Machine Learning with Clinical Data

Rami Vanguri

Biomedical Informatics / Data Science Institute
Columbia University

OSG All Hands Meeting
March 20, 2018
University of Utah, Salt Lake City, UT

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Data Science Institute

Disease Heritability using Electronic Health Records

- Heritability estimates the amount of variation in a trait due to genetics (vs environment)
  - Usually involves in-depth dedicated studies (twins, mice, etc)
  - Limited sample sizes

By using emergency contact information at Columbia University Medical Center, we can infer 4.7 million familial relationships and use them to estimate disease heritabilities.

Calculating Heritability

- Traits are assigned in electronic health records via insurance billing codes (ICD-9/10)
- Able to compute heritability for traits not typically accessible with traditional studies (such as neurological)
- Each trait (thousands) was submitted as a job on OSG

Dichotomous Disease Category	Median h_o^2	Trait with Highest Heritability: ICD9 Code	Name	Median h_o^2 (95% CI)	Trait with Lowest Heritability: ICD9 Code	Name	Median h_o^2 (95% CI)
Hematologic Diseases	0.50	287.31	Immune thrombocytopenic purpura	0.71 (0.33-0.96)	285.9	Anemia	0.20 (0.15-0.36)
Mental Health Diseases	0.41	309.28	Adjustment disorder with mixed anxiety and depressed mood	0.95 (0.36-1.00)	315.39	Other developmental speech or language disorder	0.11 (0.09-0.15)
Sense Organs Diseases	0.41	365.11	Primary open angle glaucoma	0.93 (0.52-1.00)	382.9	Unspecified otitis media	0.10 (0.06-0.16)
Endocrine and Metabolic Diseases	0.40	278.02	Overweight	0.71 (0.54-0.88)	272.4	Other and unspecified hyperlipidemia	0.23 (0.15-0.37)
Gastrointestinal Diseases	0.39	579	Celiac disease+	0.78 (0.55-0.97)	521	Dental caries	0.12 (0.07-0.18)
Infectious Diseases	0.34	111	Pityriasis versicolor	0.85 (0.50-0.94)	780.6	Fever	0.11 (0.05-0.23)
Respiratory Diseases	0.34	477.9	Allergic rhinitis, cause unspecified+	0.72 (0.25-0.93)	464.4	Croup	0.09 (0.05-0.12)
Cardiovascular Diseases	0.33	785.2	Undiagnosed cardiac murmurs	0.59 (0.42-0.84)	786.59	Other chest pain	0.18 (0.11-0.25)

Dichotomous Disease Category	Median h_o^2	Trait with Highest Heritability: ICD10 Code	Name	Median h_o^2 (95% CI)	Trait with Lowest Heritability: ICD10 Code	Name	Median h_o^2 (95% CI)
Pregnancy, Childbirth and Puerperium	0.54	O30	Multiple gestation	0.76 (0.36-1.00)	O30-O48	Maternal care related to the fetus and amniotic cavity and possible delivery problems	0.41 (0.19-0.61)
Hematologic Diseases	0.45	D57	Sickle-cell disorders*	0.97 (0.75-1.00)	D64	Other anemias	0.18 (0.11-0.30)
Injury and Poisoning	0.40	T59	Toxic effect of other gases, fumes and vapors	0.81 (0.49-0.98)	S01	Open wound of head	0.18 (0.10-0.36)
Infectious Diseases	0.40	B35	Dermatophytosis	0.81 (0.41-0.98)	B80	Enterobiasis	0.11 (0.04-0.13)
Genitourinary Diseases	0.37	N92	Excessive, frequent and irregular menstruation	0.85 (0.62-0.99)	N80-N98	Noninflammatory disorders of female genital tract	0.15 (0.09-0.20)
Respiratory Diseases	0.35	J01	Acute sinusitis	0.85 (0.61-0.98)	J02	Acute pharyngitis	0.02 (0.01-0.03)
Eye Diseases	0.34	H35	Other retinal disorders	0.55 (0.33-0.77)	H10	Conjunctivitis	0.18 (0.10-0.22)
Gastrointestinal Diseases	0.34	K90	Intestinal malabsorption	0.84 (0.69-0.98)	K02	Dental caries	0.14 (0.09-0.20)
Endocrine and Metabolic Diseases	0.34	E20-E35	Disorders of other endocrine glands	0.60 (0.28-0.89)	E84	Cystic fibrosis*	0.01 (0.01-0.02)
Cardiovascular Diseases	0.33	I15	Secondary hypertension	0.50 (0.31-0.89)	IX	Diseases of the Circulatory System	0.18 (0.10-0.28)
Skin Diseases	0.32	L70	Acne+	0.72 (0.20-0.91)	L80-L99	Other disorders of the skin and subcutaneous tissue	0.17 (0.11-0.29)
Ear and Mastoid Diseases	0.31	H61	Other disorders of external ear	0.82 (0.68-0.93)	H66	Suppurative and unspecified otitis media	0.11 (0.06-0.22)
Mental Health Diseases	0.31	F93	Emotional disorders with onset specific to childhood	0.78 (0.27-1.00)	F40-F48	Anxiety	0.02 (0.01-0.03)
External Causes of Morbidity and Mortality	0.31	V49	Car occupant injured in other and unspecified transport accidents	0.94 (0.87-0.99)	V04	Pedestrian injured in collision with heavy transport vehicle or bus	0.01 (0.00-0.01)
Signs and Symptoms	0.30	R92	Abnormal findings on diagnostic imaging of breast	0.48 (0.26-0.65)	R62	Lack of expected normal physiological development	0.07 (0.05-0.10)
Musculoskeletal Diseases	0.27	M71	Other bursopathies	0.61 (0.25-0.99)	M00-M25	Arthropathies	0.18 (0.11-0.25)
Congenital malformations	0.27	XVII	Congenital Malformations	0.73 (0.50-0.96)	Q85	Phakomatoses	0.05 (0.00-0.09)
Neoplasms	0.25	D23	Other benign neoplasms of skin	0.35 (0.20-0.53)	II	Neoplasms	0.17 (0.08-0.27)
Perinatal Diseases	0.22	XVI	Certain Conditions Originating In the Perinatal Period	0.62 (0.45-0.84)	P00-P04	Newborn affected by maternal factors and by complications of pregnancy	0.05 (0.01-0.08)
Neurological Diseases	0.17	G47	Sleep disorders+	0.31 (0.19-0.48)	G44	Other headache syndromes	0.02 (0.01-0.03)

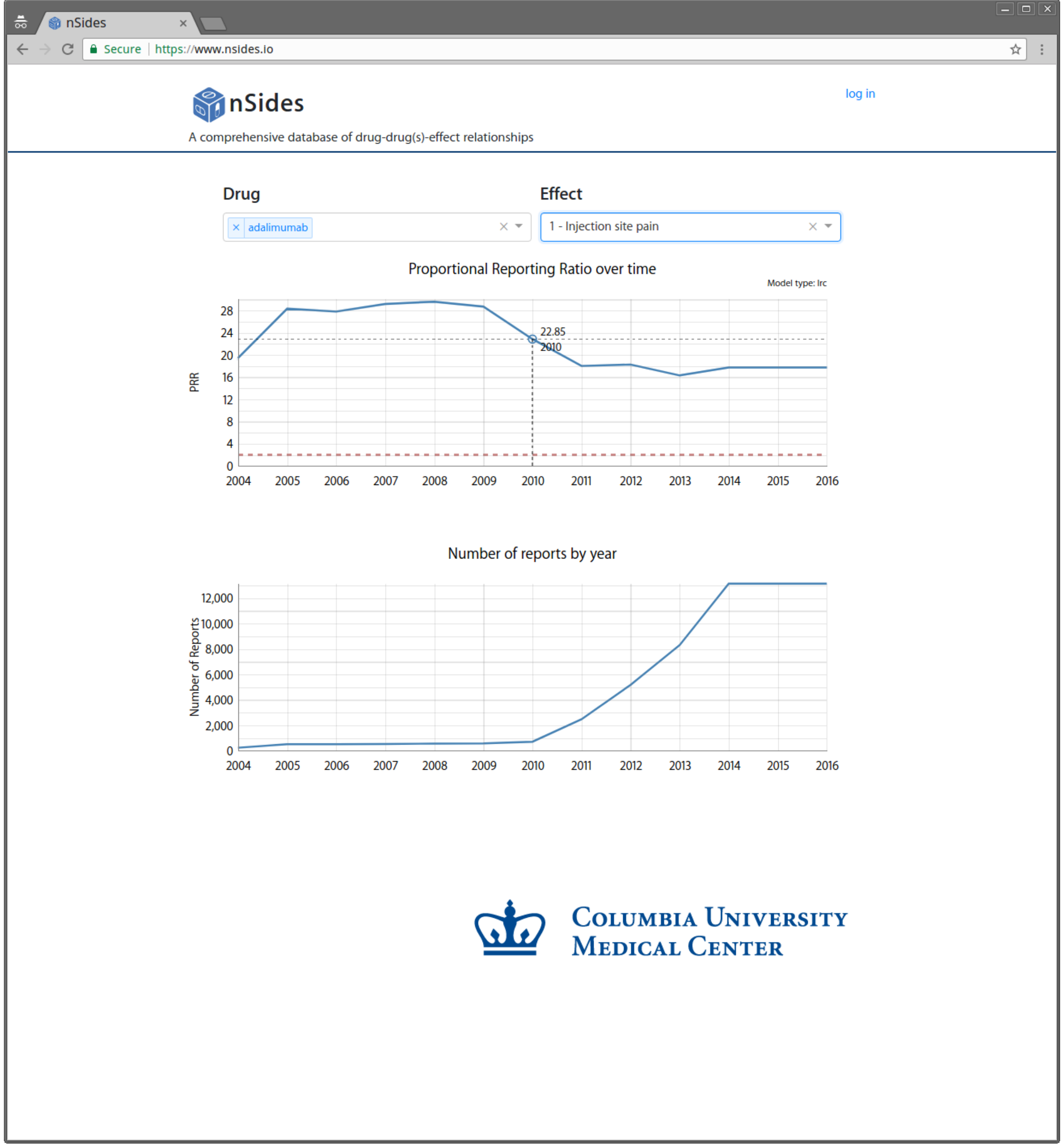
Paper just accepted to Cell!

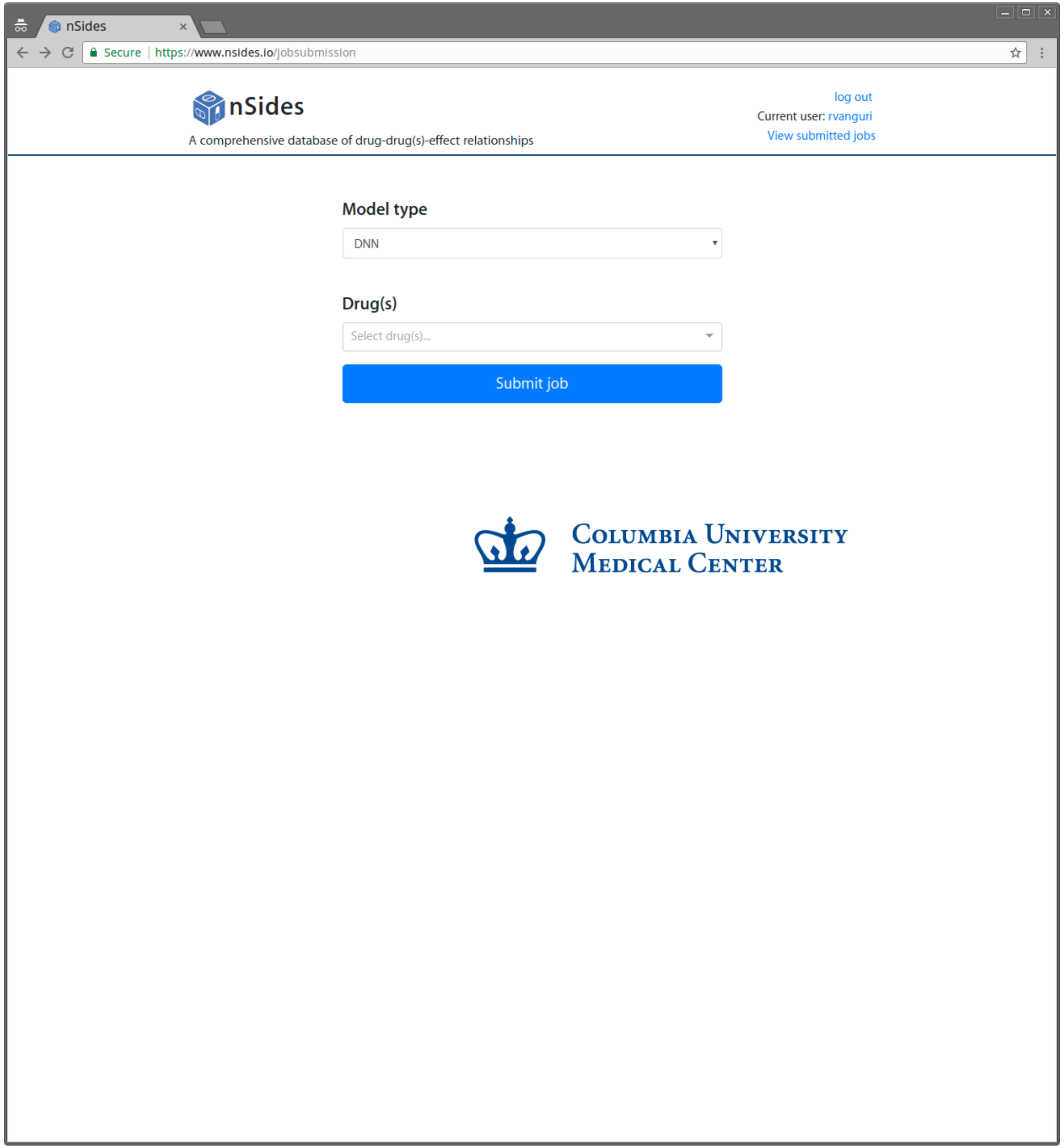
Data-Driven Drug Safety

- **Objective:** Mine the FDA Adverse Event Reporting System (FAERS) for statistically significant drug effects and interactions of multiple drugs
  - Reports from 2004-2015
- **Motivation:** Clinical trials often lack statistics to find rare drug effects, drug interactions even more difficult
- **Method:** Machine learning techniques are used to match cases/controls to calculate statistical significances
  - GPU turned out to not be that useful
- **Result:** Hypothesis generator for further investigation

nsides: Data-Driven Drug Effect Gateway

- Front-end: Public facing web gateway
- Middleware: Request drug interactions not already in database
  - Impossible to prospectively mine all possible drug interactions
  - Done via Agave with assistance from Science Gateways Community Institute (Choonhan Youn)
- Back-end: Each drug/interaction is setup as a DAG job
  - Initial population of 4500 drugs
  - Second population of prioritized drug interactions


nSides
Secure
https://www.nsides.io
nSides
log in
A comprehensive database of drug-drug(s)-effect relationships
Drug
adalimumab
Effect
1 - Injection site pain
Proportional Reporting Ratio over time
Model type: lrc
PRR
28
24
20
16
12
8
4
0
22.85
2010
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Number of reports by year
Number of Reports
12,000
10,000
8,000
6,000
4,000
2,000
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Columbia University
Medical Center


nSides
Secure
https://www.nsides.io/jobsubmission
nSides
A comprehensive database of drug-drug(s)-effect relationships
log out
Current user: rvanguri
View submitted jobs
Model type
DNN
Drug(s)
Select drug(s)...
Submit job
Columbia University
Medical Center

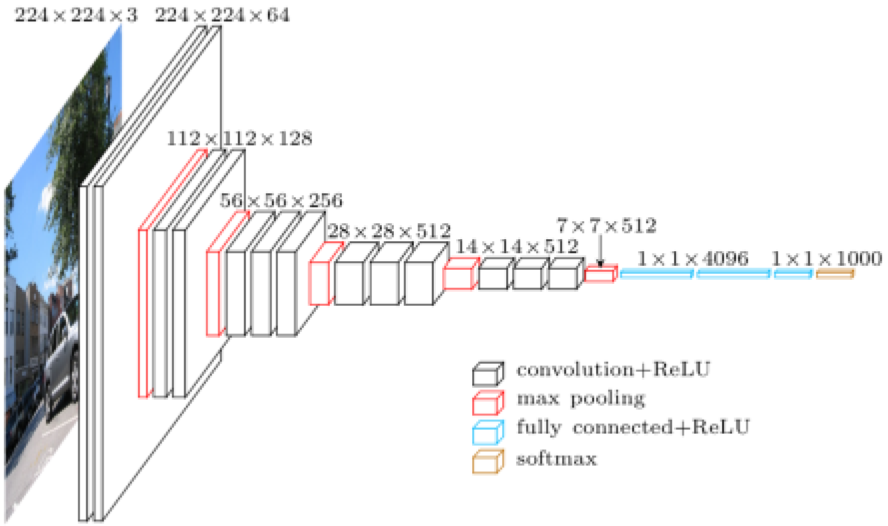
Looking Forward: Medical Imaging

- Starting July, transitioning to biomedical engineering/radiology
- Machine learning in medical imaging becoming very popular
- First ISMRM Machine Learning Workshop last week in California
  - ~60 presentations, 85 posters, full house
  - Vast majority used deep learning with GPU setups
- Variety of use cases:
  - Reconstruction: Constructing high quality imaging from undersampled data
  - Post-processing: Artifact correction
  - Clinical application: Segmentation, disease outcome and progression prediction
- Interest from clinicians, scientists and engineers!
  - Large diversity in computing abilities

Deep Learning

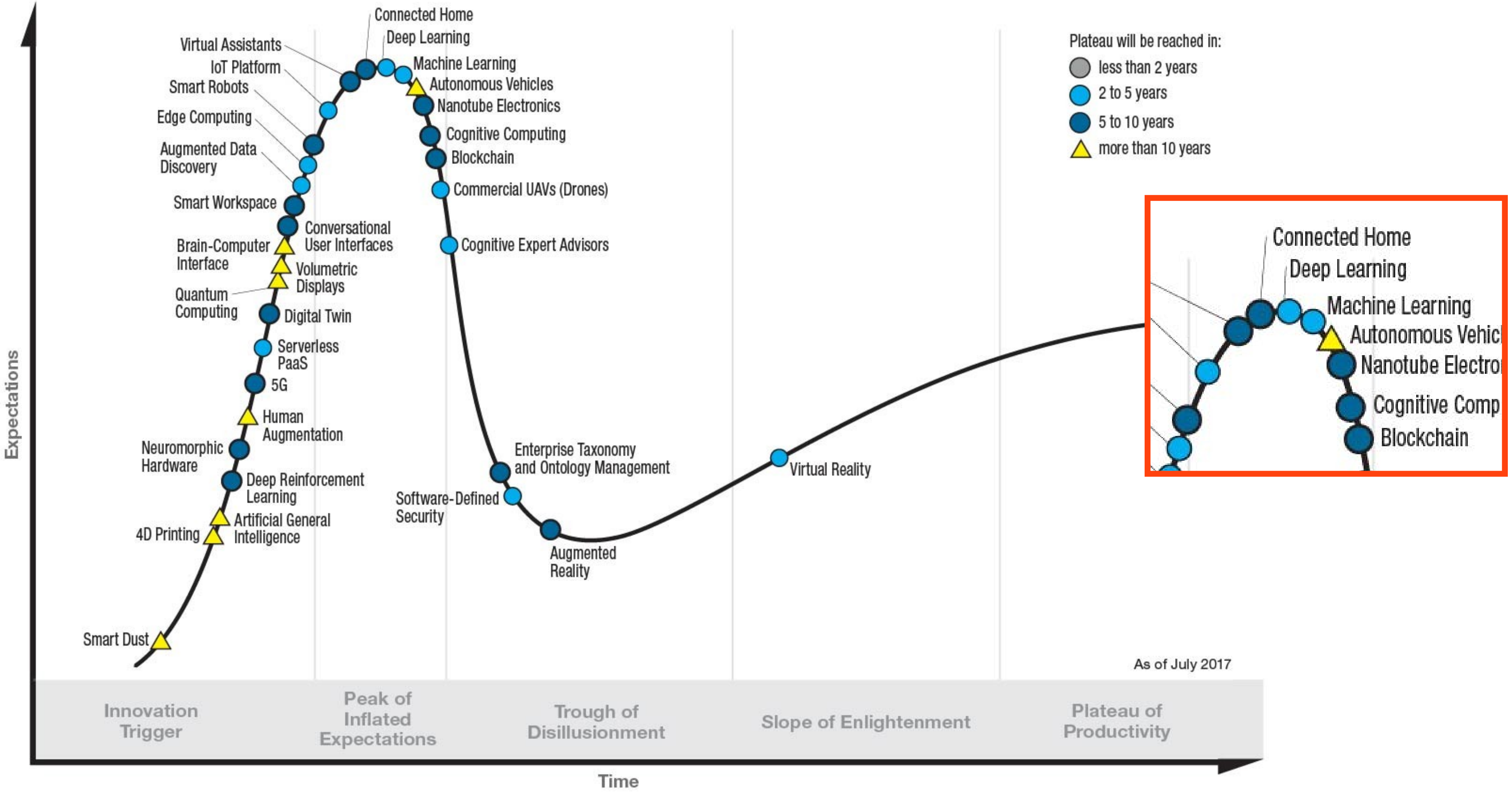
- Machine learning algorithms which uses multiple layers to extract and transform features
- Popular architectures: AlexNet, VGG Net, GoogleNet, ResNet, U-net, GAN
- Increase in performance, computing requirements and data

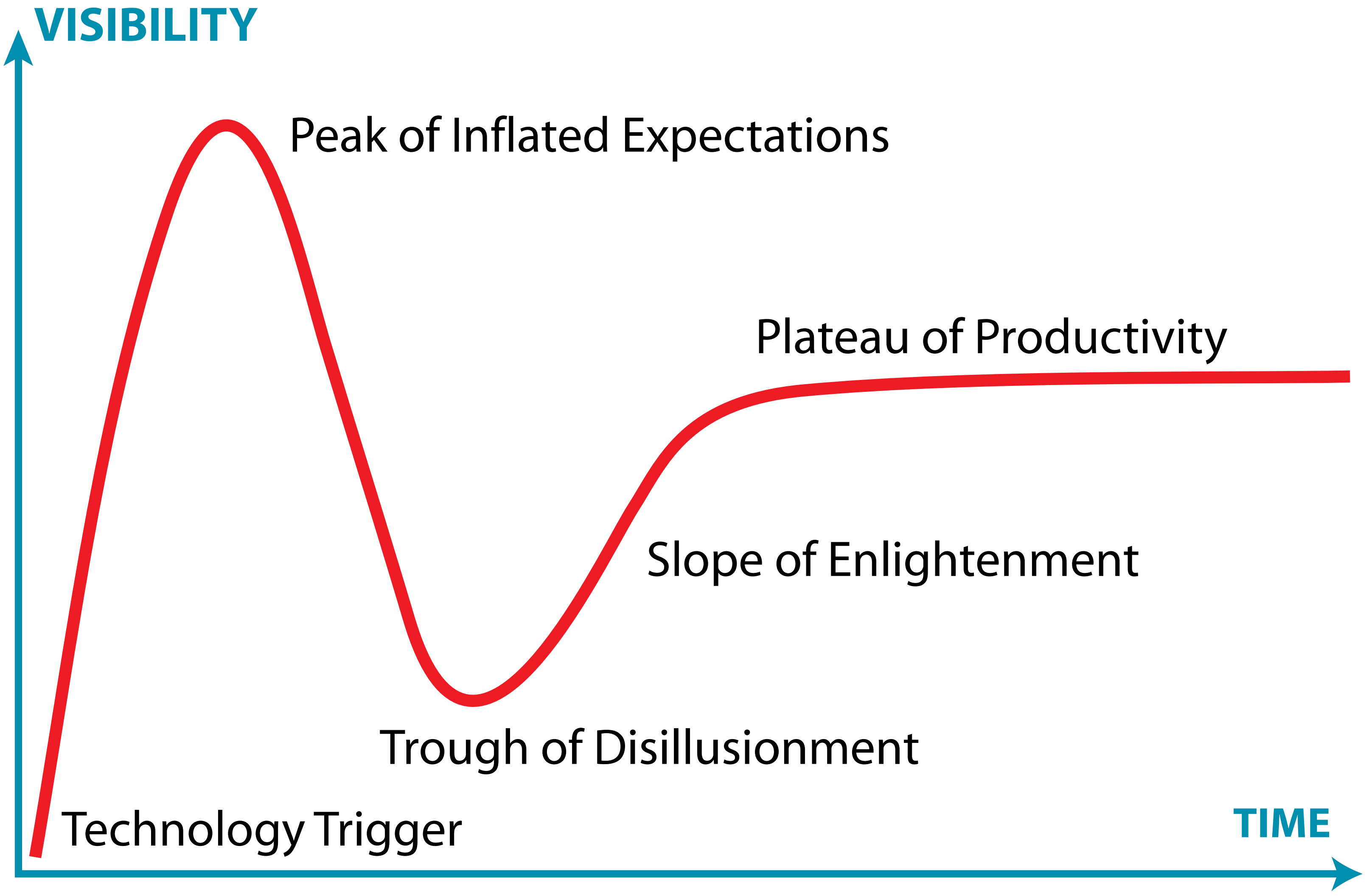
source



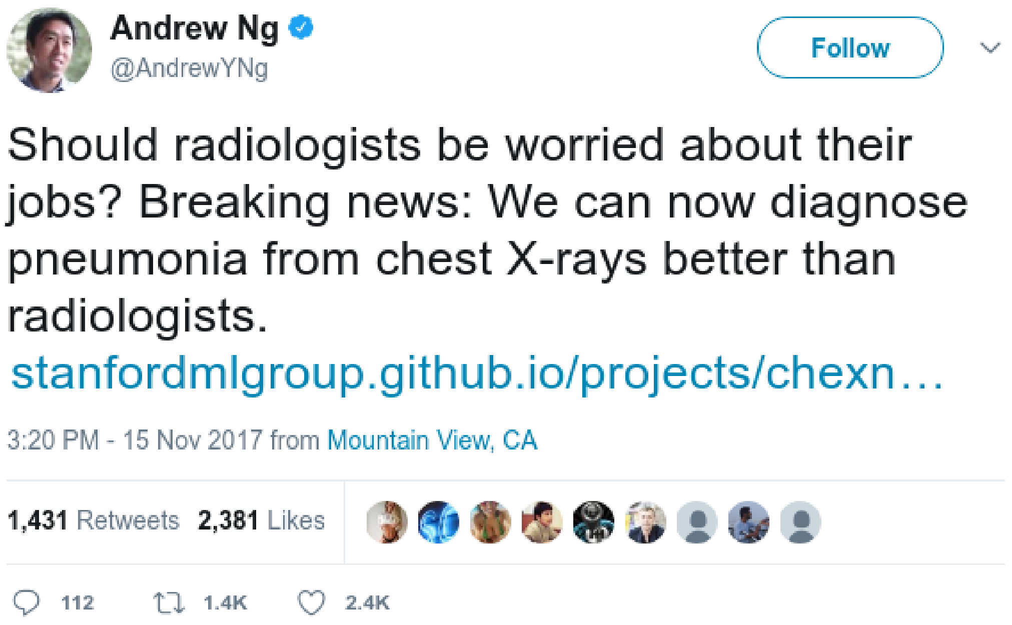
VGG-16
138M parameters!

Gartner Hype Cycle for Emerging Technologies, 2017




VISIBILITY
Peak of Inflated Expectations
Plateau of Productivity
Slope of Enlightenment
Trough of Disillusionment
Technology Trigger
TIME

Hype



Andrew Ng
@AndrewYNg
Follow

Should radiologists be worried about their jobs? Breaking news: We can now diagnose pneumonia from chest X-rays better than radiologists. stanfordmlgroup.github.io/projects/chexn…

3:20 PM - 15 Nov 2017 from Mountain View, CA

1,431 Retweets 2,381 Likes

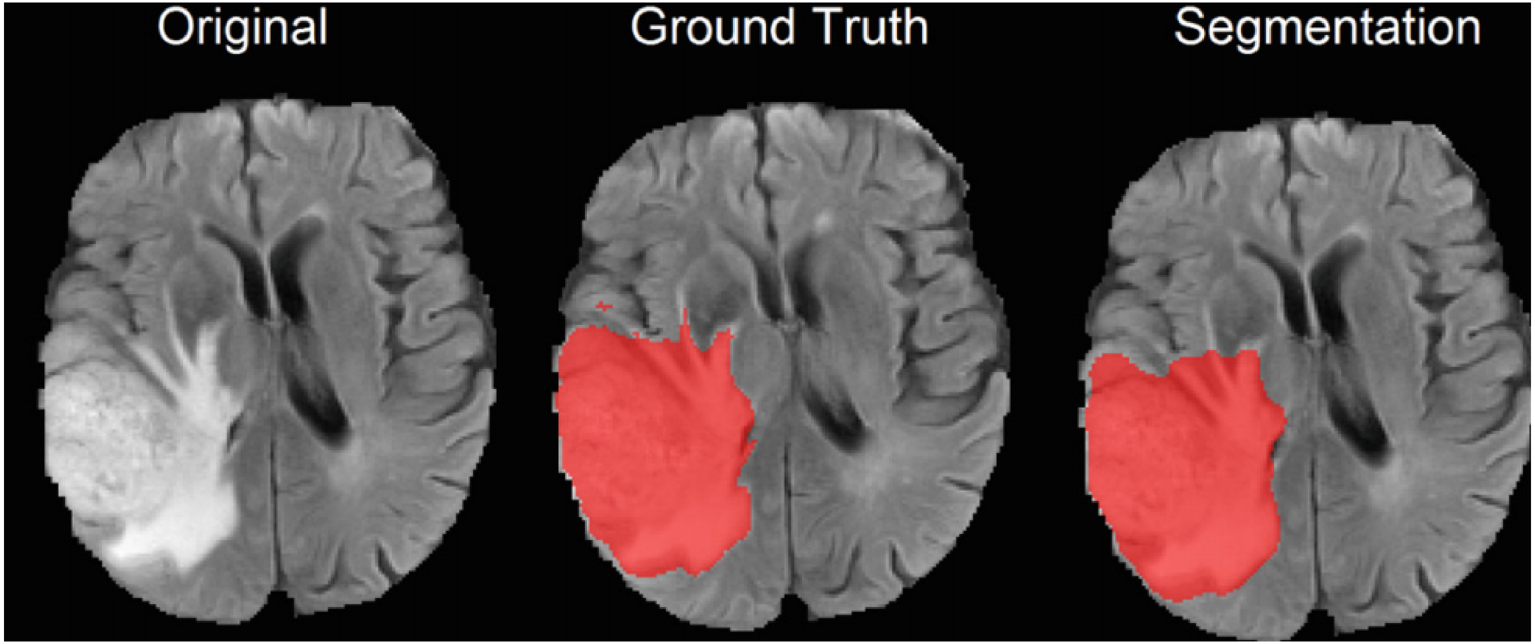
112 1.4K 2.4K

...probably not...

Example Use Cases

Clinical

- Segmentation is essential task during radiotherapy planning
- Automatic Brain Tumor Detection and Segmentation Using U-Net Based Fully Convolutional Networks
- Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, Yike Guo



Clinical

- Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks
- Used Convolutional Neural Networks to differentiate clinically significant tumors as candidates for therapy vs clinically insignificant tumors for safety surveillance

source

Science

- Elucidation of biomarkers
- Tricky with the nature of deep learning since feature importances aren't always clear
- Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects
- Used shallow machine learning to help identify Mild Cognitive Impairment patients at high risk for conversion to Alzheimers

source

Engineering

- Deep artifact learning for compressed sensing and parallel MRI
- Uses down-sampled data to reconstruct MR images
- **Acquisition with lower scan time**

source

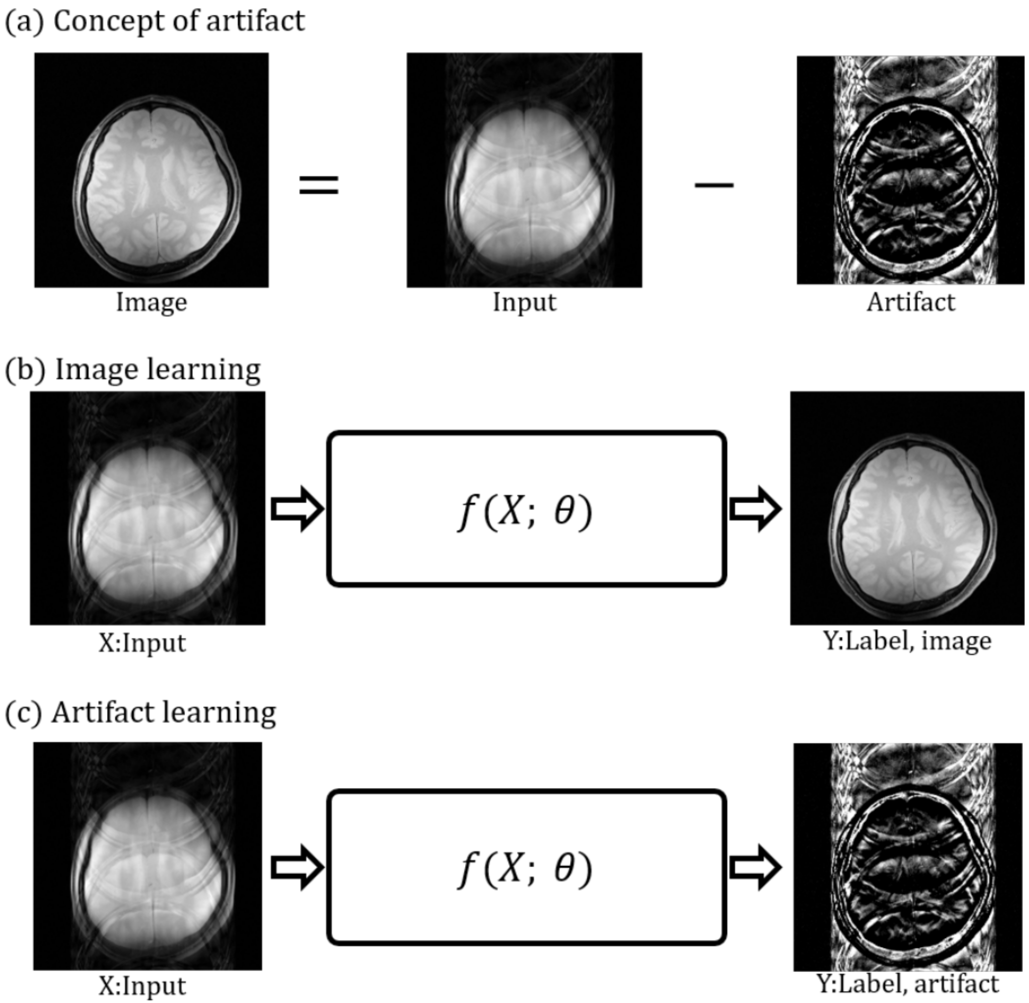


Figure 1: Concept of artifact learning. (a) The artifact image is defined as the difference between the aliased image and the artifact-free image in magnitude and phase domain. (b) Image learning: the aliased image is mapped to the artifact-free images. (c) Artifact learning: the aliased image is mapped to the artifact image. Once the artifact image is estimated, the artifact corrected image can be obtained by subtracting the estimated artifact from the input image.

Computing with Medical Imaging

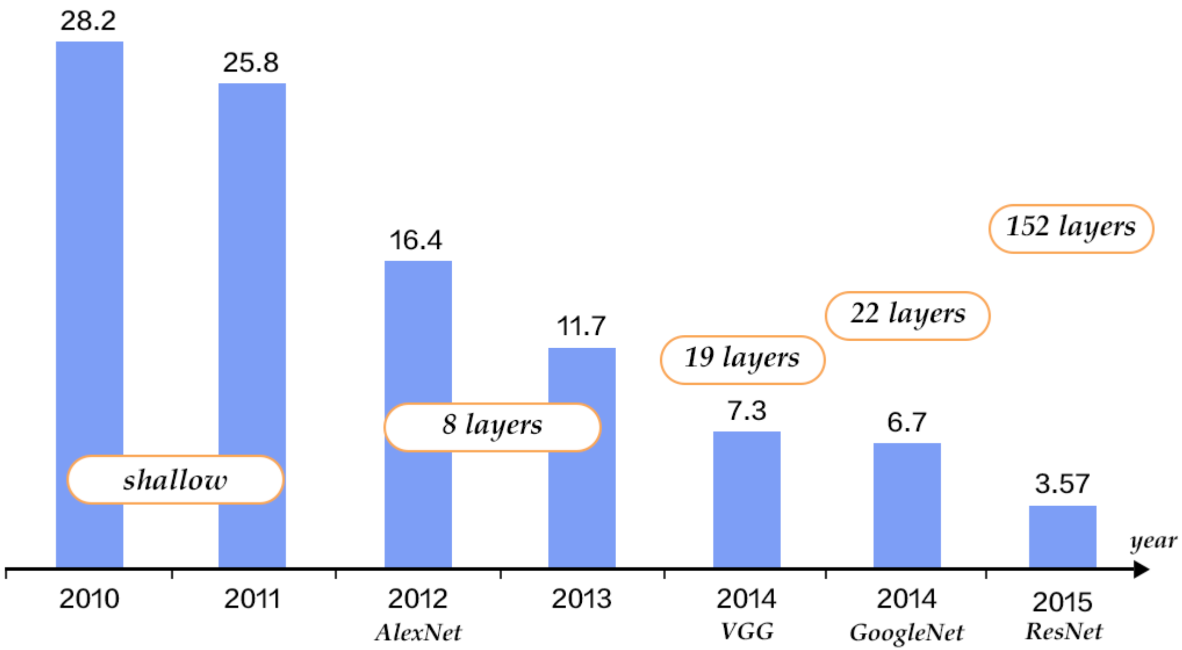
- Training machine learning networks almost always done with GPUs
- Current model is to buy a GPU machine and run locally within institute or buy time on commercial clouds
  - HIPAA compliance with clinical data available on AWS/MS
  - Knowledge of OSG’s existence is limited

Network Pre-training

- Clinical medical imaging studies often lack sufficient statistics for deep learning
  - Data augmentation helps: rotations, flipping, translation
- Overwhelming trend at workshop to use pre-trained networks
  - Decent results starting with just ImageNet
- Discussion centered around using other large public radiology data
  - Human Connectome Project
  - The Cancer Imaging Archive

ImageNet

- Database containing 14 million images which are hand-annotated



Open Science Grid

- Challenges:
  - Data involved is Protected Health Information covered by HIPAA
  - Datasets are large, especially ones typically used for pre-training
  - Jobs can be very long and not easily segmented
  - Accessibility to clinical researchers

Open Science Grid

- Pre-training can be done on OSG
  - Repository for public imaging data similar to dbGap?
  - Potential model is to pre-train on OSG and fine-tune at home institute
- Hyperparameter optimization during fine-tuning is very suitable for OSG resources
- Engineering projects could involve non-HIPAA data
- Analysis containers with Tensorflow and/or PyTorch
- Time to strike is now before trough of disillusionment

Next Steps

- Wrap up **nsides.io** in the next few months, making sure it's sustainable
- Develop public imaging deep learning analysis pipeline that can be deployed on OSG
  - Pre-training on public radiology data
  - Hyperparameter optimization
- Calculate value of adding other types of clinical data into classifiers
- Develop strategy for releasing networks and evaluation at other institutes
- Gracefully end fellowship in trough of disappointment