Deep Learning for Big Code

Martin Vechev
Spring 2018

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Today

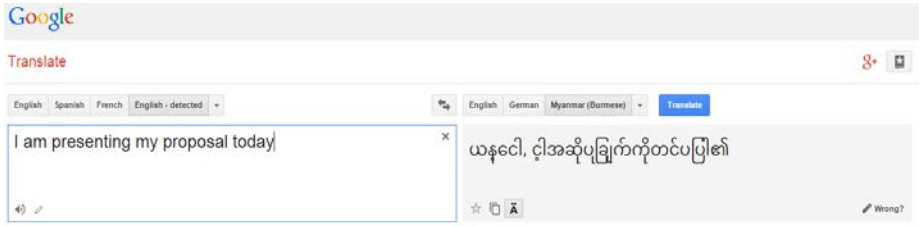
A bird's view of the area

Course organization

- objectives, your tasks, grading, howto's

Learning and probabilistic models based on Big Data have **revolutionized** entire fields

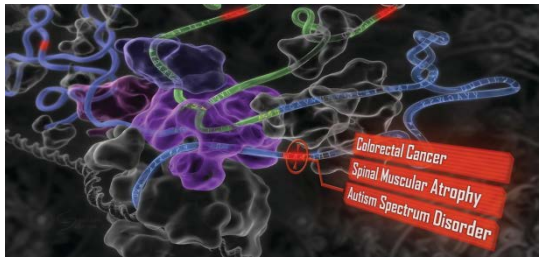
Natural Language Processing
(e.g., machine translation)



Computer Vision
(e.g., image captioning)

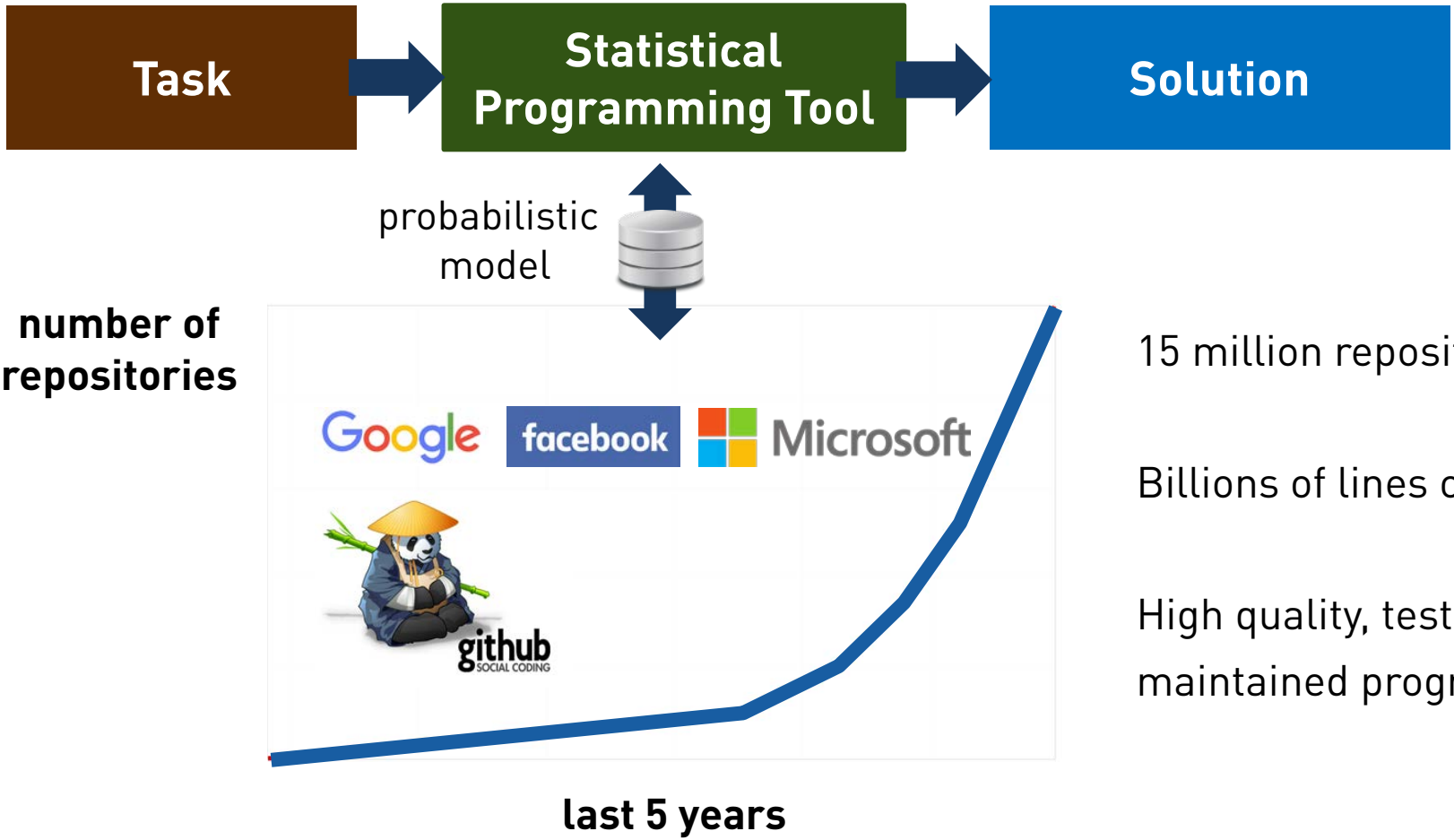
A group of people shopping at an outdoor market. There are many vegetables at the fruit stand.

Medical Computing
(e.g., disease prediction)



Can we bring this revolution to programmers?

Machine Learning for Programs



15 million repositories

Billions of lines of code

High quality, tested, maintained programs

Why now?

Advances in Programming Languages
[Automated Reasoning, Synthesis, Constraint Solving]

Confluence of streams

Advances in Machine Learning
[Deep Learning, Graphical Models, Language Models]

Data
[> 15 million public repositories]

machine learning-based programming tools
new rules, new ideas, new opportunities

Sample efforts in this space

Academia

@ETH: http://plml.ethz.ch

@Technion: http://www.cs.technion.ac.il/~yahave/prime/

@Rice: https://www.cs.rice.edu/~sc40/

@MIT: http://people.csail.mit.edu/fanl/

@Microsoft:
https://www.microsoft.com/en-us/research/project/program/

@Edinburgh: http://homepages.inf.ed.ac.uk/csutton/

Start-ups

DEEPCODE

http://deepcode.ai [ETH spin-off]

codota

http://codota.com

prodo.ai

https://prodo.ai/

NEAR

http://near.ai

DARPA's funded the 40M USD MUSE program on this topic

Seminal work in the area

Learning from Large Codebases,
PhD Thesis, ETH Zurich, 2016

Dr. Veselin Raychev
(CTO, DeepCode)

ETH Medal for Outstanding PhD thesis

ACM Doctoral Dissertation, Honorable Mention Award

Only 3rd time in 40-year history of ACM that a PhD from Europe wins this award.

"...regarded as having the potential to open up several promising new avenues of research in the years to come..."
source: https://awards.acm.org/about/2016-doctoral-dissertation

ACM SIGPLAN and Communications of the ACM Research Highlights, 2018

"...significantly advance the state-of-the-art in statistical reasoning on programs and offer the first concrete evidence of the tremendous promise of the overall approach..."
source: http://www.sigplan.org/Highlights/Papers/

@ETH

Probabilistically likely solutions to problems hard to solve otherwise

Publications

- Program Synthesis for Character Level Language Modeling, **ICLR'17**
- Learning a Static Analyzer from Data, **CAV'17**
- Statistical Deobfuscation of Android Applications, **ACM CCS'16**
- Probabilistic Mode for Code with Decision Trees, **ACM OOPSLA'16**
- PHOG: Probabilistic Mode for Code, **ACM ICML'16**
- Learning Programs from Noisy Data, **ACM POPL'16**
- Predicting Program Properties from "Big Code", **ACM POPL'15**
- Code Completion with Statistical Language Models, **ACM PLDI'14**
- Machine Translation for Programming Languages, **ACM Onward'14**

ML Engines

apk-deguard.com

DEGUARD

DEEP3

jsnice.org

JS NICE

nice2predict.org

SLANG

NICE 2 Predict

more: http://plml.ethz.ch

http://deepcode.ai

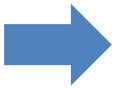
Dimensions:
Machine Learning for Programming

Applications	Code completion	Deobfuscation	Program synthesis	Feedback generation	Translation
Intermediate Representation	Sequences (sentences)	Trees	Translation Table	Feature Vectors	Graphical Models (CRFs)
Analyze Program (PL)	typestate analysis	scope analysis	control-flow analysis	alias analysis	
Train Model (ML)	Neural Networks	N-gram/back-off	SVM	PCFG	Structured SVM
Query Model (ML)	argmax P(y | x)	y ∈ Ω			Greedy MAP inference

ML to write code

```
Camera camera = Camera.open();
camera.setDisplayOrientation(90);
```

SLANG



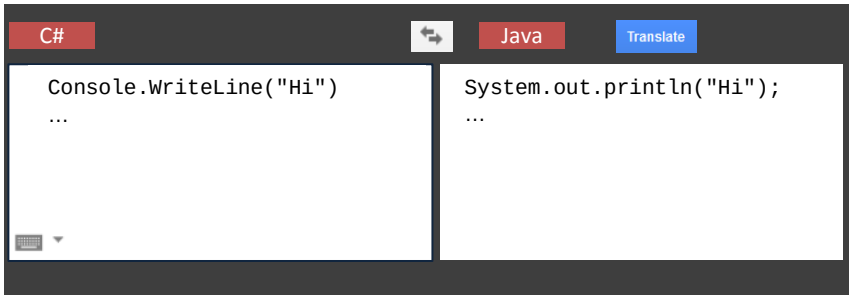
```
Camera camera = Camera.open();
camera.setDisplayOrientation(90);
camera.unlock();
SurfaceHolder holder = getHolder();
holder.addCallback(this);
holder.setType(SurfaceHolder.STP);
MediaRecorder r = new MediaRecorder();
r.setCamera(camera);
r.setAudioSource(MediaRecorder.AS);
r.setVideoSource(MediaRecorder.VS);
r.setOutFormat(MediaRecorder.MPEG4);
```

Statistical language models
Recurrent neural networks

+

Typestate analysis
Alias analysis

ML to Translate Between Languages



Phrase-based Statistical Machine Translation

+

Prefix Parsing of Context-Free Grammars

ML for code deobfuscation

```
function FZ(e, t) {   var n = [];
var r = e.length;   var i = 0;
for (; i < r; i += t)      if (i + t <
r)      n.push(e.substring(i, i +
t)); else
n.push(e.substring(i, r));
return n;
}
```

JSNice.org

```
function chunkData(str, step) {
  var colNames = [];
  var len = str.length;
  var i = 0;
  for (; i < len; i += step)
    if (i + step < len)
      colNames.push(str.substring(i, i + step));
    else
      colNames.push(str.substring(i, len));
  return colNames;
}
```

JSNice.org

JS NICE

✓ Every country ✓ 200,000 users ✓ Top ranked tool

This Page Amsterdam @thispage_ams · Jul 16
Do you write ugly JavaScript code? Not to worry. JSNice will make it look like you are a superstar coder. Yai! - buff.ly/1HR4JL7

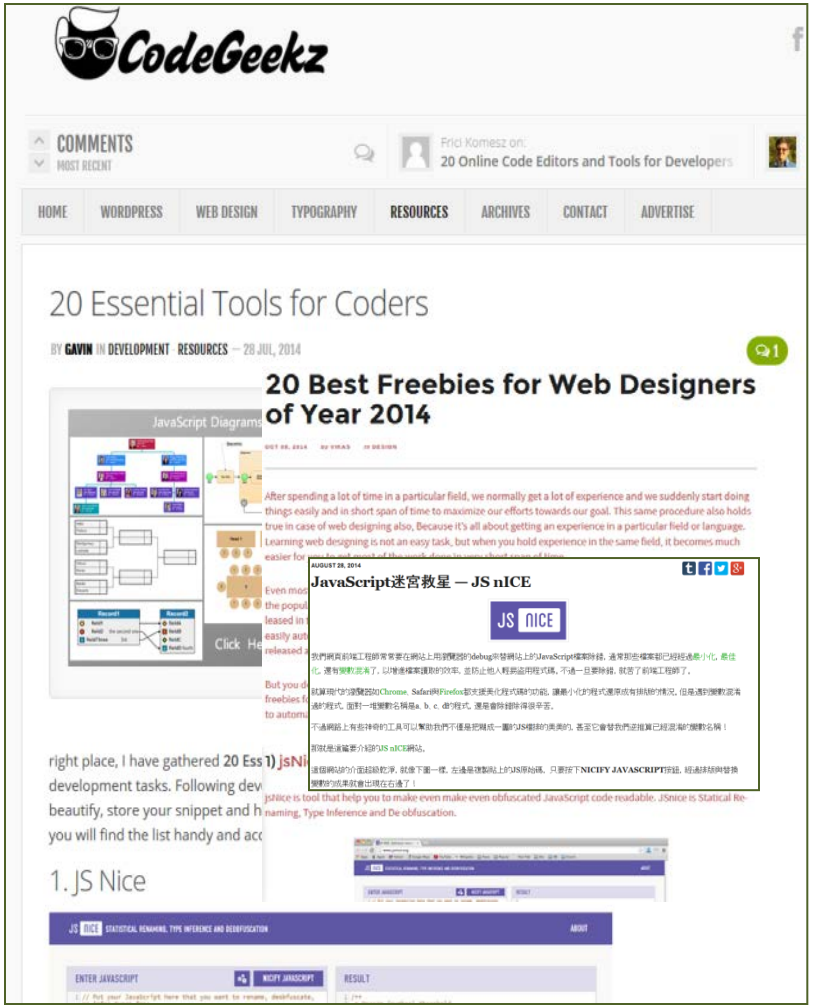
Ingvar Stepanyan @RReverser · Aug 6
JSNice.org became my must-have tool for code deobfuscation.
Expand Reply Retweet Favorite More

Brevity @seekbrevity · Jul 28
JSNice is an a amazing tool for de-minifying #javascript files. JSNice.org, Its great for #learning and reverse engineering.
Expand Reply Retweet Favorite More

Alvaro Sanchez @alvasavi · Jun 10
This is gold.
Statistical renaming, Type inference and Deobfuscation.
jsnice.org
Expand Reply Retweet Favorite More

Alex Vanston @mrvdot · Jun 7
I've been looking for this for years: JS NICE buff.ly/1pQ5qfr #javascript #unminify #deobfuscate #makeItReadable
Expand Reply Retweet Favorite More

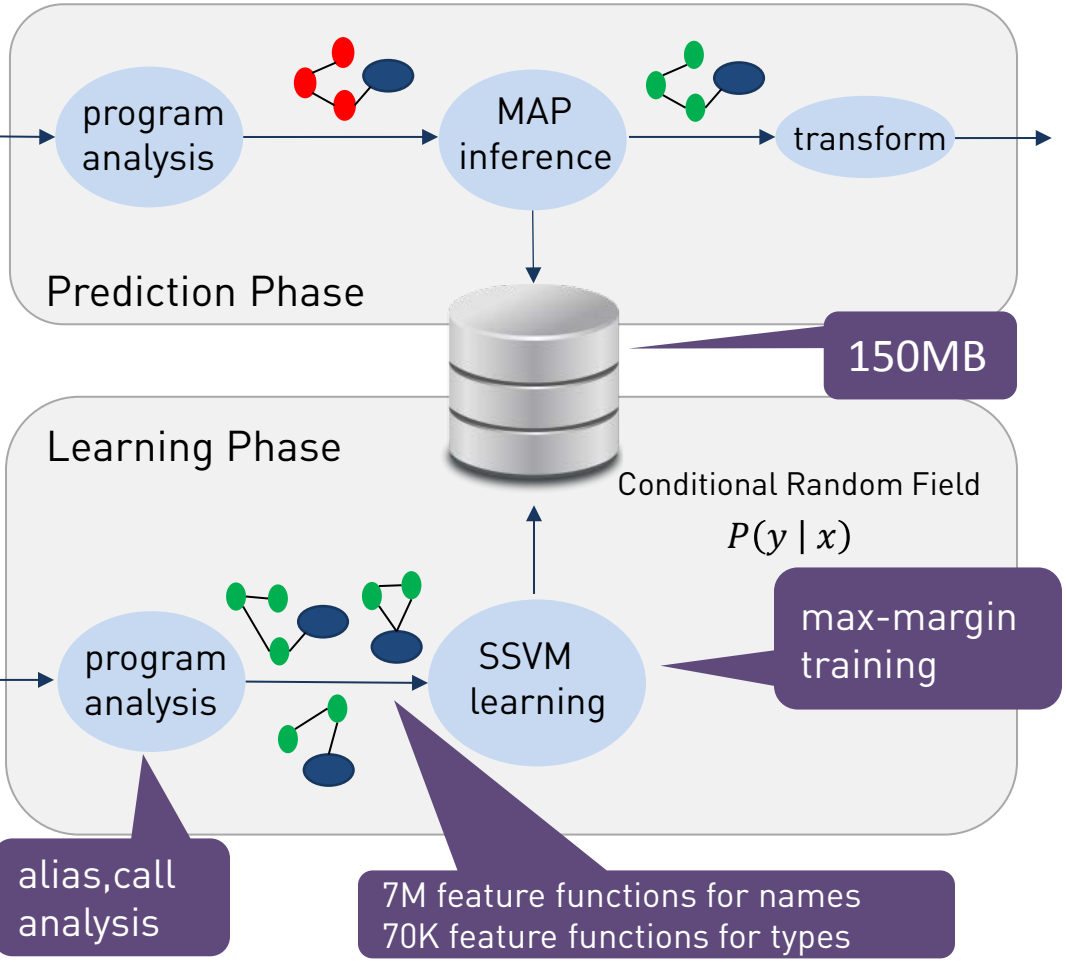
Kamil Tomšík @cztomsik · Jun 6
tell me how this works!
de-minify #jquery #javascript incl. args, vars & #jsdoc
impressive! jsnice.org



JSNice

```
var n = [];
var r = e.length;
var i = 0;
for (; i < r; i += t)
  if (i + t < r)
    n.push(e.subs(i, i + t));
  else
    n.push(e.subs(i, r));
return n;
```

Atlassian Bitbucket



```
var colNames = [];
var len = str.length;
var i = 0;
for (; i < len; i += step)
  if (i + step < len)
    colNames.push(str.subs(i, i + step));
  else
    colNames.push(str.subs(i, len));
  return colNames;
```

ML for code deobfuscation

```
package a.b.c
class a extends SQLiteHelper
{
  SQLiteDatabase b;  public
a(Context ctx) {   b =
getWritableDatabase();  }
Cursor c(String str) {
return b.rawQuery(str); }}
```

apk-deguard.com

```
package com.example.dbhelper

class DBHelper extends
SQLiteHelper {
  SQLiteDatabase db;

  public DBHelper(Context ctx) {
    db = getWritableDatabase();
  }

  Cursor execSQL(String str) {
    return db.rawQuery(str);
  }
}
```

Funny Reddit post/comment

[–] **Tycon712** • **3 points** 2 days ago

Can someone tell me what the point of using Proguard is if there are tools out there like this?

permalink embed pocket

[–] **theheartbreakpug** • **6 points** 2 days ago

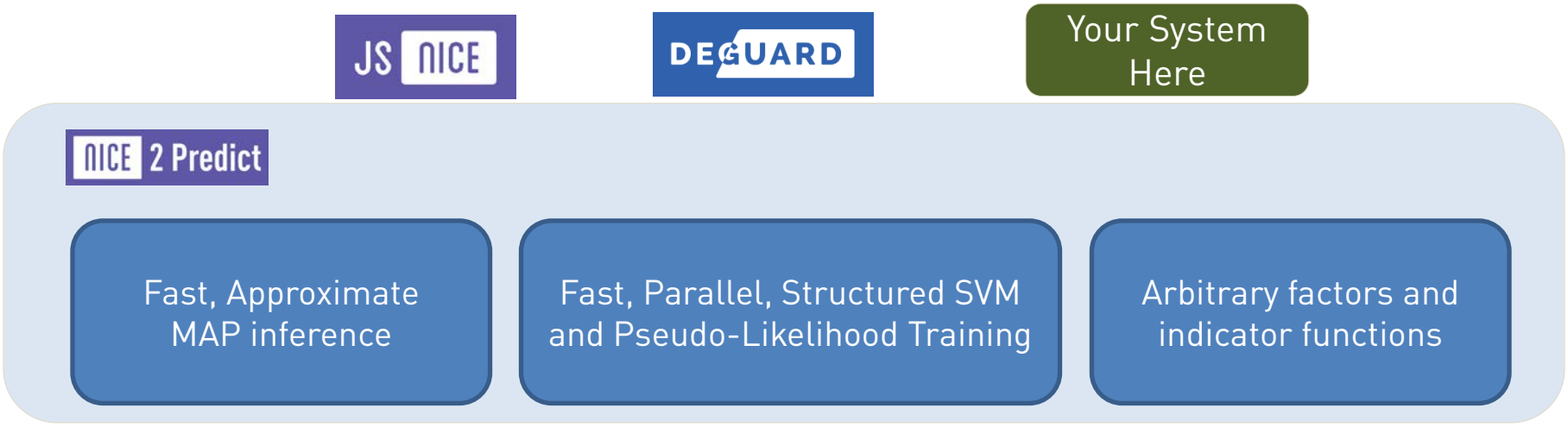
As far as I know, this is brand new. I asked the creator of ProGuard a week ago how hard it is to unobfuscate code after it's run through proguard. He said it strips all the names out of the code so it's essentially impossible. I'm super impressed by what they've done here.

Nice2Predict.org:

scalable structured prediction framework

fully, open sourced, Apache license

used by various groups worldwide



OPPORTUNITY

Advances in Programming Languages
[Automated Reasoning, Synthesis, Constraint Solving]

Advances in Machine Learning
[Deep Learning, Graphical Models, Language Models]

Data Google
[> 15 million public repositories]

Learning-based programming tools
new rules, new ideas, new opportunities

RICH PROBLEM SPACE

Applications	Code completion, Deobfuscation, Program synthesis, Feedback generation, Translation
Intermediate Representation	Sequences (sentences), Trees, Translation Table, Graphical Models (CRFs), Feature Vectors
Analyze Program (PL)	typestate analysis, scope analysis, control-flow analysis, alias analysis
Train Model (ML)	Neural Networks, N-gram language model, PCFGs, SVM, Structured SVM
Query Model (ML)	argmax P(y | x), y ∈ Ω, Greedy MAP inference

NEW PROBABILISTIC MODELS

1. Pick a structure of interest, e.g., trees:

2. Define a DSL for expressing functions:
(can be Turing complete)

```
TCond   ::=  ε | WriteOp TCond | MoveOp TCond
MoveOp  ::=  Up, Left, Right, DownFirst, DownLast,
             NextDFS, PrevDFS, NextLeaf, PrevLeaf,
             PrevNodeType, PrevNodeValue,
WriteOp ::=  WriteValue, WriteType, WritePos
```

3. Synthesize $f_{best} \in$ DSL from Dataset D:

$$f_{best} = \text{argmin cost}(D, f)$$
$$f \in \text{DSL}$$

4. Use f_{best} on new structures:

f_{best}  $\rightarrow y$

PRACTICAL IMPACT

Course objectives

Introduction to the emerging area of learning from Big Code.

Learn how to read and evaluate papers in the area

Learn how to make good technical presentations

Basic Information

- Instructors
  - Prof. Dr. Martin Vechev (martin.vechev@inf.ethz.ch)
  - Dr. Veselin Raychev (veselin.raychev@inf.ethz.ch)

- Meetings
  - once a week, 2 presentations per meeting
  - next meeting: March 5
  - Presentation schedule will be posted on course www: http://www.srl.ethz.ch/bigcode18.php

How it works: your tasks

select a paper, get date

By this Friday, send e-mail to Veselin (veselin.raychev@inf.ethz.ch) with your 5 choices

study paper

create presentation

meet advisor, get feedback

---

give final presentation: 30 min
answer questions: 15 min

Participation:
ask good questions, attend all classes

Study paper

select a paper, get date

study paper

create presentation

meet advisor, get feedback

3 'C's of reading:

Carefully: lookup unknown terms, read cited papers

Critically: find limitations, flaws

Creatively: think of improvements

Write: key ideas, try examples by hand

Consult with TA / Instructors, if questions email also fine

---

give final presentation: 30 min
answer questions: 15 min

Participation:
ask good questions, attend all classes

Create presentation

select a paper, get date

study paper

create presentation

meet advisor, get feedback

- Explain motivation for the work
- Clearly present the technical solution and results
- Use your own example, not the one in the paper
- Outline limitations / improvements
- Focus on the key/crucial concepts
- Do not present all of the details

give final presentation: 30 min
answer questions: 15 min

Participation:
ask good questions, attend all classes

Meet advisor, get feedback

select a paper,
get date

study paper

create presentation

meet advisor,
get feedback

clear technical questions on the paper

get feedback on draft presentation

meeting mandatory a week or so before presentation

give final presentation: 30 min
answer questions: 15 min

Participation:
ask good questions, attend all classes

Grading

select a paper,
get date

study paper

create presentation

meet advisor,
get feedback

quality of your final presentation

- how well you understood the material ?
- how well you presented it ?
- how well you answered the questions ?

we will take into account paper difficulty

participation

- did you ask good questions ?
- attendance will be taken (do not miss classes)

give final presentation: 30 min
answer questions: 15 min

Participation:
ask good questions, attend all classes

Useful Presentation Links

"How to give strong technical presentations", by Markus Püschel

- http://users.ece.cmu.edu/~pueschel/teaching/guides/guide-presentations.pdf
- Please read the above slides !

"Even a geek can speak" , by Joey Asher