Adversarial ML (Update)
+
Understanding Privacy Valuations

CMU lablet project w/ Matt Fredrikson, Mike Reiter (UNC)

Not a lablet project; w/ Michelle Mazurek (UMD)

Lujo Bauer

Carnegie Mellon University

January 11 2019

Adversarial Machine Learning: Curiosity, Benefit, or Threat?

Lujo Bauer

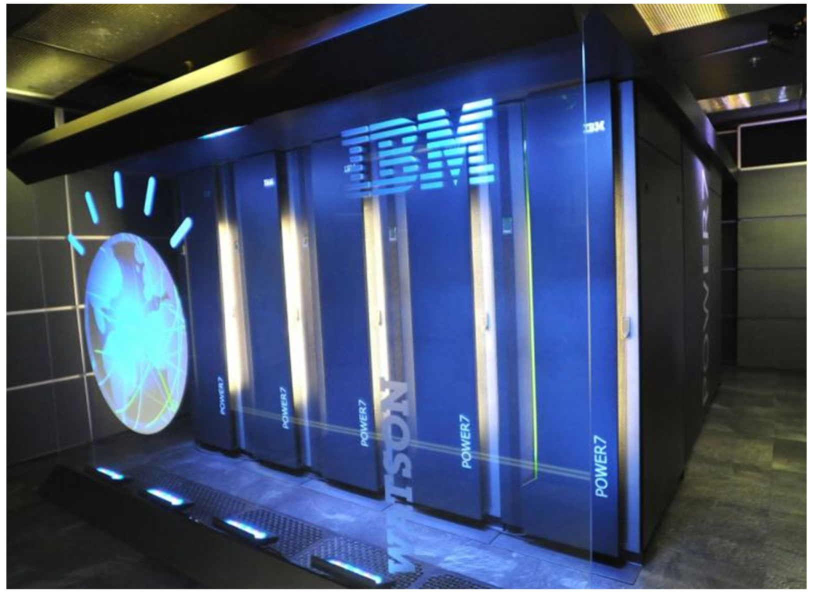
Collaborators: Mahmood Sharif,
Sruti Bhagavatula, Mike Reiter (UNC)

Carnegie Mellon University

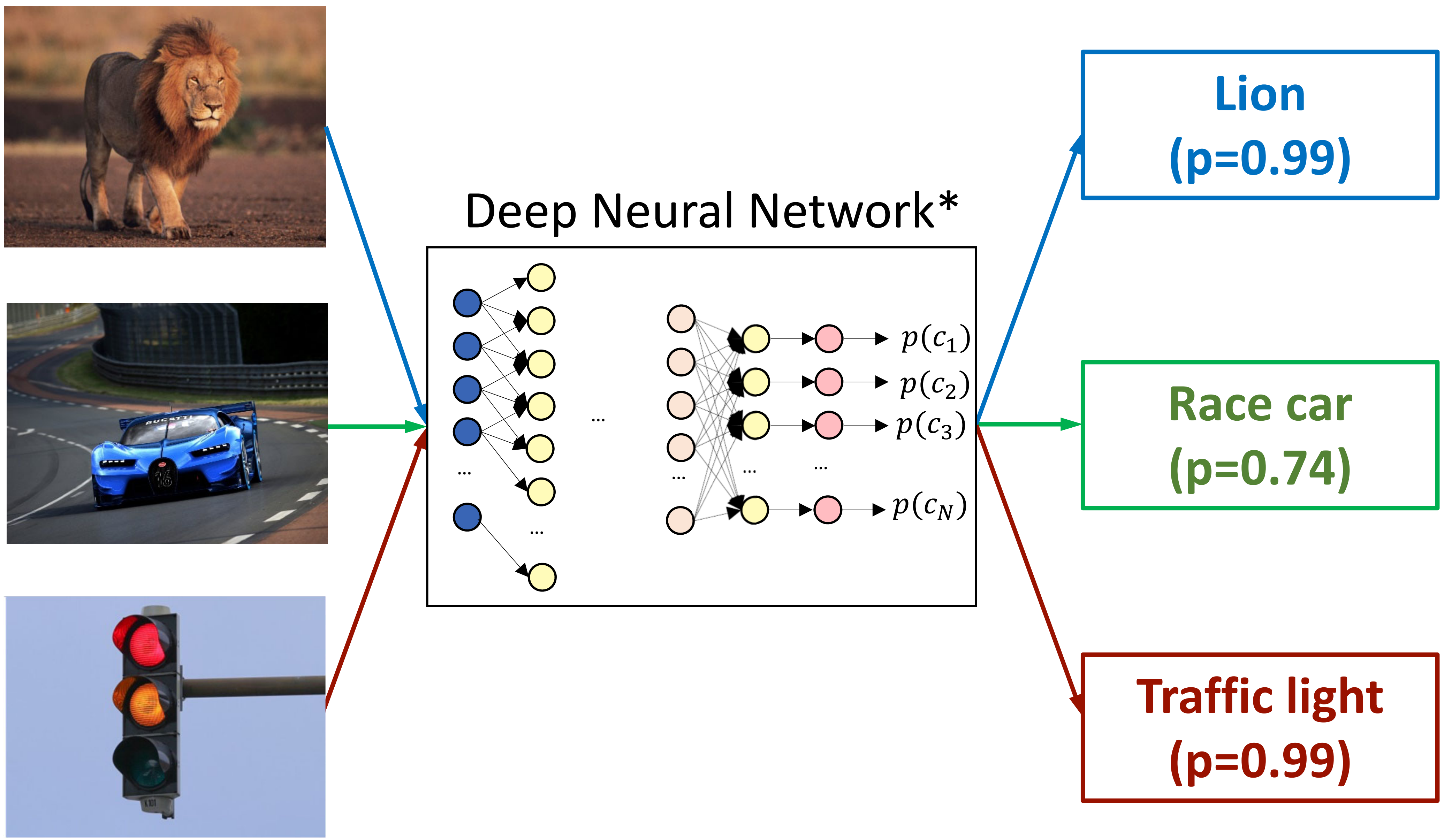
Carnegie Mellon University CyLab Security and Privacy Institute

Machine Learning Is Ubiquitous

- Cancer diagnosis
- Predicting weather
- Self-driving cars
- Surveillance and access-control

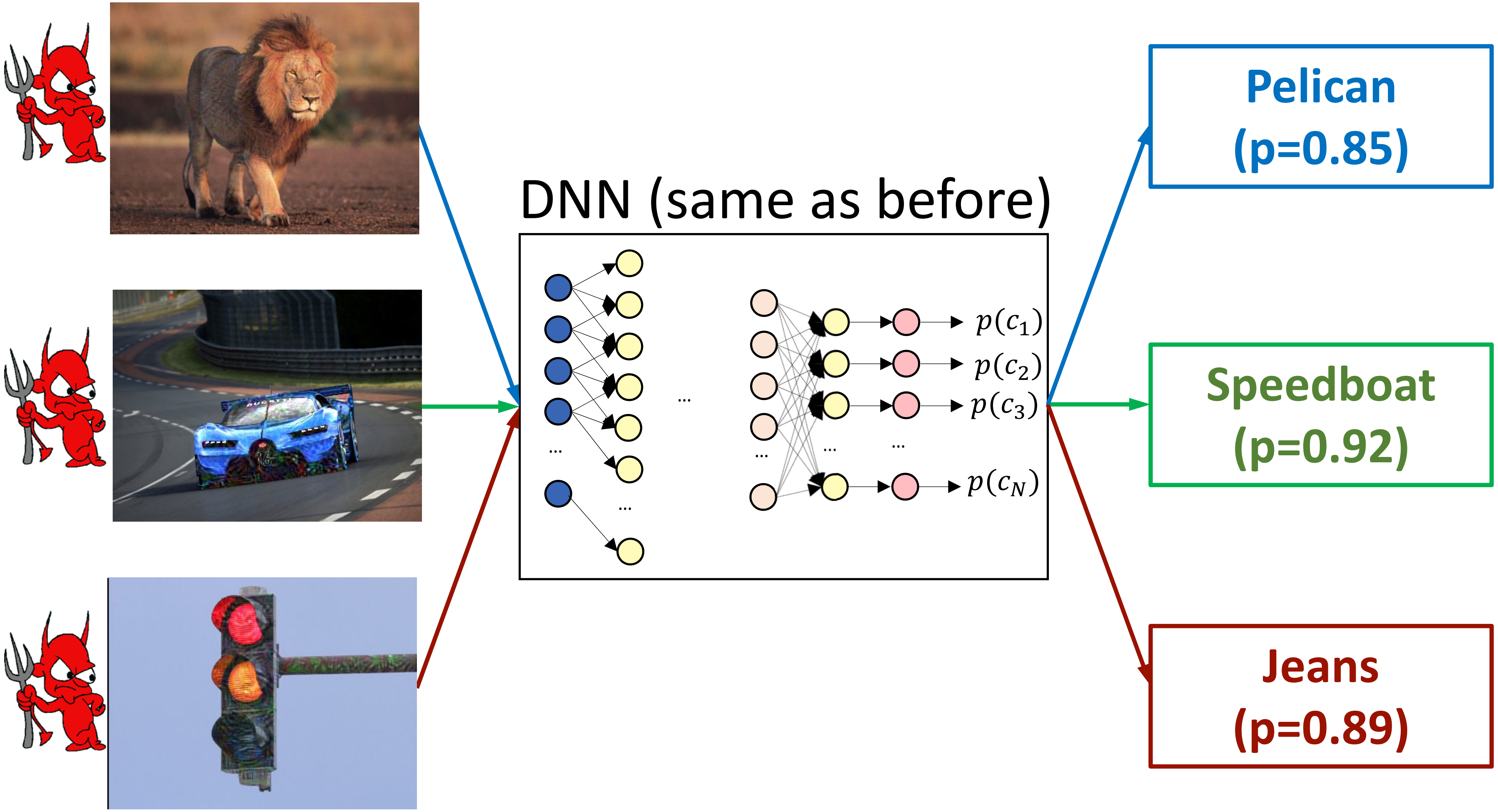


What Do You See?



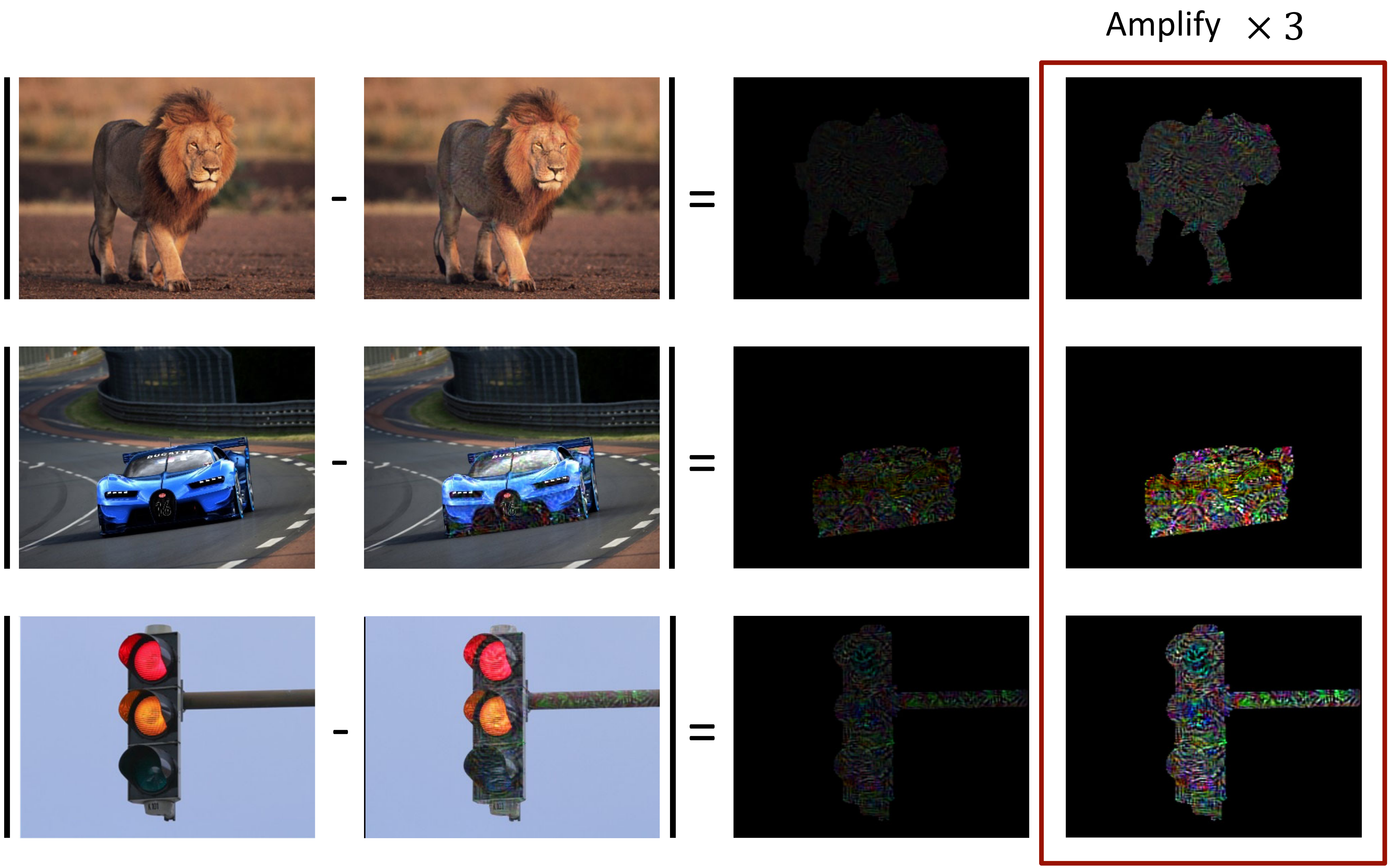
*CNN-F, proposed by Chatfield et al., “Return of the Devil”, BMVC ‘14

What Do You See Now?

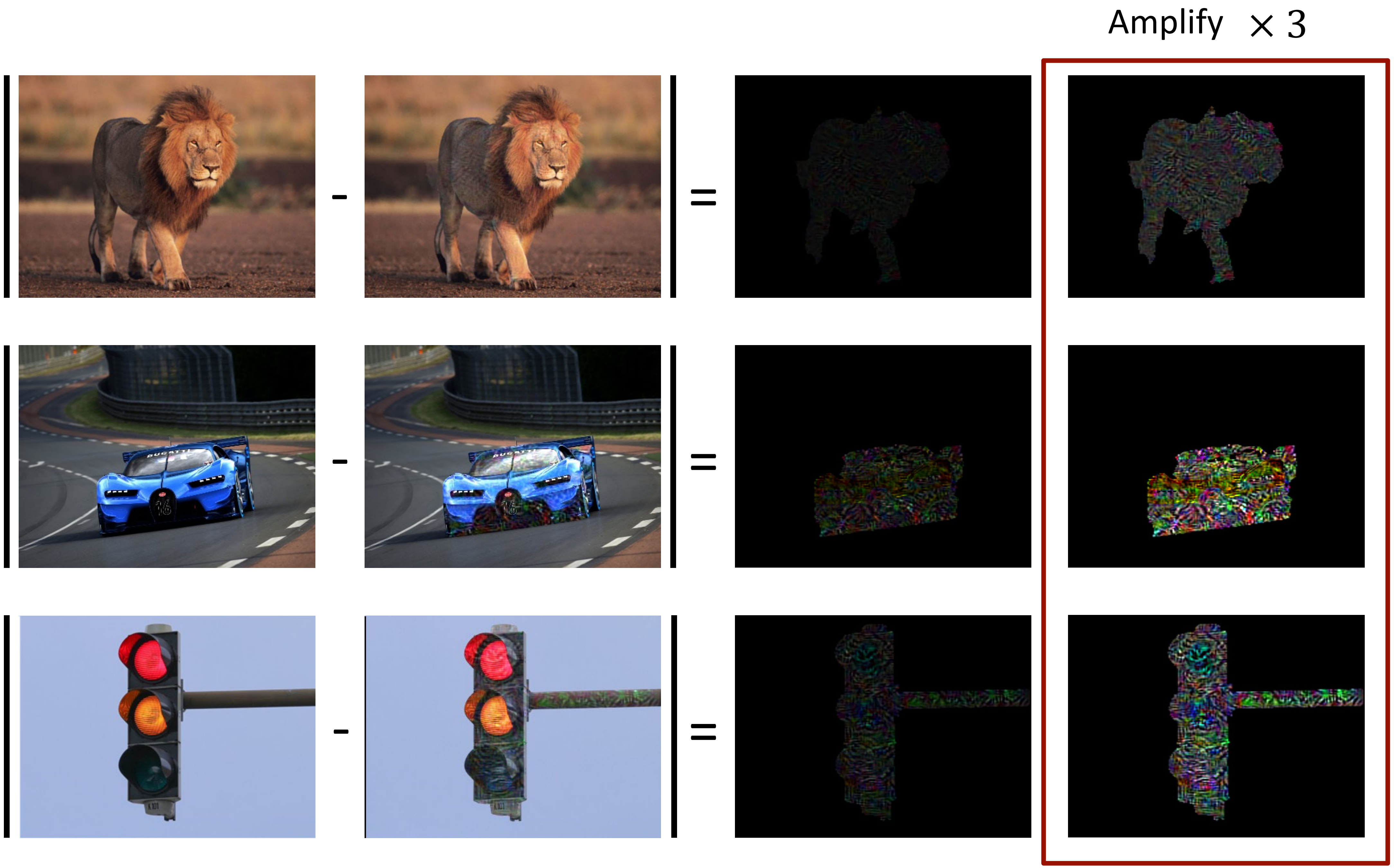


*The attacks generated following the method proposed by Szegedy et al.

The Difference



Is This an Attack?



[Sharif, Bhagavatula, Bauer, Reiter CCS'16,arXiv'17]

Can *an Attacker* Fool ML Classifiers?

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects, in a limited way

Can't control camera position, lighting

Defender / beholder doesn't notice attack
(to be measured by user study)

Carnegie Mellon University CyLab Security and Privacy Institute

Attempt #1

0. Start with Szegedy et al.'s attack
1. Restrict modification to eyeglasses
2. Smooth pixel transitions
3. Restrict to printable colors
4. Add robustness to pose

"Inconspicuousness"

Physical realizability

Carnegie Mellon University CyLab Security and Privacy Institute

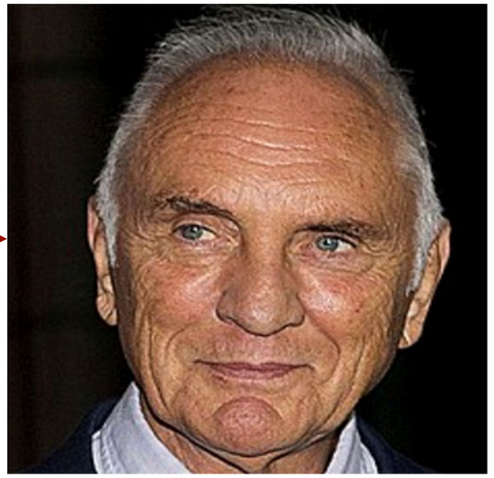
Attempt #1

0. Start with Szegedy et al.'s attack
1. Restrict modification to eyeglasses
2. Smooth pixel transitions
3. Restrict to printable colors
4. Add robustness to pose

“Inconspicuousness”

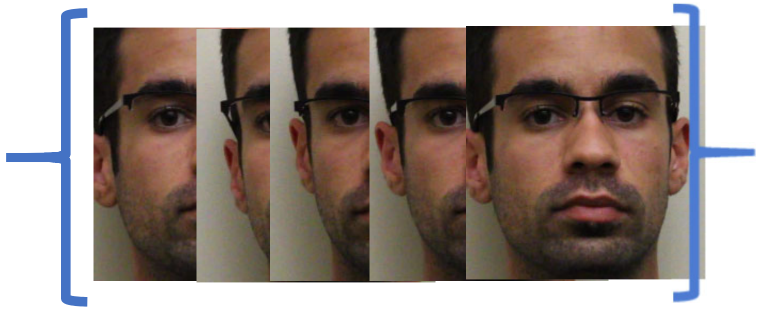
Physical realizability

Vicky McClure



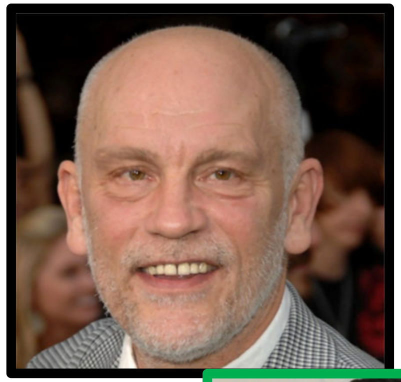
Terence Stamp

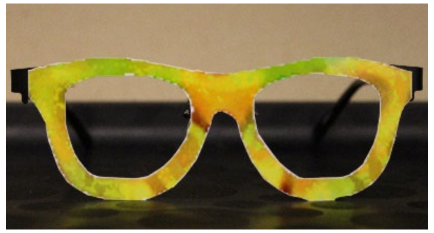
Time to Test!



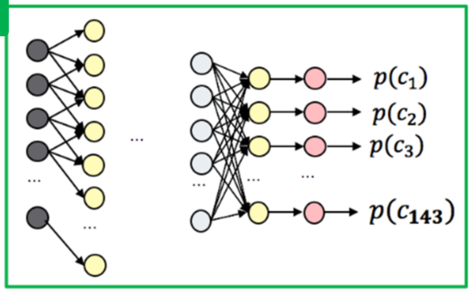
Procedure:

1. Collect images of attacker
2. Choose random target
3. Generate and print eyeglasses
4. Collect images of attacker wearing eyeglasses
5. Classify collected images





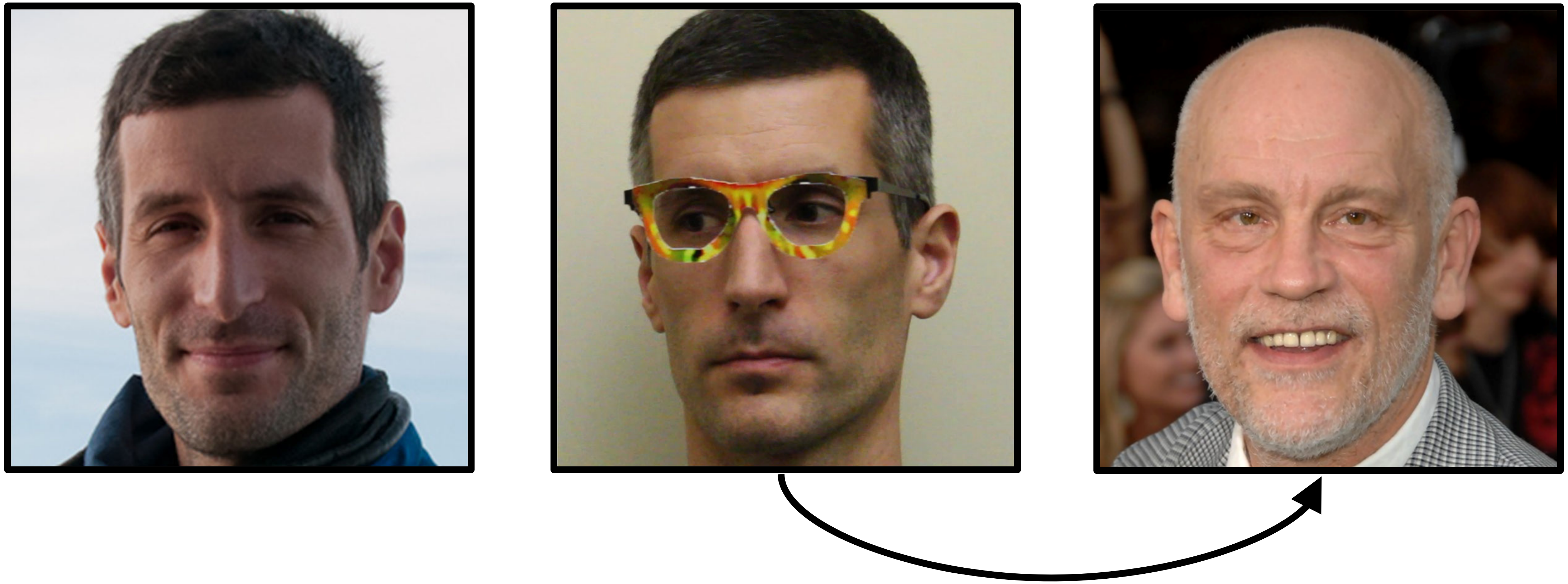
Success metric: fraction of images misclassified as target



Physically Realized Impersonation Attacks Work

Lujo

John Malkovich



100% success

Physically Realized Impersonation Attacks Work

Mahmood

Carson Daly

100% success

Carnegie Mellon University CyLab Security and Privacy Institute

Can *an Attacker* Fool ML Classifiers? (Attempt #1)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects, in a limited way ✓

Can't control camera position, lighting ?

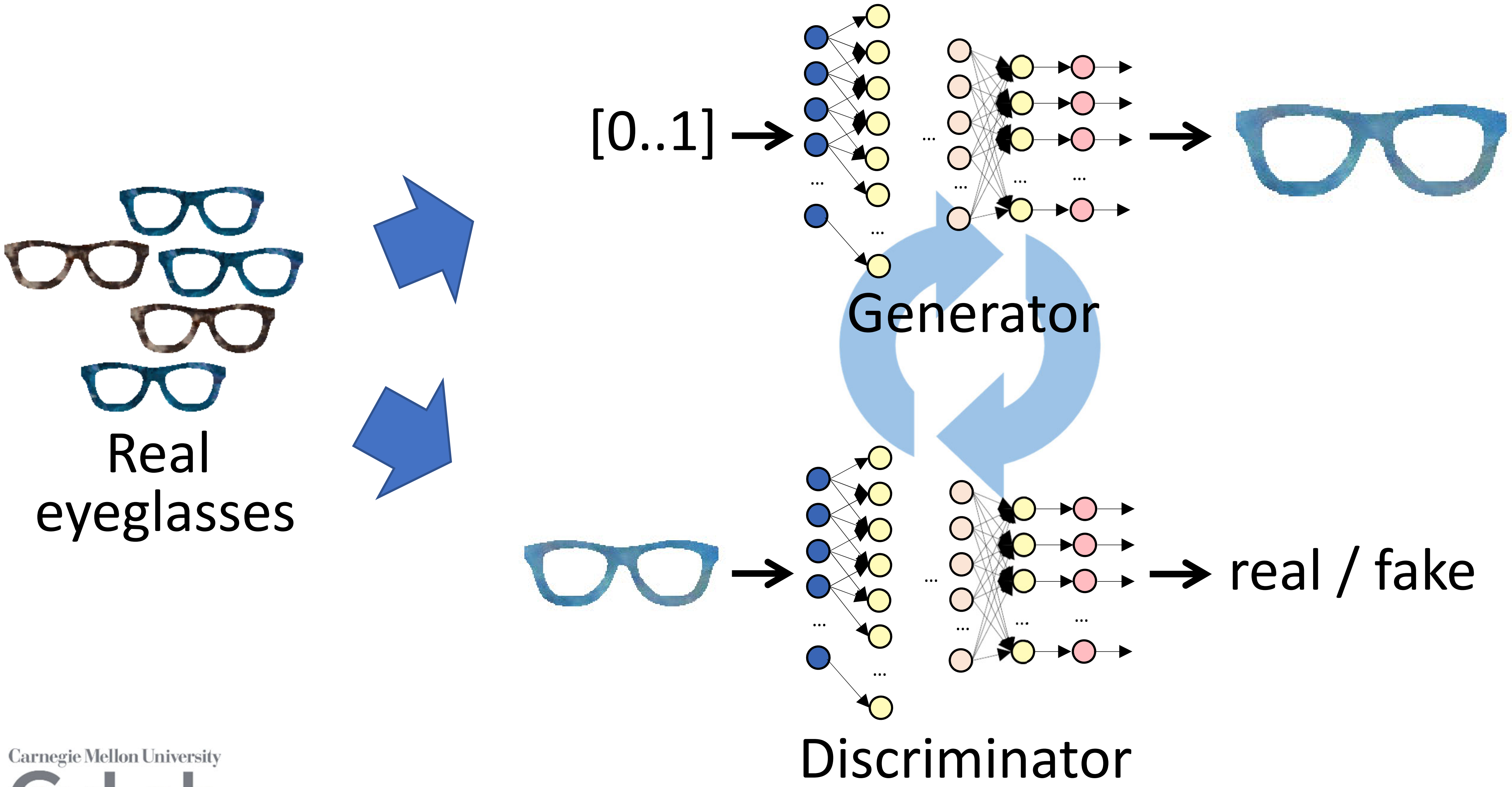
Defender / beholder doesn't notice attack (to be measured by user study) ?

Attempt #2

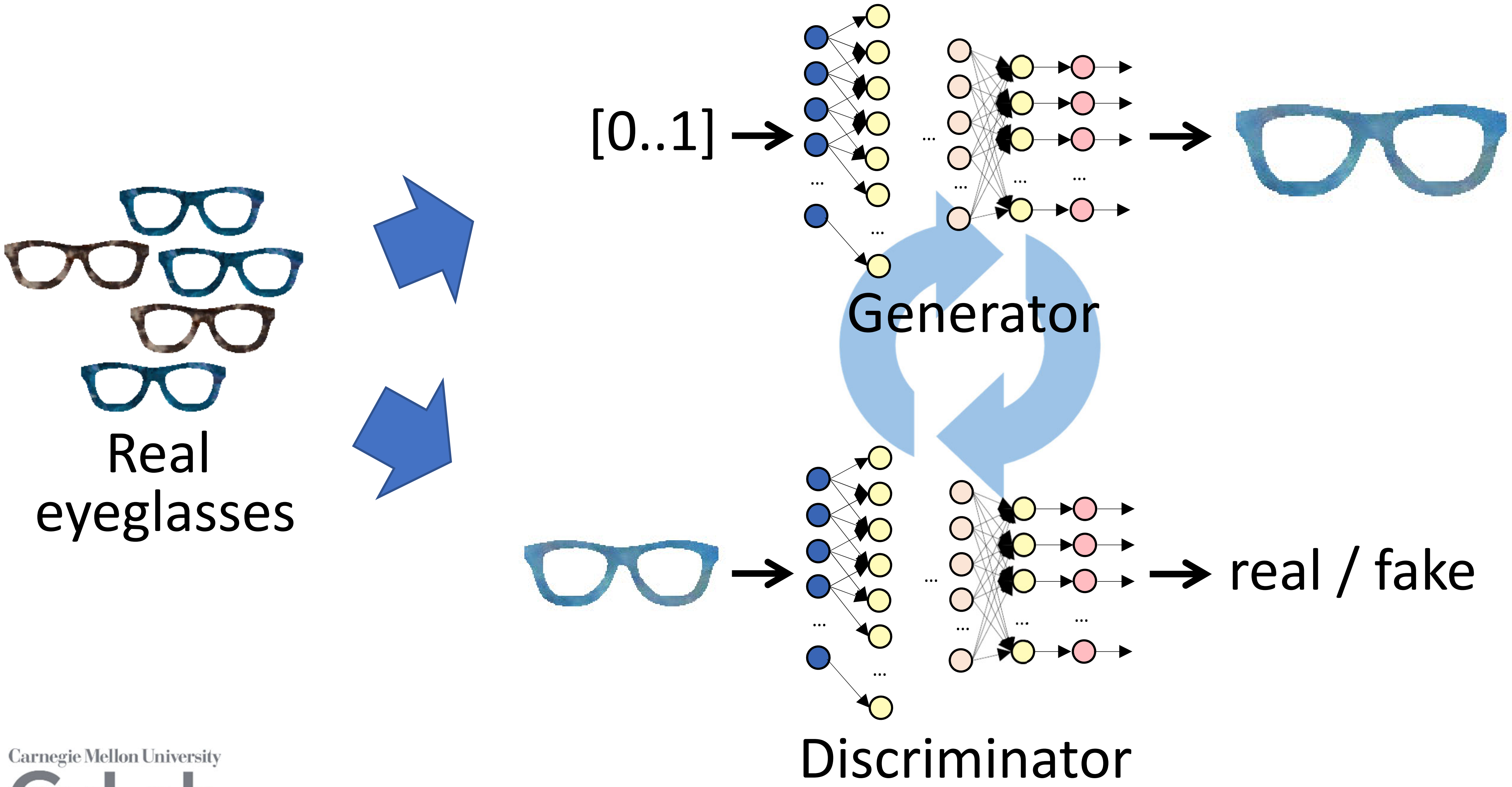
Goal: Capture hard-to-formalize constraints, i.e., "inconspicuousness"

Approach: Encode constraints using a neural network

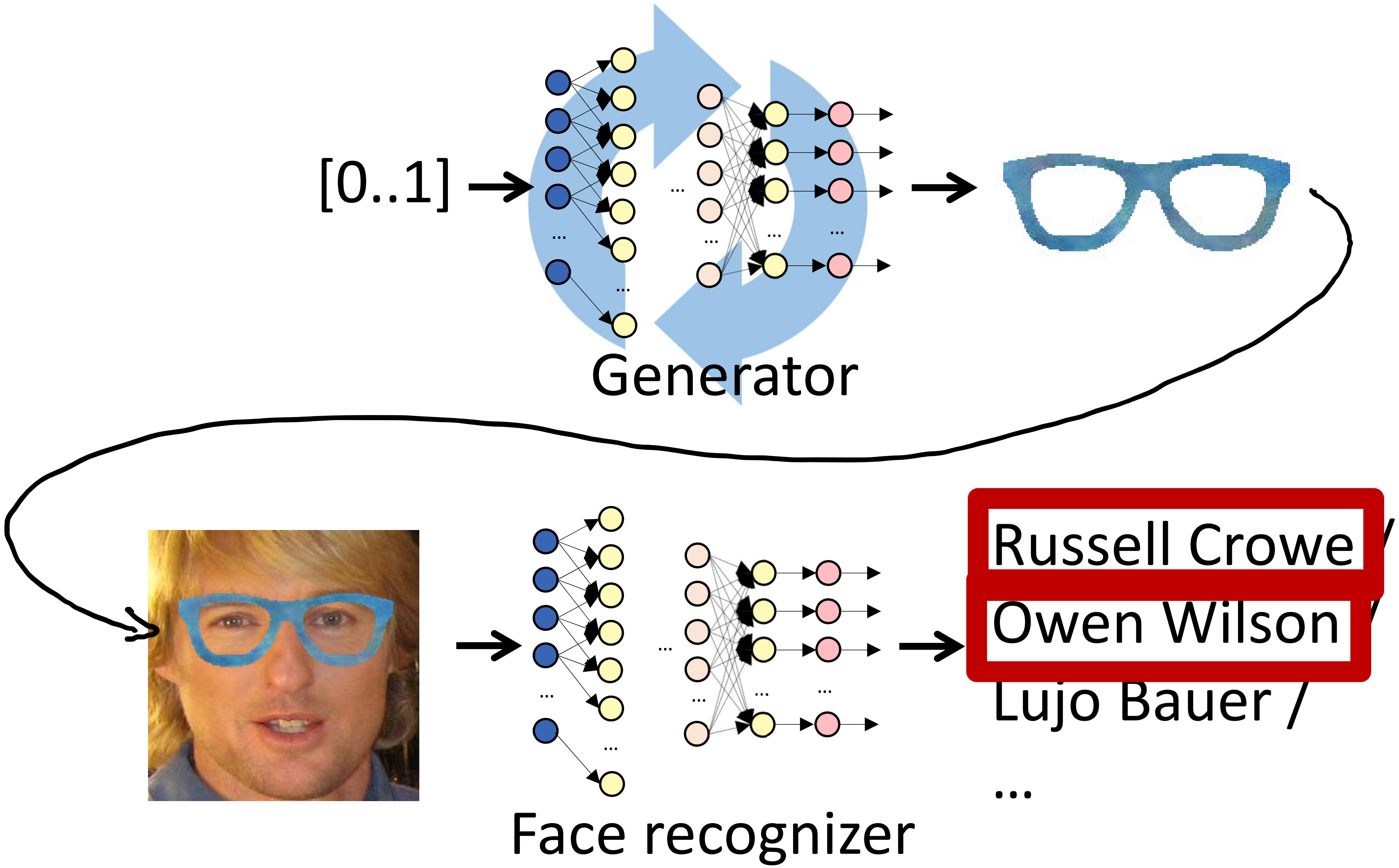
Step #1: Generate Realistic Eyeglasses

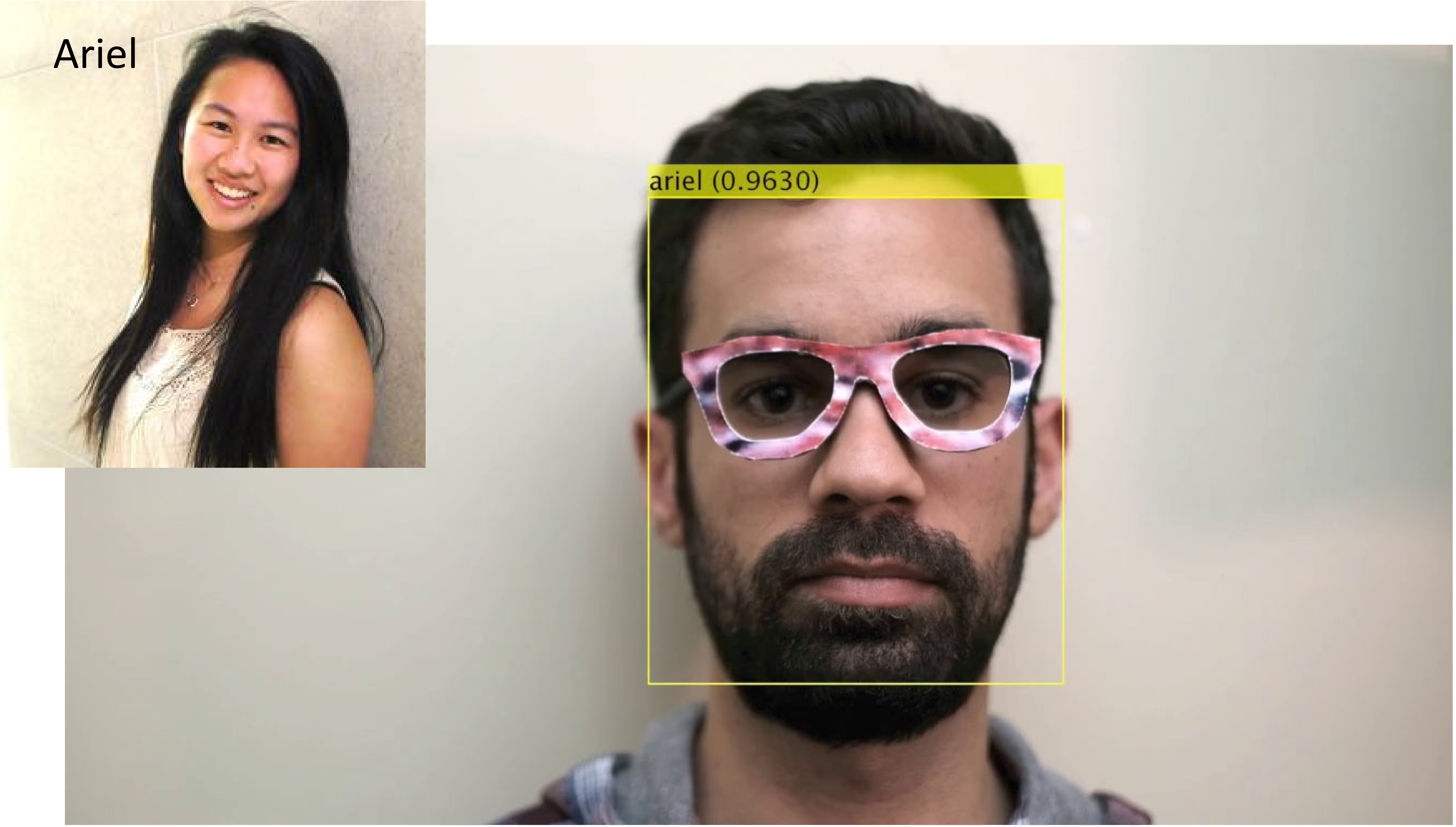


Step #2: Generate Realistic *Adversarial* Eyeglasses

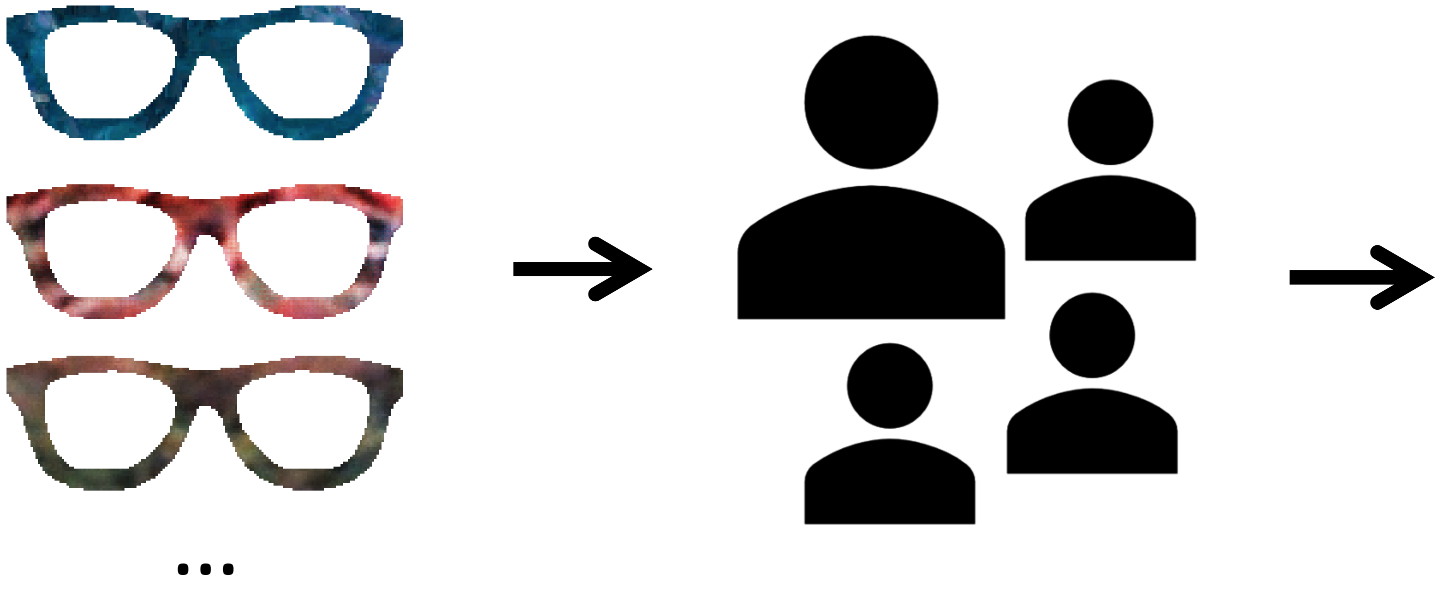


Step #2: Generate Realistic *Adversarial* Eyeglasses



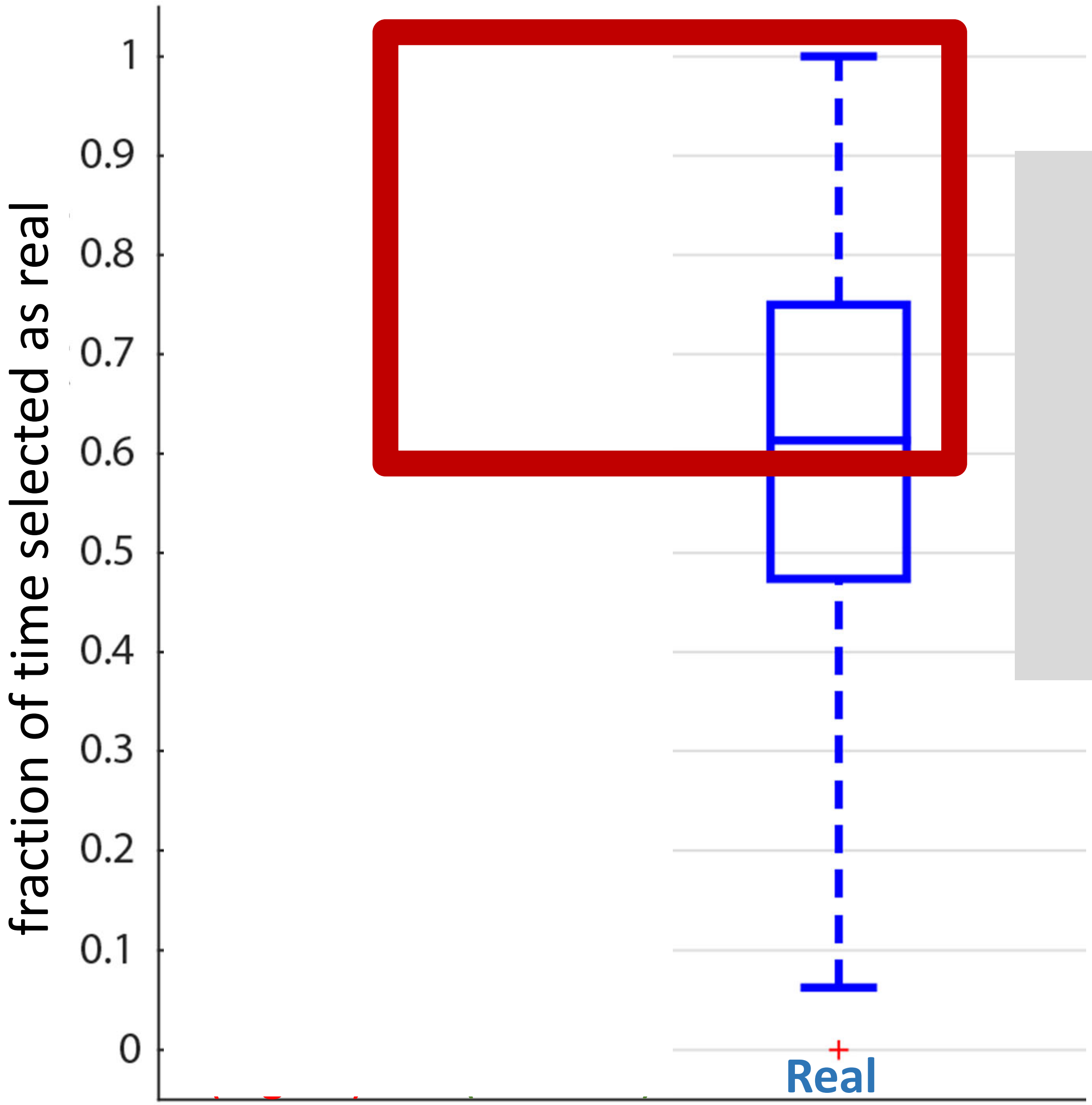

Ariel
ariel (0.9630)

Are Adversarial Eyeglasses Inconspicuous?



real / fake
real / fake
real / fake
…

Are Adversarial Eyeglasses Inconspicuous?



Most realistic 10%
of physically realized eyeglasses
are more realistic
than average real eyeglasses

Carnegie Mellon University CyLab Security and Privacy Institute

Can *an Attacker* Fool ML Classifiers? (Attempt #2)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects in a limited way

Can't control camera position, lighting

Defender / beholder doesn't notice attack
(to be measured by user study)

Considering Camera Position, Lighting

- Used algorithm to measure pose (pitch, roll, yaw)
- Mixed-effects logistic regression
  - Each 1° of pitch = 0.94x (VGG) or 1.12x (OpenFace) attack success rate
  - Each 1° of yaw = 0.94x attack success rate
- Varied luminance (add 150W incandescent light at 45°, 5 luminance levels)
  - Not included in training → 50% degradation in attack success
  - Included in training → no degradation in attack success

Carnegie Mellon University
CyLab Security and Privacy Institute

What if Defenses Are in Place?

- Already:
  - Augmentation to make face recognition more robust to eyeglasses
- New:
  - Train attack detector (Metzen et al. 2017)
    - 100% recall and 100% precision
  - Attack must fool original DNN and detector

- Result (digital environment): attack success unchanged

Can *an Attacker* Fool ML Classifiers? (Attempt #2)

Fooling face recognition (e.g., for surveillance, access control)

- What is the attack scenario?
- Does scenario have constraints?
  - On how attacker can manipulate input?
  - On what the changed input can look like?

Can change physical objects in a limited way ✓

Can't control camera position, lighting ✓ ?

Defender / beholder doesn't notice attack
(to be measured by user study) ✓

Carnegie Mellon University CyLab Security and Privacy Institute

Other Attack Scenarios?

Dodging: One pair of eyeglasses, many attackers?

Change to training process:

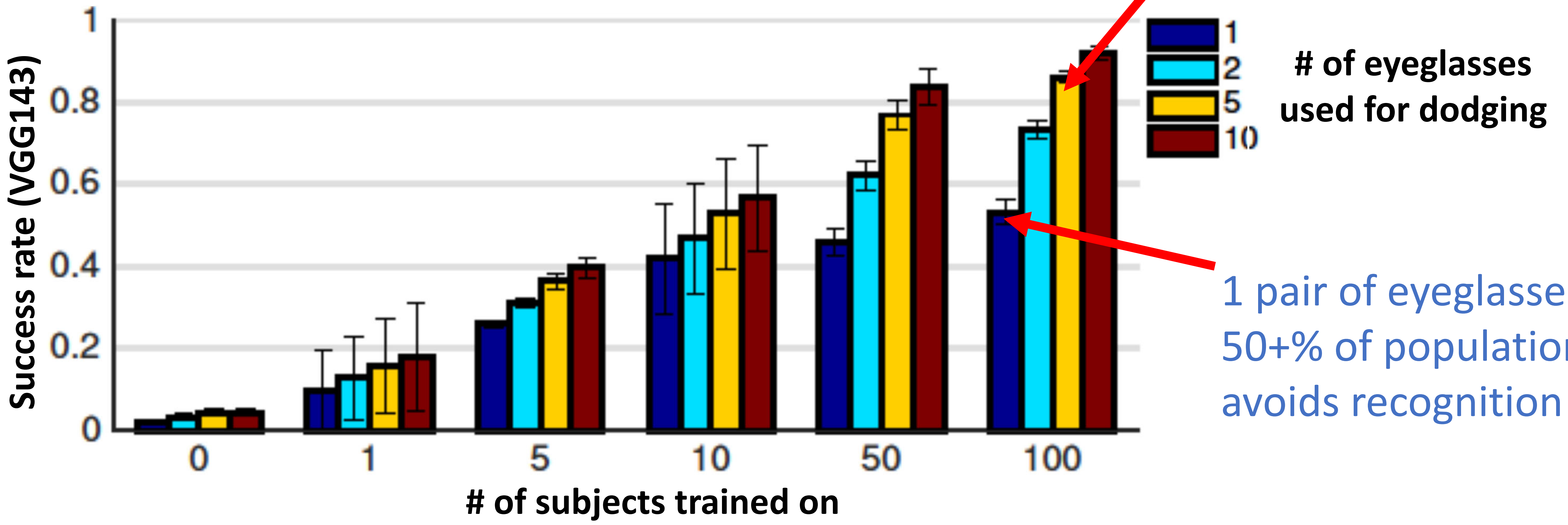
Train with multiple images of one user
$\longrightarrow$ train with multiple images of *many* users

Create multiple eyeglasses, test with large population

Carnegie Mellon University CyLab Security and Privacy Institute

Other Attack Scenarios?

Dodging: One pair of eyeglasses, many attackers?



5 pairs of eyeglasses, 85+% of population avoids recognition

1 pair of eyeglasses, 50+% of population avoids recognition

Other Attack Scenarios?
or Defense

Privacy protection?

- E.g., against mass surveillance at a political protest

Unhappy speculation: probably not

- 90% of video frames successfully misclassified
  → 100% success at defeating laptop face logon
  → 0% at avoiding being recognized at a political protest

Exception: "privacy" through denial of service

- To preserve privacy, be "identified" in many locations at once

Carnegie Mellon University CyLab Security and Privacy Institute

Fooling ML Classifiers: Summary and Takeaways

- "Attacks" may not be meaningful until we fix context
  - E.g., for face recognition:
    - Attacker: physically realized (i.e., constrained) attack
    - Defender / observer: attack isn't noticed as such
- Even in a practical (constrained) context, real attacks exist
  - Relatively robust, inconspicuous; high success rates
- Hard-to-formalize constraints can be captured by a DNN
- Similar principles about constrained context apply to other domains: e.g., malware, spam detection

For more: www.ece.cmu.edu/~lbauer/proj/advml.php

Carnegie Mellon University CyLab Security and Privacy Institute

Comparing Hypothetical and Realistic Privacy Valuations

Joshua Tan, Mahmood Sharif, Sruti Bhagavatula, Matthias Beckerle, Michelle L. Mazurek*, Lujo Bauer

Carnegie Mellon University

* University of Maryland

Why measure privacy preferences?

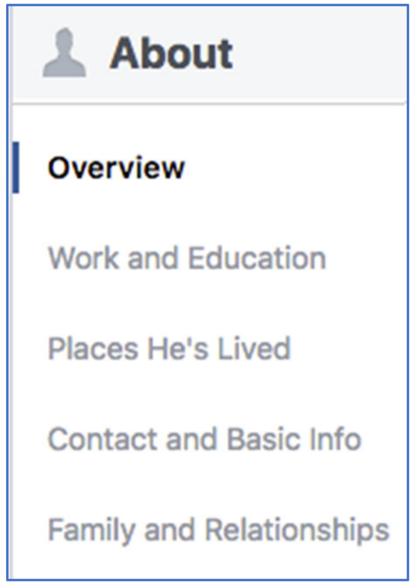
- Privacy preferences = willingness/comfort sharing personal info
- Who benefits from understanding privacy preferences?
  - System designers
    - What data are users okay sharing?
    - How much value should users receive for sharing?
  - Policy makers
    - How much “loss” do consumers incur through data breaches?
    - What kind of data sharing (if any) should be disincentivized?

Measuring privacy preferences is challenging

- Contextual factors influence users' privacy preferences and behaviors
  - E.g., willingness to share PII depends on how it will be used
- Valuations of goods (estimations of worth) influenced by framing effects and cognitive biases
  - Endowment effect = value more if own / value less if shared
  - Hypothetical bias = overestimate value in hypothetical scenario
- Stated privacy attitudes often do not align with actual behavior (privacy paradox)
- In this talk, privacy preferences are measured in $ valuations

This talk: Can we predict privacy valuations?

- Privacy valuation = willingness to sell and selling price for personal info
- How do privacy valuations depend on combinations of factors?



Attribute type

Receiving party

Scenario realism

- Does hypothetical bias explain the privacy paradox?

Methodology

- Online study with 434 Prolific participants
- Participants asked to assign selling prices to personal attributes
  - Could also choose to not sell
  - Selling scenario was information marketplace operated by CMU
  - Attributes in market are sold to buyers via an auction
  - Buyers have limited budgets and purchase lowest-priced offers first
- Collected demographics and IUIPC scores

Prices assigned to 7 attributes and 6 parties

Attributes:

- Age
- Email address
- Gender
- Relationship status
- Home address
- Occupation
- Phone number

Receiving parties:

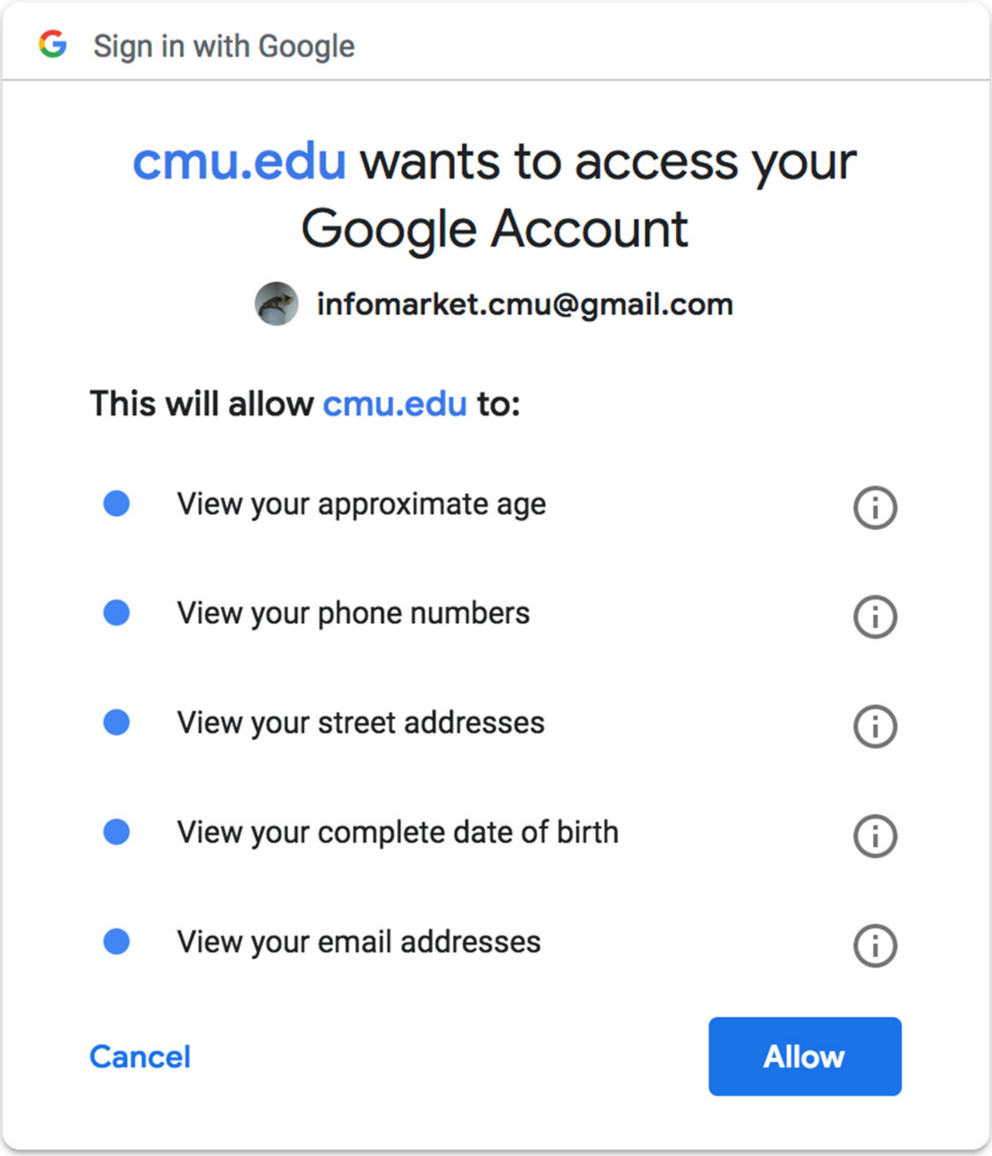
- Ad networks
- Federal agencies
- Insurance companies
- Market research companies
- Political parties
- Research pools

We varied the realism of the scenario

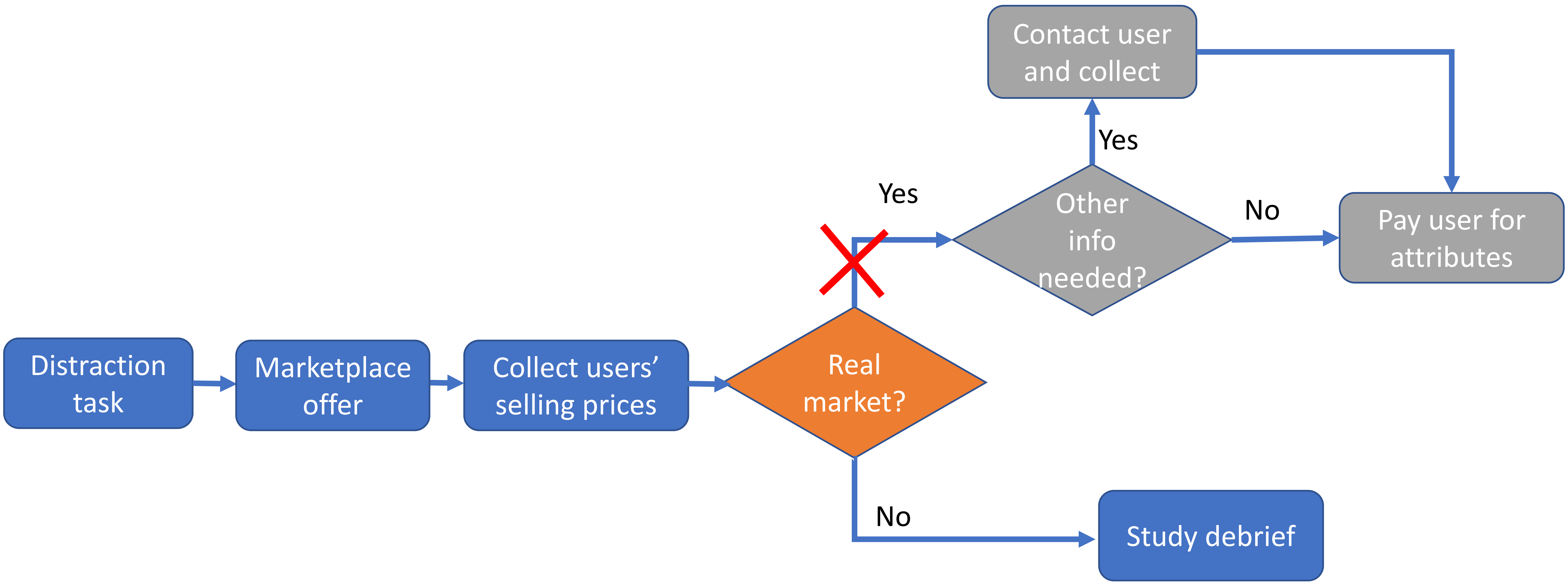
More realistic

- **Realistic with endowment (Real_{End})**
  Google SSO *before* valuations (functional market)
- **Realistic without endowment ($\text{Real}_{\text{NoEnd}}$)**
  Google SSO *after* valuations
- **Less realistic (Hyp_{Low})**
  Evaluate near-operational market
- **Even less realistic ($\text{Hyp}_{\text{Medium}}$)**
  Evaluate market concept
- **Least realistic (Hyp_{High})**
  Participate in research on buying/selling preferences

Less realistic



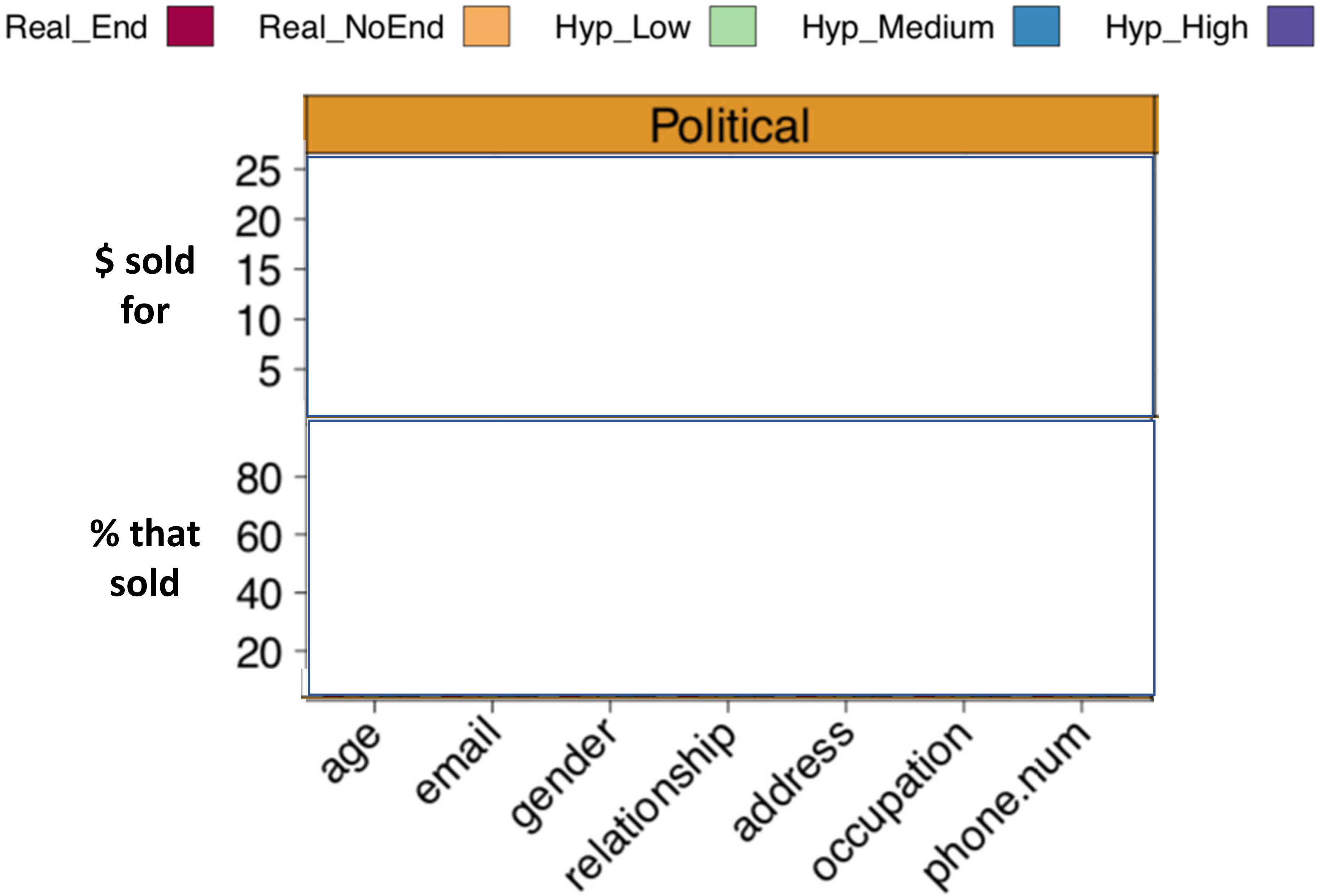
Marketplace realistic except for payment



Valuations analyzed using regressions and ML

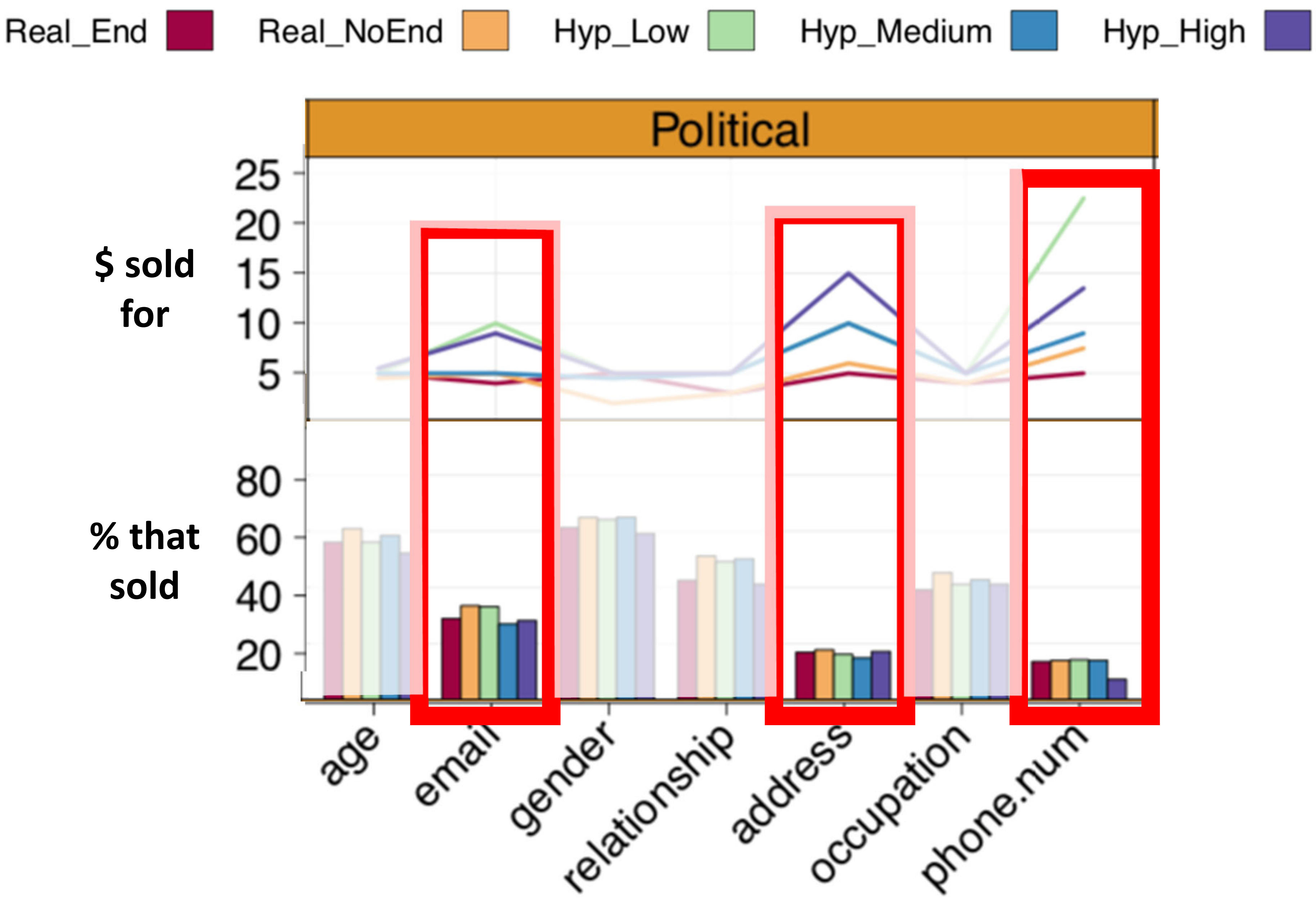
- Likelihood of selling
  - Mixed-effect logistic regression
- Dollar values
  - Mixed-effect linear regression
- Modeled two-way interactions between scenario realism, attribute type, and receiving party
  - Applied Holm-Bonferonni correction to significance tests
- Predictions of attribute rankings
  - Machine learning classifier

Comparing privacy valuations: Results

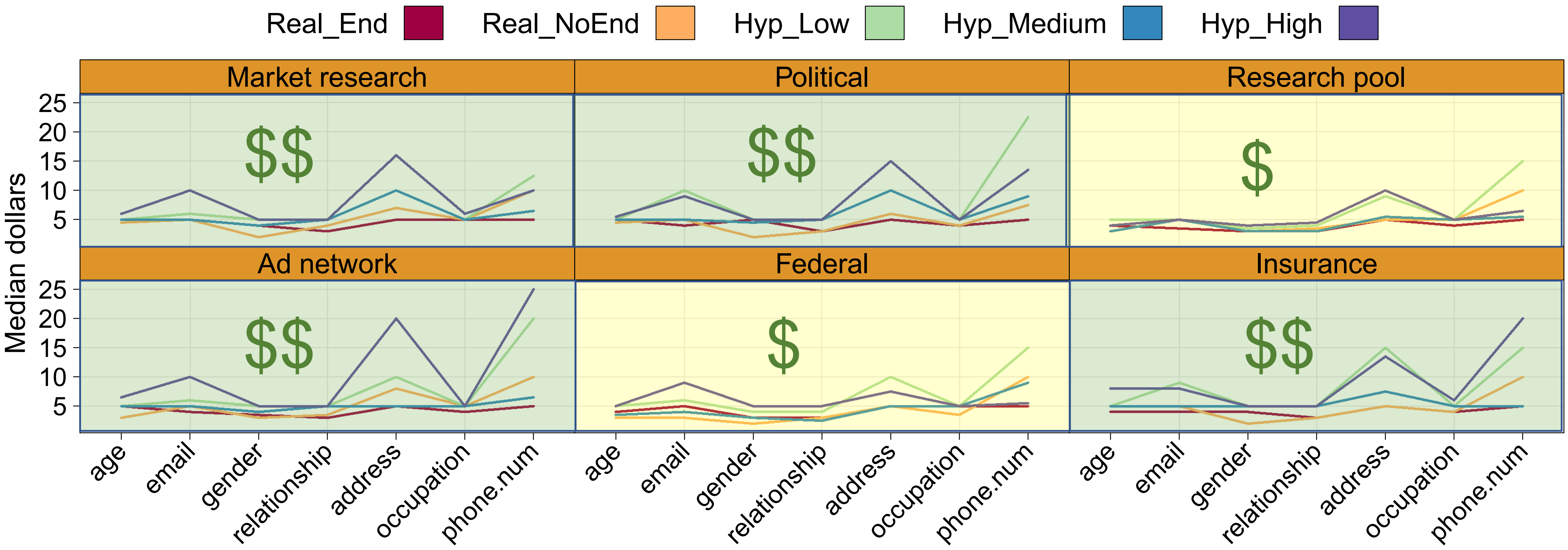


Real_End
Real_NoEnd
Hyp_Low
Hyp_Medium
Hyp_High
Political
$ sold for
25
20
15
10
5
% that sold
80
60
40
20
age
email
gender
relationship
address
occupation
phone.num

Contact info sold for more $ and less often



Selling price depends on who is buying



Can we predict valuations?

- Dollar values? Not yet.
- Scenario realism, attribute type, and receiving party insufficient for accurate prediction of absolute valuations
  - Conditional R^2 = 74.8%
  - Marginal R^2 = 13.3%
- Individual users have very different baselines in terms of $
  - Given baseline, accurate $ prediction possible

Carnegie Mellon University CyLab Security and Privacy Institute

Can we predict valuations?

- Attribute rankings? Yes.
  - Same average rankings regardless of scenario realism or buyer
- Subset of attribute rankings for hypothetical scenario further improves prediction of full rankings in realistic scenario
  - E.g., by asking a user to rank three attributes, can predict full rankings more accurately than if used average rankings

Privacy paradox often doesn't hold

- Surprisingly, *Hypothetical* values not generally different than *Realistic* values
  - Exceptions:
    - Phone number ($Real_{End}$: ~\$9, $Real_{NoEndow}$: ~\$14)
    - Home address ($Real_{End}$: ~\$8, $Real_{NoEndow}$: ~\$11)
- Calibration factor = Hypothetical / Real
  - Largest calibration factor predicted by our model was 1.61
  - List and Gallet (2001): 4.44 for public goods, 8.41 for private goods
- No significant differences in likelihood of selling by scenario realism

Carnegie Mellon University CyLab Security and Privacy Institute

Comparing privacy valuations: Takeaways

- Attribute rankings stable regardless of scenario realism and receiving party
- Selling prices can be accurately predicted based on attribute type and receiving party, given baseline price for individual person
- In contrast to other types of goods, privacy valuations not generally affected by hypothetical bias
  - Some attribute types (e.g., contact info) may not be exempt
- Privacy paradox not attributable to hypothetical bias

Joshua Tan, Mahmood Sharif, Sruti Bhagavatula,
Matthias Beckerle, Michelle L. Mazurek*, Lujo Bauer

Carnegie Mellon University CyLab Security and Privacy Institute

Carnegie Mellon University

* University of Maryland