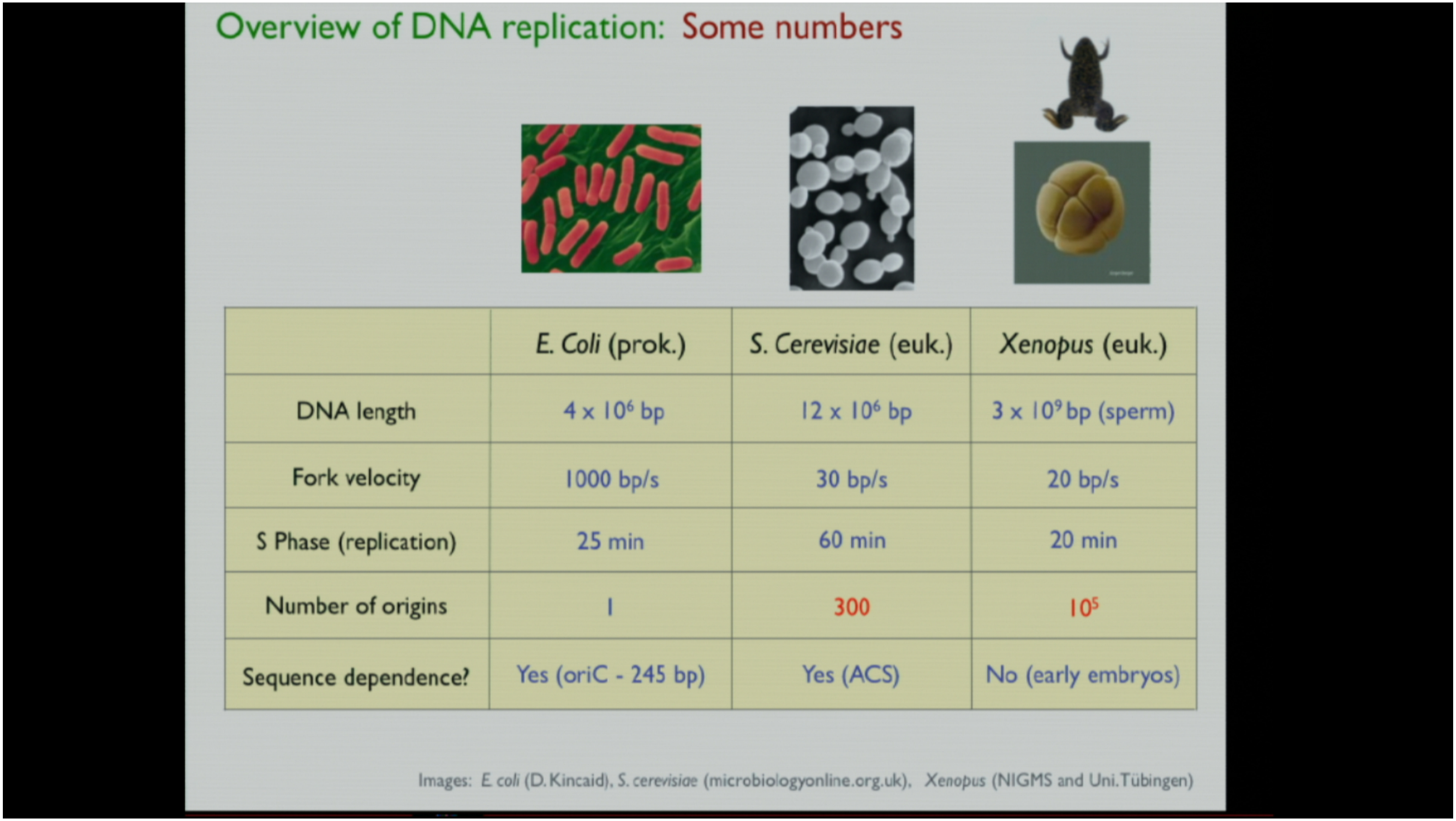
Title: Inferring the spatiotemporal DNA replication program from noisy data

Date: Dec 05, 2013 02:40 PM

URL: http://pirsa.org/13120017

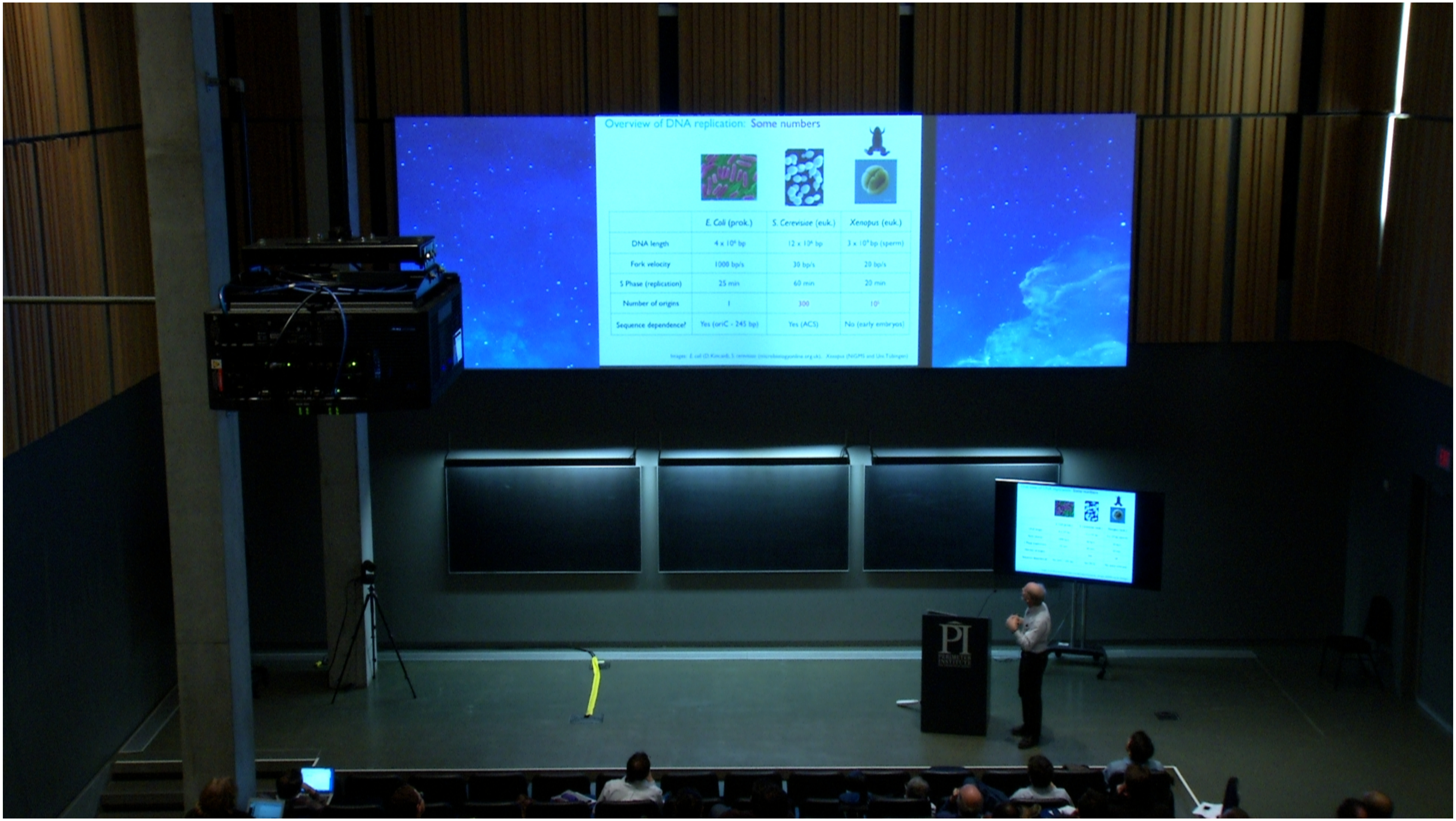
Abstract: <span>In eukaryotic organisms, DNA replication is initiated at â€œorigins,â€• launching â€œforksâ€• that spread bidirectionally to replicate the genome. The distribution and firing rate of these origins and the fork progression velocity form the â€œreplication program.â€• With Antoine Baker, I generalize a stochastic model of DNA replication to allow for space and time variations in origin-initiation rates, characterized by a function I(x,t). We then address the inverse problem of inferring I(x,t) from experimental data concerning replication in cell populations. Previous work based on curve fitting depended on arbitrarily chosen functional forms for I(x,t), with free parameters that were constrained by the data. We introduce a model-free, non-parametric method of inference that is based on Gaussian process regression, a well-known inference technique from the machine-learning community. The method replaces specific assumptions about the functional form of the initiation rate with more general prior expectations about the smoothness of variation of this rate, along the genome and in time. Using this inference method, we can recover simulated replication schemes with data that are typical of current experiments without having to know or guess the functional form for the initiation rate I(x,t). I will argue that Gaussian process regression has many other potential applications to physics.<br></span>



## Overview of DNA replication: Some numbers

| | *E. Coli* (prok.) | *S. Cerevisiae* (euk.) | *Xenopus* (euk.) |
|---|---|---|---|
| DNA length | $4 \times 10^6$ bp | $12 \times 10^6$ bp | $3 \times 10^9$ bp (sperm) |
| Fork velocity | 1000 bp/s | 30 bp/s | 20 bp/s |
| S Phase (replication) | 25 min | 60 min | 20 min |
| Number of origins | 1 | 300 | $10^5$ |
| Sequence dependence? | Yes (oriC - 245 bp) | Yes (ACS) | No (early embryos) |

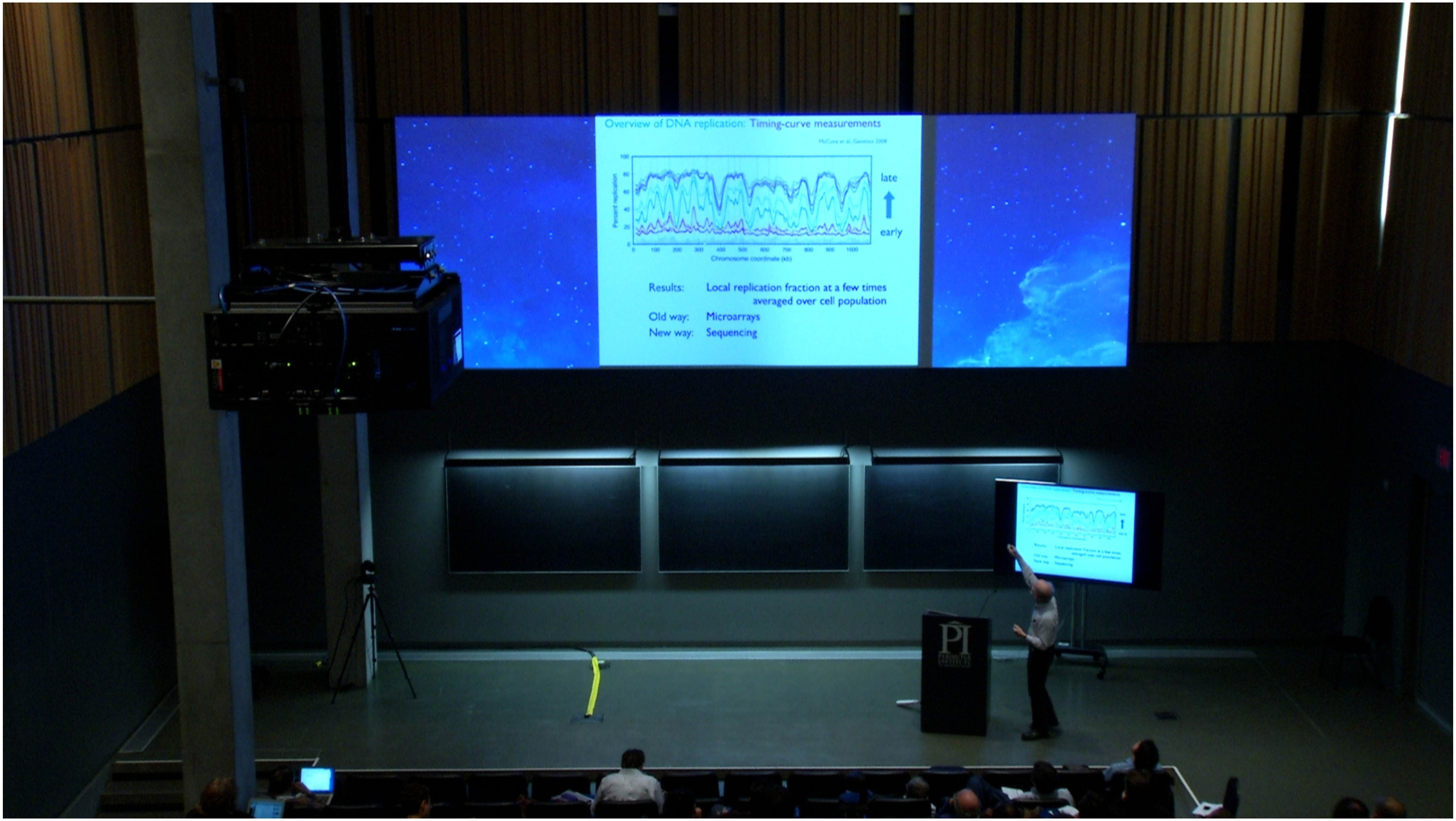
Images: *E coli* (D. Kincaid), *S. cerevisiae* (microbiologyonline.org.uk), *Xenopus* (NIGMS and Uni.Tübingen)



## Overview of DNA replication: Some numbers

| | $E. Coli$ (prok.) | $S. Cerevisiae$ (euk.) | $Xenopus$ (euk.) |
|---|---|---|---|
| DNA length | $4 \times 10^6$ bp | $12 \times 10^6$ bp | $3 \times 10^9$ bp (sperm) |
| Fork velocity | 1000 bp/s | 30 bp/s | 20 bp/s |
| S Phase (replication) | 25 min | 60 min | 20 min |
| Number of origins | 1 | 300 | $10^5$ |
| Sequence dependence? | Yes (oriC - 245 bp) | Yes (ACS) | No (early embryos) |

Images: E. coli (D. Kincaid), S. cerevisiae (microbiologyonline.org.uk). Xenopus (NIGMS and Uni Tubingen)



Overview of DNA replication: Timing-curve measurements

McCune et al., Genetics 2008

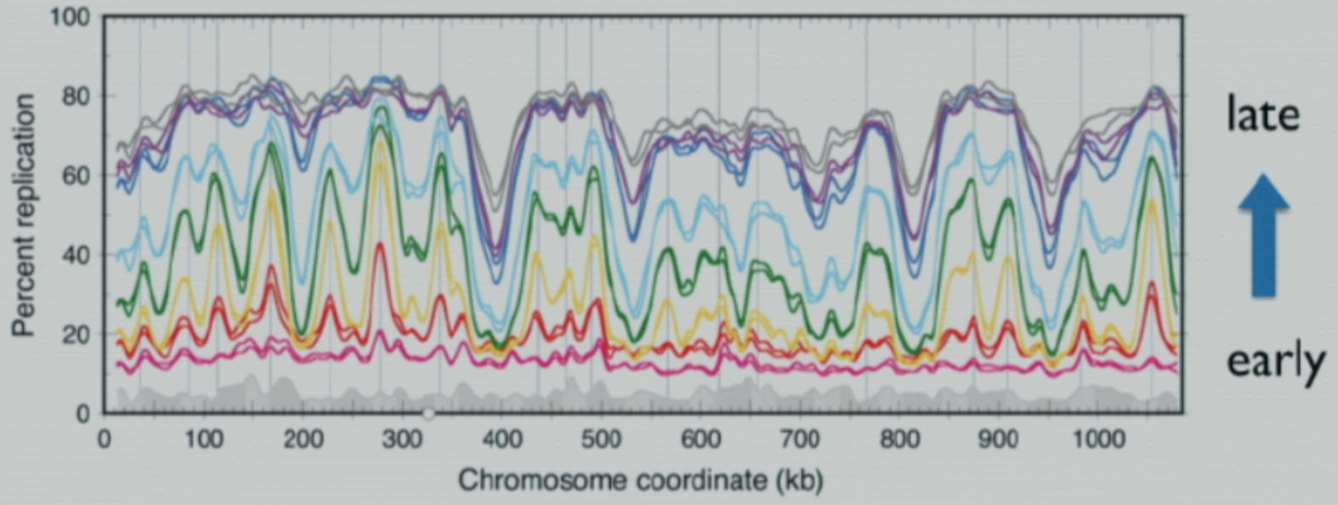
Results: Local replication fraction at a few times averaged over cell population

Old way: Microarrays

New way: Sequencing

# Overview of DNA replication: Timing-curve measurements

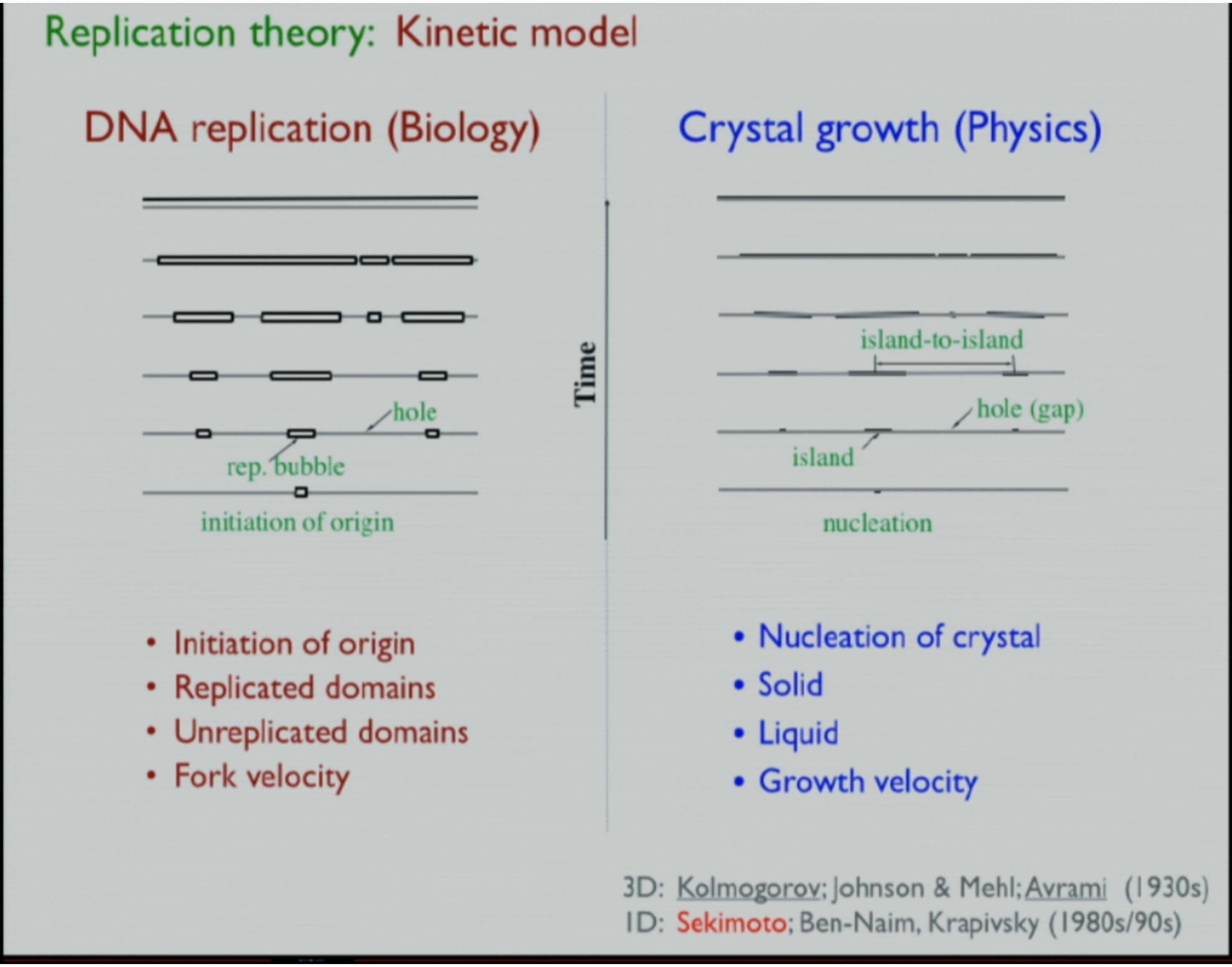
McCune et al., Genetics 2008



Results: Local replication fraction at a few times averaged over cell population

Old way: Microarrays

New way: Sequencing



## Replication theory: Kinetic model

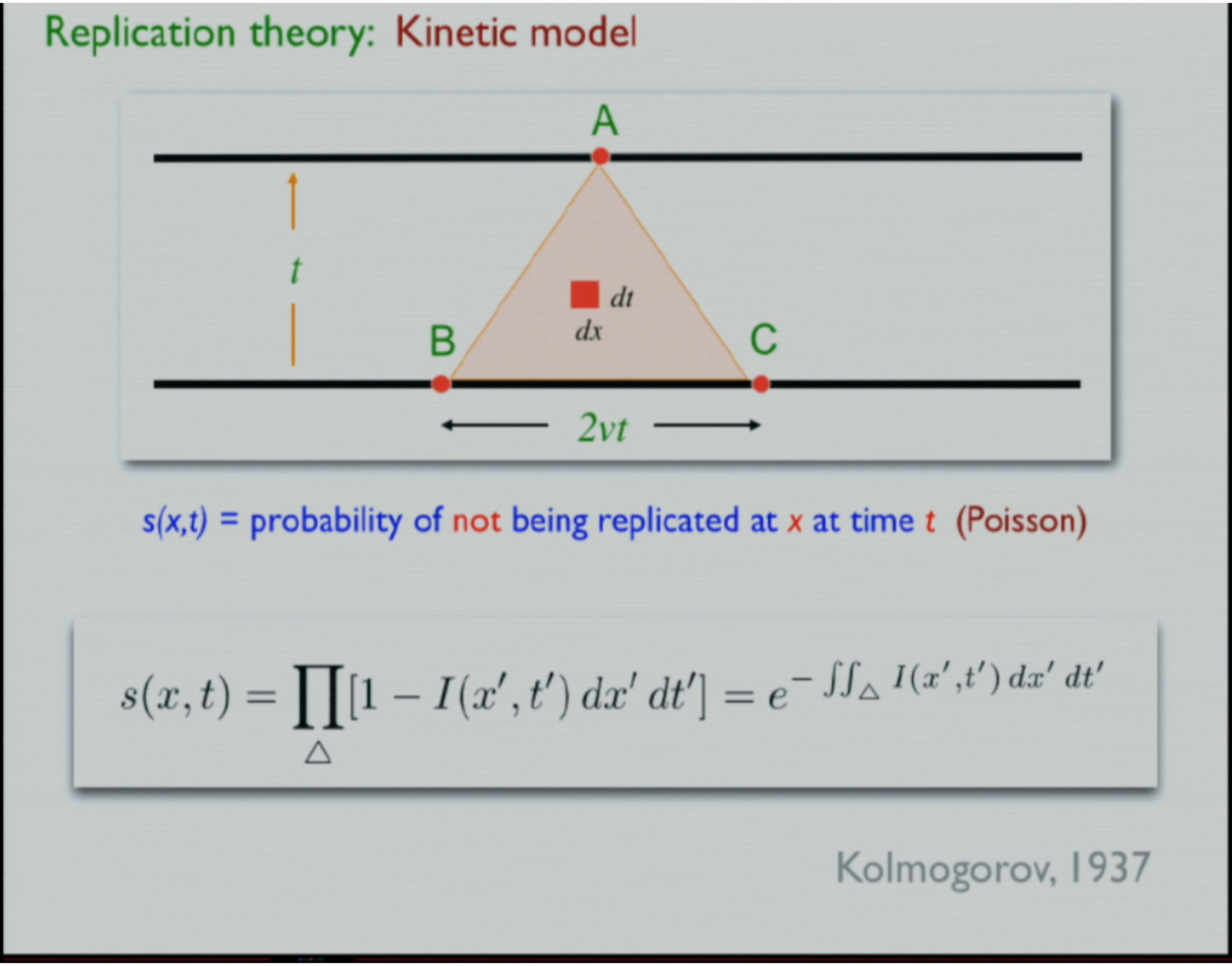
### DNA replication (Biology)

- Initiation of origin
- Replicated domains
- Unreplicated domains
- Fork velocity

### Crystal growth (Physics)

- Nucleation of crystal
- Solid
- Liquid
- Growth velocity

3D: Kolmogorov; Johnson & Mehl; Avrami (1930s)
1D: Sekimoto; Ben-Naim, Krapivsky (1980s/90s)

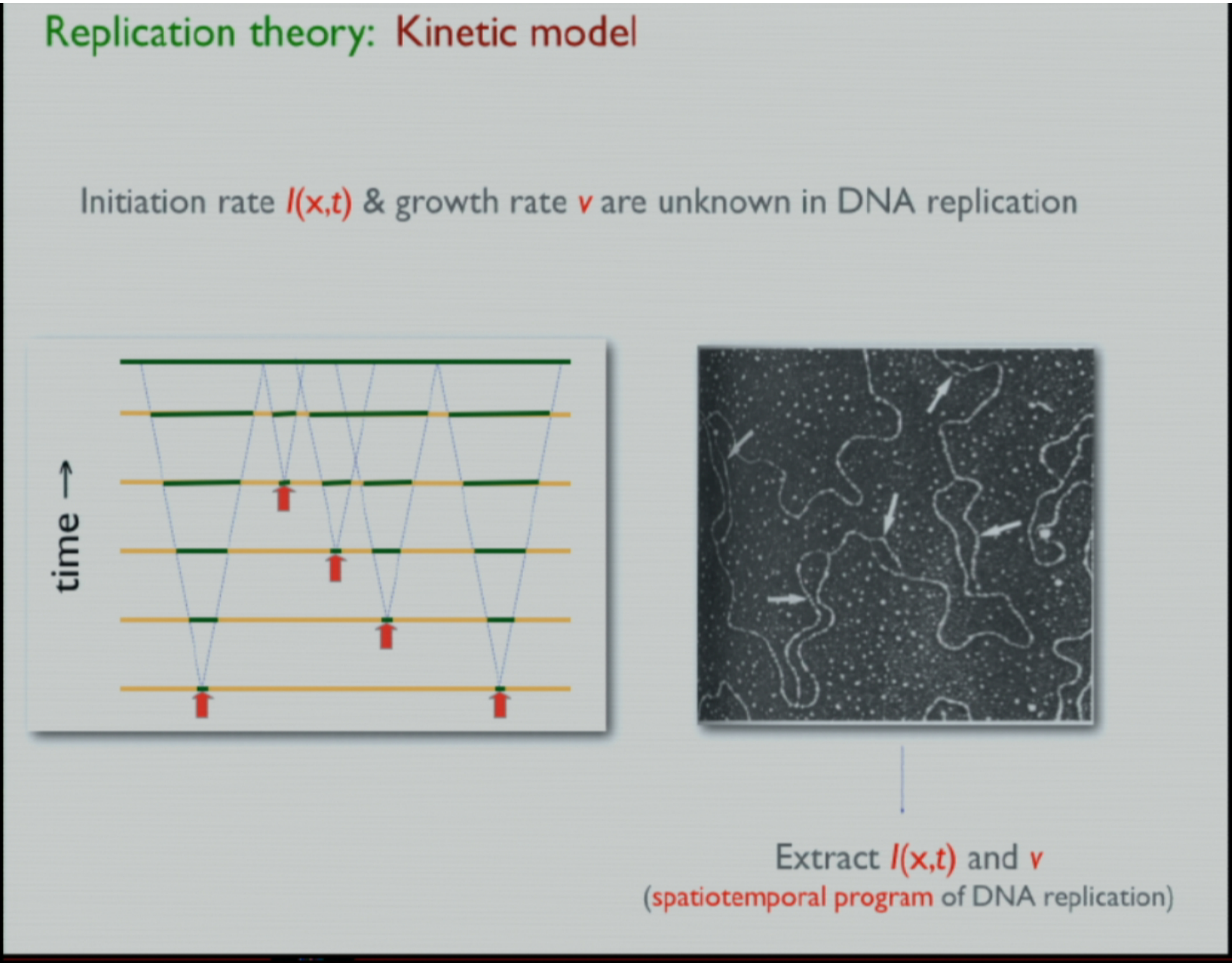


## Replication theory: Kinetic model

*s(x,t)* = probability of not being replicated at *x* at time *t* (Poisson)

$$s(x,t) = \prod_{\triangle}[1 - I(x',t')\,dx'\,dt'] = e^{-\iint_{\triangle} I(x',t')\,dx'\,dt'}$$

Kolmogorov, 1937

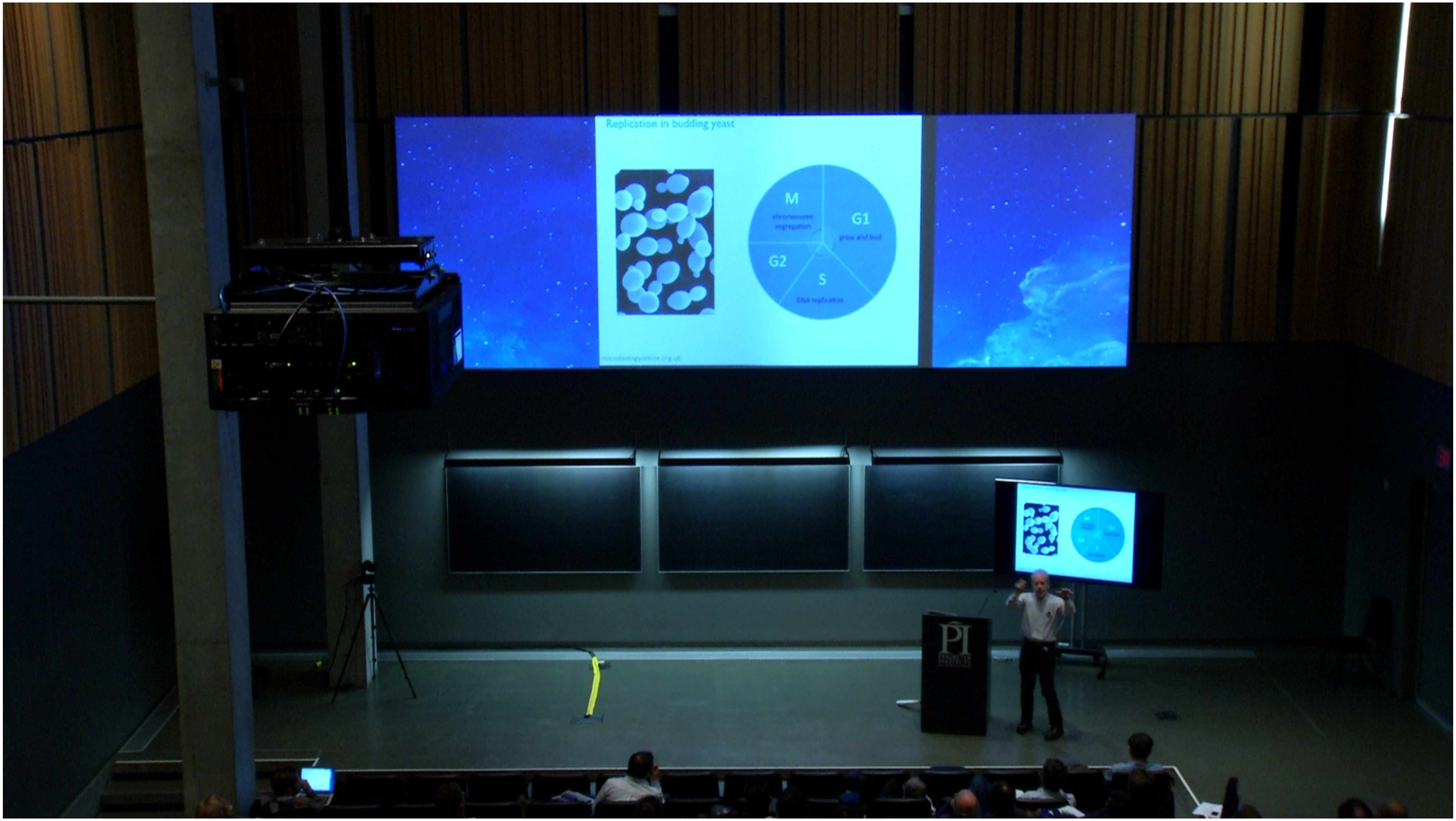


## Replication theory: Kinetic model

Initiation rate *I*(x,*t*) & growth rate *v* are unknown in DNA replication

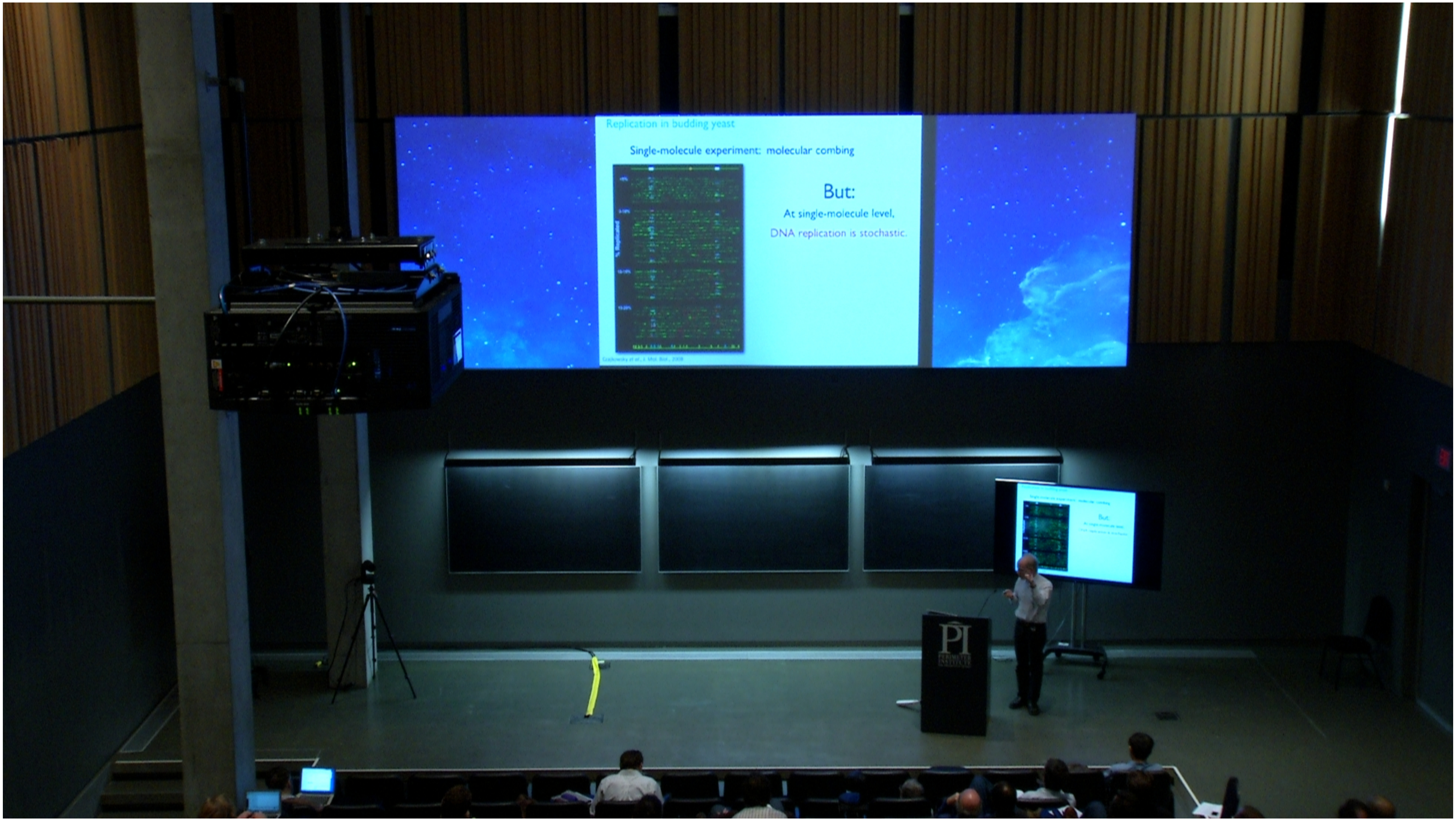
Extract *I*(x,*t*) and *v*

(spatiotemporal program of DNA replication)



Replication in budding yeast

microbiologyonline.org.uk



Replication in budding yeast

Single-molecule experiment: molecular combing

But:

At single-molecule level,

DNA replication is stochastic.

Czajkowsky et al., J. Mol. Biol., 2008

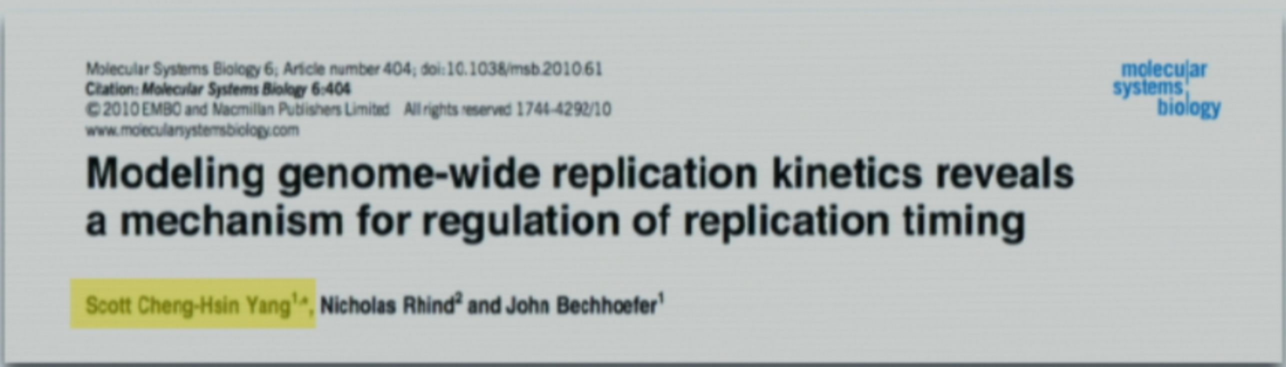
Molecular Systems Biology 6; Article number 404; doi:10.1038/msb.2010.61
Citation: Molecular Systems Biology 6:404
© 2010 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/10
www.molecularsystemsbiology.com

molecular systems biology

**Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing**

Scott Cheng-Hsin Yang[1,*], Nicholas Rhind[2] and John Bechhoefer[1]

- Global fit to all 16 chromosomes
- 337 origins,
- 3 (or 2) parameters / origin
- showed that stochasticity IS important
  - only 30% potential origins used each cycle
- $I(t)$ scales with single empirical parameter
- suggested a new view on biological mechanism for initiation
  - (loading of MCM2-7 helicase complexes)



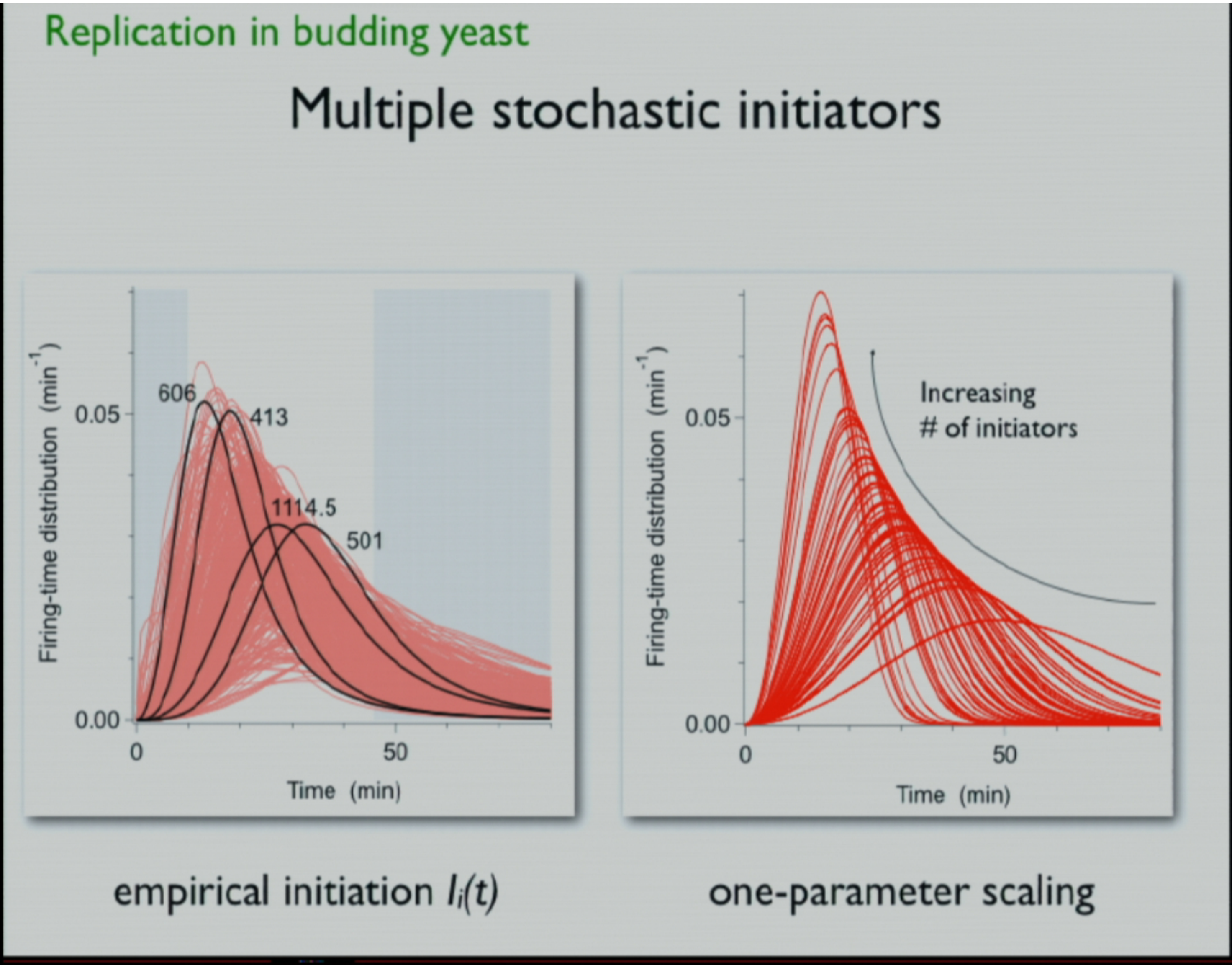
Molecular Systems Biology 6; Article number 404; doi:10.1038/msb.2010.61
**Citation:** ***Molecular Systems Biology* 6:404**
© 2010 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/10
www.molecularsystemsbiology.com

molecular systems biology

## Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing

Scott Cheng-Hsin Yang[1,*], Nicholas Rhind[2] and John Bechhoefer[1]

- Global fit to all 16 chromosomes
- 337 origins,
- 3 (or 2) parameters / origin
- showed that stochasticity IS important
  - only 30% potential origins used each cycle
- $I_i(t)$ scales with single empirical parameter
- suggested a new view on biological mechanism for initiation
  - (loading of MCM2-7 helicase complexes)

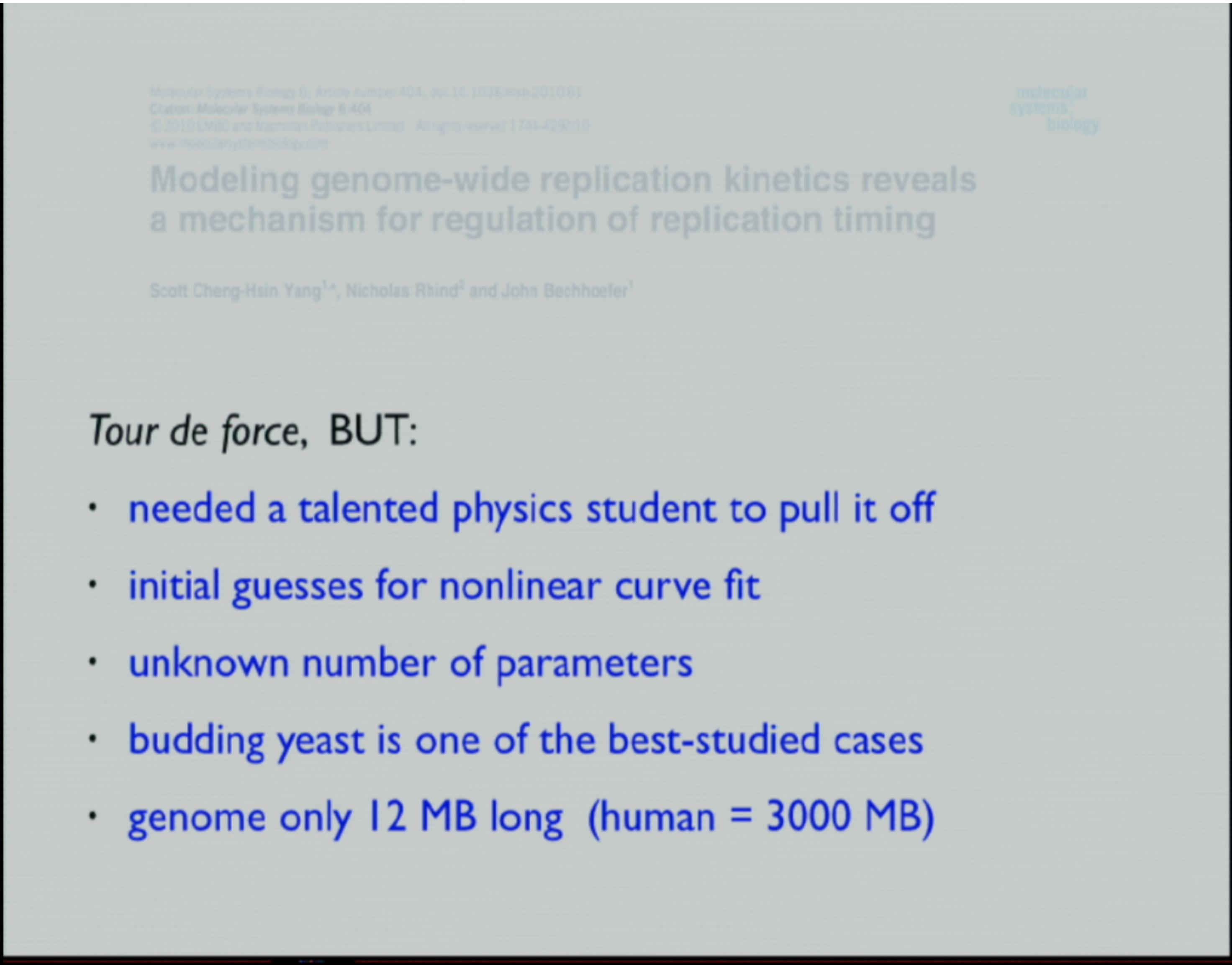


Replication in budding yeast

## Multiple stochastic initiators

empirical initiation $I_i(t)$

one-parameter scaling



Molecular Systems Biology 6; Article number 404; doi:10.1038/msb.2010.61
Citation: *Molecular Systems Biology* 6:404
© 2010 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/10
www.molecularsystemsbiology.com

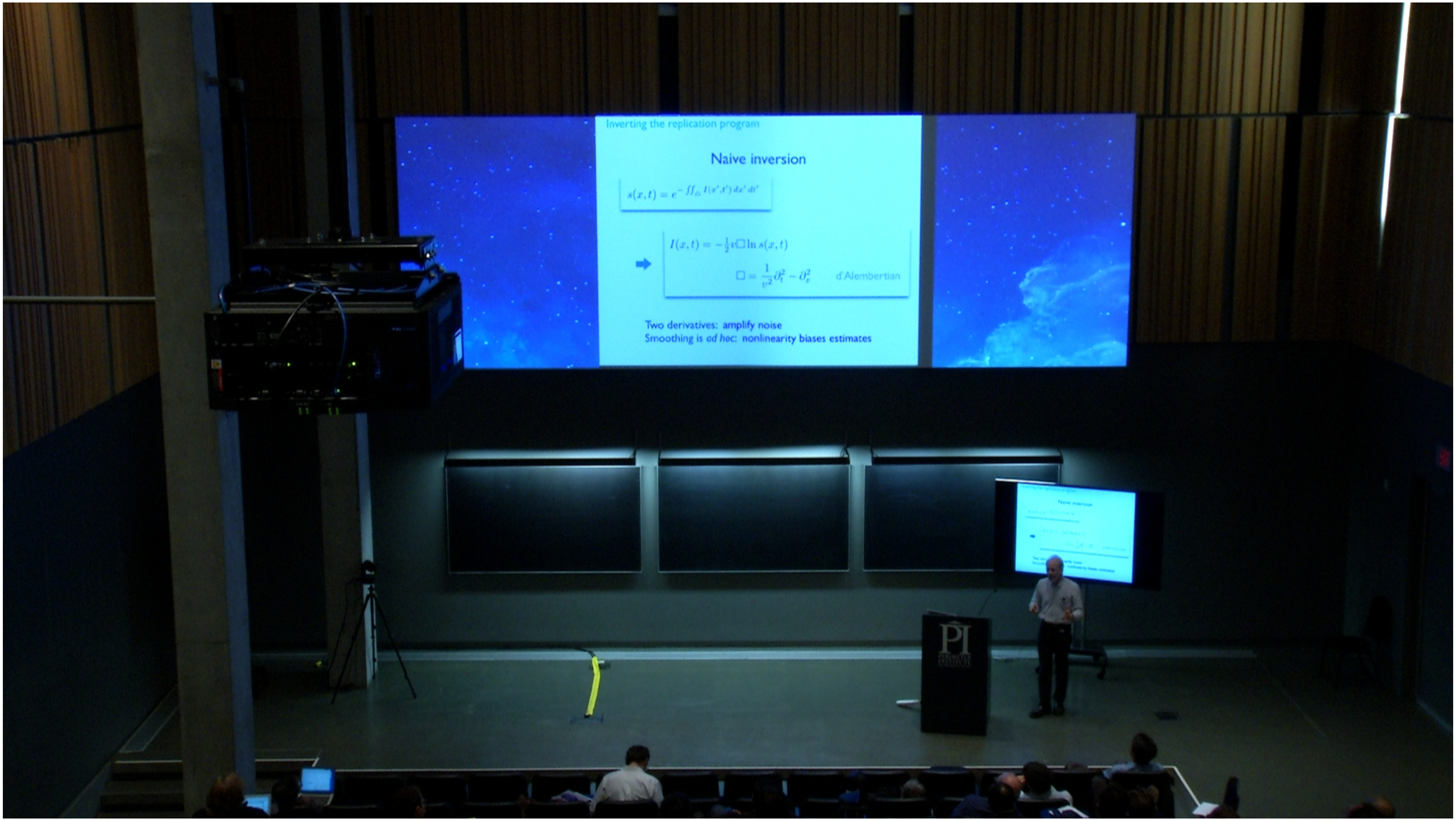
molecular
systems
biology

Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing

Scott Cheng-Hsin Yang[1],*, Nicholas Rhind[2] and John Bechhoefer[1]

*Tour de force*, BUT:

- needed a talented physics student to pull it off
- initial guesses for nonlinear curve fit
- unknown number of parameters
- budding yeast is one of the best-studied cases
- genome only 12 MB long (human = 3000 MB)



Inverting the replication program

## Naive inversion

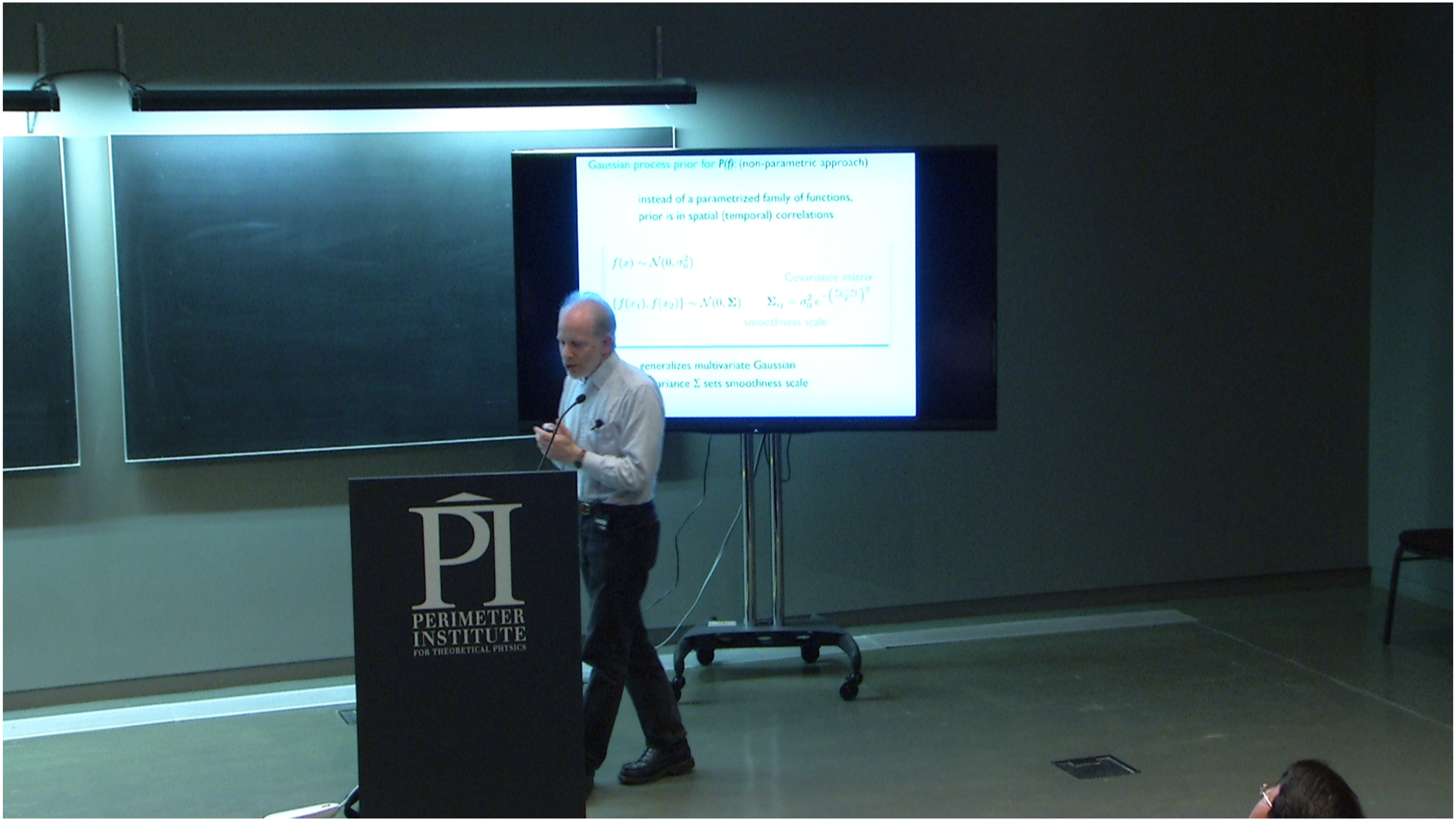
$$\boxed{s(x,t) = e^{-\iint_{\triangle} I(x',t')\,dx'\,dt'}}$$

➡

$$I(x,t) = -\tfrac{1}{2}v\Box \ln s(x,t)$$

$$\Box = \frac{1}{v^2}\partial_t^2 - \partial_x^2$$ d'Alembertian

Two derivatives: amplify noise
Smoothing is *ad hoc*: nonlinearity biases estimates



Gaussian process prior for *P(f)*: (non-parametric approach)

instead of a parametrized family of functions,
prior is in spatial (temporal) correlations

$f(x) \sim \mathcal{N}(0, \sigma_0^2)$

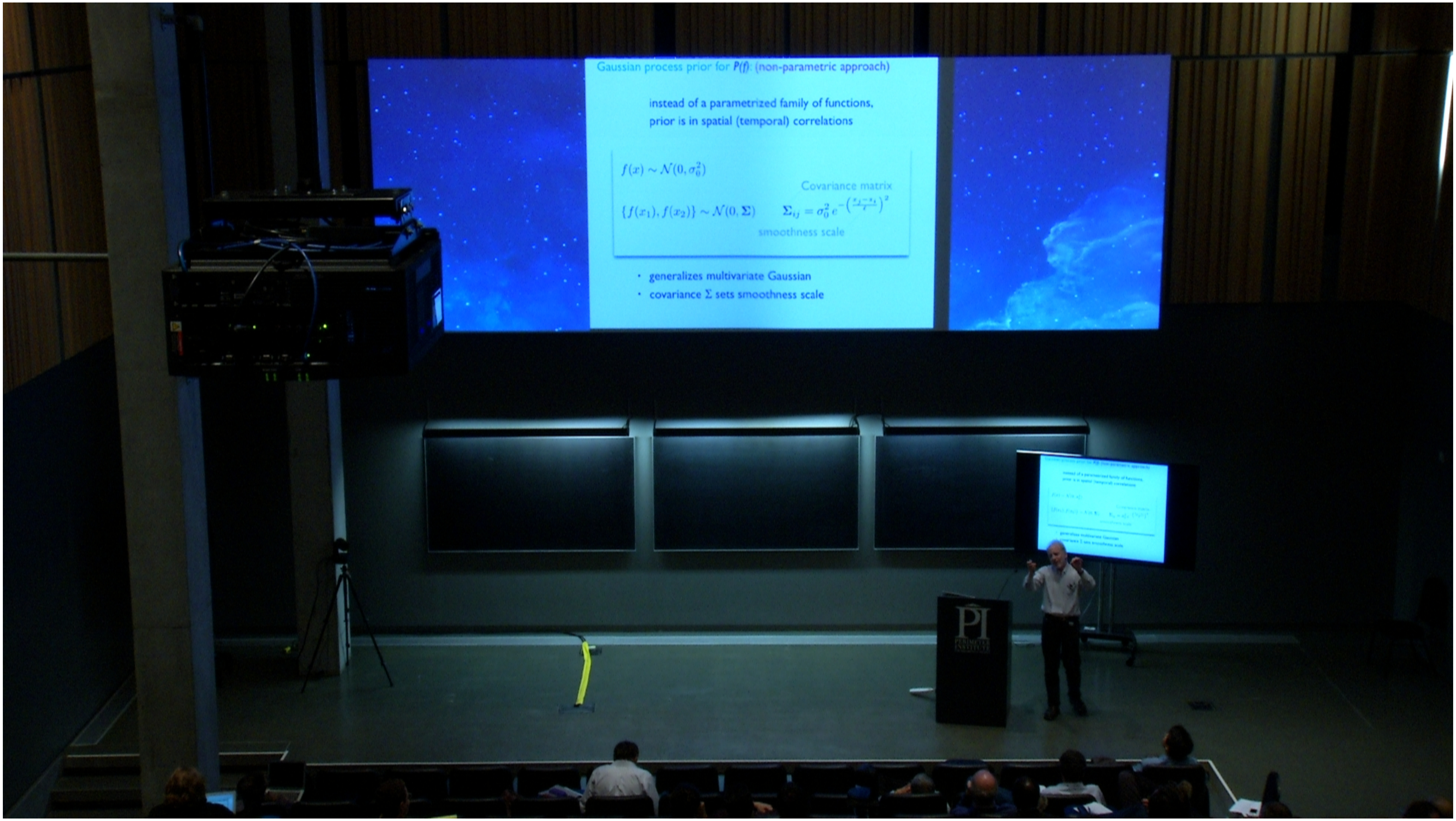
Covariance matrix

$\{f(x_1), f(x_2)\} \sim \mathcal{N}(0, \Sigma)$ $\quad \Sigma_{ij} = \sigma_0^2 \, e^{-\left(\frac{x_i - x_j}{\ell}\right)^2}$

smoothness scale

generalizes multivariate Gaussian

ariance $\Sigma$ sets smoothness scale



Gaussian process prior for *P(f)*: (non-parametric approach)

instead of a parametrized family of functions,
prior is in spatial (temporal) correlations

$f(x) \sim \mathcal{N}(0, \sigma_0^2)$

Covariance matrix

$\{f(x_1), f(x_2)\} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \qquad \boldsymbol{\Sigma}_{ij} = \sigma_0^2 \, e^{-\left(\frac{x_j - x_i}{r}\right)^2}$

smoothness scale

- generalizes multivariate Gaussian
- covariance Σ sets smoothness scale

Inverting the replication program: non-parametric approach

Simulate a replication program

initiation rate

left-moving forks

right-moving forks

initiation density

termination density