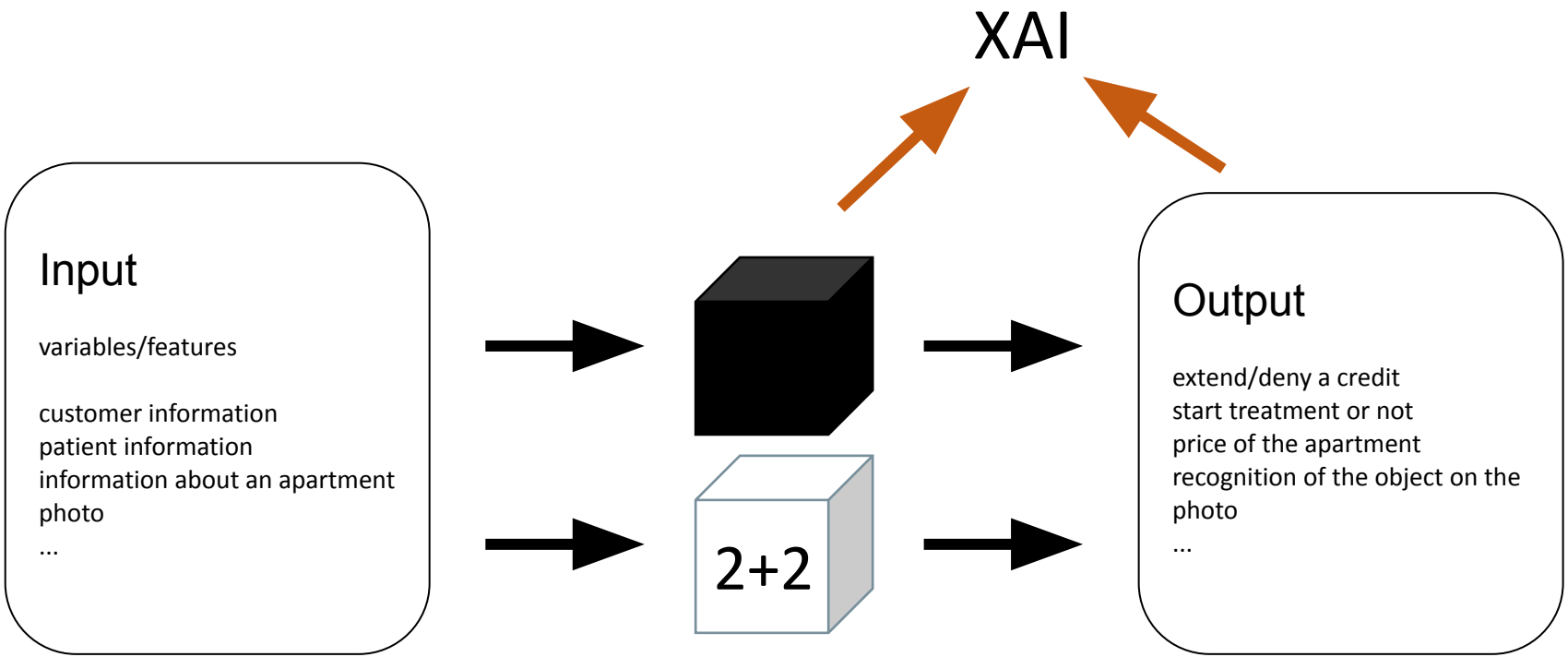
eXplainable Artificial Intelligence (XAI) - wstęp, czyli co to takiego?

Warsztaty Badawcze 2021L

Czym jest eXplainable Artificial Intelligence (XAI)?



Apple Card and Goldman Sachs accused of gender discrimination in credit card algorithm

Posted earlier today at 6:38am

PHOTO: Apple Card has caused a Twitter storm and even prompted a regulator inquiry. (Reuters: Stephen Lam)

Apple launched its own credit card in the United States a few months ago, selling it on the ability to help people keep track of their spending while protecting their privacy.

RELATED STORY: Apple wants you to read, watch and play its new services … using its new credit card

REUTERS World Business Markets Politics TV

Midterm Elections | Imprisoned In Myanmar | Sectors Up Close | Breakingviews | Investing | Future of Money | Charged: The Future of Aut

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

Amazon scraps secret AI showed bias against wom

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist

THE VERGE TECH SCIENCE C

TECH AMAZON ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

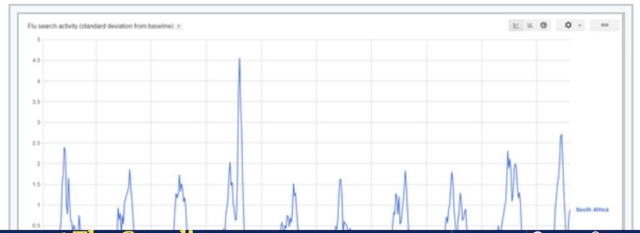
By James Vincent | @jjvincent | Oct 10, 2018, 7:09am EDT

Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offerec



Forbes Billionaires Innovation Leadership Money Con

61,215 views | Mar 23, 2014, 09:00am

Why Google Flu Is A Failu

Steven Salzberg Contributor
Pharma & Healthcare

It seemed like such a good idea at the time.

Support The Guardian
Contribute → Subscribe →

Search jobs | Sign in | Search | International edition

The Guardian

News Opinion Sport Culture Lifestyle More

World UK Science Cities Global development Football Tech Business Environment Obituaries

Google

Google Flu Trends is no longer good at predicting flu, scientists find

Researchers warn of 'big data hubris' and the importance of updating analytical models, claiming Google has made inaccurate forecasts for 100 of 108 weeks

Charles Arthur
@charlesarthur
Thu 27 Mar 2014 10.27 GMT

200 7

This article is over 4 years old

▲ Airport security personnel take a body temperature reading of a boy as he arrives at Hong Kong International Airport April 9,2013, following concerns over a deadly strain of bird flu. Photograph: Tyrone Siu/Reuters

most viewed

Dozens of Indian paramilitaries killed in Kashmir car bombing

Live Brexit: blow to May's authority as MPs reject her motion by 303 votes to 258 - as it happened

Theresa May defeated on Brexit again as ERG Tories abstain

Live Trump to sign government funding bill and declare national emergency - live

Andrew McCabe says officials discussed removing Trump after Comey firing

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

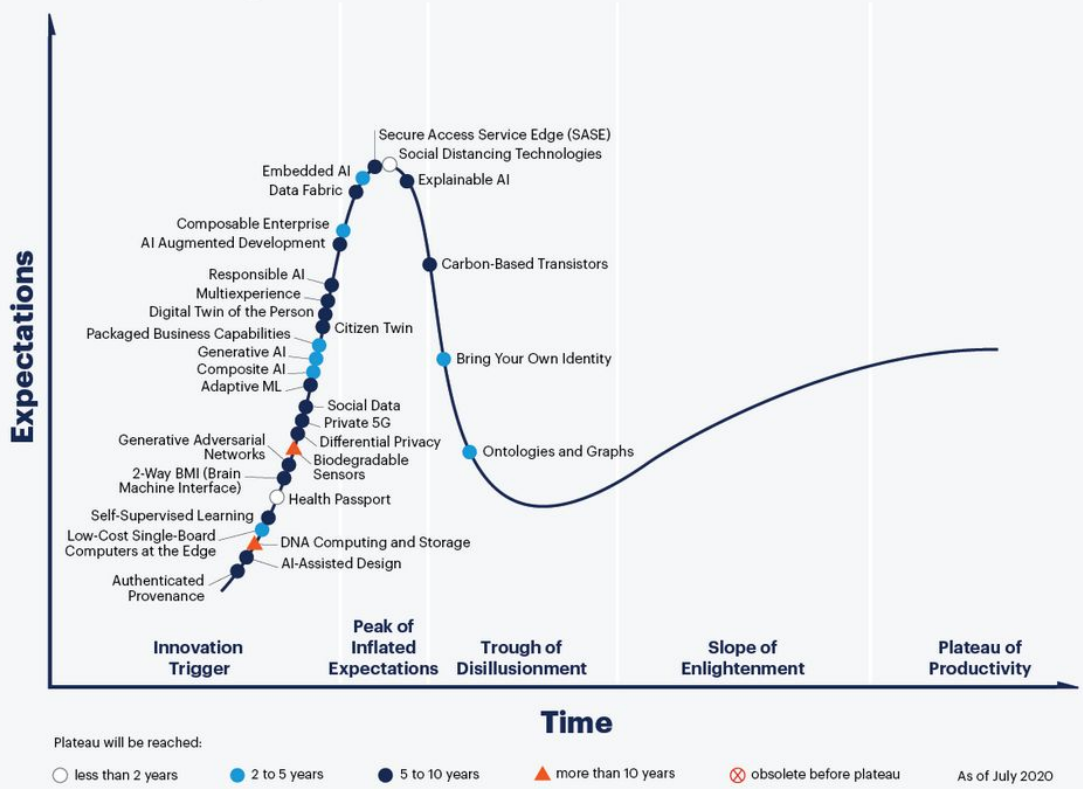
Machine Learning is Creating a Crisis in Science

The adoption of machine-learning techniques is contributing to a worrying number of research findings that cannot be repeated by other researchers.

Kevin McCaney
Wed, 02/27/2019 - 11:28

Photo credit: metamorworks/iStock

Hype Cycle for Emerging Technologies, 2020

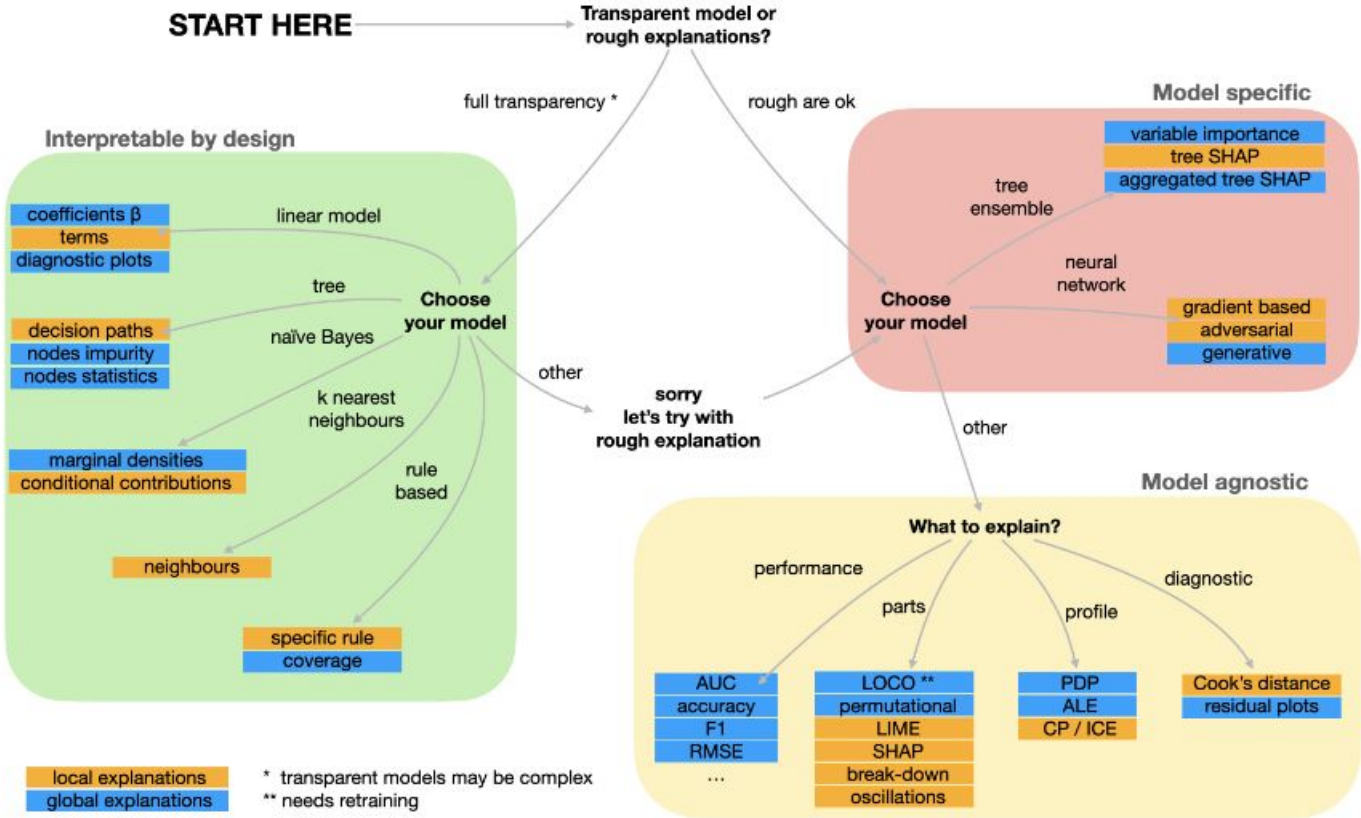


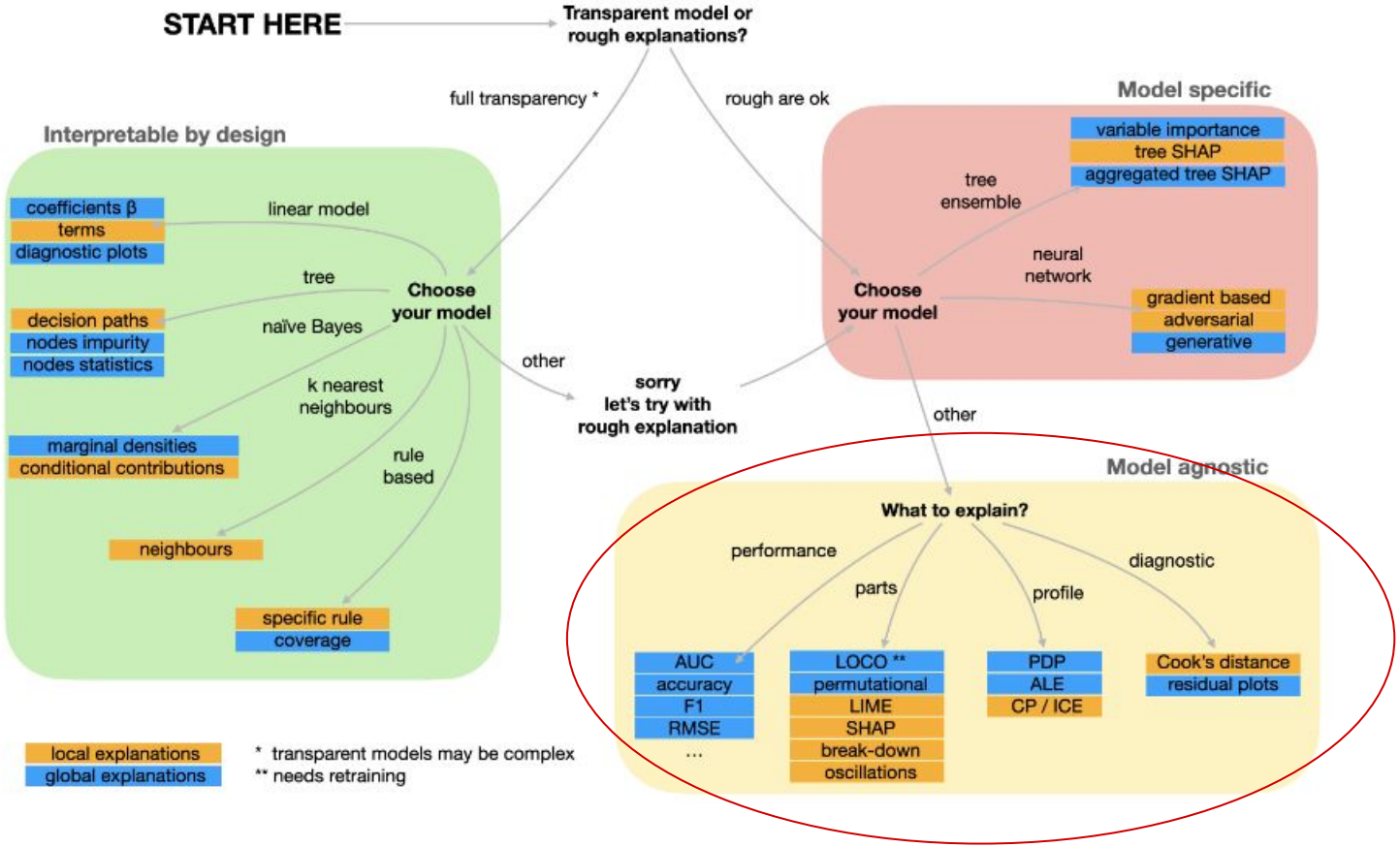
gartner.com/SmarterWithGartner

Source: Gartner

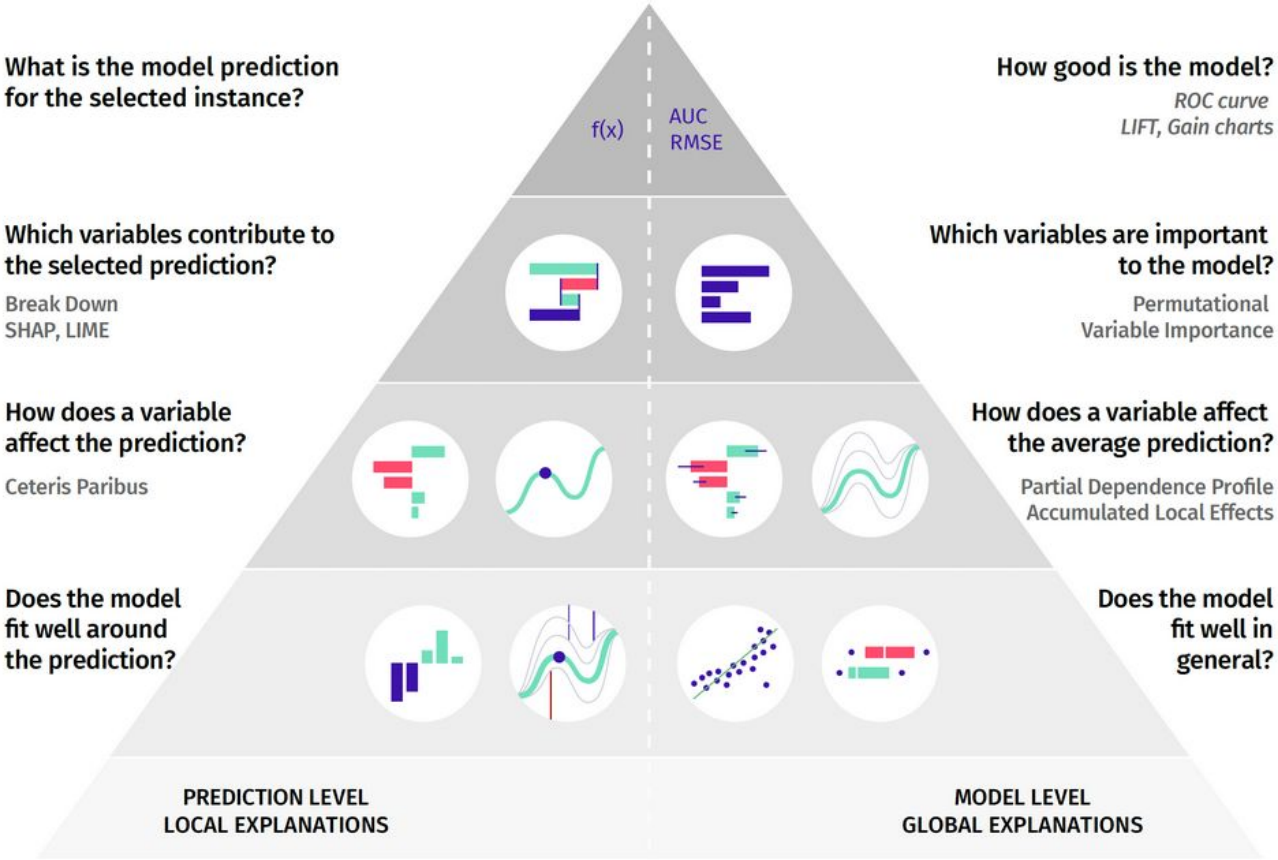
© 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.

Gartner


START HERE
Transparent model or rough explanations?
full transparency *
rough are ok
Interpretable by design
coefficients β
terms
diagnostic plots
linear model
tree
decision paths
nodes impurity
nodes statistics
naïve Bayes
Choose your model
k nearest neighbours
marginal densities
conditional contributions
rule based
neighbours
specific rule
coverage
other
sorry let's try with rough explanation
Model specific
variable importance
tree SHAP
aggregated tree SHAP
tree ensemble
neural network
Choose your model
gradient based
adversarial
generative
other
Model agnostic
What to explain?
performance
parts
profile
diagnostic
AUC
accuracy
F1
RMSE
...
LOCO **
permutational
LIME
SHAP
break-down
oscillations
PDP
ALE
CP / ICE
Cook's distance
residual plots
local explanations
global explanations
* transparent models may be complex
** needs retraining

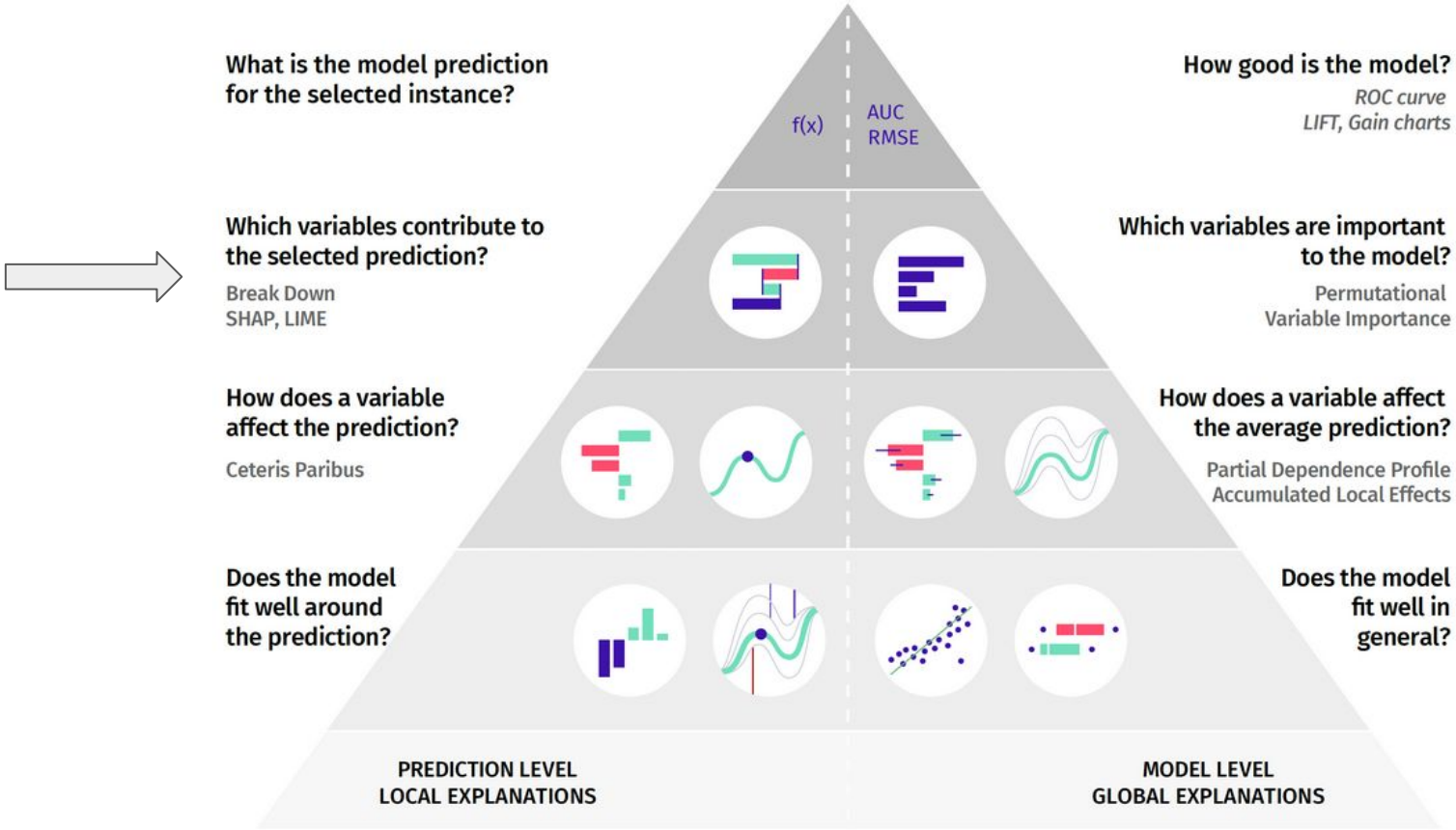

START HERE
Transparent model or rough explanations?
full transparency *
rough are ok
Interpretable by design
coefficients β
terms
diagnostic plots
linear model
tree
decision paths
nodes impurity
nodes statistics
naïve Bayes
k nearest neighbours
marginal densities
conditional contributions
rule based
neighbours
specific rule
coverage
Choose your model
other
sorry let's try with rough explanation
Model specific
variable importance
tree SHAP
aggregated tree SHAP
tree ensemble
neural network
gradient based
adversarial
generative
Choose your model
other
Model agnostic
What to explain?
performance
parts
profile
diagnostic
AUC
accuracy
F1
RMSE
...
LOCO **
permutational
LIME
SHAP
break-down
oscillations
PDP
ALE
CP / ICE
Cook's distance
residual plots
local explanations
global explanations
* transparent models may be complex
** needs retraining

Model Exploration Stack



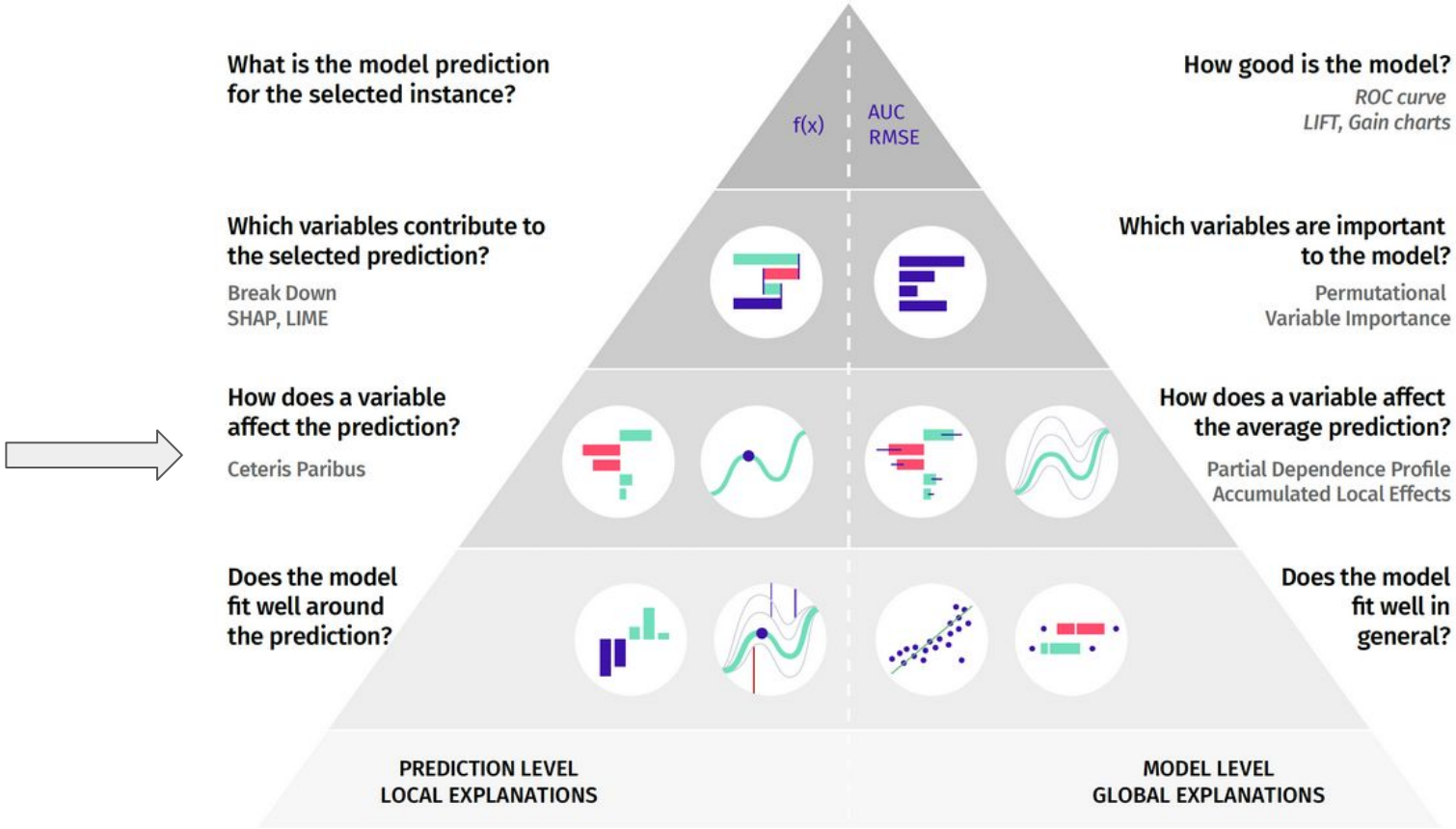
Source: Biecek, Przemysław, and Tomasz Burzykowski. 2019. *Explanatory Model Analysis*. https://pbiecek.github.io/ema/.

Model Exploration Stack



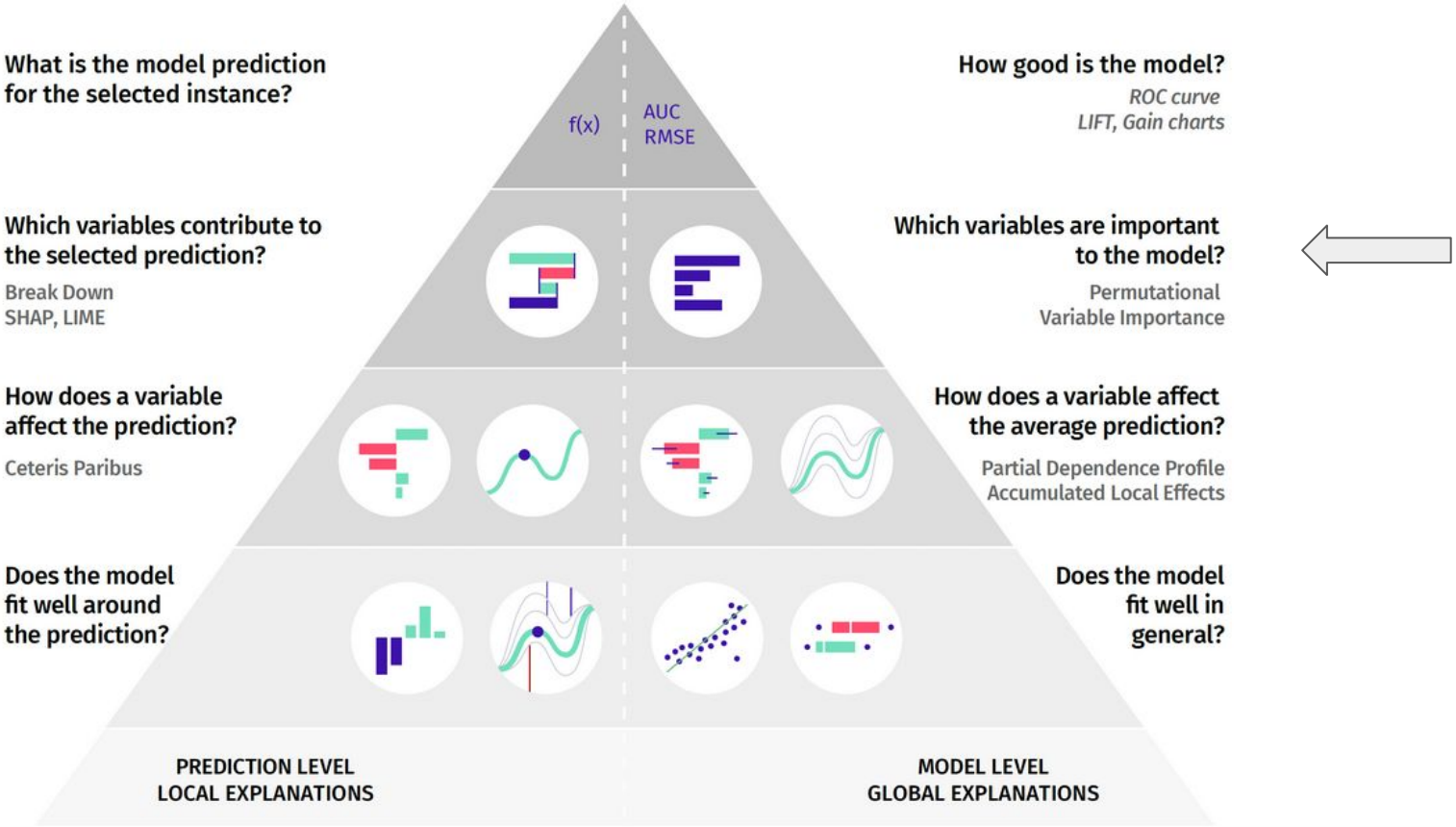
Source: Biecek, Przemysław, and Tomasz Burzykowski. 2019. *Explanatory Model Analysis*. https://pbiecek.github.io/ema/.

Model Exploration Stack



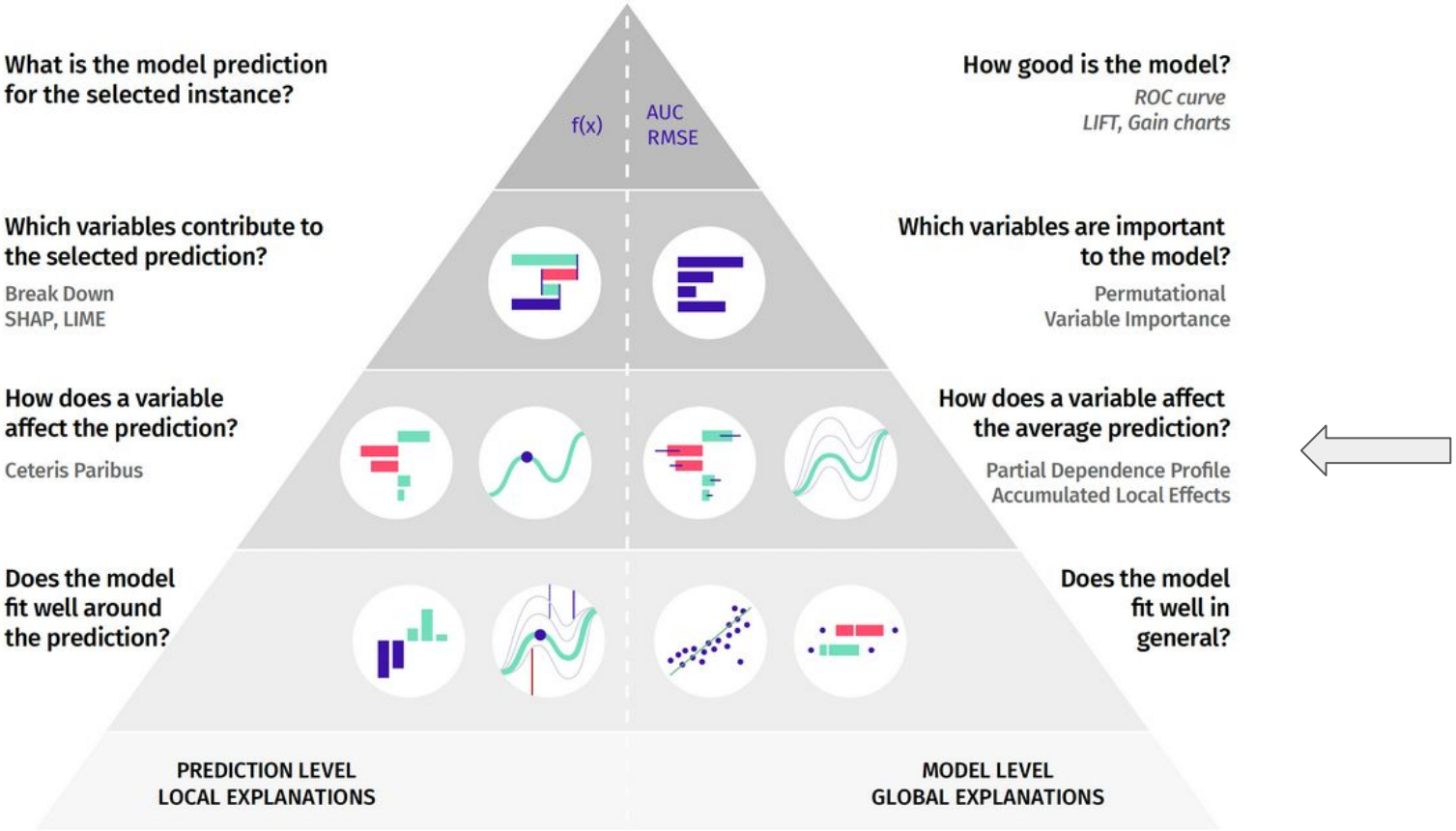
Source: Biecek, Przemysław, and Tomasz Burzykowski. 2019. *Explanatory Model Analysis*. https://pbiecek.github.io/ema/.

Model Exploration Stack



Source: Biecek, Przemysław, and Tomasz Burzykowski. 2019. *Explanatory Model Analysis*. https://pbiecek.github.io/ema/.

Model Exploration Stack



Source: Biecek, Przemysław, and Tomasz Burzykowski. 2019. *Explanatory Model Analysis*. https://pbiecek.github.io/ema/.