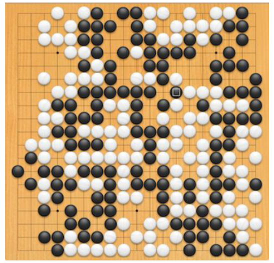
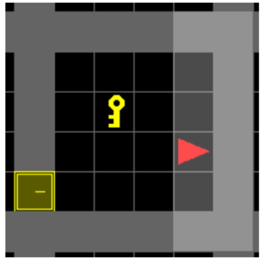
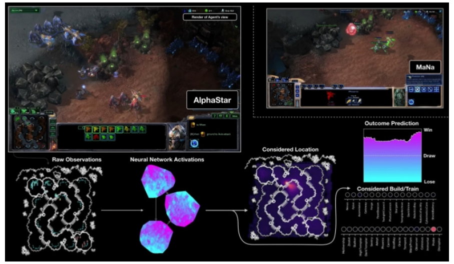
Reinforcement learning challenges for agroecology

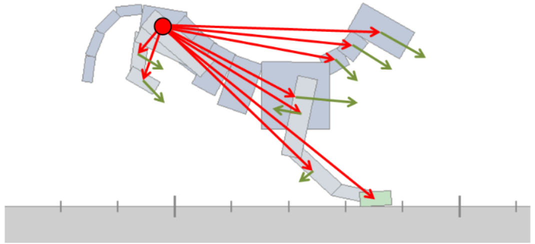
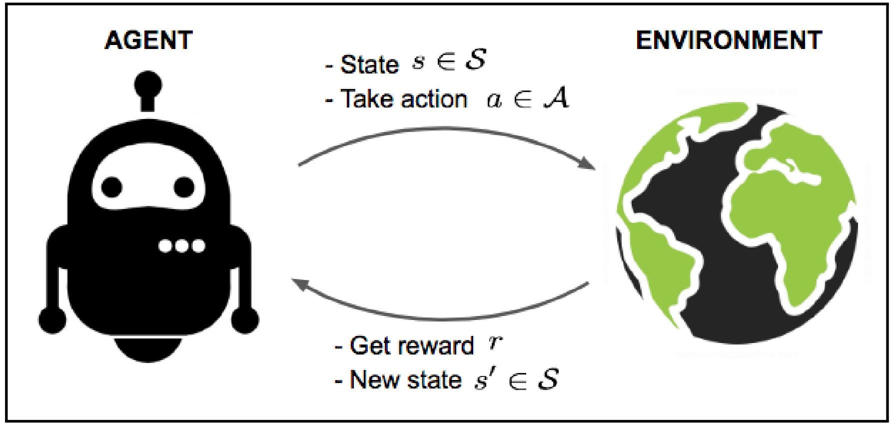
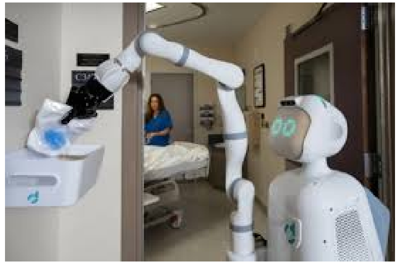
Odalric-Ambrym Maillard
Inria Lille, "SCOOL" team

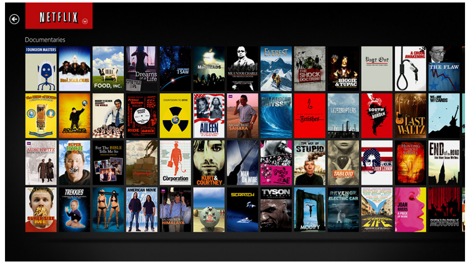
Montpellier, September 11, 2024






AlphaStar
MaNa
Outcome Prediction
Raw Observations
Neural Network Activations
Considered Location
Considered Build/Train



AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$




NETFLIX
Documentaries

Sequential COntinual and Online Learning

Reinforcement Learning Theory, Sequential Learning, "AI"

Application in Medicine / Clinical trials, Agriculture / Agroecology.

90 years ago

Foundation of Hypothesis testing :

J. Neyman, E. S. Pearson **On the problem of the most efficient tests of statistical hypotheses**. In *Philosophical Transactions of the Royal Society of London, vol 231, pp. 289–337*, 1933.

Foundation of Multi-armed bandit :

W. R. Thompson **On the likelihood that one unknown probability exceeds another in view of the evidence of two samples**. In *Biometrika, vol. 25, pp. 285–294*, 1933.

Foundation of Probability :

A. Kolgomorov **Fundamental concepts of probability**. In , 1933.

Foundation of Mathematical statistics :

Kong, W. I. **The Annals of Mathematical Statistics**. In *Ann. Math. Statist. 1 1-2.*, 1930.

1930's motivation: Agriculture, Clinical trials. $\Longrightarrow$ Today: Agroecology, Personalized medecine.

Further reading: Stigler SM **The history of statistics in 1933**. In *Statistical Science, 244-52.*, 1996.

AGRICULTURE AND AGROECOLOGY

Agriculture

- Mostly single objective, variable of interest (yield).
- Available models for variable of interests $\Longrightarrow$ Planning and Control.

Agroecology

- Diversity of objectives, practices, variables of interest. $\Longrightarrow$ Personalized, contextual
- No model available, scarce experimental data. $\Longrightarrow$ Reinforcement Learning and Bandits

CONTEXT

Stable Conditions

- SOIL: Type, Prep., Cover, etc.
- CLIMATE: T°, Sun, Rain, etc.
- USER: Tools, Worktime, etc.

EXPERIMENT: GROW BEANS

CONTEXT

Stable Conditions

- Soil: Type, Prep., Cover, etc.
- Climate: T°, Sun, Rain, etc.
- User: Tools, Worktime, etc.

POLICIES

Where to plant?

- In (plain sun) vs (morning sun) vs (evening sun).
- Near (borage) vs (tomato) vs (none) vs (both).

When to water?

- (1L per day if no rain) vs (5L per 3 days) vs (1L per 3 days until flower, then 2L per day).

$$A = 3 \times 4 \times 3$$

CONTEXT

Stable Conditions

- SOIL: Type, Prep., Cover, etc.
- CLIMATE: T°, Sun, Rain, etc.
- USER: Tools, Worktime, etc.

STOCHASTIC AgroEcoSYSTEM:
Same strategy in same context gives
Diverse outputs .

POLICIES

Where to plant?

- In (PLAIN SUN) vs (MORNING SUN) vs (EVENING SUN).
- Near (BORAGE) vs (TOMATO) vs (NONE) vs (BOTH).

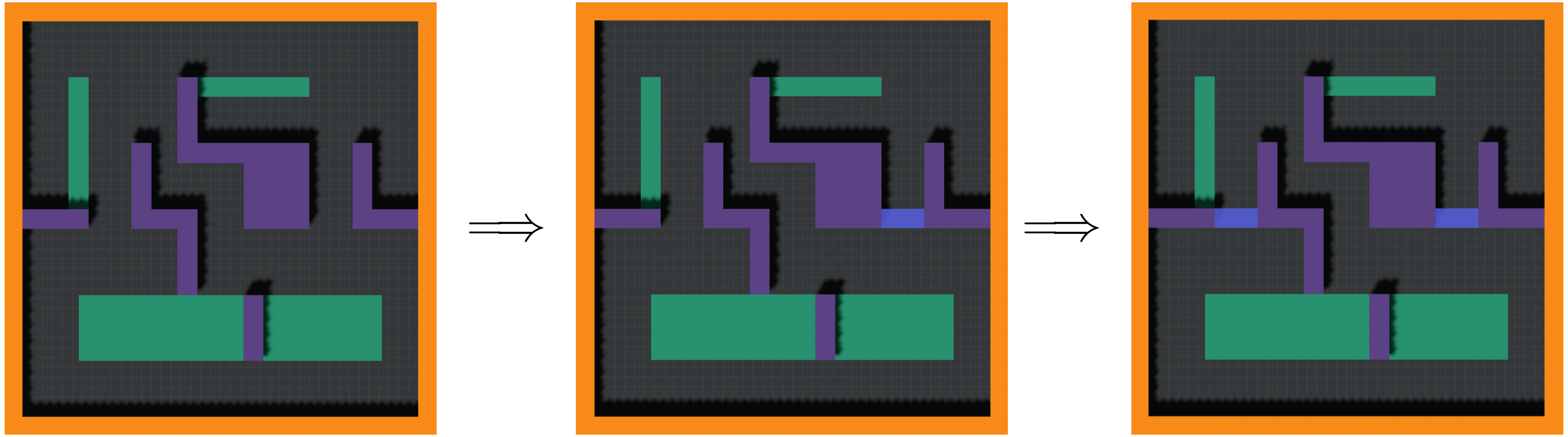
When to water?

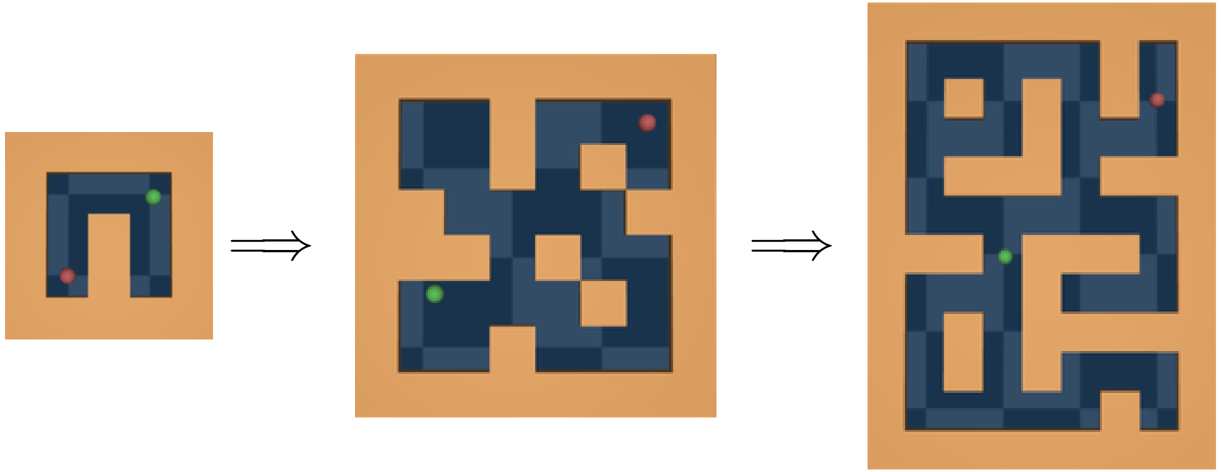
- (1L PER DAY if no rain) vs (5L PER 3 DAYS) vs (1L PER 3 DAYS until flower, then 2L PER DAY).

$A = 3 \times 4 \times 3$

Combes, R., Talebi, M. S., & Proutiere, A. **Combinatorial bandits revisited**. In *Advances in Neural Information Processing Systems 28*, 2015.

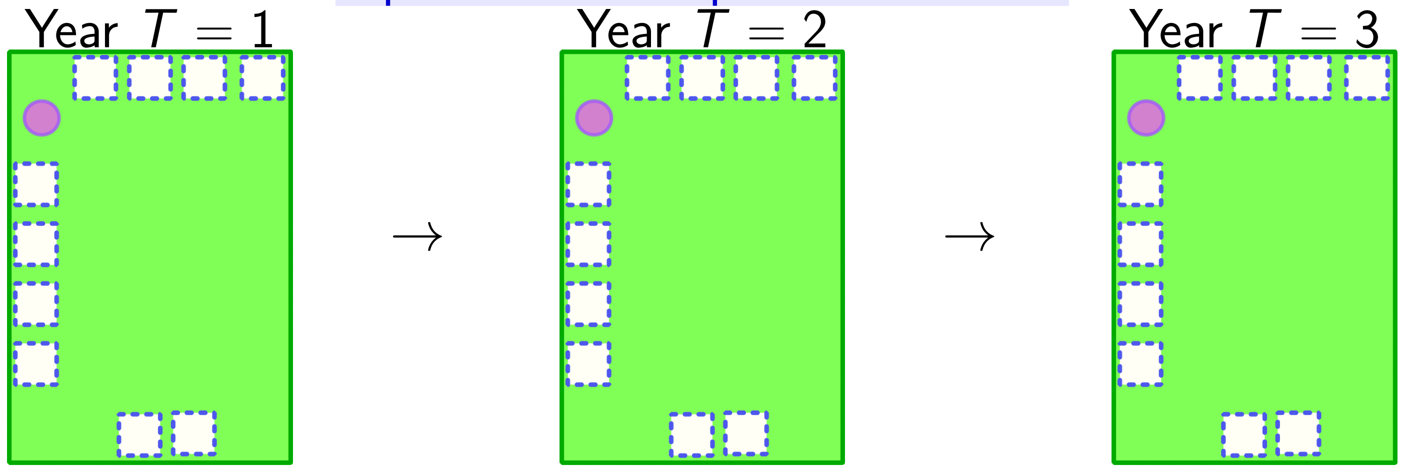
- Beyond same environment: Contextual RL, Continual RL.





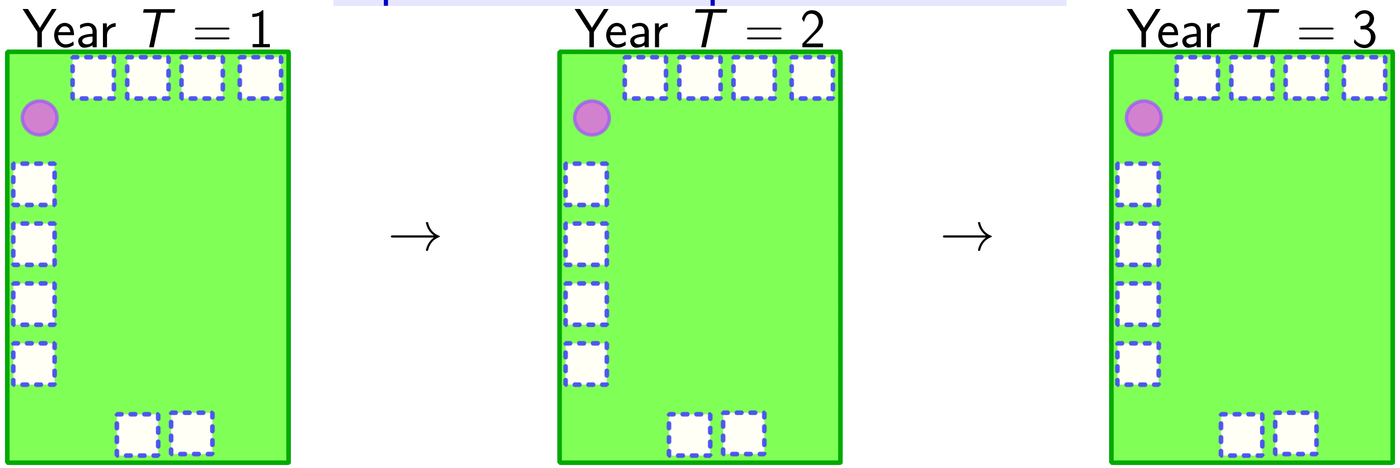
SEQUENTIAL OR GROUP EXPERIMENTS

Spatial and Temporal Allocation

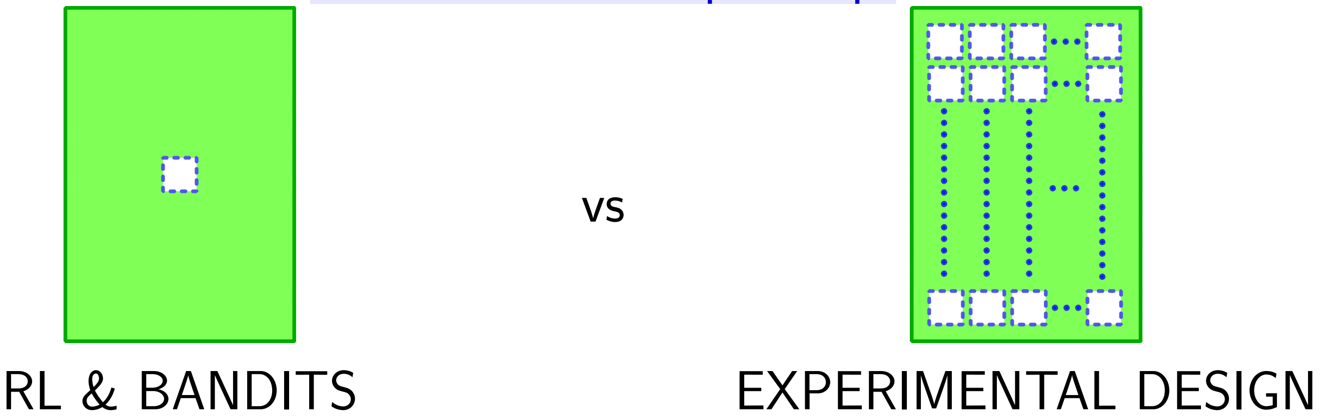


SEQUENTIAL OR GROUP EXPERIMENTS

Spatial and Temporal Allocation

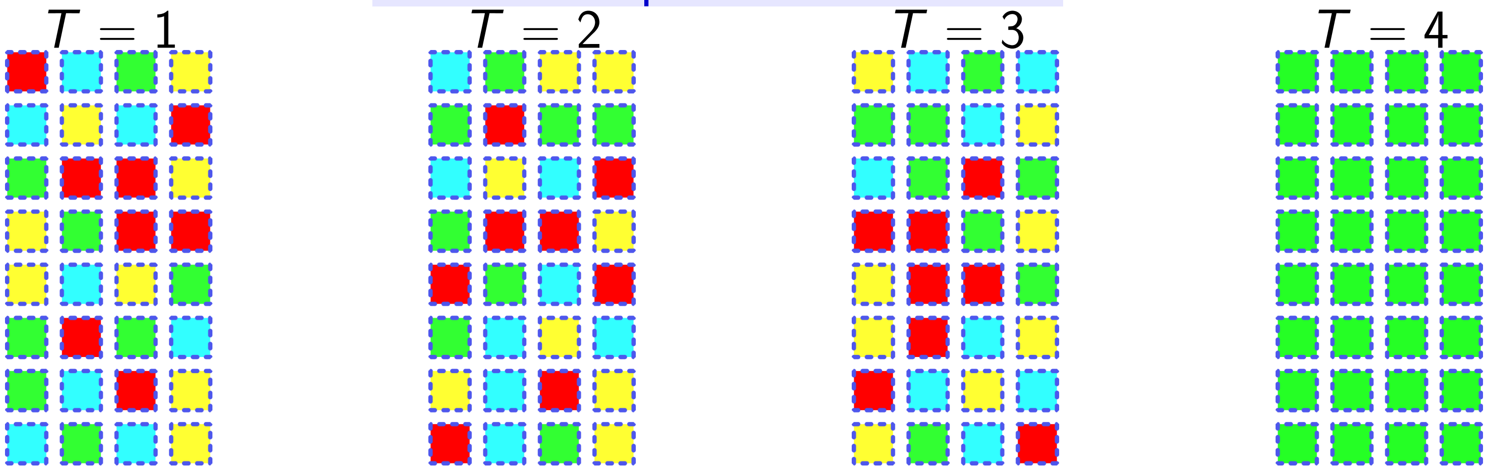


Batch constraint per step

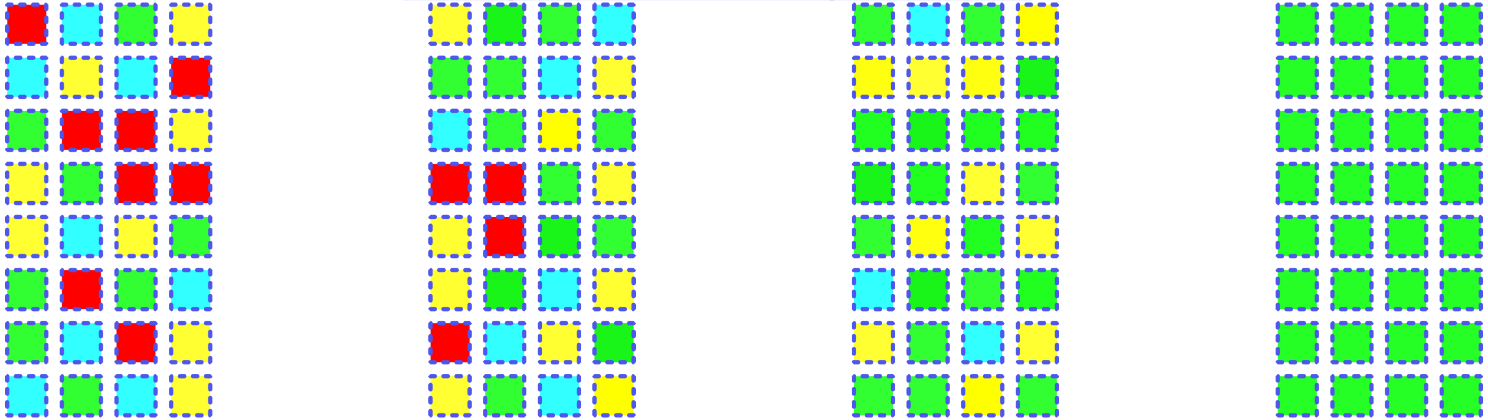


HOW TO ALLOCATE?

Random Explore then Commit

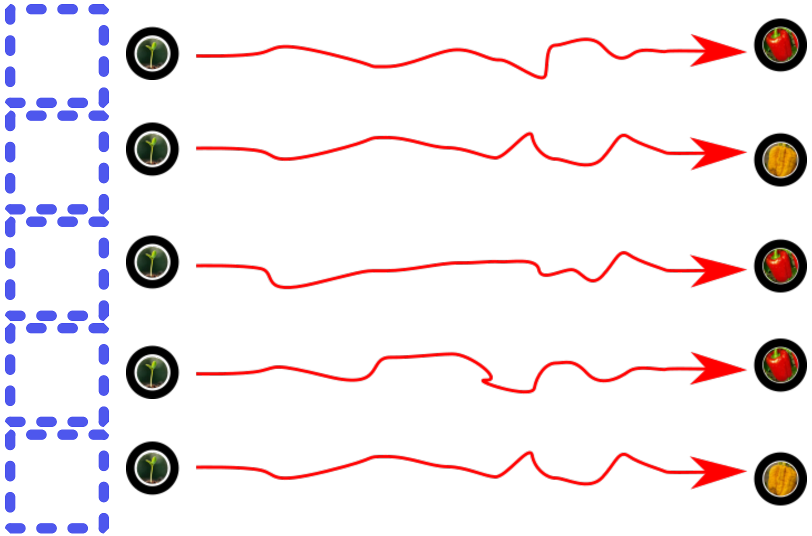


Adaptive Batch Exploration



- STUDIED EFFECTS: How many trials ?

Same strategy π:



$r_1 = 1.6$

$r_2 = 0.0$

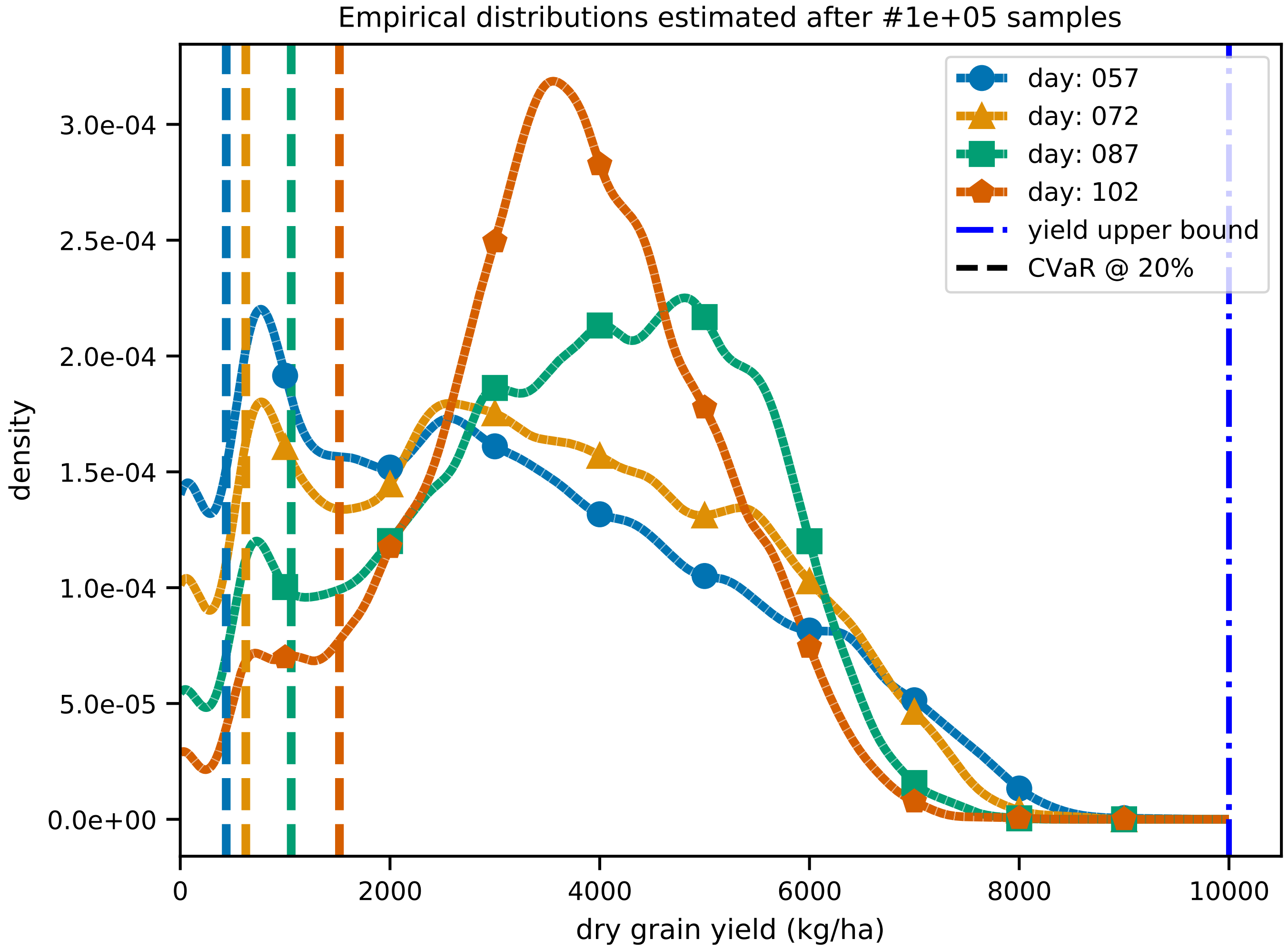
$r_3 = 1.2$

$r_4 = 1.5$

$r_5 = 0.0$

PROCESS: Apply strategy a_t at time t, receive reward r_t.

- Example: YIELD DISTRIBUTION for 4 strategies (planting date) using model DSSAT.



Expected (Risk-neutral) reward
$\mathbb{E}[r_t]$

vs

CVaR (Risk-averse) reward.
$\mathbb{C}_\alpha[r_t]$

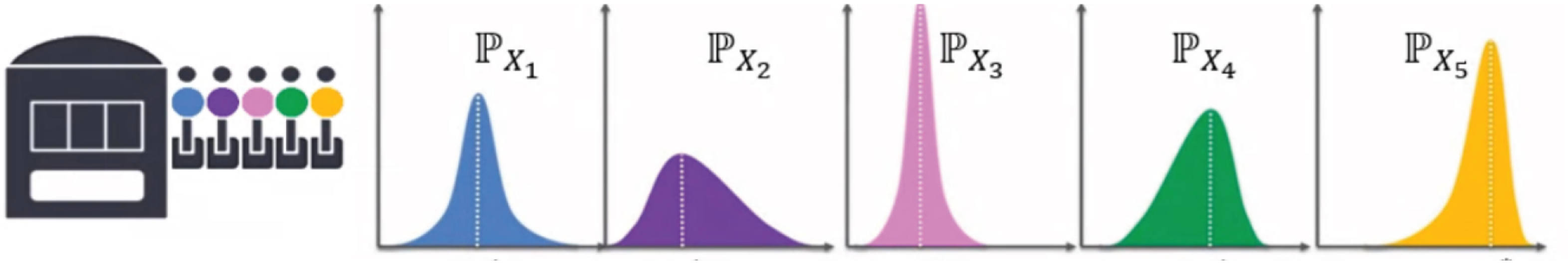
RL CHALLENGES FOR AGROECOLOGY

- Contextual RL, Continual RL
- Combinatorial policy structure
- Group Sequential RL, Adaptive experimental design
- Stochastic, Risk-averse RL

We also want:

- Learning guarantee, Reproducibility, Explainability
- Sequential Data from experiments.

Multi-armed bandits



🎯: Within-episode regret minimization.

- $\mathcal{A} = \{\pi_1, \dots, \pi_K\}$ policies with unknown mean $m_1, \dots, m_K$.
- Performance guarantee on the Cumulative Regret

$$\liminf_T \frac{\sum_{t=1}^{T} m_\star - \mathbb{E}\left[\sum_{t=1}^{T} m_{a_t}\right]}{\log(T)} \geqslant \sum_{a \in \mathcal{A}} \frac{(m^\star - m_a)}{\mathcal{K}_a(\mu^\star)}$$

OPTIMAL BANDITS STRATEGIES

State-of-the-art strategies for Expected criterion:

- **Optimistic** $\underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \widehat{m}_a(t) + B_a(t)$ where $B_a(t) \geqslant m_a - \widehat{m}_a(t)$ with high probability.

KL-UCB: Cappé, O., Garivier, A., Maillard, O. A., Munos, R., & Stoltz, G. **Kullback-Leibler upper confidence bounds for optimal sequential allocation**. In *The Annals of Statistics*, 1516-1541, 2013.

- **Bayesian** $\underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \tilde{m}_a(t)$ where $\tilde{m}_a \sim$ Posterior/Randomly reweighted mean.

TS: Thompson, W. R. **On the likelihood that one unknown probability exceeds another in view of the evidence of two samples**. In *Biometrika*, 25(3-4), 285-294, 1933.

- **Likelihood** $\underset{a\in\mathcal{A}}{\operatorname{argmin}}\, N_t(a)D(\widehat{m}_a(t), \max_a \widehat{m}_a(t)) + \ln(N_t(a))$ with divergence D.

IMED: Honda, J., & Takemura, A. **Non-Asymptotic Analysis of a New Bandit Algorithm for Semi-Bounded Rewards**. In *Journal of Machine Learning Research*, 16, 3721-3756, 2015.

- **Sub-sampling** Play all $\{a : m^\dagger_a(t) \geqslant \max_a \widehat{m}_a(t)\}$ with $\mu^\dagger_a(t)$ sub-sampled mean.

SDA: Baudry, D. and Kaufmann, E. and Maillard, O-A. **Sub-sampling for Efficient Non-Parametric Bandit Exploration**. In *Neural Information Processing System*, 2020.

CVaR Thompson Sampling

Known upper bound B on max reward.

Action $a \in \mathcal{A}$, tried $n_a(t)$ times until t, observed rewards $(X_1, \dots, X_{n_t})$

- For each a, draw a weight vector $w = (w_1, \dots, w_{n_a(t)+1}) \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{n_a(t)}, 1)$ from a Dirichlet.
- For each a, build the randomly reweighted empirical distribution:

$$\tilde{\nu}_{a,t} = \sum_{i=1}^{n_a(t)} w_i \delta_{X_i} + w_{n_a(t)+1} \delta_B \,.$$

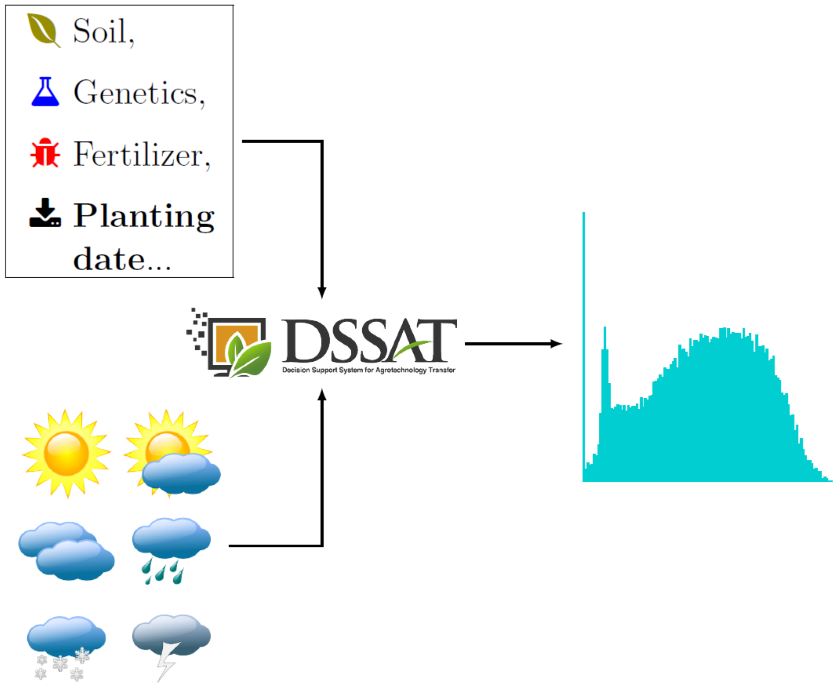
- Plays $\operatorname*{argmax}_{a \in \mathcal{A}} \mathrm{CVaR}_\alpha(\tilde{\nu}_{a,t})$

Baudry, D. and Gautron, R. and Kaufmann, E. and Maillard, O-A. **Thompson Sampling for CVaR Bandits**. In *International Conference in Machine Learning*, 2021.

Riou, C., & Honda, J. **Bandit algorithms based on thompson sampling for bounded reward distributions**. In *Algorithmic Learning Theory*, pp. 777-826, 2020.

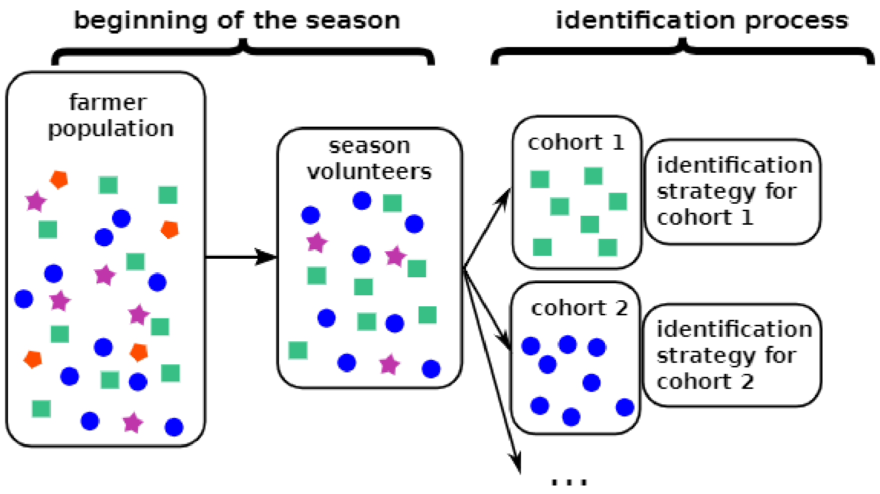
[GYM-DSSAT] SIMULATOR

- **DSSAT**: Decision Support System for AgroEcology Transfer, 30-year old internationally used Fortran simulator, integrating expertise from agronomists.

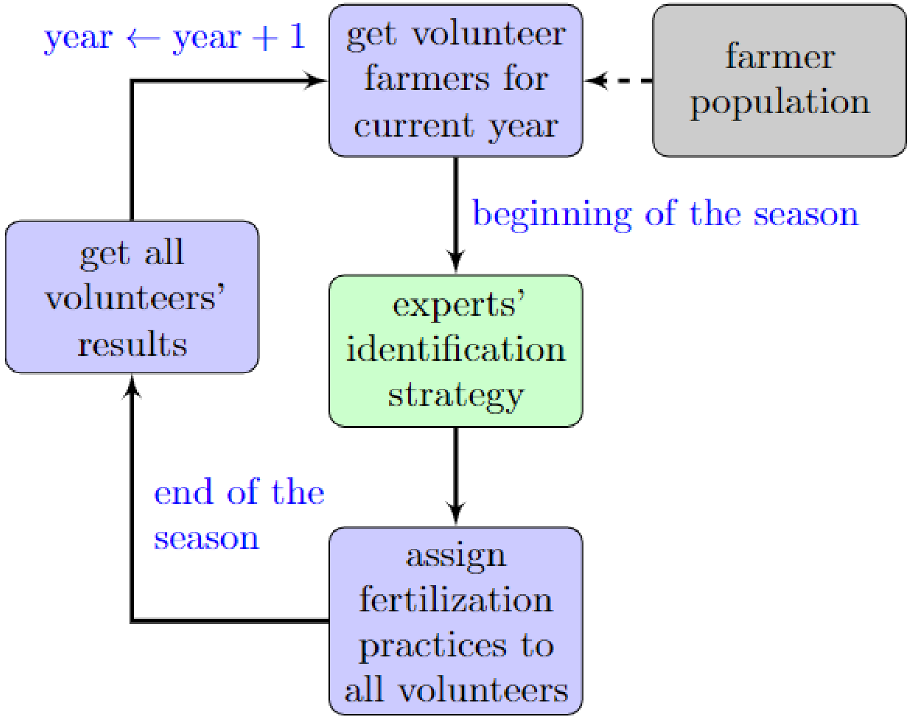


- **Gym** standardized Python for Reinforcement Learning environments.

Cohort of farmers



For T years:



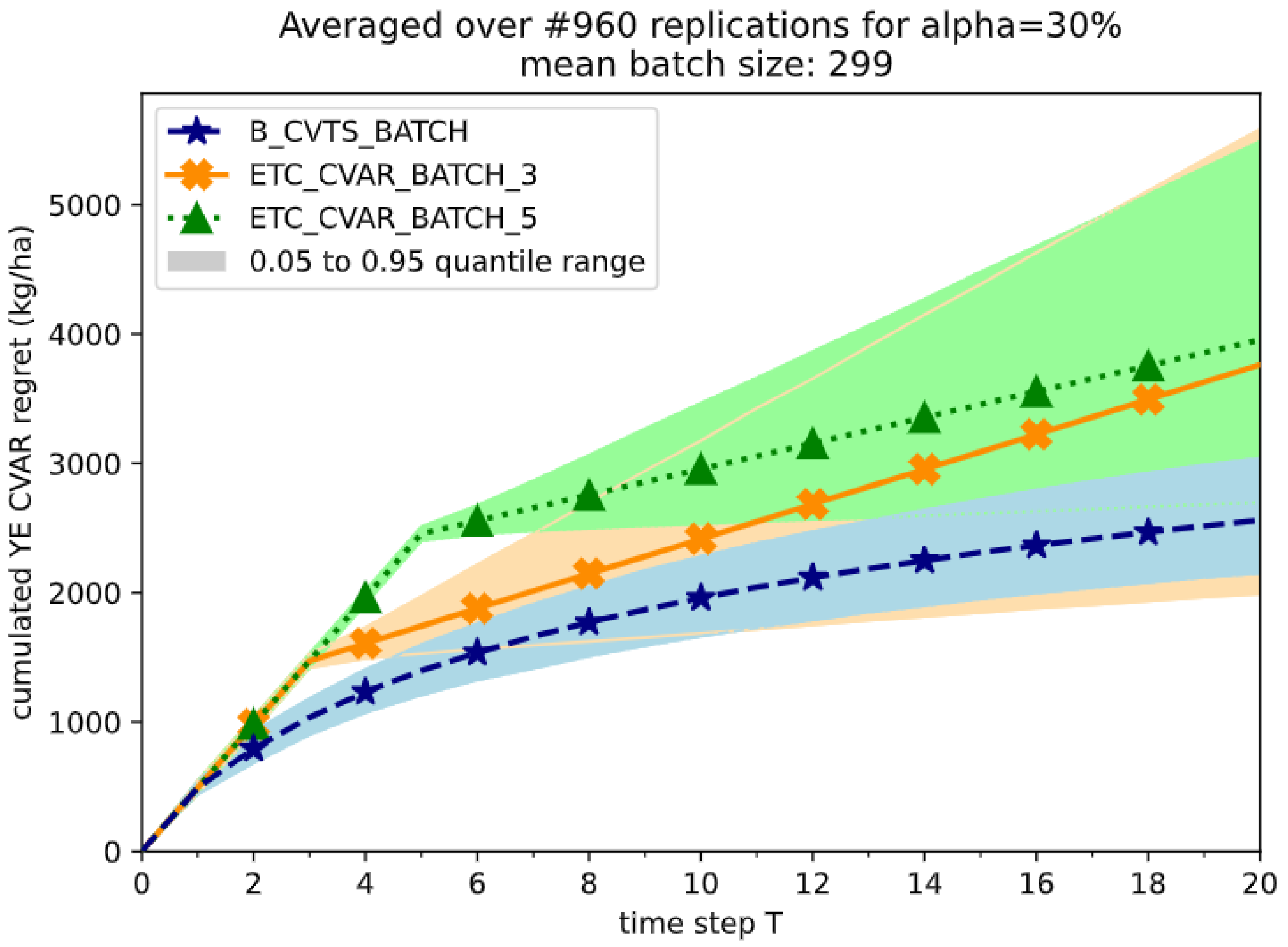
Gautron, R., Baudry, D., Adam, M., Falconnier, G. N., Hoogenboom, G., King, B., & Corbeels, M. **A new adaptive identification strategy of best crop management with farmers.**. In *Field Crops Research*, 307, 109249., 2024.

- SOIL contexts:

soil name	texture	fertility	depth	prop.
ITML840101	clay loam	low	medium	7%
ITML840102	loam	low	medium	9%
ITML840103	silty loam	low	deep	21%
ITML840104	silty clay loam	medium	medium	4%
ITML840105	silty clay loam	low	medium	24%
ITML840106	loam	low	medium	27%
ITML840107	silty clay loam	medium	medium	8%

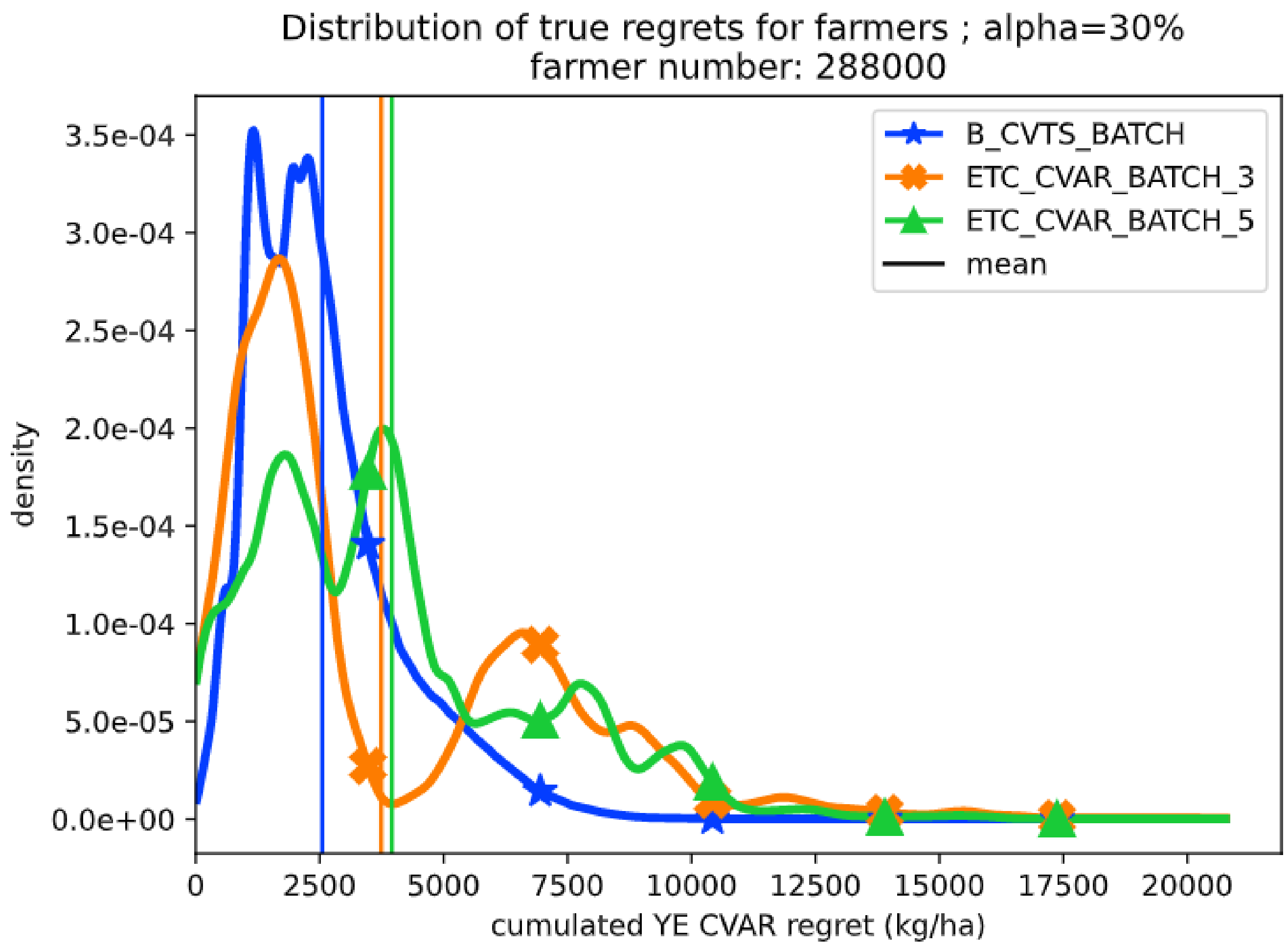
- EXPERT policies:

index	max tot. N (kg/ha)	max appl. #	rainfall thres.	NSTRES thres.	15 DAP N (kgN/ha)	30 DAP N (kgN/ha)	45 DAP N (kgN/ha)
0	135	2	No	No	15	120	0
1	135			Yes	15	120	0
2	135		Yes	No	15	120	0
3	135			Yes	15	120	0
4	135	3	No	No	15	60	60
5	135			Yes	15	60	60
6	135		Yes	No	15	60	60
7	135			Yes	15	60	60
8	70	2	No	No	23	0	47
9	180	3	No	No	60	60	60



(The lower the better: Bandit is Blue)

STRATEGY PERFORMANCE II: RISK

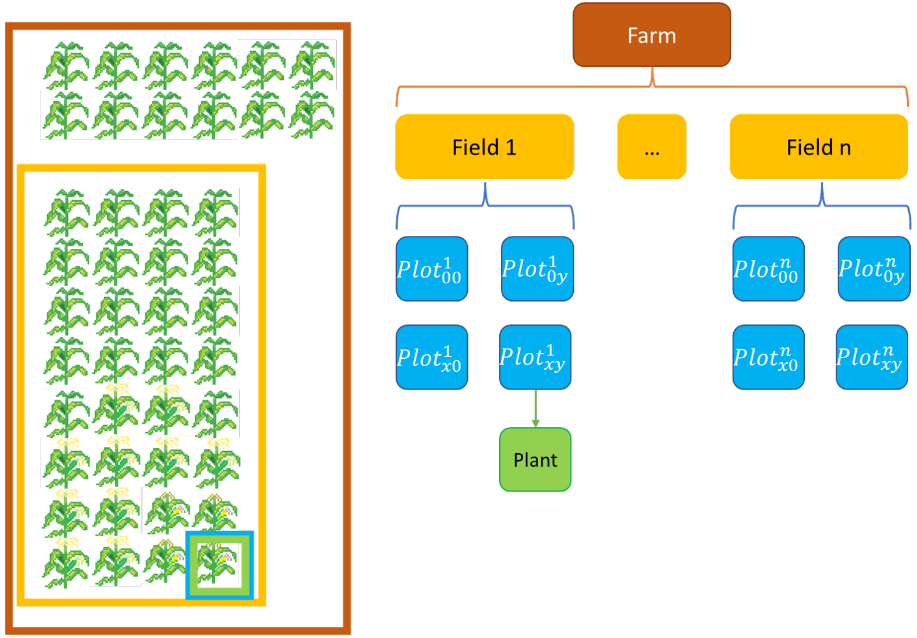


(The more mass on the left the better: Bandit is Blue)

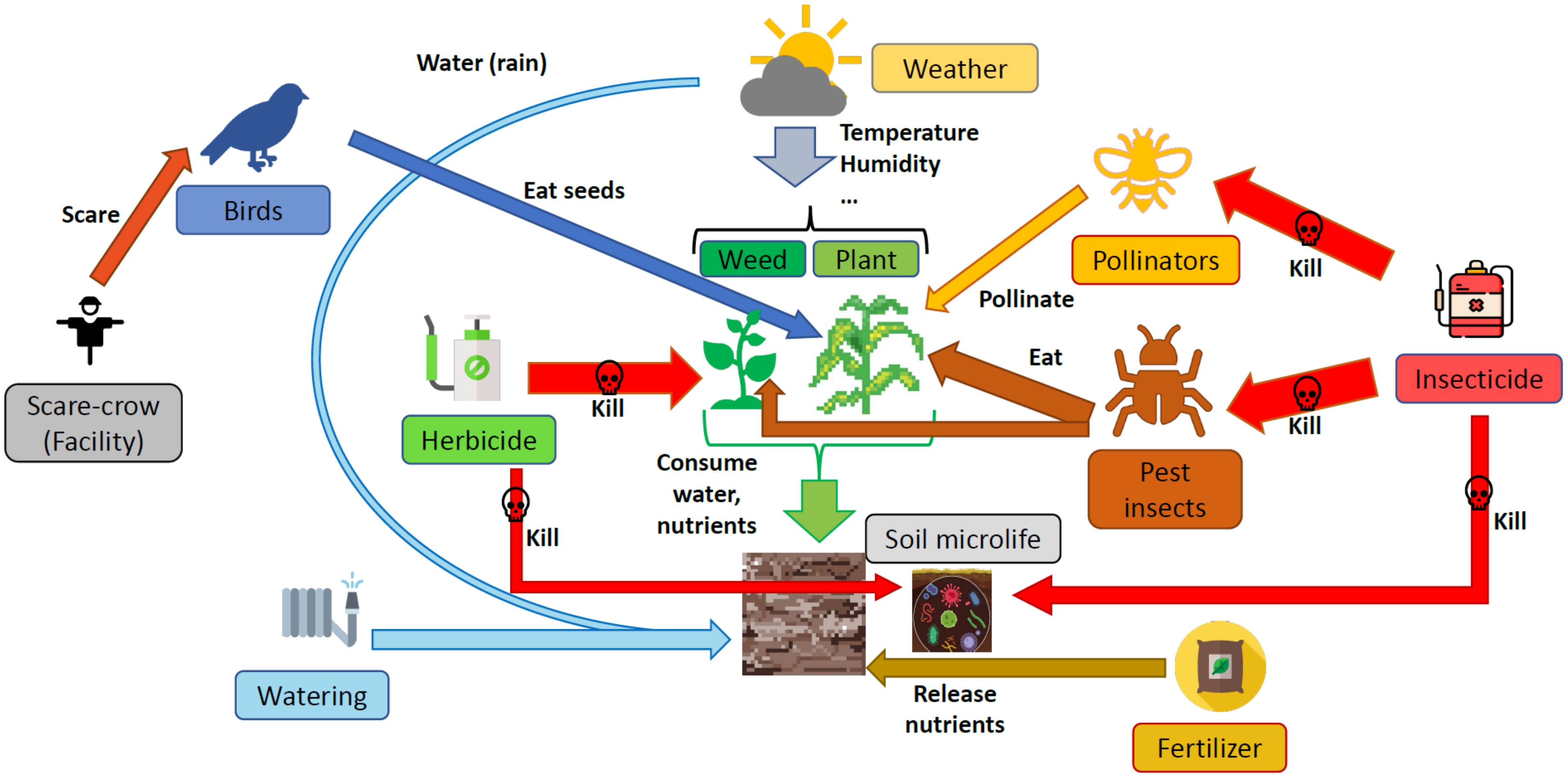
What is next?

RL PLATFORM to design and simulate gamified agroecosystems
[README] [DEMO] [DIY] [TUTO]

- To foster Reproducible research on Continual RL in stochastic environment.
- MODULAR building blocks: Farms consist of fields, farmers, entities, scoring function.

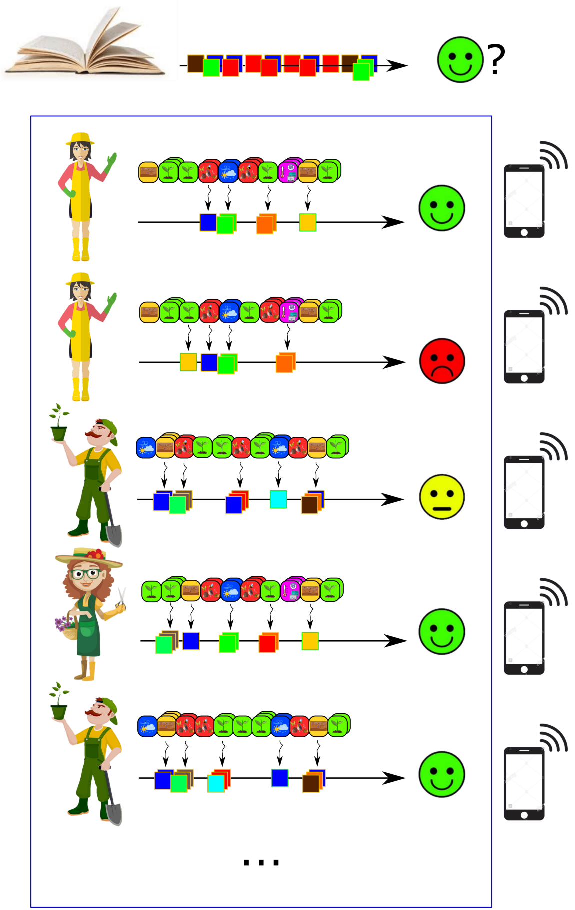


- Each entity has its own dynamic plus interacts with others.



WEGARDEN PLATFORM
[wegarden.lille.inria.fr]

- Co-identification of good practices with personalized contexts.



REINFORCEMENT LEARNING

COLLABORATIVE EXPERIMENTATION

Contexte de l'utilisateur | Suivi | Scores | Évaluation

Much remains to be done

- More interdisciplinarity between RL and Agro community
- Massify data collection, F.A.I.R. principles, reproducibility
- Improved models, simulators
- From RL algorithms to RL software
- Compliance, Appropriation, human feedback.
- ...

PEPR **AgroEcoReco** 🤞

M E R C I

“The more applied you go, the stronger theory you need”

odalric.maillard@inria.fr