FUNCTION A — name: detect-data-type / arch: mips32-clang-3.5-01
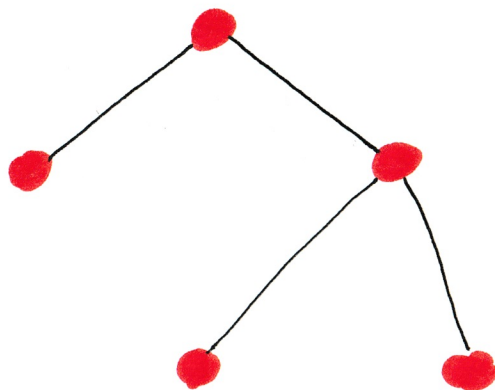
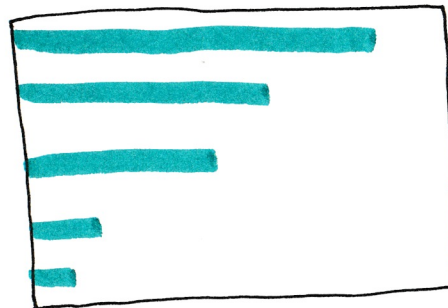**FUNCTION B** — name: detect-data-type / arch: x86

BB: 83188
len: 16
mnems: [ldr, cmp...
disasm: [ldr r2, cmp r2...
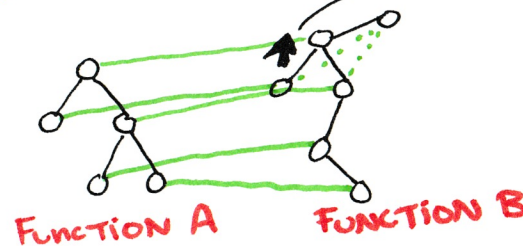heads: [83188, 8412...
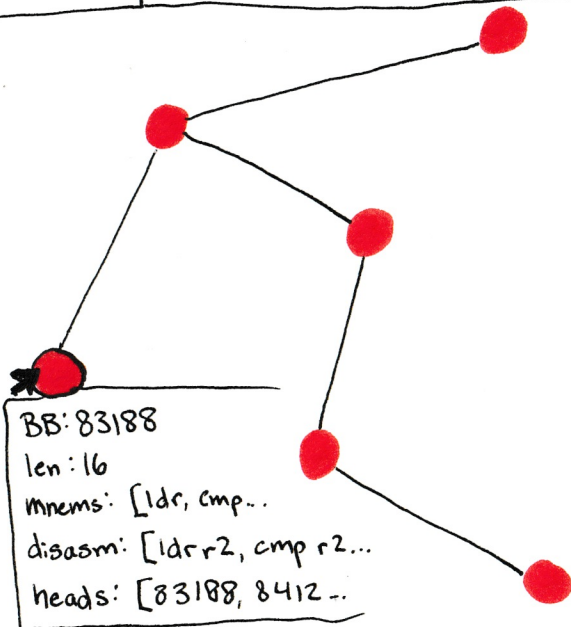
Prediction: POSITIVE

Feature Importance:

- bb-len
- edges
- dissassembly
- string
- nodes

Graph edit distance map:

FUNCTION A   FUNCTION B

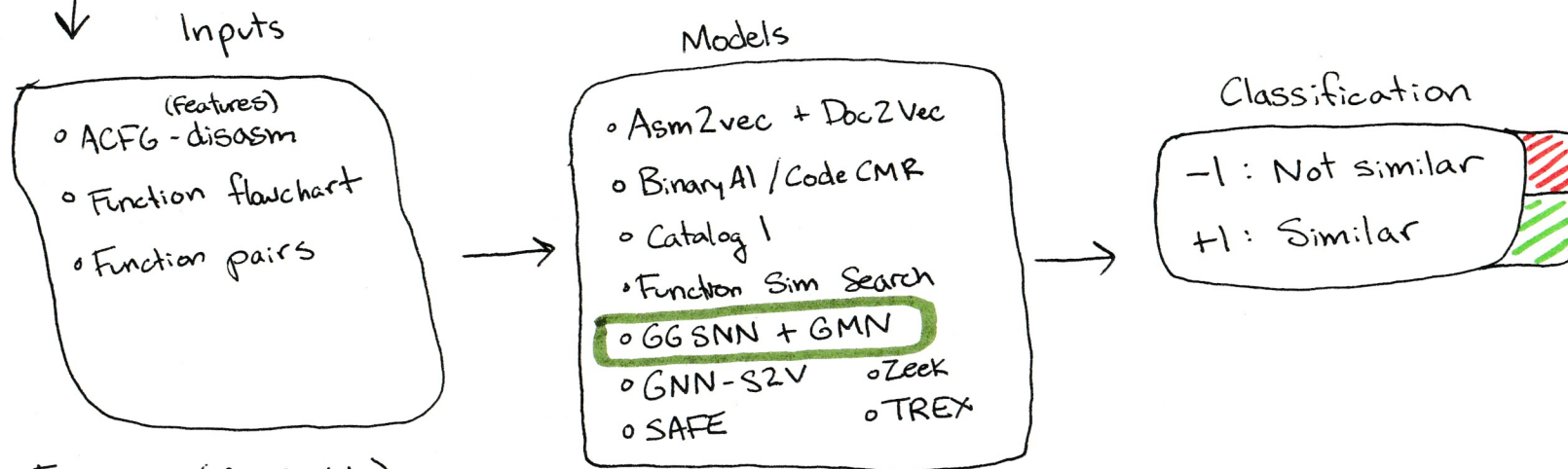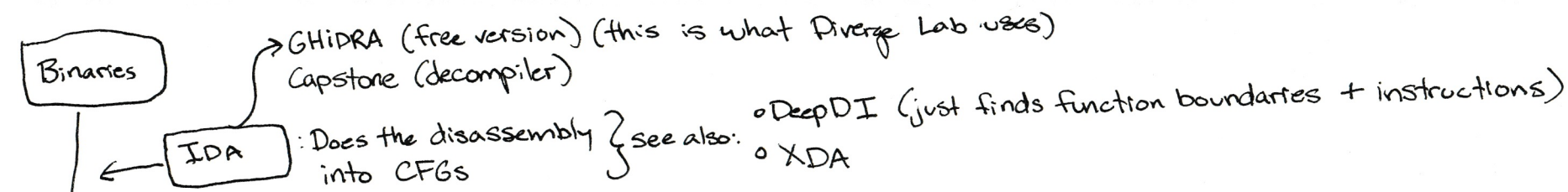| A	B
ldr	ldr
ldr	ldr
eors	eors
beq	beq

* interesting food for thought: mnemonics across architectures are different

MODEL EXPLORER

Selected: Function A vs. Function B

← Higher R^2 ———   ——— Higher R^2 →

Binaries

→ GHIDRA (free version) (this is what Diverge Lab uses)
Capstone (decompiler)

IDA : Does the disassembly into CFGs } see also: ◦ DeepDI (just finds function boundaries + instructions)
◦ XDA

Inputs

(Features)
◦ ACFG - disasm
◦ Function flowchart
◦ Function pairs

Models

◦ Asm2vec + Doc2Vec
◦ Binary AI / Code CMR
◦ Catalog 1
• Function Sim Search
◦ GGSNN + GMN
◦ GNN-S2V   ◦ Zeek
◦ SAFE   ◦ TREX

Classification

-1 : Not similar
+1 : Similar

Features: (of CFG data)
- nodes (bbs)
- edges
- bb: length (bytes)
    mnemonics
    instructions
    dissassembly
    string
    heads (addresses)

Features of function data:

idb-path
fva
start/end
bb-num (count)
bb-list