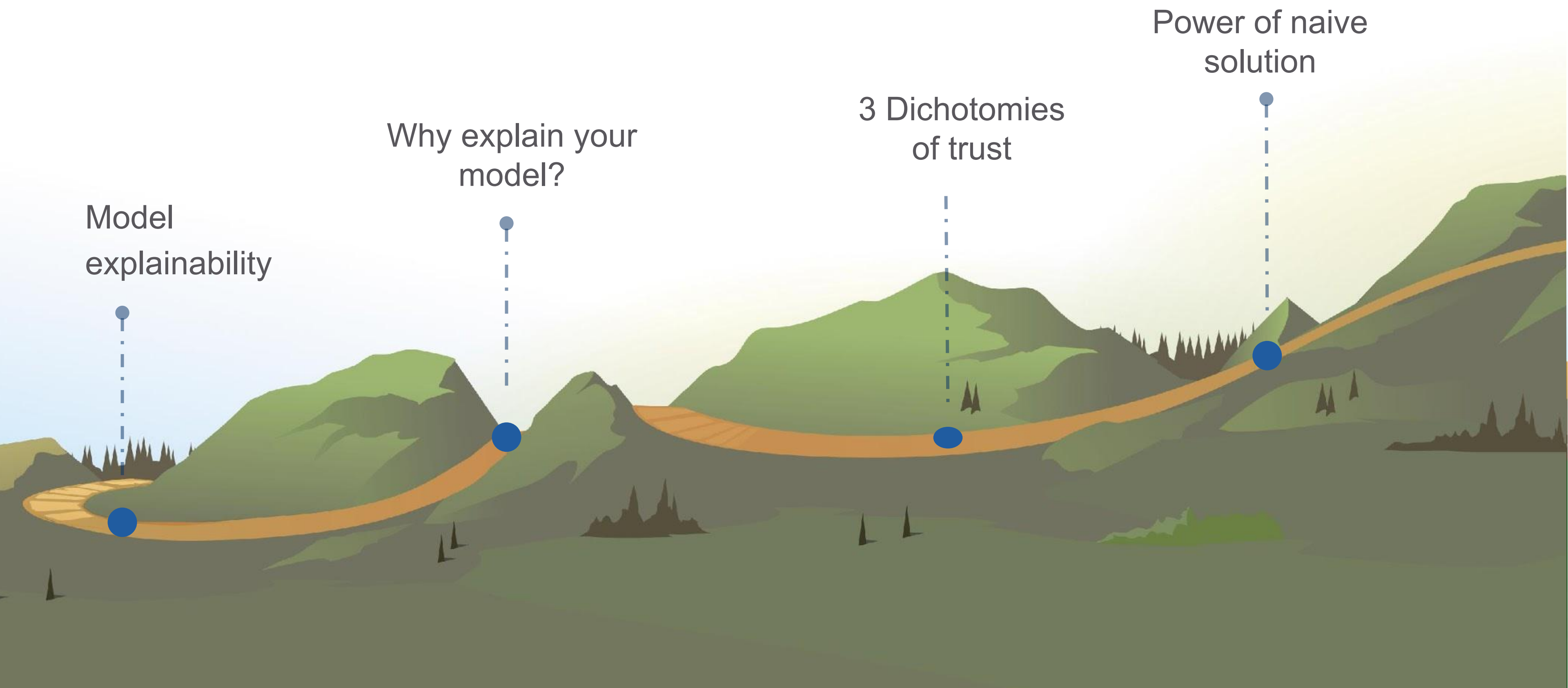
Trustworthiness of Machine Learning Applications

Mayukh Bhaowal

Director of Product Management, Salesforce Einstein

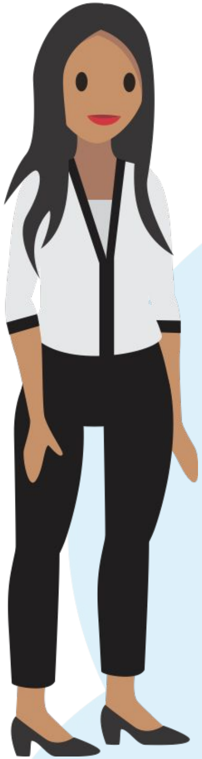
Roadmap for this talk



The Question

Why did the machine learning model make the decision that it did?

Translation #1



How do I fix this model?

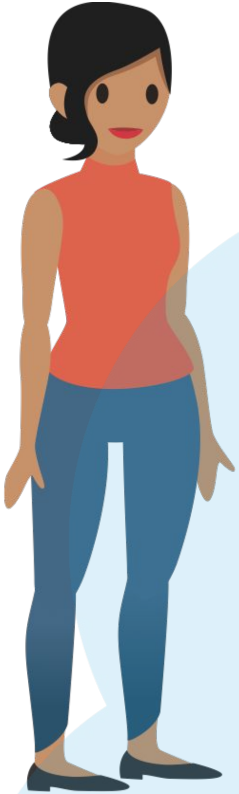
— **Data Scientist**

Translation #2

Do we have our bases covered, in case of a regulatory audit?

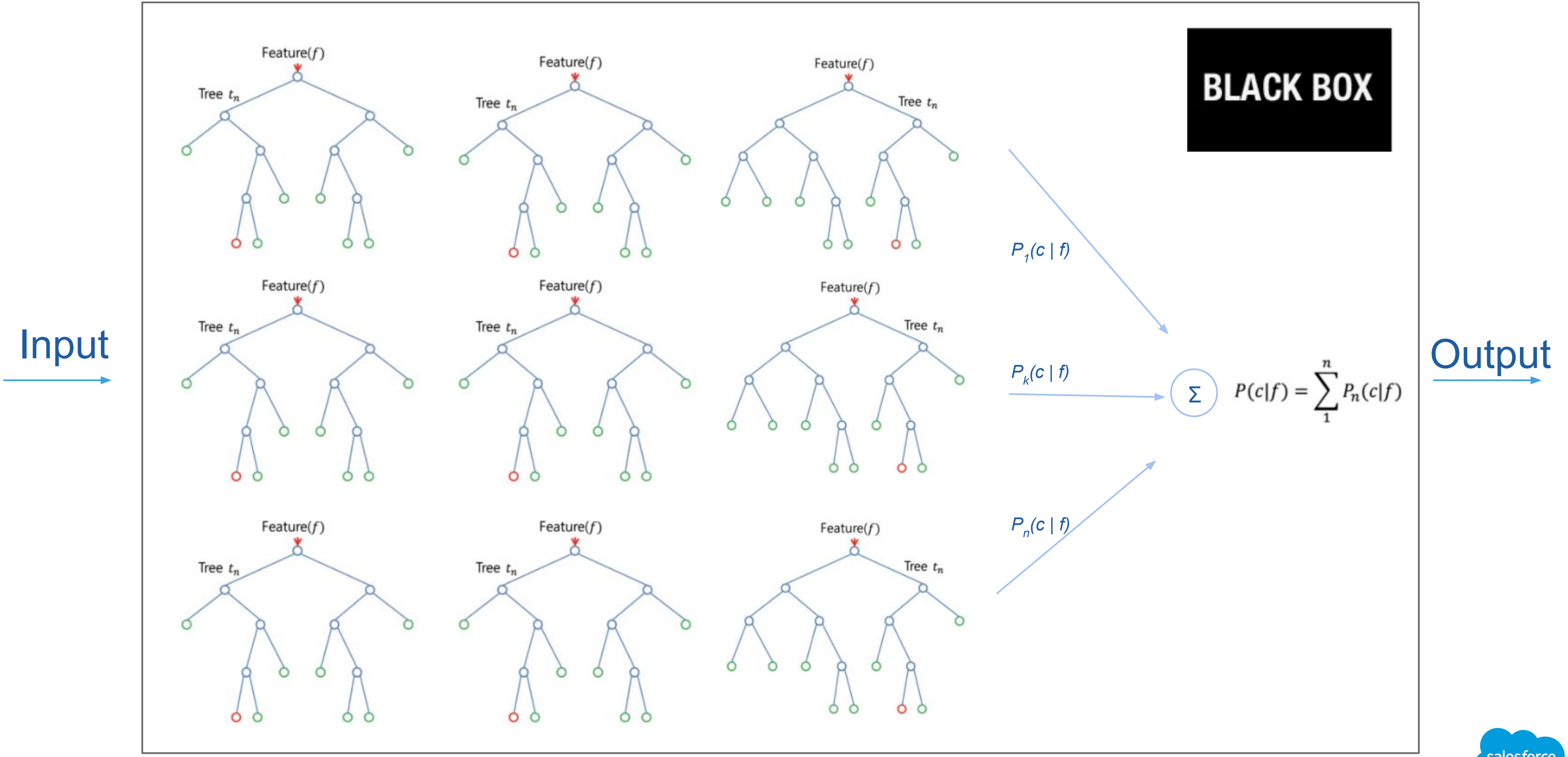
— **Legal Counsel**

Translation #3



Does Einstein know what I know? How do I use this prediction?

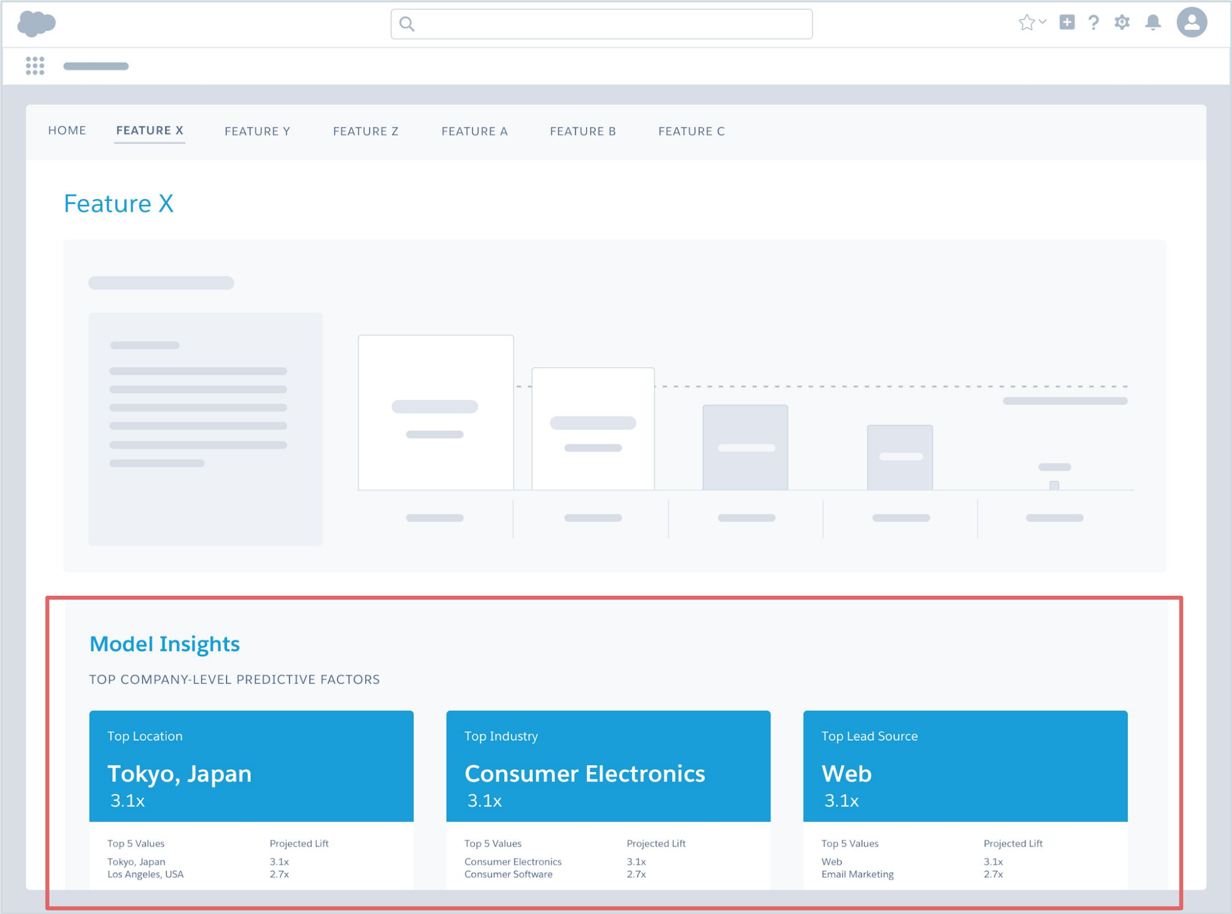
— Non Technical End User


Input
Feature(f)
Tree t_n
BLACK BOX
$P_1(c \mid f)$
$P_k(c \mid f)$
$P_n(c \mid f)$
$P(c|f) = \sum_1^n P_n(c|f)$
Output

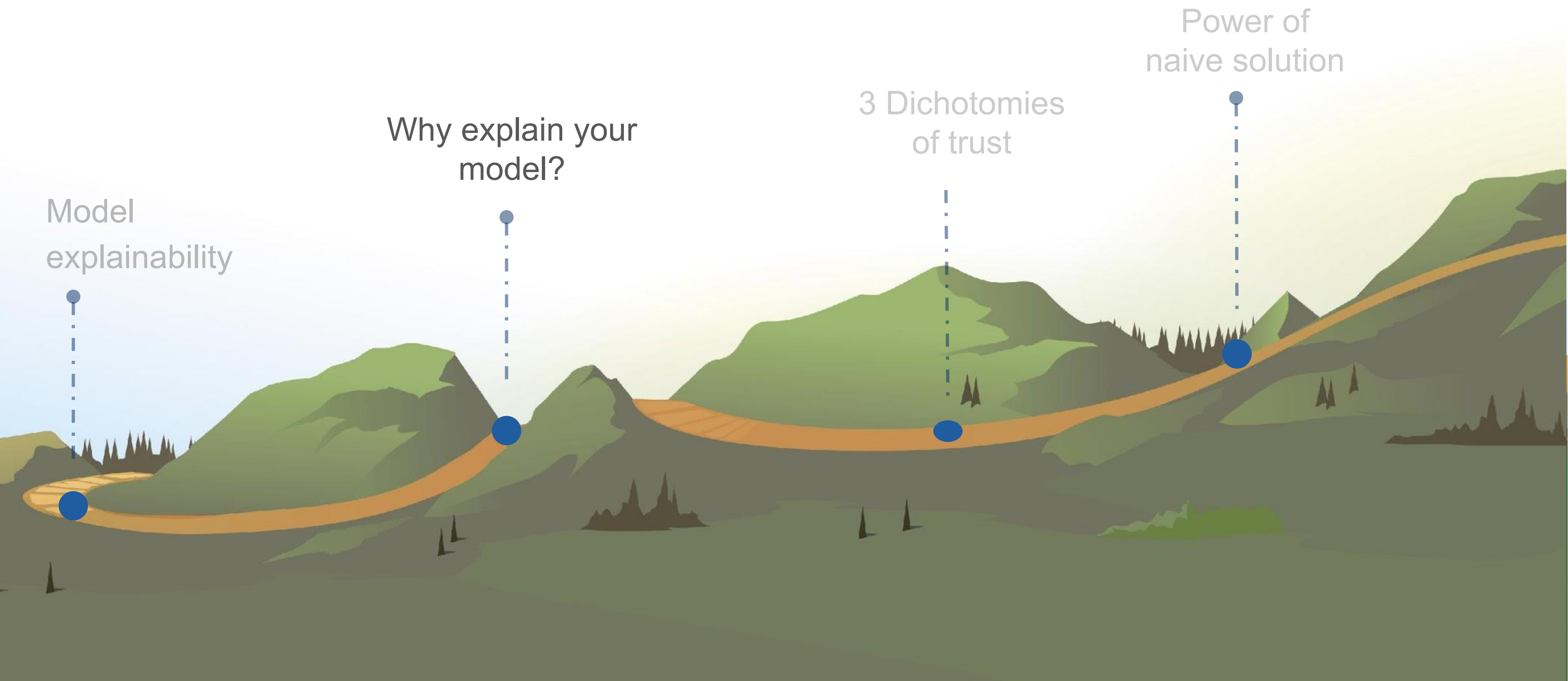
“How do I know I can trust it?”

Idea Concept:

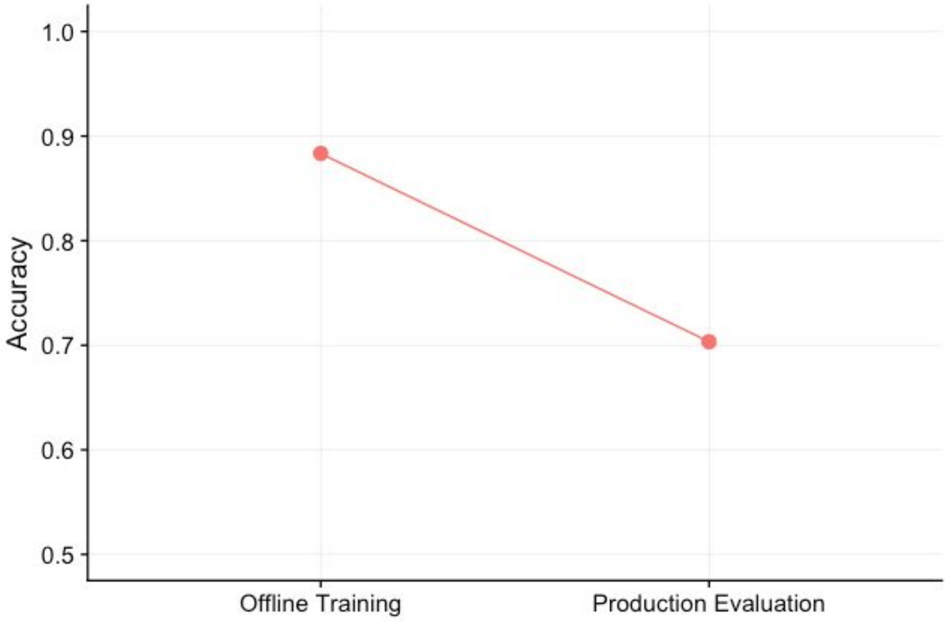
Model Insights Report



Roadmap for this talk



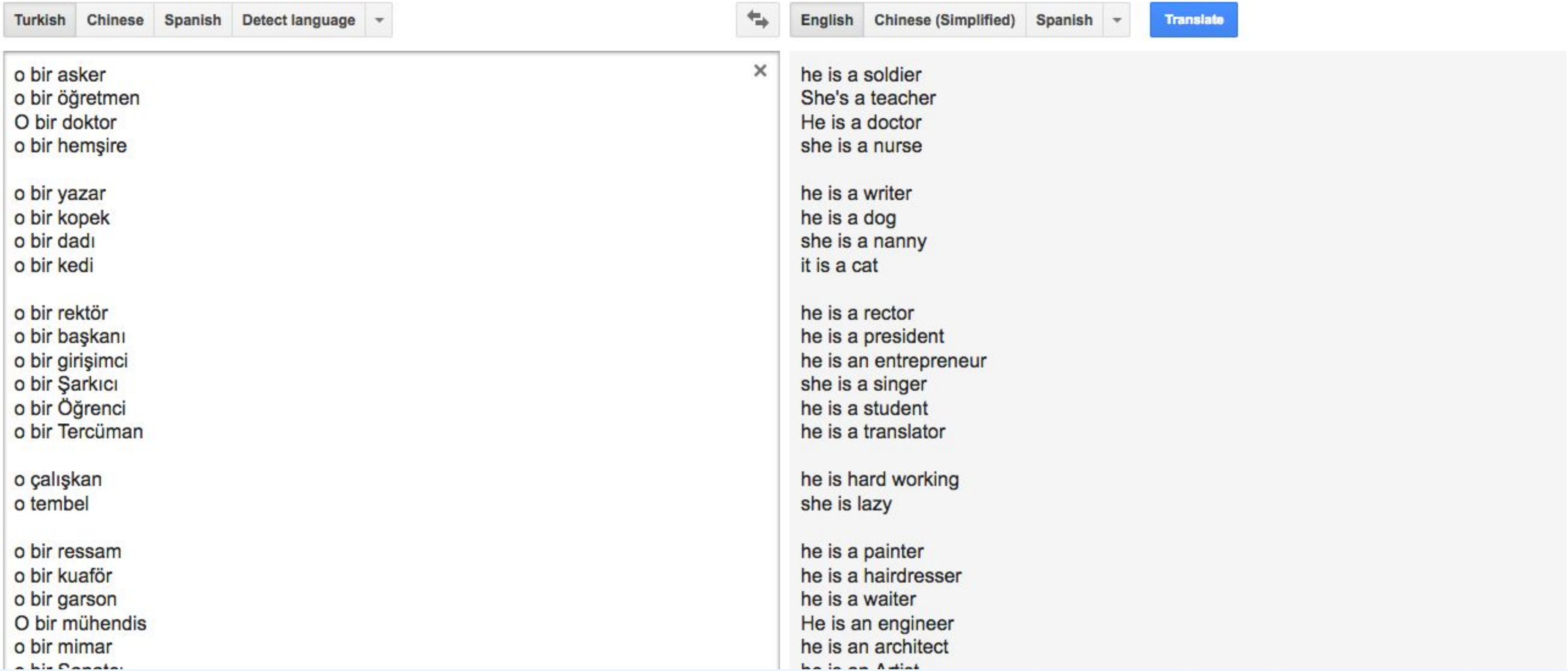
Debuggability



Top contributing features for predicting purchase:

1. Customer Interest Group
2. Thank you email
3. Customer Location

Bias


Turkish
Chinese
Spanish
Detect language
English
Chinese (Simplified)
Spanish
Translate
o bir asker
o bir öğretmen
O bir doktor
o bir hemşire
o bir yazar
o bir kopek
o bir dadı
o bir kedi
o bir rektör
o bir başkanı
o bir girişimci
o bir Şarkıcı
o bir Öğrenci
o bir Tercüman
o çalışkan
o tembel
o bir ressam
o bir kuaför
o bir garson
O bir mühendis
o bir mimar
he is a soldier
She's a teacher
He is a doctor
she is a nurse
he is a writer
he is a dog
she is a nanny
it is a cat
he is a rector
he is a president
he is an entrepreneur
she is a singer
he is a student
he is a translator
he is hard working
she is lazy
he is a painter
he is a hairdresser
he is a waiter
He is an engineer
he is an architect

Legal

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison

By Ellora Thadaney Israni

Oct. 26, 2017

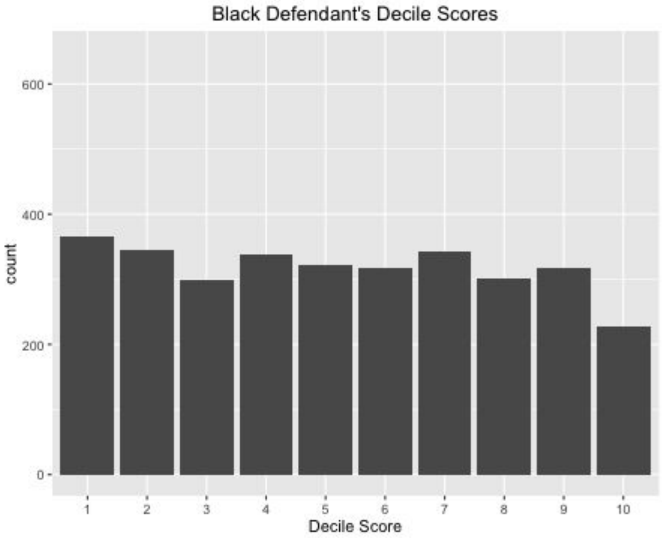
In 2013, police officers in Wisconsin arrested a man driving a car that had been used in a recent shooting. The man, Eric Loomis, pleaded guilty to attempting to flee an officer, and no contest to operating a vehicle without the owner's consent. Neither of his crimes mandates prison time.

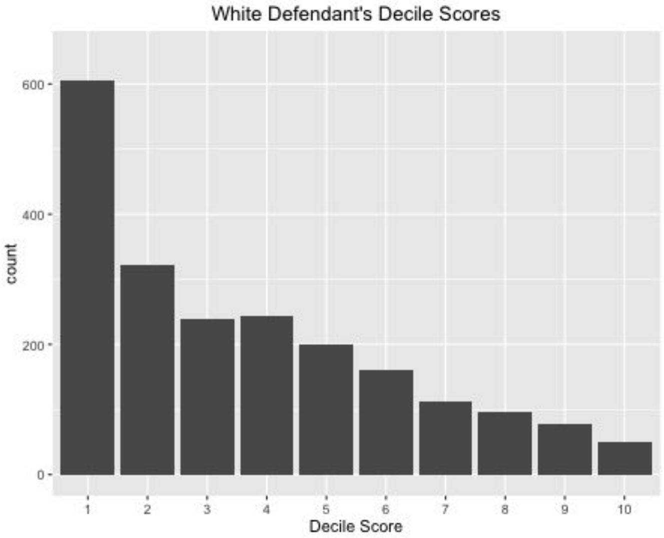
At Mr. Loomis's sentencing, the judge cited, among other factors, Mr. Loomis's high risk of recidivism as predicted by a computer program called COMPAS, a risk assessment algorithm used by the state of Wisconsin. The judge denied probation and prescribed an 11-year sentence: six years in prison, plus five years of extended supervision.

No one knows exactly how COMPAS works; its manufacturer refuses to disclose the proprietary algorithm. We only know the final risk assessment score it spits out, which judges may consider at sentencing.

Mr. Loomis challenged the use of an algorithm as a violation of his due

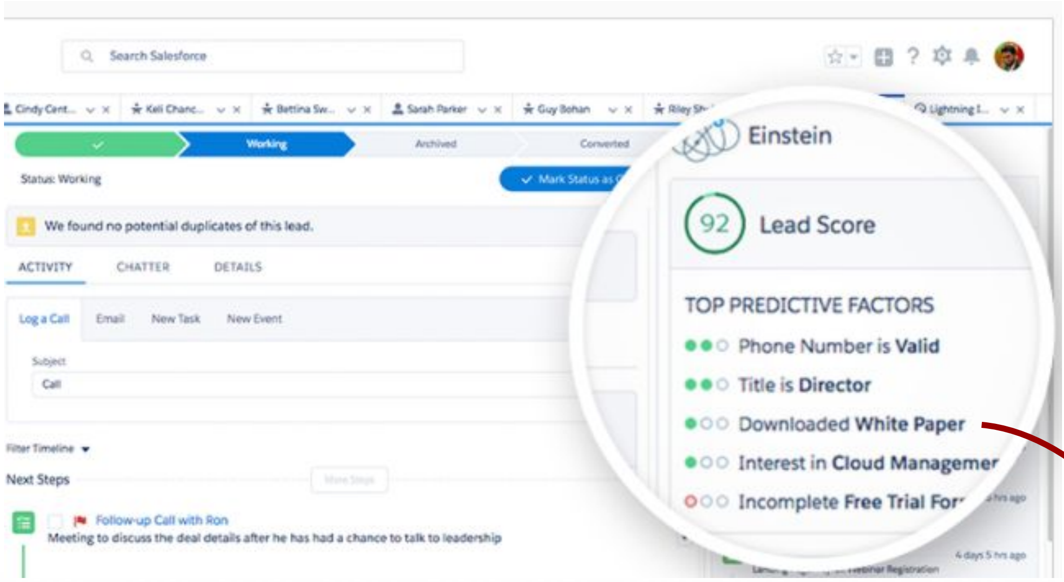
Black defendant has higher risk scores





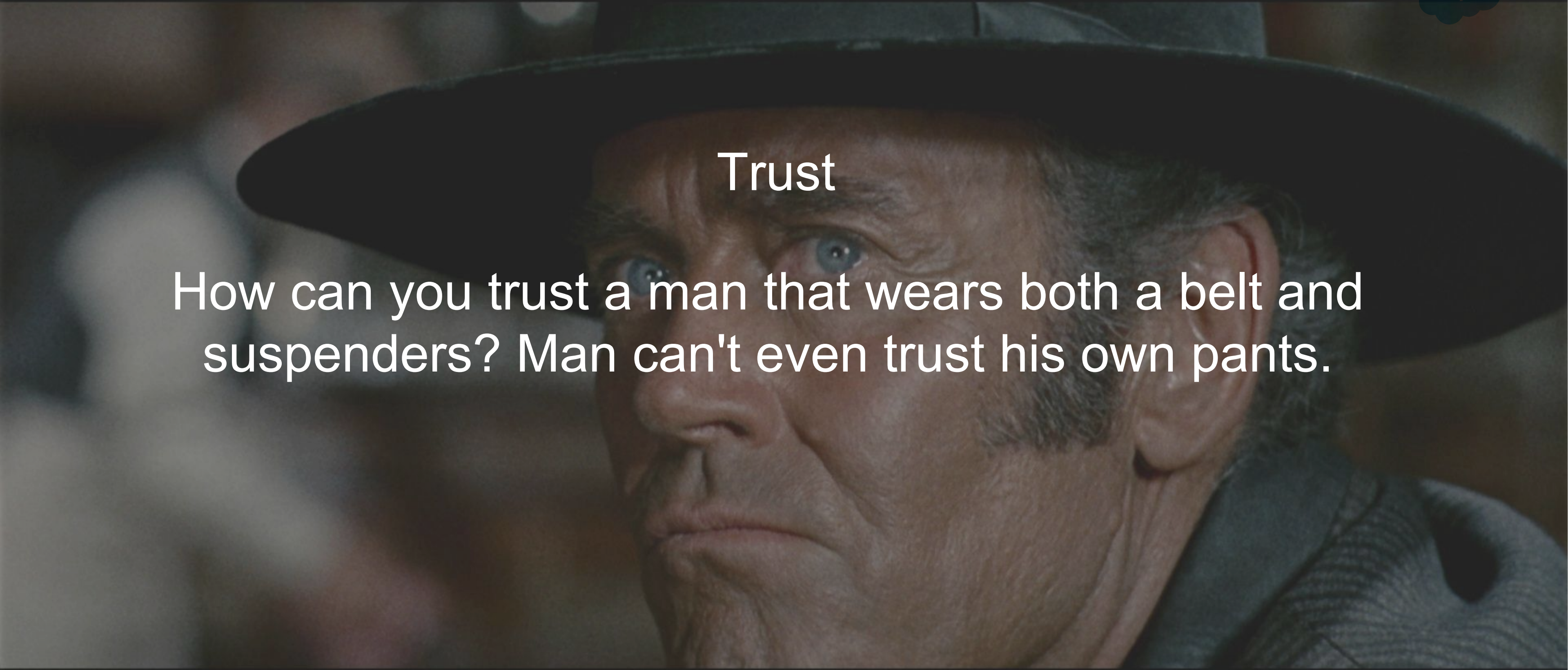
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Actionable



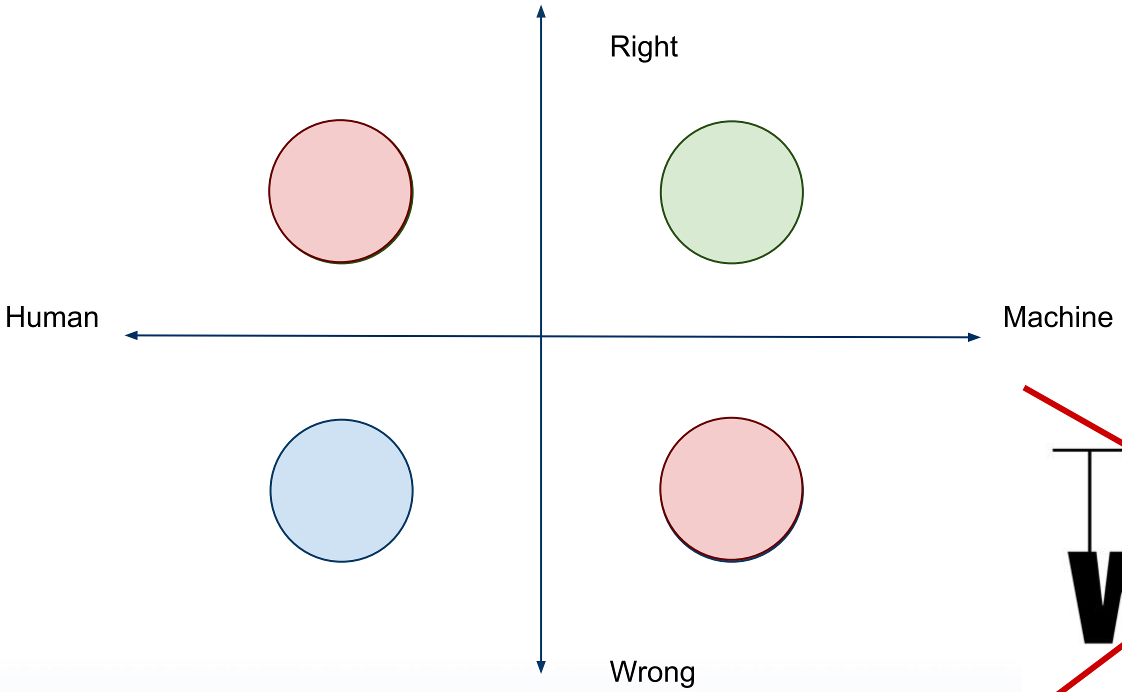
Search Salesforce
Cindy Cent...
Kell Chanc...
Bettina Sw...
Sarah Parker
Guy Bohan
Riley Sh...
Lightning I...
Working
Archived
Converted
Status: Working
Mark Status as C
We found no potential duplicates of this lead.
ACTIVITY
CHATTER
DETAILS
Log a Call
Email
New Task
New Event
Subject
Call
Filter Timeline
Next Steps
Follow-up Call with Ron
Meeting to discuss the deal details after he has had a chance to talk to leadership
Einstein
92
Lead Score
TOP PREDICTIVE FACTORS
Phone Number is Valid
Title is Director
Downloaded White Paper
Interest in Cloud Managemer
Incomplete Free Trial For
4 days 5 hrs ago

Version A
Choose
Submit
CONTACT US
Version B
Choose
Download Now



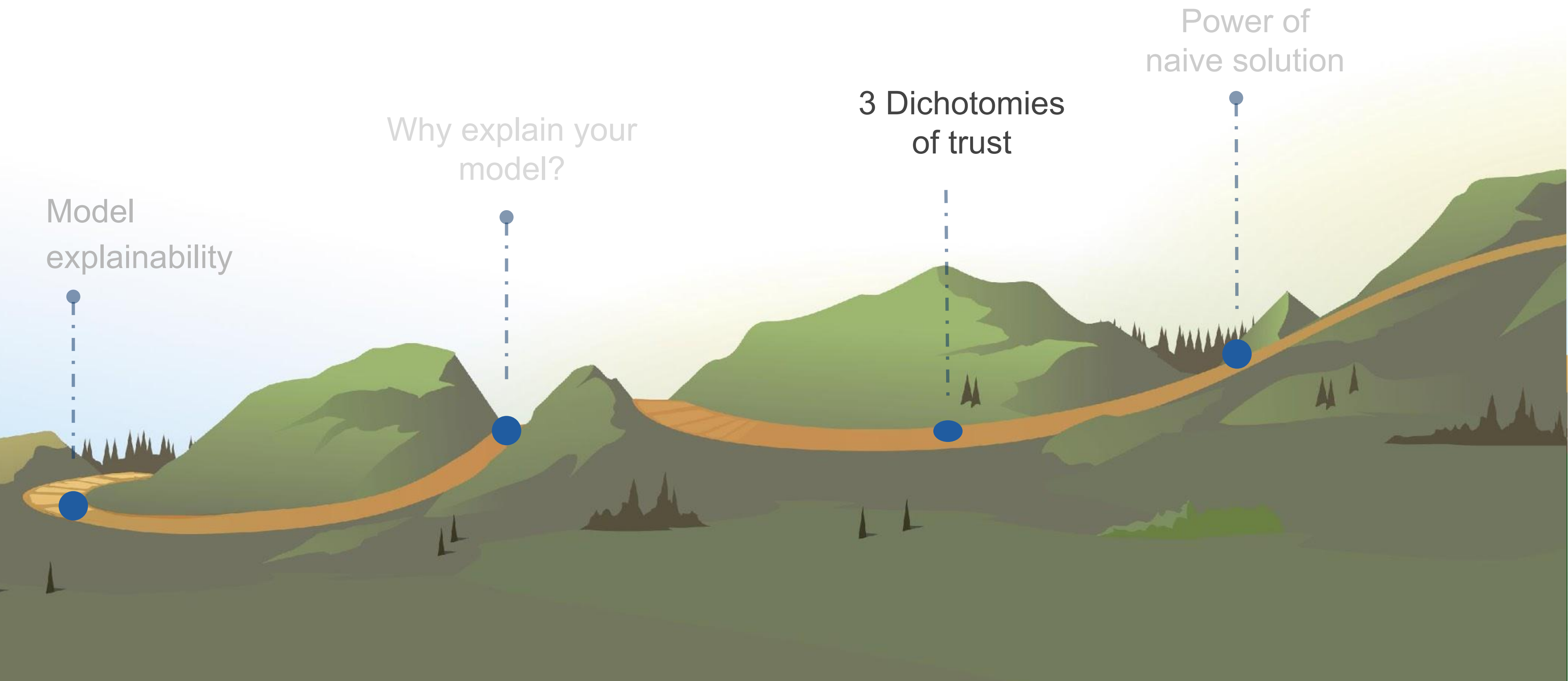
Trust

How can you trust a man that wears both a belt and suspenders? Man can't even trust his own pants.


Right
Human
Machine
Wrong


TRUST
WORTHY

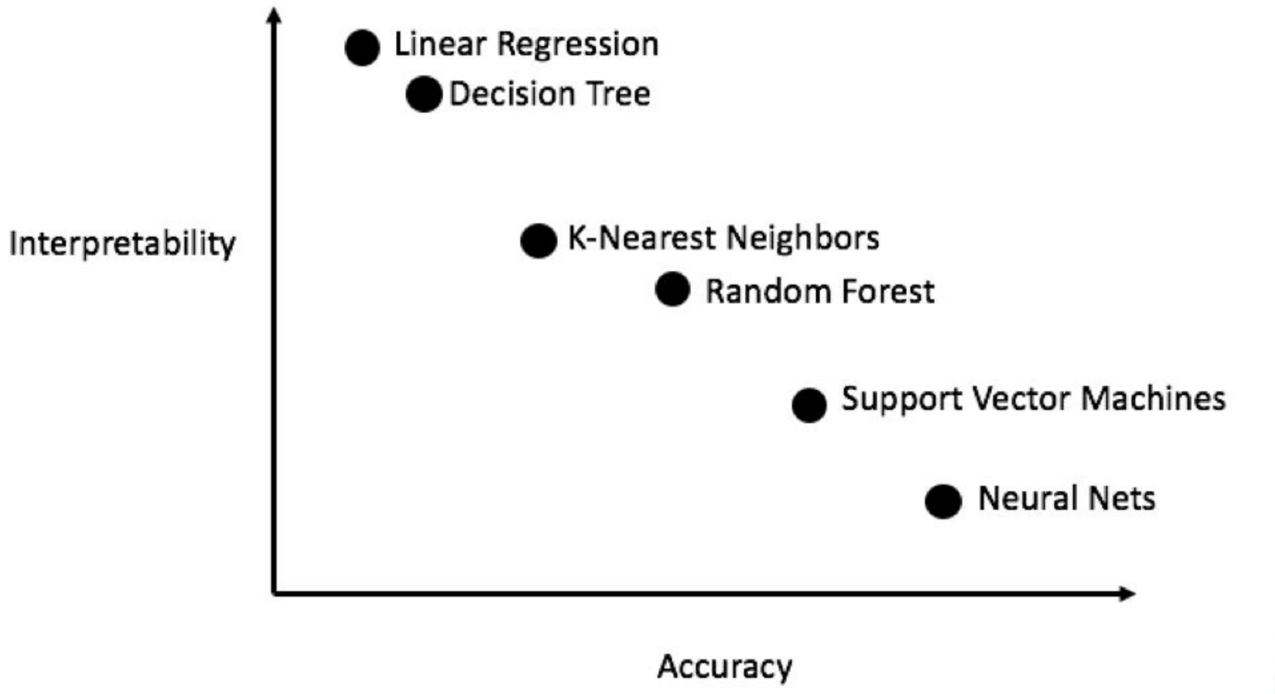
Roadmap for this talk



The 3 Dichotomies of Trust

Explainability	Accuracy
Global Explanations	Local Explanations
Model Aware	Model Agnostic

1. Explainability vs Accuracy



Example of Text Feature Engineering

Representing the word **overfitting** using various feature representations:

Morphological = [(prefix, **over-**), (root, **fit**), (suffix=imperfect tense, **-ing**)]

Unigrams = ['o', 'v', 'e', 'r', 'f', 'i', 't', 't', 'i', 'n', 'g']

Bigrams = ['ov', 've', 'er', 'rf', 'fi', 'it', 'tt', 'ti', 'in', 'ng']

Trigrams = ['ove', 'ver', 'erf', 'rfi', 'fit', 'itt', 'tti', 'tin', 'ing']

One-hot = [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Word vector = [-0.26, 0.34, 0.48, -0.06, 0.16, 0.11, 0.13, -0.15, 0.47, -0.49, 0.07, -0.39, -0.13, -0.15, 0.06, 0.09]

http://smerity.com/articles/2016/architectures_are_the_new_feature_engineering.html

2. Global vs Local Explanations

Predict House Price

Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

E.g. $h_\theta(x) = 80 + 0.1 x_1 + 0.01 x_2 + 3 x_3 - 2 x_4$

age

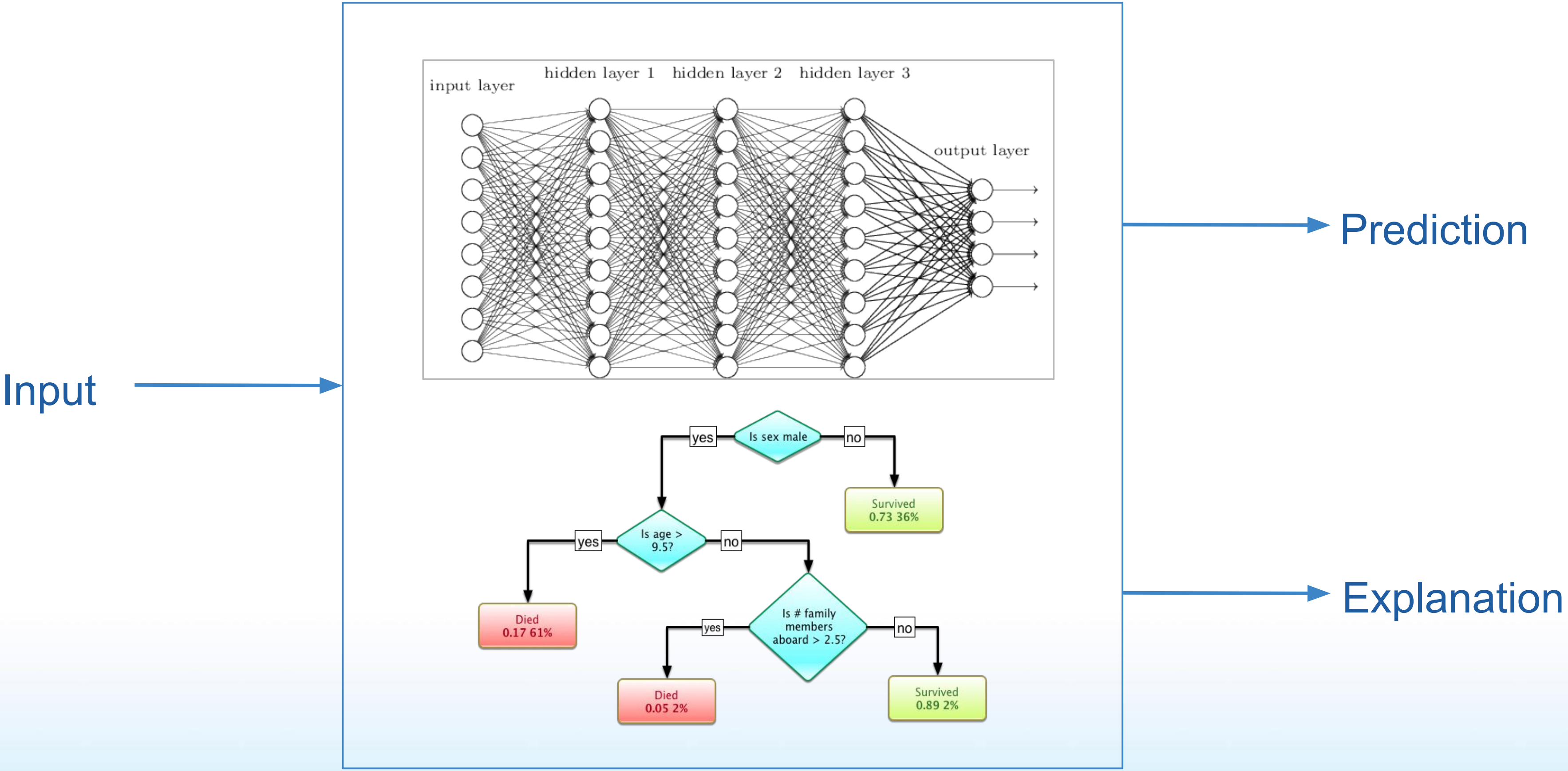
Predict House Price

Size (feet^2) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

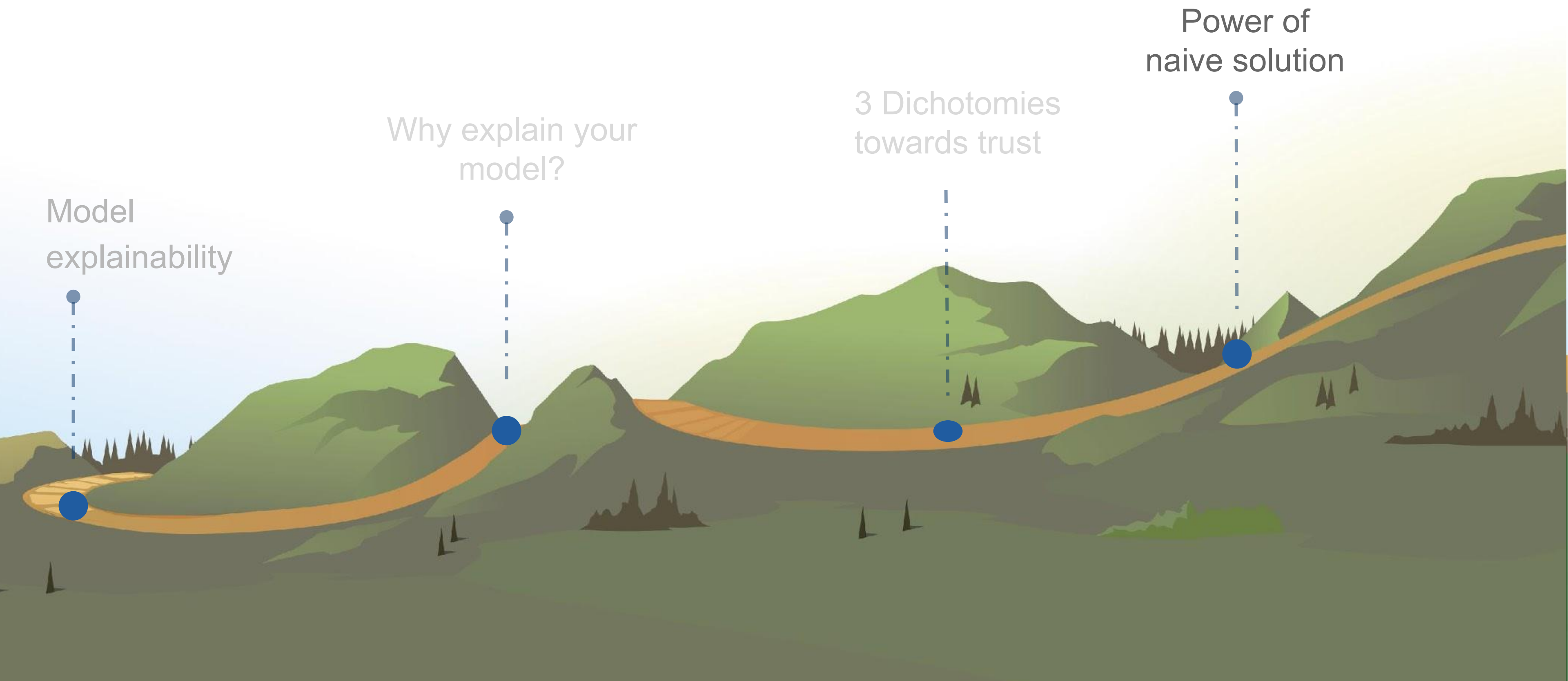
$$\text{E.g. } h_\theta(x) = 80 + 0.1 \underset{852}{x_1} + 0.01 \underset{2}{x_2} + 3 \underset{1}{x_3} - 2 \underset{36}{x_4}$$

3. Model Aware vs Model Agnostic



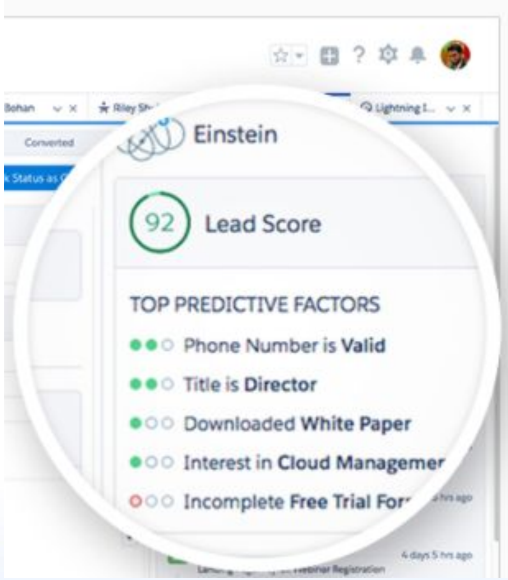
Model Agnostic Explanations

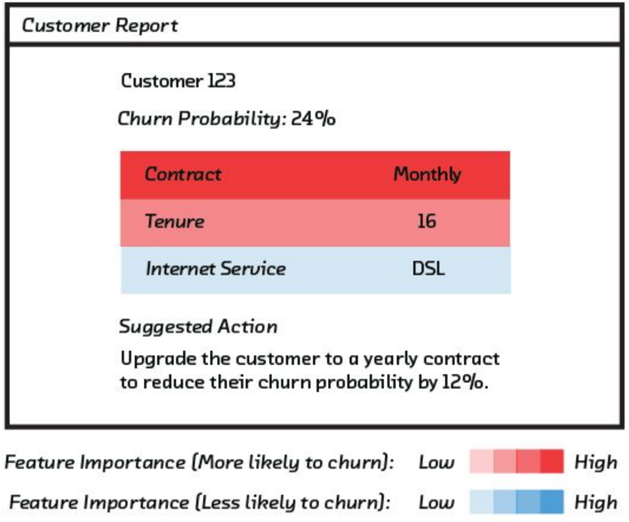
Roadmap for this talk

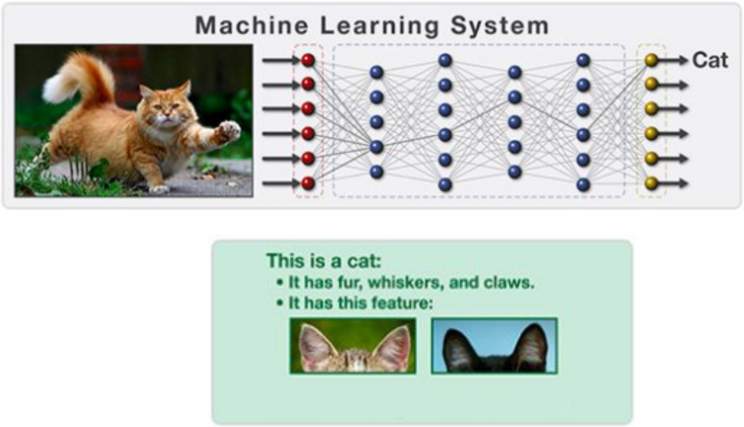


Meet 80% of Customer Needs with Naive Solution

> *The source of this lead is an inbound call and leads with such source generally have a high chance of converting*







AI

Salesforce open-sources TransmogrifAI, the machine learning library that powers Einstein

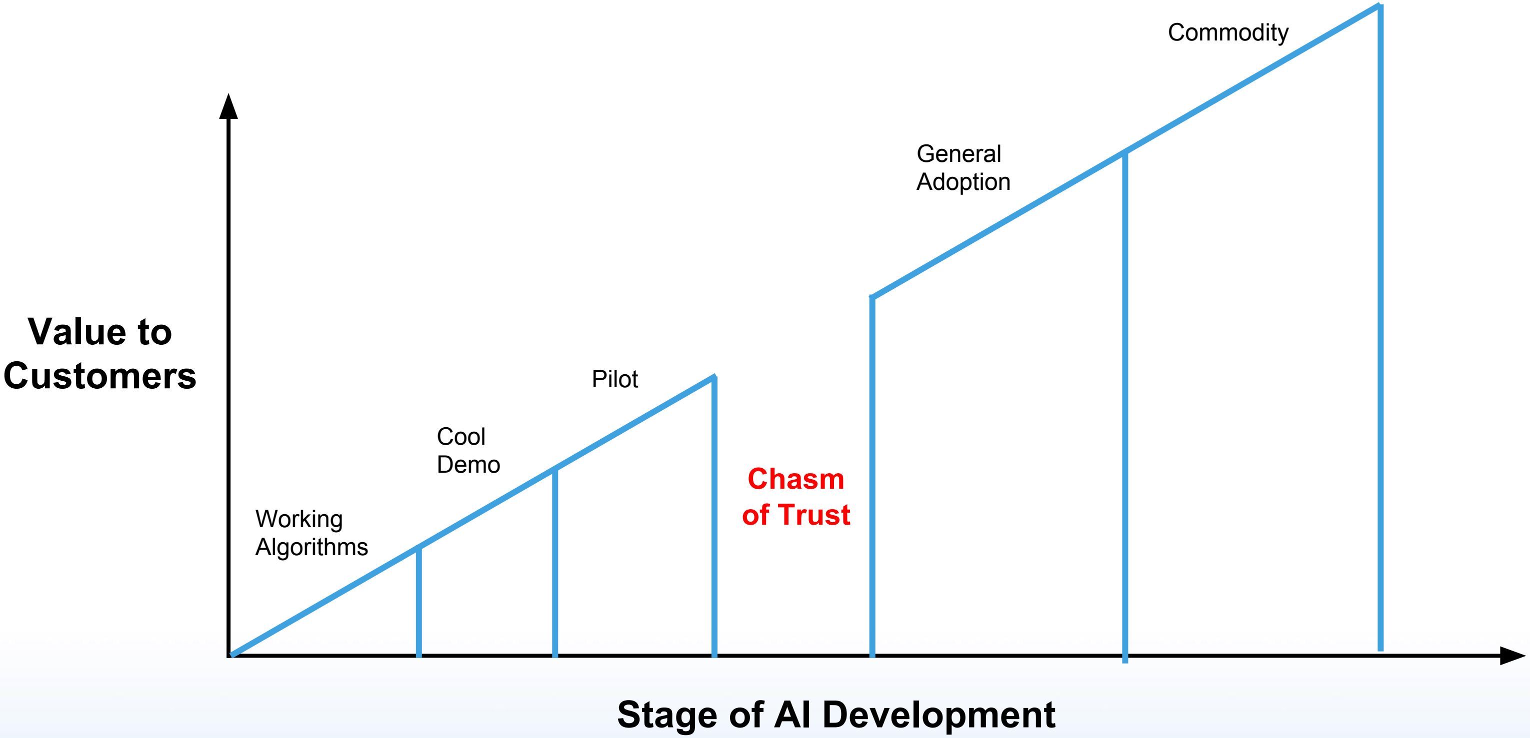
KYLE WIGGERS @KYLE_L_WIGGERS AUGUST 16, 2018 6:00 AM

Above: Salesforce Tower in Indianapolis is the company's largest hub outside its global headquarters in San Francisco. It opened on May 20, 2016.

Image Credit: Salesforce

Machine learning models — artificial intelligence (AI) that identifies relationships among hundreds, thousands, or even millions of data points — are rarely easy to architect. Data scientists spend weeks and months not only preprocessing the data on which the models are to be trained, but extracting useful features (i.e., the data types) from that data, narrowing down algorithms, and ultimately building (or attempting to build) a system that performs well not just within the

Lack of Trust is a Barrier to Adoption



Thank You!

Twitter: @mayukhb
Linkedin: linkedin.com/in/mayukhbhaowal
Email: mbhaowal@salesforce.com