Humanizing AI: Filling the Gaps with Multi-Faceted Research

Joel W. Branch, Ph.D.

Chief Product Officer, Lucd

Introduction

- Problem
  - AI is permeating our society with profound impact, at an astoundingly fast rate
  - Many AI deployments have been hastily developed, with little thought of real-world consequences
- What do I mean by *humanizing* AI?
  - Creating governance frameworks, involving a broad array of research competencies, for democratizing the development of safe, effective, and game-changing AI solutions
  - *Not* creating pleasantly convincing human-like interfaces, e.g., voice assistants (though techniques can be extended)

Agenda

- Short summary on the evolution of AI
- AI deployment challenges
- Humanizing AI with a fresh approach to governance

Evolution of AI

Short history of AI milestones

IBM's Deep Blue beats chess champion Gary Kasparov

IBM's Watson defeats former Jeopardy champions

OpenAI's GPT-3 exhibits language capabilities nearly indistinguishable from humans'

1997 | 2000 | 2011 | 2016 | 2020

Honda's ASIMO robot displays incredible human-like action

Google DeepMind's AlphaGo defeats Go champion Lee Sedol

Increasing AI commoditization (APIs)

Trends in AI value and applications

- AI Augmentation predicted to create $2.9T of business value in 2021[1]
- AI engineering a top tech strategy: focusing on operationalization for business[2]
- Increasingly democratized AI development via no/low-code frameworks (will permeate AI engineering)

AI deployment challenges

Why enterprise AI efforts are failing

- Some stats[3, 4]…
  - 7 out of 10 companies report little to no impact from AI projects
  - 40% of companies that made significant investments in AI have yet to report gains
  - 87% of data science projects do not make it to production

Why enterprise AI efforts are failing

- Major adoption challenges
  - AI bias
    - Fueled by over-technically driven solution development
  - Black box decision-making
    - Lack of transparency and rationale in AI output
  - Disparate hard-to-access data
    - Approx. 50% of AI development spent on data access and cleaning[5]

Notable examples[6]

- Race and gender bias in job recruitment software
- Race bias in online ads
- Race bias in facial recognition software

AI bias in depth

- *Gender Shades*[7] project example
  - Project evaluates the accuracy of popular AI-powered gender classification products
  - Contributed new benchmark image dataset attempting a balance among gender and skin types
  - Exposure of biased/insensitive performance resulted in internal reviews within Microsoft and IBM

Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

Gender Classifier	Darker Subjects Accuracy	Lighter Subjects Accuracy	Error Rate Diff.
Microsoft	87.1%	99.3%	12.2%
FACE++	83.5%	95.3%	11.8%
IBM	77.6%	96.8%	19.2%

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Why enterprise AI efforts are failing

- Insufficient organizational support
  - Executives do not champion (and fund) AI-based strategies
  - AI is limited to "projects," and/or isolated within "innovation labs"
  - AI development teams are nearly exclusively comprised AI and data –related talent
  - Limited supply of AI and data science talent

Why enterprise AI efforts are failing

- Immature development process

Deployed system ← AI model development ← Data collection

Deployed system

- Unfair
- Inaccurate
- Non-competitive, non-trustworthy solutions


AI model development

- No explainability
- No governance
- No multi-disciplinary involvement

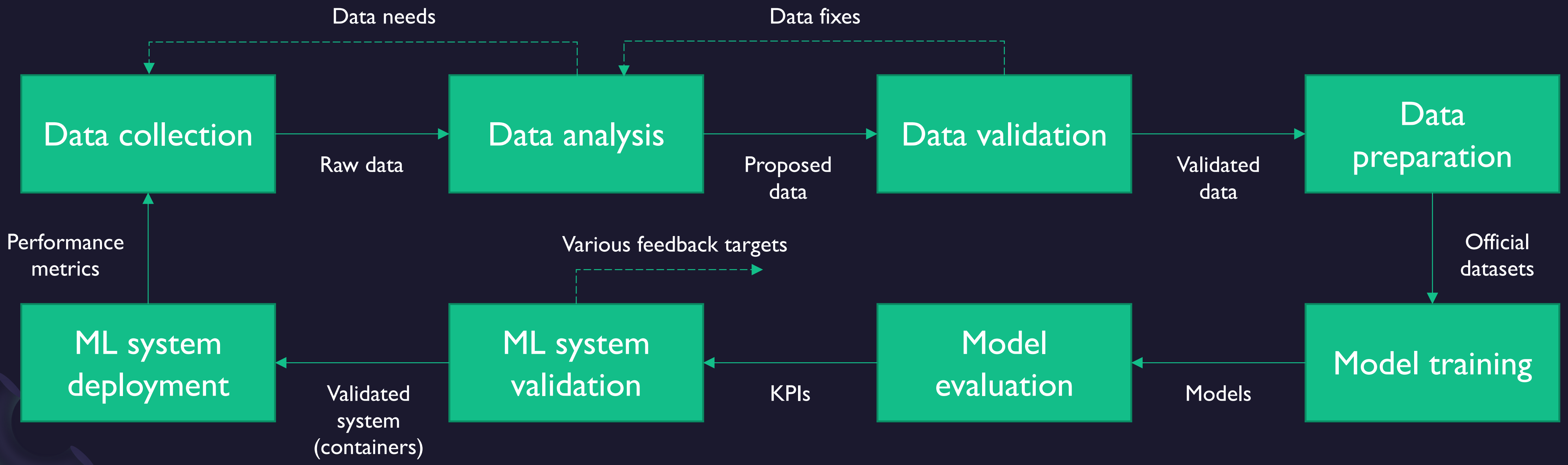
Data collection

- Careless data collection
- Biased data
- Insufficient data

Humanizing AI

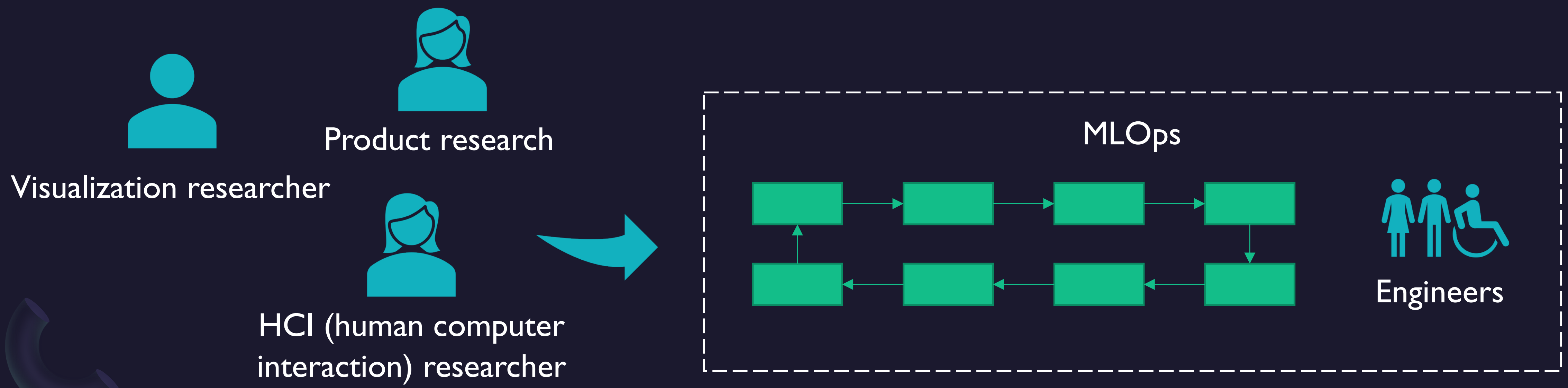
Improving the AI lifecycle with MLOps

- MLOps
  - Disciplined enterprise process (modeled off DevOps)
  - Traditionally ML engineers (data scientists, data engineers, ML engineers, SW engineers)



Improve MLOps to enable “humanization”

- How?
  - Expand MLOps to broaden participation among different research competencies, to ***democratize*** the development of ***safe***, ***effective***, and ***game-changing*** AI solutions



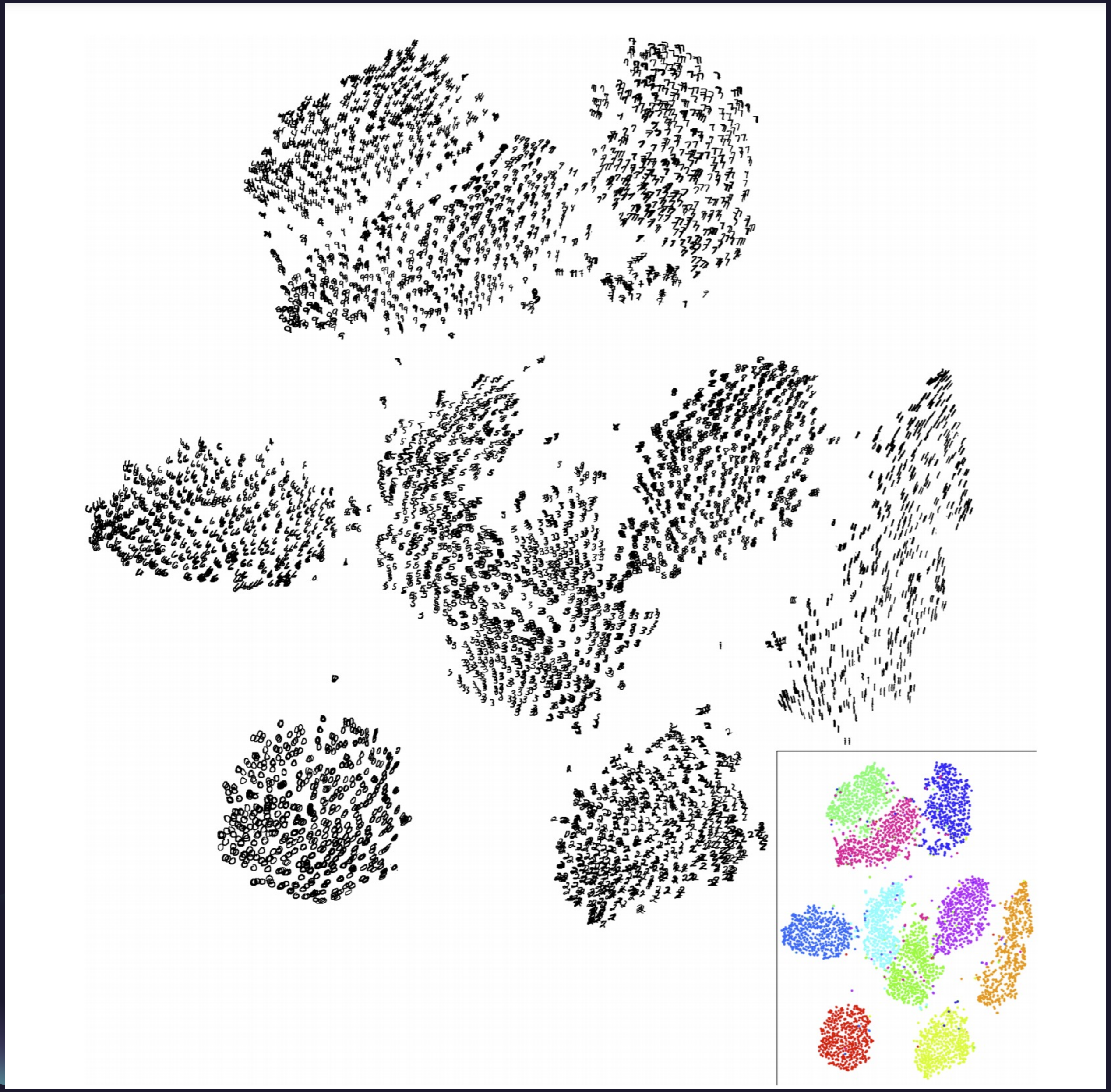
Visualization research

- Some relevant AI challenges
  - Exploding data volume and dimensionality limit interpretability in *data analysis* and *validation* phases
  - *Intersectional bias* (hardest to identify, most prevalent) limits *data collection* through *validation* phases
    - E.g., determining if data has bias toward (a) *tall*, (b) *males*, and (c) *blonde hair*, as opposed to just *men*

Example Solutions	**Details**	**Humanization**
Latent space (compressed dimension) data exploration	Need new ways to intuitively visualize distinct data groupings and relationships among them[8]	Helps stakeholders w/ little domain knowledge assess semantic data features
Data subgroup performance analysis	Identification and vis. of subgroups for comparative AI perf. analysis; needs collaboration with data scientists[9]	Helps reduce bias in deployed models; helps non-tech stakeholders participate in bias detection

Visualization research

Example visualization of latent space data



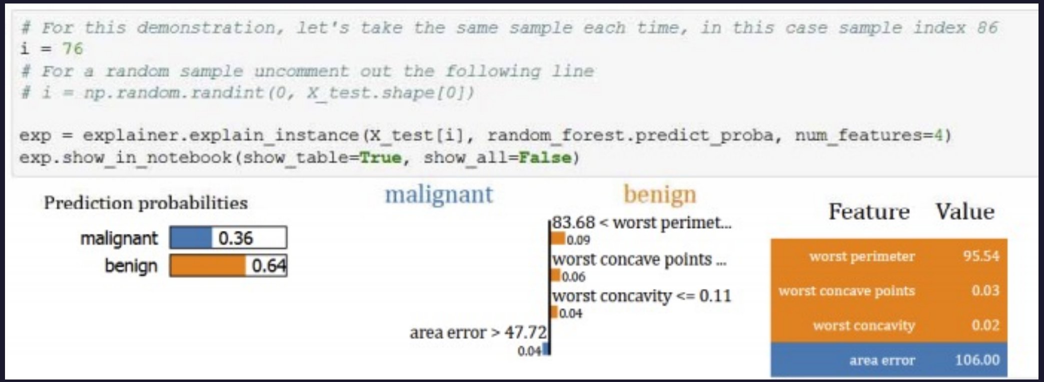
HCI research

- Some relevant AI challenges
  - (Deep learning) AI model explainability limit trust in *model evaluation* through *deployment* phases
  - Scaling up AI deployments will require automating parts of the MLOps workflow, a process which is still premature and requires AI domain knowledge

Example Solutions	**Details**	**Humanization**
Decision and structure –based AI model explainability	Need new ways to express decision rationale using causality and natural language techniques (where possible); needs collaboration w/ ML engineers	Helps non-tech MLOps stakeholders participate in model evaluation; increases trust and engagement among end-users
Opensource frameworks for MLOps	Need new policy languages to express automation and governance rules to make MLOps easier to use; collaboration with ML engineers	Innovates and simplifies frameworks for expressing and enforcing AI ethics

HCI (and product) research

From AI explainability to causality



Feature-based decision explanation[10]

Causality (human)

Explainability interface framework

Explainability (machine)

Metrics of explainability (hard, open research)

- “Goodness” of explanations
- “Satisfaction” of the user
- “Understandability”
- “Trust”

Potential techniques

- Integrate AI explanations with knowledge graphs to express “evidence” of decisions
- Augment training data w/ meaningful human explanations[11]

Summary

- Humanizing AI
  - Necessary for safe, enjoyable, competitive AI solutions
  - Importance will increase as (governmental) AI ethics policies start to mature
  - Many open problems still exist, and require cross-discipline collaboration in among academia and industry

Thank You

References

1. J-D Lovelock *et al.* "Forecast: The Busines Value of Artificial Intelligence, Worldwide, 2017-2025." Gartner, 2018.
2. E. Brethenoux. "Top Strategic Technology Trends for 2021: AI Engineering." Gartner, 2021.
3. J. Vanian. "Why Most Companies Are Failing at Artificial Intelligence: Eye on A.I." Fortune, 2019.
4. "Why do 87% of Data Science Projects Never Make it into Production." VentureBeat, 2019. https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/.
5. "2020 State of Data Science." Anaconda, 2020.
6. N. T. Lee *et al.* "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms." Brookings, 2019.
7. J. Buolamwini and T. Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Conference on Fairness, Accountability, and Transparency, 2018.
8. Y Liu *et al.* "Latent Space Cartography: Visual Analysis of Vector Space Embeddings." EuroVis, 2019.
9. A. A. Cabrea *et al.* "FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning." IEEE VAST, 2019.
10. M. T. Ribeiro *et al.* "Why Should I Trust You?: Explaining the Predictions of Any Classifier." ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining, 2016.
11. N. C. F. Codella *et al.* "Teaching Meaningful Explanations." 2018.