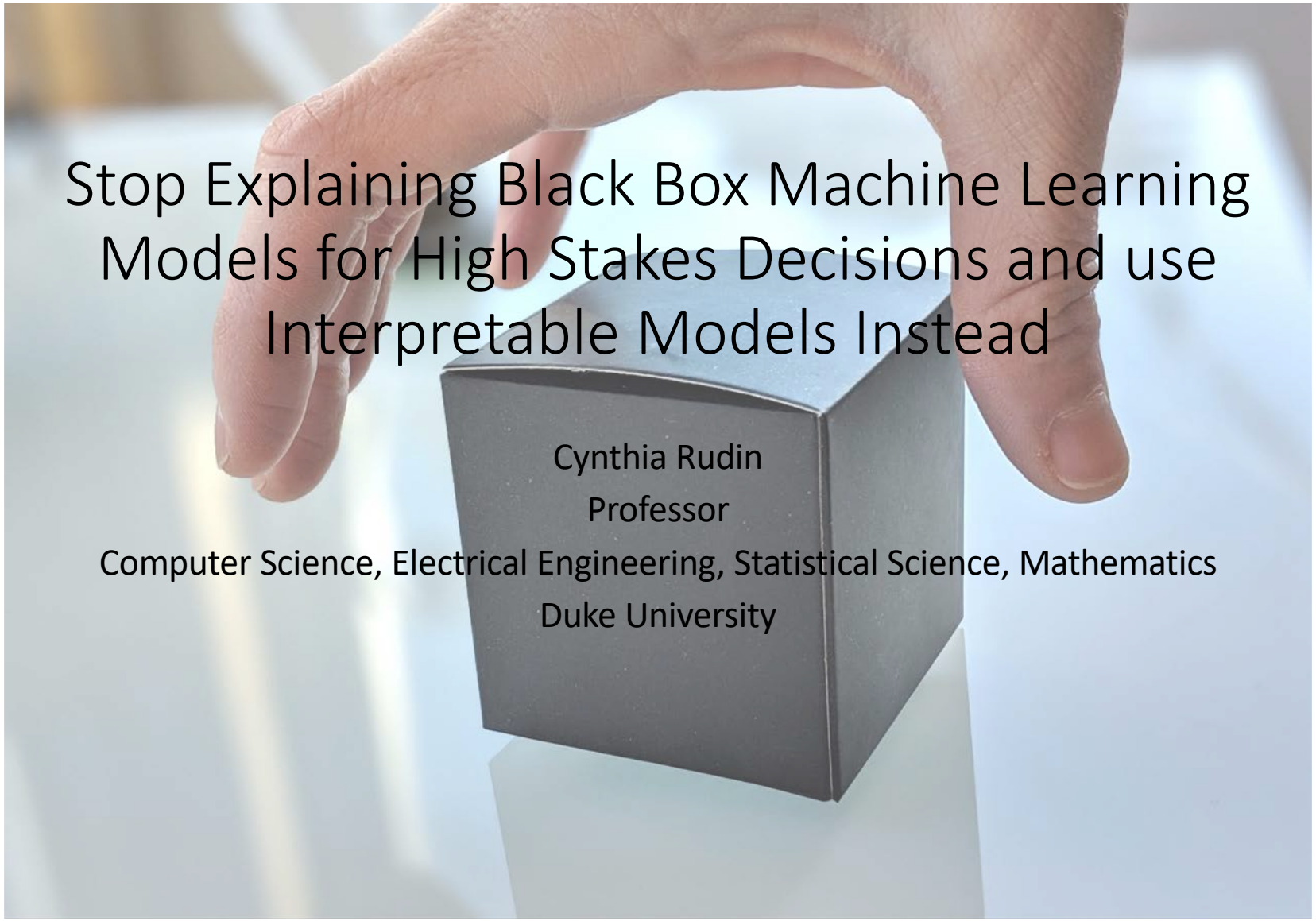


Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead

Cynthia Rudin
Professor
Computer Science, Electrical Engineering, Statistical Science, Mathematics
Duke University

A black box predictive model is a formula that is either too complicated to understand or proprietary.

What happens when you use a black box?

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

Glenn Rodriguez was denied parole because of a miscalculated “COMPAS” score.

A typographical error in a COMPAS score can lead to years of extra prison time.

Glenn Rodriguez

How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say

BY MICHAEL MCGOUGH
AUGUST 07, 2018 09:26 AM, UPDATED AUGUST 07, 2018 09:26 AM

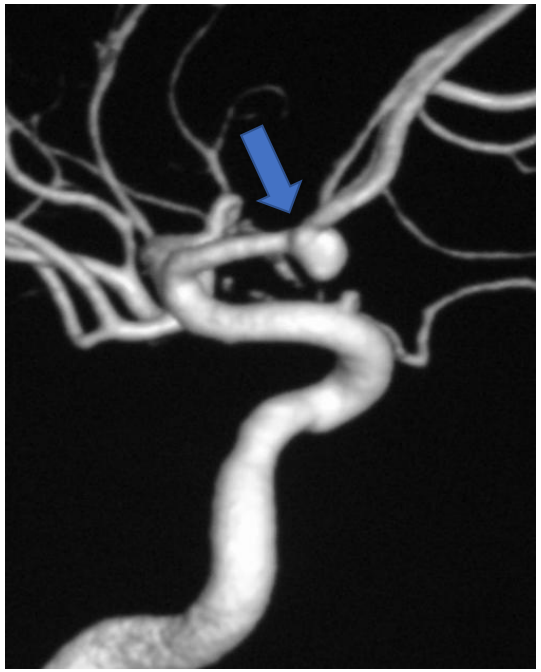
Smoke is affecting air quality all over California. Here's what it looks like at the Carr Fire, north of Redding, on July 31, 2018.
BY PAUL KITAGAKI JR.

Where did Breezometer go wrong?
We'll never know…

And this is the tip of the iceberg…

- An interpretable machine learning model obeys a domain-specific set of constraints that makes its computations easier to understand.
- My technical definition: An interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge.
- There's a spectrum.

Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

- Seizure are common (20%)
- Seizure→ Brain Damage
- Need EEG to detect seizures

Need to use EEG data to predict seizures to determine EEG duration

EEG is expensive and limited: 24hrs of monitoring is $1600-$4000

- 2HELPS2B was not created by doctors
- It is a ML model
- It is just as accurate as black box models.
- Doctors can decide themselves whether to trust it
- Doctors can calibrate the score with information not in the database

2HELPS2B

1.	Any cEEG Pattern with Frequency **2 H**z	1 point	$\cdots$
2.	**E**pileptiform Discharges	1 point	$+\ \cdots$
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	$+\ \cdots$
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	$+\ \cdots$
5.	Prior **S**eizure	1 point	$+\ \cdots$
6.	**B**rief Rhythmic Discharges	**2** points	$+\ \cdots$
		SCORE	$=\ \cdots$

SCORE	0	1	2	3	4	5	6+
RISK	$<$5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

There are many variables to choose from.

Variable
PDR
BRDs
Unreactive background
Prior Sz
GRDA
LRDA
GPDs
LPDs
BIPDs
Infection
Inflammation
Neoplasm
ICH
Metabolic encephalopathy
Stroke
SAH
SDH
TBI
Hypoxic/ischemic
IVH
Hydrocephalus
Discharges
Frequency (>2Hz)[c]

Risk-Calibrated Supersparse Linear Integer Models (Risk-SLIM)

(Ustun, R, 2019)

Logistic Loss

Model Size

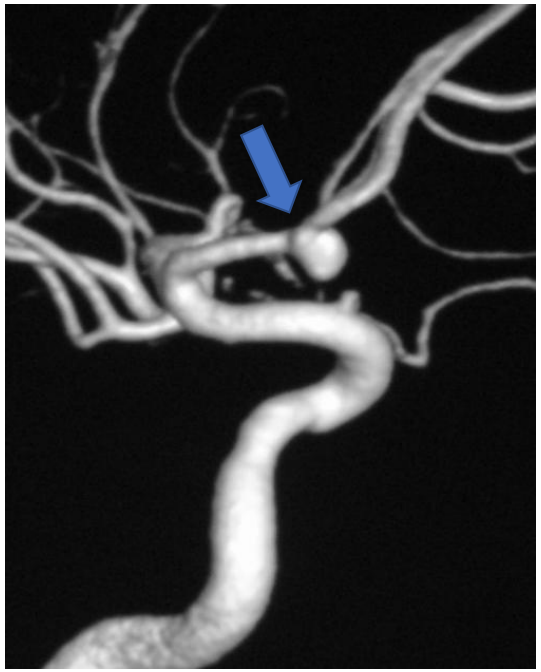
$$\min_{\lambda \in L} \sum_{i=1}^{n} \log\left(1 + e^{-y_i x_i^{\intercal} \lambda}\right) + C\|\lambda\|_0$$

$$\lambda \in L \text{ means that } \forall j,\ \lambda_j \in \{-10,-9,\ldots,0,\ldots,9,10\}$$

Small Integer Coefficients

MINLP – really hard…

Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

1-Hour Screening EEG

2HELPS2B=3 (high-risk)

- Placed on Continuous EEG for >72H
- Start on preventative medications

So far…

- 2HELPS2B validated on independent multicenter cohort (N=2111)

- Implemented: University of Wisconsin, Massachusetts General Hospital/Harvard Medical School
- Ongoing implementation: Emory University, Duke University, Medical University of South Carolina, Free University of Brussels (Belgium)

- Resulted in **63.6%** reduction in duration of EEG monitoring per patient
  - $1,134.831 saving per patient[1]
- **2.82 X** More Patients Monitored
- **$6.1M** estimated savings in FY 2018 at MGH,UW

[1]2016 Medicare Reimbursement Most Common Professional Code

- So that’s how interpretable models are supposed to work…
  but don’t they lose accuracy?

The New York Times

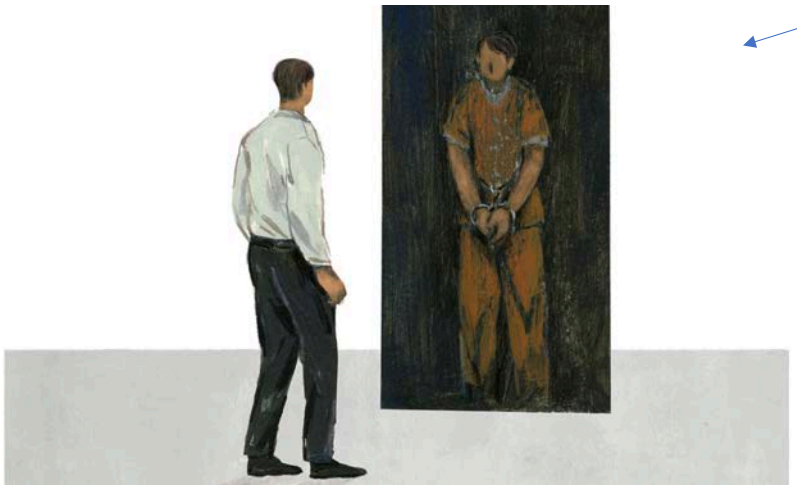
OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

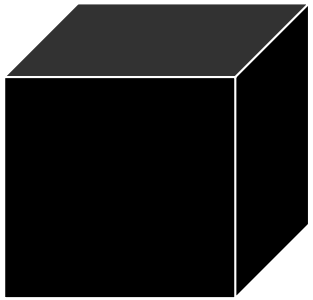


Glenn Rodriguez was denied parole because of a miscalculated "COMPAS" score.

How accurate is COMPAS?
Data from Florida can tell us…

COMPAS vs. CORELS

COMPAS: (Correctional Offender Management Profiling for Alternative Sanctions)

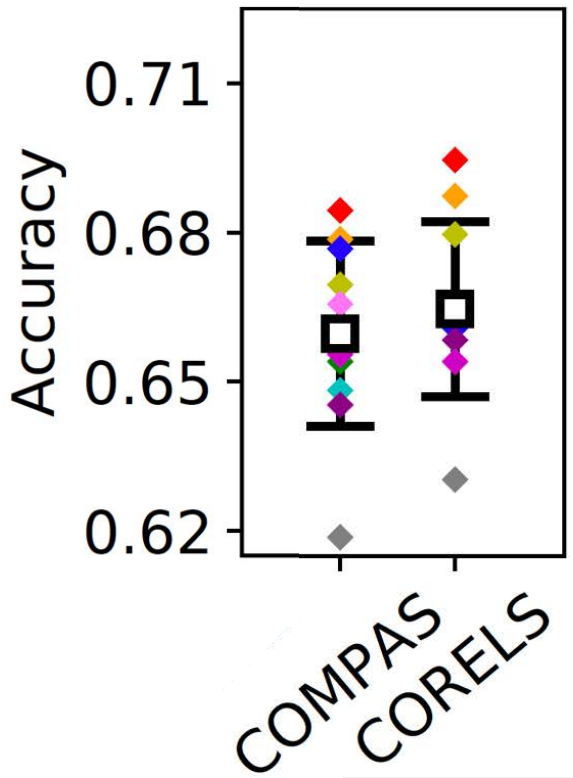


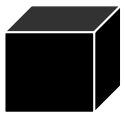
CORELS: (Certifiably Optimal RulE ListS, with Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, and Margo Seltzer, KDD 2017 & JMLR 2018)

Here is the machine learning model:

If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

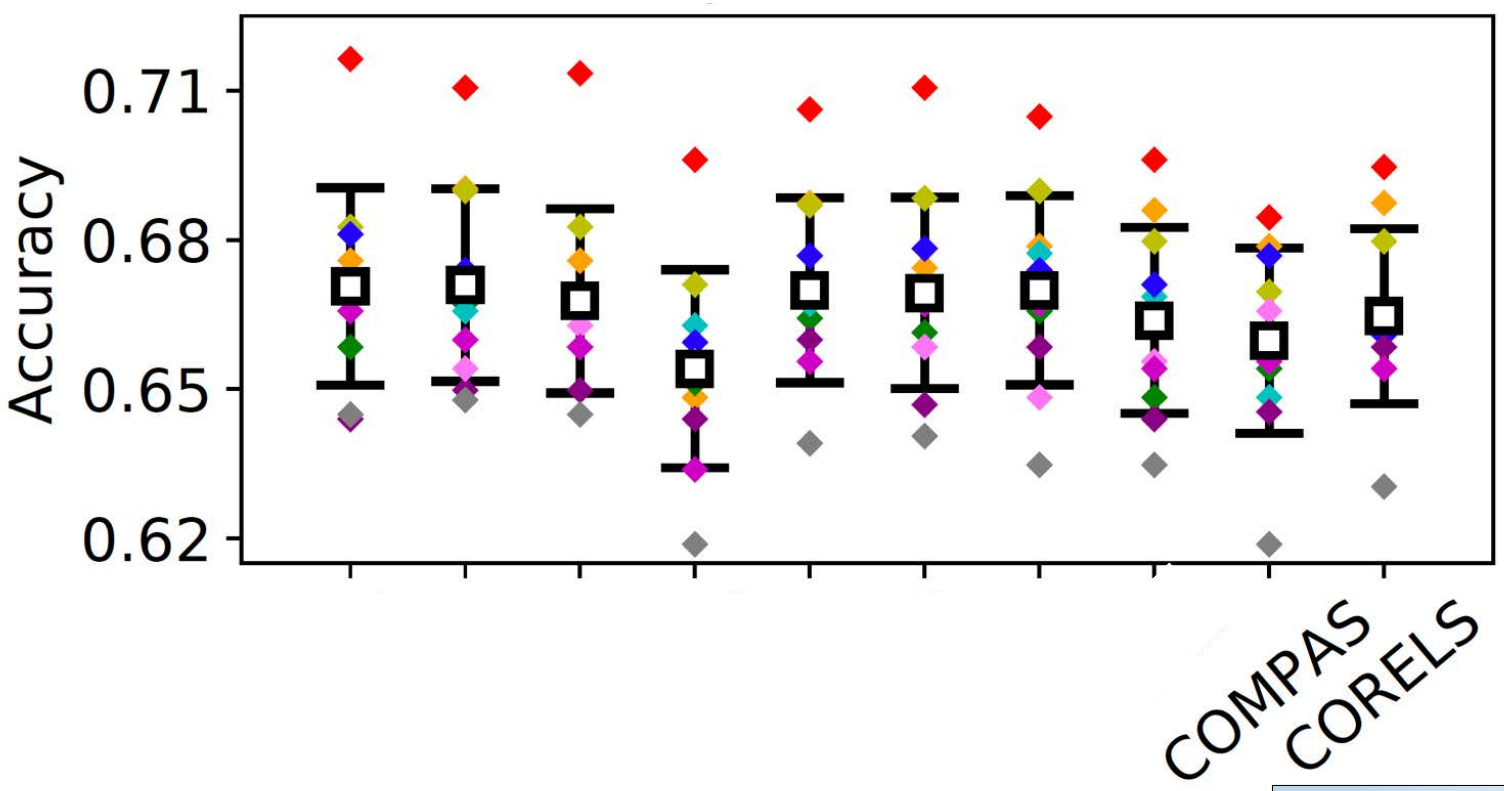
Prediction of re-arrest within 2 years

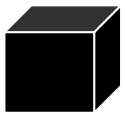




If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

Prediction of re-arrest within 2 years





If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

- Interpretable ML – When you use a model that is not black box.

- Explainable ML – When you use a black box and explain it afterwards (posthoc)
  - Start with a black box.
    - Create another model that approximates it.
    - Compute derivatives of it.
    - Visualize what part of the input the model is paying attention to.
    - :

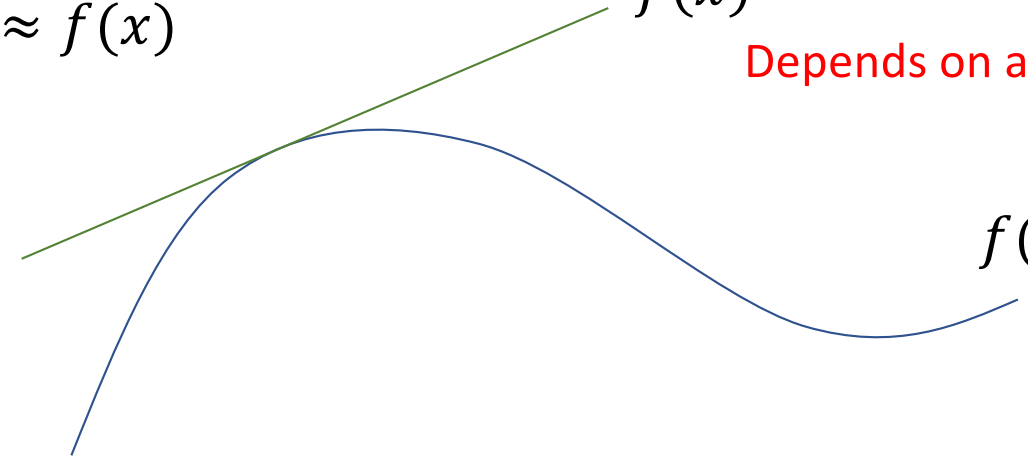
Interpretable Models ≠ Explanations of Black Box Models

- Double Trouble: Forces you to rely on two models instead of one. Those models necessarily disagree with each other
  - An explanation that is right 90% of the time is wrong 10% of the time.
- Typos are a problem when inputting data into black box models.
- If you can produce an accurate interpretable model, why explain a black box? (e.g., COMPAS vs CORELS)

Interpretable Models $\neq$ Explanations of Black Box Models

- "Explanations" are not actually explanations of what the model is doing. **Approximations are not explanations**! Gets variable importance wrong.

$f(x) \approx \tilde{f}(x)$



$\tilde{f}(x)$ An approximation

Depends on age, priors and race

$f(x)$ COMPAS

Depends on age, number of prior crimes

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

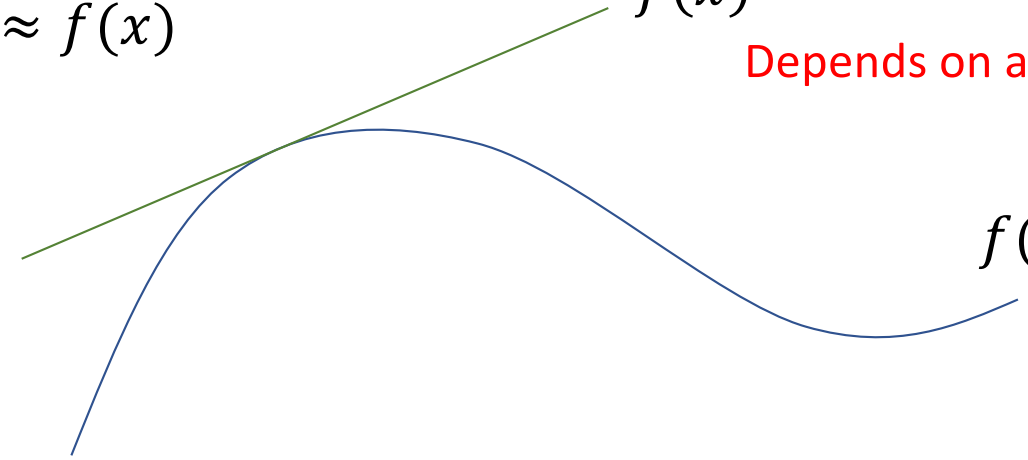
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Interpretable Models $\neq$ Explanations of Black Box Models

- "Explanations" are not actually explanations of what the model is doing. **Approximations are not explanations**! Gets variable importance wrong.

$f(x) \approx \tilde{f}(x)$



$\tilde{f}(x)$ An approximation

Depends on age, priors and race

$f(x)$ COMPAS

Depends on age, number of prior crimes

What ProPublica Did

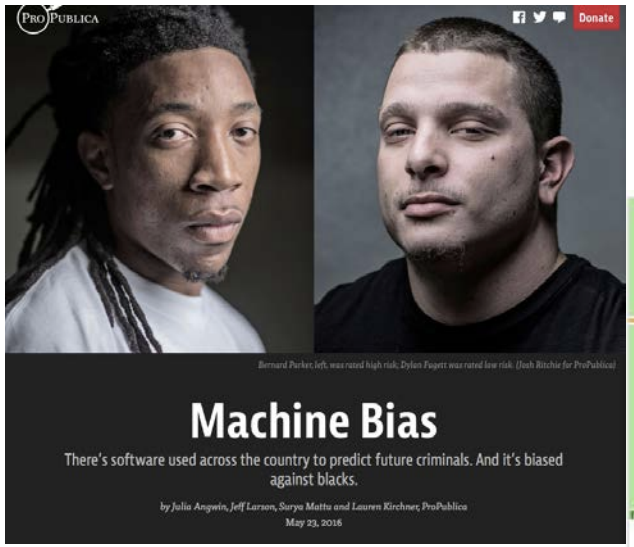
- They showed that FPR and FNR varied by race.

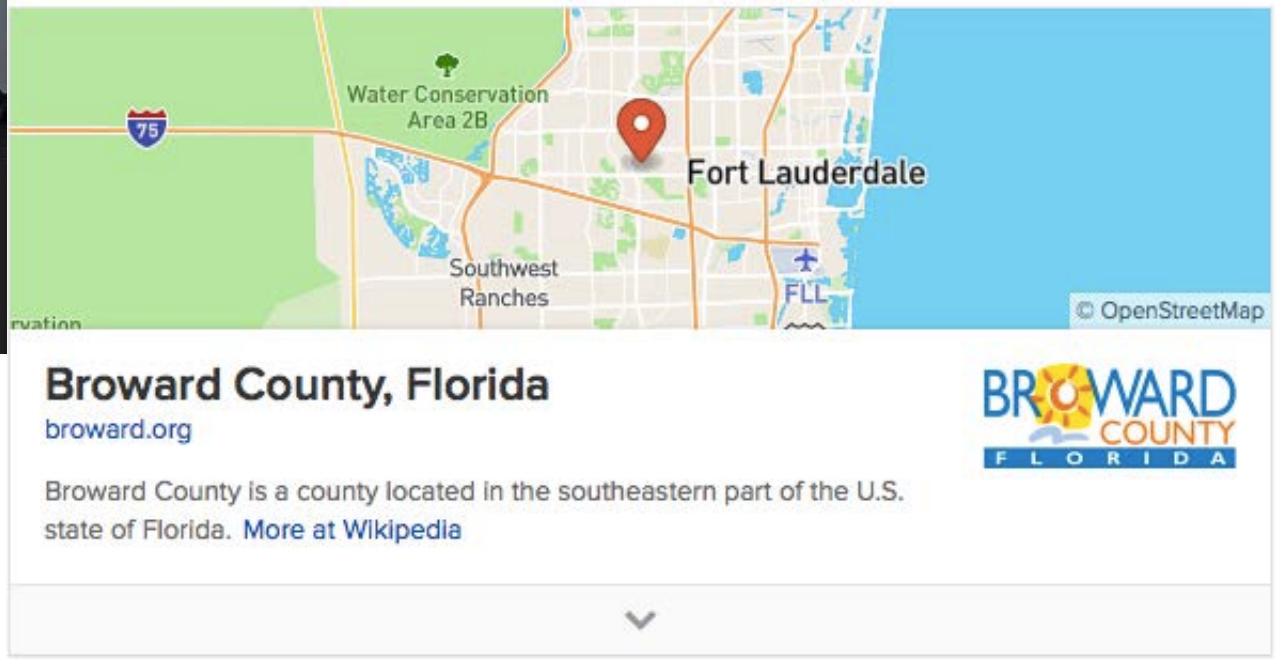
- They suggested maybe this might not be a good comparison, we should condition on age and number of priors and reexamine.

- After conditioning on age and number of priors, still found a linear approximation to COMPAS with a low pvalue for the race covariate.
  - We don't think COMPAS is linear
- Concluded that COMPAS depends on race.
  - Bad idea

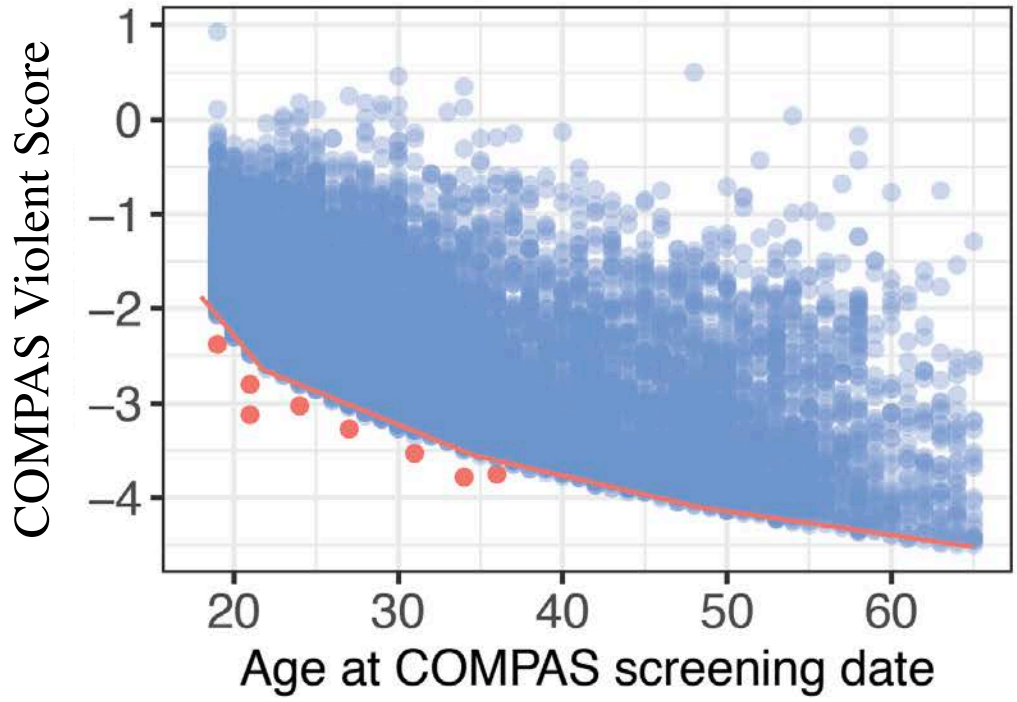
Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2020

A peek inside COMPAS?





A peek inside COMPAS?



Scatter plot of COMPAS violent scores vs age for all individuals in Broward County FL.

Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2020

A peek inside COMPAS?

Does COMPAS $- f_{\text{age}}$ depend on race?

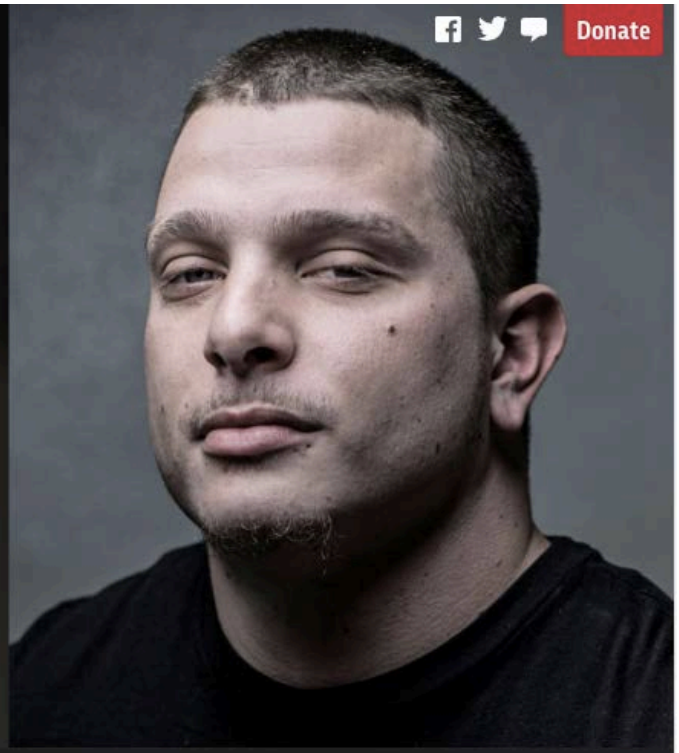
It doesn't seem to.

(We ran machine learning methods *with and without race* to see if they need race to predict COMPAS well. They performed similarly.)

Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2020

PRO PUBLICA

Donate



arker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

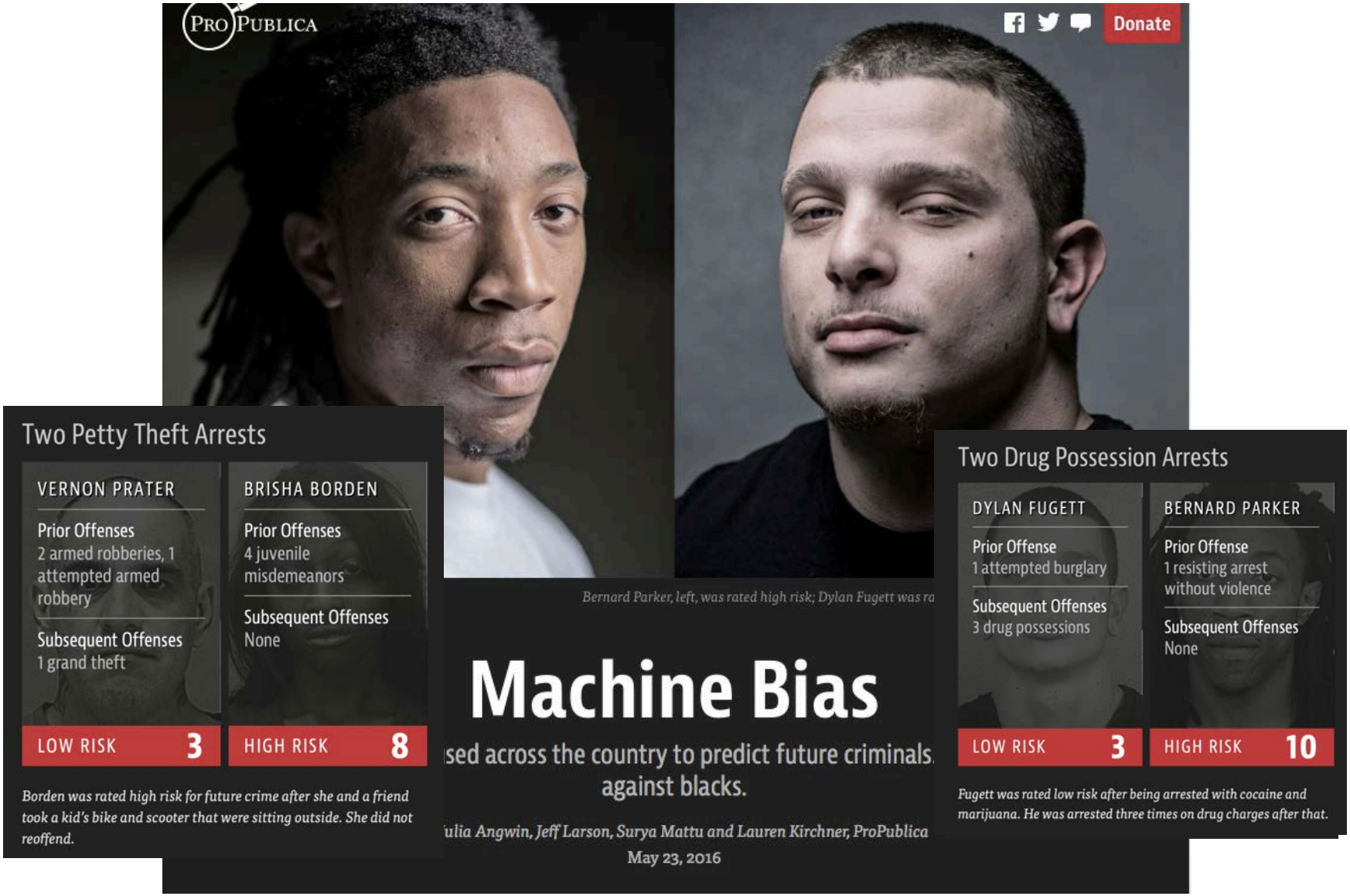
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

ine Bias

try to predict future criminals. And it's biased inst blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016


PRO PUBLICA
Donate
Two Petty Theft Arrests
VERNON PRATER
Prior Offenses
2 armed robberies, 1 attempted armed robbery
Subsequent Offenses
1 grand theft
LOW RISK 3
BRISHA BORDEN
Prior Offenses
4 juvenile misdemeanors
Subsequent Offenses
None
HIGH RISK 8
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.
Bernard Parker, left, was rated high risk; Dylan Fugett was ra
Machine Bias
sed across the country to predict future criminals.
against blacks.
ulia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
Two Drug Possession Arrests
DYLAN FUGETT
Prior Offense
1 attempted burglary
Subsequent Offenses
3 drug possessions
LOW RISK 3
BERNARD PARKER
Prior Offense
1 resisting arrest without violence
Subsequent Offenses
None
HIGH RISK 10
Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

137 factors entered by hand for each survey

1% error rate → 75% chance of at least one typo on a survey

This is a serious disadvantage to complicated or proprietary models.

In Florida….?

Name	COMPAS Violent Decile	# Arrests	# Charges	Selected Prior Charges	Selected Subseq. Charges
Shirley Darby	1	2	4	Aggravated Battery (F,1), Child Abuse (F,1), Resist Officer w/Violence (F,1)	
Joseph Salera	1	8	14	Battery on Law Enforc Officer (F,3), Aggravated Assault W/Dead Weap (F,1), Aggravated Battery (F,1), Resist/obstruct Officer W/viol (F,1)	
Bart Sandell	1	9	15	Attempted Murder 1st Degree (F,1), Resist/obstruct Officer W/viol (F,1), Agg Battery Grt/Bod/Harm (F,1), Carrying Concealed Firearm (F,1)	Armed Sex Batt/vict 12 Yrs + (F,2), Aggravated Assault W/dead Weap (F,3), Kidnapping (F,1)
Miguel Wilkins	1	11	22	Aggrav Battery w/Deadly Weapon (F,1), Driving Under The Influence (M,2), Carrying Concealed Firearm (F,1)	
Jonathan Gabbard	1	7	28	Robbery / Deadly Weapon (F,11), Poss Firearm Commission Felony (F,7)	
Brandon Jackel	1	22	40	Resist/obstruct Officer W/viol (F,3), Battery on Law Enforc Officer (F,2), Attempted Robbery Deadly Weapon (F,1), Robbery 1 / Deadly Weapon (F,1)	
Fernando Galarza	2	2	6	Murder in the First Degree (F,1), Aggrav Battery w/Deadly Weapon (F,1), Carrying Concealed Firearm (F,1)	

Continued on next page

Name	COMPAS Violent Decile	# Arrests	# Charges	Selected Prior Charges	Selected Subseq. Charges
Nathan Keller	2	8	17	Aggravated Assault (F,5), Aggravated Assault W/dead Weap (F,2), Shoot/throw Into Vehicle (F,2), Battery Upon Detainee (F,1)	
Zachary Campanelli	2	11	21	Armed Trafficking In Cocaine (F,1), Poss Weapon Commission Felony (F,1), Carrying Concealed Firearm (F,1)	
Aaron Coleburn	2	16	25	Attempt Murder in the First Degree (F,1), Carrying Concealed Firearm (F,1), Felon in Pos of Firearm or Amm (F,1)	
Bruce Poblano	2	22	39	Aggravated Battery (F,3), Robbery / Deadly Weapon (F,3), Kidnapping (F,1), Carrying Concealed Firearm (F,2)	Grand Theft in the 3rd Degree (F,3)
Phillip Sperry	3	11	16	Aggravated Assault W/dead Weap (F,1), Burglary Damage Property>$1000 (F,1), Burglary Unoccupied Dwelling (F,1)	
Dylan Azzi	3	11	17	Aggravated Assault W/dead Weap (F,2), Aggravated Assault w/Firearm (F,2), Discharge Firearm From Vehicle (F,1), Home Invasion Robbery (F,1)	Fail Register Vehicle (M,2)
Russell Michaels	3	9	23	Solicit to Commit Armed Robbery (F,1), Armed False Imprisonment (F,1), Home Invasion Robbery (F,1)	Driving While License Revoked (F,3)
Bradley Haddock	3	15	25	Attempt Sexual Batt / Vict 12+ (F,1), Resist/obstruct Officer W/viol (F,1), Poss Firearm W/alter/remov Id# (F,1)	
Randy Walkman	3	24	36	Murder in the First Degree (F,1), Poss Firearm Commission Felony (F,1), Solicit to Commit Armed Robbery (F,1)	Petit Theft $100-300$ (M,1)
Carol Hartman	4	5	16	Aggrav Battery w/Deadly Weapon (F,1), Felon in Pos of Firearm or Amm (F,4)	Resist/Obstruct W/O Violence (M,1), Possess Drug Paraphernalia (M,1)

Possibly typos in the COMPAS documentation from Northpointe?

COMPAS Documentation

> Violent Recidivism Risk Score
> $= (\text{age} * -\text{w}) + (\text{age-at-first-arrest} * -\text{w}) + (\text{history of violence} * \text{w})$
> $+ \quad (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w})$

Corrected version?

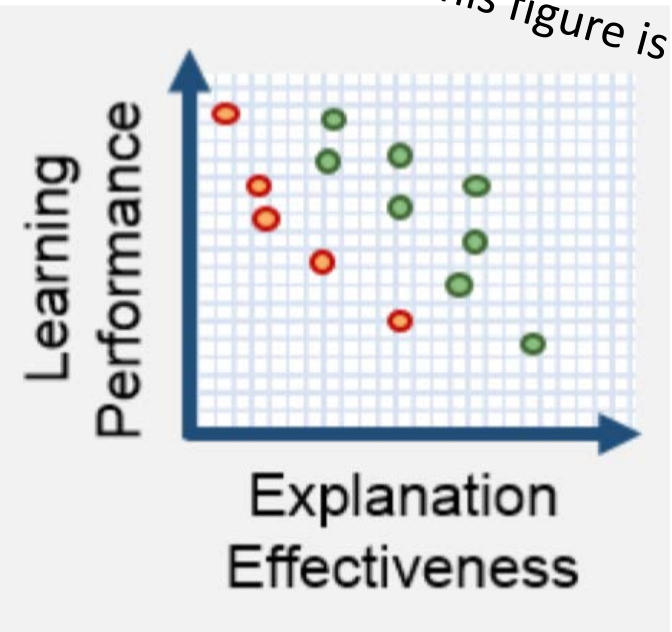
> Violent Recidivism Risk Score
> $= (f(\text{age}) * -\text{w}) + (g(\text{age-at-first-arrest}) * -\text{w}) + (\text{history of violence} * \text{w})$
> $+ \quad (\text{vocation education} * \text{w}) + (\text{history of noncompliance} * \text{w}),$

where f and g are proprietary transformations of age, such as linear splines?

Back to Interpretable vs Explainable…

This figure is phony baloney

The tradeoff doesn't happen like this



Static dataset?
Fixed evaluation metric?

Are they talking about explaining black boxes?

From the DARPA XAI BAA, 2016

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

There has been an increasing trend in healthcare and criminal justice to leverage machine learning (ML) for high-stakes prediction applications that deeply impact human lives. Many of not. There is a spectrum between fully transparent models (where we understand how all the variables are jointly related to each other) and models that are lightly constrained in model form (such as models

- Typos (e.g., Glenn Rodriguez's COMPAS calculation)
- Black box models *still* force you to trust the dataset.
- Double trouble: Forces you to rely on two models instead of one.

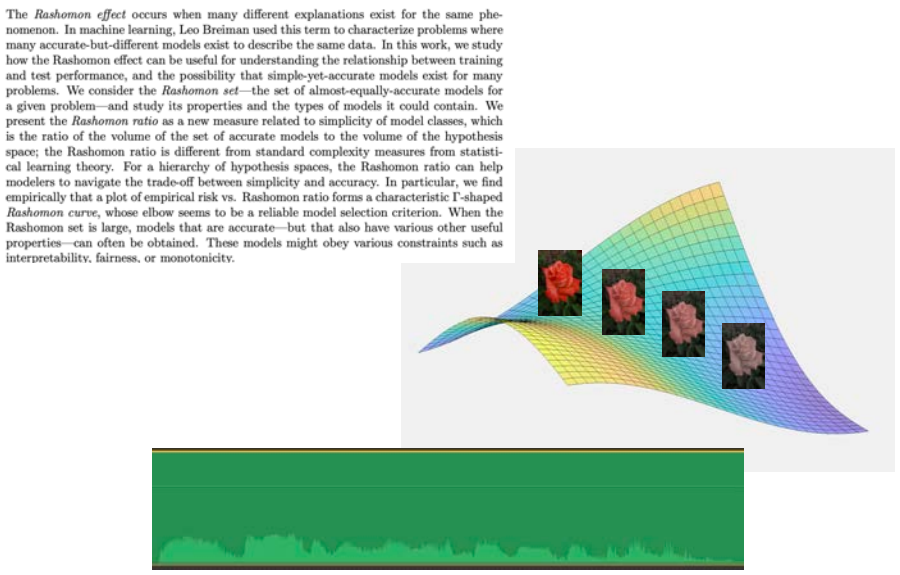
  Those models necessarily disagree with each other
  - An explanation that is right 90% of the time is wrong 10% of the time.
- The explanations are not really explanations, they don't use the same variables.

(Propublica scandal: They said COMPAS depends on age, criminal history, and *race*. But their analysis is wrong - it possibly *only* depends on race through age and criminal history.)

- If you can produce an interpretable model, why explain black boxes? Do you really want to extend the authority of the black box?

Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes



Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

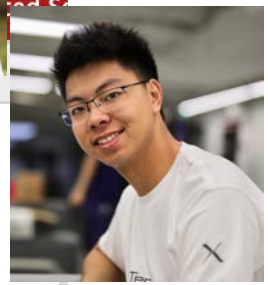
Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

- …But don’t they lose accuracy?

  The answer seems to be no.
  But you must be able to choose a good definition of interpretability.
  In most cases, interpretability helps accuracy.

An interpretable deep neural network?



Zhi

Webster

arXiv.org > cs > arXiv:2002.01650

Computer Science > Machine Learning

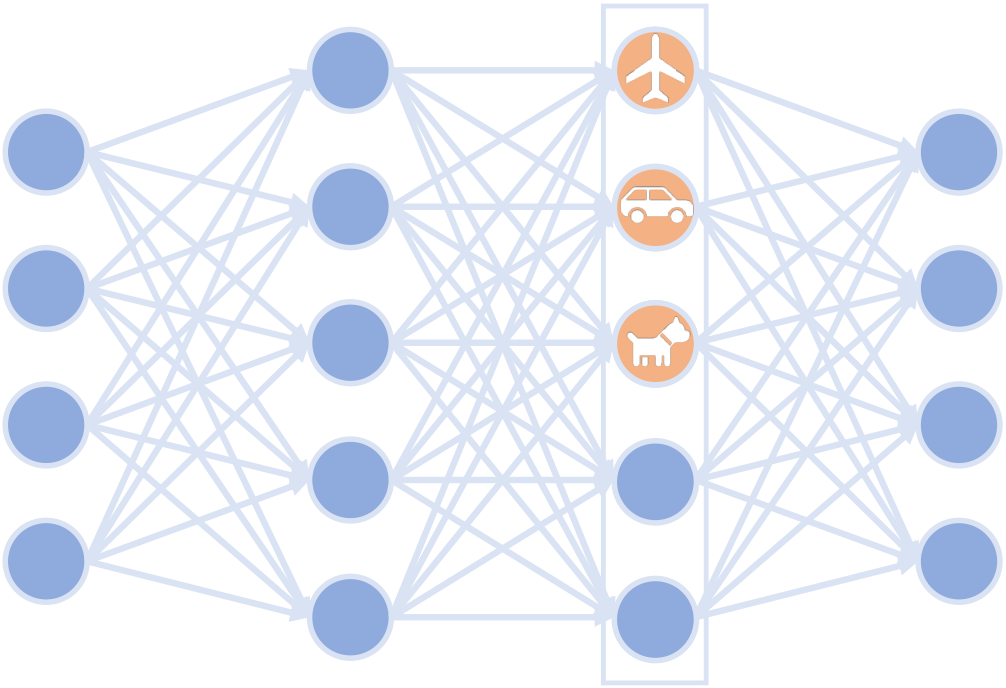
[Submitted on 5 Feb 2020 (v1), last revised 19 Oct 2020 (this version, v4)]

Concept Whitening for Interpretable Image Recognition

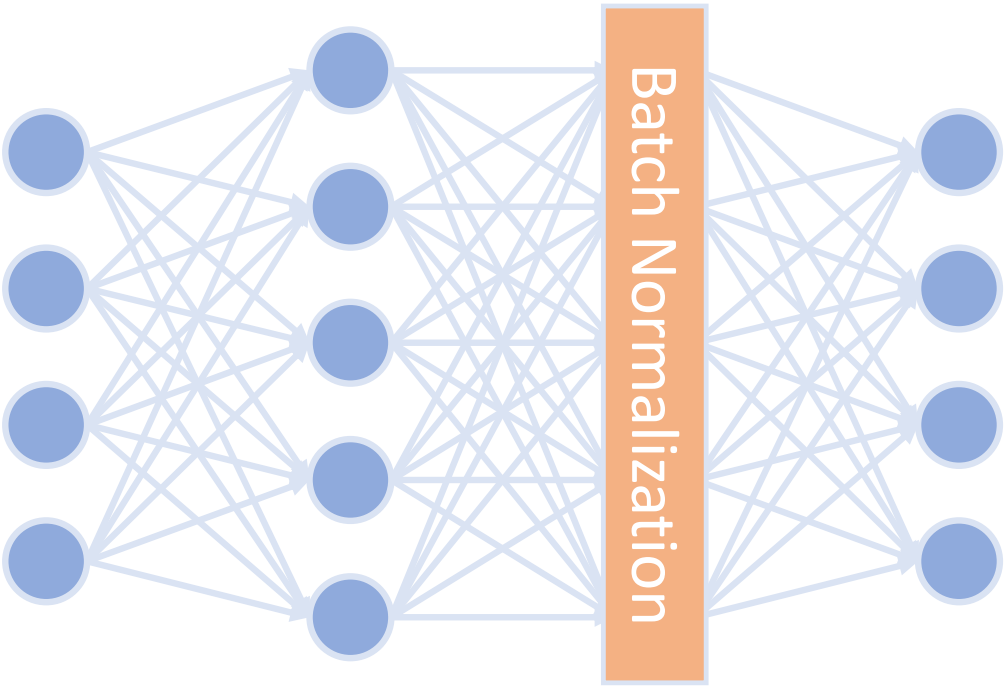
Zhi Chen, Yijie Bei, Cynthia Rudin

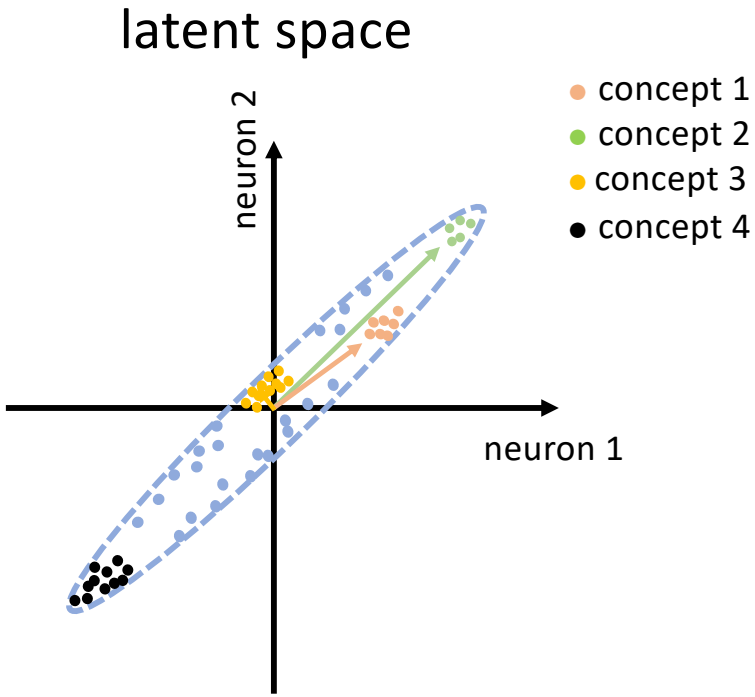
What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning is undoubtedly important, but the calculations of neural networks are very challenging to understand. Attempts to see inside their hidden layers can either be misleading, unusable, or rely on the latent space to possess properties that it may not have. In this work, rather than attempting to analyze a neural network posthoc, we introduce a mechanism, called concept whitening (CW), to alter a given layer of the network to allow us to better understand the computation leading up to that layer. When a concept whitening module is added to a CNN, the axes of the latent space are aligned with known concepts of interest. By experiment, we show that CW can provide us a much clearer understanding for how the network gradually

Nature Machine Intelligence, 2020



CNN’s are not naturally disentangled

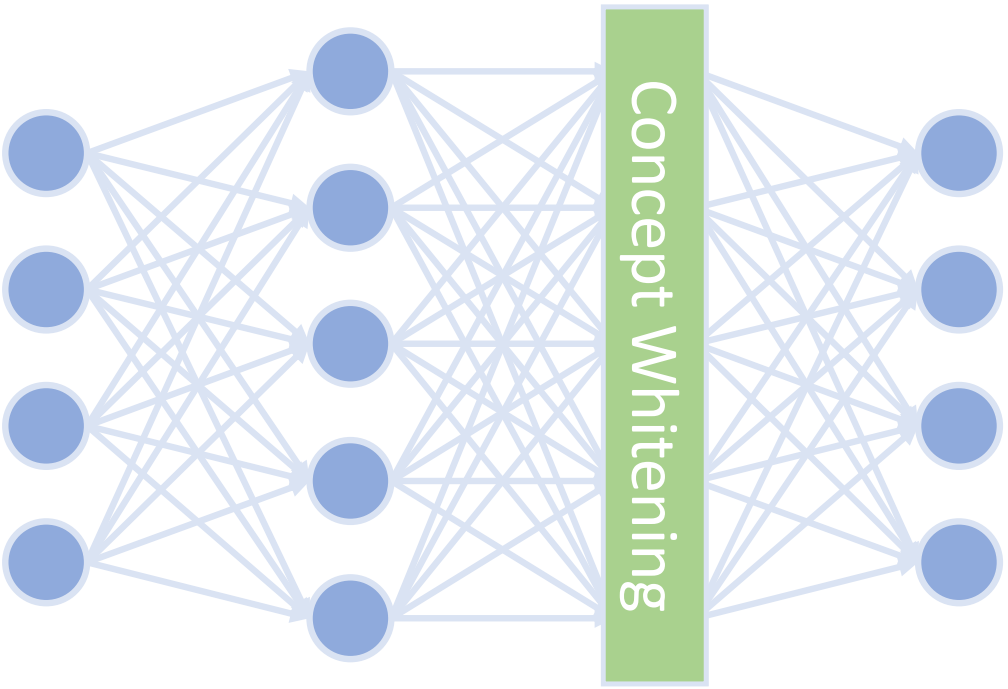


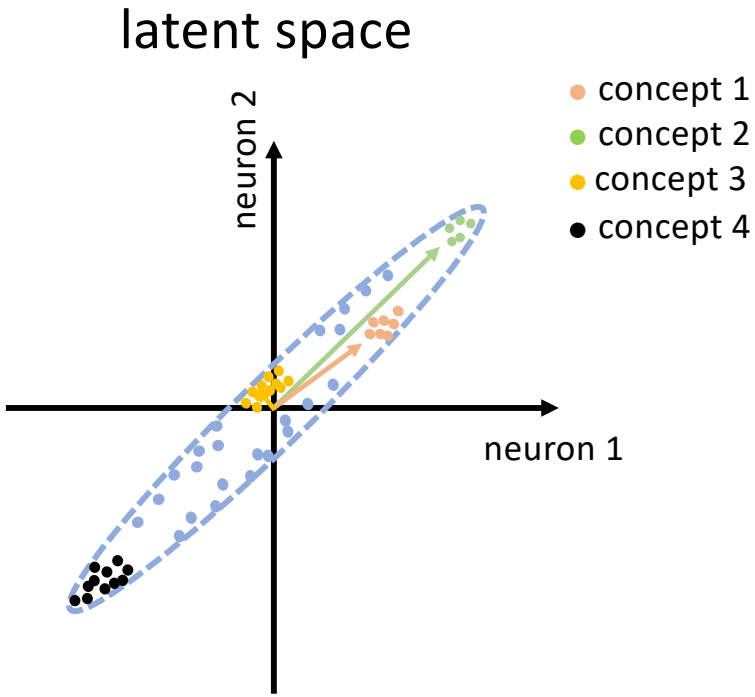


CNN's are not naturally disentangled

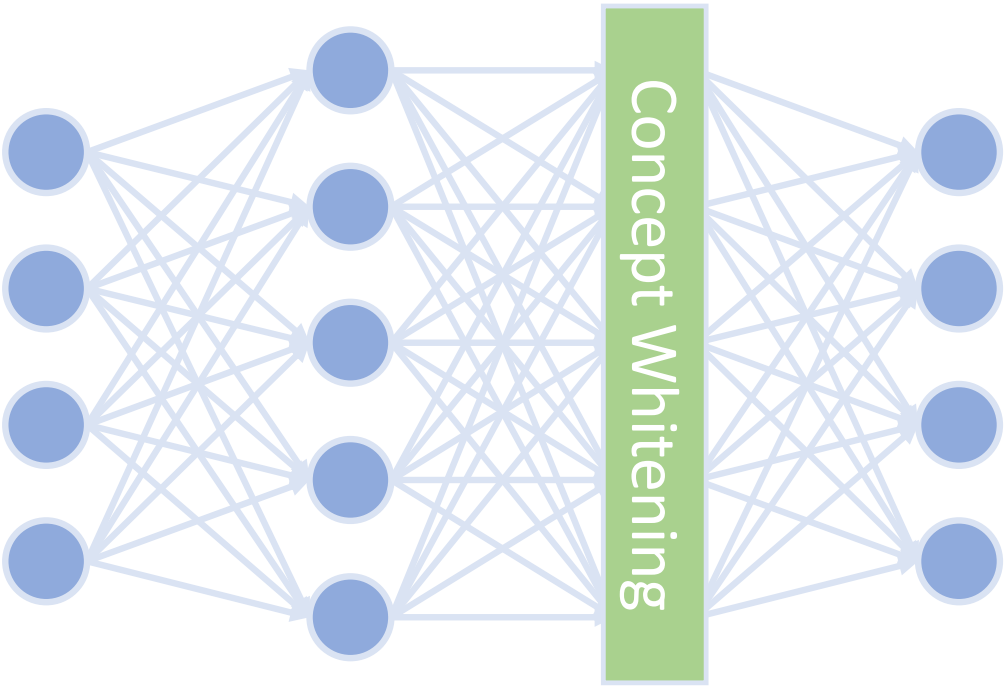
Consider the latent space of the Batch Norm layer

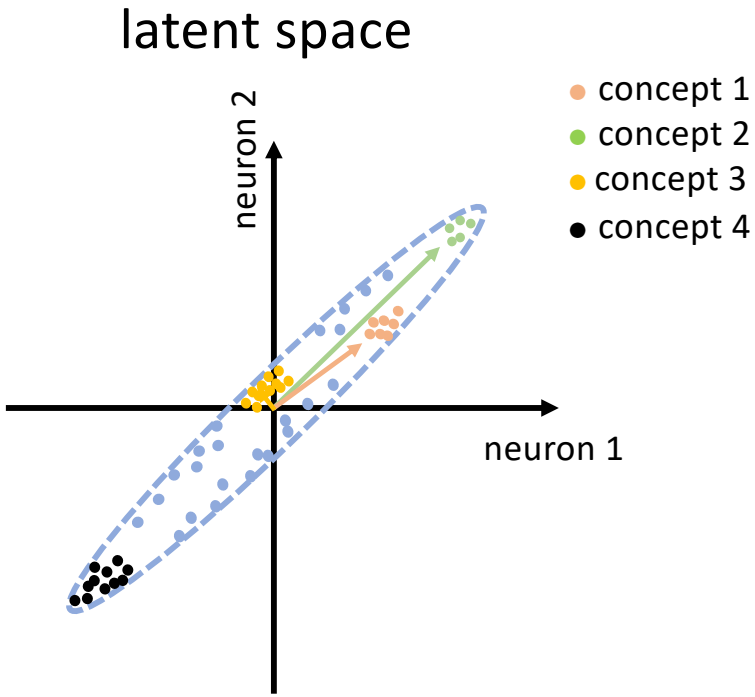
Create a vector pointing towards each concept. They are not naturally orthonormal.





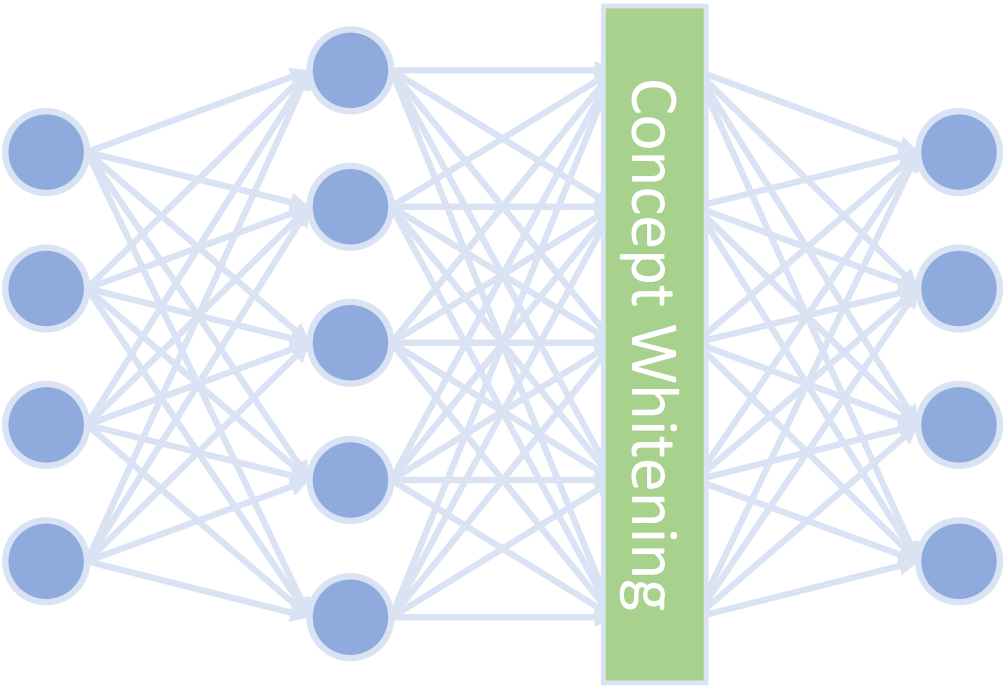
Concept Whitening (CW) disentangles this space.

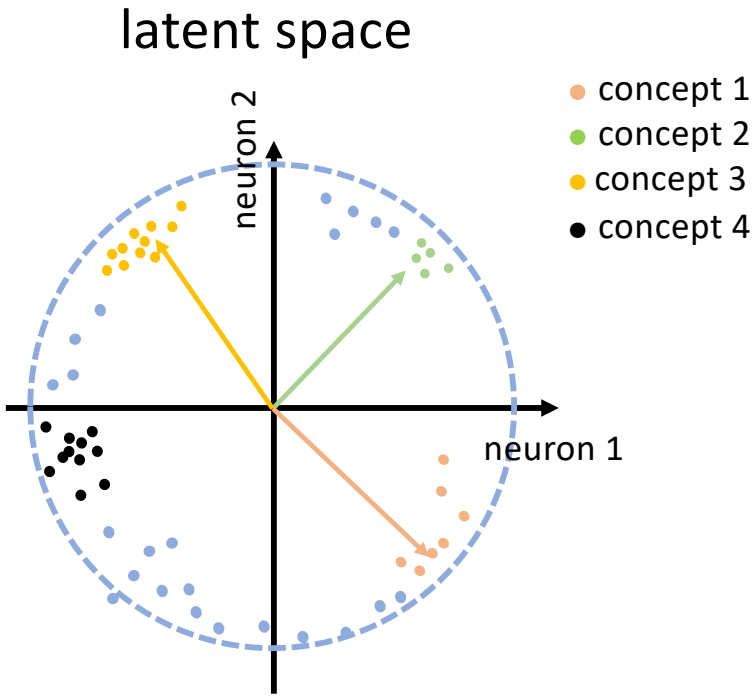




When a CW module is added to a CNN,

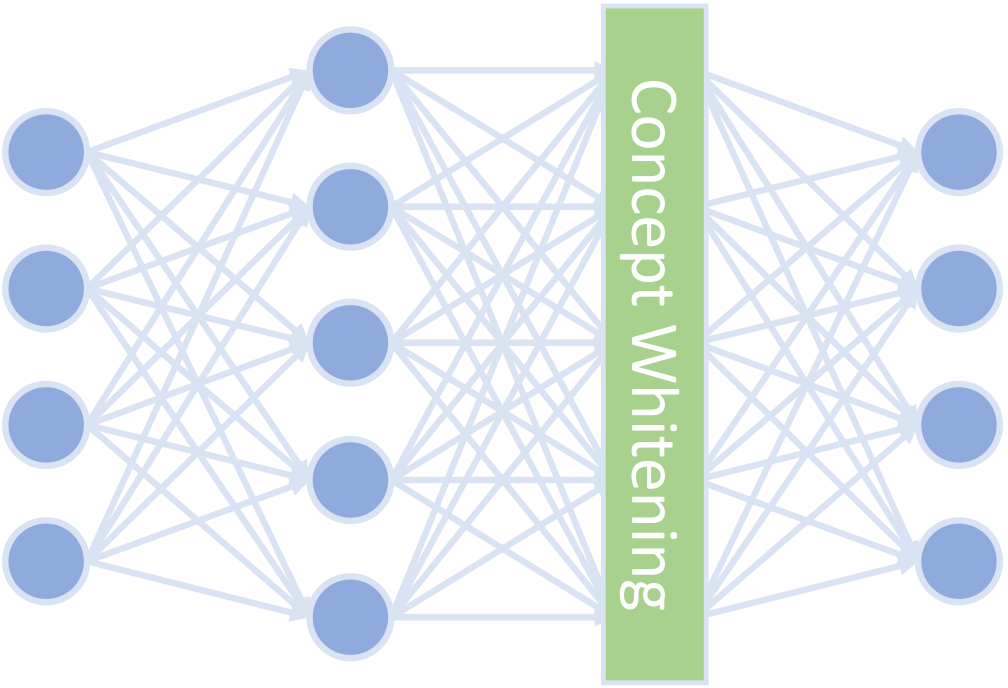
- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest

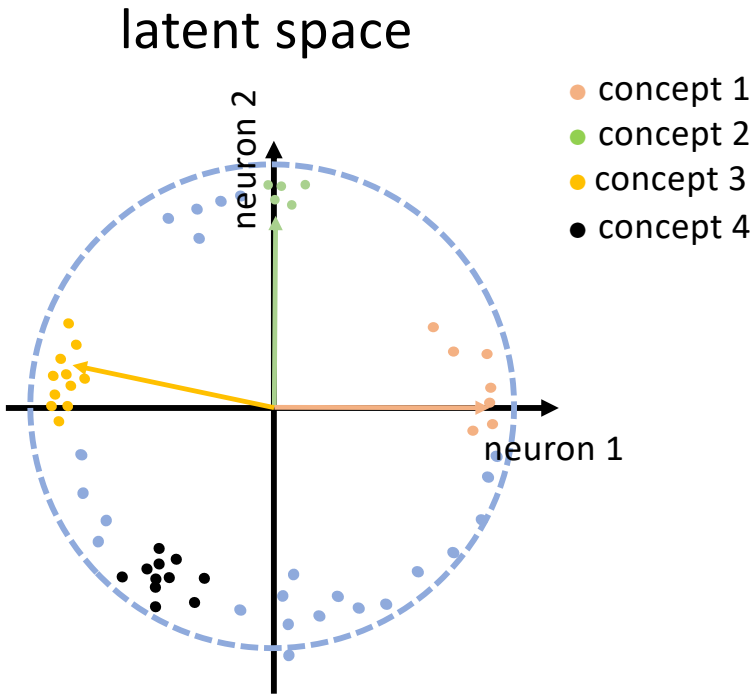




When a CW module is added to a CNN,

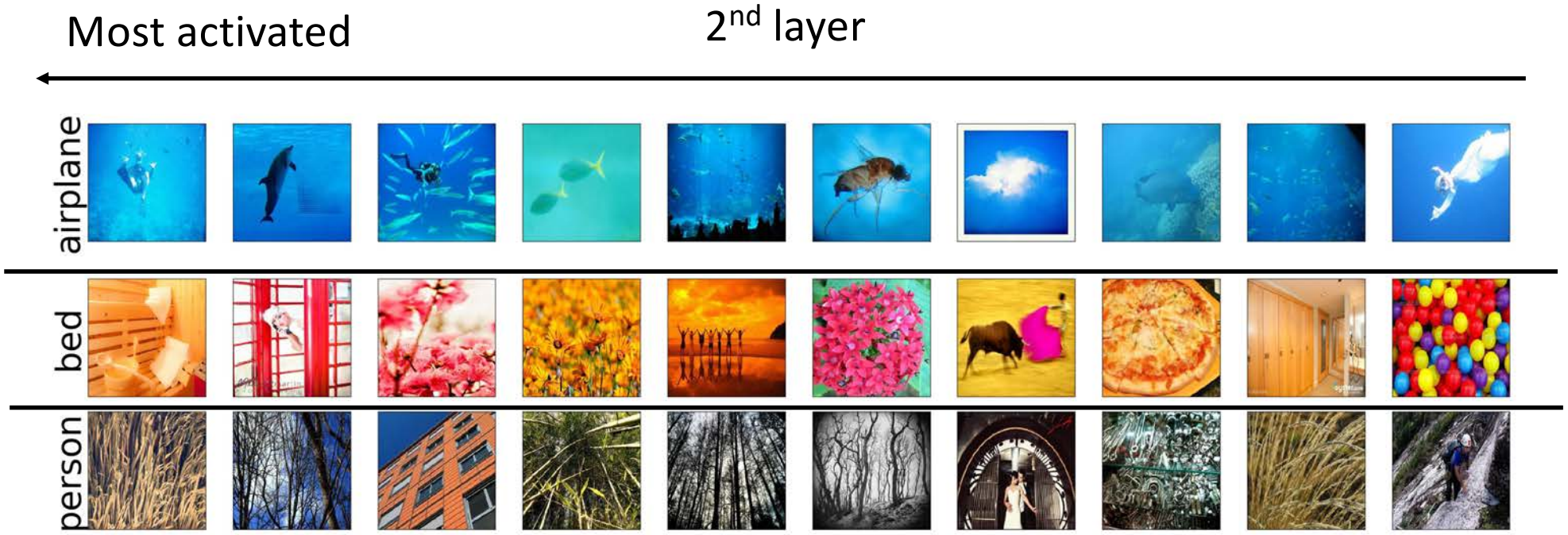
- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest





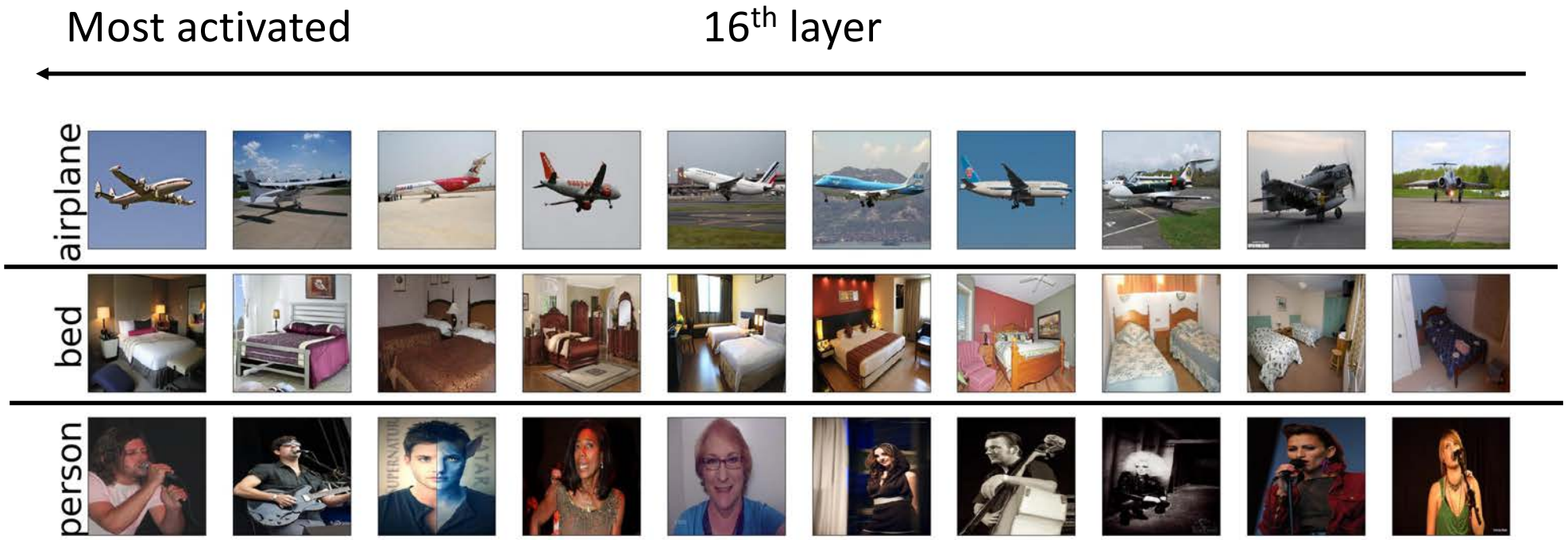
When a CW module is added to a CNN,

- the latent space is whitened (decorrelated and normalized)
- the axes of the latent space are aligned with concepts of interest



When CW is added to different layers…

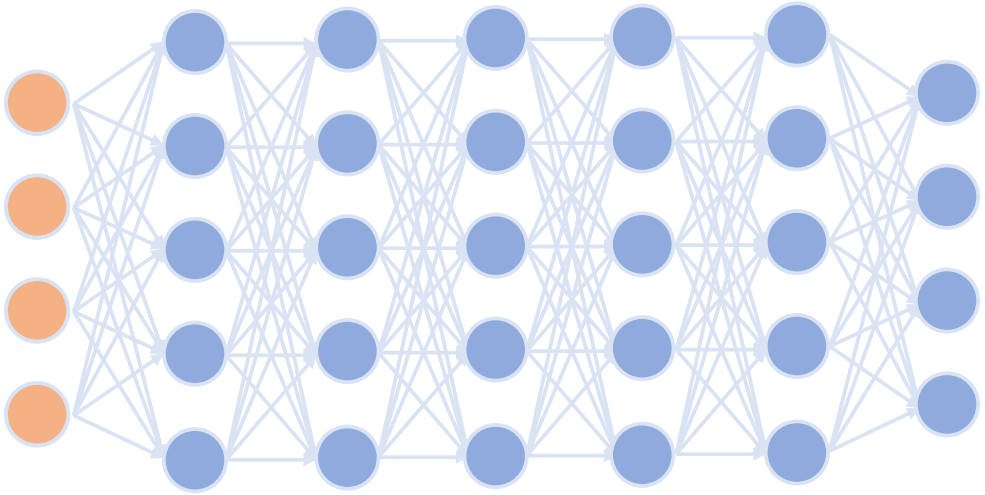
In earlier layers, color and texture information related to the concepts are represented along the axes

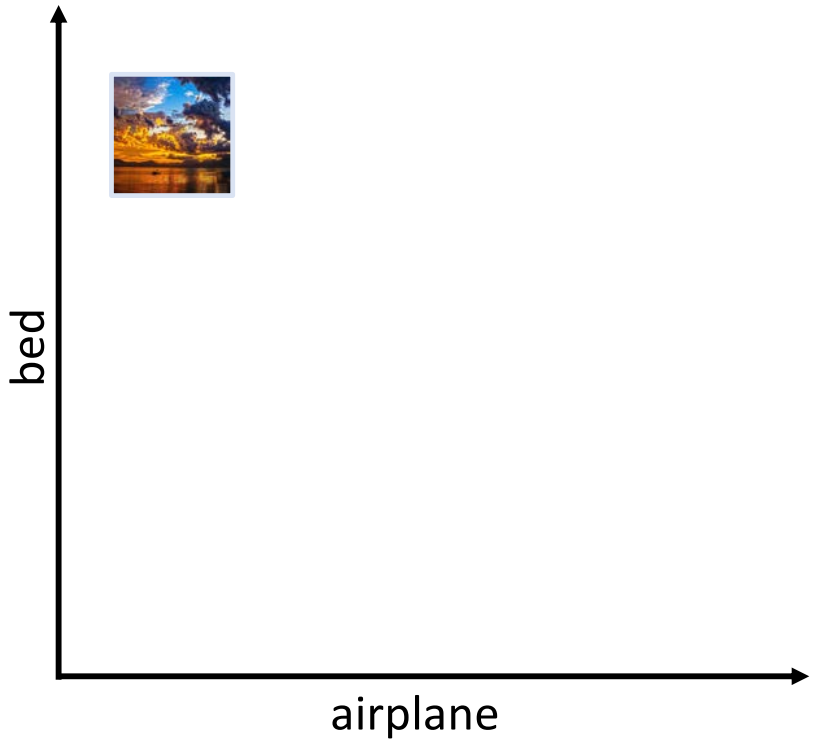


When CW is added to different layers…

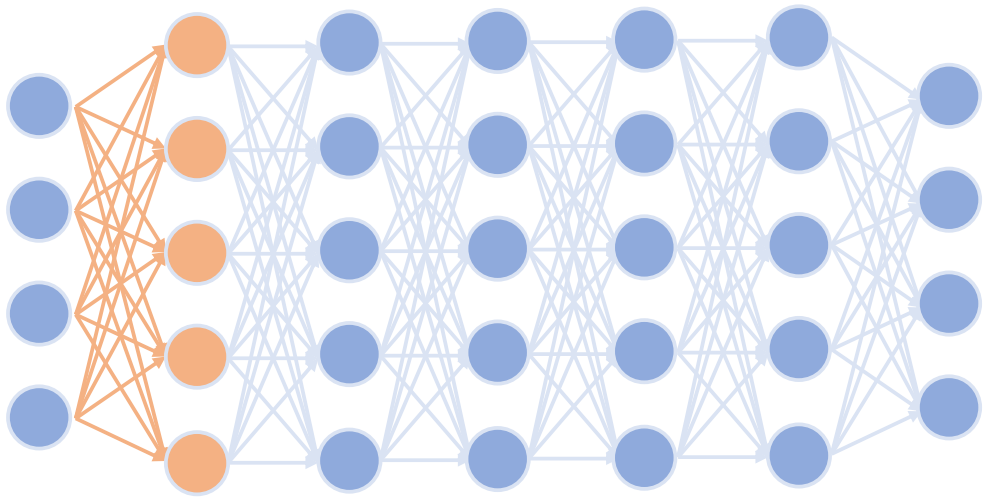
In deeper layers, pure high-level semantic meaning of target concepts is captured by the axes

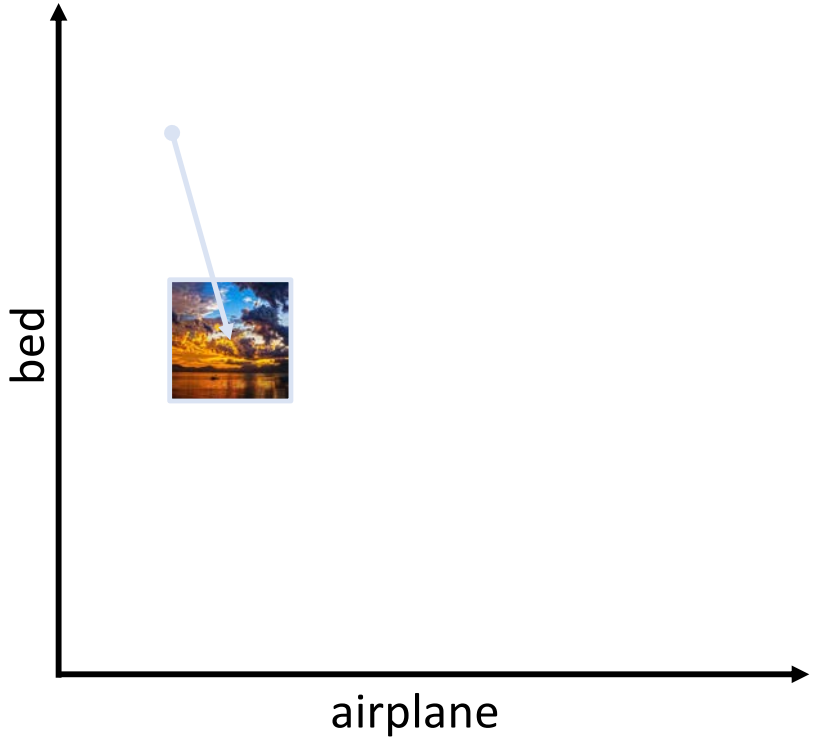
Because this image has warm colors, it lies mainly along the bed axis at layer 1



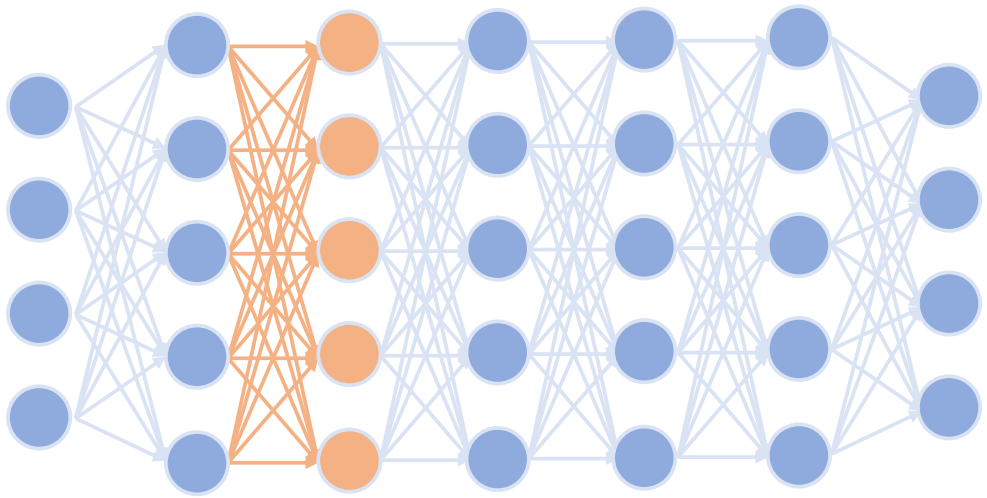


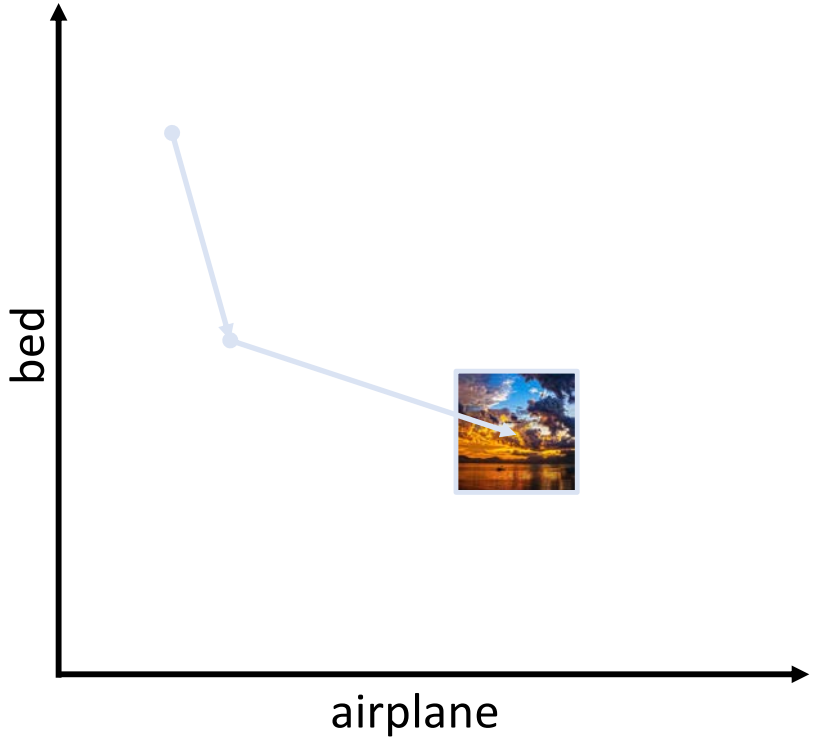
See how an image travels through the layers



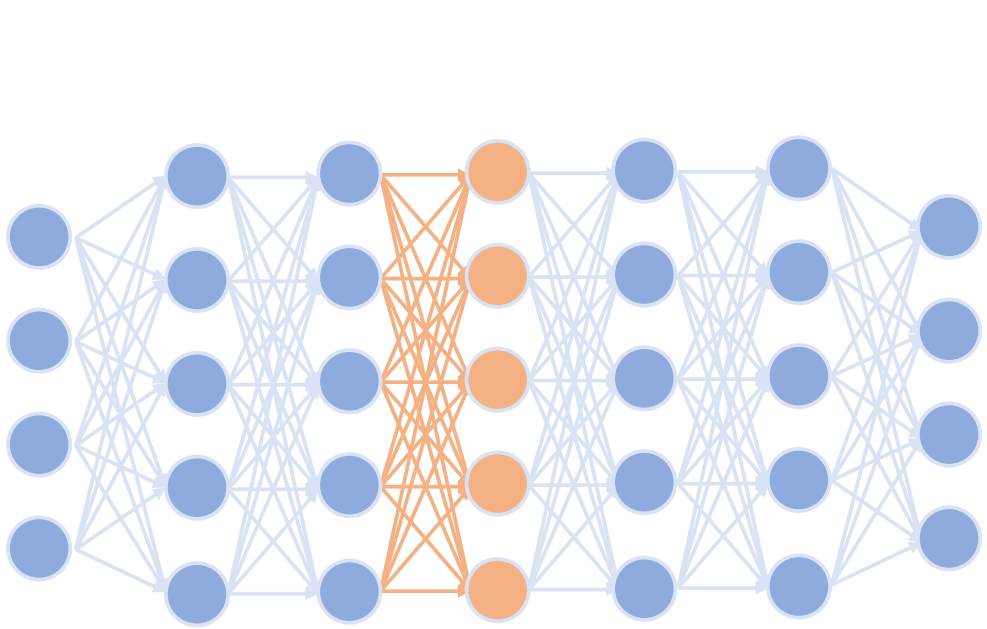


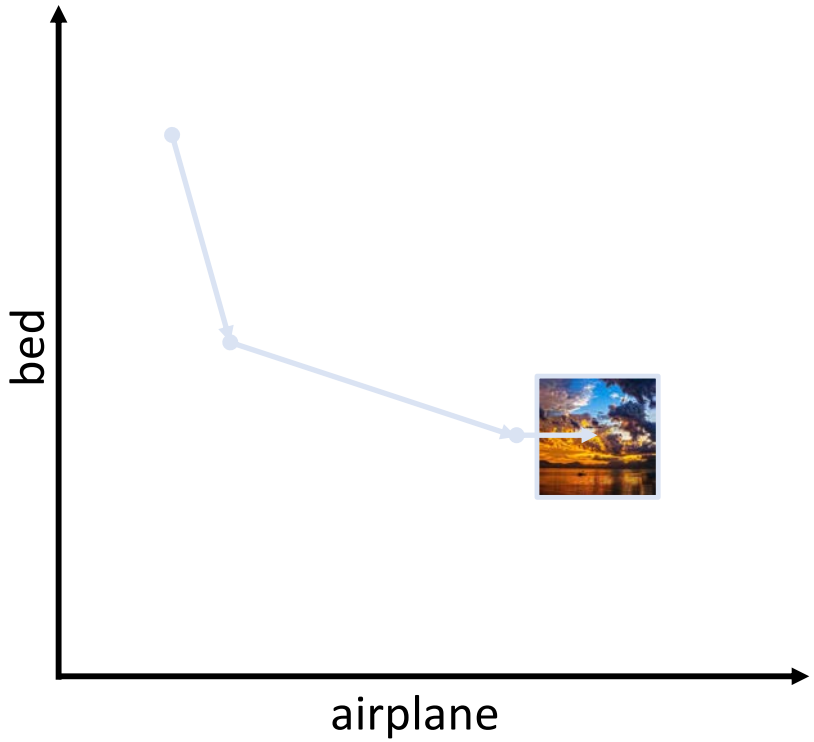
See how an image travels through the layers



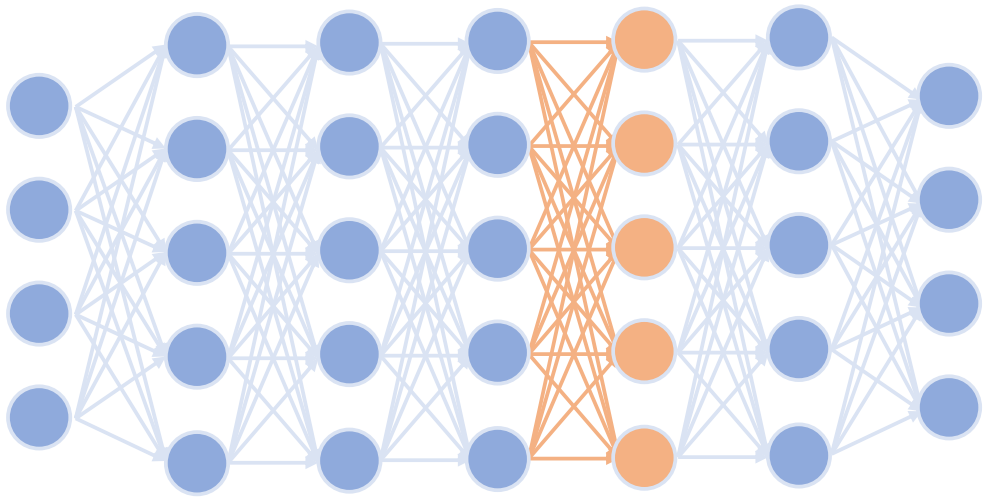


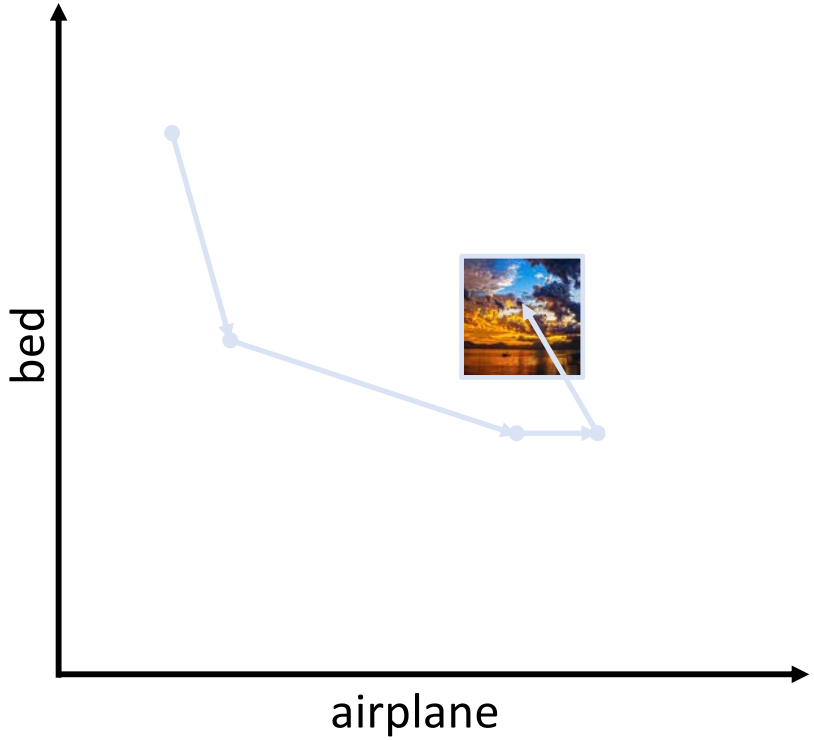
See how an image travels through the layers



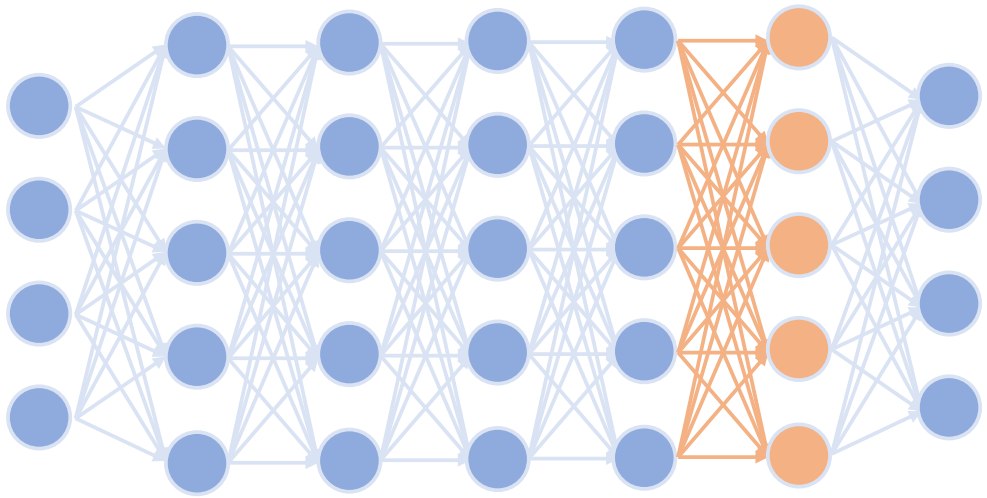


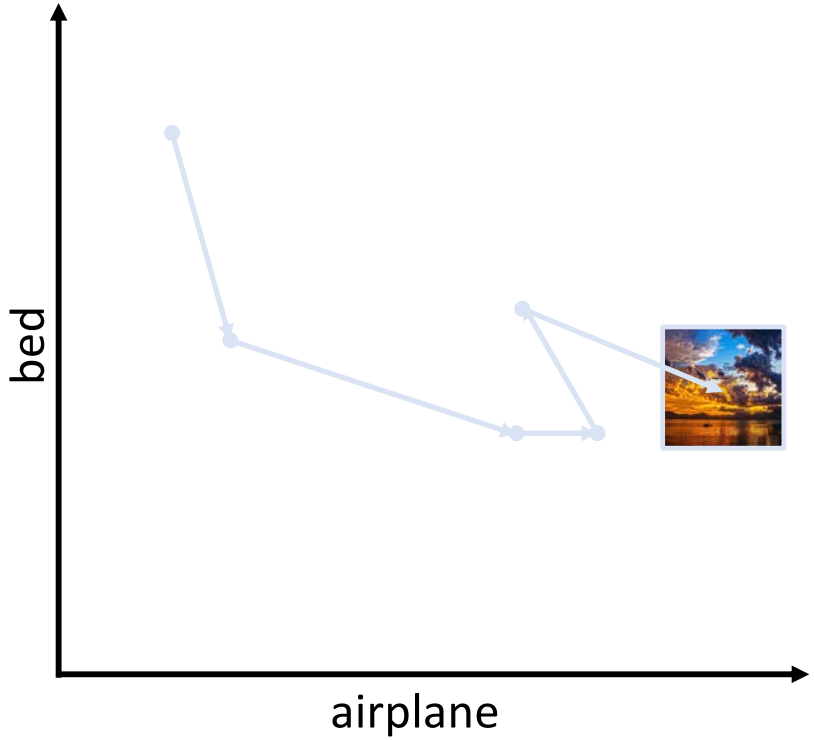
See how an image travels through the layers



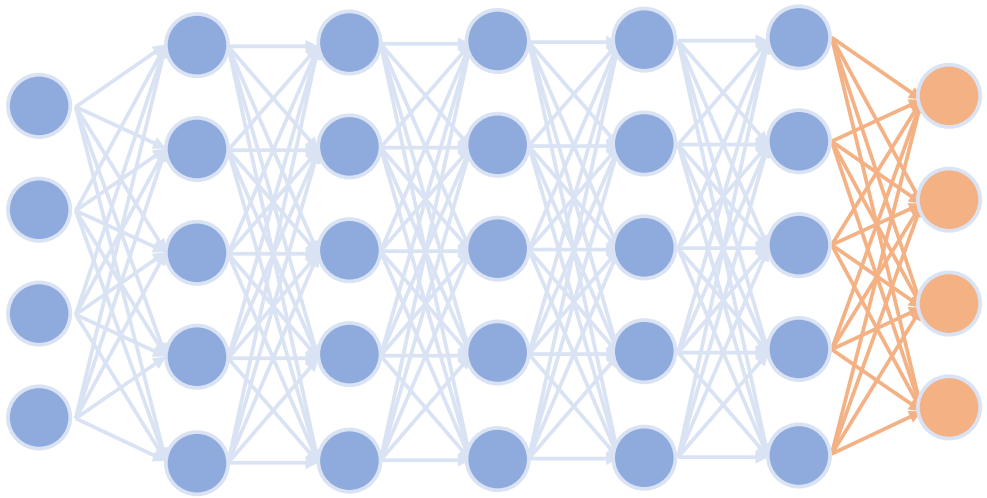


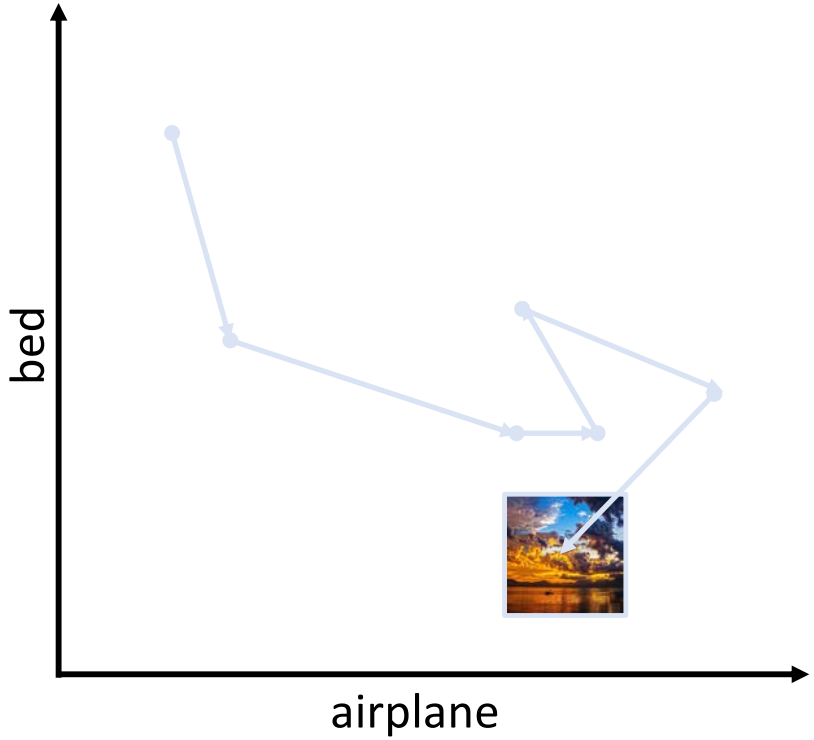
See how an image travels through the layers





See how an image travels through the layers





See how an image travels through the layers

Advantages of CW over BatchNorm

- ➢ No sacrifice in accuracy

  - accuracy is on par with standard CNNs

- ➢ Easy to use

  - warm-start from pretrained model requires only one additional epoch of further training
  - Note: requires training data for the concepts to define the axes

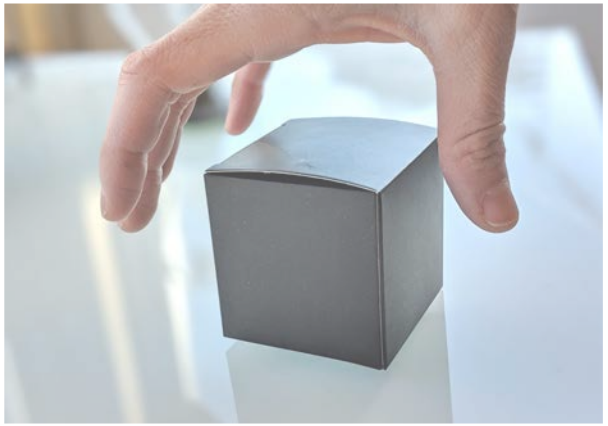
- ➢ Disentangles the latent space

Summary /Perspective

- Predictive models are everywhere
  - Criminal justice, air pollution, allocation of health services, health treatment, recommender systems, automated driving systems

- Don’t we want to be able to troubleshoot them?
- Don’t we want to detect if they biased in a harmful way?
- Don’t we want to be able to ensure a typo didn’t make a decision for us?
- Don’t we want to be able to use them in high stakes decisions?
- Don’t we want to understand them?

Thanks



Interpretability vs Explainability

Cynthia Rudin Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead,
Nature Machine Intelligence, 2019.

RiskSLIM

Berk Ustun and Cynthia Rudin
Learning Optimized Risk Scores.
Journal of Machine Learning Research, 2019

2HELP2B

Aaron F. Struck, Berk Ustun, ….., Cynthia Rudin, M Brandon Westover.
Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients. JAMA Neurology, 2017

CORELS

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin Learning Certifiably Optimal Rule Lists for Categorical Data
Journal of Machine Learning Research, 2018.

Analysis of COMPAS

Rudin, Wang, and Coker. The *Age* of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2020

Concept Whitening for Neural Disentanglement

Zhi Chen, Yijie Bei, and Cynthia Rudin
Concept Whitening for Interpretable Image Recognition.
Nature Machine Intelligence, accepted, 2020.