

Analysis of Robustness of a Large Game Corpus

Mahsa Bazzaz, Seth Cooper
Northeastern University







MOTIVATION

- Formalize the characteristics of game level data
 - Hard constraints in structured data
 - Sensitivity to input change
- Introduce a large dataset of 2D tile-based game levels

HARD CONSTRAINTS IN STRUCTURED DATA

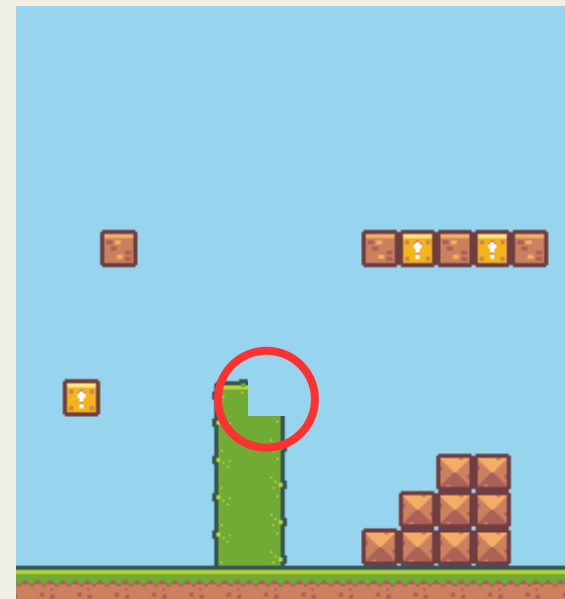
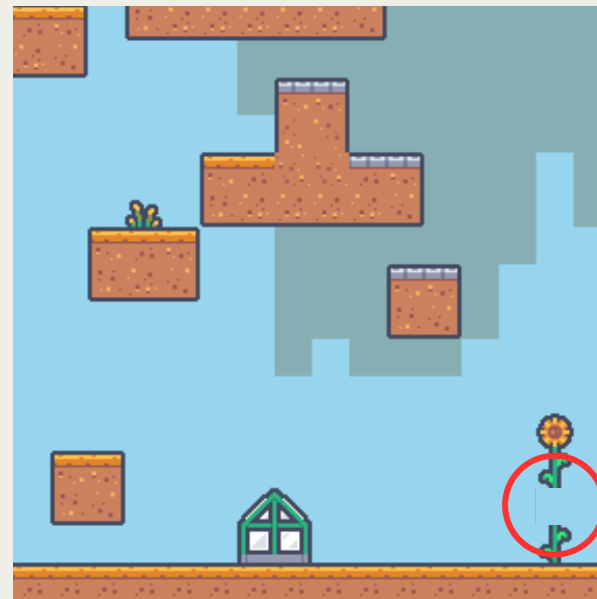
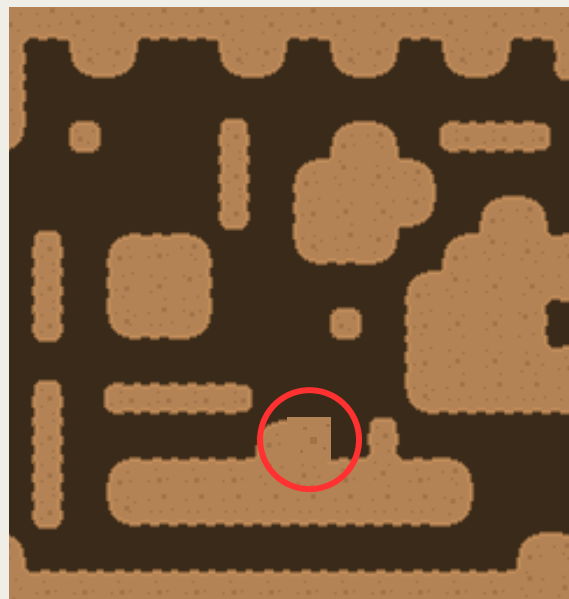
- Hard constraints impose mandatory conditions that the variables must satisfy.
In contrast, soft constraints allow some flexibility.
- hard **global** constraints:
 - **solvability**: A solvable game level is one which is possible for the player to complete.
 - having keys and corresponding doors or treasures
 - incorporating a specific number of enemies or obstacles at varying difficulty levels.
- hard **local** constraints: local structures that have multiple pieces and can be missed or arranged incorrectly

Dataset	Tiles
cave	
platform	
crates	
vertical	

ACCEPTABILITY VS SOLVABILITY

- **Solvability:** The key global constraint. A solvable game level is one which is possible for the player to complete
- **Acceptability:** In addition to global constraints, there are also local constraints in game levels that need to be satisfied to produce acceptable levels.

solvable **and** acceptable levels can be directly used
while unsolvable **or** unacceptable levels cannot.



RELATED WORK

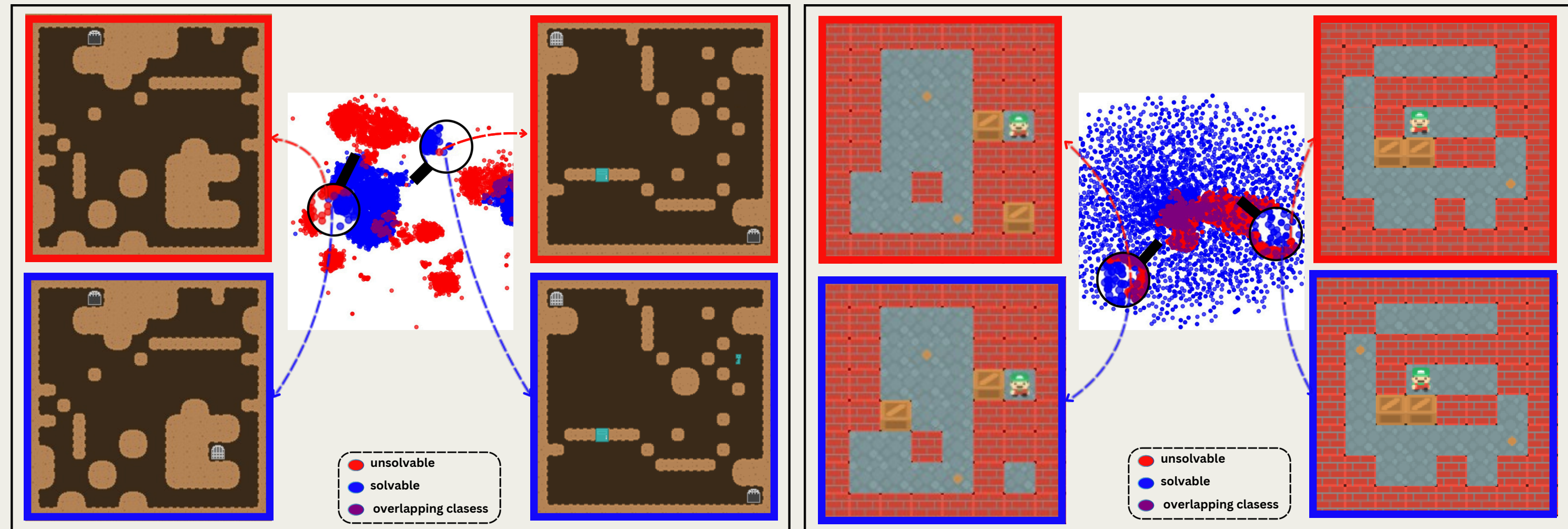
- Research by Sturtevant et al. has previously explored how minor modifications to existing levels can significantly increase the complexity and solution length of puzzles that may not be easily anticipated by human designers.
- Their quantitative analysis and user study revealed that even small incremental design adjustments can greatly affect the overall experience of playing a level.
- The results of this prior user study that showcases this phenomenon were a major inspiration for this work.

- Nathan Sturtevant, Nicolas Decroocq, Aaron Tripodi, and Matthew Guzdial. 2020. The Unexpected Consequence of Incremental Design Changes. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment 16, 1 (Oct. 2020), 130–136. <https://doi.org/10.1609/aiide.v16i1.7421>

SENSITIVITY TO INPUT CHANGE

Small changes in the input of structured data can cause noticeable changes in the output.

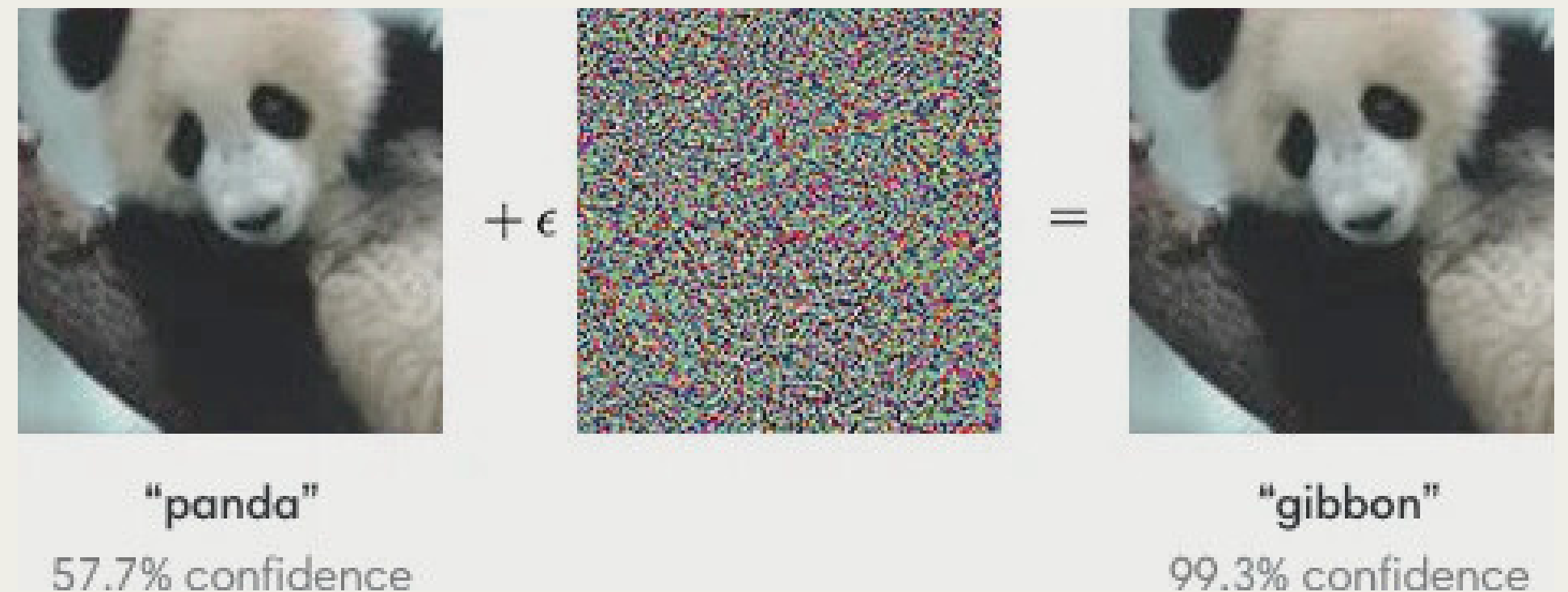
In game levels, solvable and unsolvable levels can be only different by a single tile change or swap.



ROBUSTNESS OF CLASSIFIERS

Captures how consistently the classifier behaves under small perturbations of its inputs. It is formally defined as the probability that any two points x and x' drawn from distribution \mathcal{D} , which lie within a distance r of each other, receive the same predicted label from classifier f .

$$A_r(f, \mathcal{D}) = \mathbb{P}_{x \sim \mathcal{D}}[f(x) = f(x') | \forall x', d(x, x') \leq r]$$



- Robi Bhattacharjee and Kamalika Chaudhuri. 2020. When are Non-Parametric Methods Robust?. In Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119), Hal Daumé III and Aarti Singh (Eds.). PMLR, 832–841.
- <https://openai.com/index/attacking-machine-learning-with-adversarial-examples/>

ROBUSTNESS OF DATA

In this context, we are interested in quantifying the robustness of the **data**.

We replace $f(x)$ by the label of x .

$$\mathbf{D}_r(\mathcal{D}) = \mathbb{P}_{x \sim \mathcal{D}}[\text{Label of } x = \text{Label of } x' | \forall x', d(x, x') \leq r]$$

For a given radius r , the value of $\mathbf{D}_r(\mathcal{D})$ represents how sensitive the data is to changes in the input.

Similarly, the opposite --**non-robustness**-- can be formally defined as:

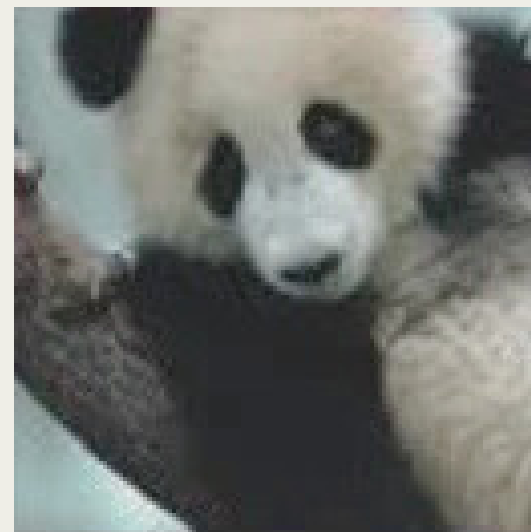
$$\mathbf{ND}_r(\mathcal{D}) = \mathbb{P}_{x \sim \mathcal{D}}[\text{Label of } x \neq \text{Label of } x' | \exists x', d(x, x') \leq r]$$

and by computing $\mathbf{ND}_r(\mathcal{D})$ we can understand how far from being robust the data is.

ROBUSTNESS OF DATA

The important difference between the robustness of data and the robustness of the classifier in adversarial training:

the **true label** of data and the **predicted label of data** when using a classifier.

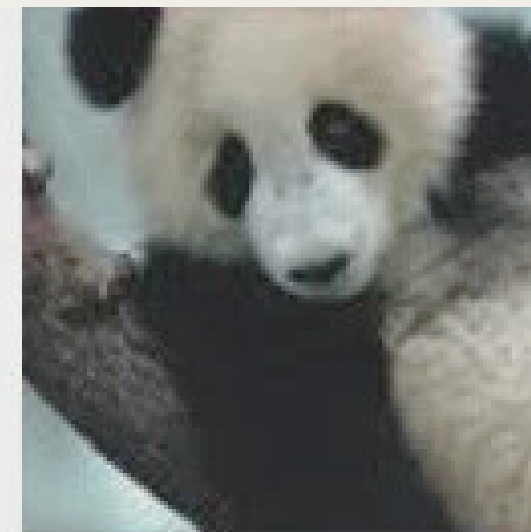


panda

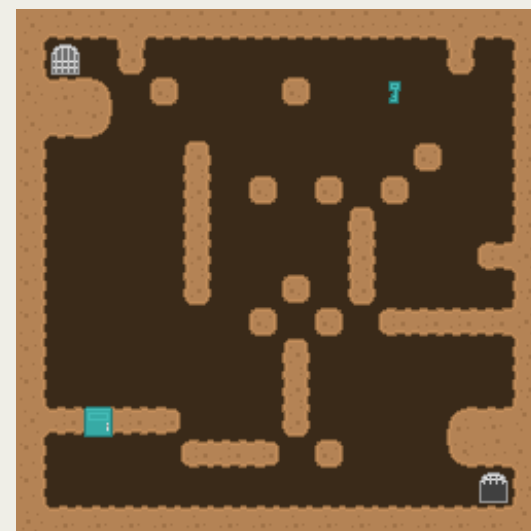
+ ϵ



=



noisy but still panda

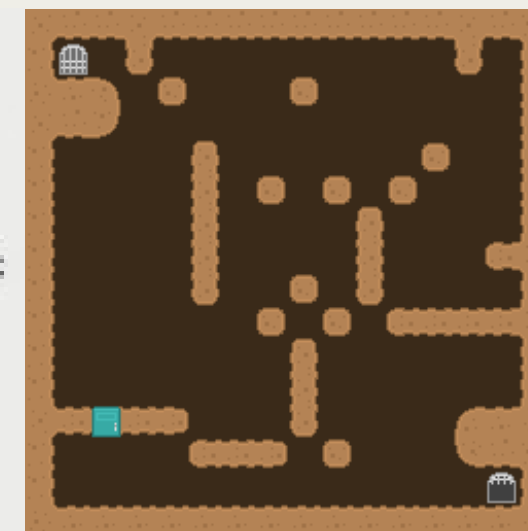


solvable level

+ ϵ



=

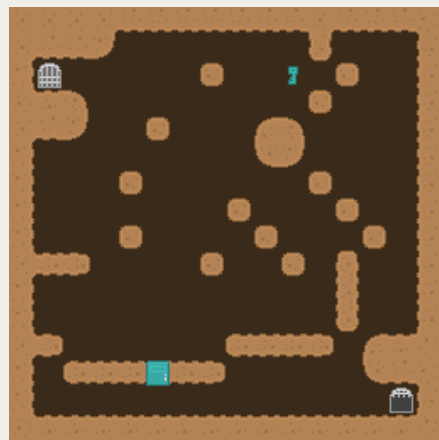


unsolvable level

DOMAIN

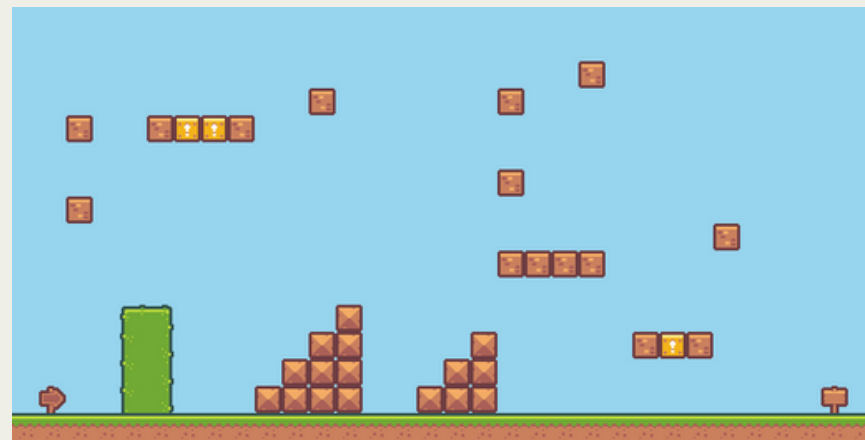
cave

top-down maze



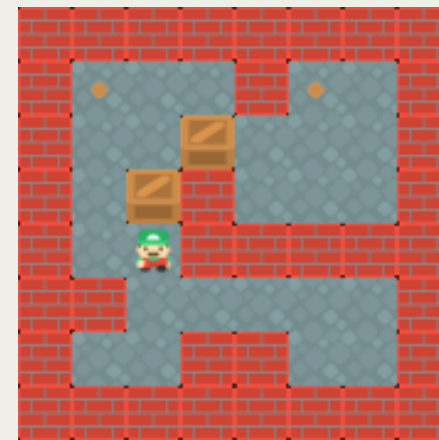
platform

platformer inspired
by Super Mario Bros.



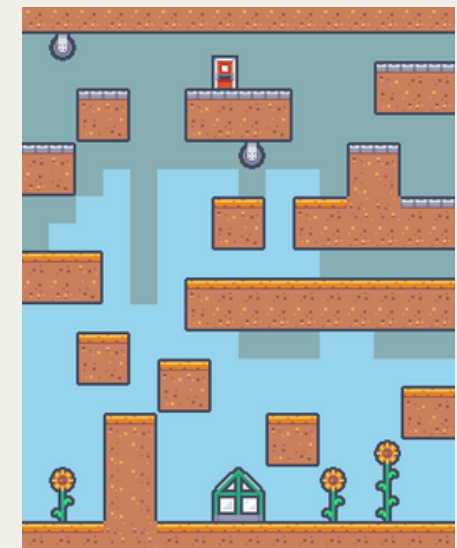
crates

puzzle inspired by
Sokoban



vertical

platformer inspired
by Super Cat Tales



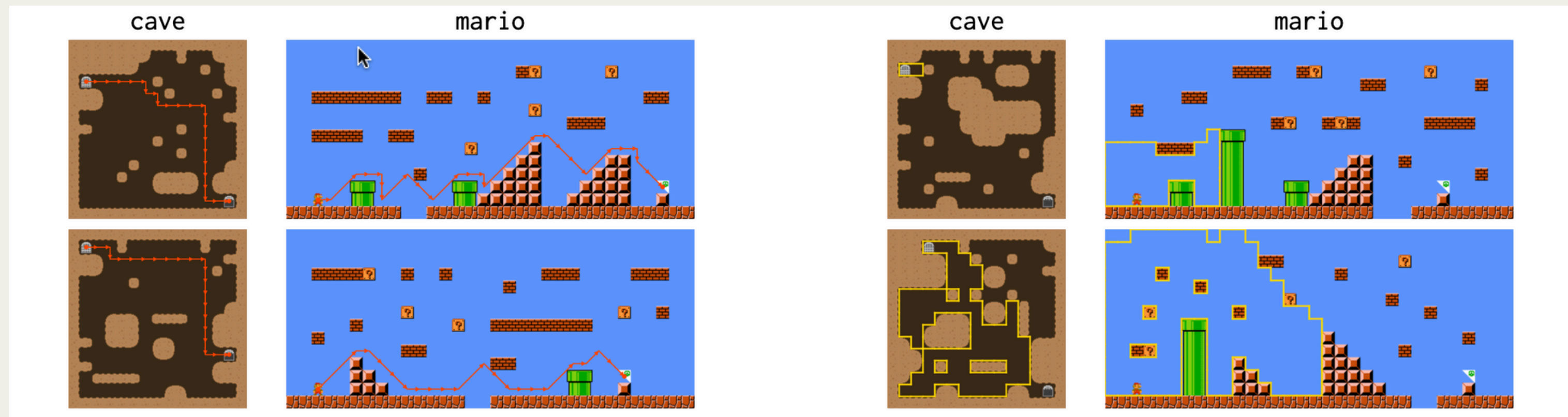
- Kenney. 2022. Free game assets. <https://www.kenney.nl/assets>.
- Nintendo. 1985. Super Mario Bros. Game [NES].
- Thinking Rabbit. 1928. Sokoban. Game.
- Neutronized. 2016. Super Cat Tales. Game [iPhone].

GENERATED GAME LEVEL CORPUS (GGLC)

- This work introduces a large dataset of **solvable** and **unsolvable** four 2D tile-based games that are created based on popular classic genres.
- Custom games are generated instead of generating the original games to ensure that the dataset can be **freely accessible to everyone without any copyright concerns**.
- The **solutions** of all solvable levels are included as part of the metadata.
- The GGLC, along with setup documentation, is publicly available on GitHub (<https://github.com/TheGGLC>).

STURGEON

- Levels are generated using the **Sturgeon** constraint-based level generator.
- The **unsolvable** levels are mainly generated using Sturgeon's unreachability constraint that uses constraints that the level's goal is not reachable from its start.



- Seth Cooper. 2022. Sturgeon: tile-based procedural level generation via learned and designed constraints. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Vol. 18. 26–36.
- Seth Cooper and Mahsa Bazzaz. 2024. Literally Unplayable: On Constraint-Based Generation of Uncompletable Levels. In Proceedings of the 19th International Conference on the Foundations of Digital Games. 1–8
- <https://github.com/crowdgames/sturgeon-pub>

MEASURING NON-ROBUSTNESS

- Distribution Non-robustness
 - measurement of Non-robustness of game levels in **discrete** form
 - distance metric: 1 tile changes
- Sample Non-robustness
 - measurement of Non-robustness of game levels in **continues** form
 - uses image embedding (CLIP) and dimensionality reduction to a 2D space (UMAP)
 - distance metric: Euclidean distance

DISTRIBUTION NON-ROBUSTNESS

Different games exhibit varying degrees of sensitivity, influenced by their gameplay mechanics and level design structures.

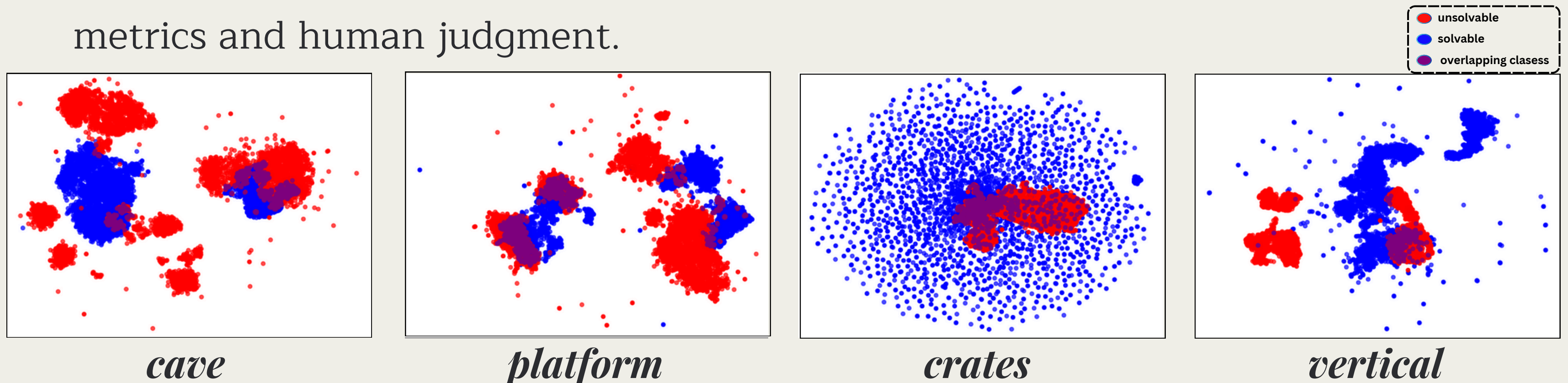
- **Crates** shows the highest sensitivity, where small changes, such as moving a box into a corner, often result in unsolvable levels due to its mechanics’ reliance on precise box placement.
- **Vertical** shows high sensitivity in terms of acceptability, as its visual design features multiple tile types with the same functionality but distinct decorative appearances.
- **Cave** shows sensitivity to input change that both results in unsolvability and unacceptability. This is inherited from the rigid framework of mazes and the versatility of tiles in this game.

	Cave	Platform	Crates	Vertical
Changed Solvability	43.1%	17.1%	78.9%	17.5%
Changed Acceptability	43.0%	0.1%	0%	42.5%

Percentage of single-tile changes that altered the solvability of the levels.

SAMPLE NON-ROBUSTNESS

- We use the level images that enable us to **compare** with the common non-structured machine learning datasets such as CIFAR-10.
- We used previous work on the alignment of similarity metrics with human judgment on tile-based video game levels to choose our embedding. Their findings indicate that **CLIP** exhibited one of the highest overall agreement between the metrics and human judgment.



- Sebastian Berns, Vanessa Volz, Laurissa Tokarchuk, Sam Snodgrass, and Christian Guckelsberger. 2024. Not All the Same: Understanding and Informing Similarity Estimation in Tile-Based Video Games. In Proceedings of the CHI Conference on Human Factors in Computing Systems. 1–23.

SAMPLE NON-ROBUSTNESS

- The results are compatible with true distribution results, indicating that the created dataset is a good representation.
- Even with inflated values for MNIST and CIFAR-10 (non-optimal embedding) some game datasets still exhibit greater sensitivity to small changes compared to

Dataset	Radius									avg
	0.00001	0.00005	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	
CIFAR-10	0.0	0.0	0.1	2.9	6.7	12.4	12.8	14.4	18.9	7.6
MNIST	0.0	0.0	0.1	1.6	3.7	5.8	6.2	6.1	10.1	3.7
Cave	0.0	0.7	2.3	24.3	29.6	32.6	32.6	35.2	36.2	21.5
Platform	0.0	0.0	0.0	2.0	5.9	12.5	14.1	21.2	28.1	9.3
Crates	0.0	0.0	0.3	2.5	6.0	13.1	15.3	20.8	25.8	9.3
Vertical	0.0	0.0	0.1	1.1	3.2	5.6	6.7	10.3	15.9	4.8

Percentage of samples with different solvability label of sample data in distribution for radius r in [0.00001, 0.1].

Values in different datasets are normalized to a common range [0, 1] using the minimum and maximums.

Thank you!

MAHSA BAZZAZ, SETH COOPER



<https://github.com/TheGGLC/>

