Introduction to Machine Learning

Amit Sethi

Faculty member at

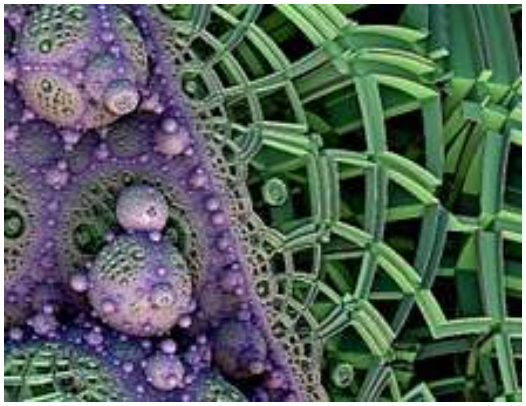
Indian Institute of Technology Bombay

Questions answered in this talk

- Is ML a fad?
- What is ML, and how is it related to AI, NN, and DL?
- What makes a problem good or bad for ML?
- How should one critique an ML model?
- What questions should be asked beyond the performance of ML?
- What are some myths and realities related to ML?

Is AI / ML just a fad?

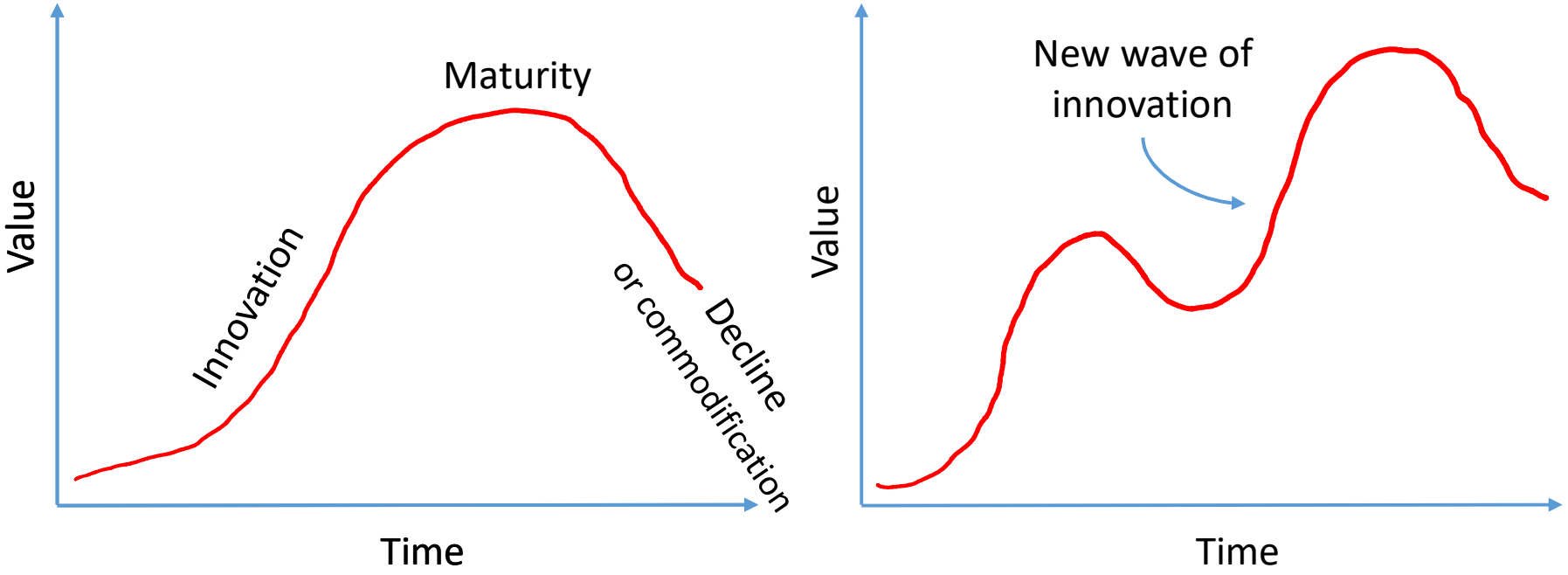
- Other hot buzzwords for colleges, VCs, and grant agencies over the last 30 years:
  - Programming, data structures, databases, IT
  - Computer networks, wireless communication
  - Web 2.0
  - Nano technology



- Some of these have deeply affected our economies and society, but have reached maturity from the PoV of ongoing innovation
- Others have made a limited impact as challenges to their promises became better understood

Image courtesy: Pixabay.com

Technology life cycle has to be understood critically



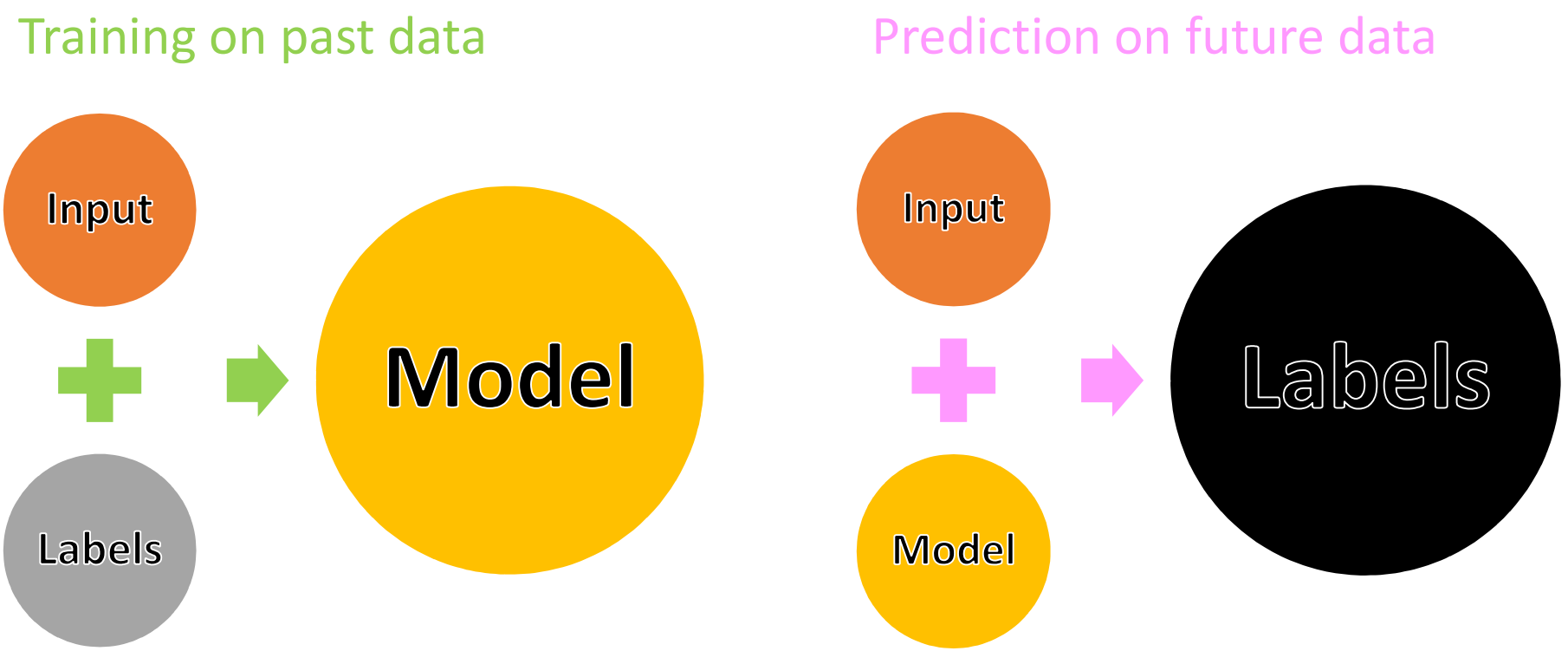
ML is being deeply embedded in our economy and society

- Automated pattern recognition is already here
  - Images and videos: find people, objects, diseases …
  - Voice: convert to text, …
  - Text: queries, chat bots, translation, …
  - Time series: stocks, power consumption, …

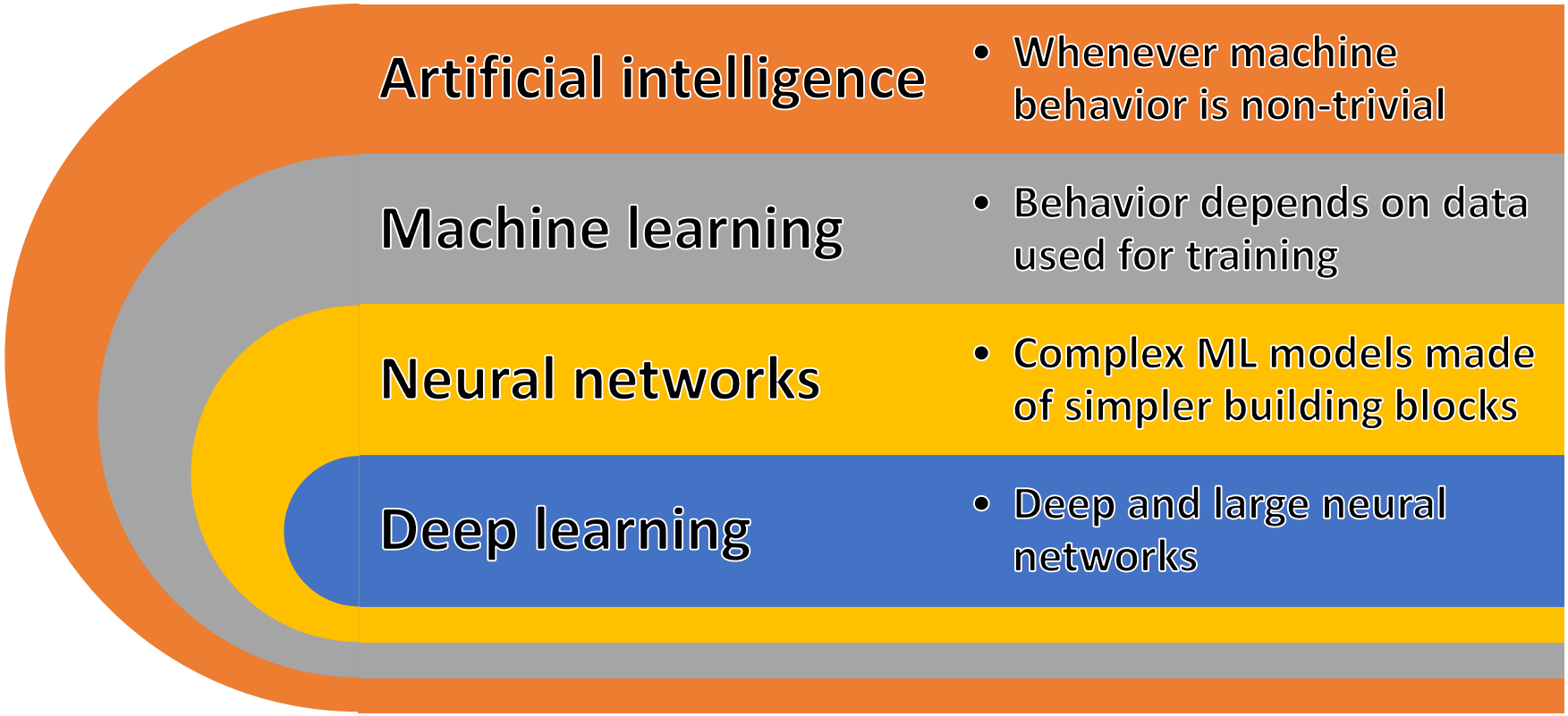
- Automated decision-making is coming
  - Economic decisions: customer targeting, credit approval, …
  - Autonomous machines: Driverless cars, drones, robots, …
  - Critical decisions: medical diagnosis, criminal forensics …

Image courtesy: Pixabay.com

ML model training and deployment



Relation between AI, ML, NN, and DL



Why ML now?

- Lots of data
  - Inexpensive collection, storage, transmission

- Lots of compute
  - Server farms, GPUs (thousands of cores)

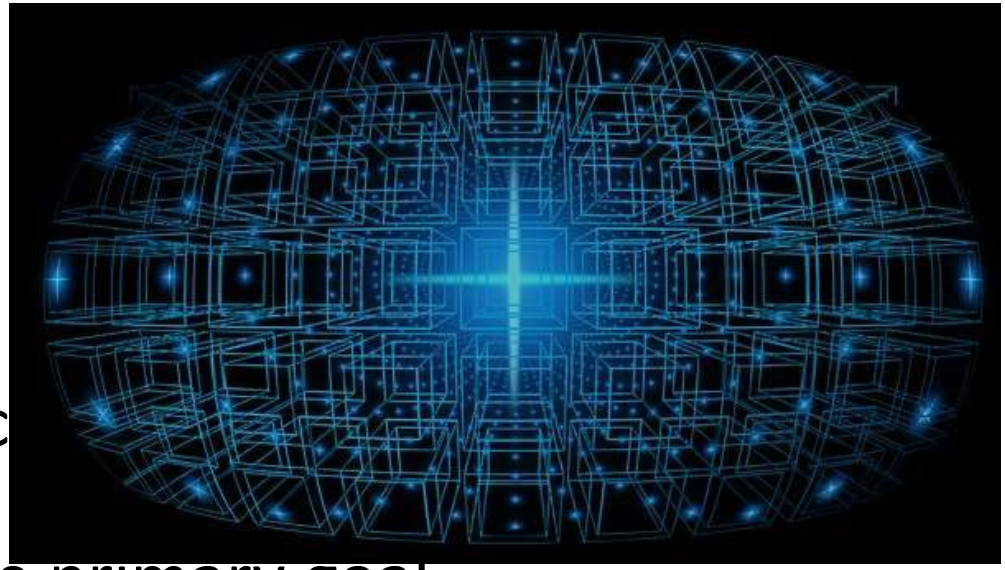
- New frameworks and algorithms that can use lots of data and compute
  - Deep learning (e.g. ResNet, GPT)

Images courtesy: Pixabay.com

When ML is not a silver bullet

- Not sure what the desired output looks like
  - Not been that successful with unsupervised learning
- Too little data
  - ML does not easily generalize to new domains
- Data too unstructured
  - State-of-the-art ML exploits structure in data
- Model is simple or explainability is desired
  - Mathematics, physics, and statistics may be better

Sweet spot for ML



- Lots of structured data
- Explainability is not critic
- Prediction accuracy is the primary goal
- Underlying model is complex but stationary

Image courtesy: Pixabay.com

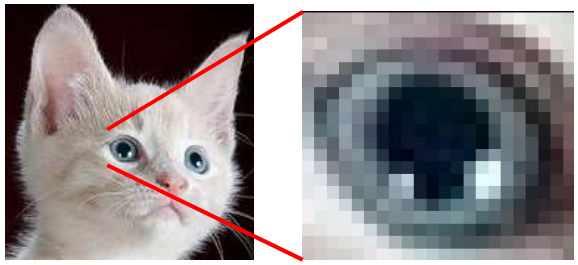
Examples of structure in the data

- Records

Product SKU	Price	Margin	Volume
A123ajkhdf	$ 120	30%	1,000,000
B456ddsjh	$200	10%	2,000,000

- Temporal order

- Spatial order



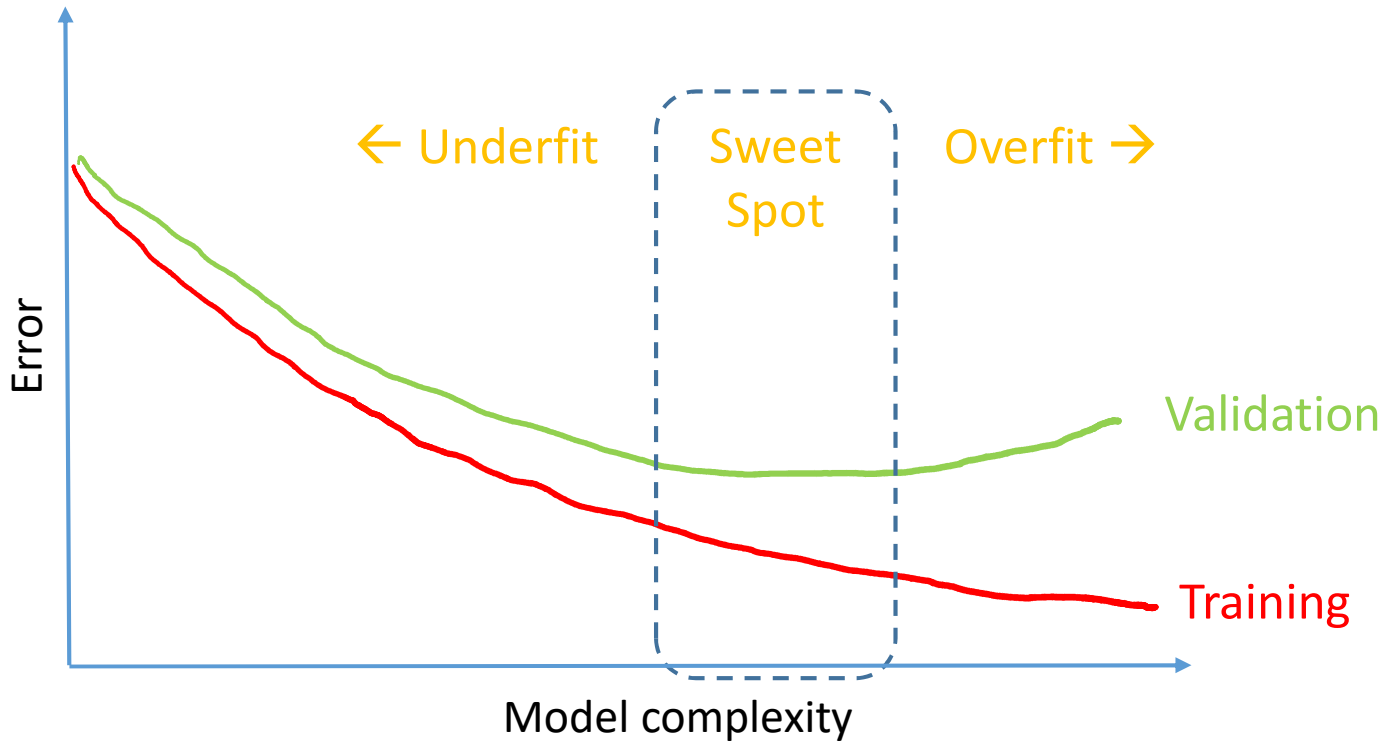
- Web of relationships

Images courtesy: Pixabay.com

Examples of ideal and non-ideal problems for ML

Ideal	**Not ideal**
• Recognize a thing, place, or person in photographs • Recreate speech from just lip movement • Divide music into genres and recommend based on past likes • Learn inverse models (performance → design) from forward models (design → performance) • Predict properties of new molecules	• Predict the next solar eclipse • Make a chat bot for an under-documented language • Predict which company will become the next unicorn • Develop a comprehensive expert system to replace doctors • Predict the onset of the next argument with spouse ☺

Model choice and rigorous validation are very important



What is model complexity anyway?

Modern ML models are mathematical functions with…

- Thousand of inputs, and
- Millions of parameters that are learned during training

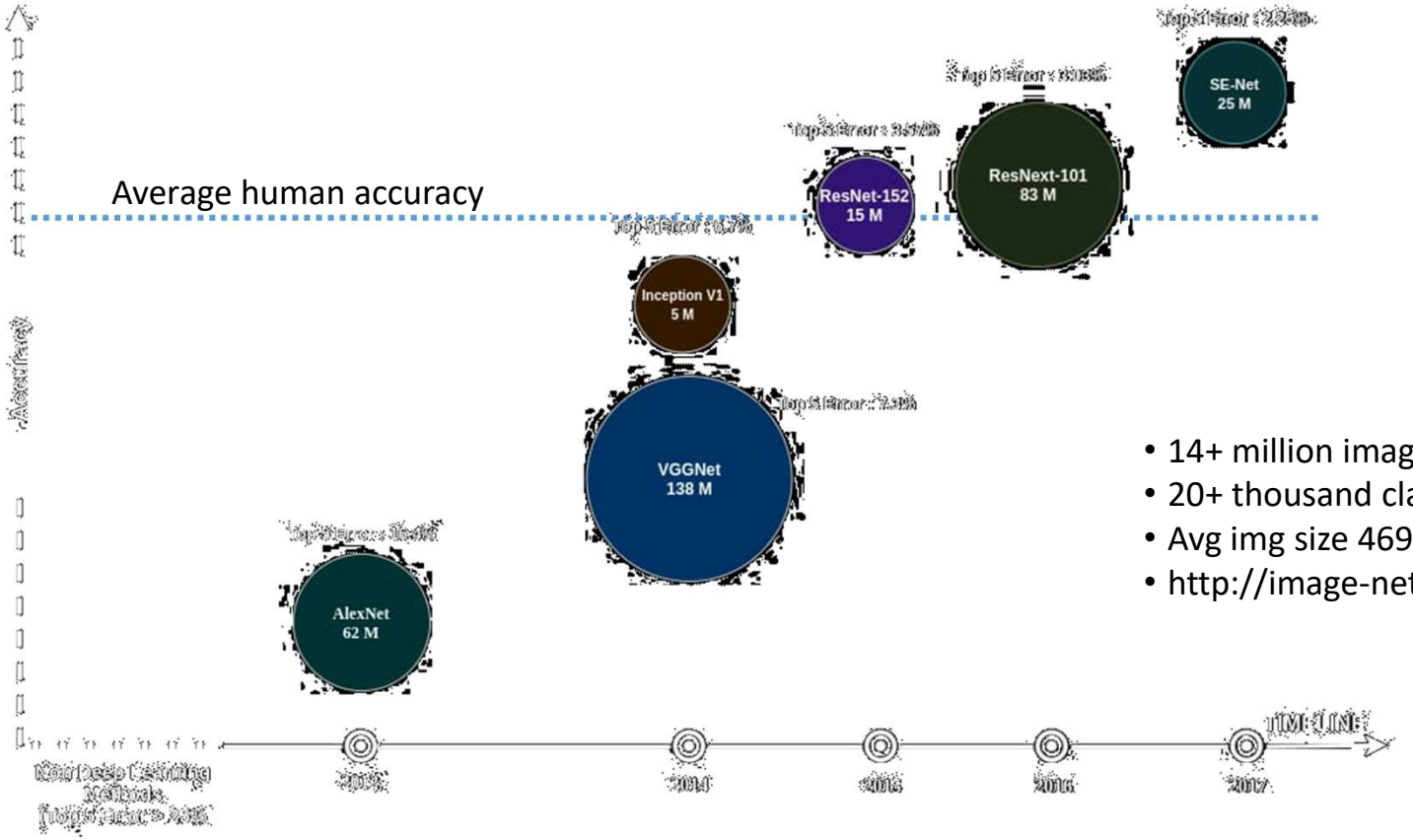
Increasing complexity means …

- More inputs to exploit more complex data interrelationships
- More parameters to model more complex relationships

Example of increasingly complex models:

- Simple model – one input, two unknowns:
  - $y = w_1 x_1 + w_0$
- Adding another input x_2:
  - $y = w_2 x_2 + w_1 x_1 + w_0$
- Adding nonlinear functions:
  - $y = \log(w_2 x_2 + w_1 x_1) + w_0$
- Adding more unknowns (w's):
  - $y = \log(w_2 x_2 + w_1 x_1) + w_3 x_1^2 + w_4 x_2^2 + w_0$
- Adding functions of functions:
  - $y = \log(w_2 \log(w_1 x_2 + w_4 x_1) + w_3 x_1^2$ …

Progress of DL on ImageNet

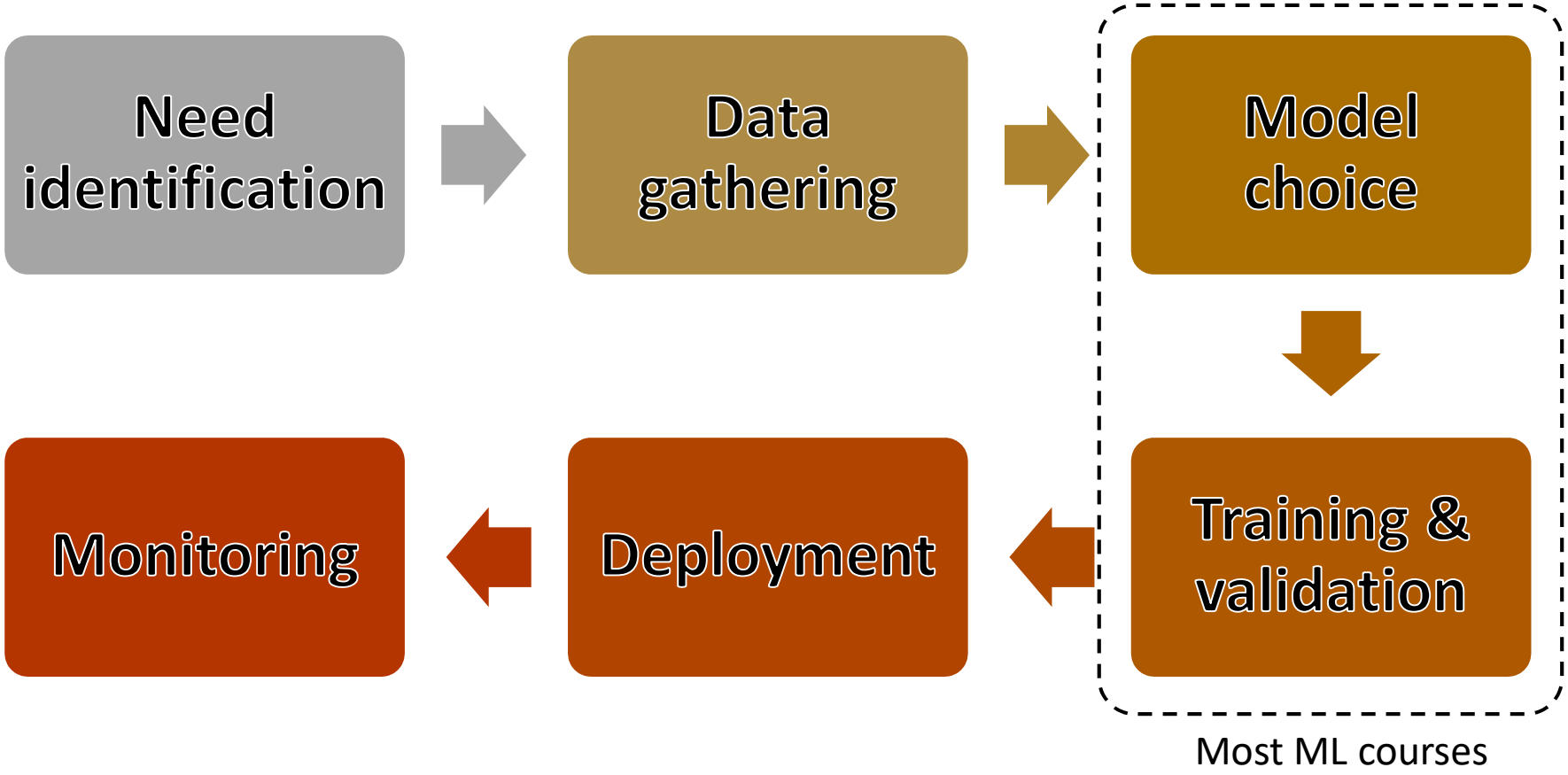


- 14+ million images
- 20+ thousand classes
- Avg img size 469x387
- http://image-net.org/

Other considerations

- Memory
- Transferability
- Computations and power
- Speed and parallelizability

ML life stages



Overlooked questions

- Problem
  - Right business or societal need
- Data and provenance
  - Relevant and enough
  - Diverse and representative
  - Ethically gathered, stored
  - Meticulously recorded
- Model
  - Exploits structure in data
  - Sufficiently complex
  - Not too complex
  - Meets deployment constraints
- Validation
  - Realistic; not too optimistic
  - Covers diversity of use cases
  - Meaningful performance metrics
- User guidance
  - Declaration of intended use case
  - Description of data used
- Monitoring in-field performance
  - Data differences, concept drift
  - Unethical and unauthorized use

Try to ask critical questions in response to the following statements

“We need to store every keystroke and mouse movement of use of company computers so that we can run ML on it later.”

“Our product can give biometric access based on face recognition. No need for cumbersome finger, ID, or iris scans.”

“We can detect pneumonia in chest x-ray with 99% accuracy.”

“We can recognize people’s emotions with 91% accuracy from their faces as they watch videos on our website.”

“Our video call plug-in can tell when people are lying.”

“Our autonomous vehicle is 20 times safer than an average driver.”

Directions in ML research

- Train with unlabeled data
- Train with fewer labeled data
- Train with high-level data labeling
- Take learning from one task to a new task
- Design models that can explain their decisions
- Make models more cautious by recognizing new scenarios

Myth or reality?

- AI is constantly learning and improving on its own
- AI can be self-aware
- One needs large investments to get started in AI
- AI is only for whiz-kids
- AI will only replace mundane jobs
- AI will completely replace humans in most jobs
- AI cannot be fooled easily
- AI can make sense of messy data
- AI can predict anything very accurately

Takeaways

- ML is not a silver bullet
  - Not necessary for simple models
  - Does not work well with ambiguous, general, and unstructured scenarios
  - Do not expect solid explanations for ML predictions any time soon
  - Our best bets are provenance, rigorous validation, and monitoring
- ML still useful in many scenarios
  - ML can also go beyond human-level capabilities in well-defined tasks
- Data is the currency of ML
  - Relevant
  - Clean
  - Realistic
- Critically design and accept ML models