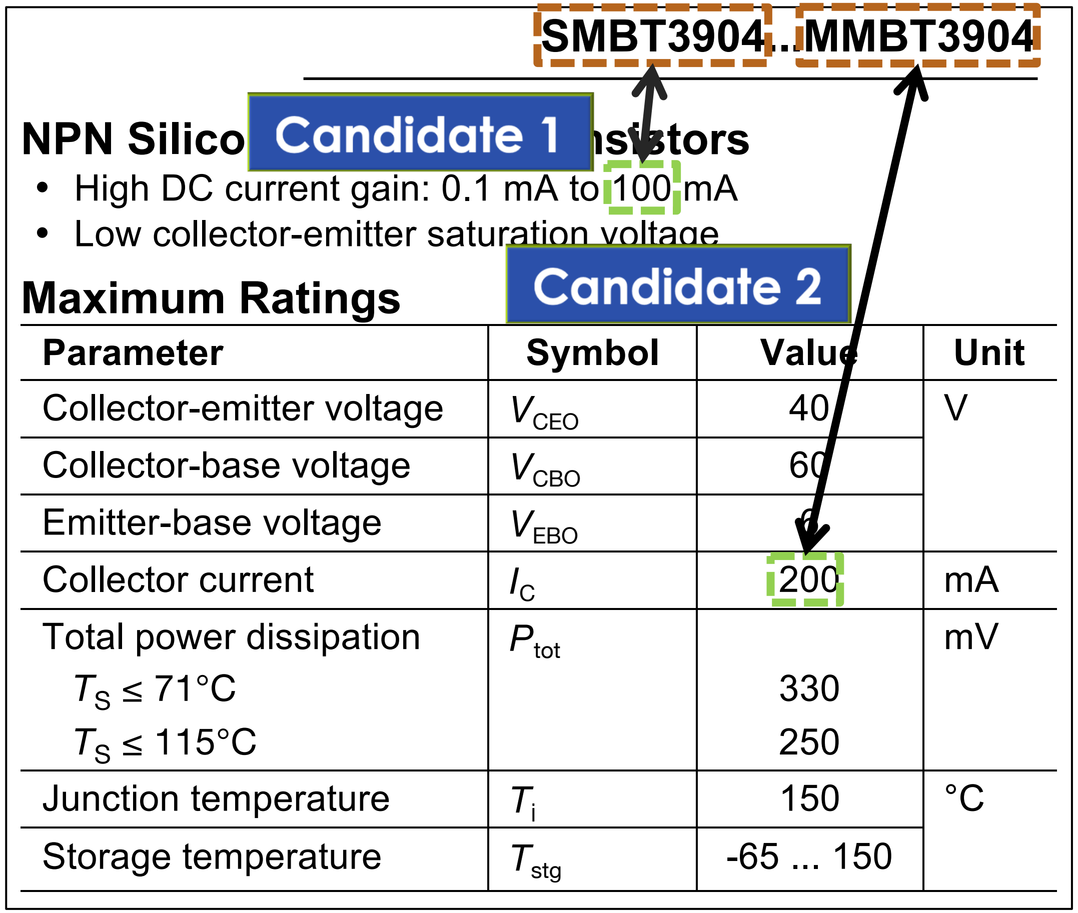
HasCollectorCurrent

Transistor Part	Current
SMBT3904	200mA
MMBT3904	200mA

Generating richly formatted training data

Multimodal weak supervision

Transistor Datasheet



SMBT3904...MMBT3904

NPN Silico Candidate 1 nsistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings **Candidate 2**

Parameter	**Symbol**	**Value**	**Unit**
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation $T_S \leq 71°C$ $T_S \leq 115°C$	P_{tot}	 330 250	mV
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Doc. level Candidates		Supervision	
		Manual	Labeling function
SMBT3904	100	✗	✗
MMBT3904	200	✓	✓

Weak supervision: express any supervision signal via labeling functions to generate training data

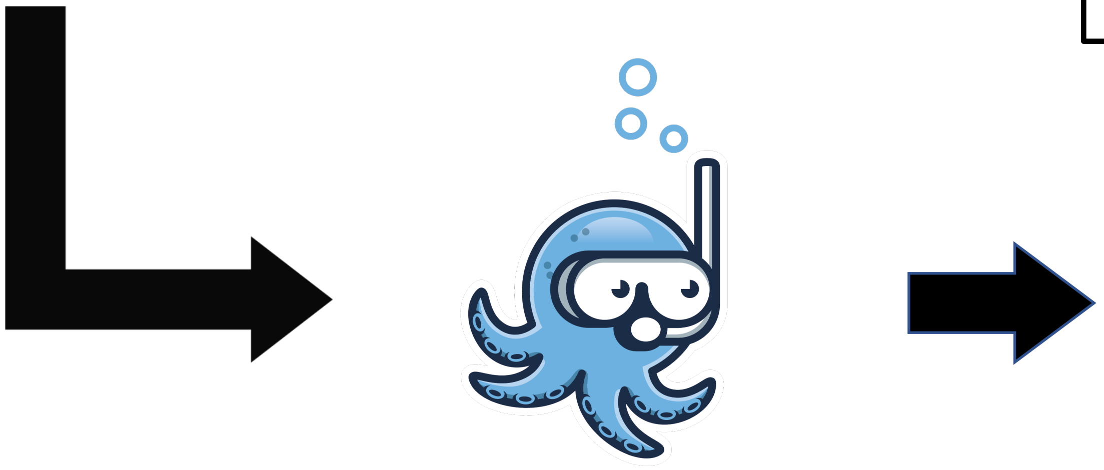
```
# Check if current is in the same row with keyword `collector`
def in_the_same_row_with(candidate):
  if 'collector' in row_ngrams(candidate.current):
    return 1
  else: return -1
```

Modeling Weak Supervision

Doc. level Candidates		Multimodal Supervision		
		Vertically aligned with 'Value'	Row ngrams contain 'mA'	'current' in sentence
SMBT3904	100	✗	∅	✓
SMBT3904	200	✓	✓	✗
SMBT3904	150	✓	✗	✗

∅ =Abstain

Intuition: Use agreements / disagreements to learn the accuracy of LFs without ground truth



Data programming/MeTal

Output: Probabilistic Training Labels

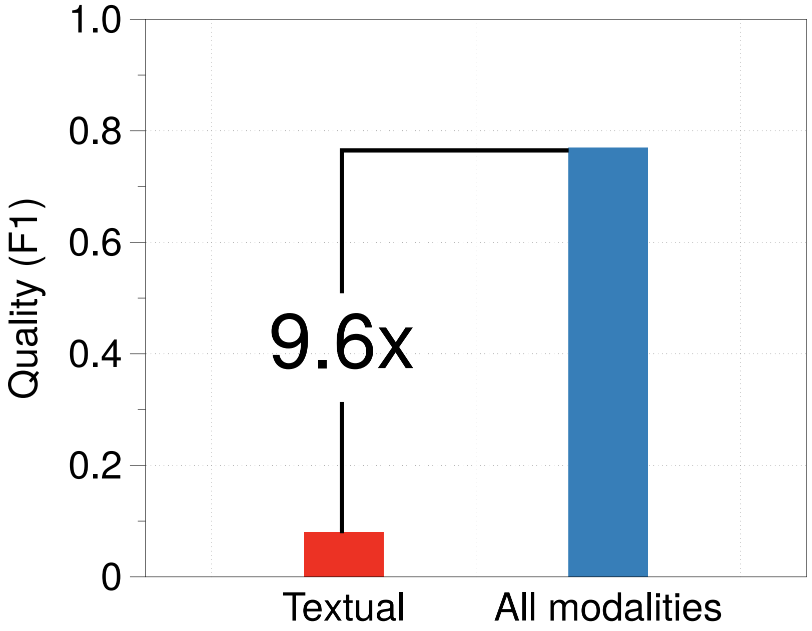
SMBT3094	100	0.5
SMBT3094	200	0.85
SMBT3094	150	0.15

Multimodal supervision is key to quality

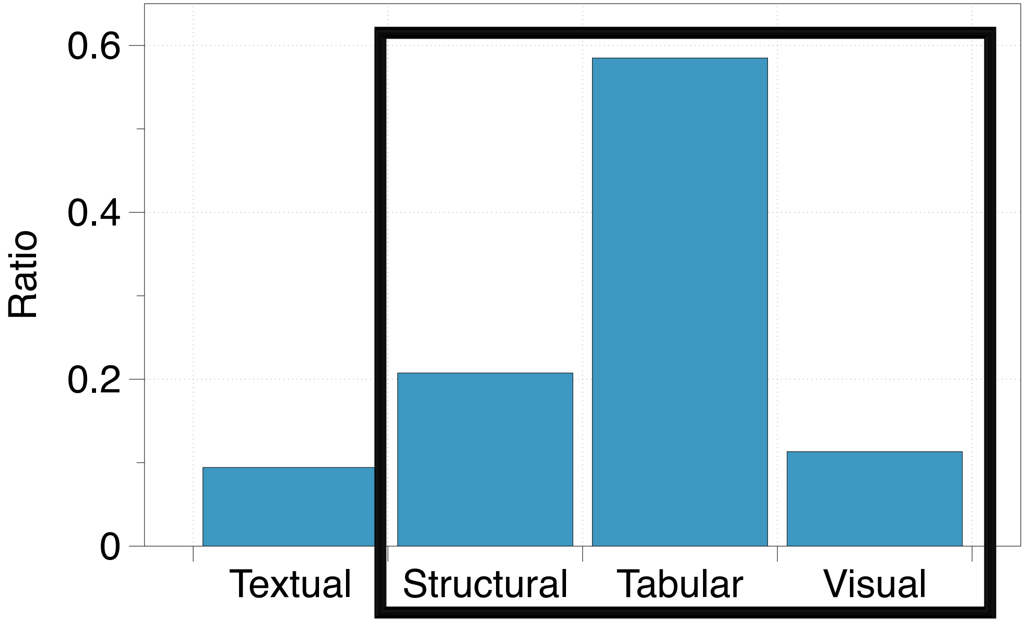
FONDUER

For transistor datasheets…

Different supervision resources' effect



Modality distribution of supervision

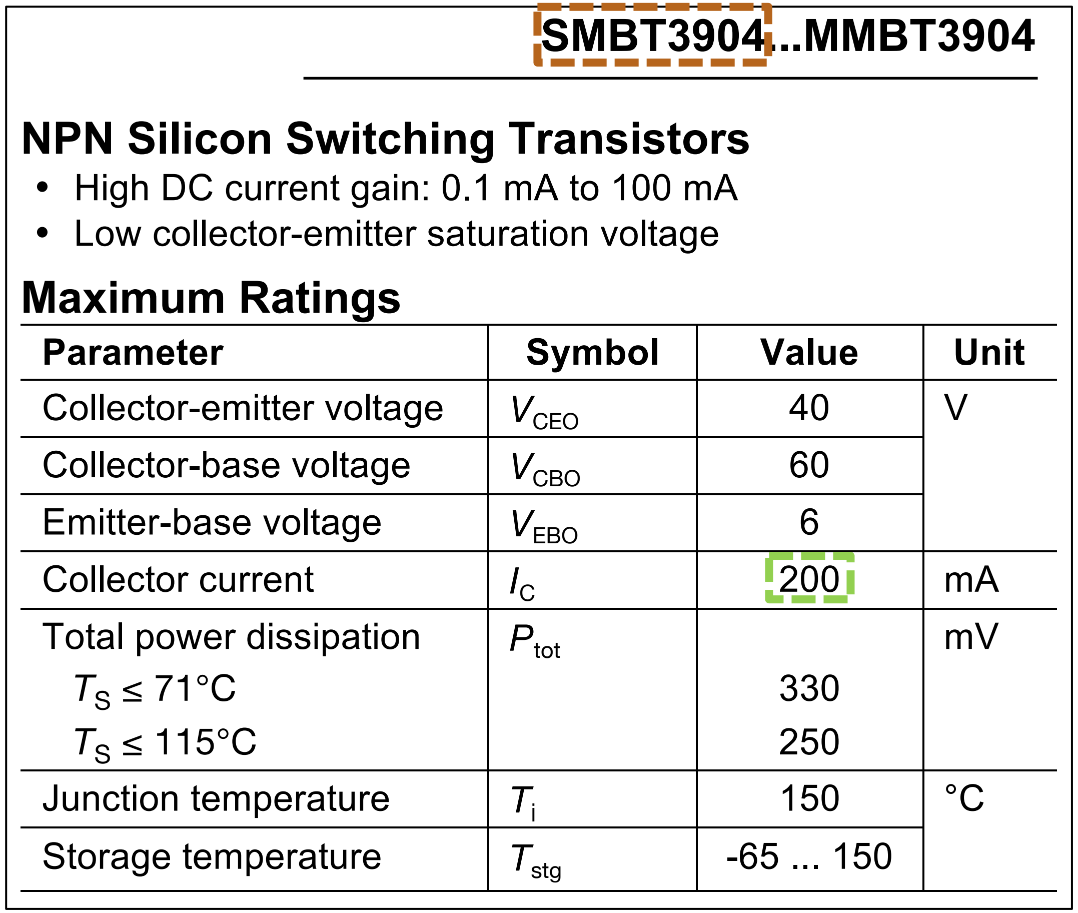


Users intuitively rely on multimodal information for supervision

Featurization and Classification for Richly Formatted Data

LSTM for Textual Information

Transistor Datasheet



SMBT3904...MMBT3904

NPN Silicon Switching Transistors

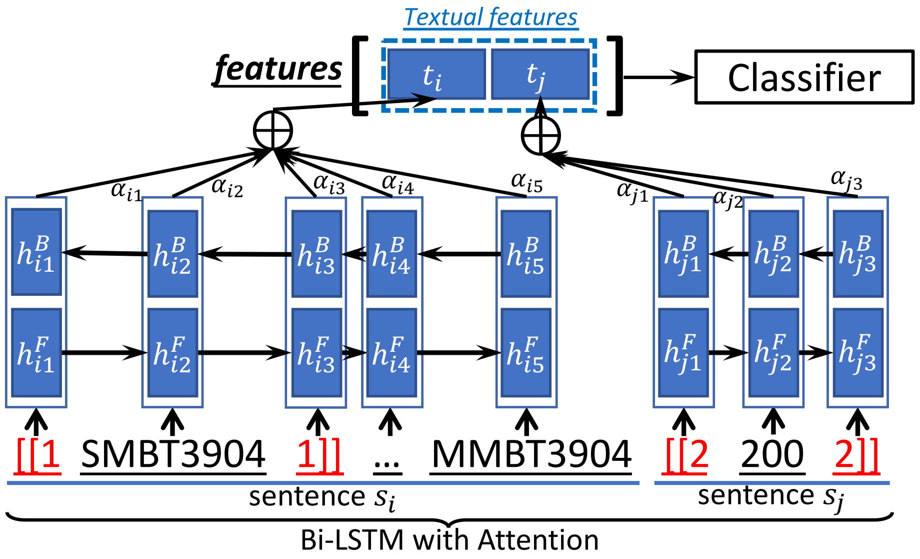
- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \le 71$°C		330	
$T_S \le 115$°C		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

LSTM excels at relation extraction from text

Xu et al., 2015; Miwa et al., 2016; Zhang et al., 2016

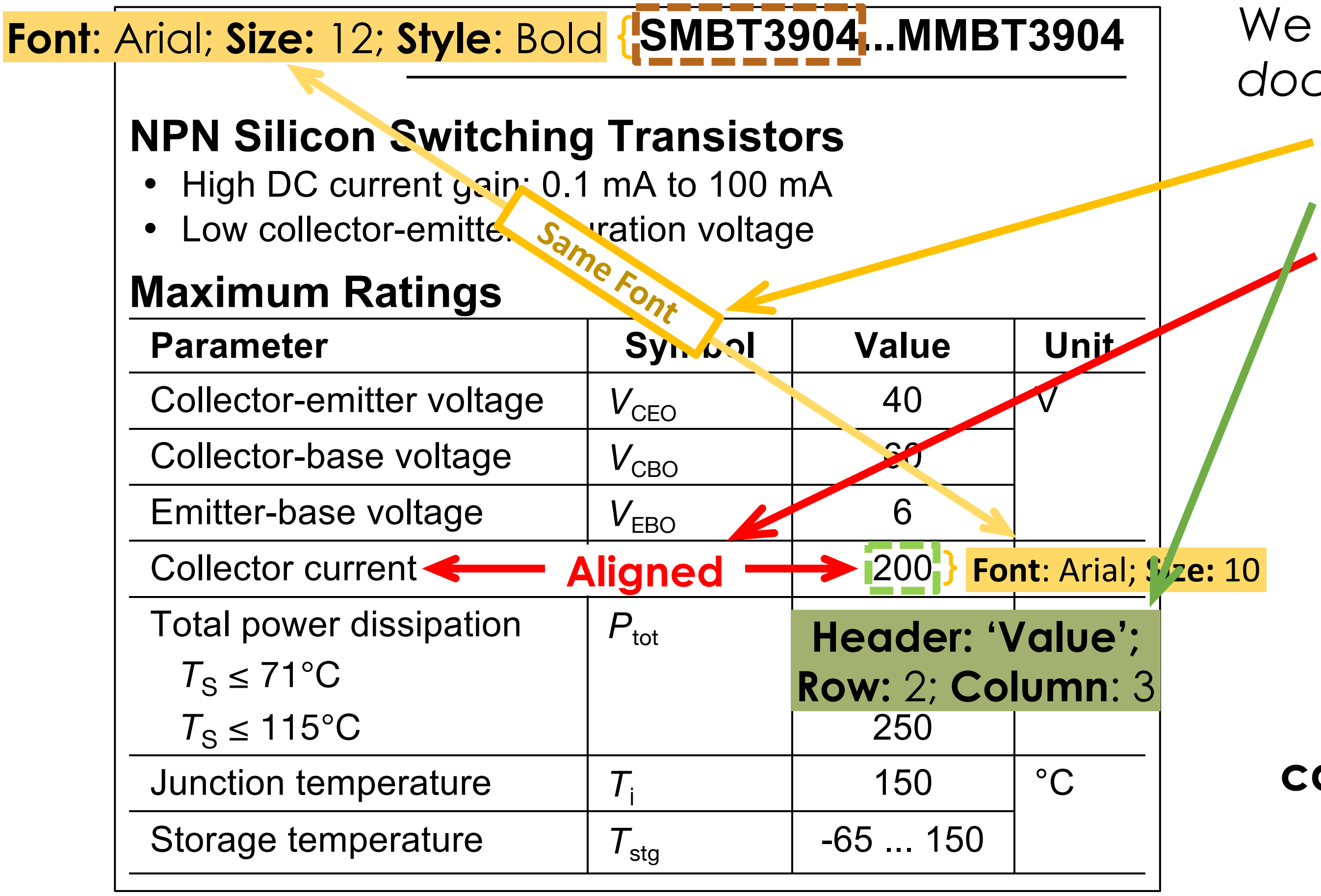

Textual features
features
t_i
t_j
Classifier
α_{i1}
α_{i2}
α_{i3}
α_{i4}
α_{i5}
α_{j1}
α_{j2}
α_{j3}
h^B_{i1} h^B_{i2} h^B_{i3} h^B_{i4} h^B_{i5} h^B_{j1} h^B_{j2} h^B_{j3}
h^F_{i1} h^F_{i2} h^F_{i3} h^F_{i4} h^F_{i5} h^F_{j1} h^F_{j2} h^F_{j3}
[[1 SMBT3904 1]] ... MMBT3904 [[2 200 2]]
sentence s_i
sentence s_j
Bi-LSTM with Attention

Problem: LSTM networks struggle to capture the multimodal characteristics of richly formatted data.

Augmenting LSTM with Multimodal Features

FONDUER

Transistor Datasheet


Font: Arial; Size: 12; Style: Bold
Same Font
Aligned
Font: Arial; Size: 10
Header: ‘Value’; Row: 2; Column: 3

SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current		200	
Total power dissipation $T_S \le 71°C$ $T_S \le 115°C$	P_{tot}	 250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

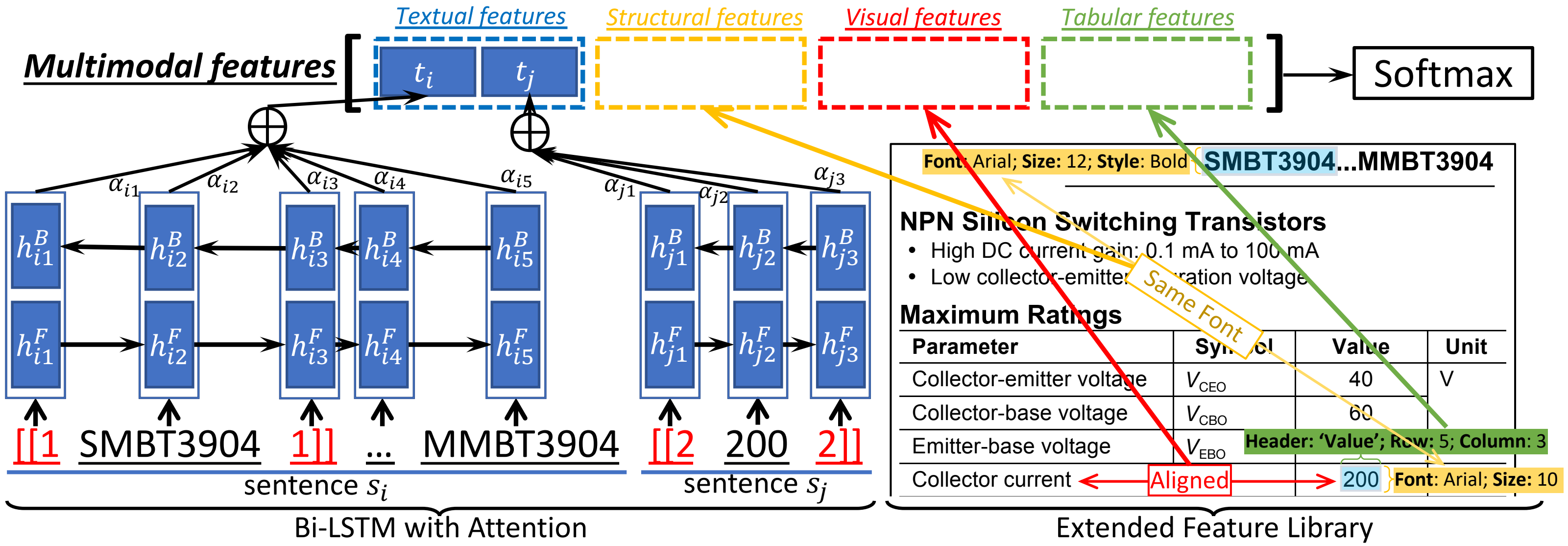
We use the multimodal information stored in the *document* to extract basic multimodal features:

- ❑ Structural features
- ❑ Tabular features
- ❑ Visual features

Augmentation with multimodal features captures signals a traditional LSTM would miss.

Fonduer's Multimodal LSTM

Signals from different modalities can be useful to find the information.



SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

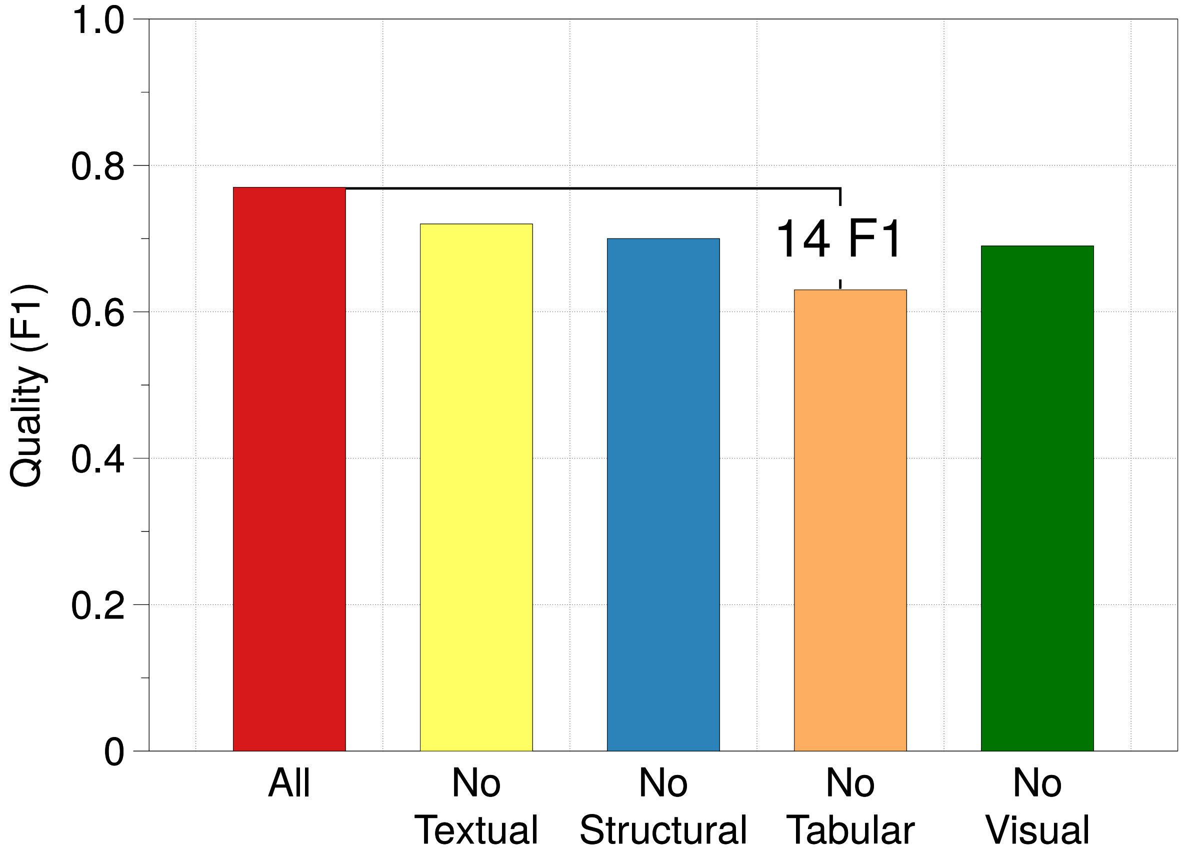
Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}		
Collector current		200	

Fonduer: a KBC system that takes advantage of both techniques to reason about all available signals.

The impact of multimodal features

For transistor datasheets…



Multimodal features significantly impact the quality of extraction

Fonduer in the wild

Empirical results & real-world uses

Fonduer vs. Human-curated Knowledge Bases

Same set of documents

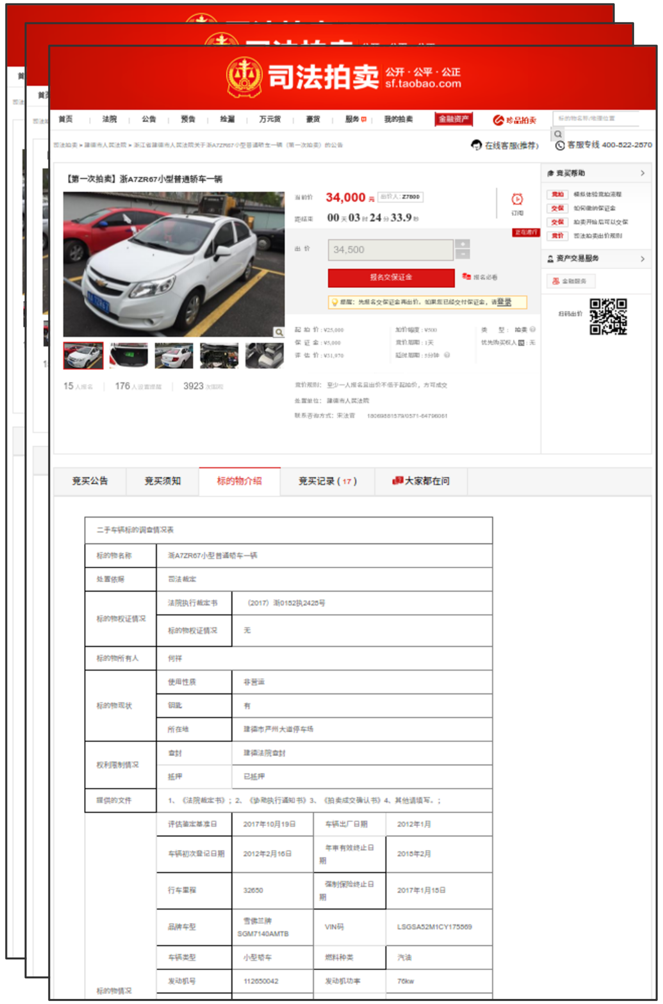
GWAS Catalog	**Fonduer**
Human-created	Machine-created
10 years	**<6** months
1.0x extractions	**1.59x** extractions
	Precision **0.89**

How people use Fonduer in industry

Input: User-customized HTML auction pages → **Output:** Structured knowledge base

Extract key facts (make, model, license, etc.)

Improve auction search quality and UX



司法拍卖 公开·公平·公正 sf.taobao.com

首页 | 法院 | 公告 | 预售 | 捡漏 | 万元贷 | 豪贷 | 服务 | 我的拍卖 | 金融资产 | 珍品拍卖

在线客服(推荐) 客服专线 400-822-2570

【第一次拍卖】浙A7ZR67小型普通轿车一辆

当前价 34,000 元 出价人：Z7800

距结束 00天03时24分33.9秒

出价 34,500

报名交保证金

起拍价：¥25,000 加价幅度：¥500 类型：拍卖
保证金：¥5,000 竞价周期：1天 优先购买权人：无
评估价：¥31,970 延时周期：5分钟

15 人报名 | 176 人设置提醒 | 3923 次围观

竞价规则：至少一人报名且出价不低于起拍价，方可成交
处置单位：建德市人民法院
联系咨询方式：宋法官 18069881579/0571-64796061

竞买帮助
竞拍 模拟体验竞拍流程
交保 如何缴纳保证金
交保 拍卖开始后可以交保
竞价 司法拍卖出价规则

资产交易服务

金融服务

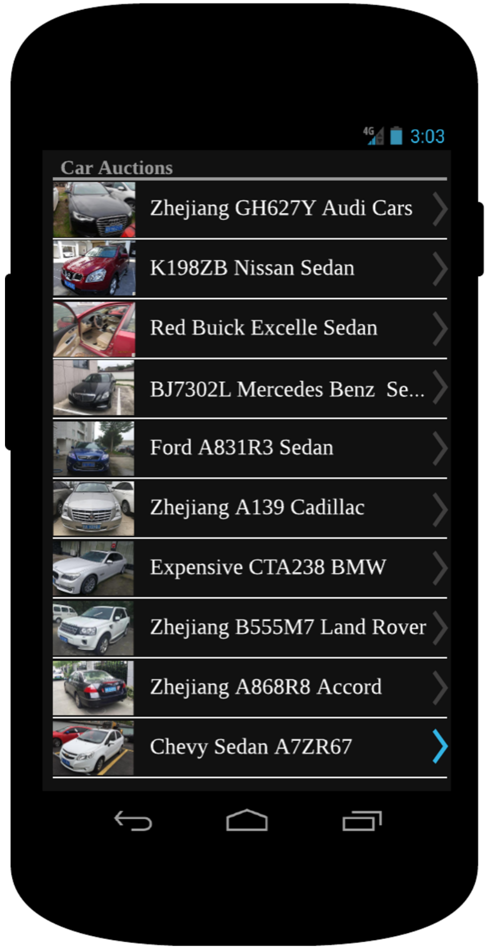
扫码出价

竞买公告 | 竞买须知 | 标的物介绍 | 竞买记录（17） | 大家都在问

二手车辆标的调查情况表				
标的物名称	浙A7ZR67小型普通轿车一辆			
处置依据	司法裁定			
标的物权证情况	法院执行裁定书	（2017）浙0182执2428号		
	标的物权证情况	无		
标的物所有人	何祥			
标的物现状	使用性质	非营运		
	钥匙	有		
	所在地	建德市严州大道停车场		
权利限制情况	查封	建德法院查封		
	抵押	已抵押		
提供的文件	1、《法院裁定书》；2、《协助执行通知书》3、《拍卖成交确认书》4、其他请填写。；			
标的物情况	评估鉴定基准日	2017年10月19日	车辆出厂日期	2012年1月
	车辆初次登记日期	2012年2月16日	年审有效终止日期	2018年2月
	行车里程	32650	强制保险终止日期	2017年1月15日
	品牌车型	雪佛兰牌 SGM7140AMTB	VIN码	LSGSA52M1CY175569
	车辆类型	小型轿车	燃料种类	汽油
	发动机号	112650042	发动机功率	76kw

Fonduer

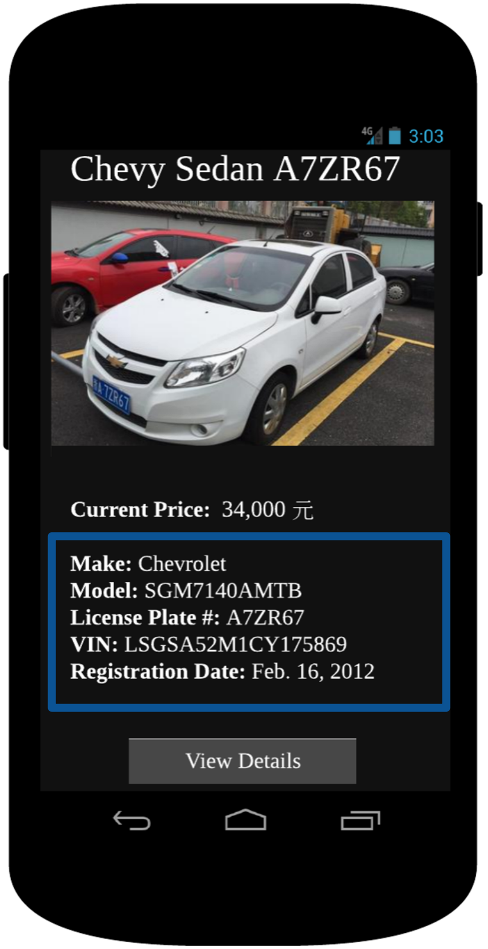
Alibaba.com



3:03

Car Auctions

- Zhejiang GH627Y Audi Cars
- K198ZB Nissan Sedan
- Red Buick Excelle Sedan
- BJ7302L Mercedes Benz Se...
- Ford A831R3 Sedan
- Zhejiang A139 Cadillac
- Expensive CTA238 BMW
- Zhejiang B555M7 Land Rover
- Zhejiang A868R8 Accord
- Chevy Sedan A7ZR67



3:03

Chevy Sedan A7ZR67

Current Price: 34,000 元

Make: Chevrolet
Model: SGM7140AMTB
License Plate #: A7ZR67
VIN: LSGSA52M1CY175869
Registration Date: Feb. 16, 2012

Knowledge Base Construction from Richly Formatted Data

- Fonduer helps build high-quality KBC from richly formatted data
- Allows users to leverage multimodal signals
- Augments deep learning model with features from each data modality to achieve high quality
- Fonduer is supporting real world applications

Thank you!
Sen Wu
(senwu@cs.stanford.edu)

https://github.com/HazyResearch/fonduer