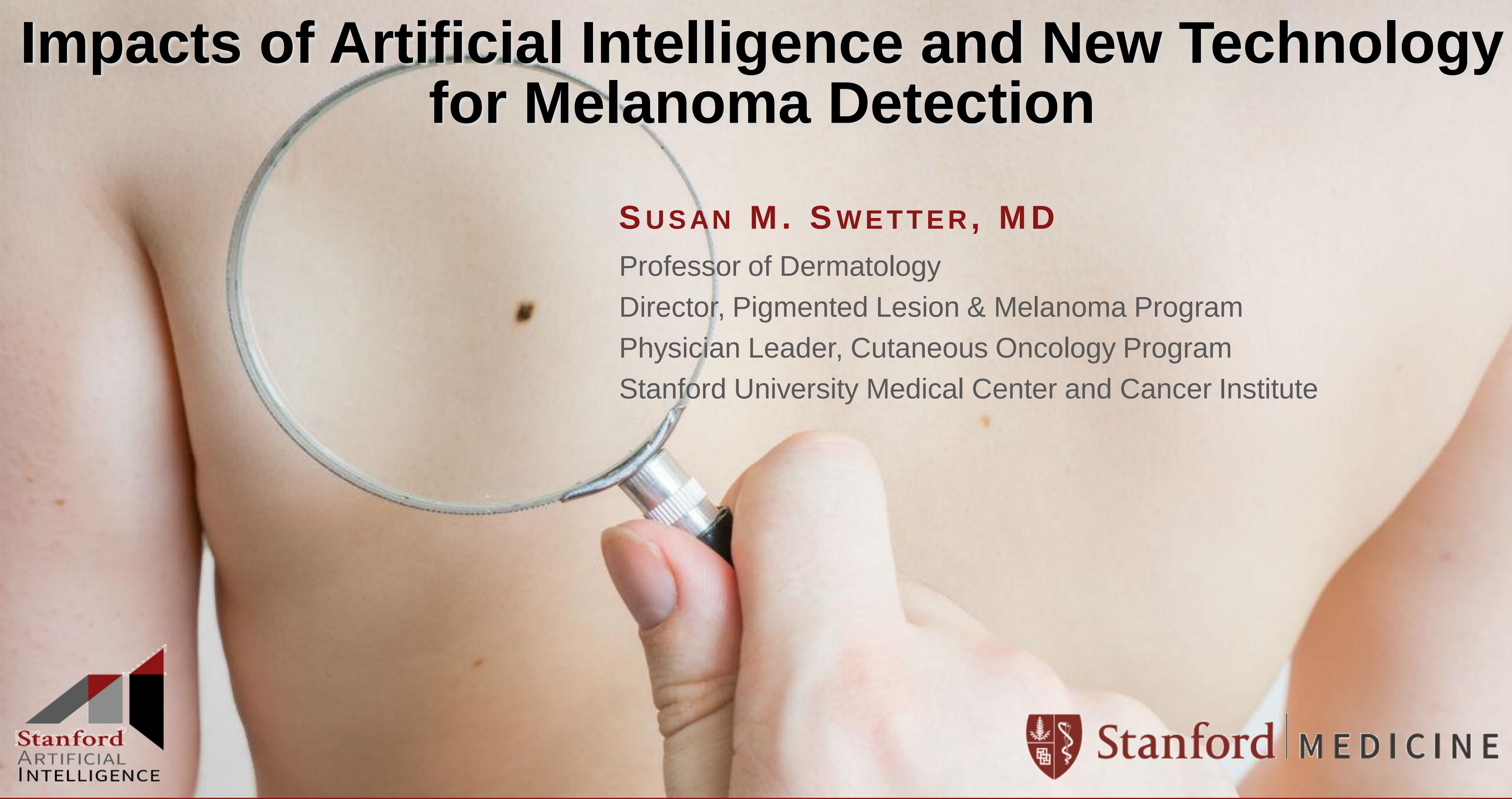


Impacts of Artificial Intelligence and New Technology for Melanoma Detection

Susan M. Swetter, MD

Professor of Dermatology

Director, Pigmented Lesion & Melanoma Program

Physician Leader, Cutaneous Oncology Program

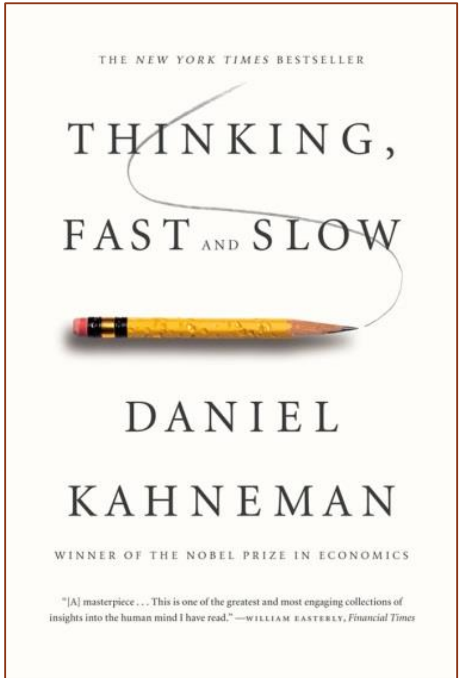
Stanford University Medical Center and Cancer Institute

Stanford Artificial Intelligence

Stanford MEDICINE

How do we know what we know?

- Unconscious pattern matching
- Rules-based systems



<u>Systems drive the way we think</u>:
System 1 – fast, intuitive, emotional
System 2 – slower, more deliberate and logical

Stanford
MEDICINE

Dual-process theory applied to Dermatologists

System 1 (intuitive) **"visual intelligence"** – rapid, operates below level of perceptible consciousness ("gut feeling")

- focuses on pattern recognition: "blink, think" and "10-second rule" (*Giuseppe Argenziano, Naples*)

- why "ugly duckling" rule discriminates melanoma more accurately than ABCD clinical warning signs (*Gaudy-Marquest C et al. JAMA Dermatol 2017*)

System 2 (analytical) – deliberate judgment, based on conscious applications of rules acquired through learning

Transition from intuitive to analytical reasoning can hinder clinical reasoning and increase diagnostic error

Stanford ARTIFICIAL INTELLIGENCE

Pelaccia T et al. *Med Educ Online* 2011. Stolper E et al. *BMC Fam Pract* 2009.
Norman GR, Eva KW. *Med Educ* 2010.

Stanford MEDICINE

Skin Cancer Facts

- Skin cancer - most common cancer in the US
- 1 in 5 Americans will develop skin cancer in their lifetime
- Latest estimate (2012): >5.4 million cases BCC/SCC treated in >3.3 million persons in the US (*Rogers HW et al. Arch Dermatol 2015*)
- In 2019 –
  - estimated >96K new cases of invasive melanoma and >95K melanomas in situ in the US
  - nearly 8000 melanoma-associated deaths
- Survival rate for melanoma is >95% if detected early

Stanford Artificial Intelligence

Siegal RL et al. Cancer Statistics, 2019. *CA Cancer J Clin* 2019

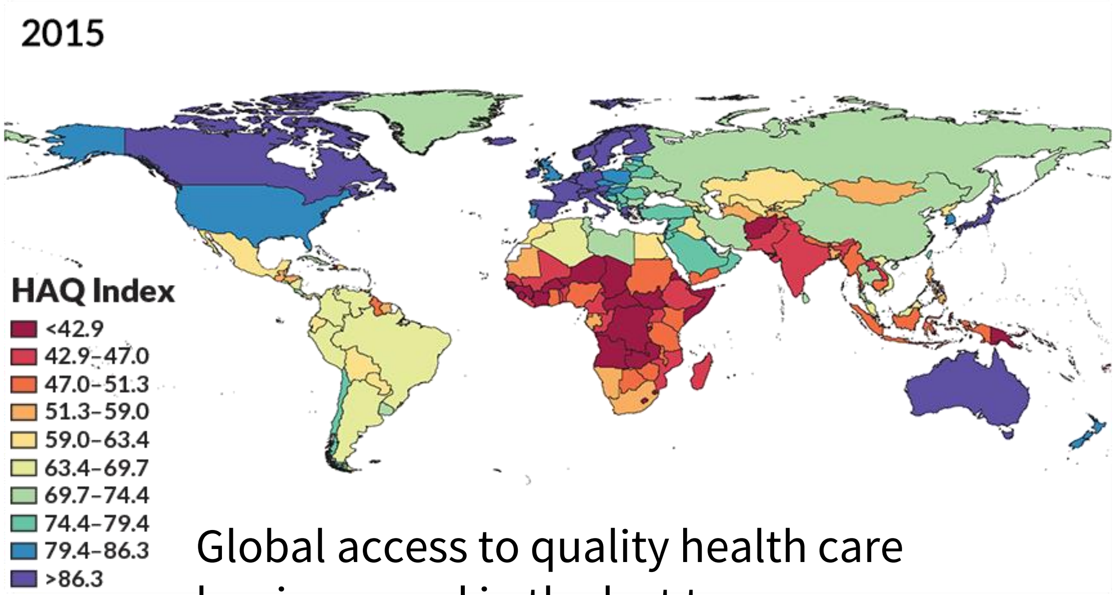
Stanford MEDICINE

“MELANOMA WRITES ITS MESSAGE IN THE SKIN WITH ITS OWN INK AND FOR ALL OF US TO SEE”

-Neville Davis, Queensland, Australia

…so why is early detection so hard?

Global Health Care Accessibility



Global access to quality health care has improved in the last two decades, but the gap between countries with the most and least effective treatments has grown

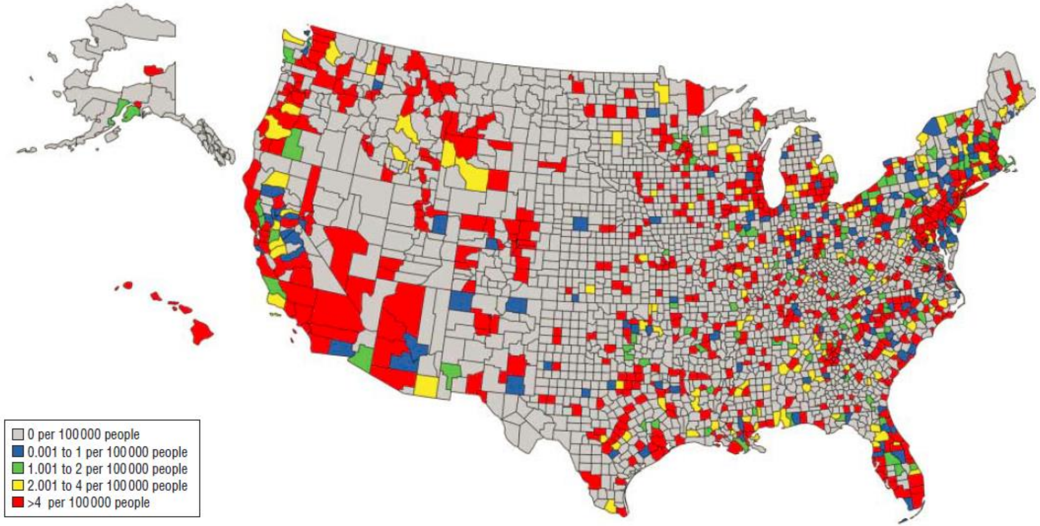
Stanford Artificial Intelligence

Stanford MEDICINE

Barber R et al. *The Lancet* 2017

How do we democratize health care access?

- **6.3 billion smartphones globally by 2021**
- Data collection at scale
- Diagnostics at scale
- Diminish health disparities



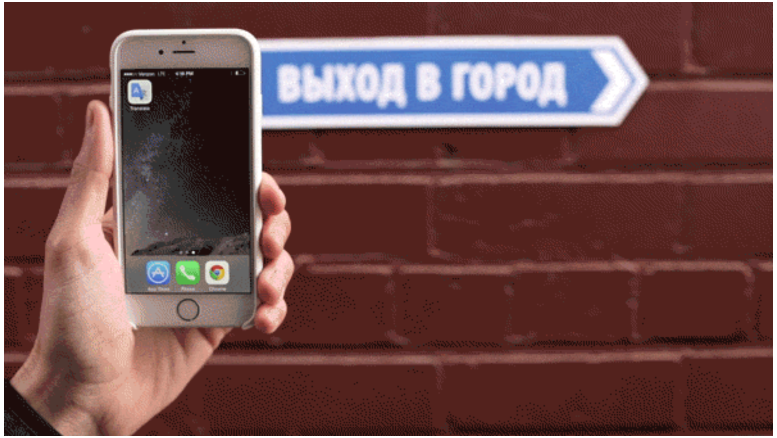
Can we use AI to expand access to dermatologists?

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Aneja S et al. *Arch Dermatol* 2012

AI has changed our world



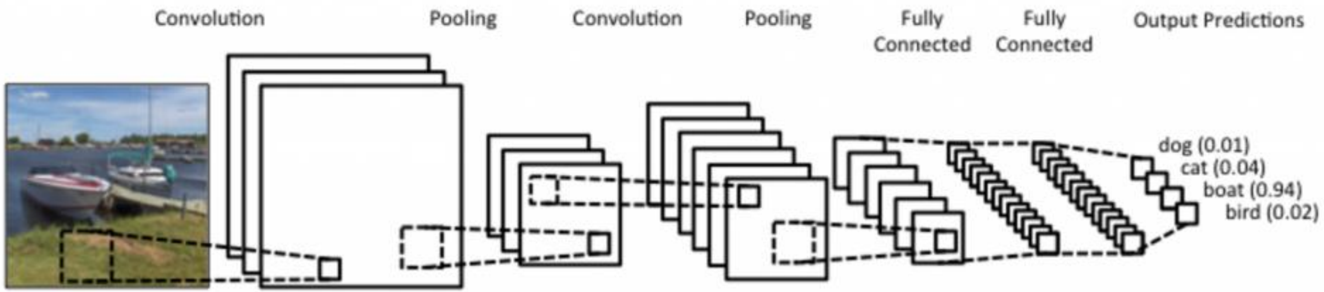
- Driverless cars
- Translation capabilities
- Mortgage lending
- Financial markets

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Convolutional Neural Networks

Multilayer neural networks: involve local connections, shared weights, pooling



- Capture relationship of pixels to each other using filter as a matrix
- Run algorithms over and over until errors are minimized across all images
- Allows for profound level of pattern recognition beyond human brain capability

Stanford ARTIFICIAL INTELLIGENCE

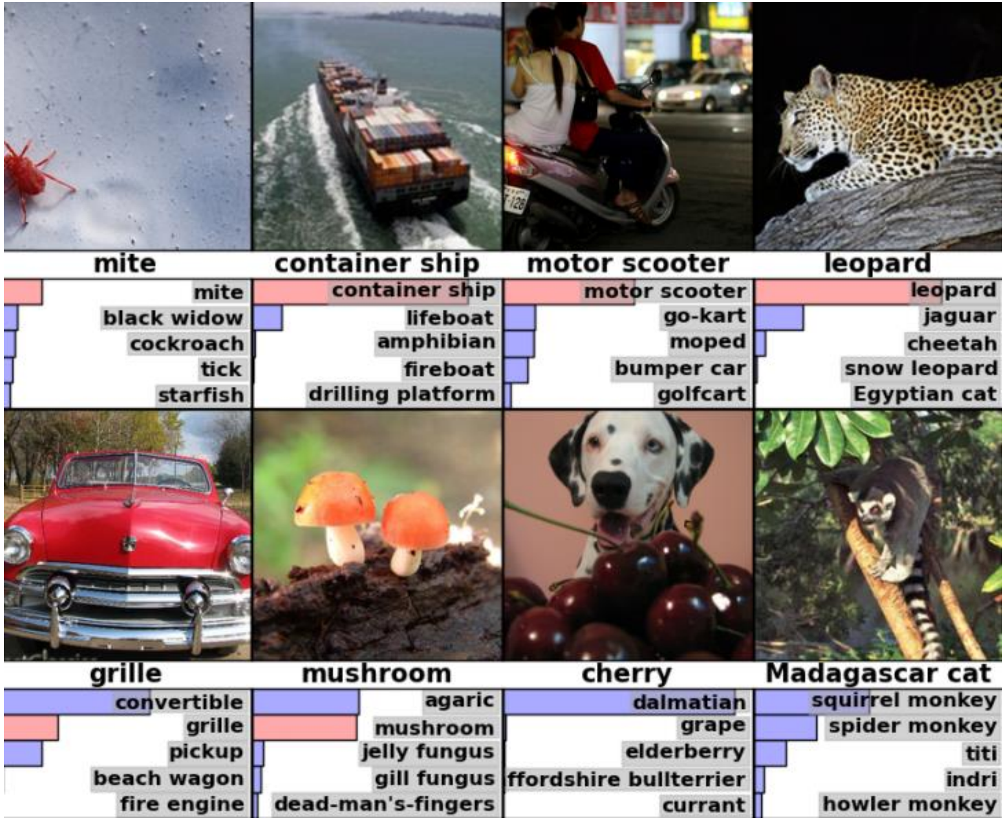
https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

Stanford MEDICINE

Advances in AI/Machine Learning/Computer Vision

- Breakthroughs in artificial intelligence over past 7 years

- Convolutional neural networks + large databases + processing power = **deep learning**



Stanford Artificial Intelligence

Krizhevsky, Sutskever, Hinton et al. ImageNet Classification with Deep Convolutional Neural Networks, NIPS 2012

Algorithms train themselves to recognize what is important and what is not, without human intervention

Fig. 1a: Malamute

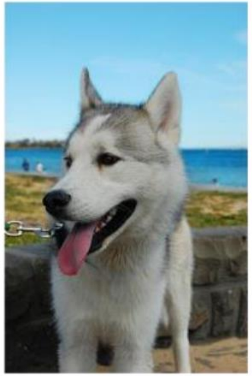


Fig. 1b: Eskimo Dog

Fig. 1c: Husky

- ImageNet project: CNNs trained to classify 1.4 million high-resolution images into 1000 different classes with low error rates

- With sufficient examples of huskies, chihuahuas, basal cell carcinomas or melanomas, algorithms learn the relevant patterns of these categories

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Clinical and Dermoscopic Images Spanning Breadth of Disease

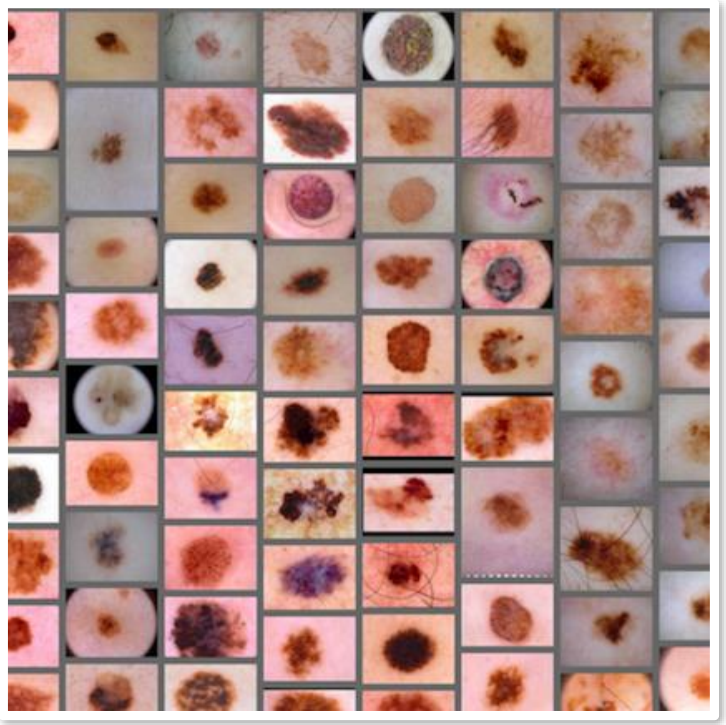
Constructing the dataset:

Nearly 130,000 images:
- clinician-labeled and/or biopsy-proven from 18 different, open-access online repositories

- clinical data from Stanford Dermatology clinics

Biopsy proven images:
- University of Edinburgh library
- ISDIS ISBI challenge (dermoscopic)

Images comprised more than 2,000 dermatologic diseases



Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Visual Taxonomy – Disease Partitioning Algorithm

129K images
2K diseases

Stanford
ARTIFICIAL
INTELLIGENCE

Stanford
MEDICINE

Our Objectives

Evaluate performance of deep learning algorithms (namely, a single CNN) on classification of cutaneous malignancies

Compare to dermatologists

2 main (binary) questions:

- Lesion benign or malignant?
- Biopsy/treat or reassure?

Methods

Recruited 21 board-certified dermatologists

- Stanford University
- University of Pennsylvania
- Massachusetts General Hospital/Harvard
- University of Iowa

Three 100+ question tests

a. Epidermal tumors (BCC/SCC vs SKs)
b. Clinical images of melanocytic lesions (melanoma vs benign nevi, excluded SKs)
c. Dermoscopic images of melanocytic lesions (separate dataset)

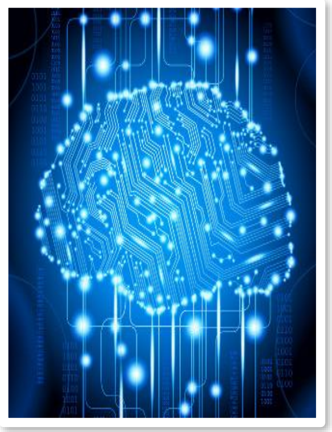
Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Training

Objective: build a system that could accommodate significant variation inherent in photographic images (e.g. lighting, zoom, angle) with no pre-processing or lesion segmentation

Goal: feed any captured skin image directly into the system and output a classification

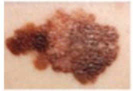


Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Deep Convolutional Neural Network (CNN)

Skin lesion image



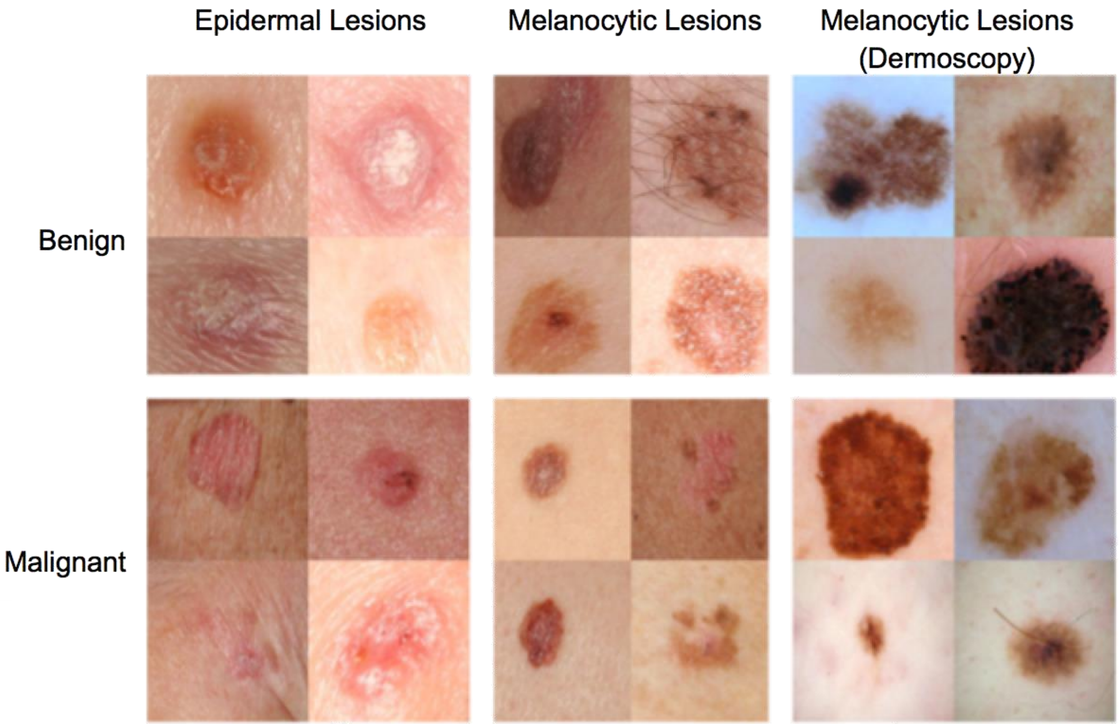
- Used GoogleNet Inception v3 CNN architecture pretrained on ImageNet dataset
- Trained and fine-tuned our CNN using transfer learning
- Resulted in probability distribution over clinical classes of skin
- Applied a partitioning algorithm to our taxonomy to define training classes

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

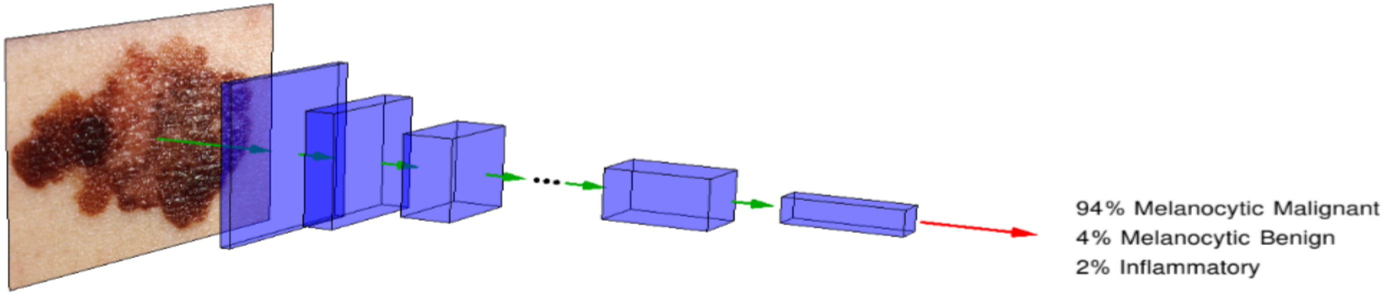
Sample Test Images

Used new dermatologist-labeled dataset of >129K clinical images, including >3300 separate dermoscopic images



Algorithm ready for testing within 1 year

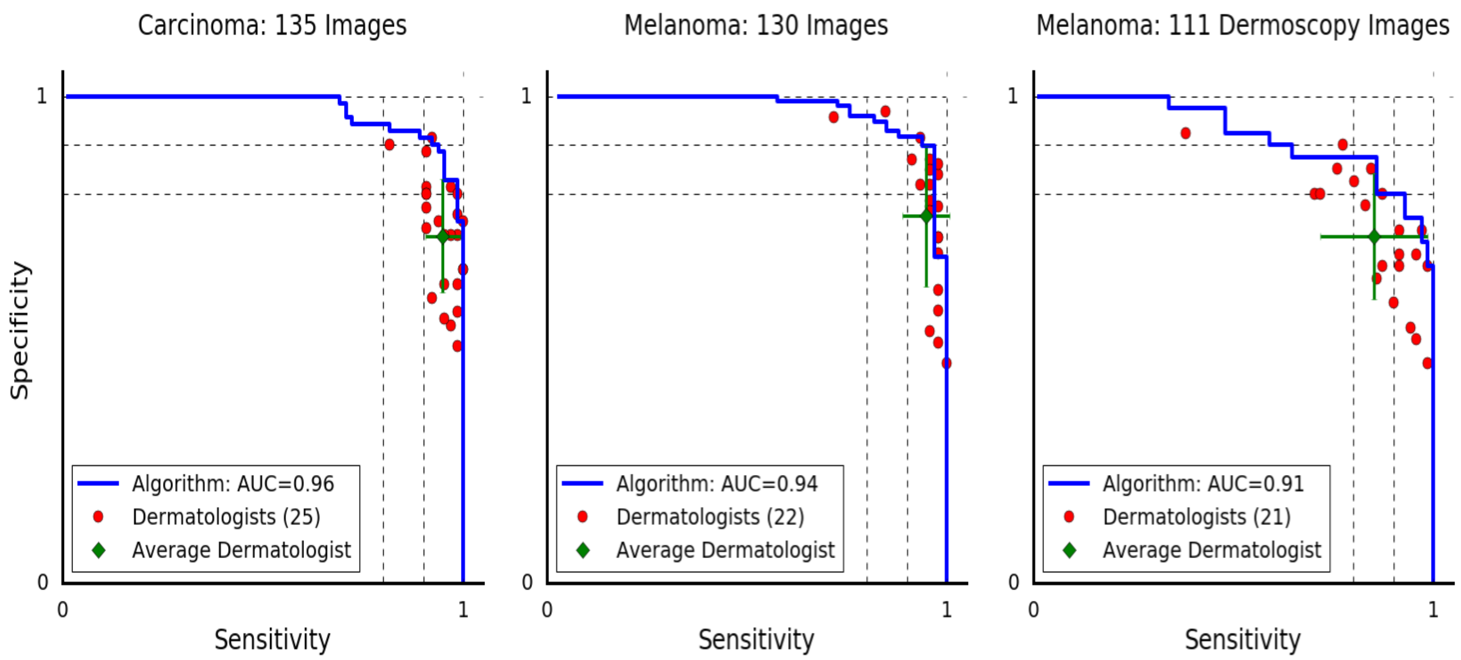
CNN outputs a malignancy probability for each image



Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Results for CNN vs Dermatologists



CNN performed at least as well as dermatologists as a whole

(Limitation of dermoscopy test as most dermatologists were not experts in pigmented lesions or dermoscopy)

Stanford Artificial Intelligence

Esteva A et a. *Nature*. 2017;542:115-118.

Stanford Medicine

LETTER

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1*], Brett Kuprel[1*], Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

There are 5.4 million new cases of skin cancer in the United States[2] every year. One in five Americans will be diagnosed with a cutaneous malignancy in their lifetime. Although melanomas represent fewer than 5% of all skin cancers in the United States, they account for approximately 75% of all skin-cancer-related deaths, and are responsible for over 10,000 deaths annually in the United States alone. Early detection is critical, as the estimated 5-year survival rate for melanoma drops from over 99% if detected in its earliest stages to about 14% if detected in its latest stages. We developed a computational method which may allow medical practitioners and patients to proactively track skin lesions and detect cancer earlier. By creating a novel disease taxonomy, and a disease-partitioning algorithm that maps individual diseases into training classes, we are able to build a deep learning system for automated dermatology.

Previous work in dermatological computer-aided classification[12,14,15] has lacked the generalization capability of medical practitioners owing to insufficient data and a focus on standardized tasks such as dermoscopy[16–18] and histological image classification[19–22]. Dermoscopy images are acquired via a specialized instrument and histological images are acquired via invasive biopsy and microscopy; whereby both modalities yield highly standardized images. Photographic images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and very large datasets[25], have recently been shown to exceed human performance in visual tasks such as playing Atari games[26], strategic board games like Go[27] and object recognition[6]. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification. We restrict the comparisons to image-based classification.

We utilize a GoogleNet Inception v3 CNN architecture[9] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge[6], and train it on our dataset using transfer learning[28]. Figure 1 shows the working system. The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labelled images organized in a tree-structured taxonomy of 2,032 diseases, in which the individual diseases form the leaf nodes. The images come from 18 different clinician-curated, open-access online repositories, as well as from clinical data from Stanford University Medical Center. Figure 2a shows a subset of the full taxonomy, which has been organized clinically and visually by medical experts. We split our dataset into 127,463 training and validation images and 1,942 biopsy-labelled test images.

To take advantage of fine-grained information contained within the taxonomy structure, we develop an algorithm (Extended Data Table 1) to partition diseases into fine-grained training classes (for example, amelanotic melanoma and acrolentiginous melanoma). During inference, the CNN outputs a probability distribution over these fine classes. To recover the probabilities for coarser-level classes of interest (for example, melanoma) we sum the probabilities of their descendants (see Methods and Extended Data Fig. 1 for more details).

We validate the effectiveness of the algorithm in two ways, using nine-fold cross-validation. First, we validate the algorithm using a three-class disease partition—the first-level nodes of the taxonomy, which represent benign lesions, malignant lesions and non-neoplastic

[1]Department of Electrical Engineering, Stanford University, Stanford, California, USA. [2]Department of Dermatology, Stanford University, Stanford, California, USA. [3]Department of Pathology, Stanford University, Stanford, California, USA. [4]Dermatology Service, Veterans Affairs Palo Alto Health Care System, Palo Alto, California, USA. [5]Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA. [6]Department of Computer Science, Stanford University, Stanford, California, USA.
*These authors contributed equally to this work.

Limitations of our study

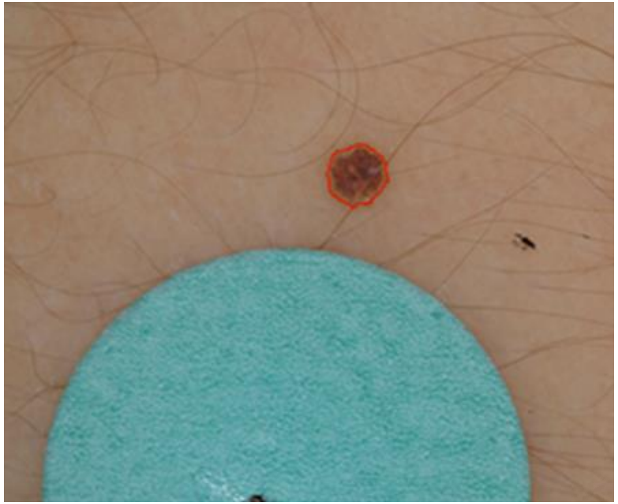
- Retrospective
- Potential spectrum bias
  - Study design→ 2 disease categories
- Differences between in-person exam and telederm ("in silico")
  - We "blink, think, and compare" and use dermoscopy to aid diagnosis
- Transparency on pathology labels/ dermatopathology accuracy/ wide variability for melanocytic neoplasms (*Elmore JG et al. BMJ 2017*)
- Opportunities for bias/confounding
  - Need to ensure extensive representation by varied skin types
  - Extensive additional work is needed
    - Prospective clinical validation studies needed

Stanford Artificial Intelligence

Stanford Medicine

Limitations of Deep Learning

- System is opaque; we don't know why it calls an image benign or malignant
- Dots, rulers, marks, etc. may introduce bias into dataset
- Investigators may not be aware of them or extent
- Impact on MD cognition/learning?



Strengths

Computer can assess image data imperceptible to human eye

- Only looking at specific lesions, not whole-body or sequential change detection

Broadly applicable across disciplines

- Algorithms will improve with transfer learning
- Edges, shapes remain important
- Show a skin cancer app cat pictures and it will improve at classifying skin cancer

Capable of running on a smartphone device

- App created to study this, prospective “real-time” validation in clinical setting

Stanford ARTIFICIAL INTELLIGENCE

What’s happened with AI in Dermatology since then?

- At least 8 publications pertaining to AI (CNNs) for:
  - dermoscopy (with or without clinical images), dermatopathology (nod BCC, dermal nevi, SK), melanocytic lesions, melanoma, acral melanoma dermoscopy, nonmelanoma skin cancer (BCC/SCC), onychomycosis
- CNN consistently performs as well if not better than “expert” clinicians
- Limitations: datasets lack full spectrum of skin phenotypes and lesions (particularly less common banal lesions/disorders)
- No real-world prospective use/validation as yet

Stanford ARTIFICIAL INTELLIGENCE

Gilmore SF. *PLoS One* 2018; Han SS. *PLoS One* 2018; Marchetti MA. *J Am Acad Dermatol* 2018; Yu C. *PLoS One* 2018; Haenssle HA. *Ann Oncol* 2018; Brinker TJ. *Eur J Cancer* 2019; Fujisawa Y. *Brit J Dermatol* 2019; Marka A. *BMC Med Imaging* 2019

Stanford MEDICINE

Can AI be used to quantify and monitor skin disease severity?

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Context matters

Dermatologist’s clinical impression based on factors beyond visual and dermoscopic examination of a lesion in isolation

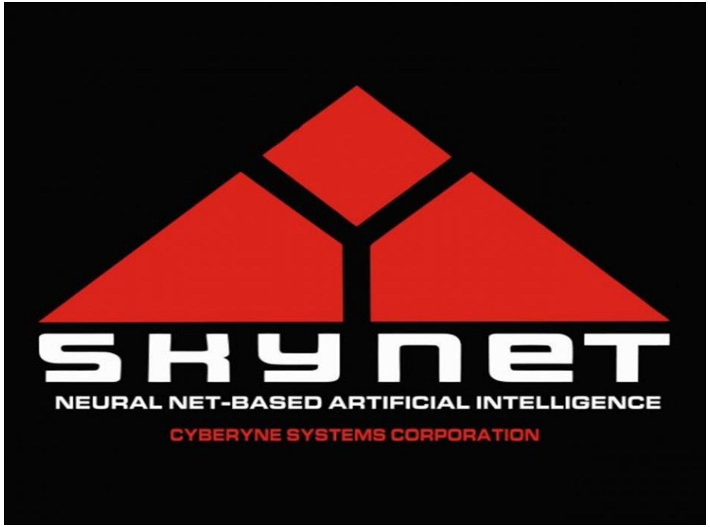
- Sequential imaging, change detection, incorporation of clinical metadata?

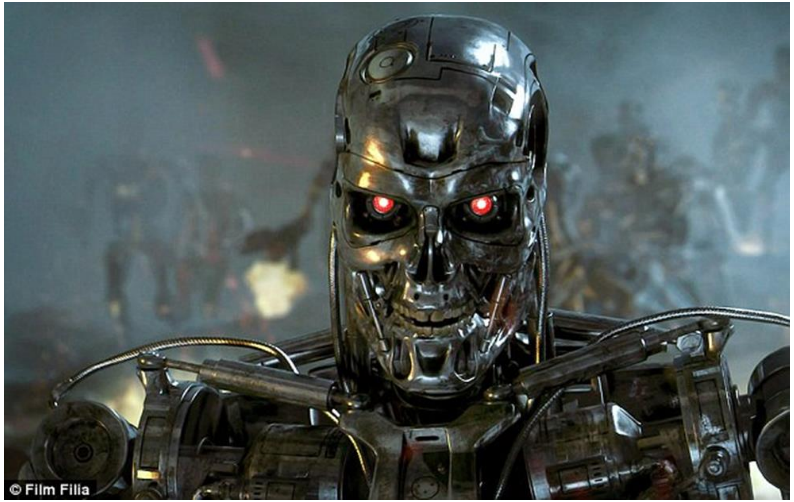
Are we comfortable with AI as black box?

A trained neural net does not necessarily mimic the decision-making approach of humans

Identifies its own criteria for informative patterns associated with a disease

Is AI a Pandora's box?





- **Probably not, and it won't replace dermatologists!**
- Human decisions made based on machine output of probability and other factors!
- Real vs imaginary dangers:
- "AI is a fundamental existential risk for human civilization." *(Elon Musk, 2017)*
- "AI software will help us understand biology, understand how to intervene and improve lives very dramatically." *(Bill Gates, 2018)*
- **Roles in health care and education**
- **Regulation and liability need to be addressed proactively**

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Real-world, clinical validation of app in progress

"Smart dermoscope"

"Augmented clinician"

Stanford MEDICINE

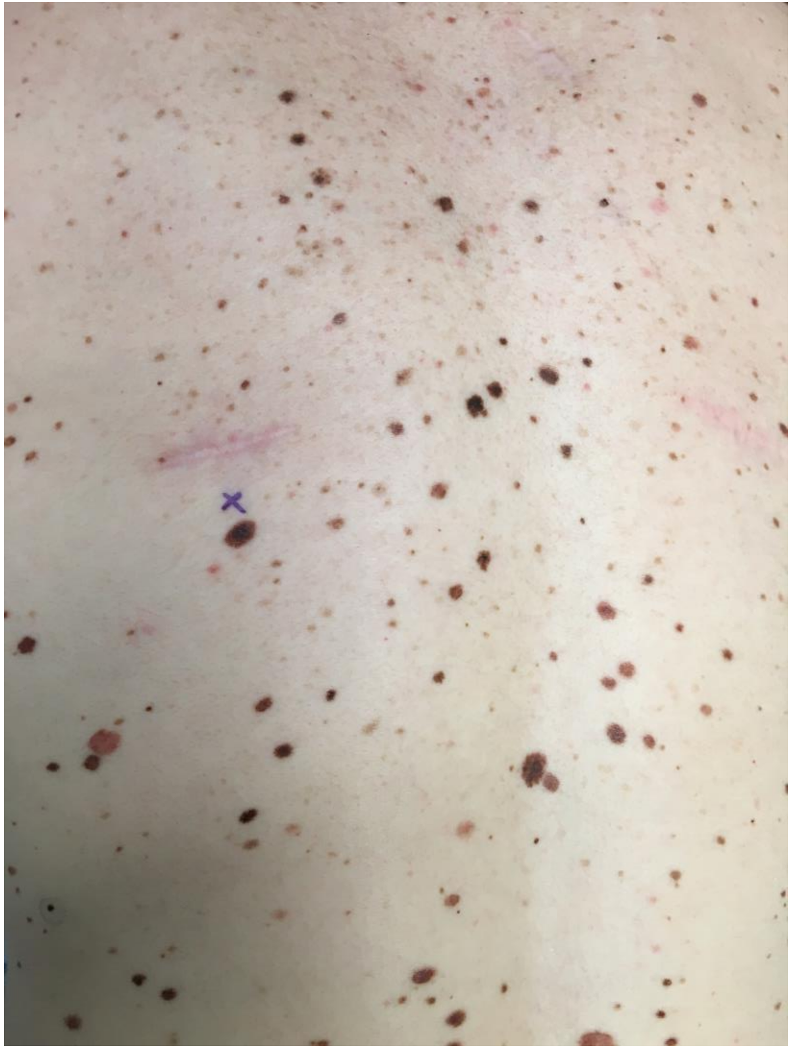
Can AI help dermatologists?

-35 Y female with atypical mole syndrome, seven cutaneous melanomas (3 melanomas in situ, 3 T1a invasive melanomas and 1 T2a SLNB-positive melanoma) and ten severely dysplastic nevi diagnosed since 2016

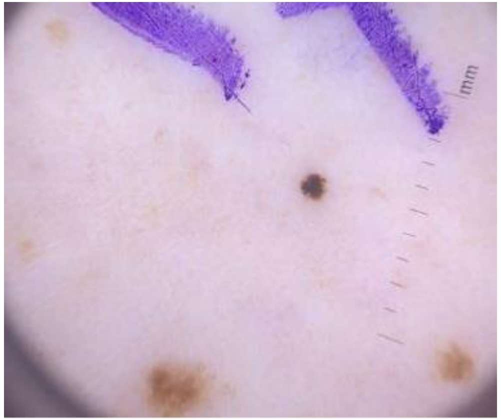
-negative for p16/CDKN2A mutation

-followed q 3 mos with total body photography and digital dermoscopy

Stanford Artificial Intelligence



A tool for surveillance? – NOT YET

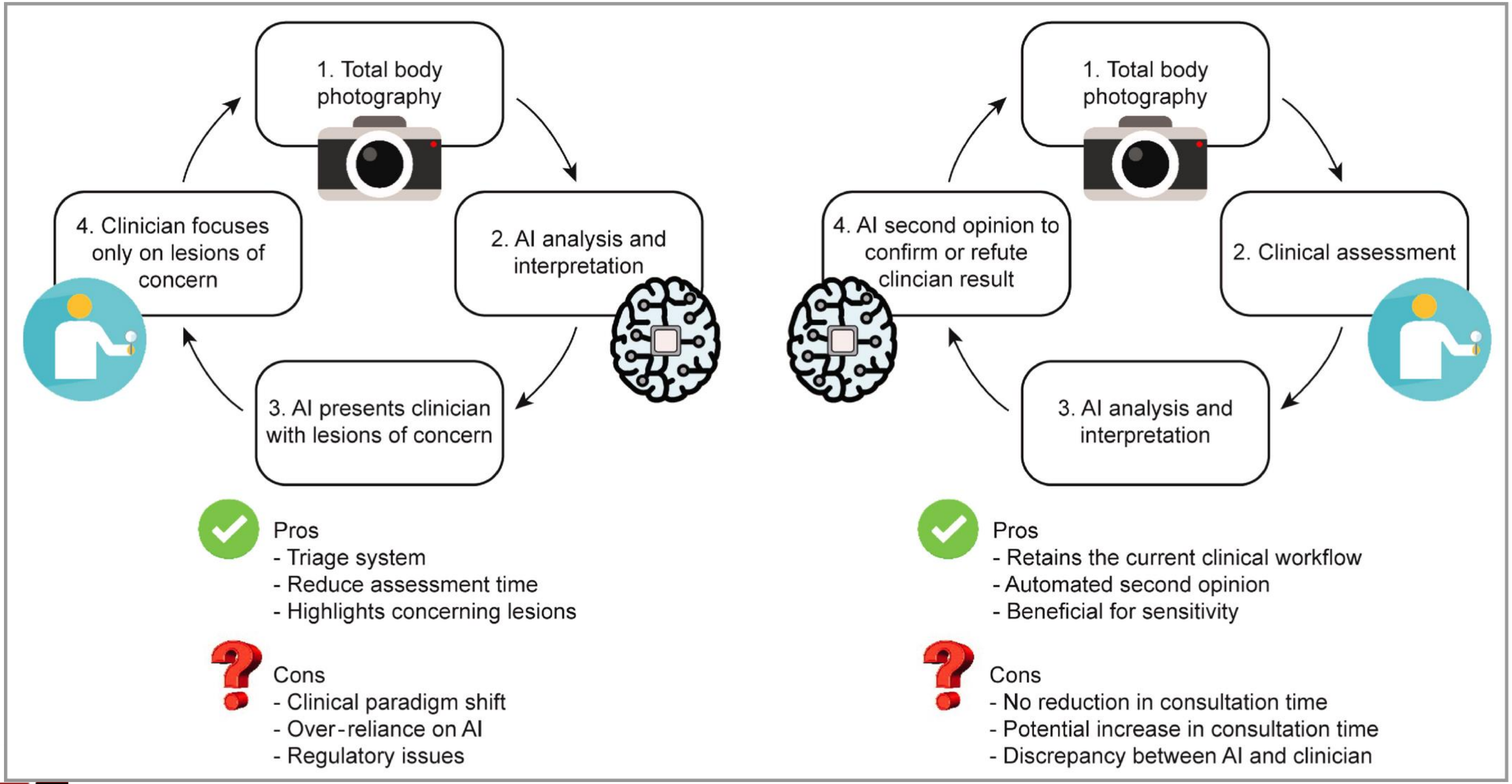


New moderately dysplastic nevi with focal severe atypia

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

Can AI enhance clinical decision-making?

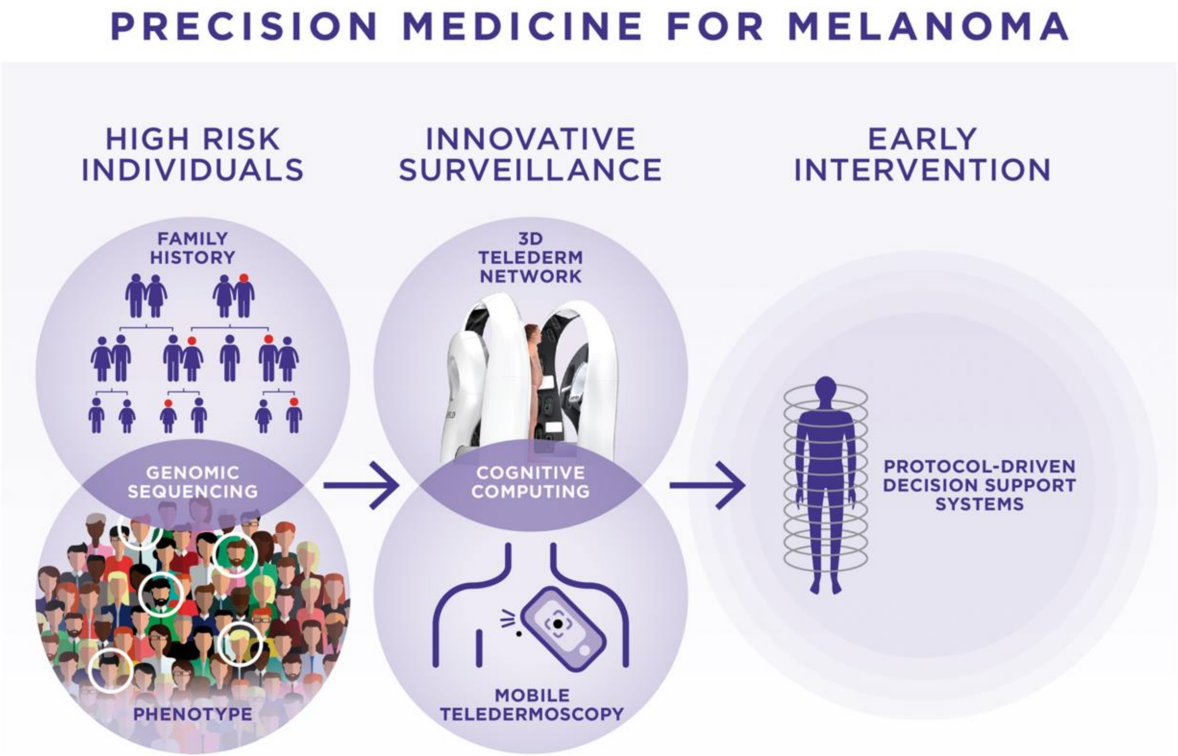


Stanford ARTIFICIAL INTELLIGENCE

Janda M, Soyer HP. *Brit J Dermatol*. 2019:180: 247-248.

Stanford MEDICINE

Future of melanoma early detection



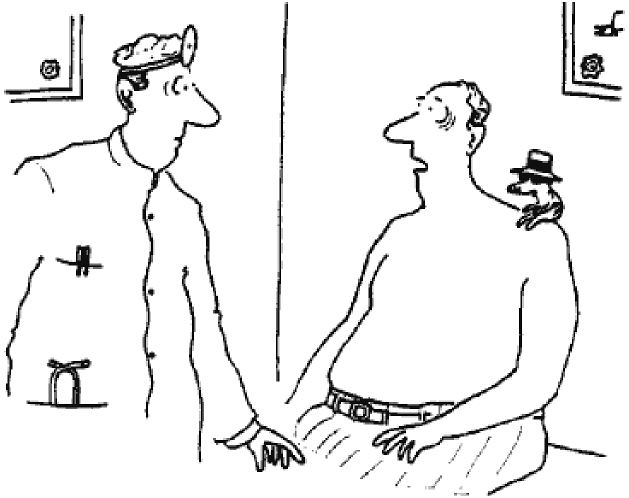
Targeted surveillance of high risk individuals with high resolution 3D imaging and integration of AI

Mar VJ, Soyer HP. *Ann Oncol*. 2018;29:1625-1628. (Schematic from Smithers BM, Dunn J and Soyer HP. Whither melanoma in Australia? Med J Aust 2017;207:330–331.)

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE

AI can augment, but not replace decision-making and human interactions in medicine



"Doctor, I have a suspicious looking mole on my shoulder."

Stanford ARTIFICIAL INTELLIGENCE

Stanford MEDICINE