INTRO

Class imbalance is prevalent across all problems.

→ How to do anomaly detection.

- Medical
- Net. Security
-

CARLETON DATA SCIENCE

Focus on data science

↳ for Masters Dgr.

2 New Programs of D.S

→ w- PHD program

Data Day on Mar 31st !

BENJAMIN FUNG PRESENTATION (on Authorship & Malware Analysis).

Data-mine & ML for
Funded by CFI.

Today focus on National Defense Sector

Authorship Analysis

feature-engineering w-embedding

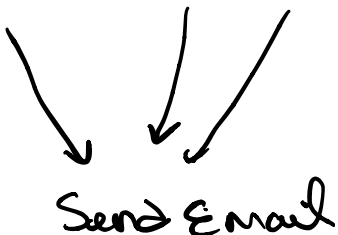→ can be used for different problems

— deep learning not so reusable.

Cheque Scam

① Send cheque to someone

② Ask for return on some money

③ Bounce the cheque

How to prove the scammer is the author of the email?

Authorship Identification

Candidate X   CY   C$_z$

Send Email

There is networking proof but no link between keyboard & typer

Handcrafted Stylometric Features

- lexical (caps/no caps)
- syntactic (spaces, commas)
- structural
- content-specific
- idiosyncratic features

Why Models don't work:
Naïve Bayes
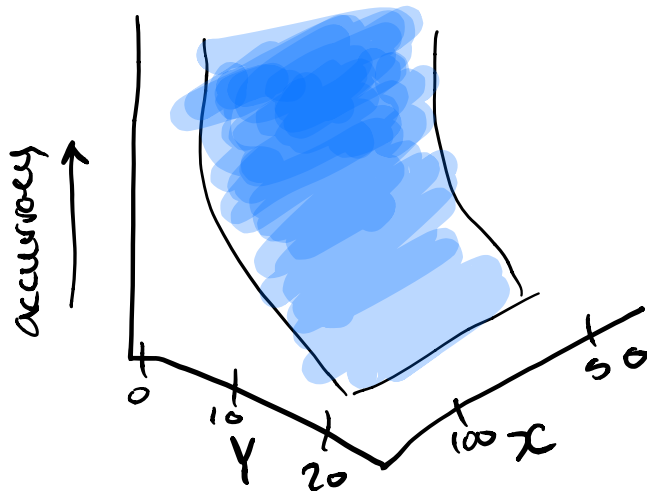- low classification accuracy.

Decision Trees
- accuracy ≈ SVM

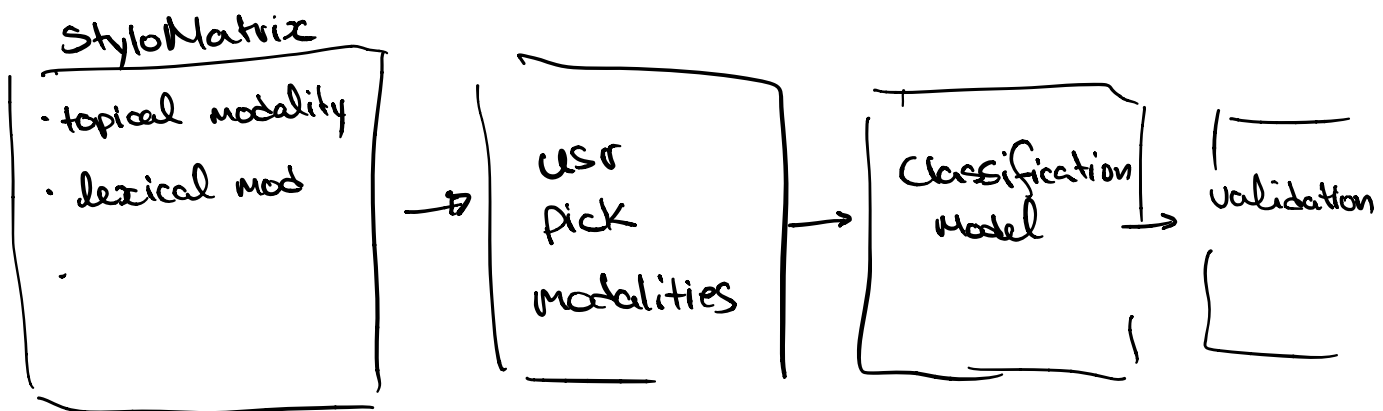Support Vector Machine
.

- small data

x: # writing samples
y: number of candidates

StyloMatrix
- topical modality
- lexical mod
  .

→ usr pick modalities → Classification model → validation

topic-lexic-2-vec

1. find topic
2. find context   } find in emails.
3 find bias

Also is <mark>Authorship Verification,</mark> Auth. Characterization.

age   political   → sex.
      orientation

(Assemby — Reverse Engineering)
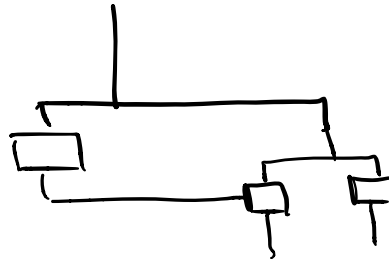
Given binary code — how to understand its behaviour.

push eax
. . . . -

. . . -

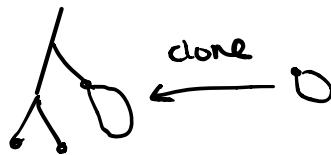test . . .

jz - - -

$\Rightarrow$

Control Flow Graph

PROBLEM — how to find subgraph clone

clone

Approach
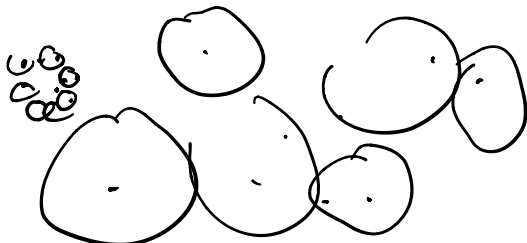   when same
  ① Hash code to compare

  ② diff operators
      normalize & hash

  ③ almost same idiosyncracies
      can't hash ⇢ do subgraph clone w- search
                              engine.
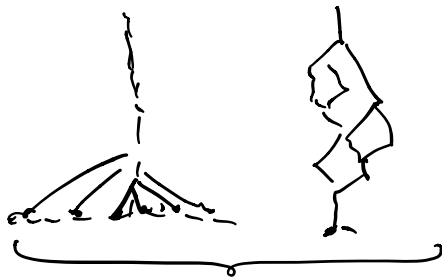   <mark>locality sensing hashing</mark>

Adaptive Locality Sens Hashing

• find circular reference
• expand circ. ref as much possible.

shrink scope for dense distrib.

Obfuscation & Optimization

Same Code — how detect?
↳ use ML/AI (Kamin∅).

Fighting Rats

Origin

[Agr.]
↑
social
interaction
zone

corn.   corner
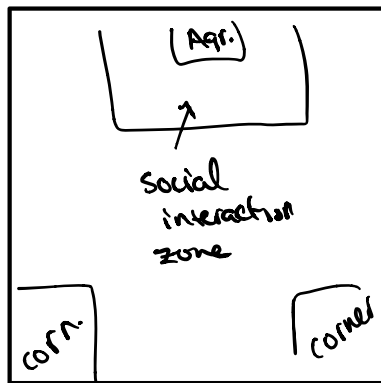
Session #1

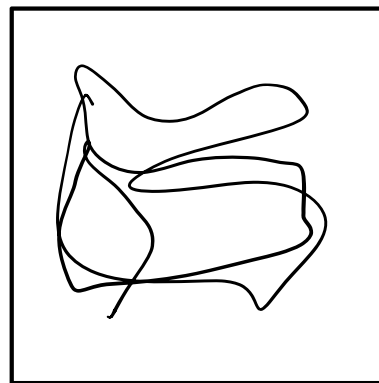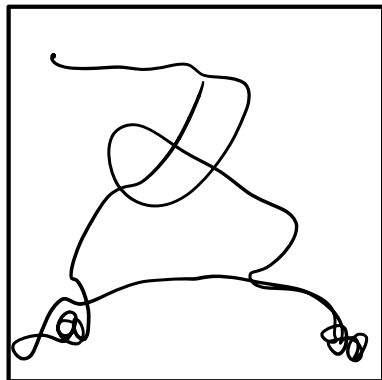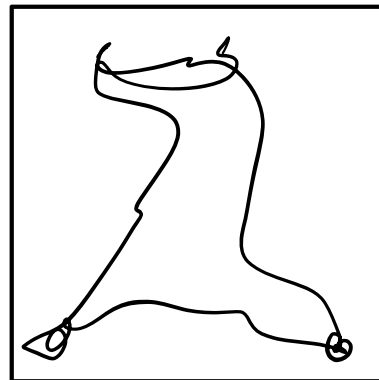### Session n/2

Session # n

(✳) Could classify political sentiment using 2013 twitter data.
→ StyloMatrix. App on GitHub