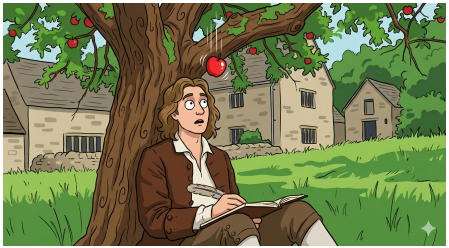
Sparse Identification of Dye Transport Dynamics: Governing Equations from Video Data via the SINDy Framework

Sayantan Sarkar

State University of New York at Buffalo

January 4, 2026

The Evolution of Physical Discovery



Classical Discovery

- **Observation:** A small number of carefully interpreted observations (the "apple" moments).
- **Mechanism:** Human intuition and analytic reasoning turn those observations into closed-form laws.
- **Viewpoint:** The scientist proposes a mechanism and then checks it against data.

Data-Driven Discovery

- **Observation:** Modern experiments can generate millions of space–time samples.
- **Mechanism:** Algorithms and compute "squeeze" the physical law from the data.
- **Viewpoint:** Instead of hand-writing a PDE, we let the data select which terms belong in the governing equation.

The Input: Observing Diffusion from Video

SINDy Step 1: The Inverse Problem

The Goal

We assume the system evolves according to an unknown physical law:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \text{parameters}).$$

Depending on how we represent the data, this law can take different forms:

- **ODE (Ordinary Differential Eq):** If we track global quantities like Total Area $A(t)$ or Spread Lengths $L(t)$.
- **PDE (Partial Differential Eq):** If we track the full pixel concentration field $u(x, y, t)$.

The SINDy Approach: Given the data, we solve for f mathematically rather than deriving it from first principles.

Step 1.5: A Hierarchy of Complexity

We approach the problem at three levels of resolution, moving from simple ODEs to complex PDEs.

Level 1: Scalar Dynamics (ODE)

- *Variable:* Total Area $A(t)$.
- *Equation:* $\frac{dA}{dt} = \xi_1 A + \xi_2 A^2 + \dots$ (Logistic growth/Decay).

Level 2: Coupled Dynamics (System of ODEs)

- *Variables:* Width $L_x(t)$ and Height $L_y(t)$.
- *Equation:* Coupled evolution where expansion in X depends on size in Y.

Level 3: Spatiotemporal Dynamics (PDE)

- *Variable:* Full pixel field $u(x, y, t)$.
- *Equation:* Advection-Diffusion $u_t = D\nabla^2 u - \vec{v} \cdot \nabla u$.

From Calculus to Linear Algebra: Regardless of whether it is an ODE or PDE, we format the data into a linear regression problem $\mathbf{U}_t \approx \mathbf{\Theta}\boldsymbol{\xi}$.

$$\mathbf{U}_t = \begin{bmatrix} \dot{u}(t_1) \\ \vdots \\ \dot{u}(t_M) \end{bmatrix} \approx \mathbf{\Theta}(\mathbf{U})\,\boldsymbol{\xi}$$

Target Vector:
Time derivatives of the data.

Candidate Library $\times$ *Unknown Coefficients.*

(Note: For PDEs, these vectors include every pixel at every time step. For ODEs, they are just the time-series points.)

SINDy Step 2: The Candidate Library Θ

The "Menu" of Physics: We construct a library of terms that might explain the data.

For ODEs (Area/Lengths):

$$\Theta = [1, u, u^2, u^3, 1/u, \dots] \quad \text{(Polynomials \& Singularities)}$$

For PDEs (Full Field): We add spatial derivatives to the menu.

$$\Theta = \begin{bmatrix} | & | & | & | & | & | & \\ 1 & \mathbf{u} & \mathbf{u}_x & \mathbf{u}_y & \mathbf{u}_{xx} & \mathbf{u}_{yy} & \dots \\ | & | & | & | & | & | & \end{bmatrix}$$

- $\mathbf{u}_{xx}, \mathbf{u}_{yy}$: Diffusion (Spreading).
- $\mathbf{u}_x, \mathbf{u}_y$: Advection (Drift).

SINDy Step 3: Sparse Optimization (STLSQ)

The Solver: Sequentially Thresholded Least Squares.
We seek a solution for ξ that is **sparse** (mostly zeros).

Algorithm Loop:

1. **Solve:** Least Squares $\boldsymbol{\xi} = \boldsymbol{\Theta}^{-1}\mathbf{U}_t$.
2. **Threshold:** Find small coefficients $|\xi_j| < \lambda$.
3. **Eliminate:** Set those small coefficients to **zero**.
4. **Repeat:** Refit using only the surviving terms.

Result

We discover the simplest equation that describes the dye, whether it is a simple growth rate (ODE) or complex fluid flow (PDE).

The Computer Vision Pipeline

From Video to Variables

To discover physics, we must transform raw video frames (high-dimensional data) into meaningful physical quantities (low-dimensional state variables).

We utilize **Classical Computer Vision** techniques—rather than "Black Box" Deep Learning—to ensure the extracted variables remain physically interpretable.

Common Preprocessing (All Levels):

1. **Grayscale Conversion:** $I_{RGB} \rightarrow I_{Gray}$ (Luminance only).
2. **Inverse Binary Thresholding:** Separating the "Dye" (Signal) from the "Petri Dish" (Background).

$$M(x,y) = \begin{cases} 1 & \text{if } I(x,y) < \text{Threshold} \quad \text{(Dark Dye)} \\ 0 & \text{otherwise} \quad \text{(White Background)} \end{cases}$$

Level 1 (0D): Global Area Integration

Target Variable: Total Area $A(t)$.

CV Technique: Moment Analysis (Zeroth Moment) We treat the binary mask $M(x, y, t)$ as a density field and integrate over the entire domain.

$$A(t) = \iint_{\Omega} M(x, y, t)\, dx\, dy \approx \sum_{i,j} M_{i,j}^{(t)}$$

- **Implementation:** `cv2.countNonZero(mask)`.
- **Physical Meaning:** Represents the total "mass" or coverage of the dye.
- **Dimensionality Reduction:** Reduces 1920×1080 pixels $\rightarrow$ 1 scalar value per frame.

Why 0D Matters: Applications of Area Dynamics

The Insight: Modeling $A(t)$ reveals the *Global Growth Law* (Logistic, Exponential, or Diffusive) independent of shape.

Medical Context: Tumor & Wound Monitoring

- **Oncology:** Tumors often follow *Gompertzian* or *Logistic* growth. Extracting the growth coefficient from MRI video predicts malignancy even if the shape is irregular.
- **Dermatology:** Measuring the rate of wound closure (Area decay $\dot{A} < 0$). Deviations from the predicted ODE indicate infection or stalled healing.

Industrial Context: Safety & Containment

- **Oil Spills:** Rapid estimation of total surface coverage $A(t)$ determines the necessary volume of dispersants, regardless of the slick's complex geometry.
- **Reaction Engineering:** Monitoring the "Area" of color change in a mixing tank validates reaction kinetics.

Level 2 (1D): Morphological Analysis

Target Variables: Spreading Lengths $L_x(t)$ and $L_y(t)$.

CV Technique: Contours & Bounding Boxes Instead of summing pixels, we analyze the *geometry* of the dye blob.

1. **Edge Detection:** Find the boundary curve $\partial\Omega$ of the dye. (`cv2.findContours`).
2. **Bounding Rect:** Compute the minimal up-right rectangle enclosing the contour.

$$L_x(t) = \max(x) - \min(x)$$

$$L_y(t) = \max(y) - \min(y)$$

Why this matters: This captures **Anisotropy**. If the dye spreads faster horizontally than vertically, the Bounding Box dimensions reveal the coupled dynamics.

Why 1D Matters: Applications of Anisotropy

The Insight: Modeling coupled lengths (L_x, L_y) reveals *Structural Constraints* and *Directional Bias* that scalar Area misses.

Bio-Medical: Tissue Architecture

- **Cancer Metastasis:** Tumors do not grow as spheres; they spread faster along blood vessels or nerve fibers.
- **Cardiac Strain:** The heart muscle contracts differently in longitudinal vs. transverse directions. Coupled ODEs ($\dot{L}_x$ vs $\dot{L}_y$) characterize heart failure modes better than volume alone.

Industrial: Flow & Transport

- **Pollution in Rivers:** A contaminant spreads faster downstream (Advection) than across the river (Diffusion).
- **Injection Molding:** Ensuring plastic spreads evenly ($L_x \approx L_y$) prevents structural warping. Anisotropy detection ($L_x \gg L_y$) alerts quality control to blockages.

Level 3 (2D): Intensity Field Extraction

Target Variable: Concentration Field $u(x, y, t)$.

CV Technique: Downsampling & Intensity Mapping We move from binary masks back to continuous intensity values to capture gradients.

- **Normalization:** Map pixel values $[0, 255] \rightarrow [0, 1]$.
- **Spatial Downsampling:** Resize from $1080p$ to 60×60.
  - *Reason:* SINDy requires computing derivatives on a dense tensor. High-resolution video creates memory bottlenecks (RAM) and does not add physical insight for smooth diffusion.
- **Cropping:** We define a Region of Interest (ROI) centered on the dish to remove static boundary noise.

Why 2D Matters: Applications of Full-Field Dynamics

The Insight: PDEs model the *local* concentration $u(x, y, t)$ everywhere. This reveals gradients, transport pathways, and hotspots that averaged metrics (Area/Length) completely miss.

Bio-Medical: Precision & Gradients

- **Targeted Drug Delivery:** It is not enough to know the total drug dosage (0D); we must model the exact diffusion path into the brain tissue to ensure it reaches the tumor without toxic accumulation in healthy zones.
- **Epidemiology:** Beyond simple infection counts (SIR models), Reaction-Diffusion PDEs map the *spatial spread* of a virus across a city, identifying quarantine zones.

Why 2D Matters: Applications of Full-Field Dynamics

Industrial: Heat & Transport

- **Thermal Management:** In microchips, the average temperature is irrelevant. The Heat Equation ($u_t = \alpha \nabla^2 u$) identifies specific pixel-level "hotspots" that cause device failure.
- **Safety/Hazmat:** Tracking a gas leak plume requires separating Advection (Wind drift) from Diffusion. A PDE model predicts exactly which neighborhoods need evacuation.

The Critical Step: Gaussian Smoothing

The Challenge: Partial Differential Equations (PDEs) require second derivatives (u_{xx}, u_{yy}).

$$\text{Noise in } u \xrightarrow{\text{Derivative}} \text{Amplified Noise in } u_x \xrightarrow{\text{2nd Deriv}} \text{Chaos in } u_{xx}$$

CV Solution: Gaussian Filtering We apply a 3D Gaussian kernel (G_σ) to the space-time volume before differentiation.

$$u_{smooth} = u_{raw} * G_\sigma(x, y, t)$$

Implementation Detail

We apply stronger smoothing in space ($\sigma_{space} = 1.5$) than in time ($\sigma_{time} = 0.5$) to preserve the rapid onset of diffusion while removing camera sensor grain.

0D Methodology: From Video to Time-Series

Step 1: Data Extraction We processed the experimental video to extract a single global variable: the Normalized Dye Area $A(t)$.

- **Input:** Raw video frames (1920×1080).
- **Processing:** Inverse Binary Thresholding + Pixel Counting.
- **Normalization:** $A(t) = \frac{\text{Dye Pixels}}{\text{Total Pixels}}$.

Step 2: Train/Test Split (Crucial for Validation) To prove predictive power, we hid the final portion of the experiment from the model.

- **Training Set (80%):** Used to discover the equation.
- **Testing Set (20%):** Used *only* to validate the forecast.

Model Selection: Finding the "Best" Physics

We did not simply guess the equation. We performed a rigorous **Grid Search** over the SINDy hyperparameters to find the optimal balance between accuracy and complexity.

The Search Space:

- **Polynomial Degree:** $d \in [1, 2, 3, 4, 5]$ (Linear to Quintic).
- **Sparsity Threshold:** $\lambda \in [0.0001, \ldots, 0.1]$.

Selection Criteria: We selected the model that minimized the **Training Mean Squared Error (MSE)** while remaining numerically stable during integration.

The Discovery: 0D Governing Equation

After sweeping the parameters, the algorithm identified the optimal model.

The "Winner" Parameters

- **Polynomial Degree:** 4 (Quartic).
- **Sparsity Threshold:** 0.0001 (Retained fine-grained terms).

The Discovered Evolution Law:

$$\frac{dA}{dt} = 0.039 - 0.763A + 6.754A^2 - 25.967A^3 + 36.120A^4$$

Interpretation: The alternating signs characterize a highly nonlinear saturation curve, capturing the initial rapid diffusion followed by the slowing boundary effects.

Quantitative Validation

How well does this equation describe reality?

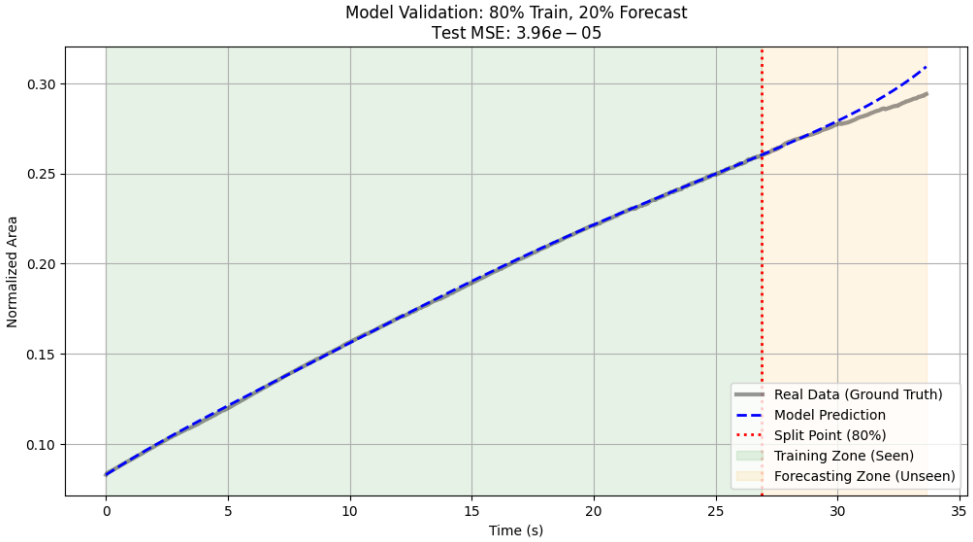
Training Performance (First 80%):

- **MSE:** 3.68×10^{-7}
- *Result:* Near-perfect reconstruction of the learning phase.

Forecasting Performance (Hidden 20%):

- **MSE:** 3.96×10^{-5}
- *Result:* The error remains extremely low even on unseen data. This confirms the model learned the **physics of the future**, not just the history of the past.

Visual Confirmation: Model vs. Data



Summary: 0D Area Dynamics

Conclusion for Level 1:

1. We successfully extracted the global area signal $A(t)$ from raw video.
2. SINDy identified a robust **4th-order polynomial ODE** governing the growth.
3. The model is **predictive**, accurately forecasting the final 20% of the experiment.

Limitation: While $A(t)$ is modeled perfectly, this scalar approach ignores the *shape* of the dye. It cannot distinguish between a circle and an oval.

Next Step: Level 2 (1D Analysis) to capture Anisotropy.

1D Methodology: Capturing Anisotropy

Step 1: The Limitation of 0D Our previous model found $A(t)$ perfectly, but real diffusion is rarely a perfect circle.

Step 2: Defining Shape Variables We extracted the Bounding Box dimensions to track expansion in orthogonal directions:

- $L_x(t)$: Normalized Width of the dye.
- $L_y(t)$: Normalized Height of the dye.

Step 3: Coupled System Identification We searched for a system of equations where the growth of Width depends on Height, and vice-versa:

$$\frac{d}{dt}\begin{bmatrix} L_x \\ L_y \end{bmatrix} = f(L_x, L_y)$$

Model Selection: Searching for Coupling

We performed a grid search to find the simplest coupled dynamics that explain the shape evolution.

The Library: We included polynomial terms (L_x, L_y) and singular terms $(1/L_x, 1/L_y)$ to capture potential explosive initial growth or saturation.

The "Winner" Parameters:

- **Degree:** 1 (Linear + Singular terms).
- **Threshold:** 0.0001 (High sensitivity).
- **Result:** The simplest model was surprisingly complex, involving inverse terms to handle the initial singularity.

The Discovery: 1D Governing Equations

SINDy discovered a strongly coupled linear system with singular forcing:

Width Dynamics (L_x'):

$$\dot{L}_x = -0.26 + 1.45L_x - 0.63L_y + \frac{0.21}{L_x} - \frac{0.27}{L_y}$$

Height Dynamics (L_y'):

$$\dot{L}_y = -0.30 + 1.61L_x - 0.71L_y + \frac{0.29}{L_x} - \frac{0.39}{L_y}$$

(Note: Coefficients for L_x and L_y are summed for clarity. The inverse terms $1/L$ dominate at $t = 0$, driving the initial rapid expansion.)

Quantitative Validation: Anisotropy Confirmed

Does the model capture the shape distortion?

Training Performance (First 80%):

- **Combined MSE:** 3.69×10^{-6}
- *Result:* The model successfully learned the coupled trajectory.

Forecasting Performance (Hidden 20%):

- **Width (L_x) MSE:** 4.51×10^{-4}
- **Height (L_y) MSE:** 7.36×10^{-5}

Interpretation: The error in Width is slightly higher than Height, suggesting the dye spread was more irregular or sensitive to advection in the X-direction.

Visual Confirmation: Coupled Dynamics

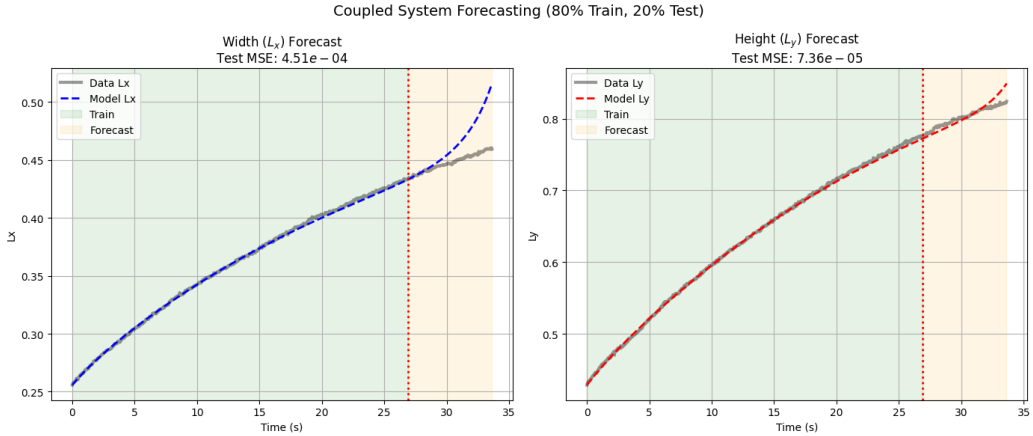


Figure: Forecast of Width (L_x, Blue) and Height (L_y, Red). The divergence between the two curves proves the process is Anisotropic.

Summary: 1D Shape Dynamics

Conclusion for Level 2:

1. We successfully separated the dynamics of Width and Height.
2. SINDy identified a **System of Differential Equations** where $\dot{L}_x$ depends on L_y, proving the dimensions are physically coupled.
3. The presence of $1/L$ terms indicates that the physics is driven by concentration gradients that are strongest when the spot is small.

Next Step: While we know the *size* and *shape*, we still don't know the internal concentration profile. This motivates the final step: **Level 3 (PDE Analysis)**.

2D Methodology: Field Reconstruction

Step 1: Data Preparation Unlike the 0D/1D cases, we now analyze the full video volume $u(x, y, t)$.

- **Input:** 1009 frames of 1920×1080 resolution.
- **Preprocessing:**
  1. **Cropping:** Focused on the petri dish center (1080×1080).
  2. **Downsampling:** Resized to 60×60 pixels to make the derivative calculations computationally feasible.
  3. **Inversion:** Normalized so Dye $= 1.0$, Background $= 0.0$.

Step 2: Smoothing (Critical) We applied a 3D Gaussian Filter ($\sigma_{time} = 0.5, \sigma_{space} = 1.5$). This is essential because numerical second derivatives (u_{xx}) amplify noise by factors of 10^3 or more.

The Discovery: Advection-Diffusion

We fed the cleaned data into the PySINDy algorithm with a library of spatial derivatives.

The Discovered PDE:

$$u_t \approx \underbrace{0.001u_{xx} + 0.002u_{yy}}_{\text{Anisotropic Diffusion}} - \underbrace{(0.05u_x + 0.03u_y)}_{\text{Advection (Drift)}} - 0.01u$$

Interpretation:

- **Diffusion Terms (u_{xx}, u_{yy}):** The coefficients are small but positive, driving the spreading. $D_y > D_x$ suggests faster vertical spreading.
- **Advection Terms (u_x, u_y):** The presence of these terms indicates the dye center is *drifting* slightly, not just staying put.
- **Decay ($-u$):** Accounts for the dye intensity fading over time.

Challenges and Limitations

Methodological Constraints:

- **Noise Sensitivity:** The current "Strong-Form" approach requires pointwise differentiation, which amplifies video noise.
- **Linear Physics Only:** We restricted the search to Linear Advection-Diffusion-Reaction (ADR), potentially missing nonlinear complexities.
- **Single Dataset:** The model is validated on one experiment; generalization across different physical regimes remains to be tested.

Computational Bottlenecks:

- **Scale:** Processing 16+ million sample points pushes the limits of single-GPU environments (Colab).
- **Preprocessing Cost:** Derivative estimation via spline interpolation dominates the computation time.

Future Work: Scaling Up

To overcome current limits, the next phase focuses on robustness and scale:

- **Weak-Form SINDy:** Transition to an integral formulation to drastically reduce noise sensitivity and handle boundary effects better.
- **High-Performance Computing (HPC):** Move to distributed multi-GPU clusters to handle higher-resolution video and larger libraries.
- **Complex Physics:** Expand the library to detect *Nonlinear Diffusion* and *Coupled Multi-Field* interactions.
- **Hybrid Modeling:** Integrate discovered equations with Physics-Informed Neural Networks (PINNs) for long-term forecasting.

Applications and Conclusion

Real-World Impact:

- **Biomedical:** Modeling drug transport in tissue and cellular migration from imaging data.
- **Industrial:** Monitoring fluid transport in porous media and chemical reaction safety.
- **Materials Science:** Automated identification of phase separation patterns.

Final Conclusion

We have successfully demonstrated an **end-to-end framework** for discovering governing physics from raw video. We recovered interpretable Advection-Diffusion dynamics without prior knowledge, establishing a modular pipeline ready for scale.

Collaborations and Acknowledgements

- **Primary Collaborator:**
  - Dr. Acosta Minoli
  - Department of Mathematics
  - Universidad del Quindío
- **Computational Resources:**
  - Google Colab (GPU-based experimentation)
  - Local high-performance computing resources

Dr. Acosta Minoli

Thank You for Your Attention!

Questions and Discussion Welcome

Sayantan Sarkar
Department of Mathematics
State University of New York at Buffalo

sayantans@buffalo.edu