Advances in Computer Aided Text Analysis Using Machine and Deep Learning

Tim Hubbard, PhD
University of Notre Dame

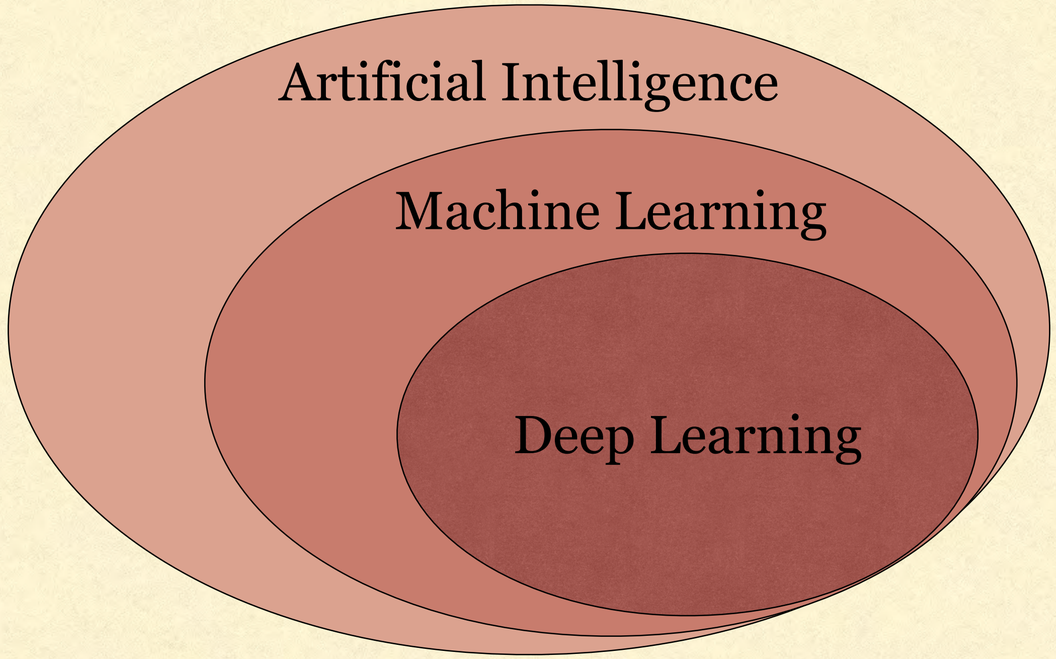
UNIVERSITY OF NOTRE DAME
Mendoza College of Business

"Keep in mind throughout this [presentation] that none of these deep-learning [or machine learning] models truly understand text in a human sense; rather, these models can map the statistical structure of written language, which is sufficient to solve many simple textual tasks."

–Chollet & Allaire (2018: 165)

Difference Between Machine & Deep Learning

- Machine Learning:
  - Artificial intelligence using algorithms that can change on its own, by feeding it structured data.
- Deep Learning:
  - Same, but there are numerous layers of algorithms that extract various features of the data and pass them to the next layer. The algorithms work to determine the weights of the paths between the layers. The input need not be structured.



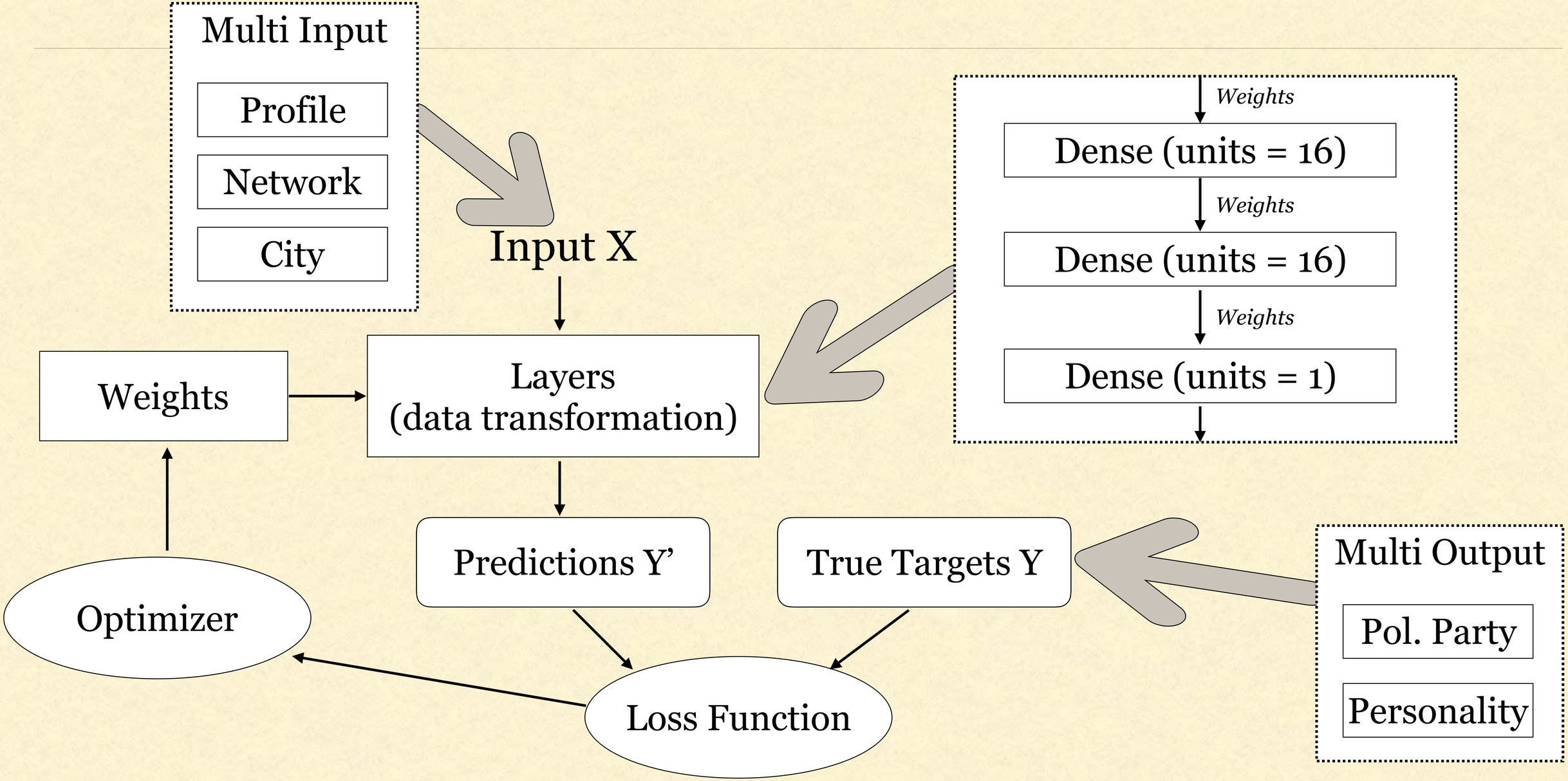
Chollet & Allaire (2018: 165)

EXAMPLES OF SEVERAL OPTIONS

Manual Coding	LIWC	Machine Learning	Deep Learning
Assessing Feature Articles	Assessing level of non-conforming language	Determining 0/1 category of press release	Determining strategic decisions
2 human raters per article	Snowball dictionary with clear words	Expert trained small sample (75)	Direct measurement of constructs
6,260 articles	6,006 articles	49,436 press releases	1,000 participants
7 hours on MTurk (re-ran 795 unclear)	Experts validated dictionary	Validation across several raters & computer	Holdout sample to be used

How do we hypothesize deep learning?

- **Hypothesis:** My outputs can be predicted given my inputs.
- **Hypothesis:** My available data is sufficiently informative to learn the relationship between the inputs and outputs.



Chollet & Allaire (2018)

Text Pre-Processing

STEP	TEXT
Starting text	Tom, all 9 of the boy's cars are different colors!
Strip white space	Tom, all 9 of the boy's cars are different colors!
Remove numbers	Tom, all of the boy's cars are different colors!
Remove case	tom, all the boy's cars are of different colors!
Remove puncutation	tom all the boys cars are of different colors
Remove stop words	tom boys cars different colors
Stemming words	tom boy car differ color
Remove sparse terms	boy car differ color

Tokens & N-Grams

Text:
“The cat sat on the mat.”

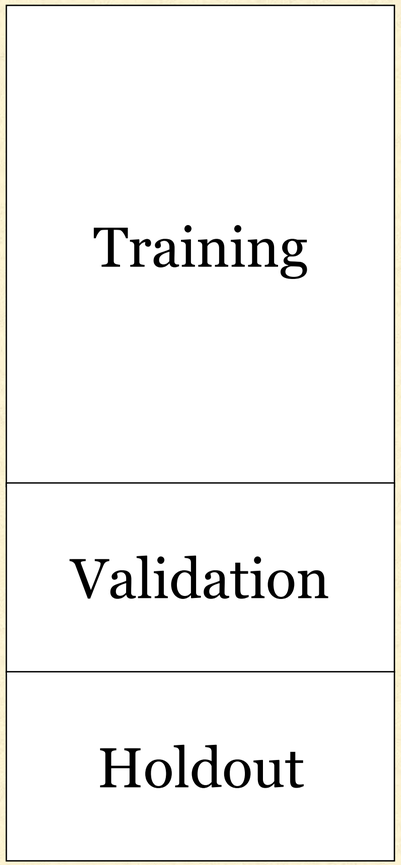
Tokens:
“the”, “cat”, “sat”, “on”, “the”, mat”, “.”

One-hot Vectorization:
2, 1, 1, 1, 1

2-grams:
“the”, “the cat”, “cat”, “cat sat”, “sat”, “sat on”, “on”, “on the”, “the”, “the mat”, “mat”

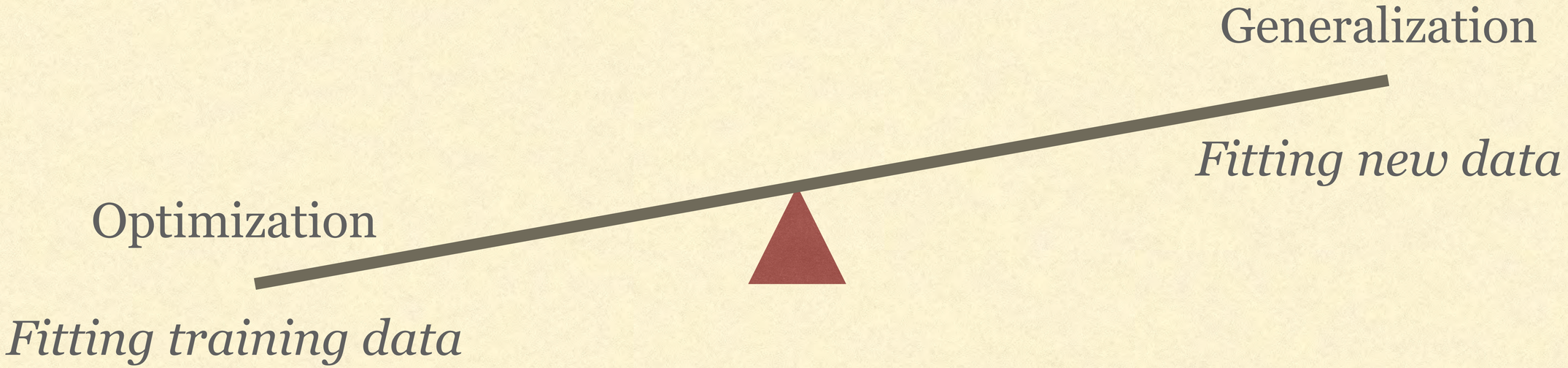
Training, Validation & Test Sets

Basic:



K-Fold:

Overfitting



How to actually do this?

- R & Python: Top level, easy programming
- Keras: Interface between top level and TensorFlow
- TensorFlow: Does the hard work

- NVIDIA GPU: Faster than CPU

- Can be done on computing arrays: Google CloudML and Amazon EC2

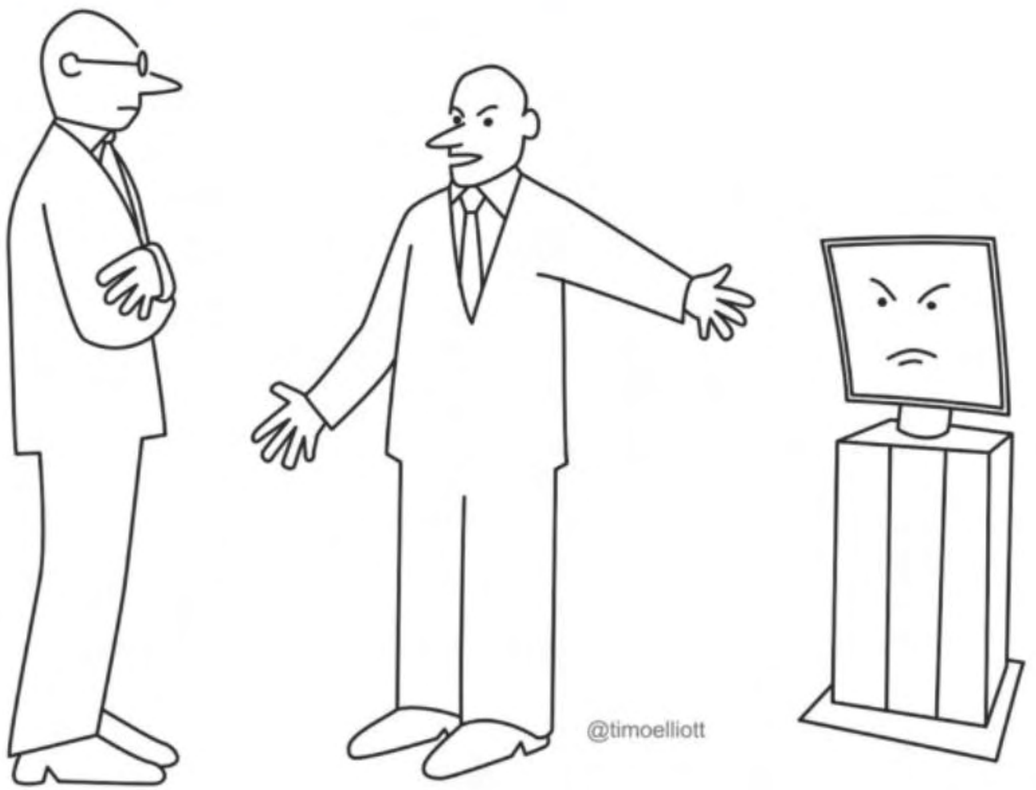
- Kaggle for competitions

A Few Closing Thoughts

- Accurate, direct measurement of inputs/outputs is critical—garbage in, garbage out
- Humans are best

- Democratization of Deep Learning:
  - Must fully describe algorithms, software (including versions) and empirical choices
  - Should post data use for training — the algorithm output is not enough!

- If we want to get better at a text task, we need to be able to replicate exactly what others did, but improve:
  - Add data —> More data, better data, and different types of data
  - Update software —> We are getting better everyday
  - Try different configurations —> Change the properties of the layers, optimizers, etc.



His decisions aren't any better than yours
— but they're WAY faster...

THANK YOU!

A final recommendation of a book if you want to learn how to do this: