A Principled Component Analysis (PCA) of Open Source AI

julia ferraioli
Open Source AI/ML Strategist @ AWS
@juliaferraioli@floss.social

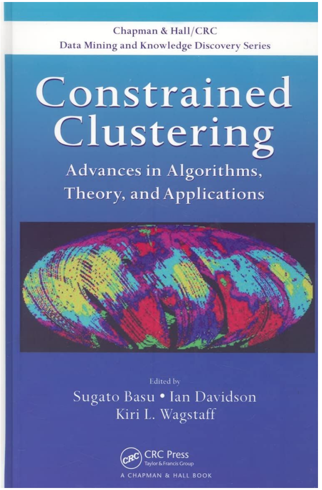
hi, I'm julia

- research background in artificial intelligence (both symbolic and statistical)
- open source practitioner, engineer, and storyteller
- LaTeX and Lego enthusiast (but hey, at least one of them is free!)
- at AWS focusing on the intersection of open source and AI/ML
- creator of (bad) puns

@juliaferraioli@floss.social

yes, the title is a pun

everyone needs an origin story



Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Constrained Clustering

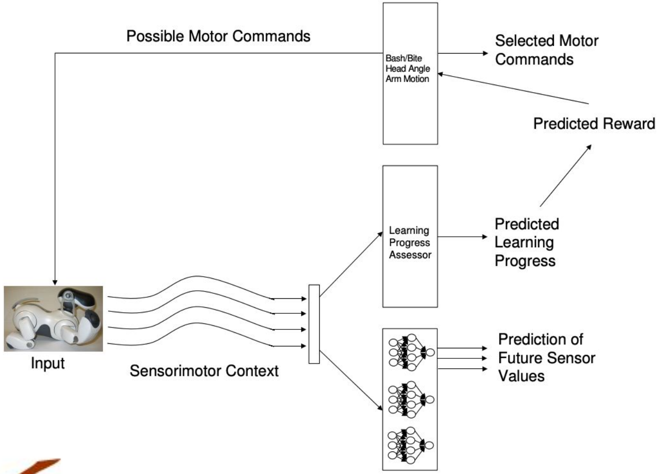
Advances in Algorithms, Theory, and Applications

Edited by
Sugato Basu • Ian Davidson
Kiri L. Wagstaff

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

WEKA
The University of Waikato

CC-BY-SA logo courtesy of WEKA



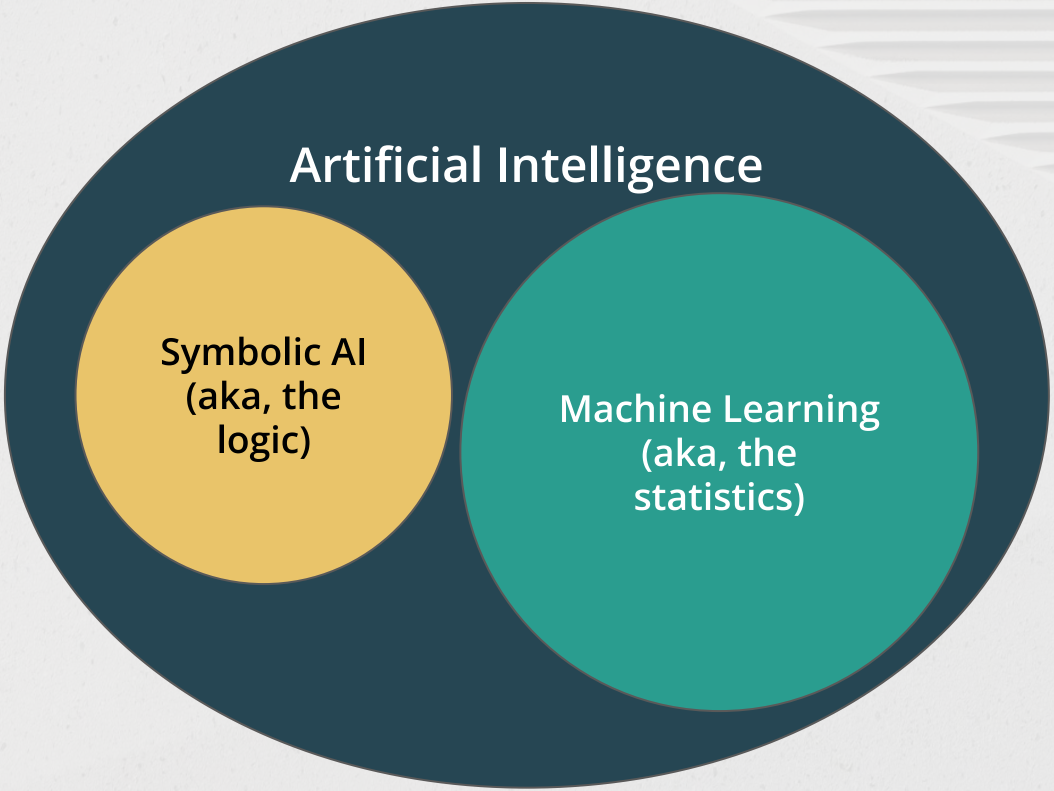
PYRO
Python Robotics

@juliaferraioli@floss.social

(some) foundations of artificial intelligence

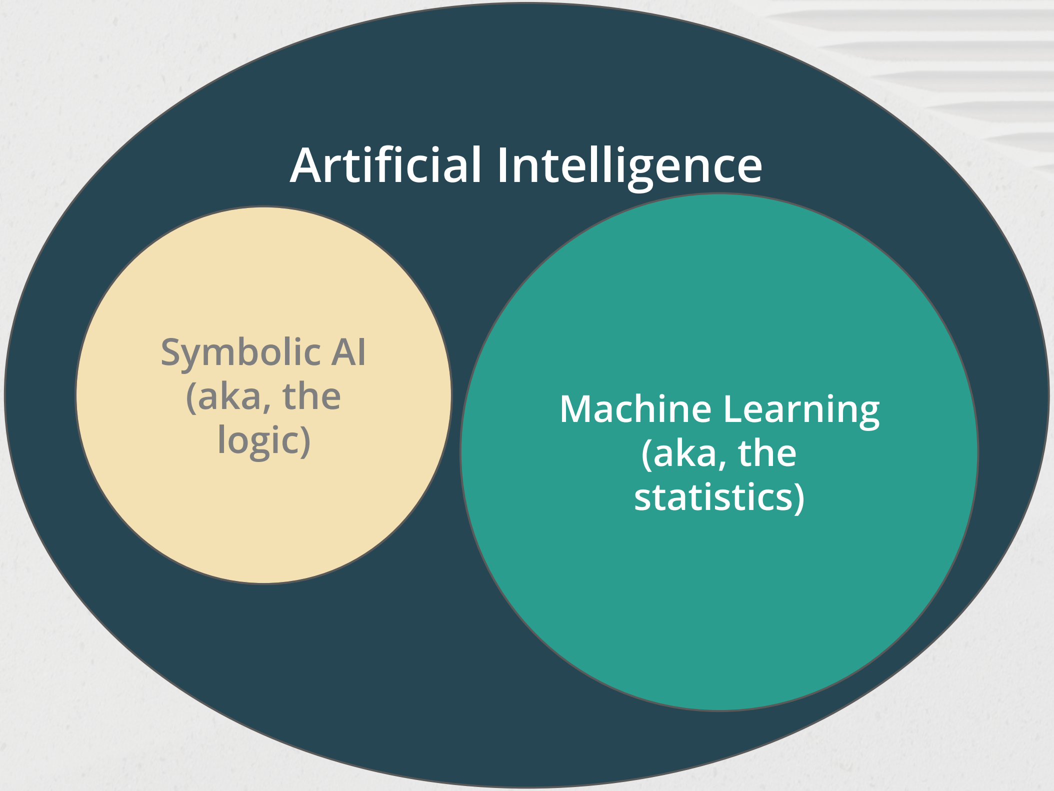
- ethics
- psychology
- computer science
- linguistics
- mathematics
- formal logic
- cognitive science
- philosophy
- economics

a very (very) simplified diagram of AI


Artificial Intelligence
Symbolic AI (aka, the logic)
Machine Learning (aka, the statistics)

@juliaferraioli@floss.social

a very (very) simplified diagram of AI



Artificial Intelligence

Symbolic AI
(aka, the
logic)

Machine Learning
(aka, the
statistics)

@juliaferraioli@floss.social

(some possible) elements of machine learning

data	transformation	cleaning
training	task / prompt	parameters
algorithm(s)	feedback / tuning	evaluation
model	interface	hardware

but if we reduce the dimensionality…

(some possible) elements of machine learning

data

- training
- validation
- test
- task / prompt
- feedback

code

- cleaning
- transformation
- parameters
- algorithm(s)

other stuff

- hardware
- duration
- evaluation

output

- model
- interface

@juliaferraioli@floss.social

it becomes easier to reason about

data

data (the big problem)

known problems include…

- historic unknown provenance
- need for deidentification and anonymization
- privacy of feedback (if applicable)
- split of corpus (if applicable)

data (the big problem)

is the...

- entire data set
- data description
- data sheet
- data collection methodology

required to "recompile" a model?

data (the big problem)

unsolved problems…

- equitable hosting costs
- transparency to avoid GIGO
- attribution and opting out
- rollback of data (and dependent models)

code

code (the known quantity)

it's open source software; I know this!

- majority of machine learning fits in here
- governed by same requirements as open source software
- may produce a model or interface
- intersects with data
    - data cleaning / processing
    - moral and ethical value judgements codified

the other stuff

other stuff (the unknown quantity)

are these...

- hardware specifications
- disclosure of training time
- additional configuration (if applicable)
- definition of correctness (if applicable)

required to "recompile" a model?

other stuff (the unknown quantity)

unsolved problems…

- equitable compute cost
- access to hardware
- feedback loop(s)
- attribution

@juliaferraioli@floss.social

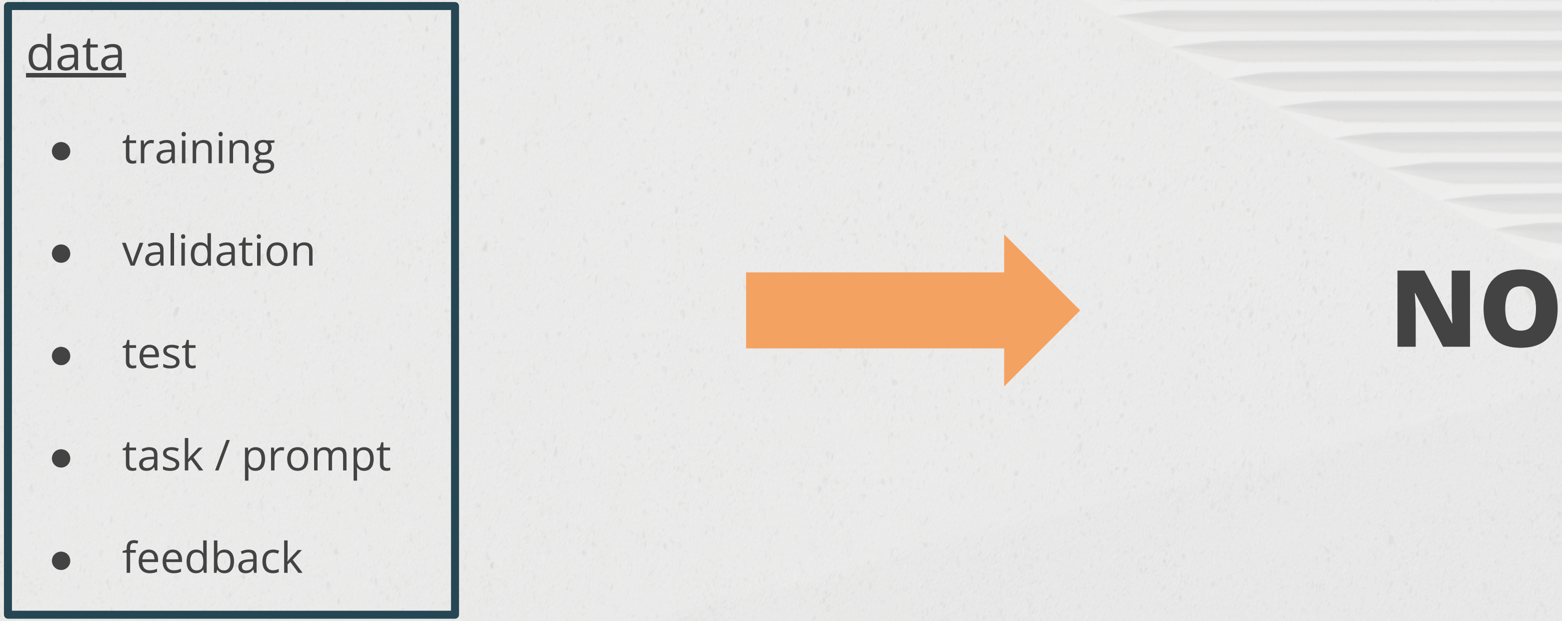
output

output (the *mostly* great unknown)

$$\left\{\begin{matrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{matrix}\right\}$$

the open source litmus test

can I do {*machine learning task*} with *just this*?

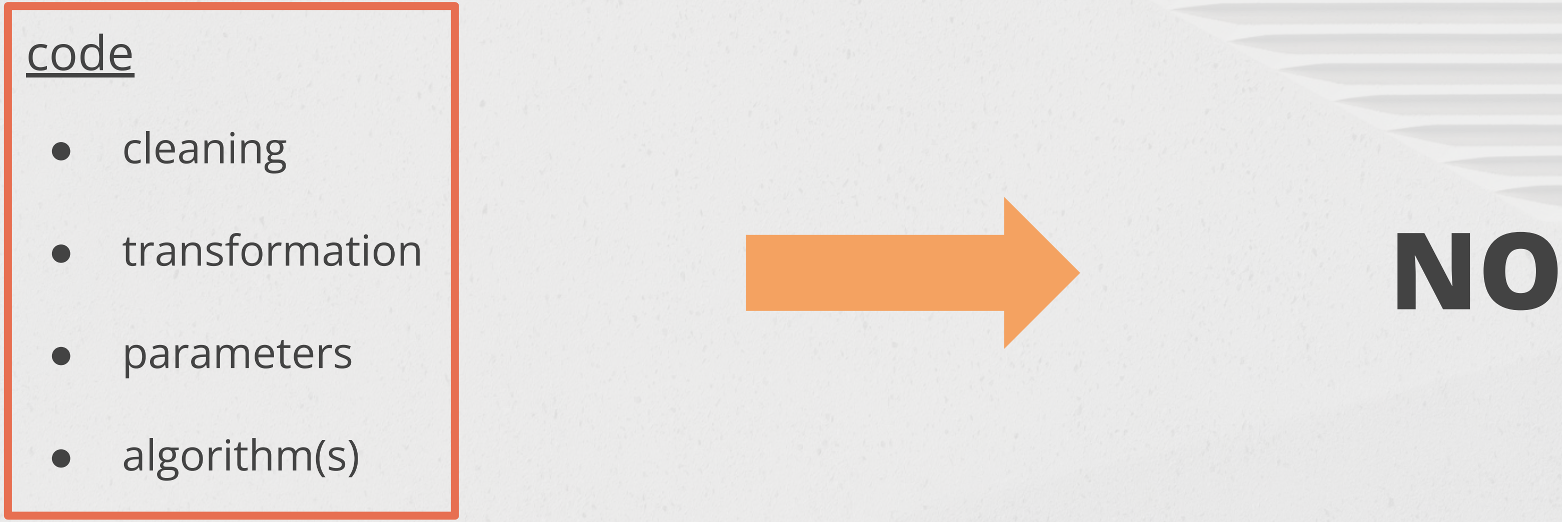


<u>data</u>

- training
- validation
- test
- task / prompt
- feedback

NO

can I do {*machine learning task*} with *just this*?



code

- cleaning
- transformation
- parameters
- algorithm(s)

NO

can I do {*machine learning task*} with *just this*?

other stuff

- hardware
- duration
- evaluation

→

NO

@juliaferraioli@floss.social

can I do {*machine learning task*} with *just this*?

output

- model
- interface

NO

can I do {*machine learning task*} with all of this? **yes!**

data

- training
- validation
- test
- task / prompt
- feedback

code

- cleaning
- transformation
- parameters
- algorithm(s)

other stuff

- hardware
- duration
- evaluation

output

- model
- interface

@juliaferraioli@floss.social

other questions…

thanks!