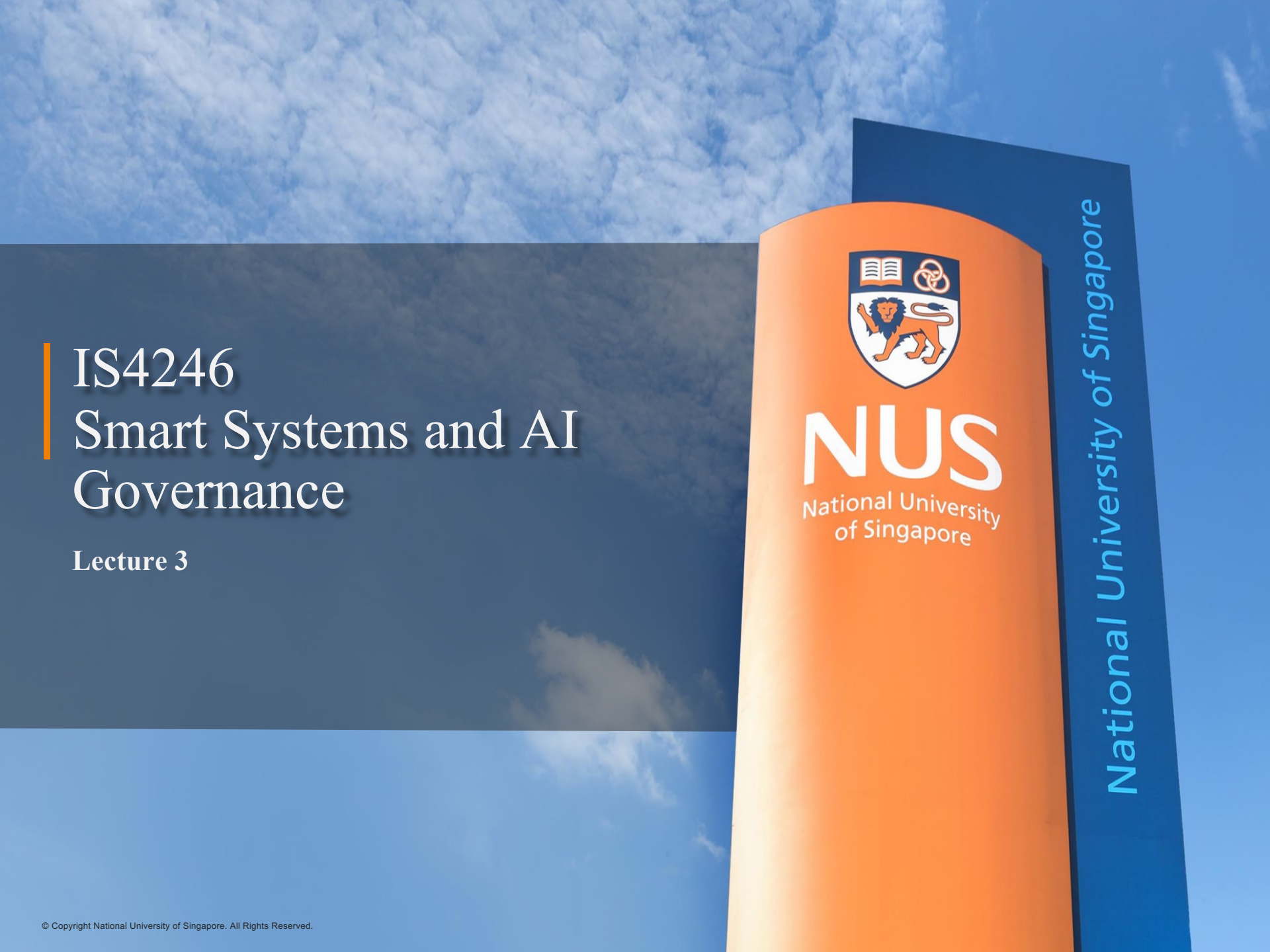
IS4246
Smart Systems and AI Governance

Lecture 3



Agenda

- Review From Last Time
- Trusting AI
- Lab: Case Study #1

Review from Last Time

Questions You Had From Last Week

-

Agenda

- Review from Last Week
- Trusting AI
  - **Augmented Analytics**
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Components of Trust
  - Explainability
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door

Lecture 3 Learning Objectives

Explain the importance of building trust between user and AI.

Identify best practices for creating explanability and control when designing AI systems.

Augmented Analytics

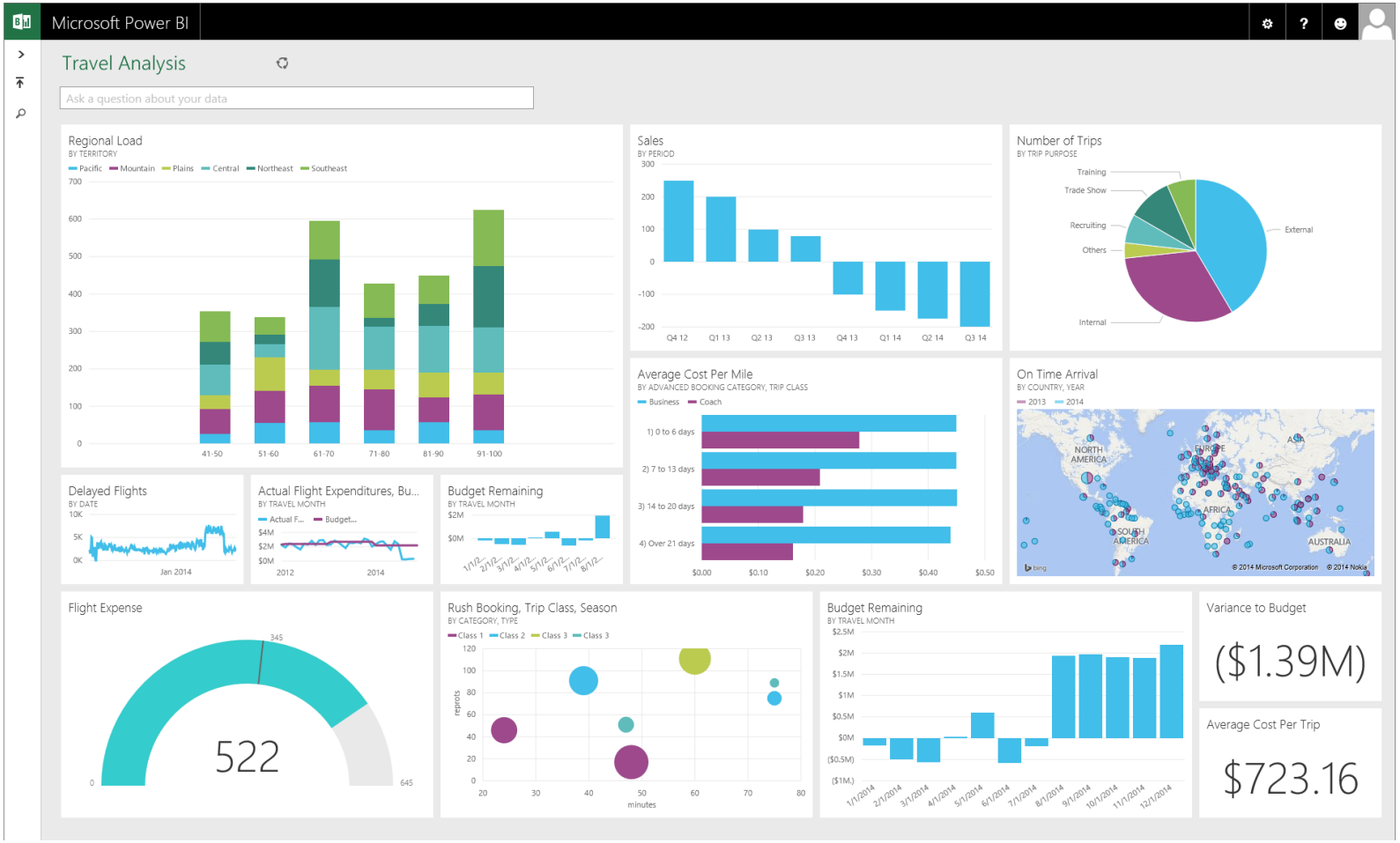
- **Augmented Analytics (AA)** is an advanced method for transforming data into insights for decision-making
  - It is a combination of Business Intelligence (BI) and the advanced features of Artificial Intelligence (AI).
  - Leveraging the full potential of AA makes analytics more accessible to non-coders and reduces time to insights
  - AA enhances analysis, data preparation, visualization, modelling, and generation of insights
  - Human interaction and perspectives are still necessary; decision-makers play an important role in operationalizing findings

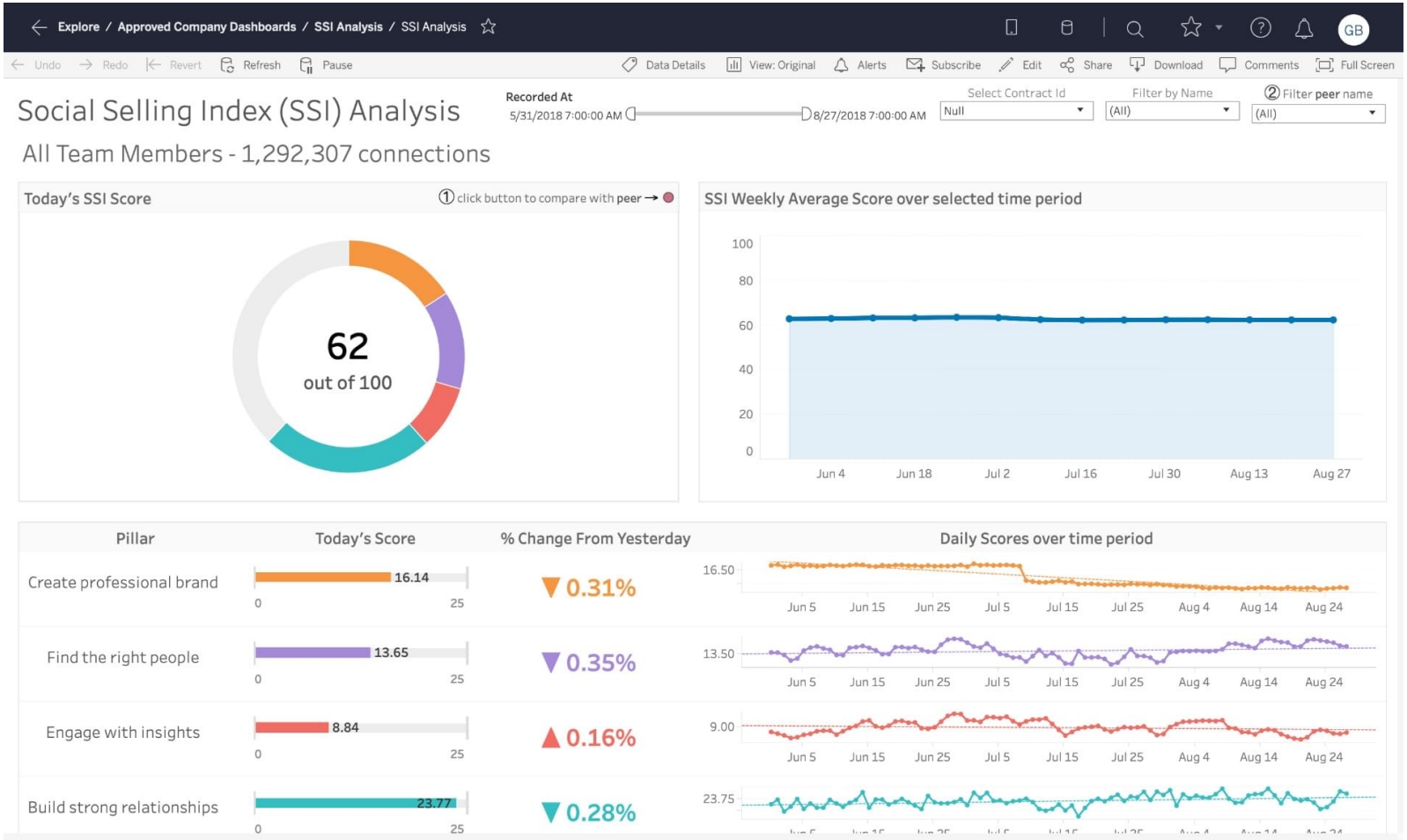
Alghamdi, N. A., & Al-Baity, H. H. (2022). Augmented Analytics Driven by AI: A Digital Transformation beyond Business Intelligence. *Sensors*, *22*(20), 8071. Chicago

Augmented Analytics

Tool	Description
Qlik	Qlik's Sense is a high-performance tool that allows users with different analytical levels to search and analyze any dataset. Qlik's AA component, Insight Advisor, facilitates data exploration by automatically generating insights based on data analysis, which automates and speeds up the data preparation process. Its search-based visual analysis displays hidden insights as powerful visuals that can be modified and adjusted to create effective dashboards. Furthermore, NLP is used in conversational analytics, which allows users to evaluate data in a conversational manner [5].
Power BI	Power BI allows analysts to perform data preparation, data discovery, and building of dashboards using similar design techniques. The platform works with Excel and Office 365 and consists of an active user community that builds the tool's potential. Power BI's analytical capabilities are enhanced by the availability of powerful AA capabilities and ML algorithms. Furthermore, features such as Quick Insights and Q&A visualizations allow users to easily examine and interpret data. Other elements, such as text analytics and visual analytics, enable customers to successfully employ the analytics capabilities in their data analysis [20].
Tableau	Tableau fully integrates Einstein analytics to leverage AI technologies to examine and analyze data in order to make predictions and recommendations based on those findings. The presence of features such as Ask Data and Explain Data demonstrates that the industry is moving beyond traditional visualization-based solutions. In addition, Tableau uses smart analytics tools such as NLP and NLG to give customers a better data analysis experience [14].
ThoughtSpot	ThoughtSpot is a BI and analytics company known for its highly scalable and relational analytics search engine, which allows business users to interact easily with data. It is considered one of the first BI suppliers to deliver AI-generated insights throughout the user experience, from a smart homepage to search, dashboards, and datasets. It has a user-friendly interface for providing automated insights and allows users to ask questions and execute queries [21].

Augmented Analytics




Explore / Approved Company Dashboards / SSI Analysis / SSI Analysis
GB
Undo
Redo
Revert
Refresh
Pause
Data Details
View: Original
Alerts
Subscribe
Edit
Share
Download
Comments
Full Screen
Social Selling Index (SSI) Analysis
Recorded At
5/31/2018 7:00:00 AM
8/27/2018 7:00:00 AM
Select Contract Id
Null
Filter by Name
(All)
Filter peer name
(All)
All Team Members - 1,292,307 connections
Today's SSI Score
click button to compare with peer
62
out of 100
SSI Weekly Average Score over selected time period
100
80
60
40
20
0
Jun 4
Jun 18
Jul 2
Jul 16
Jul 30
Aug 13
Aug 27
Pillar
Today's Score
% Change From Yesterday
Daily Scores over time period
Create professional brand
16.14
0
25
▼ 0.31%
16.50
Find the right people
13.65
0
25
▼ 0.35%
13.50
Engage with insights
8.84
0
25
▲ 0.16%
9.00
Build strong relationships
23.77
0
25
▼ 0.28%
23.75
Jun 5
Jun 15
Jun 25
Jul 5
Jul 15
Jul 25
Aug 4
Aug 14
Aug 24

- Power BI uses the existing Microsoft systems like Azure, SQL, and Excel to build data visualizations that don't break the bank.

- Tableau specializes in making beautiful visualizations, but much of its advertising is focused on corporate environments with data engineers and bigger budgets.

https://technologyadvice.com/blog/information-technology/power-bi-vs-tableau/

Augmented Analytics

- **Augmented Analytics (AA)** is an advanced method for transforming data into insights for decision-making
  - It is a combination of Business Intelligence (BI) and the advanced features of Artificial Intelligence (AI).
  - Leveraging the full potential of AA makes analytics more accessible to non-coders and reduces time to insights
  - AA enhances analysis, data preparation, visualization, modelling, and generation of insights
  - **Human interaction and perspectives are still necessary; decision-makers play an important role in operationalizing findings**

Agenda

- Review from Last Week
- Trusting AI
  - Augmented Analytics
  - **Factors of Fairness**
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Components of Trust
  - Explainability
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door

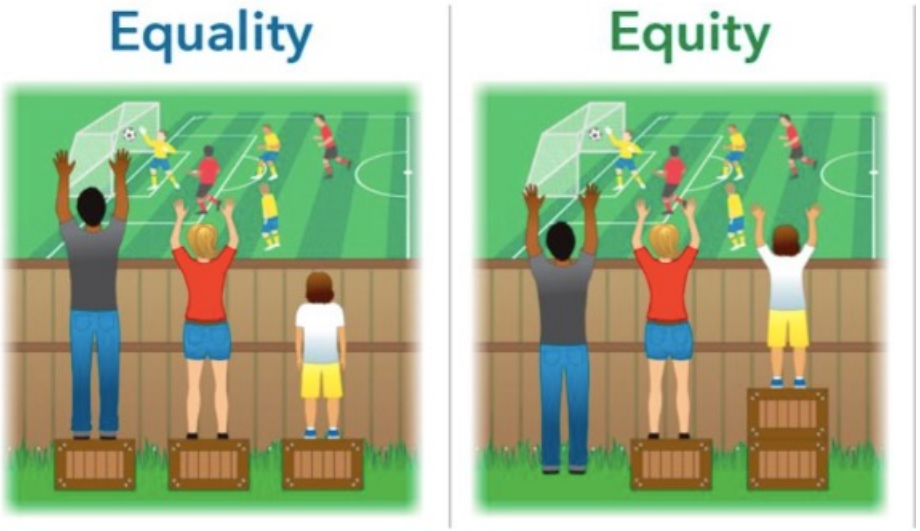
Factors in Fairness

- Distributive – Is the *outcome* fair?
- Procedural – Was the decision *process* fair?
- Interactional – Was the *interaction* with me fair?

Distributive Justice

- Equity Theory: outcomes proportional to inputs

$$\frac{O_1}{I_1} = \frac{O_2}{I_2}$$



Get ready – I will ask you for your own examples in a few minutes

Distributive Justice

- Examples of outcomes
  - College Admissions

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications

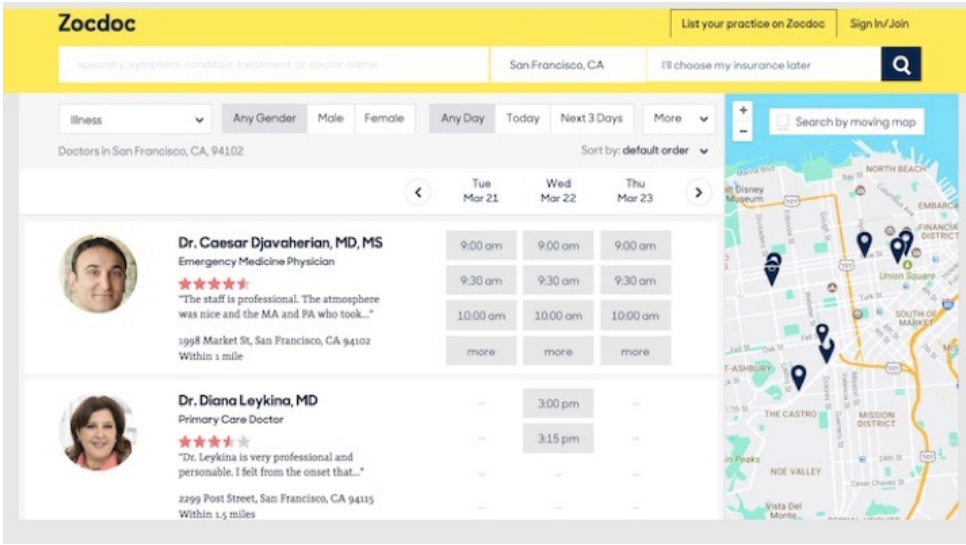
17

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications

Distributive Justice

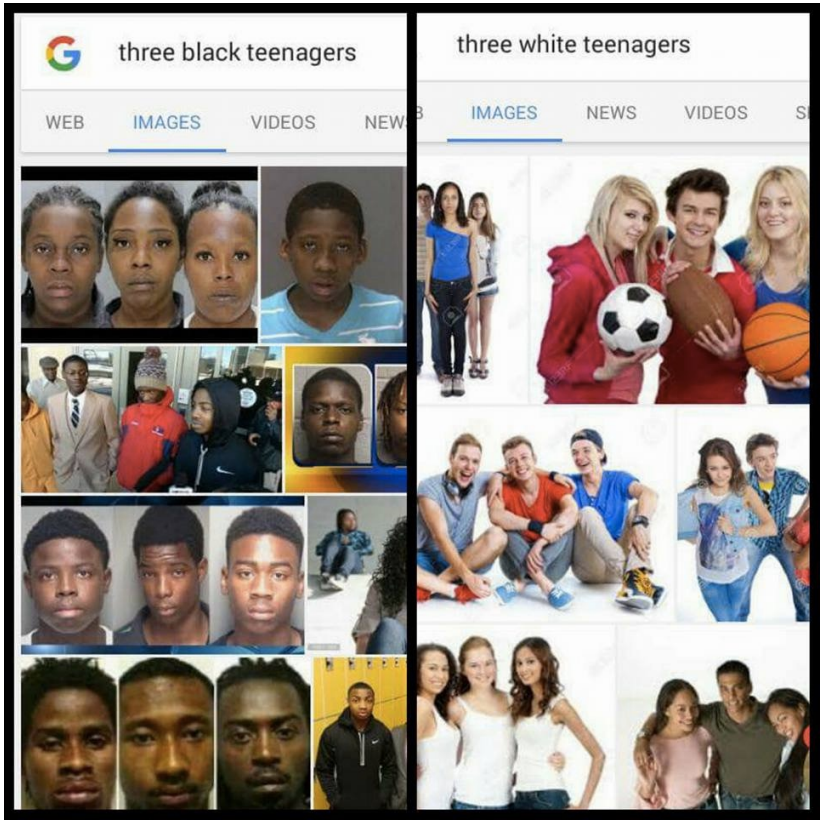
- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings



19

Distributive Justice

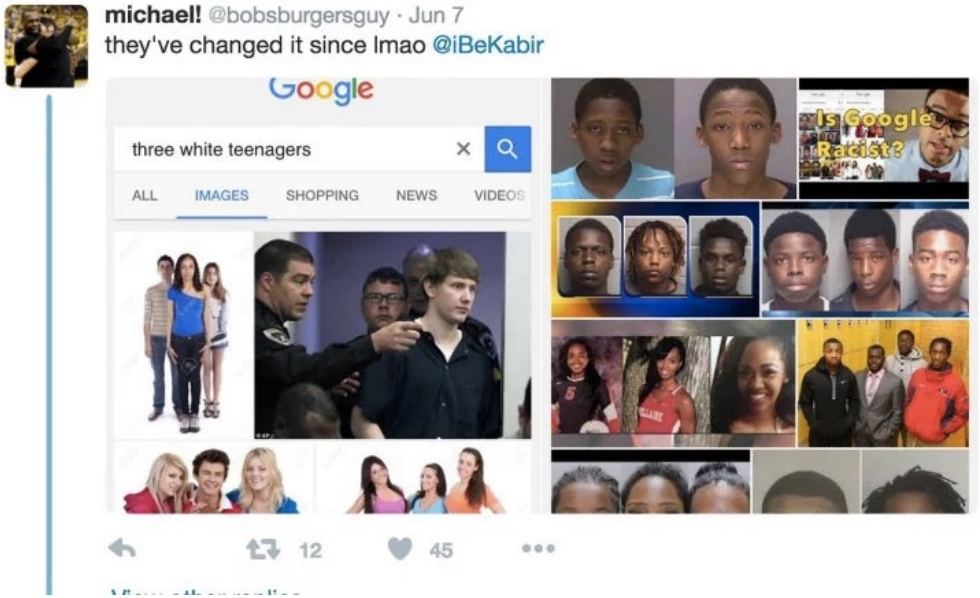
- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results



Note: **What** is being distributed is not always clear cut. In this case one could say “respect” for the different groups is being distributed.

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results

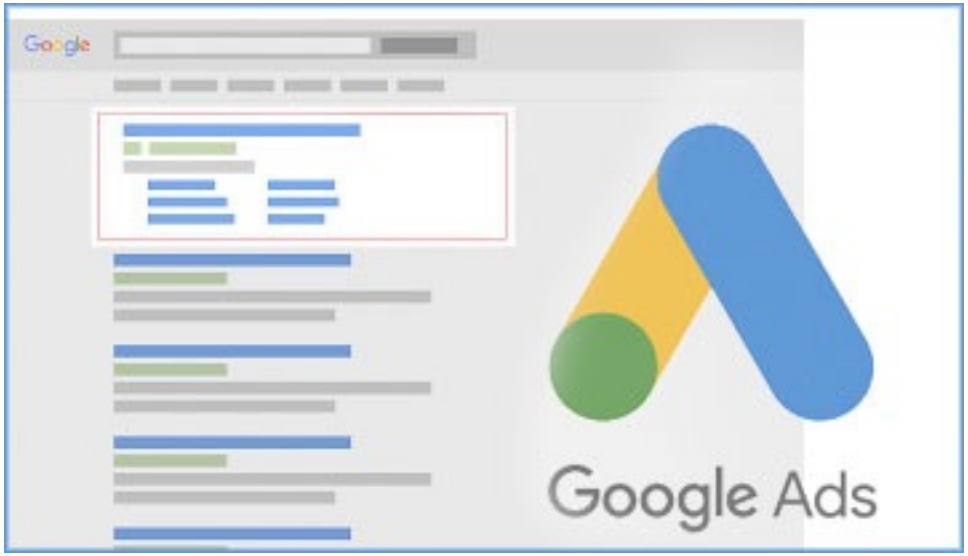


Note: **What** is being distributed is not always clear cut. In this case one could say “respect” for the different groups is being distributed.

21

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results
  - Accompanying ads



22

Ad Auctions

Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Anja Lambrecht,[a] Catherine Tucker[b]

[a] Marketing, London Business School, London NW1 4SA, United Kingdom; [b] Marketing, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142

Contact: alambrecht@london.edu, http://orcid.org/0000-0001-6766-1602 (AL); cetucker@mit.edu, http://orcid.org/0000-0002-1847-4832 (CT)

Received: November 28, 2017
Revised: March 2, 2018
Accepted: March 13, 2018
Published Online in Articles in Advance: April 10, 2019

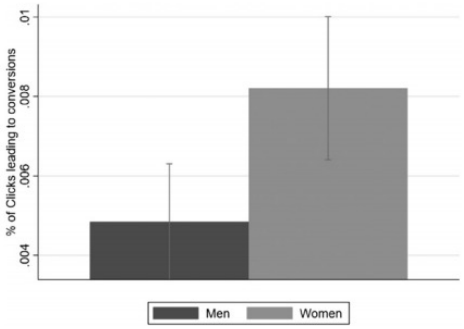
https://doi.org/10.1287/mnsc.2018.3093

Abstract. We explore data from a field test of how an algorithm delivered ads promoting job opportunities in the science, technology, engineering and math fields. This ad was explicitly intended to be gender neutral in its delivery. Empirically, however, fewer women saw the ad than men. This happened because younger women are a prized demographic and are more expensive to show ads to. An algorithm that simply optimizes cost-effectiveness in ad delivery will deliver ads that were intended to be gender neutral in an apparently discriminatory way, because of crowding out. We show that this empirical regularity extends to other major digital platforms.

History: Accepted by Joshua Gans, business strategy.
Funding: Supported by a National Science Foundation Career Award [

Figure 2. Ad Targeting Settings: Ad Intended to Be Shown to Both Men and Women Aged 18+ Years

Figure 3. Women Are More Likely than Men to Convert After Clicking



23

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results
  - Accompanying ads
  - Pay

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results
  - Accompanying ads
  - Pay
  - Authority & Status

25

Distributive Justice

- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results
  - Accompanying ads
  - Pay
  - Authority & Status
  - Pricing

Distributive Justice

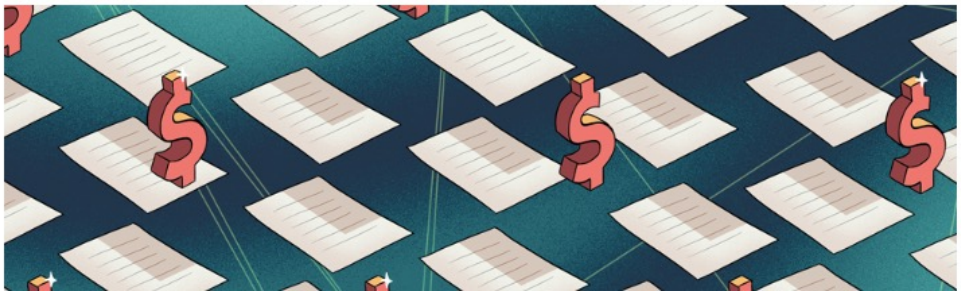
- Examples of outcomes
  - College Admissions
  - Job Applications
  - Loan Applications
  - Search Rankings
  - Search Results
  - Accompanying ads
  - Pay
  - Authority & Status
  - Pricing
  - What else…?

27

BUSINESS • STARTUPS • VENTURE

HR Tech Startup TalentGuard Snags $4M To Help Companies Keep Employees Longer

Mary Ann Azevedo July 31, 2019



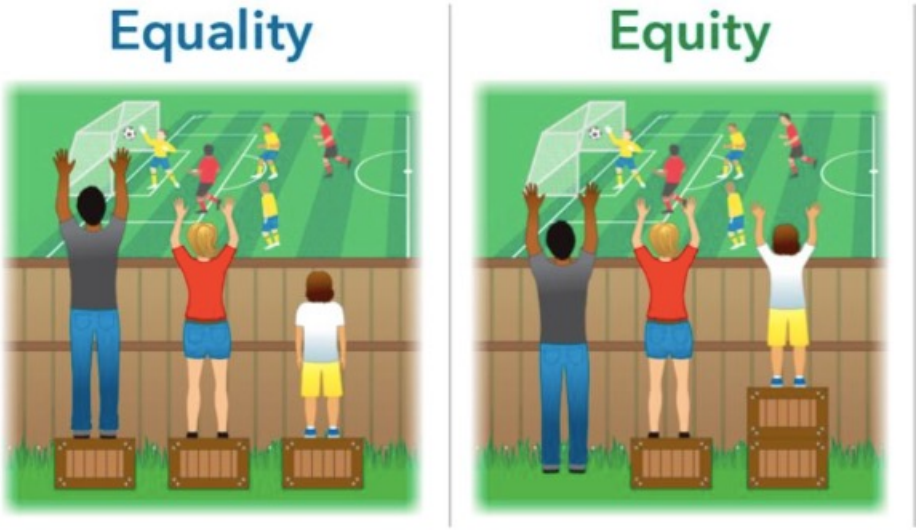
The company also plans to use the money to add more artificial intelligence and machine learning bots into its software "to make it smarter," Ginac said, especially in the area of predictive people development. What does that mean?

"We want to be able to take lots of information and predict a targeted skill development plan," she explained."The goal is to give employees a way to develop a curated running development plan that they can execute on and thus help increase the chance of them staying on longer at a company."

Distributive Justice

- Equity Theory: outcomes proportional to inputs
  - *Which* inputs should be counted*?*
  - *How* do you count inputs?

$$\frac{O_1}{I_1} = \frac{O_2}{I_2}$$

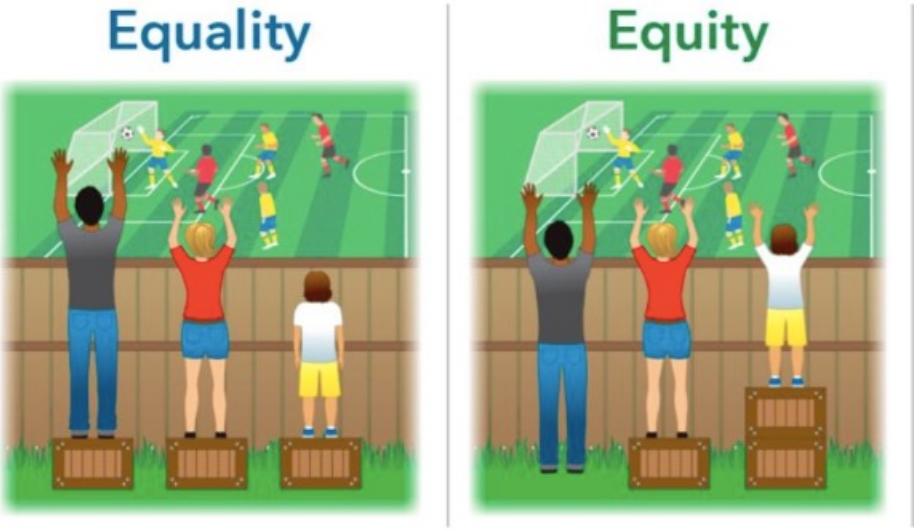


Distributive Justice

- Equity Theory: outcomes proportional to inputs
  - *Which* inputs should be counted?
  - *How* do you count inputs?

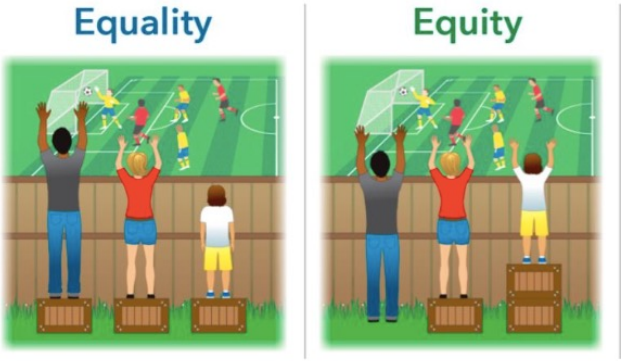
Procedural Justice

$$\frac{O_1}{I_1} = \frac{O_2}{I_2}$$



Distributive Justice

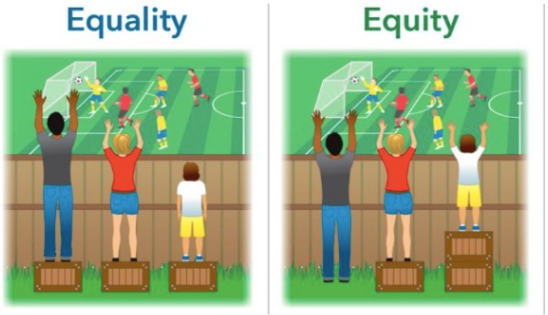
- Three (incommensurate) Allocation Rules
  - **Equality**: equal distribution (of resources or outcomes)
  - **Equity**: outcomes proportional to inputs (e.g. talent / effort)
  - **Need**: distribution accorded to urgency



Distributive Justice

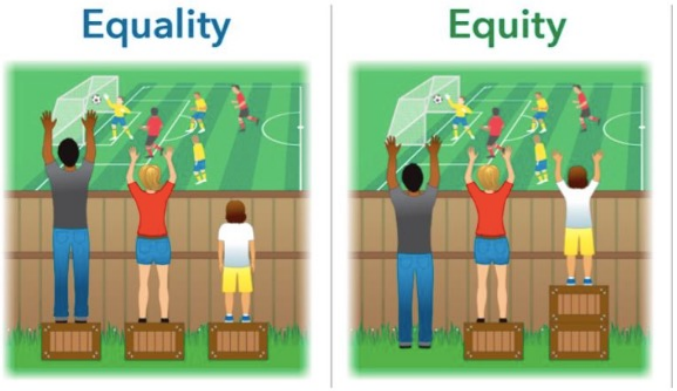
- Three (incommensurate) Allocation Rules
  - **Equality**: equal distribution (of resources or outcomes)
  - **Equity**: outcomes proportional to inputs (e.g. talent / effort)
  - **Need**: distribution accorded to urgency

Problems with the graphics?



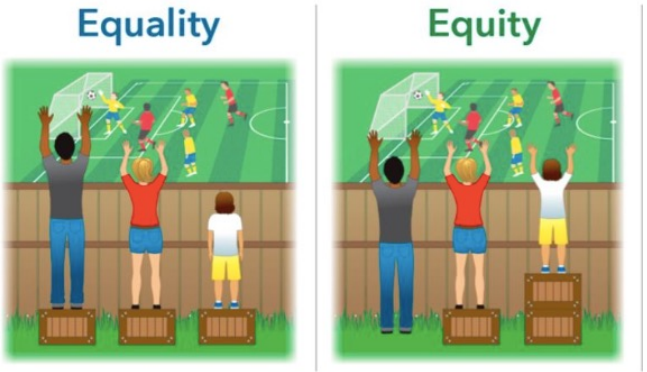
Distributive Justice

- Three (incommensurate) Allocation Rules
  - **Equality**: equal distribution (of resources or outcomes)
  - **Equity**: outcomes proportional to inputs (e.g. talent / effort)
  - **Need**: distribution accorded to urgency



Distributive Justice

- Three (incommensurate) Allocation Rules
  - **Equality**: equal distribution (of resources or outcomes)
  - **Equity**: outcomes proportional to inputs (e.g. talent / effort)
  - **Need**: distribution accorded to urgency



Distributive Justice

35

Come up and write ONE example of an outcome or decision made with an IT product or company.

Write whether it should be distributed via Need, Equality, Equity, a Combination, or “not sure”

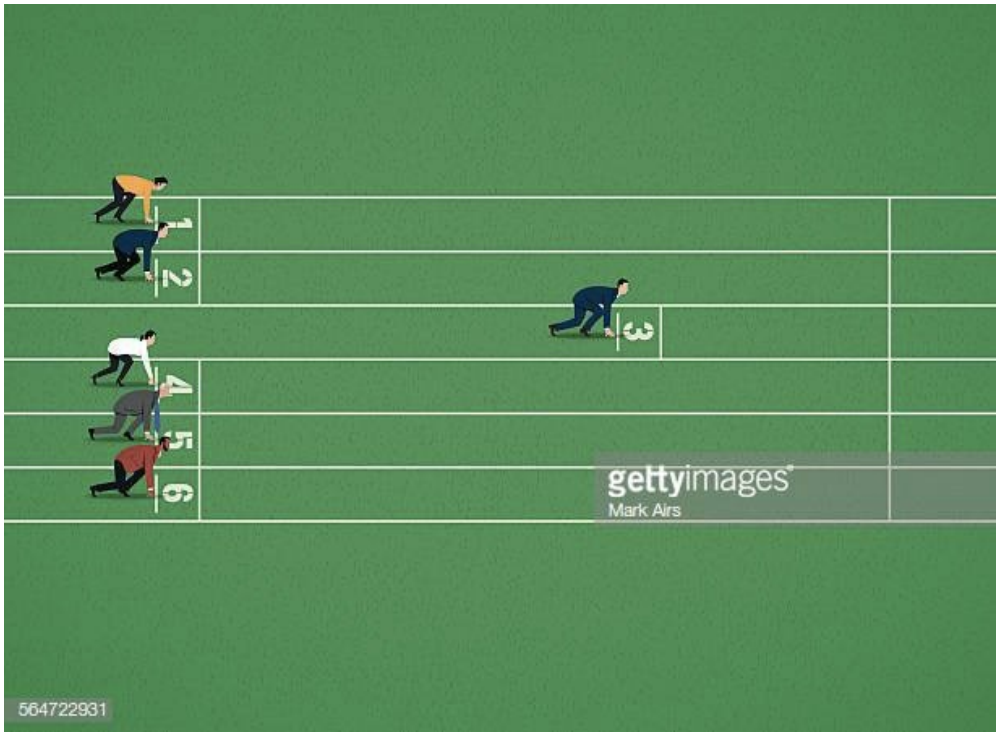
Procedural Justice

- The PROCESS by which decisions are made
  - **Consistency**: Is the same process applied to everyone?
  - **Lack of Bias**: Was there inherent bias? (against equity, equality, *and* need) Was a person or group discriminated against?
  - **Accuracy**: is correct information being used?
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?
  - **Representation of all concerned**: Appropriate stakeholders have input into the deliberation of a decision
  - **Ethics**: norms of acceptability are not violated

37

Procedural Justice

- The PROCESS by which decisions are made
  - **Consistency**: Is the same process applied to everyone?



38

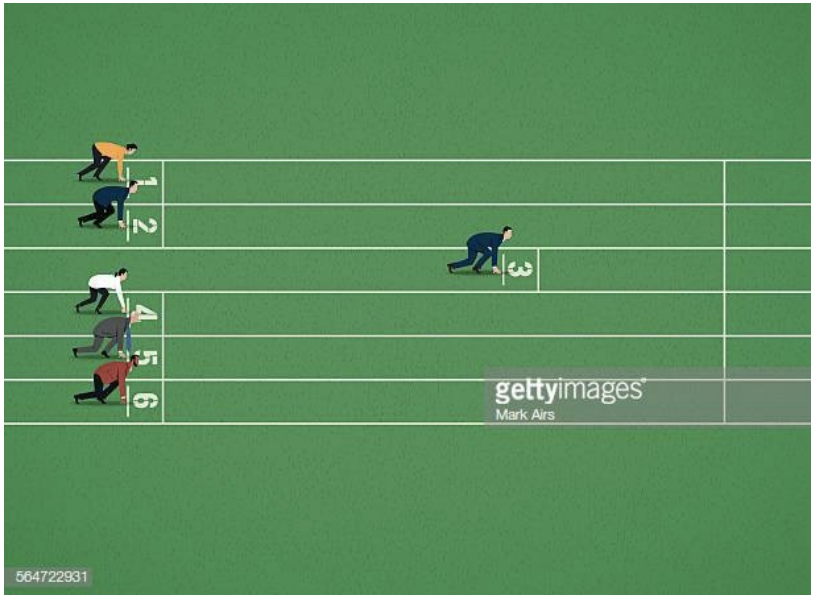
Procedural Justice

- The PROCESS by which decisions are made
  - **Consistency**: Is the same process applied to everyone?

39

Procedural Justice

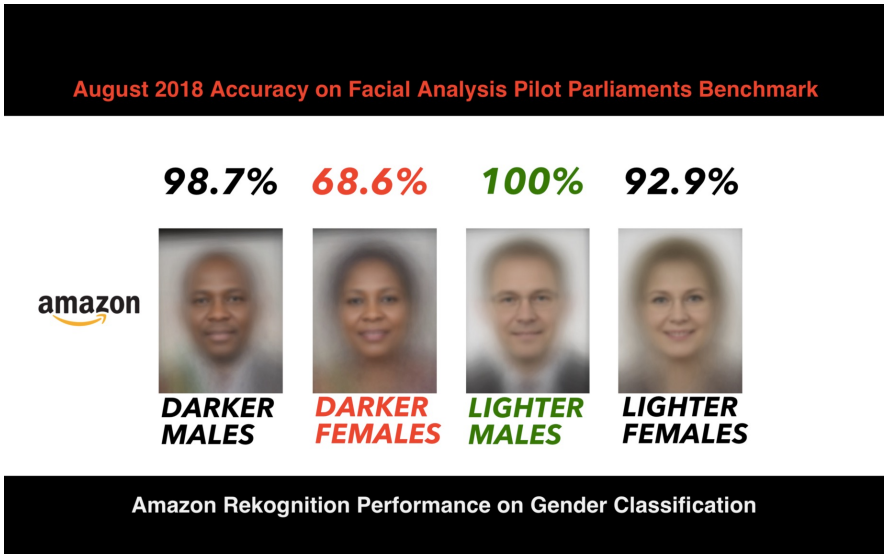
- The PROCESS by which decisions are made
  - **Consistency**: Is the same process applied to everyone?
  - How could this *accidentally* be violated?



40

Procedural Justice

- The PROCESS by which decisions are made
  - **Consistency**: Is the same process applied to everyone?
  - How could this *accidentally* be violated?



41

Procedural Justice

- The PROCESS by which decisions are made
  - **Lack of Bias**: Was there inherent bias? (against equity, equality, *and* need) Was a person or group discriminated against?

A Former Cop Describes Racist Police Quotas in New York

A former officer for the NYPD breaks down the class-action lawsuit led by minority cops against the largest police force in America.

By **John Surico**

Apr 4 2016, 12:00am

How predictive policing technology can lead to discrimination and profiling

Procedural Justice

- The PROCESS by which decisions are made
  - **Lack of Bias**: Was there inherent bias? (against equity, equality, *and* need) Was a person or group discriminated against?
  - How could this *accidentally* be violated?

Procedural Justice

- The PROCESS by which decisions are made
  - **Lack of Bias**: Was there inherent bias? (against equity, equality, *and* need) Was a person or group discriminated against?
  - How could this *accidentally* be violated?
    - Experts: Out-dated values or ideas

Procedural Justice

- The PROCESS by which decisions are made
  - **Lack of Bias**: Was there inherent bias? (against equity, equality, *and* need) Was a person or group discriminated against?
  - How could this *accidentally* be violated?
    - Experts: Out-dated values or ideas
    - Machine Learning: Incomplete Data

Clip slide

CASE #1
Boston's StreetBump Pothole App

- App that citizens download to phone
- Accelerometer and GPS detect when a car hit a pothole
- Automatically reports potholes to city
- City knows where to go to fix potholes

CASE #1
Boston's StreetBump Pothole App*

- **Voice & Diversity**
- **Discrimination**
- **Technological Solutionism**

*In our presentation of the StreetBump Pothole app we emphasize that these were ethical issues that were considered and anticipated, rather than not.

Discussion Questions

- The text was based on ***employee*** perceptions. How much of this applies to:
  - Customers
  - Suppliers
  - Third-Party Community Groups

48

Procedural Justice

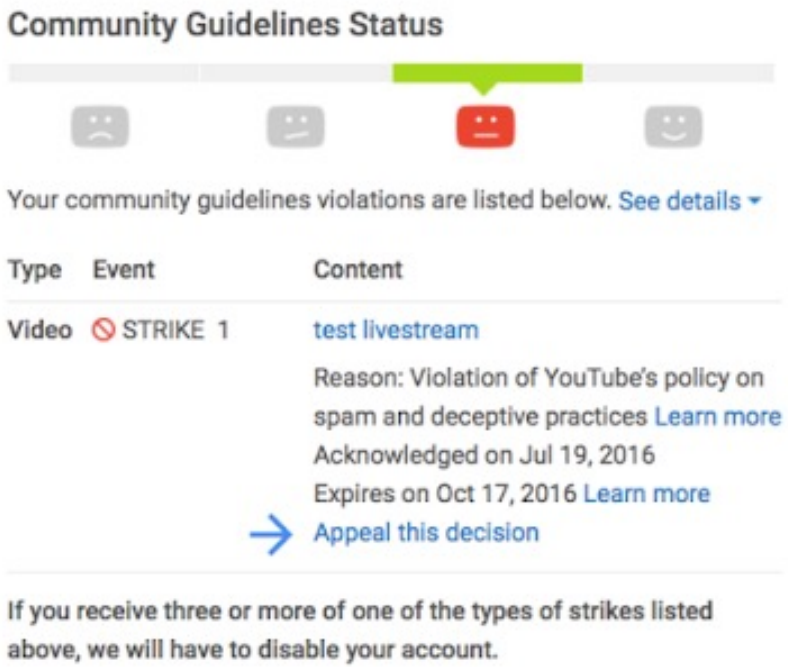
- The PROCESS by which decisions are made
  - **Accuracy**: is correct information being used?

Procedural Justice

- The PROCESS by which decisions are made
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?

Procedural Justice

- The PROCESS by which decisions are made
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?



51

Procedural Justice

- The PROCESS by which decisions are made
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?
    - People generally want to know *why* they got a bad outcome

52

Procedural Justice

- The PROCESS by which decisions are made
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?
    - People generally want to know *why* they got a bad outcome
      - "**Why**" is not always clear-cut
      - **Why =** "what is the smallest number of things I can most easily change that will have the biggest chance of getting a good outcome next time."
      - Knowing *why* might give them a chance to provide information to correct the decision.

53

Procedural Justice

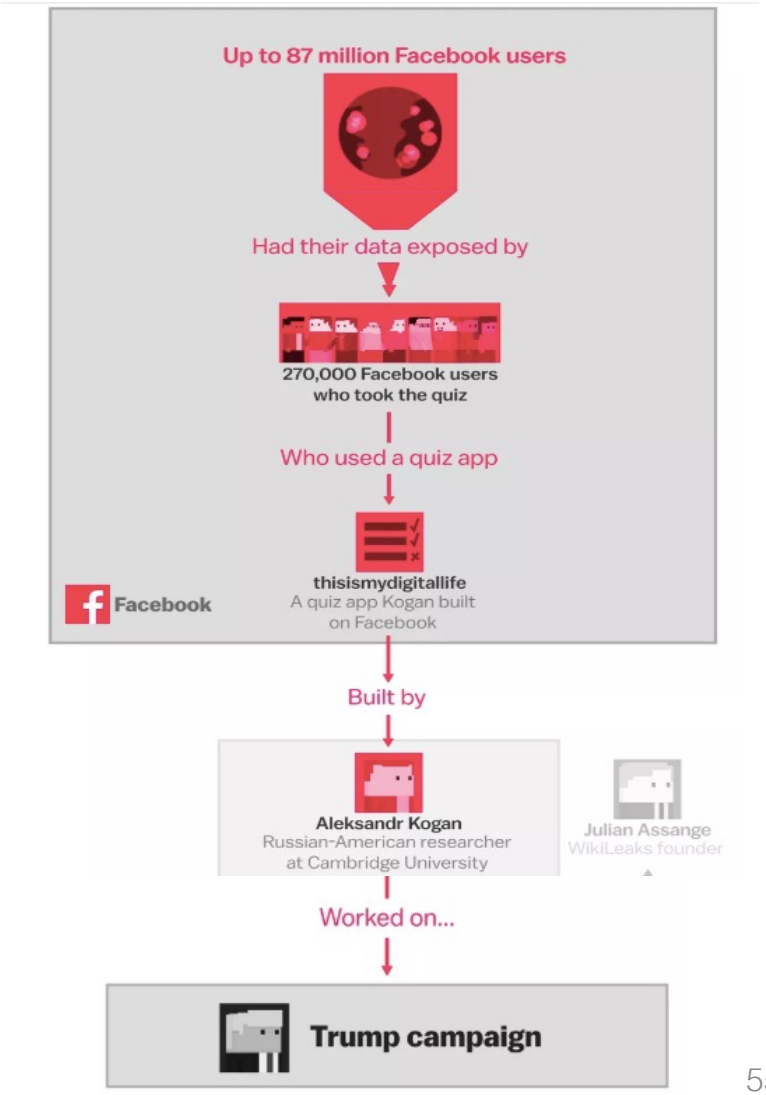
- The PROCESS by which decisions are made
  - **Correction**: Did users have visibility and input in the process to correct errors, appeal mistakes, or input additional information?
    - People generally want to know *why* they got a bad outcome
      - "**Why**" is not always clear-cut
      - **Why =** "what is the smallest number of things I can most easily change that will have the biggest chance of getting a good outcome next time."
      - Knowing *why* might give them a chance to provide information to correct the decision.

Having interpretable decision processes and flexible input procedures into the process can be expensive and difficult.

54

Procedural Justice

- The PROCESS by which decisions are made
  - **Representation of all concerned**: Appropriate stakeholders have input into the deliberation of a decision



55

Procedural Justice

- The PROCESS by which decisions are made
  - **Ethics**: norms of acceptability are not violated

Procedural Justice

- The PROCESS by which decisions are made
  - **Ethics**: norms of acceptability are not violated

57

Procedural Justice

- The PROCESS by which decisions are made
  - **Ethics**: norms of acceptability are not violated

58

Interactional Justice

- **Informational Justice**: Sharing Relevant Info
- **Interpersonal Justice**: Person is treated with dignity, courtesy, and respect

59

Interpersonal Justice

- Person is treated with dignity, courtesy, and respect
  - As if they “mattered”
    - Give full attention
    - Let them talk to someone with power to do something
    - Having “memory” of past interactions
    - Give positive feedback for good behavior or accomplishments
    - Have “Fun” – Allow a little humor – recognize which errors are serious and which are minor
    - Go the extra mile through small, unexpected actions

Informational Justice

- Provide adequate justifications and **explanations** when things go badly
- Provide *correct* information
- Help the person what to do next, or do next time

61

Informational Justice

- Provide adequate justifications and **explanations** when things go badly
- Provide *correct* information
- Help the person what to do next, or do next time

Explainability Is the Key to Trust in AI

They say that familiarity breeds contempt. However, in an AI context, it's more likely to breed trust and acceptance. When voice recognition systems were first developed, it seemed like science fiction. Nowadays, our children are unfazed by talking to smart gadgets and not in the least freaked out by the fact that the gadget can correctly identify them.

62

Informational Justice

- Provide adequate justifications and **explanations** when things go badly
- Provide *correct* information
- Help the person what to do next. or do next time

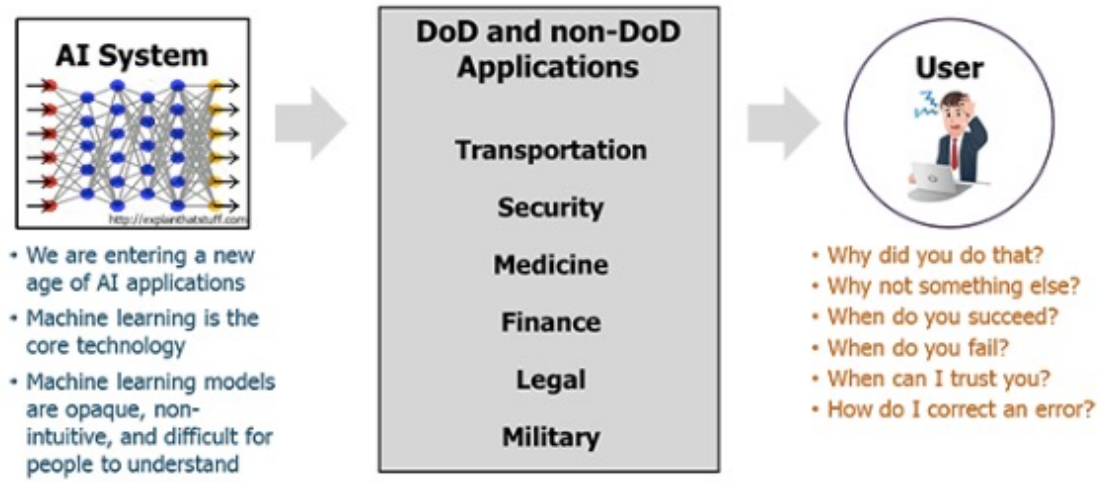
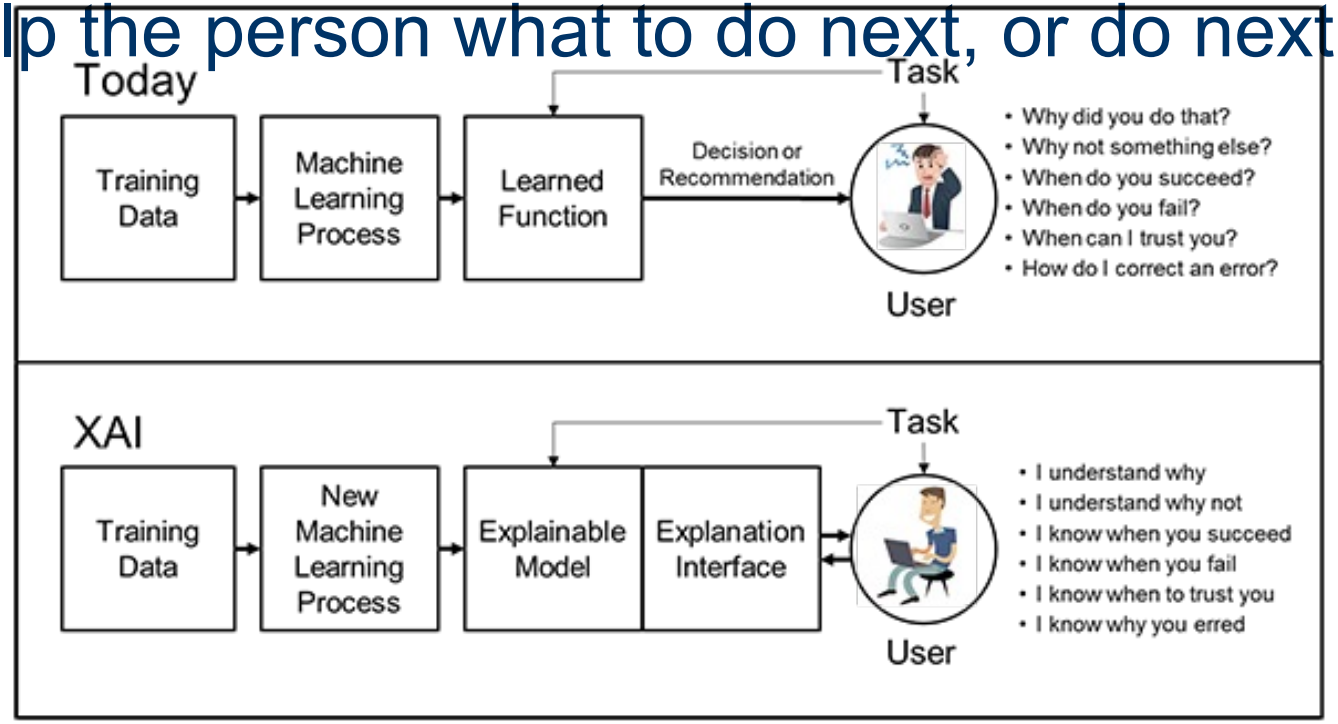


Figure 1. The Need for Explainable AI

63

Informational Justice

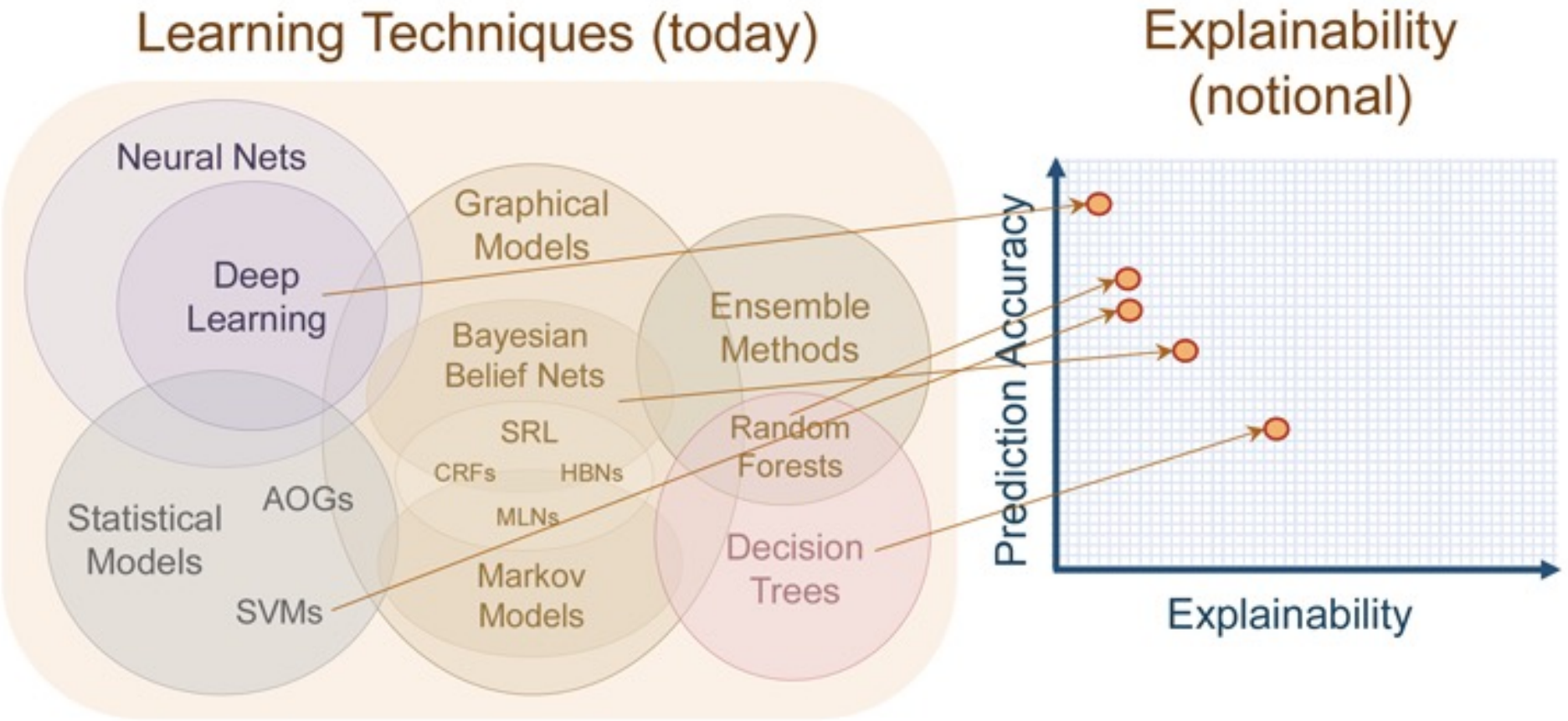
- Provide adequate justifications and **explanations** when things go badly
- Provide *correct* information
- Help the person what to do next, or do next time

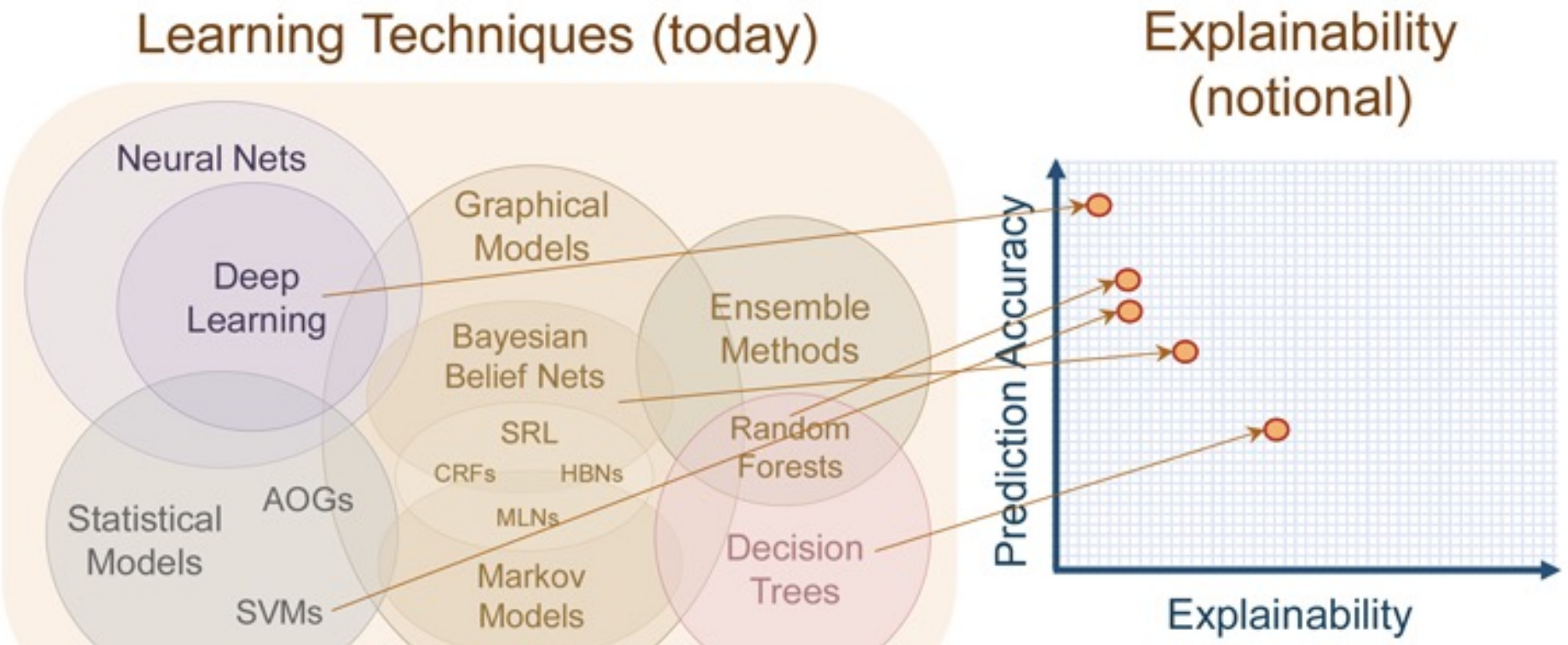


64

Informational Justice

- Provide adequate justifications and **explanations** when things go badly





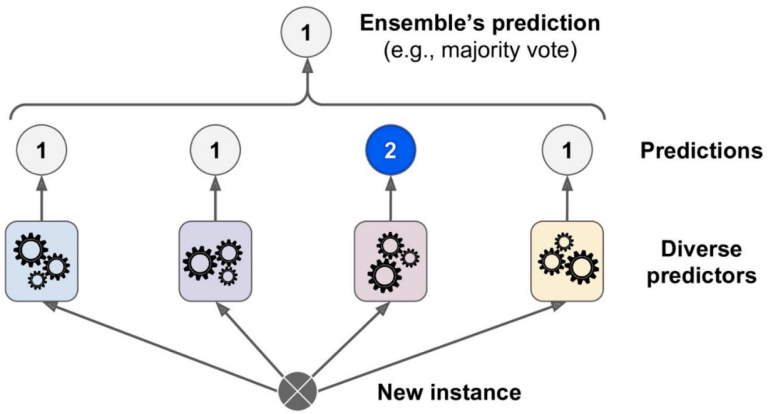
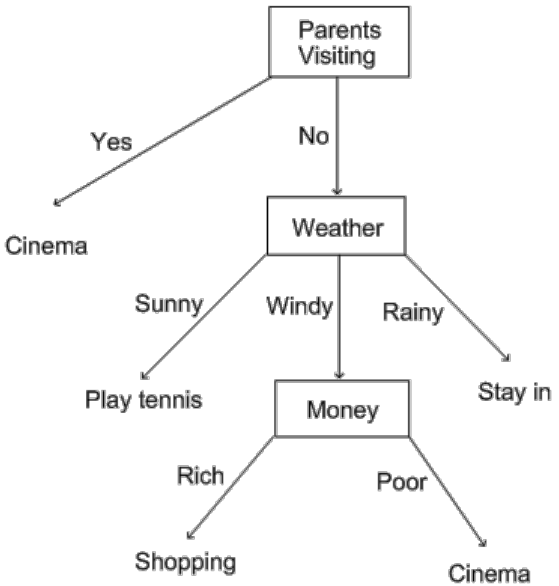


Figure 7-2. Hard voting classifier predictions



66

Agenda

- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - **Building Trust**
  - Explainability
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door

Building Trust

Trust is Essential for Successful Human Interactions

- Brands that are Trusted Cost More
- Invest in Building Brand for Trust
- Assurance of Trust with Doctors
- Depositing Savings with Trusted Bank
- No Big Purchases on Un-Trusted Websites
- Kids Play with Trusted People

Trust in AI

AI systems are not perfect. They are probabilistic and learn from past data, requiring oversight & trust for success

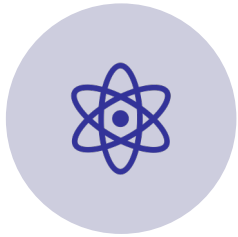
AI systems generate predictions, insights & actions affecting stakeholders directly or indirectly (how?)

Components of Trust

COMPETENCE

RELIABILITY

PREDICTABILITY



BENEVOLENCE

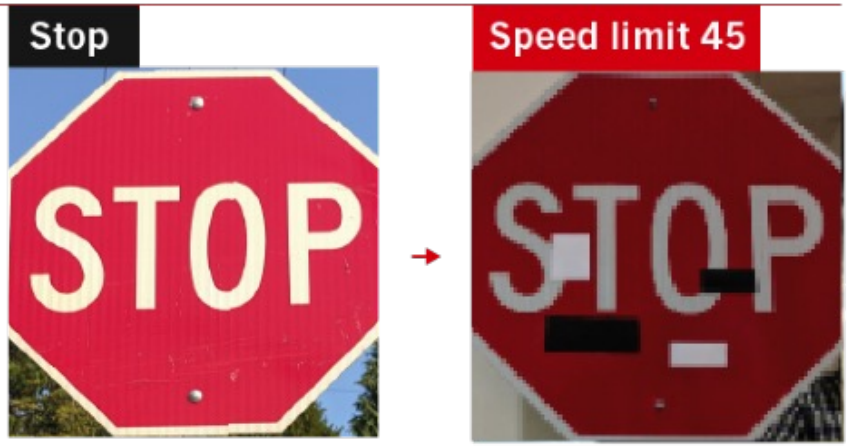
Competence

- The ability to get job done
- Clear Value: Improvement to user experience & satisfaction
- Example: Google Search (satisfactory results)

Reliability

- Ensures consistent experiences
- Low risk of failure or breakdown
- Consistently delivers on abilities

These stickers made an artificial-intelligence system read this stop sign as 'speed limit 45'.



Predictability

If humans already can perform the task quickly with skill, introducing AI with wide-ranging outputs can throw them off

Especially necessary for critical, time-sensitive tasks

More dynamic AI-based solutions can be used for open-ended goals like exploration.

Benevolence

- Belief that the trusted party wants to do good for the user
- Honesty and transparency in the relationship

How to Build Trust?

- Explainability
  - Users understand how the AI systems work
  - Setting the right expectations
  - Calibrating trust in AI's recommendations
  - Optimizing explanations for user understanding
- Control
  - Allow users to second-guess the AI's predictions
  - Allow users to edit data, choose types of results, ignore recommendations, and correct mistakes
  - Allow users to stop and mistakes, and correct them

Agenda

- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - **Explainability**
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door

Explainability

- Critical for gaining trust in AI decisions that heavily impact people’s lives
- An AI system providing a correct recommendation for the wrong reasons is a fluke; it is not trustworthy.
- Humans can benefit from a "theory of mind" when it comes to understanding AI systems
- Showing reasoning behind AI decisions is necessary for gaining user trust and credibility

Agenda

- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Explainability
  - **Who Needs an Explanation?**
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door

Who Needs an Explanation?

1. **Decision-makers**
2. **Affected users**
3. **Regulators**
4. **Internal stakeholders**

Decision-Makers

They use AI systems to make decisions (e.g., issuing a loan, detecting cancer, triaging patients, employee screening)

Most decision-makers need simplified descriptions and have a low tolerance for complex explanations.

They need to understand how the system works

Sometimes insights into the model help improve their own future decisions

Affected Users

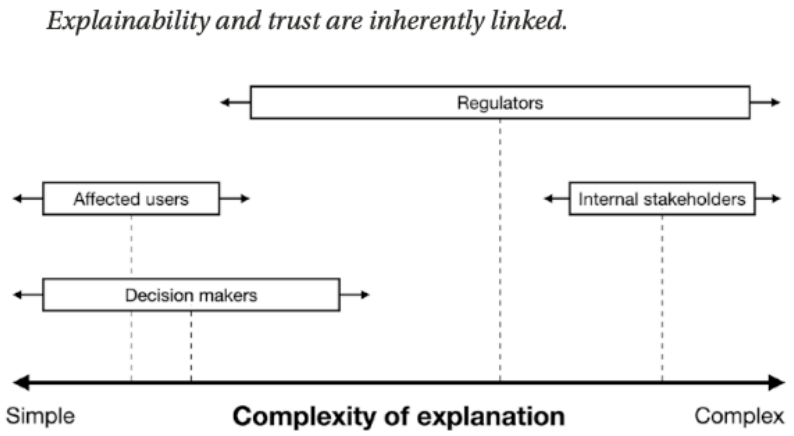
- The people impacted by the decision
  - e.g., loan applicant or hospital patient
- They need explanations that can help them understand
  - if they were treated fairly
  - What factors could be changed to get a different result
- Low tolerance for complex explanations

Regulators

- Internal: auditing committees
- External: government agencies
  - Enforce policies such as EU's General Data Protection Regulation (GDPR)
- Need explanations to ensure decisions are made safely and fairly
- Explanations can show the overall process: the training data used, and the level of confidence in the algorithm
- They may or may not have a high tolerance for complex explanations

Internal Stakeholders

- Builders of the systems:
  - ML engineers, Product managers, Designers, Data scientists, Developers
- Need explainability for debugging & improving system
- Require detailed system explanations



Agenda

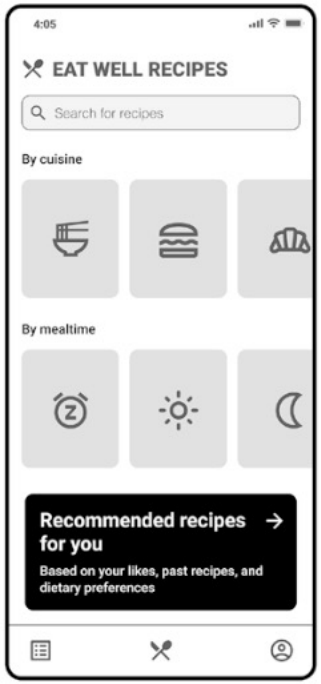
- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Components of Trust
  - Explainability
  - Who Needs an Explanation?
  - **Guidelines for Designing AI Explanations**
  - Types of Explanations
  - Evaluating Explanations
- Lab #1 – Next Door Analysis

Guidelines for Designing AI Explanations

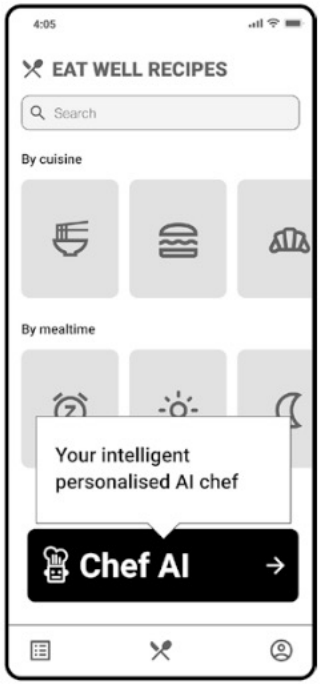
- Make clear what the system can do
- Make clear how well the system does its job
- Set expectations for adaptation
- Plan for calibrating trust
- Be transparent
- Build cause-and-effect relationships
- Optimize for understanding

Make Clear What the System Can Do

- Make sure users understand the capabilities of the system
- Provide contextual information to build trust
- Enumerate search possibilities
- Clarify how input influences results
- Avoid open-ended interaction for high-stakes situations



Aim for



Avoid

Make Clear How Well the System Does Its Job

- Set expectations of performance out of the box
- Clarify mistakes that will happen
- Use uncertain language when appropriate
- Consider providing a help context
- Update expectations when things change
- Don't expect deterministic behavior from a probabilistic system

Aim for

Avoid

Set Expectations for Adaptation

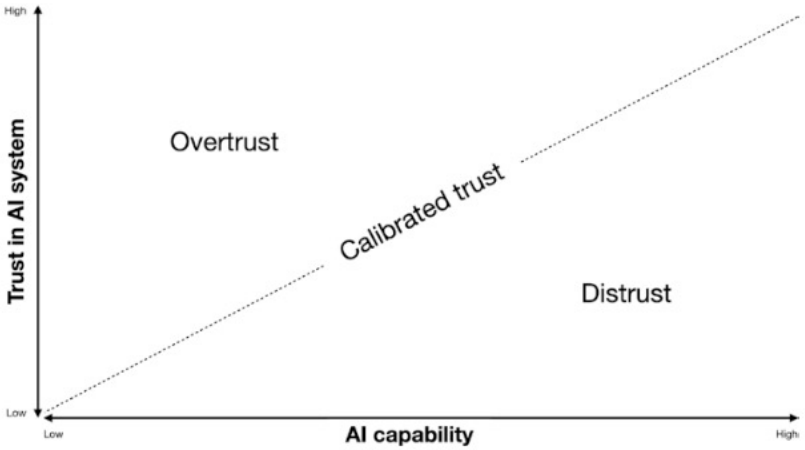
- AI systems can learn and adapt
- Clarify how input influences results
- Communicate changes in accuracy when conditions change
- Examples: Netflix recommendations, navigation system predicting time to arrival

Aim for

Avoid

Plan for Calibrating Trust

- Calibrate trust in AI by explaining & displaying confidence
- The user must know when to trust & use own judgment
- Trust calibration occurs over time: AI adapts, user preferences change, explanations should too
- Design better explanations & workflows to plan for trust calibration



Example Workflows that help in calibrating trust

- Communicate what data is used to train the AI (e.g. locality)
- Offering a ‘sandbox’ to trial product
- Showing accuracy levels or changes in accuracy when analyzing assembly line defects
- Displaying *reasoning* for product recommendations
  - e.g. “customers also bought” vs. “recommended for you”

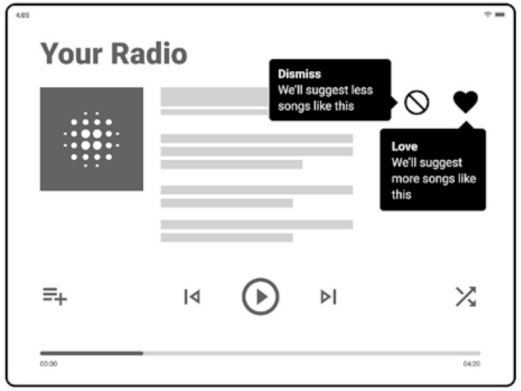


Figure 4-6. Music recommendation. *The messaging highlights how the AI improves over time*

Be Transparent

- Make users aware of any data collected, tracked, or monitored
- Allow users to choose how their data is used
- Make explanations of predictions understandable
- Possibly used non-black box models for additional transparency

Transparency means operating in ways that make it easy for others to understand the actions of a system.

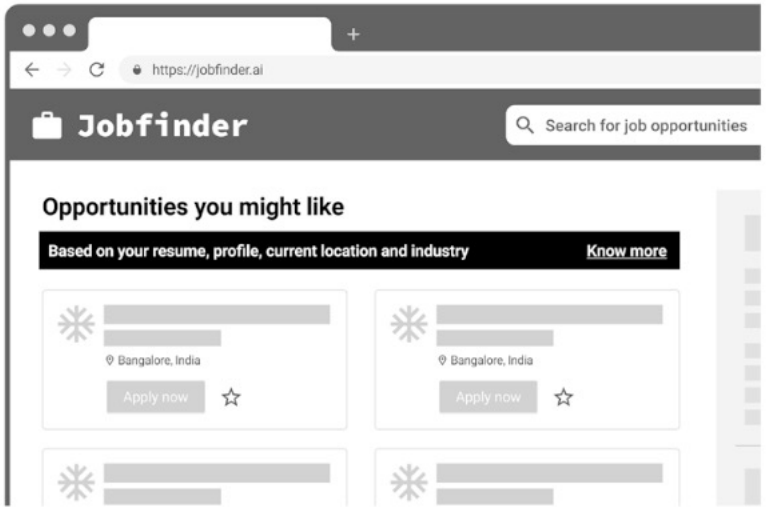


Figure 4-7. *Job recommendations. The callout explains how the system generated the recommendations*

Build Cause-and-Effect Relationships

- Allow users to identify a cause-and-effect relationship between their actions and the system's response
- Examples
  - Alexa lights up when spoken to

Building trust is a long-term process. A user's relationship with your product can evolve over time through back-and-forth interactions that reveal the AI's strengths, weaknesses, and behaviors.

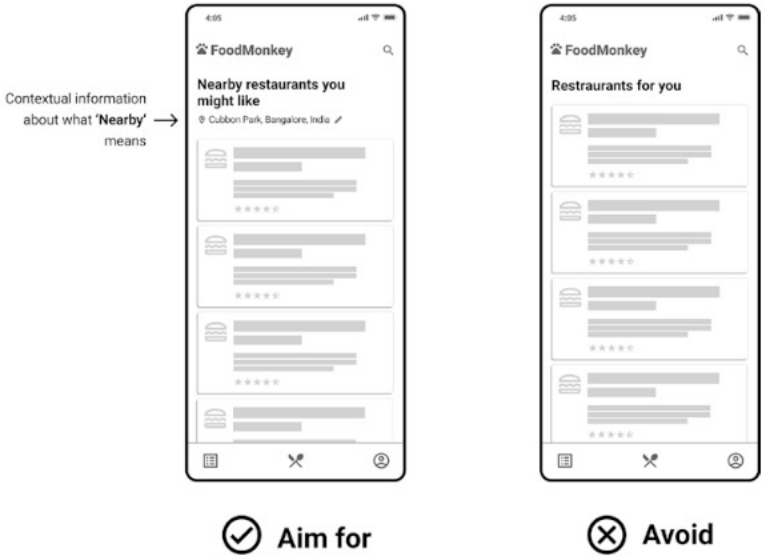
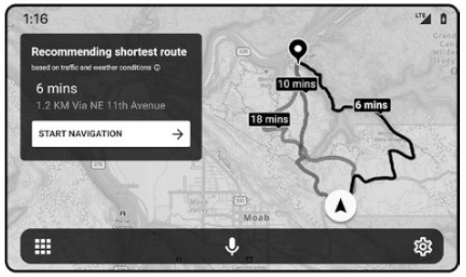


Figure 4-8. Restaurant recommendations: Build cause-and-effect relationships. *(Left) The restaurant recommendations include contextual information of the user's current location, which helps the user understand why certain restaurants are recommended. (Right) The recommendation on the right is ambiguous*

Optimize for Understanding

- Different stakeholders in your AI system need different levels of explanations
- Don’t try to explain everything unless it affects user trust and decision-making
- Accuracy vs. explainability is a challenge in high-stakes environments



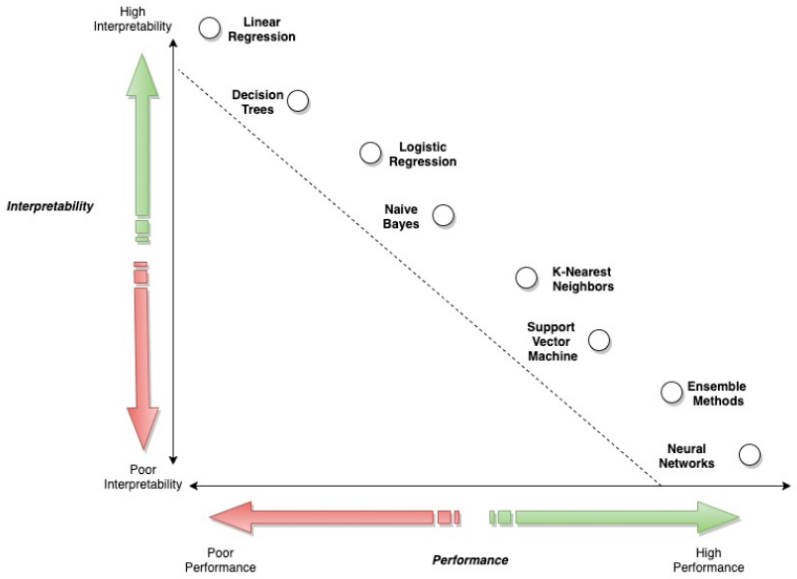
Aim for

Avoid

Figure 4-9. *Don’t explain everything. (Top) Informing the user that the shortest route is recommended is an easy explanation that most users can understand and act on. (Bottom) In this case, a detailed explanation of how the AI system works is not useful*

Accuracy vs. explainability

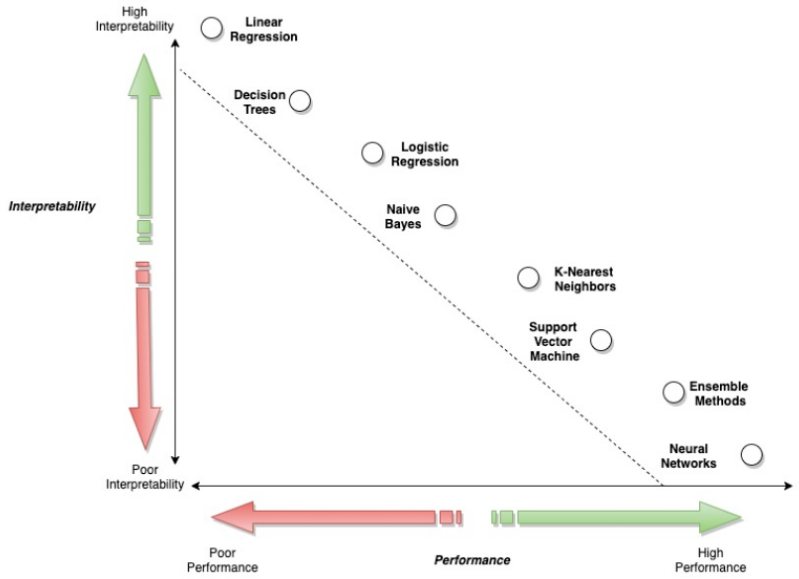
- Is interpretability is required?
- Model explainability can be used in any AI/ML use case
- If detailed transparency is required, then your AI/ML method selection becomes limited



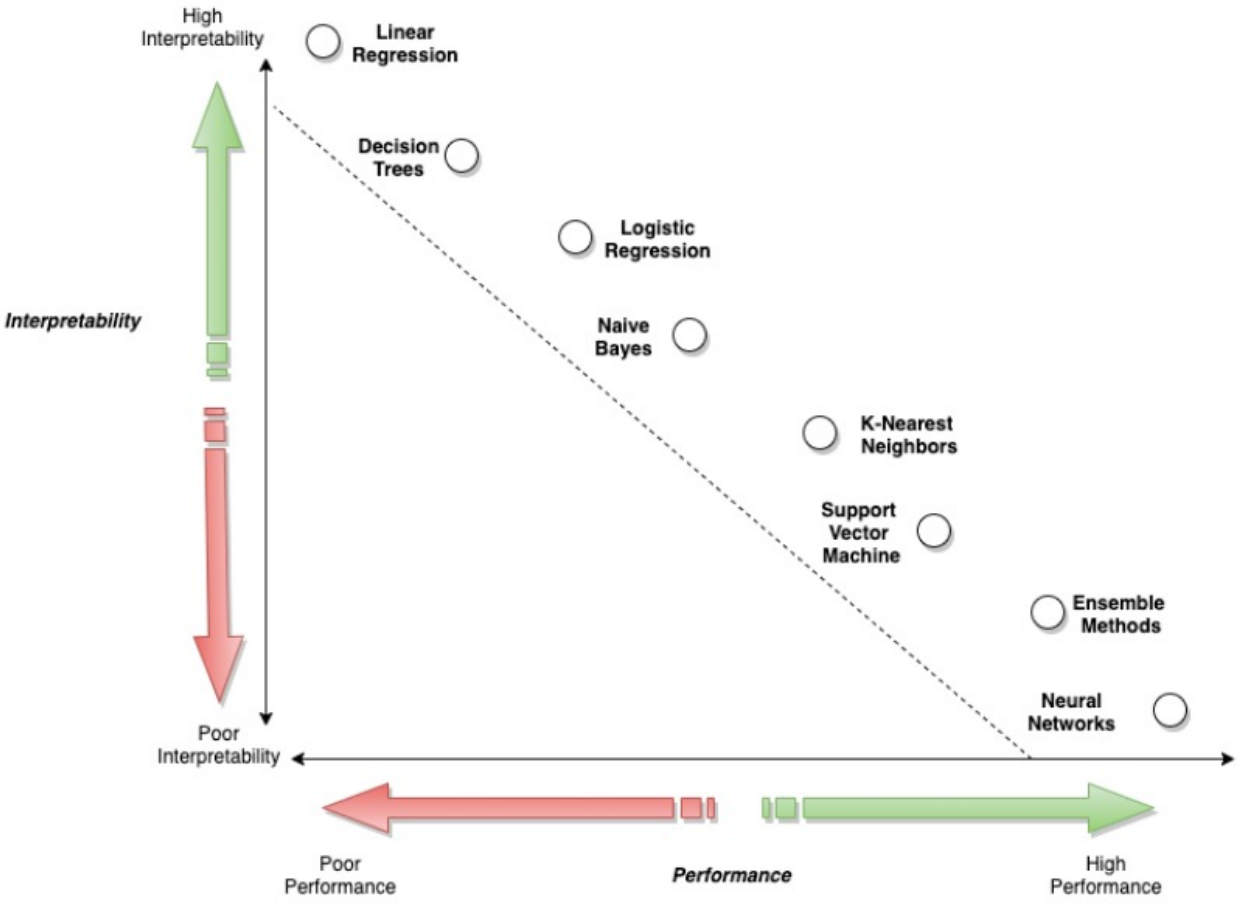
Is interpretability a hard business requirement?

Regulations or business requirements for complete model transparency:

- Select an interpretable model.
- Document how the inner mechanisms of the model impact the output and explain the model in human terms.



Accuracy vs. explainability



Agenda

- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Components of Trust
  - Explainability
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - **Types of Explanations**
  - Evaluating Explanations
- Lab #1 – Next Door

Types of Explanations – What is the Question?

The following are some of the most common types of questions:

1. What did the system do?
2. Why did the system do it?
3. Why did the system not do this?
4. What would the system do if this happened?
5. How does it do it?
6. What is the overall model of how the system works?
7. What data does the system learn from?
8. How confident is the system about a prediction or an outcome?
9. What can I do to get a different prediction?
10. What changes are permitted to keep the same prediction?

Types of Explanations

Data use explanations

Descriptions

Confidence-based explanations

Explaining through experimentation

No explanation

Data Use Explanations

What Data does the system learn from?

How is it collected? From sensors, users, or …?

This prevents both over-trust or distrust

How is the data used?

What decisions or service is enabled by using this data?

How does it do it?

Highlight which datasets, training data points or key variables were most important to the decision (e.g., Shapley Values)

Privacy Implications

Be explicit about what is used for AI and what isn't

Provide optionality to opt-out

Data Use Explanations

Scope of Data Use:

Include which *subsets* of data were used to increase trust and transparency.

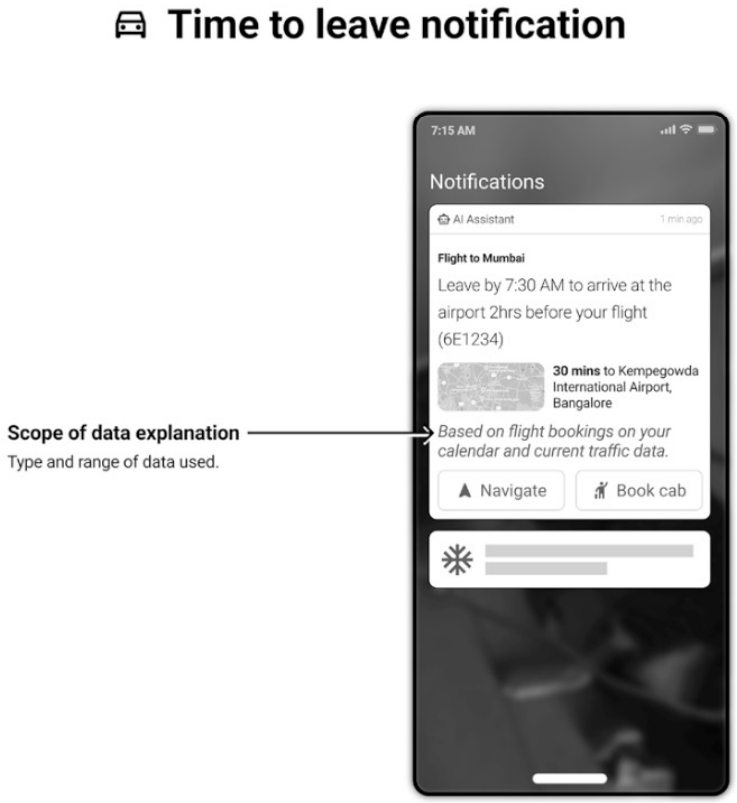


Figure 4-10. Scope of data use. *The notification has an explanation of how the AI knows it is time to leave for a flight, for example, based on flight bookings on your calendar and current traffic data*

Data Explanations: Examples-Based

Using **specific examples** from training data (e.g., of similar points to the query) can increase trust in the system.

Useful even if the model is too black-boxed to explain its reasoning

Figure 4-13. Specific examples-based explanation. *When it detects a dog breed, the dog classification system shows similar images from its training data along with the result. This helps users gauge how much they can trust the result. For example, if the similar images for a "poodle" prediction were photos of cats, you wouldn't trust the system's results*

Data Explanations: Examples-Based

Showing **generic examples** of where the model performs well vs poorly can also help

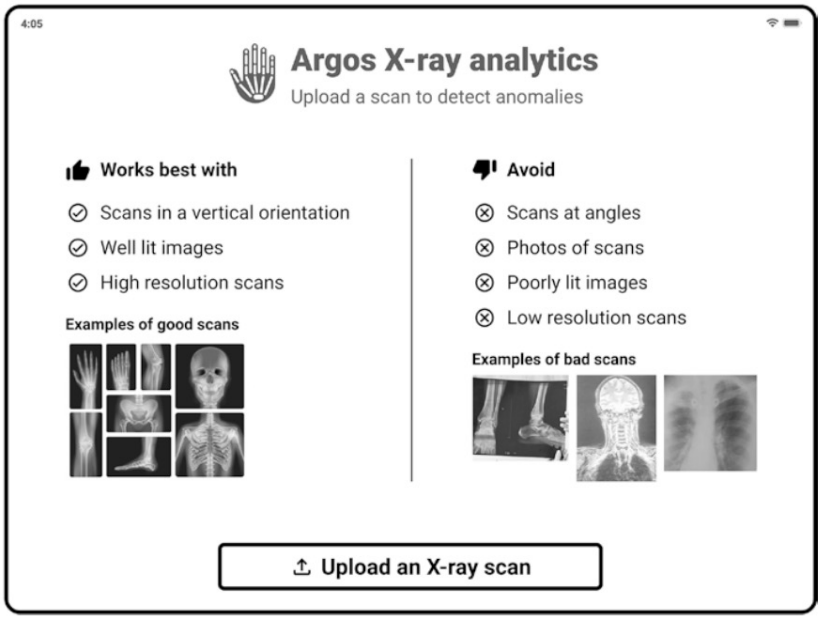


Figure 4-12. Generic examples-based explanation. *X-ray analytics interface shows examples of images the system performs well on and where it doesn't*

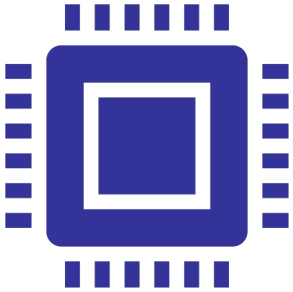
Better Descriptions

Explain the benefit to users, not the technology or algorithm

For example, it is better to say "We help you remove spam from your inbox" than "We've created an email filtering technique that categorizes email into spam and non-spam by using deep learning."

Better Descriptions: More human-scale, less technical

More understandable, intuitive answers

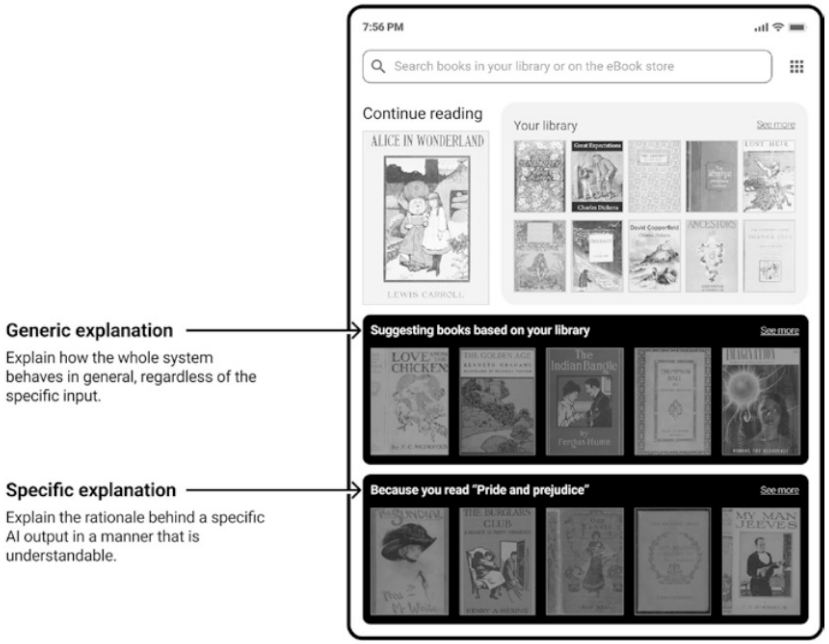


For example, typing “area of Poland” into the search engine Bing returns the literal answer (120,728 square miles) along with the note “About equal to the size of Nevada.”

Partial vs. Full Descriptions

Partial Descriptions:

- Usually best for UX
- Generic
  - How the overall system behaves
  - Inputs and outputs
- Specific
  - Explaining a particular output
  - “This dog is most likely a German shepherd because of XYZ features.”
  - “Books recommended because you read Pride and Prejudice”

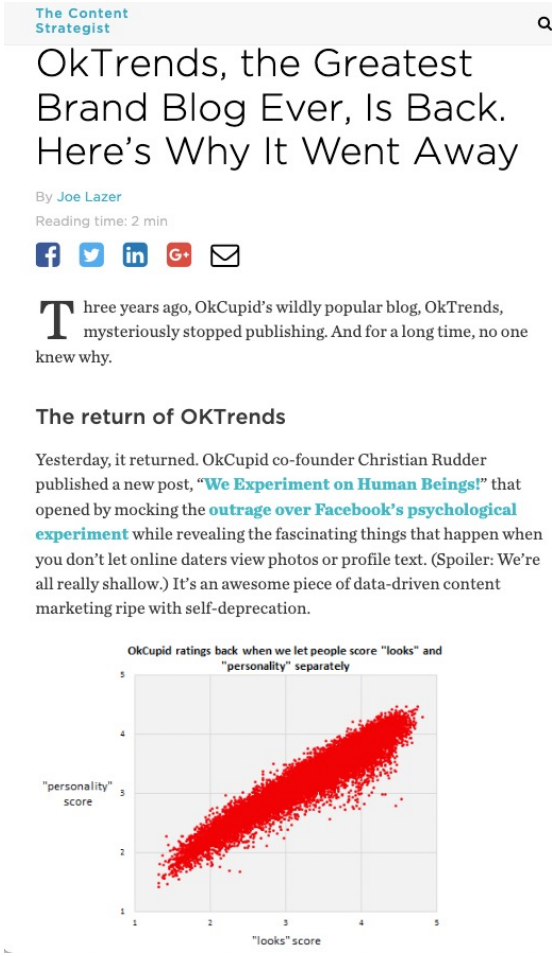


Partial vs. Full Descriptions

Full Descriptions

- Can be too much info in many cases

- Can be kept on a company blog or marketing materials and linked to with tooltips



The Content Strategist

OkTrends, the Greatest Brand Blog Ever, Is Back. Here's Why It Went Away

By Joe Lazer

Reading time: 2 min

Three years ago, OkCupid's wildly popular blog, OkTrends, mysteriously stopped publishing. And for a long time, no one knew why.

The return of OKTrends

Yesterday, it returned. OkCupid co-founder Christian Rudder published a new post, "**We Experiment on Human Beings!**" that opened by mocking the **outrage over Facebook's psychological experiment** while revealing the fascinating things that happen when you don't let online daters view photos or profile text. (Spoiler: We're all really shallow.) It's an awesome piece of data-driven content marketing ripe with self-deprecation.

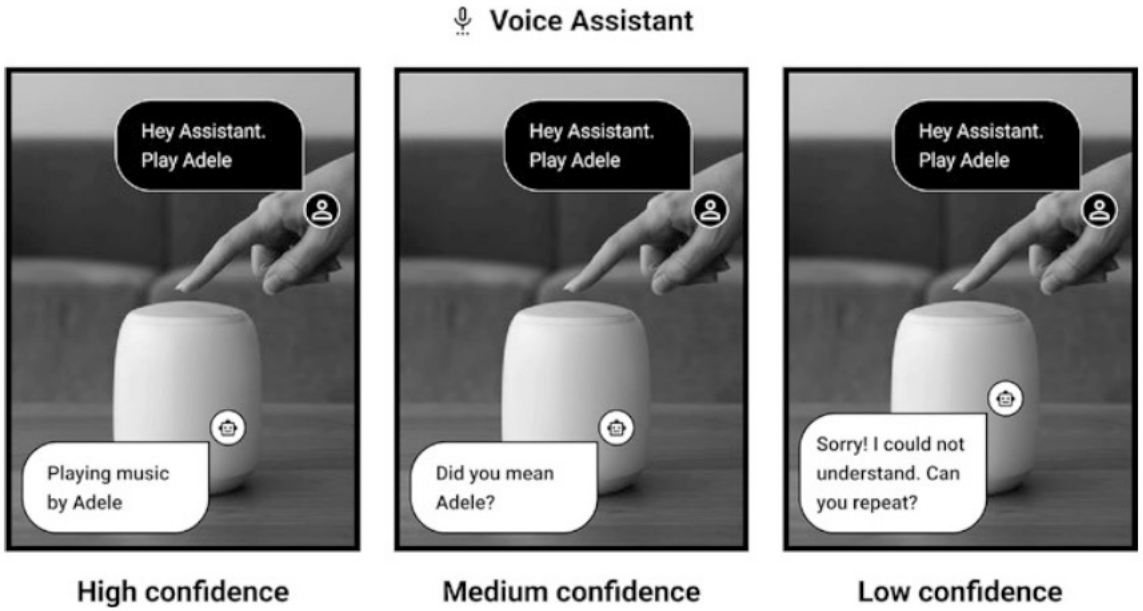
Confidence-Based Explanations

Image	Confidence value	Confidence level	System output
	98%	High	*"That's a Donut"*
	65%	Medium	*"Maybe this is a Donut"*
	15%	Low	*"Not a Donut"*

Figure 4-15. *Examples of confidence levels and outputs for a system that identifies donuts*

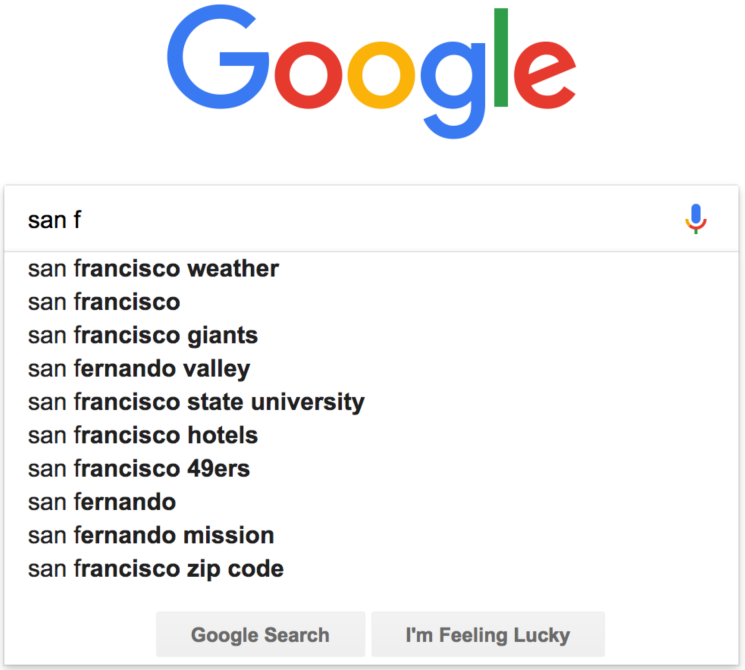
Confidence-Based Explanations

- Should you use confidence scores?
  - Test with users to tell if it is helpful or annoying
  - Any risk of creating blind trust in the system?
  - Using natural language may be more intuitive



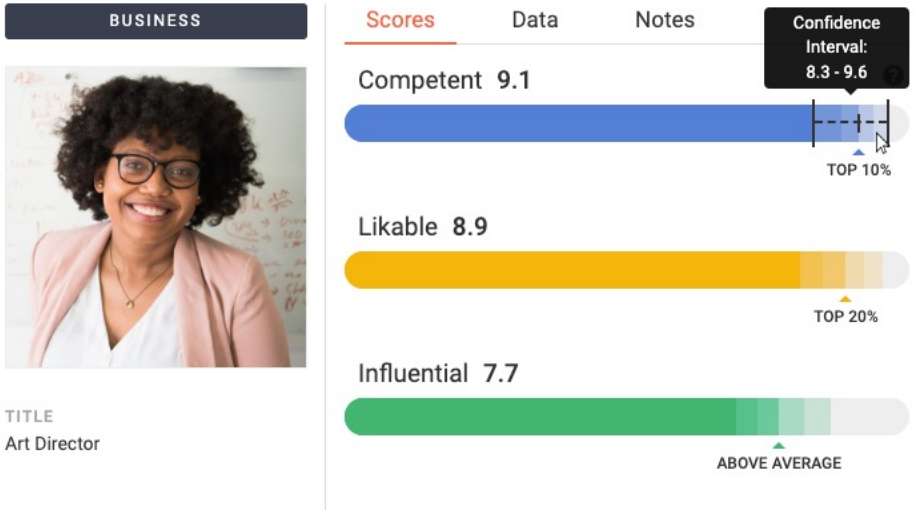
Confidence-Based Explanations

- N-Best Results
  - You can sort guesses by confidence scores without showing the scores themselves



Confidence-Based Explanations

- Numeric Scores
  - Can show the overall accuracy of the system, or the confidence for specific queries



No Explanation at all?

- In some cases, no explicit explanation is needed.
  - If too distracting or repetitive
  - If reveals private info or technologies

Agenda

- Review from Last Week
- Trusting AI
  - Factors of Fairness
    - Distributional
    - Procedural
    - Interactional
  - Building Trust
  - Components of Trust
  - Explainability
  - Who Needs an Explanation?
  - Guidelines for Designing AI Explanations
  - Types of Explanations
  - **Evaluating Explanations**
- Lab #1 – Next Door

Evaluating Explanations

- Internal Assessment
- User Validation
  - Qualitative Methods
  - Quantitative Methods

Evaluating Explanations: Internal Assessment

Evaluate your explanations with your product managers, designers, machine learning scientists, and engineers on your team

- Which type of explanation is best for the user and product?
- Observer team-members interacting with the explanations
- What parts of the explanation are (ir)relevant to the user?
- Are there privacy, IP, or security risks?

Evaluating Explanations: User Validation - Qualitative

Interviews

Surveys & Customer feedback forms

Task Completion rates/speed

Silent observer (fly on the wall)

Evaluating Explanations: User Validation - Quantitative

Tracking usage logs, product funnels, and usage metrics can provide some signals of where to look for issues

These should be paired with qualitative methods

Thank You!