



# Machine Learning is...

**Key Battleground for Free Technology**

# Freedom in Process and Results of Machine Learning (ML)

- ❖ Process
  - Tools used in research, development, deployment
  - **“Are the process and tools of ML free?”**
  - What does “free” mean in the question above?
- ❖ Results
  - Models, decisions, inferences
  - **“Are the eventual models and decisions free?”**
  - What does “free” mean in the question above?
- ❖ Why do we care?
  - “Free” as in “free software”

Today's focus: **Process**



# Overview

1. Crash intro: ML stack
2. History is still here: ML's proprietary legacy
3. Status quo: where are we now?
4. Proprietary is coming (again): Show me the ecosystem
5. Proposals: what worked, tips, and what more we can do

The background of the slide is a decorative pattern of squares in three colors: blue, yellow, and cyan. The squares are arranged in a somewhat irregular, grid-like pattern, with some squares missing or overlapping, creating a modern, abstract look. The colors are vibrant and high-contrast.

# ML Stack

# What is, and Why Machine Learning?

- ❖ Why?
  - A slow but gradual revival in R&D since “AI winter”
  - Exploding amount of data
  - Automated decision-making
- ❖ What?
  - “Statistical learning”
  - Buzzwords vs. Reality



# What ML is, and is NOT

## ❖ ML is...

- Andrew Ng (CS Professor at Stanford, former Chief Scientist at Baidu)

“[Machine learning is] the science of getting computers to act without being explicitly programmed”

- Trevor Hastie, Rob Tibshirani (Statistics Professors at Stanford)

“[A statistician’s job is]to extract important patterns and trends, and to understand “what the data says”. We call this learning from data.”

## ❖ ML is not...

- **...the magical and mysterious blackbox with a silver bullet that solves every problem there ever was, is, and will be.**

# History: Hypes and Buzzwords, 1958 version

## NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo  
of Computer Designed to  
Read and Grow Wiser

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human be-

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

### Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

1958 New York  
Times...

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

### Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

# Buzzwords != Reality

## ❖ ML Types

- Supervised: training data are all labeled
  - Deep learning belongs to this category (and the very resource intensive kind)
- Semi-supervised
- Unsupervised
- Human-in-the-Loop
- Game-theoretic
- ...etc.

## ❖ “No free lunch” (NFL) theorem

- “Two algorithms are equivalent when their performance is averaged across all possible problems.” (David Wolpert, 2005)
- AKA, “Your problems could be better solved by ML techniques that are not deep learning”

## ❖ “...But we already have all the free libraries for ML!”

- **...but libraries matter much less than you think**



# Five-step Process to Large-scale ML System

- ❖ Research
  - Interactive analysis, proof-of-concept model, building, training, evaluation, validation
- ❖ Development
  - Build user-facing applications, systems, infrastructures for the ML models from Research
- ❖ Scale
  - Distributed computing, querying, data/feedback ingestion and processing (streaming, batching, etc.)
- ❖ Deployment
  - Scheduling, logging, messaging, resource management, storage of data/results/feedback
- ❖ Maintenance
  - Re-train, re-tune, improve metrics and distance functions, add new features...
- ❖ You need an **ecosystem** to ML at scale (not just a library)

The background of the slide is an abstract pattern of overlapping squares in three colors: yellow, blue, and cyan. The squares are arranged in a non-uniform, grid-like fashion, creating a vibrant and modern aesthetic. A large white rectangular area is centered on the slide, containing the main title.

# Proprietary Legacy

# A Long-time Proprietary Legacy

## ❖ 1950s and on:

- Fortran
- Most optimized basic linear algebra subprograms (BLAS) through the years
- Still prevalent in legacy code of research, development, deployment, and maintenance
- Examples: National labs and research institutions (e.g. SLAC, NASA)

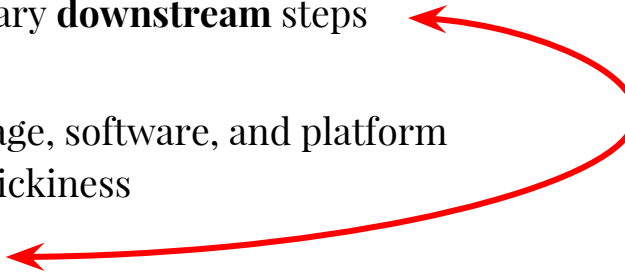
## ❖ 1970s and on

- “AI Winter”, big and small ones
- MATLAB
- Matrix manipulation and numerical computing
- Now often seen in research, development, and education
- Examples: Education, research prototypes

# It is still here, on a massive scale

- ❖ Education: from the roots...
  - Real-life classrooms: [Stanford ML](#), [Brown ML](#), [U Penn ML](#), [Columbia ML](#), [U Toronto ML](#), [U Utah ML](#), [U Pittsburgh ML](#), [UIUC Computer Vision](#), [UT Austin CV](#), [UMass ML](#), etc. etc.
  - Almost exclusive use of MATLAB
  - MOOCs: [Vanderbilt programming intro](#), [Rice EE](#), [Davidson College “Linear Algebra” X 2](#), [Berkeley EE](#), [EPFL MATLAB and Octave](#), etc.
  - **And yes, this one, too:** [Stanford ML](#), intro class by Andrew Ng
- ❖ Research prototypes
  - Famous ones: [t-SNE](#) (U Toronto), [local linear embedding](#) (NYU)
  - [Many papers in highly visible journals & conferences](#)
  - Fun fact: Deep autoencoders (a kind of neural network) was first and most famously [prototyped](#) in MATLAB circa 2006

# From the Roots of the Process...

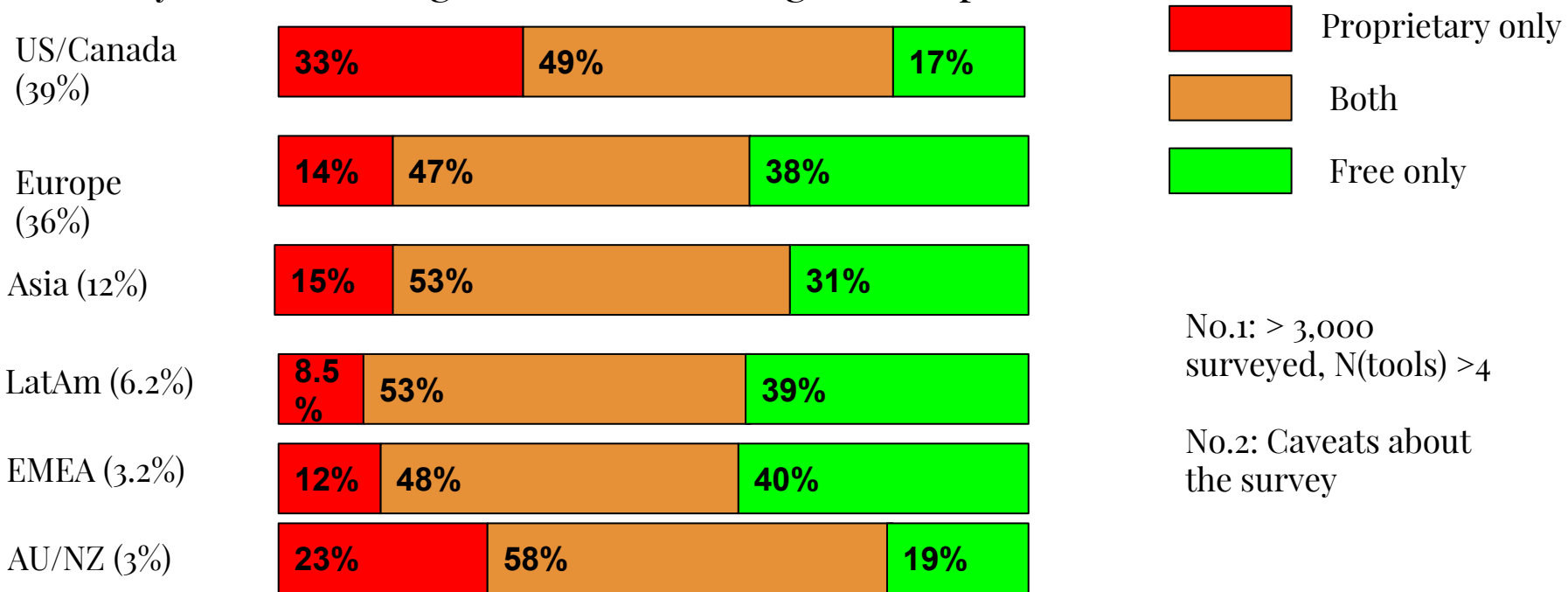
- ❖ Proprietary still has strongholds on the ML upstream steps
    - Education & Research
    - Why the upstream matters
      - First bait
      - Ecosystems start to form
    - Easier to integrate with proprietary **downstream** steps
  - ❖ Case in point: SAS
    - Proprietary programming language, software, and platform
    - On “honey traps” & consumer stickiness
  - ❖ Free software in **downstream**
    - Excellent individual software solutions (e.g. Apache family)
    - Ecosystem vs. Combinations of software solutions
- 

The background of the slide is an abstract pattern of overlapping squares in three colors: yellow, blue, and cyan. The squares are arranged in a somewhat regular grid but with some missing or overlapping, creating a textured, mosaic-like effect. The colors are vibrant and saturated.

**Where are we now?**

# Silver Lining? Or ... (1)

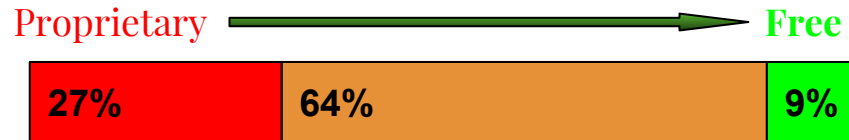
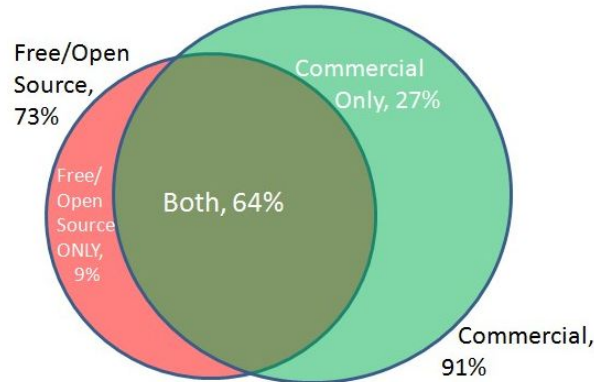
- ❖ FLOSS vs. Proprietary in Data Mining: [2014. KDNuggets Tool Survey](#)
- ❖ Proxy for measuring free software usage in ML process



# Silver Lining? Or ... (2)

- ❖ FLOSS vs. Proprietary in Data Mining: [2015. KDNuggets Tool Survey](#)
- ❖ Changes and caveats
  - Free vs. Proprietary breakdown **by region** disappeared
  - Free vs. Proprietary breakdown **entirely** disappeared in [2016 KDNuggets Tool Survey](#)
- ❖ “Do people actually care about free software anymore? Especially in ML?”

Analytics, Data Mining, Data Science  
Software Usage, 2015



Aggregate Shares of ML/Data Mining tool types, 2015.  
(Source: KDNuggets 2016 software pool. N = 2895)



# Silver Lining? Or ... (3)

- ❖ 2014 - 2016 KD Nuggets Tools Survey (multiple choice) aggregate. Based on top 10 tools from 2016

Name	2016 (%)	2015 (%)	2014 (%)	Color
R	49	46.9	38.5	Green
Python	45.8	30.3	19.5	Green
SQL	35.5	30.9	25.3	Orange
Excel	33.6	22.9	25.8	Red
RapidMiner	32.6	31.5	44.2	Red
Hadoop	22.1	18.4	12.7	Green
Spark	21.6	11.3	2.6	Green
Tableau	18.5	12.4	9.1	Red
KNIME	18.0	20	15	Green
scikit-learn	17.2	8.3	N/A	Green

Green going up & top!

Languages, library, & software alike!

Excel???

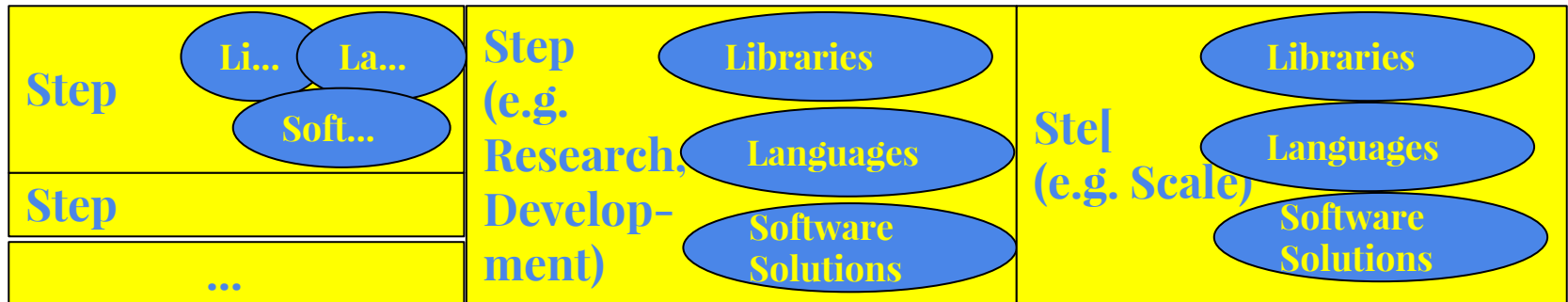
SQL...What kind?

What else?

# Words of Caution

- ❖ Bits and pieces != Whole fish
- ❖ Libraries and programming languages != Ecosystem
- ❖ What is an ML **ecosystem**?

**Ecosystem: where all steps work together, flow seamlessly**



The background of the slide is an abstract pattern of overlapping squares in three colors: yellow, blue, and cyan. The squares are arranged in a non-uniform, grid-like fashion, creating a vibrant and modern aesthetic. A white rectangular area is centered on the slide, containing the main text.

# **Proprietary Strikes Back**

# Major Proprietary Pushes into ML Ecosystems

## Big “Honey Traps”

[Amazon AWS AI](#): Any organization using Amazon’s cloud services

[Google Cloud Platform](#): Any Google enterprise platform or cloud service user

[Microsoft Azure](#): **See a pattern here???**


### PLAYBOOK:

Step 1: Hook you on one thing first...

Step 2: Build the proprietary **ecosystem** of ML around you...

Step 3: Profit. Rinse. Repeat.

## Small “Odd Couples/Trios”

1. FLOSS license-ed version
  2. FLOSS version+ technical support
  3. Build proprietary ecosystem around FLOSS version
- 

### PLAYBOOK:

Step 1: Modularized software solutions

Step 2: Build **ecosystem** with other proprietary ML software solutions

Step 3: Trapped by big honey traps

# Case Study: Revolution Analytics

- ❖ Individual software solutions → Ecosystem components
- ❖ Pre-Microsoft Acquisition
  - Start: FLOSS + tech support for R (GPL ver. 2/3)
  - CEO from SPSS » Pushed for ecosystem development
  - Initial ecosystem: **Independent** R Open (GPL ver. 2/3), R Open + technical support/services
- ❖ Post-Microsoft Acquisition
  - Current ecosystem: MRAN, R Server platform (servers, APIs, packages, apps, clients), R Open
  - **ALL** integrates with proprietary Azure, SQL Server, etc. (e.g. run Hadoop [ASL 2.0])
- ❖ Survival of the most free?
  - TIBCO got R a proprietary implementation
  - ...and built a honey trap ecosystem, too
  - ...and sells it big time

# Where is Freedom in 5-step ML process?

## ❖ Current status

- Libraries, programming languages: **Mostly, “YES”**
- Individual software solutions: **SOME**
- Ecosystem: **NO (not yet, at least)**

## ❖ Pro and Contra

- Lots of free ML libraries and software solutions
- What about an ecosystem that holds them together in the ML process?

## ❖ Where to find freedom in ML now?

- **“What should be”** vs. **“What is”**
- **Everywhere in the Process** vs. **Downstream**
- **Holistic Ecosystem** vs. **Individual Software Solutions + Libraries**

The background of the slide is a decorative pattern of overlapping squares in three colors: blue, yellow, and cyan. The squares are arranged in a somewhat irregular, grid-like fashion, creating a vibrant and modern aesthetic. A white rectangular area is centered on the slide, containing the text.

**So...What do we do?**

# Tips and Steps

## ❖ Have a plan

- Build free alternatives
- Find free alternatives, use free alternatives
- Collaboration: developers, technical users, non-technical users
- Starting free from the upstream: Education, R&D
- Build an **ecosystem!**

## ❖ Know the facts

- What ML is, is not, and what is ML good for
- Find ML techniques and types that work for **YOUR** particular problems

## ❖ Process before models

- “Garbage in, garbage out”

## ❖ Patience and persistence

- Step by step, over time, things will change



# More Thoughts on Freedom in ML

## ❖ Complex issues

- Licensing: “free” as in free software
- Development coordination: building an ecosystem is hard
- Viable business models to sustain FLOSS development
- Free process, should/can results be free? Should/can decisions be free?
- Capabilities and limits of ML
- ...etc.

## ❖ **Ecosystem** matters!

- Collaborate to build, use, maintain
- Don't take the first baits from the honey traps

## ❖ Start small, start from the start

- Education » Research » Development: free from the roots
- Building your own cluster
- Build free ecosystems from free components

# Resources (GPL or GPL-compatible Free Alternatives)

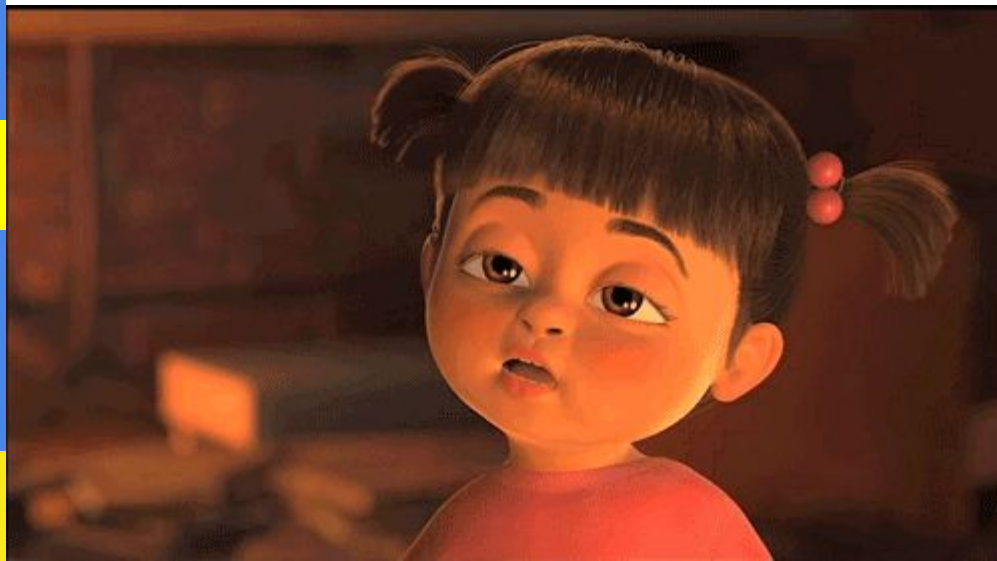
- ❖ Education: [Sage](#)
- ❖ Research: R/Python notebooks, Spark
- ❖ Development: [R Studio](#), [Shiny](#), Zeppelin
- ❖ Scale: Spark, Hadoop (Yarn, Hive), [Arvados](#)
- ❖ Deployment: Kafka, Lucene, ES stack, Storm, Samza, Flink, etc.
- ❖ Maintenance: Zookeeper, your choices of whatever kinds of free databases you need

# Pokemon or Big Data?

Let's play!



**Thank You!**



# Appendix

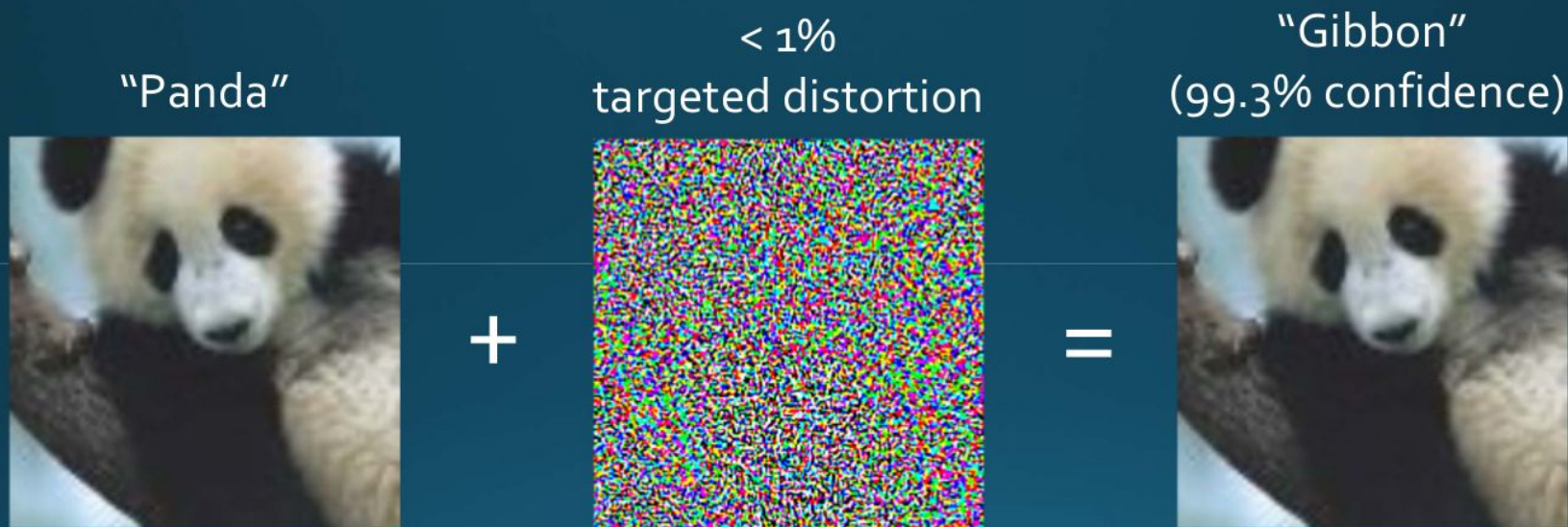


a young boy is holding  
a baseball bat

Statistically impressive,  
but individually unreliable

- ❖ “Statistically impressive but individually unreliable” (DARPA Perspective on AI, 2016)

# Appendix



Inherent flaws can be exploited

- ❖ An example of ML model poisoning (DARPA Perspective on AI, 2016)

# Appendix



Internet trolls cause the AI bot, Tay, to act offensively

Skewed training data creates maladaptation

- ❖ Another example of ML model poisoning (DARPA Perspective on AI, 2016)