LEARNING FROM SEQUENCES WITH POINT PROCESSES

Massil Achab

CMAP, Ecole polytechnique

October 9th, 2017

SEQUENCES ?

Discrete events in **continuous** time: **timestamps**

SEQUENCES ?

Discrete events in continuous time: timestamps

- Social network activity

SEQUENCES ?

Discrete events in continuous time: timestamps

- Social network activity

- Neural spike trains

SEQUENCES ?

Discrete events in **continuous** time: **timestamps**

- Social network activity

- Earthquakes and aftershocks

- Neural spike trains

SEQUENCES ?

Discrete events in **continuous** time: **timestamps**

- Social network activity

- Earthquakes and aftershocks

- Neural spike trains

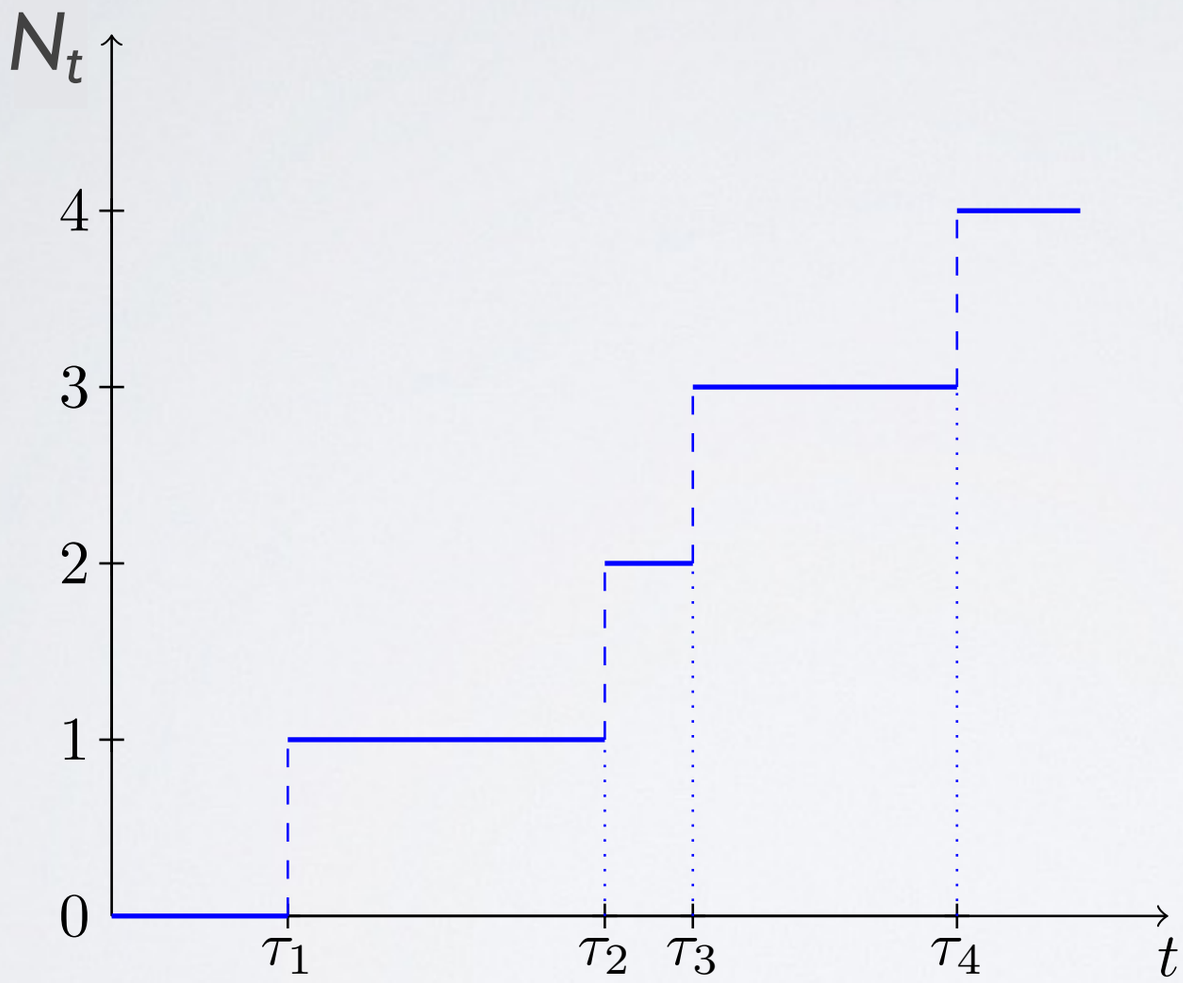
- Financial transactions

POINT PROCESS ?

- The process counting occurrences of those events $(\tau_1, \tau_2, \ldots)$: temporal point process.

POINT PROCESS ?

- The process counting occurrences of those events $(\tau_1, \tau_2, \ldots)$: temporal point process.



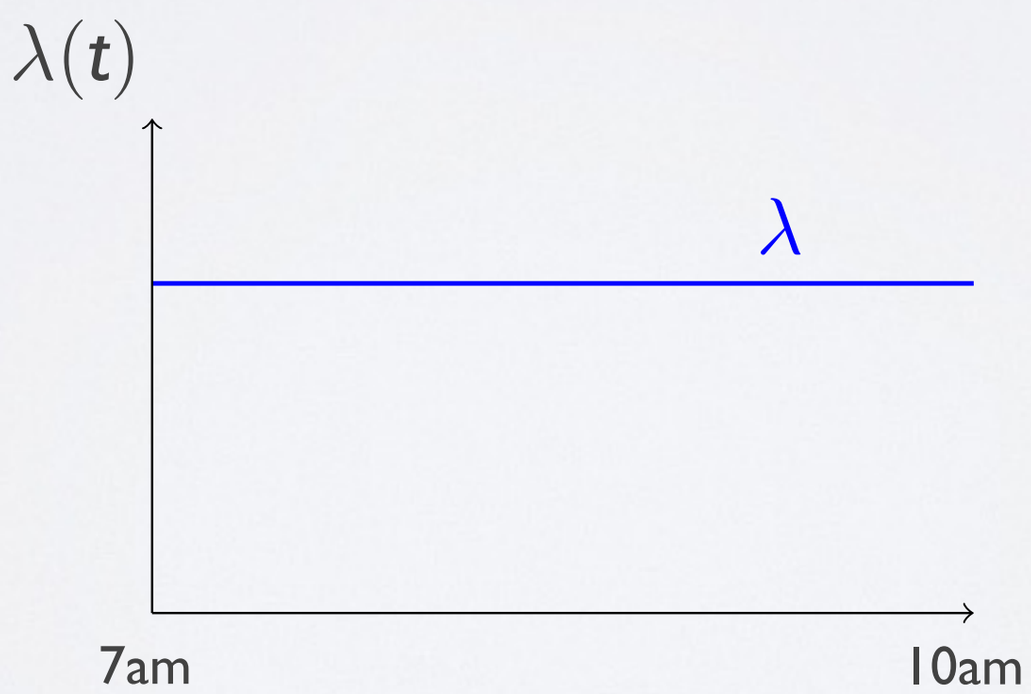
$$N_t = \sum_{\tau \in Z} \mathbf{I}_{\{\tau \leq t\}}$$

- We rather characterize the point process via its arrival rate.

POINT PROCESS ?

POINT PROCESS ?

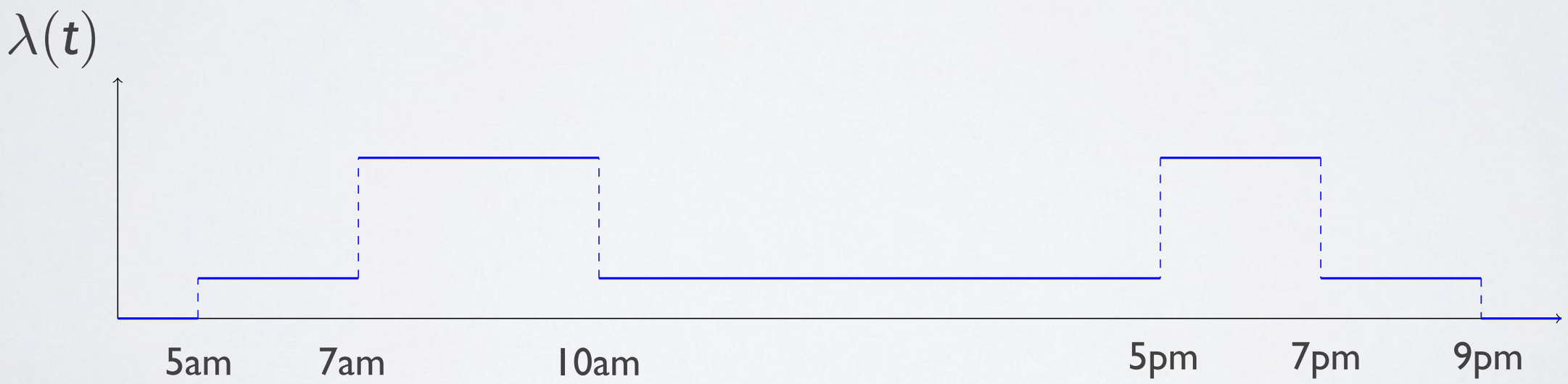
- Between 7am and 10am, the arrival rate is constant.
- $\lambda(t) = \dfrac{\mathbb{E}(N_{t+h} - N_t)}{h} = \lambda > 0$
- Homogeneous Poisson process



POINT PROCESS ?

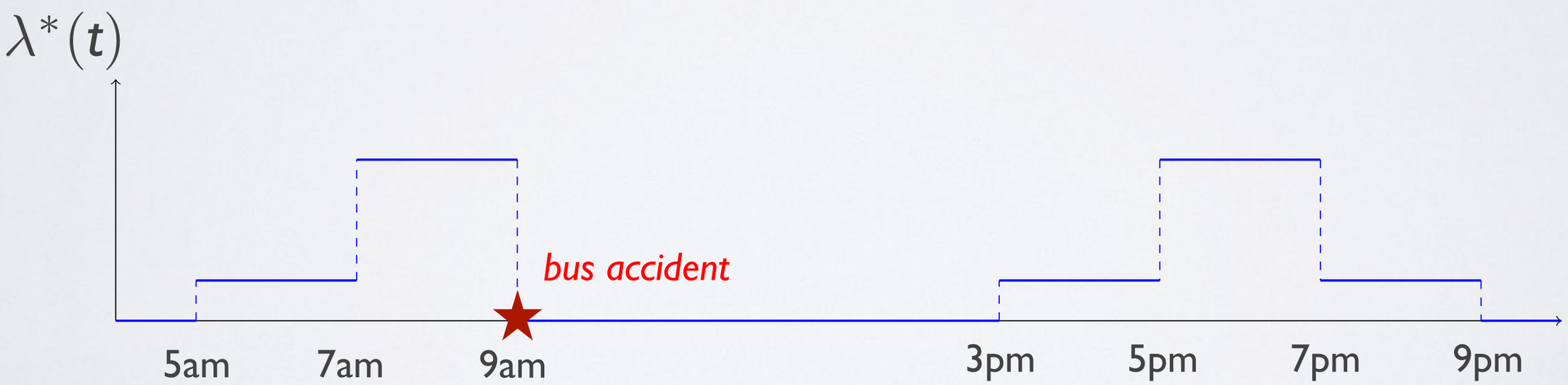
- Over an entire day, the arrival rate is piecewise-constant.
- $$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{E}(N_{t+h} - N_t)}{h} = \lim_{h \to 0} \frac{\mathbb{P}(N_{t+h} - N_t = 1)}{h}$$

- Inhomogeneous Poisson process



POINT PROCESS ?

- When a *bus accident* happens, the arrival rate is impacted.
- $\lambda^*(t) = \lim_{h \to 0} \frac{\mathbb{P}(N_{t+h} - N_t = 1 | \mathcal{F}_{t^-})}{h}$
- $\lambda^*(t)$ is the conditional intensity of the point process.



LEARNING ?

- We observe where and when, and we aim at understanding how and why from the data.

LEARNING ?

- We observe where and when, and we aim at understanding how and why from the data.
- The conditional intensity characterizes the point process.
- The conditional probability density function on the next event writes as a function of the conditional intensity:

$$f^*(t) = \lambda^*(t) \exp\left(-\int_\tau^t \lambda^*(s) ds\right)$$

LEARNING ?

- We observe where and when, and we aim at understanding how and why from the data.

- The conditional intensity characterizes the point process.

- The conditional probability density function on the next event writes as a function of the conditional intensity:

$$f^*(t) = \lambda^*(t) \exp\left(-\int_\tau^t \lambda^*(s) ds\right)$$

- With $(\tau_1, \tau_2, \ldots, \tau_n)$ a realization of the point process on $[0, T]$, the likelihood writes:

$$L = \left[\prod_{i=1}^{n} \lambda^*(\tau_i)\right] \exp\left(-\int_0^T \lambda^*(t) dt\right)$$

QUESTIONS ADDRESSED IN THIS THESIS

- Cox proportional hazards model:
  - Used in medical studies to relate the failure time to individuals' covariables.
  - We observe at most one event (= failure).
  - Can we adapt the estimation algorithm to the large-scale setting ?

QUESTIONS ADDRESSED IN THIS THESIS

- Cox proportional hazards model:
  - Used in medical studies to relate the failure time to individuals' covariables.
  - We observe at most one event (= failure).
  - Can we adapt the estimation algorithm to the large-scale setting ?
- Hawkes process:
  - Originally introduced to model earthquakes occurrences, now also applied in social network analysis, finance and neuroscience.
  - How many events of one type are triggered by one event of another type ?
    - Can we design a non-parametric method that answers to this question ?
    - Can we better understand order book dynamics using such method ?

LARGE-SCALE COX MODEL

LARGE SCALE COX-MODEL

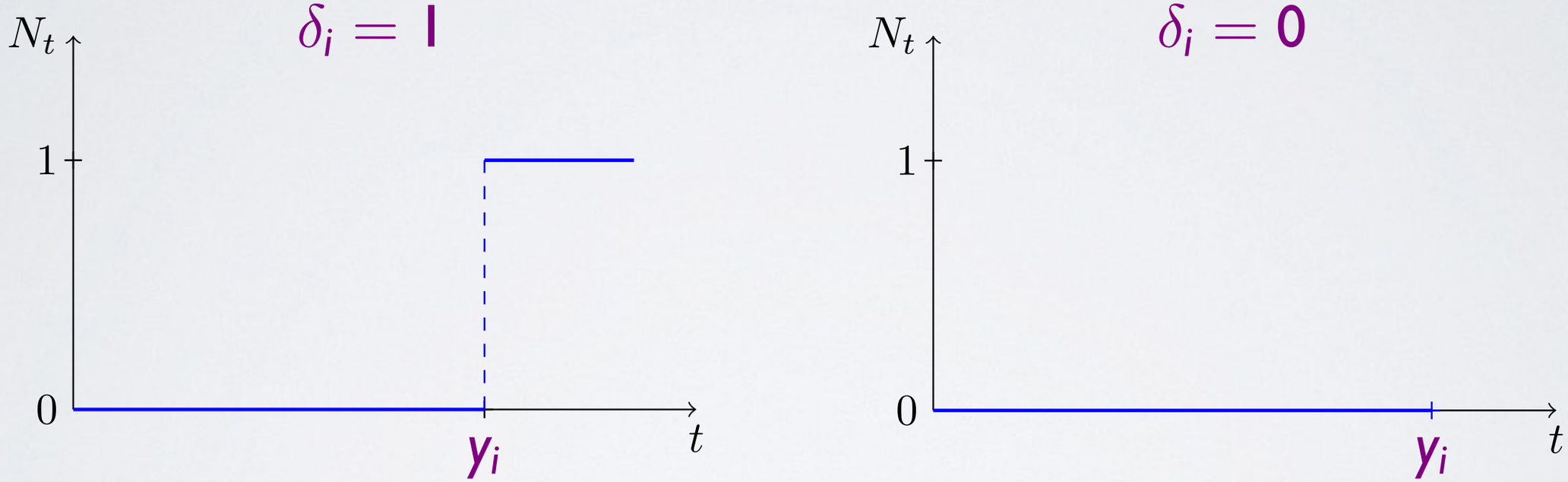
- Medical study with **n** patients (covariates: age, sex, risk factors, etc.)

- We observe typical survival data: $(y_i, x_i, \delta_i)_{i=1}^{n}$.

LARGE SCALE COX-MODEL

- Medical study with **n** patients (covariates: age, sex, risk factors, etc.)

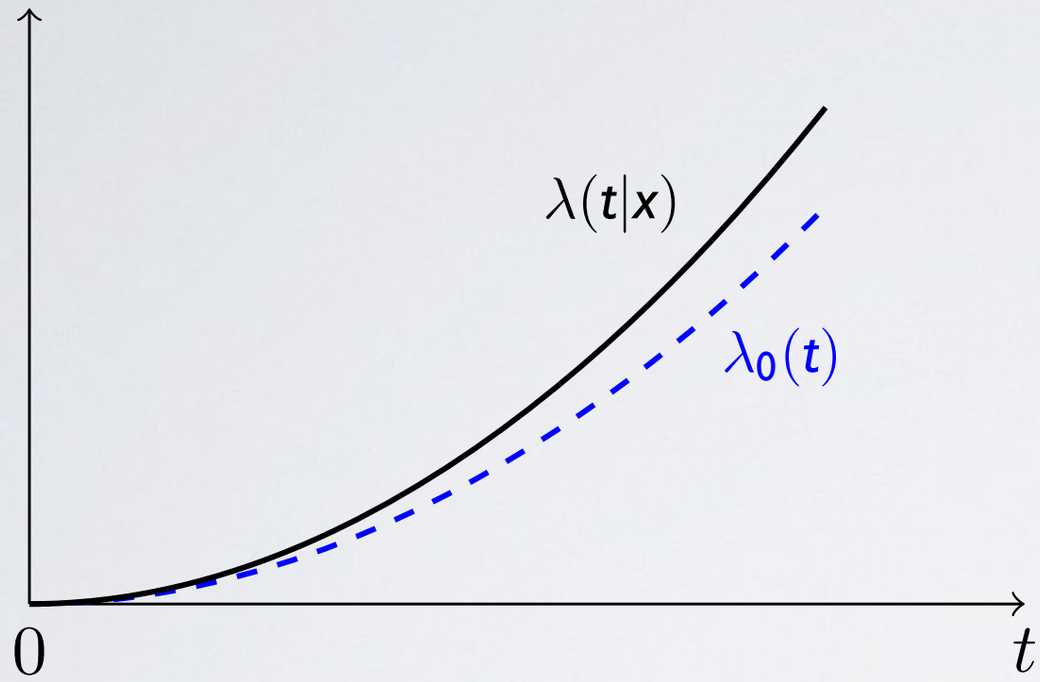
- We observe typical survival data: $(y_i, x_i, \delta_i)_{i=1}^{n}$.



- **Goal:** relate patients' covariates to their survival time, using patients' data with observed and not observed failure time.

LARGE SCALE COX-MODEL

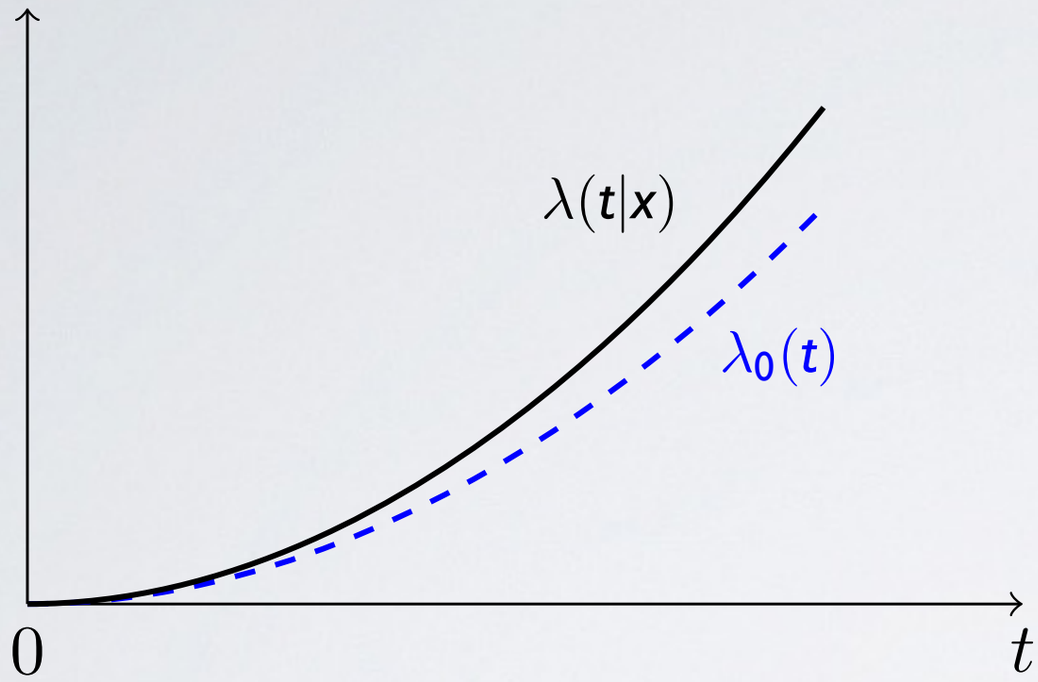
- Cox model: semi-parametric intensity (= hazard ratio)



- $\lambda(t|x) = \lambda_0(t) \exp(\theta^\top x)$

LARGE SCALE COX-MODEL

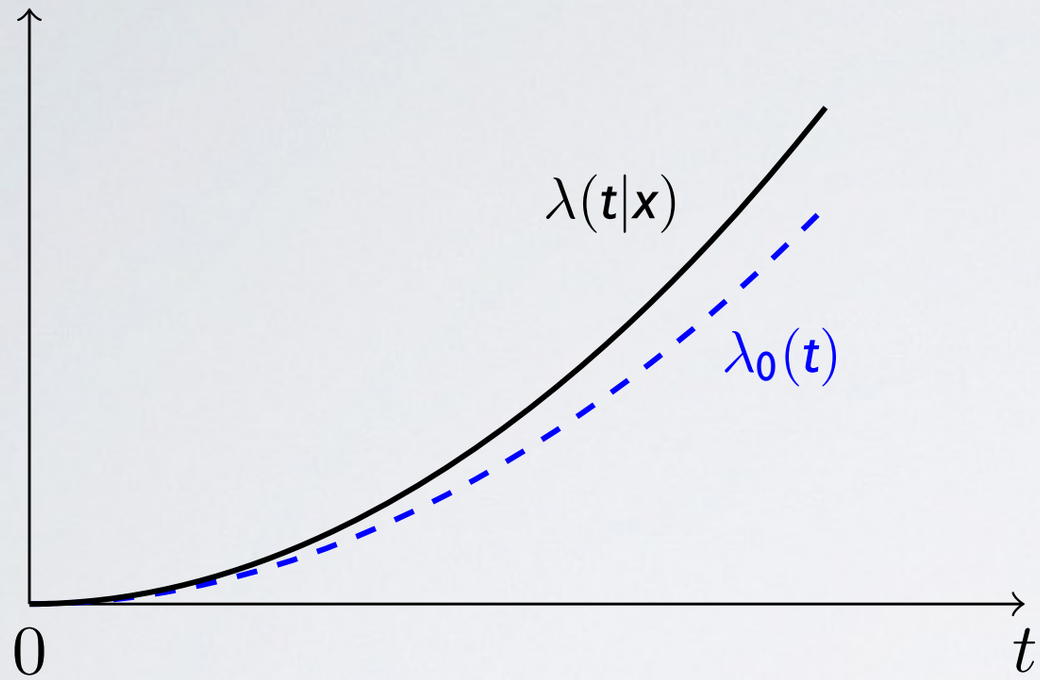
- Cox model: semi-parametric intensity (= hazard ratio)



- $\lambda(t|x) = \lambda_0(t)\exp(\theta^\top x)$
- $\dfrac{\lambda(t|x_i)}{\lambda(t|x_j)} = \exp(\theta^\top(x_i - x_j))$
  $= \exp(\theta_k) \quad \text{if} \quad x_i - x_j = e_k$

LARGE SCALE COX-MODEL

- Cox model: semi-parametric intensity (= hazard ratio)



- $\lambda(t|x) = \lambda_0(t)\exp(\theta^\top x)$
- $\dfrac{\lambda(t|x_i)}{\lambda(t|x_j)} = \exp(\theta^\top(x_i - x_j))$
  $= \exp(\theta_k) \quad \text{if} \quad x_i - x_j = e_k$

- In Regression models and life tables, D.R. Cox enables estimating θ while considering $\lambda_0(t)$ a nuisance parameter by maximizing the (convex) partial likelihood $L(\theta)$.

$$\ell(\theta) = \log L(\theta) = \frac{1}{|D|}\sum_{i\in D}\log\left(\frac{\exp(\theta^\top x_i)}{\sum_{j\in R_i}\exp(\theta^\top x_j)}\right)$$

LARGE SCALE COX-MODEL

$$\ell(\theta) = \log L(\theta) = \frac{1}{|D|} \sum_{i \in D} \log \left(\frac{\exp(\theta^\top x_i)}{\sum_{j \in R_i} \exp(\theta^\top x_j)} \right)$$

- $D = \{i \in [n] \mid \delta_i = 1\}$: patients whose failure time is observed.
- $R_i = \{j \in [n] \mid y_j \geq y_i\}$: patients at risk at time y_i.

LARGE SCALE COX-MODEL

$$\ell(\theta) = \log L(\theta) = \frac{1}{|D|} \sum_{i \in D} \log \left(\frac{\exp(\theta^\top x_i)}{\sum_{j \in R_i} \exp(\theta^\top x_j)} \right)$$

- $D = \{i \in [n] \mid \delta_i = 1\}$: patients whose failure time is observed.
- $R_i = \{j \in [n] \mid y_j \geq y_i\}$: patients at risk at time y_i.
- Maximizing $\log \left(\frac{\exp(\theta^\top x_i)}{\sum_{j \in R_i} \exp(\theta^\top x_j)} \right)$ discriminates patient i versus the other patients in R_i.

LARGE SCALE COX-MODEL

$$\ell(\theta) = \log L(\theta) = \frac{1}{|D|} \sum_{i \in D} \log \left(\frac{\exp(\theta^\top x_i)}{\sum_{j \in R_i} \exp(\theta^\top x_j)} \right)$$

- $D = \{i \in [n] \mid \delta_i = 1\}$: patients whose failure time is observed.
- $R_i = \{j \in [n] \mid y_j \geq y_i\}$: patients at risk at time y_i.
- Maximizing $\log \left(\frac{\exp(\theta^\top x_i)}{\sum_{j \in R_i} \exp(\theta^\top x_j)} \right)$ discriminates patient i versus the other patients in R_i.
- Cox regression: regression on survival time that takes into account patients whose failure time is not observed.

CONVEX OPTIMIZATION 101

- Most machine learning estimation problems can be expressed as a convex optimization problem:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) + g(\theta)$$

- $f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$ is a convex data fitting term (usually smooth), and $g(\theta)$ is convex penalty term on the predictor (smooth or not).

- Example (Lasso): $f_i(\theta) = (y_i - \theta^\top x_i)^2$

$$g(\theta) = \alpha ||\theta||_1$$

CONVEX OPTIMIZATION 101

- Vanilla algorithm to find $\hat{\theta}$: proximal gradient descent.

- The proximal operator of h is defined by

$$\text{prox}_h(y) = \arg\min_{x \in \mathbb{R}^d} h(x) + \frac{1}{2}||y - x||_2^2.$$

- Algorithm:
  - given θ^0 and $\eta > 0$ small enough
  - Until convergence, do

$$\theta^{t+1} \leftarrow \text{prox}_{\eta g}(\theta^t - \eta \nabla f(\theta^t))$$

STOCHASTIC GRADIENT DESCENT

- Modern settings:
  - large-scale: n, number of observations, is large
  - high-dimensional: d, dimension of each observation, is large
- Consequence: computation of $\nabla f(\theta^t) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta^t)$ is time-consuming.
- Idea behind Stochastic Gradient Descent (SGD): replace $\nabla f(\theta^t)$ with a noisy descent direction d^t faster to compute,

$$d^t = \nabla f(\theta^t) + \epsilon^t \quad \text{with} \quad \mathbb{E}(\epsilon^t) = 0$$

STOCHASTIC GRADIENT DESCENT

- The usual version of SGD from Robbins and Monro (1951) writes:
  - sample $i_t \sim \mathcal{U}[n]$
  - compute $d^t = \nabla f_{i_t}(\theta^t) = \nabla f(\theta^t) + \left(\nabla f_{i_t}(\theta^t) - \frac{1}{n}\sum_{j=1}^{n} f_j(\theta^t)\right)$
- For many problems (linear regression, logistic regression, support vector machines, etc.),

$$f_i(\theta) = c(y_i, \theta^\top m(x_i))$$

$$\nabla f_i(\theta) = \partial_2 c(y_i, \theta^\top m(x_i)) m(x_i)$$

- Then, computing $\nabla f_i(\theta)$ is *n* times faster than computing $\nabla f(\theta)$.

SGD'S HIGH VARIANCE

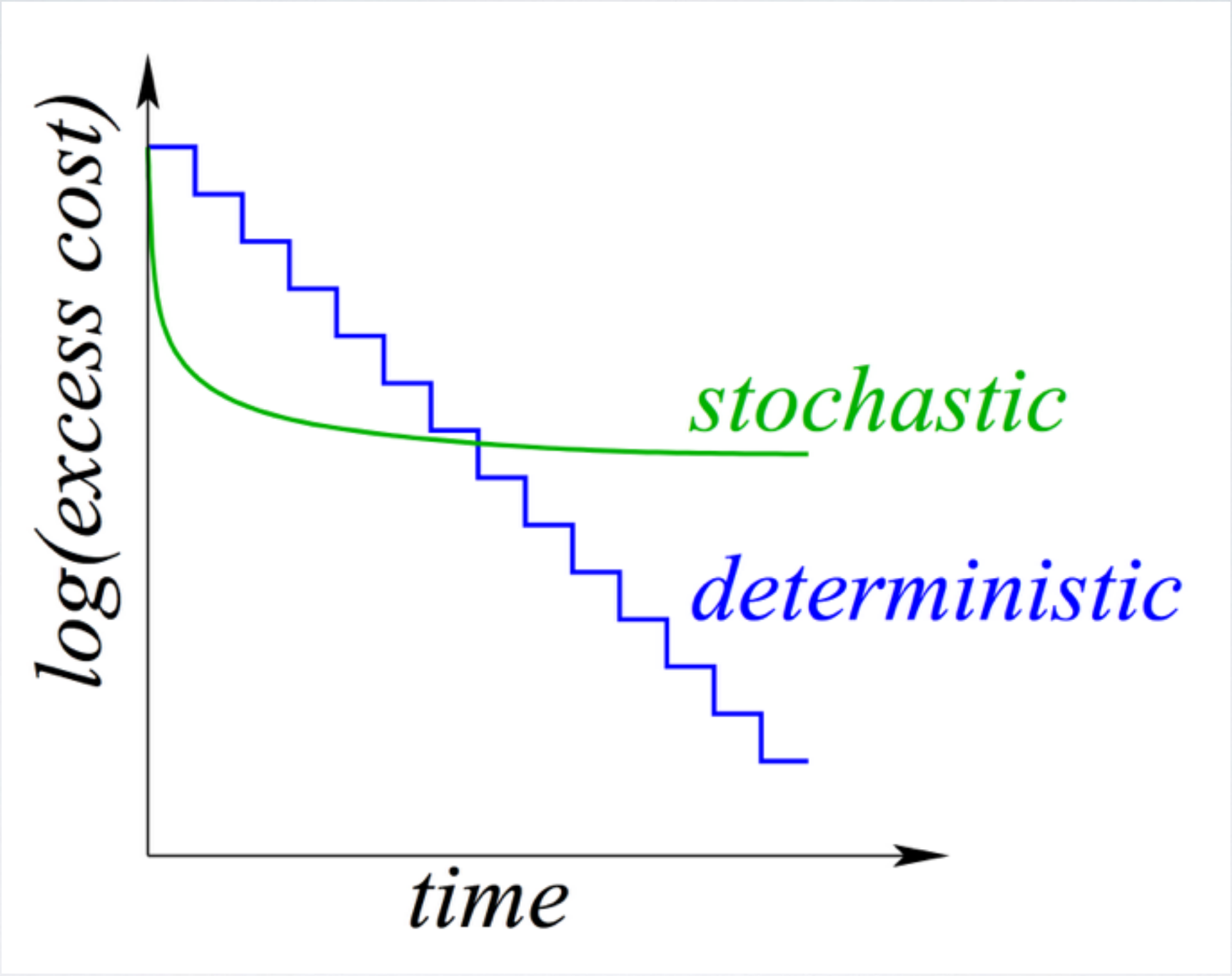


Figure: Picture borrowed from Francis Bach's presentations.

SGD'S HIGH VARIANCE

- **Assumptions:**
  - f is L-smooth i.e.

$$\forall x, y : ||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2,$$

  - f is μ-strongly convex i.e.

$$\forall x, y : f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}||x - y||_2^2.$$

SGD'S HIGH VARIANCE

- **Assumptions:**
  - f is L-smooth i.e.

$$\forall x, y : ||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2,$$

  - f is μ-strongly convex i.e.

$$\forall x, y : f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}||x - y||_2^2.$$

- Convergence rates:

$$\mathbb{E}F(\theta^t) - F(\theta^*) = O(1/t) \text{ for Prox-SGD}$$

$$= O(\rho^t) \text{ with } \rho < 1 \text{ for Prox-GD}$$

- The latter rate is called linear convergence rate.

SGD WITH VARIANCE REDUCTION

- Recent findings - (M. Schmidt , 2013), (A. Defazio, 2014) and (T. Zhang, 2013) - proved that reducing the variance in SGD enables reaching a linear convergence rate.

- Descent directions

$$\text{(SAG)} \qquad \theta \leftarrow \theta - \eta \left(\frac{\nabla f_i(\theta) - y_i}{n} + \frac{1}{n} \sum_{j=1}^{n} y_j \right),$$

$$\text{(SAGA)} \qquad \theta \leftarrow \theta - \eta \left(\nabla f_i(\theta) - y_i + \frac{1}{n} \sum_{j=1}^{n} y_j \right),$$

$$\text{(SVRG)} \qquad \theta \leftarrow \theta - \eta \left(\nabla f_i(\theta) - \nabla f_i(\tilde{\theta}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\tilde{\theta}) \right).$$

- SAG's descent direction is biased $(\alpha = 1/n)$, while SAGA's and SVRG's ones are unbiased $(\alpha = 1)$.

REMARKS

- For Cox model,

$$f(\theta) = -\ell(\theta) = -\frac{1}{|D|}\sum_{i\in D}\log\left(\frac{\exp(\theta^\top x_i)}{\sum_{j\in R_i}\exp(\theta^\top x_j)}\right) = \frac{1}{|D|}\sum_{i\in D} f_i(\theta)$$

$$\nabla f_i(\theta) = \sum_{j\in R_i}\pi_\theta^i(j)(x_j - x_i) \quad \textbf{with} \quad \pi_\theta^i(j) = \frac{\exp(\theta^\top x_j)}{\sum_{k\in R_i}\exp(\theta^\top x_k)}$$

- The gradient $\nabla f_i(\theta)$ writes as an expectation and its exact computation may be expensive.

REMARKS

- For Cox model,

$$f(\theta) = -\ell(\theta) = -\frac{1}{|D|}\sum_{i\in D}\log\left(\frac{\exp(\theta^\top x_i)}{\sum_{j\in R_i}\exp(\theta^\top x_j)}\right) = \frac{1}{|D|}\sum_{i\in D} f_i(\theta)$$

$$\nabla f_i(\theta) = \sum_{j\in R_i}\pi^i_\theta(j)(x_j - x_i) \quad \textbf{with} \quad \pi^i_\theta(j) = \frac{\exp(\theta^\top x_j)}{\sum_{k\in R_i}\exp(\theta^\top x_k)}$$

- The gradient $\nabla f_i(\theta)$ writes as an expectation and its exact computation may be expensive.

	comp. $\nabla f_i(\theta)$	comp. $\nabla f(\theta)$
lin. / log. reg., SVMs, neural	1	n
Cox model	$\lvert R_i \rvert$	n

OUR APPROACH: HYBRID SVRG

- We designed a new SGD-like algorithm,
  - replacing $\nabla f_i(\theta)$ with an estimate $\widehat{\nabla} f_i(\theta)$, using a MCMC method.
  - adding the SVRG term $-\nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$ to the descent direction.

- ApproxMCMC(θ, i, N) outputs an approximation of $\nabla f_i(\theta)$ using N iterations of a MCMC. We focused on three implementations:
  - Independent Metropolis-Hastings (IMH) (unif. and adap. proposals)
  - Adaptative Importance Sampling (AIS).

Algorithm Hybrid SVRG

1: **for** $k = 1$ **to** K **do**
2: **for** $t = 0$ **to** $m - 1$ **do**
3: Pick $i \sim \mathcal{U}[n]$
4: $\widehat{\nabla} f_i(\theta^t) \leftarrow \textsc{ApproxMCMC}(\theta^t, i, N_k)$.
5: $d^t = \widehat{\nabla} f_i(\theta^t) - \nabla f_i(\tilde{\theta}) + \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(\tilde{\theta})$
6: $\omega^{t+1} \leftarrow \theta^t - \gamma d^t$
7: $\theta^{t+1} \leftarrow \text{prox}_{\gamma g}(\omega^{t+1})$
8: **end for**
9: Update $\tilde{\theta} \leftarrow \frac{1}{m} \sum_{t=1}^{m} \theta^t$, $\theta^0 \leftarrow \tilde{\theta}$
10: Compute $\nabla f_i(\tilde{\theta})$ for $i = 1, \ldots, n$
11: **end for**

THEORETICAL GUARANTEES

- **Assumption:** Denoting $\eta^t = \widehat{\nabla} f_{i_t}(\theta^{t-1}) - \nabla f_{i_t}(\theta^{t-1})$, we assume that the bias and the expected squared error can be bounded this way:

$$||\mathbb{E}_t \eta^t||_2 \leq \frac{C_1}{N_k} \quad \text{and} \quad \mathbb{E}_t||\eta^t||_2^2 \leq \frac{C_2}{N_k},$$

where N_k is the length of the Markov chain.

- **Proposition:** Suppose there exists $M > 0$ such that the proposal Q and the stationary distribution π satisfy $\pi(x) \leq M Q(x)$, for all x in the support of π. Then, the error η^t obtained by the algorithm IMH satisfies the previous assumption.

THEORETICAL GUARANTEES

- **Theorem:** Suppose that $F = f + g$ is μ-strongly convex. Consider algorithm **HSVRG**, with a phase length m and a step-size $\gamma \in (0, \frac{1}{16L})$, satisfying

$$\rho = \frac{1}{m\gamma\mu(1-8L\gamma)} + \frac{8L\gamma(1+1/m)}{1-8L\gamma} < 1.$$

Assuming there exists $B > 0$ such that $\sup_{t\geq 0} ||\theta^t - \theta^*||_2 \leq B$, then:

$$\mathbb{E}F(\tilde{\theta}^K) - F(\theta^*) \leq (F(\theta^0) - F(\theta^*))\rho^K + \frac{3\gamma C_2 + BC_1}{1-8L\gamma} \sum_{k=1}^{K} \rho^{K-k} \frac{1}{N_k}.$$

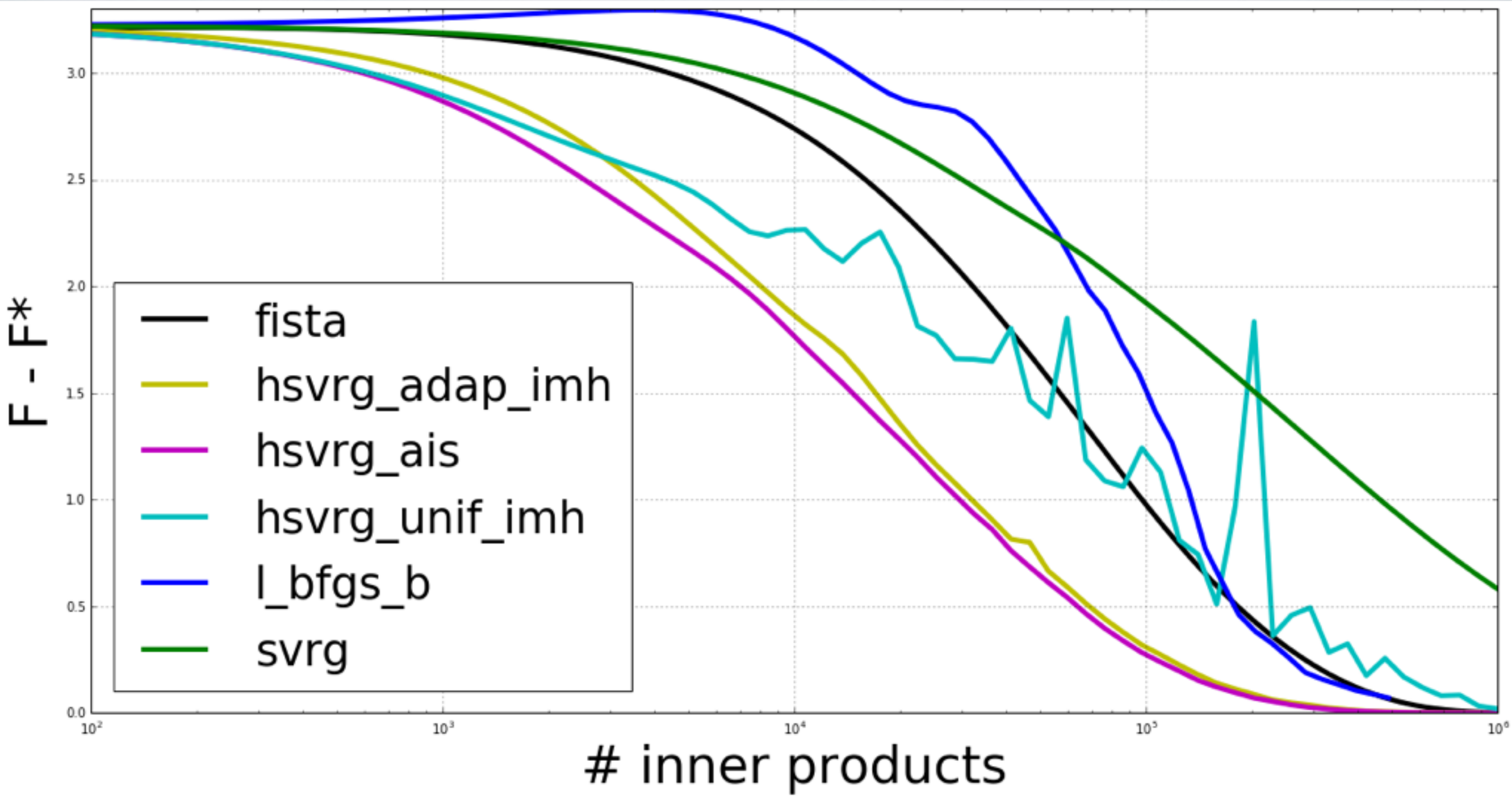
- **Corollary:** The choice $N_k = k^\alpha / \rho^k$ with $\alpha > 1$ gives

$$\mathbb{E}F(\tilde{\theta}^K) - F(\theta^*) \leq \left(F(\theta^0) - F(\theta^*) + \frac{3\gamma C_2 + BC_1}{1-8L\gamma}\right) \rho^K.$$

Then, **HSVRG** achieves a linear rate under strong convexity.

EXPERIMENTS

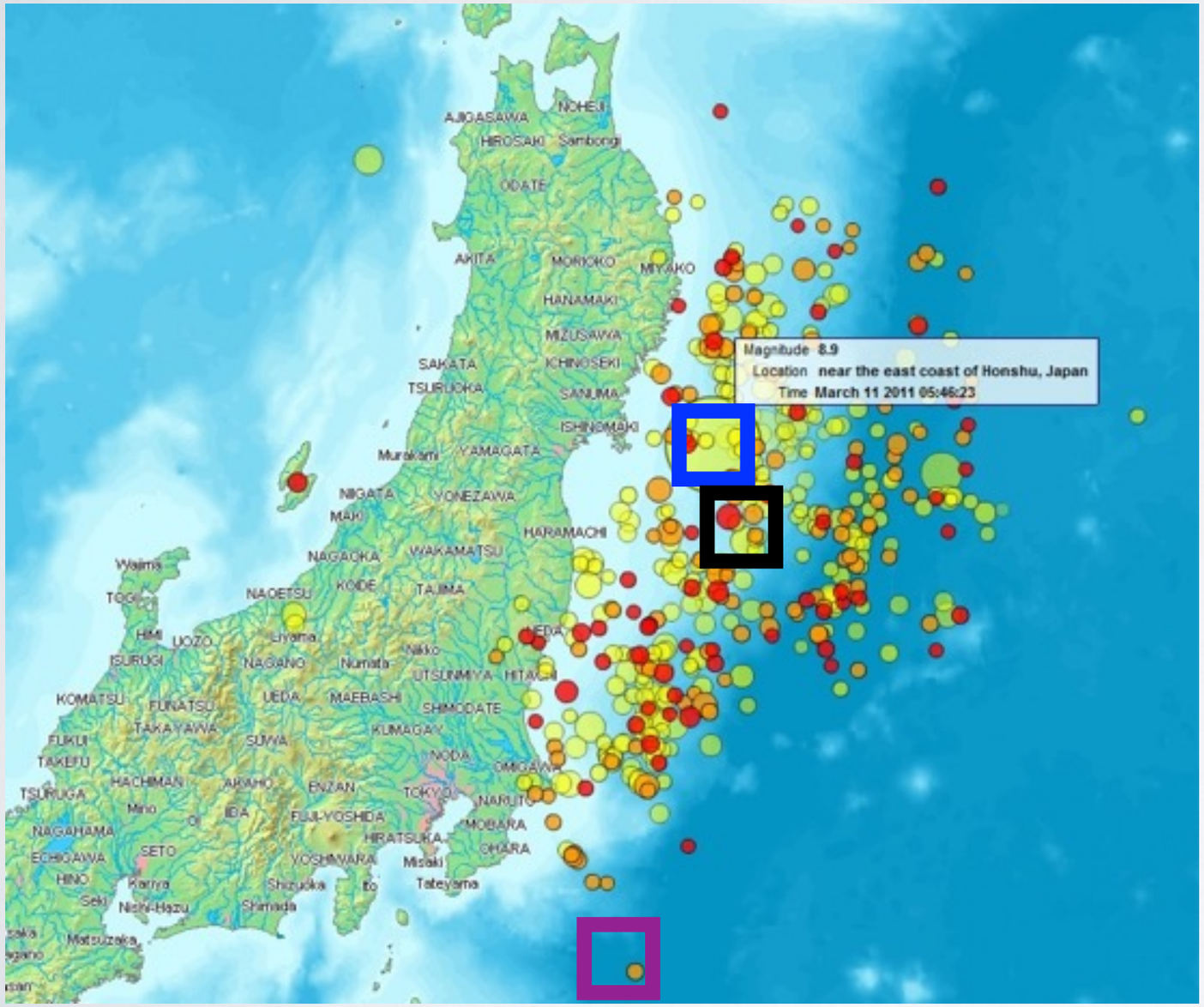
Experiments on a real dataset, with the three implementations of ApproxMCMC.



NON-PARAMETRIC ESTIMATION OF HAWKES CAUSALITY

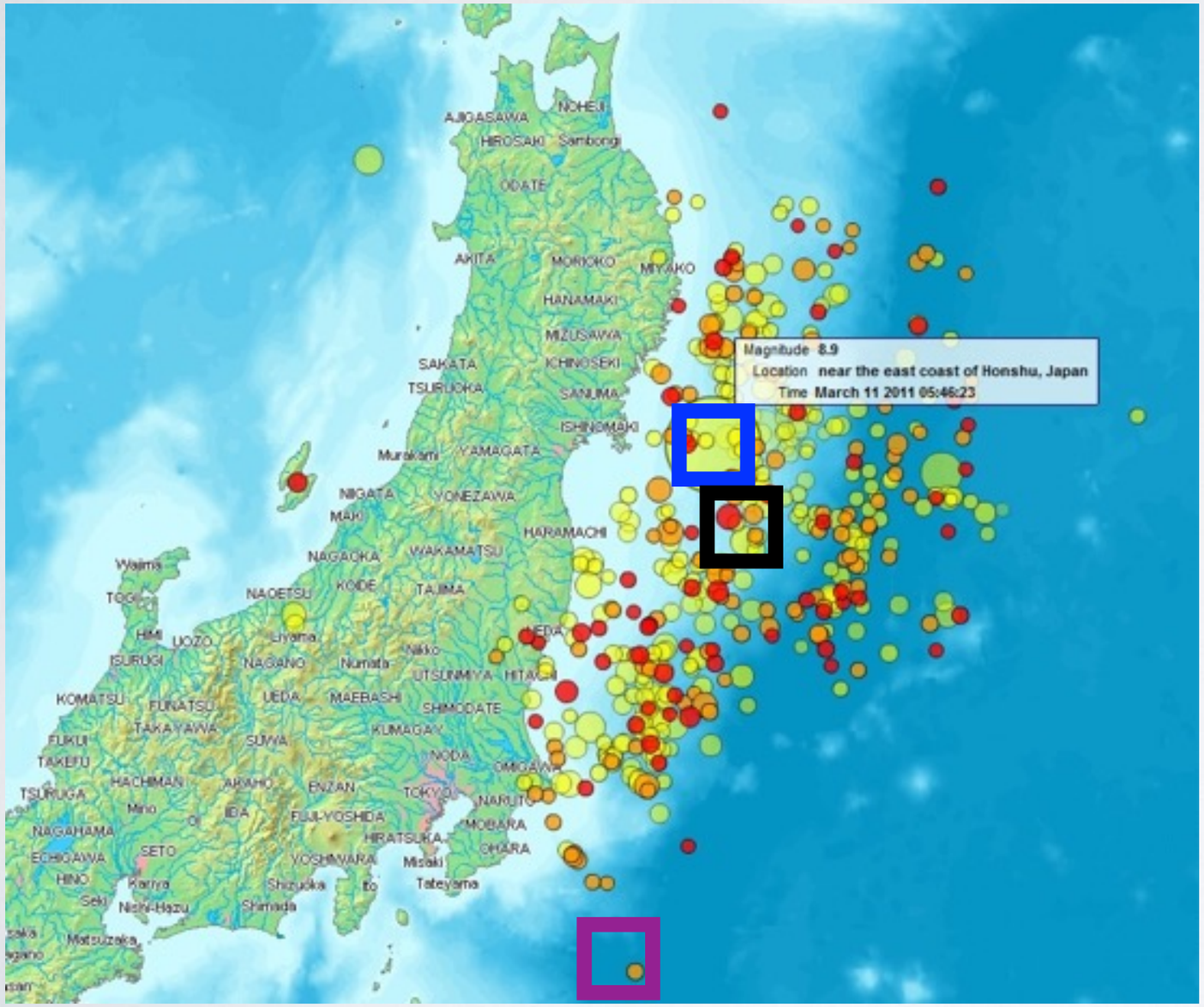
EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:



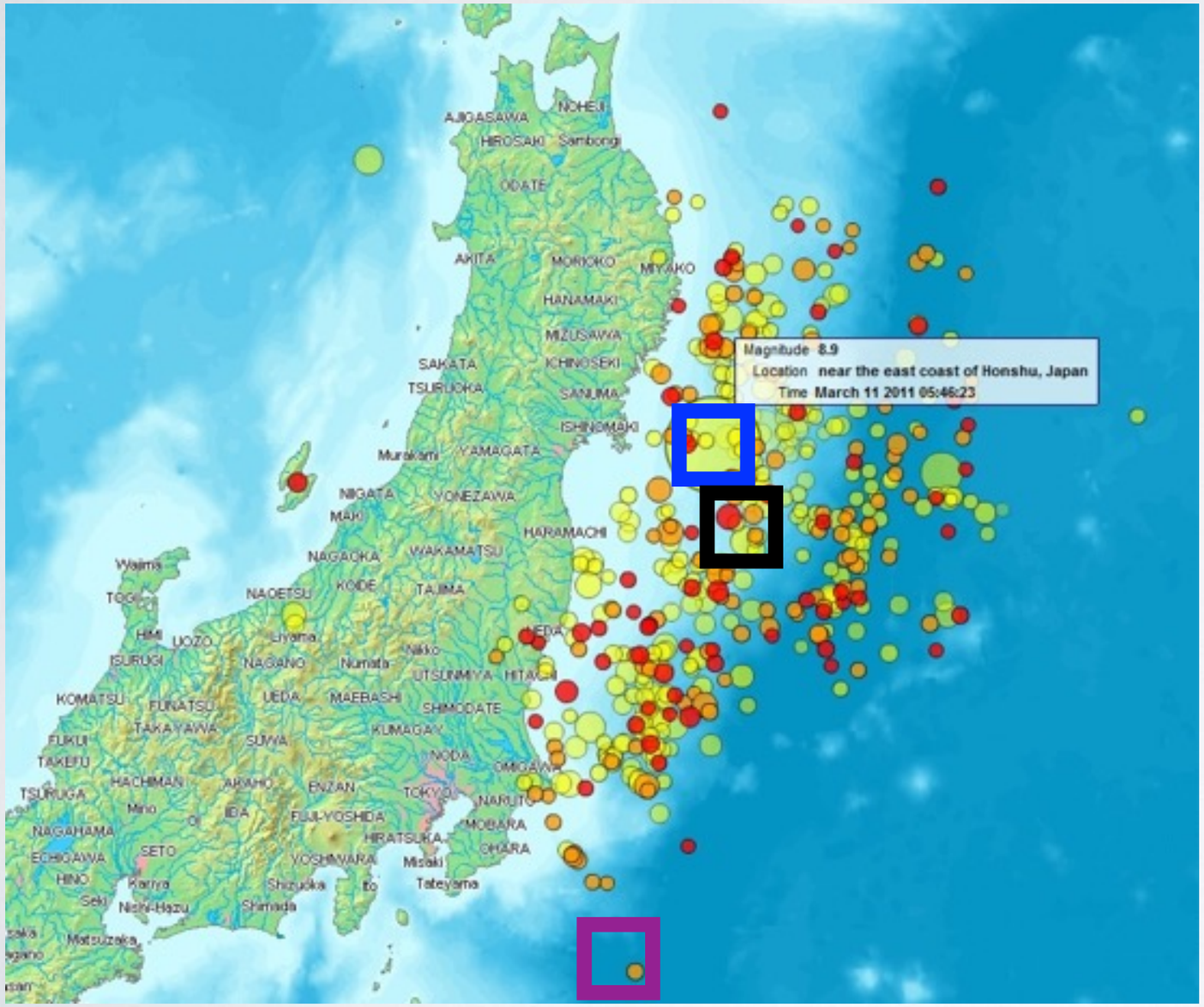
EARTHQUAKES

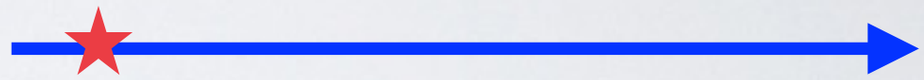
- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

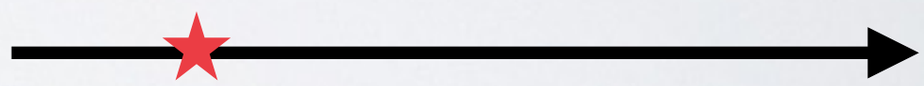


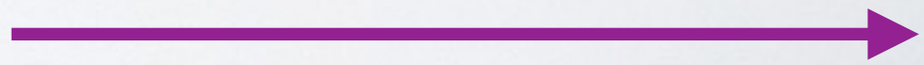
EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:



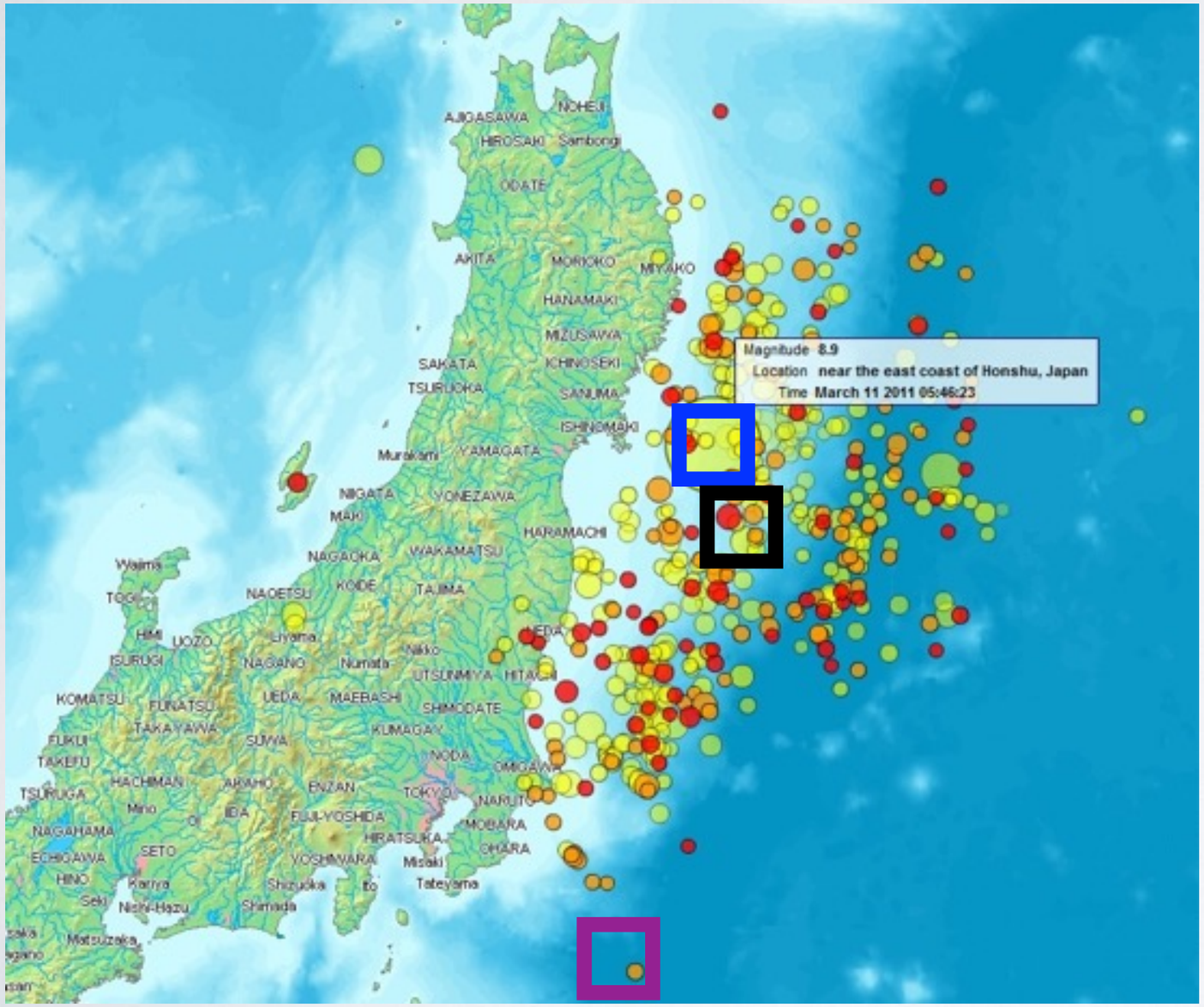


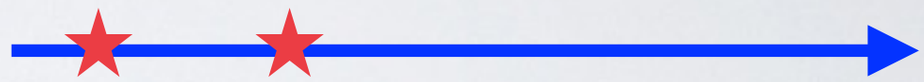


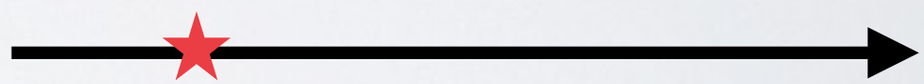


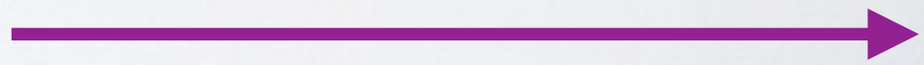
EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:



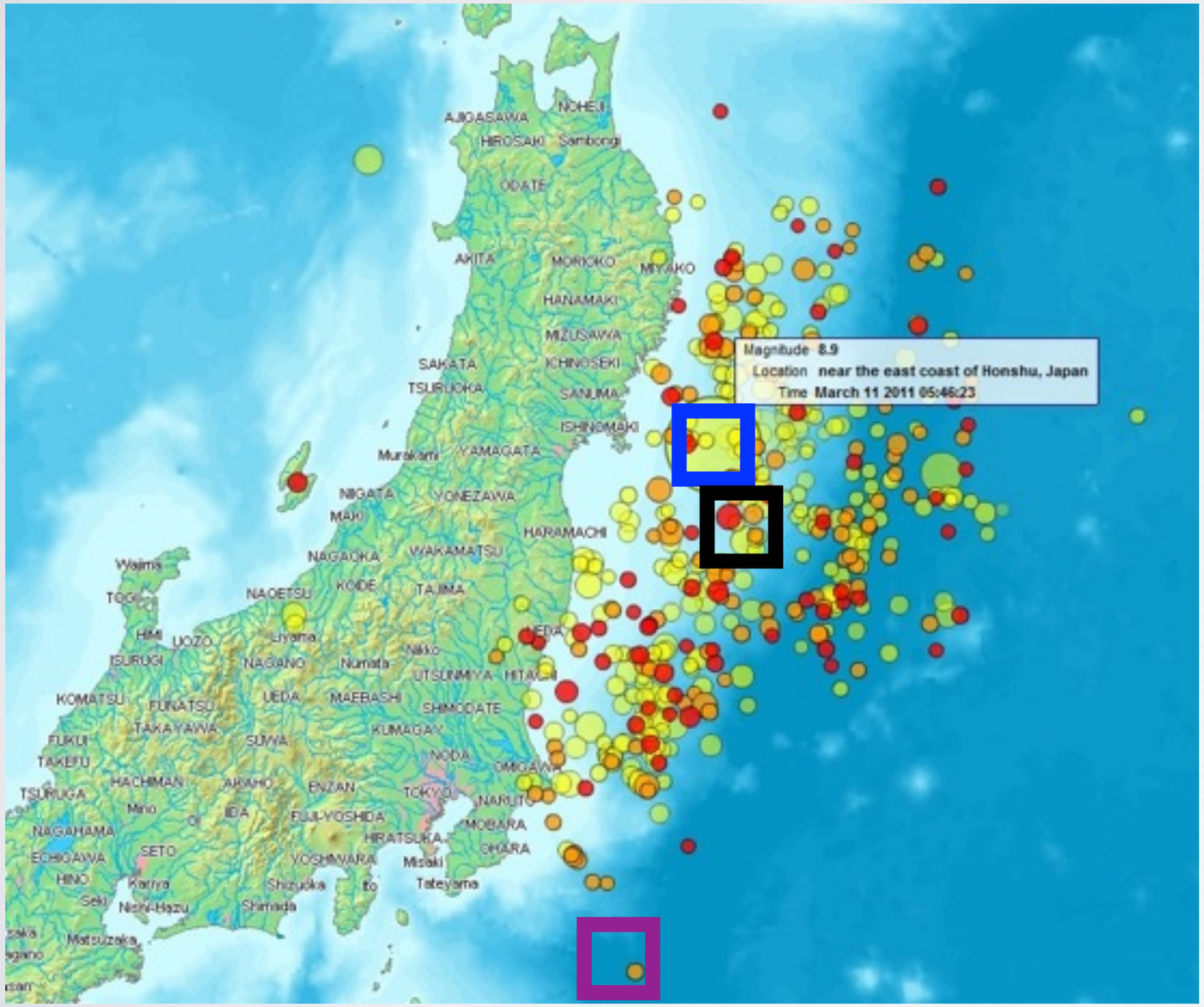


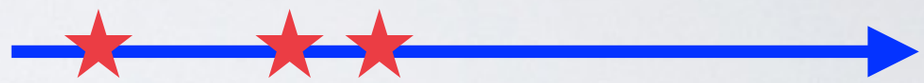


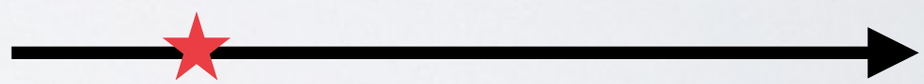


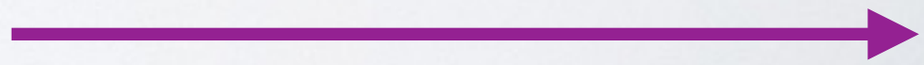
EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:



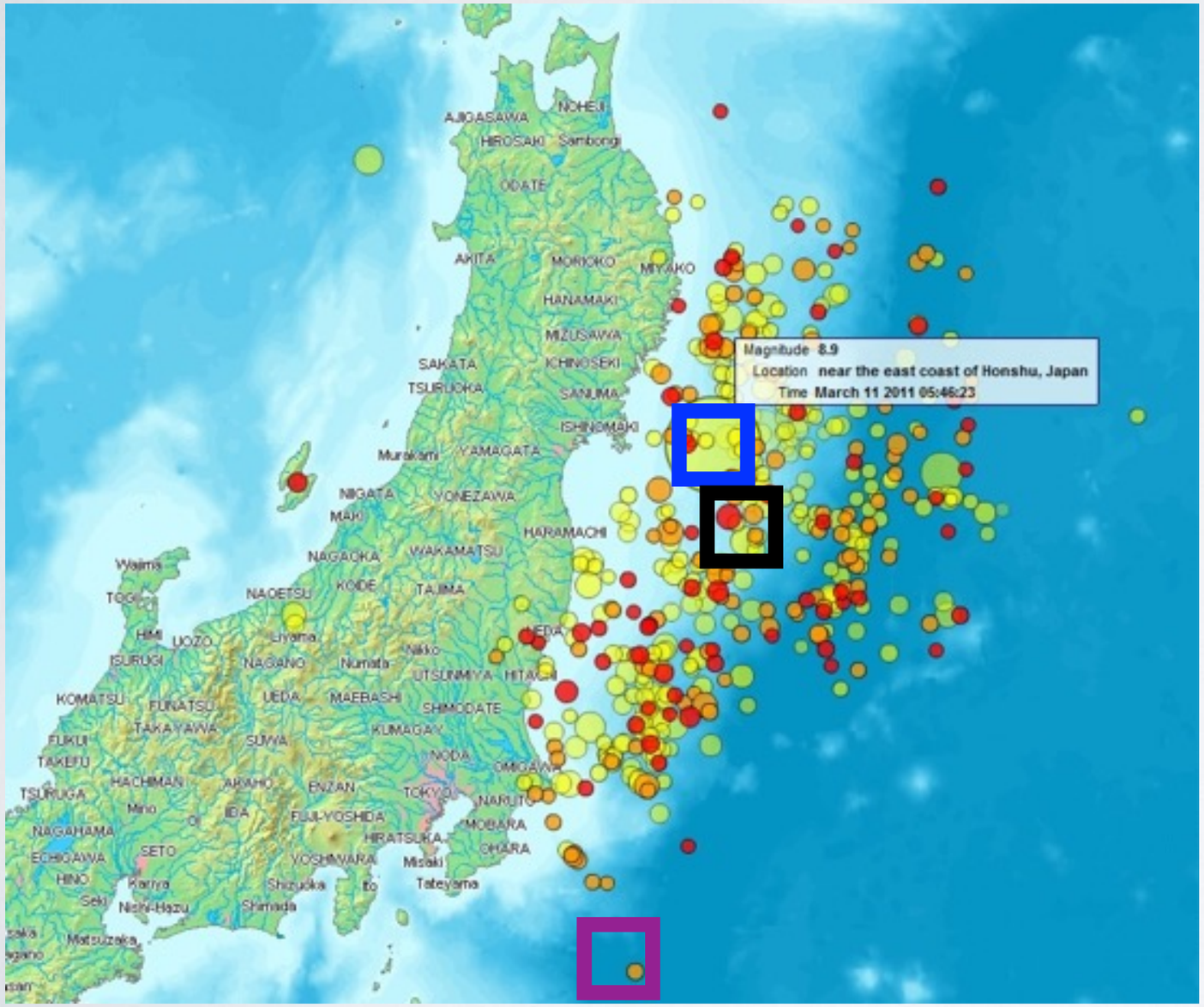


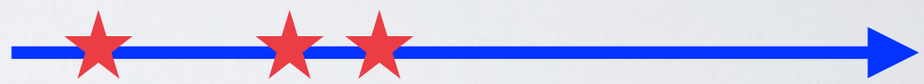


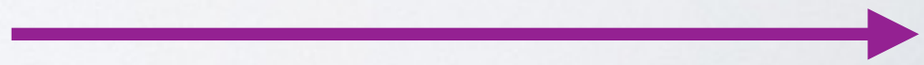


EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

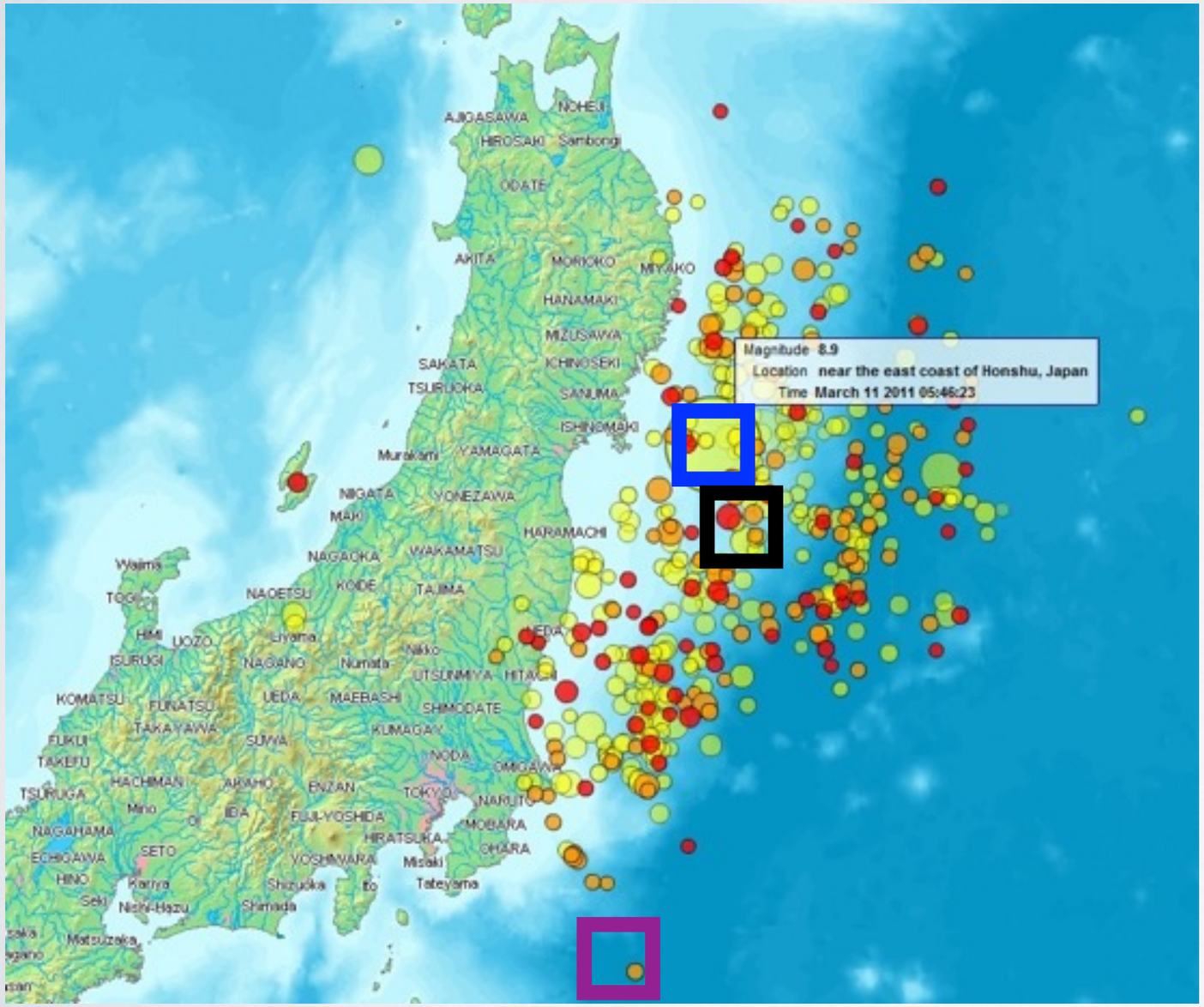


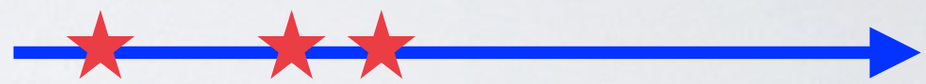


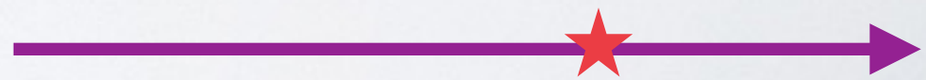


EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

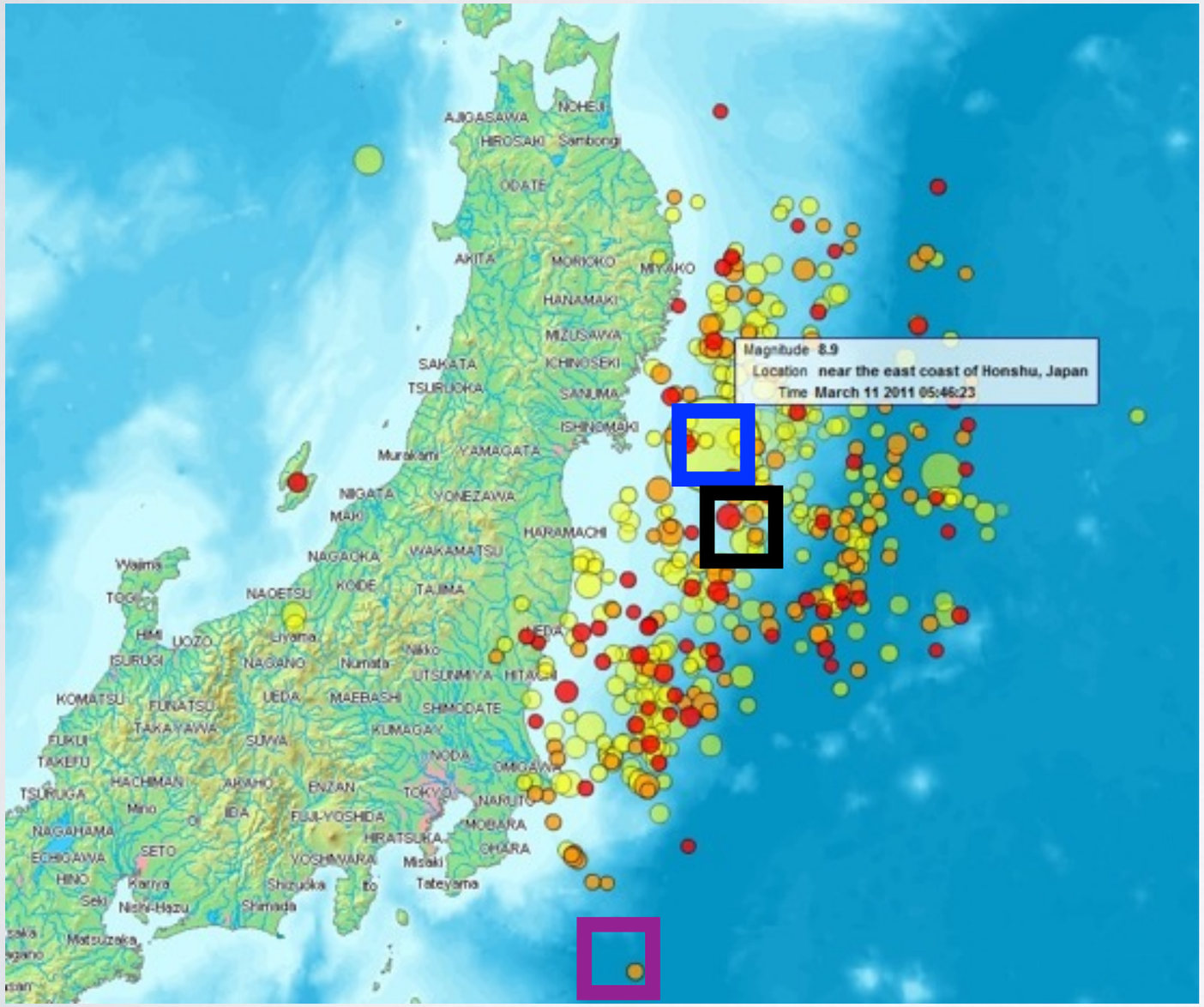


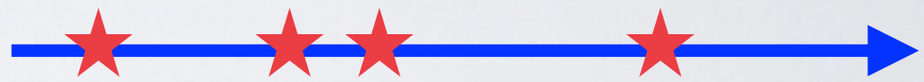


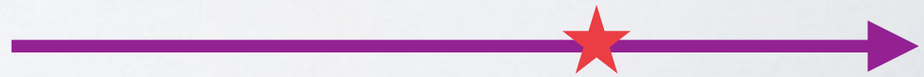


EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

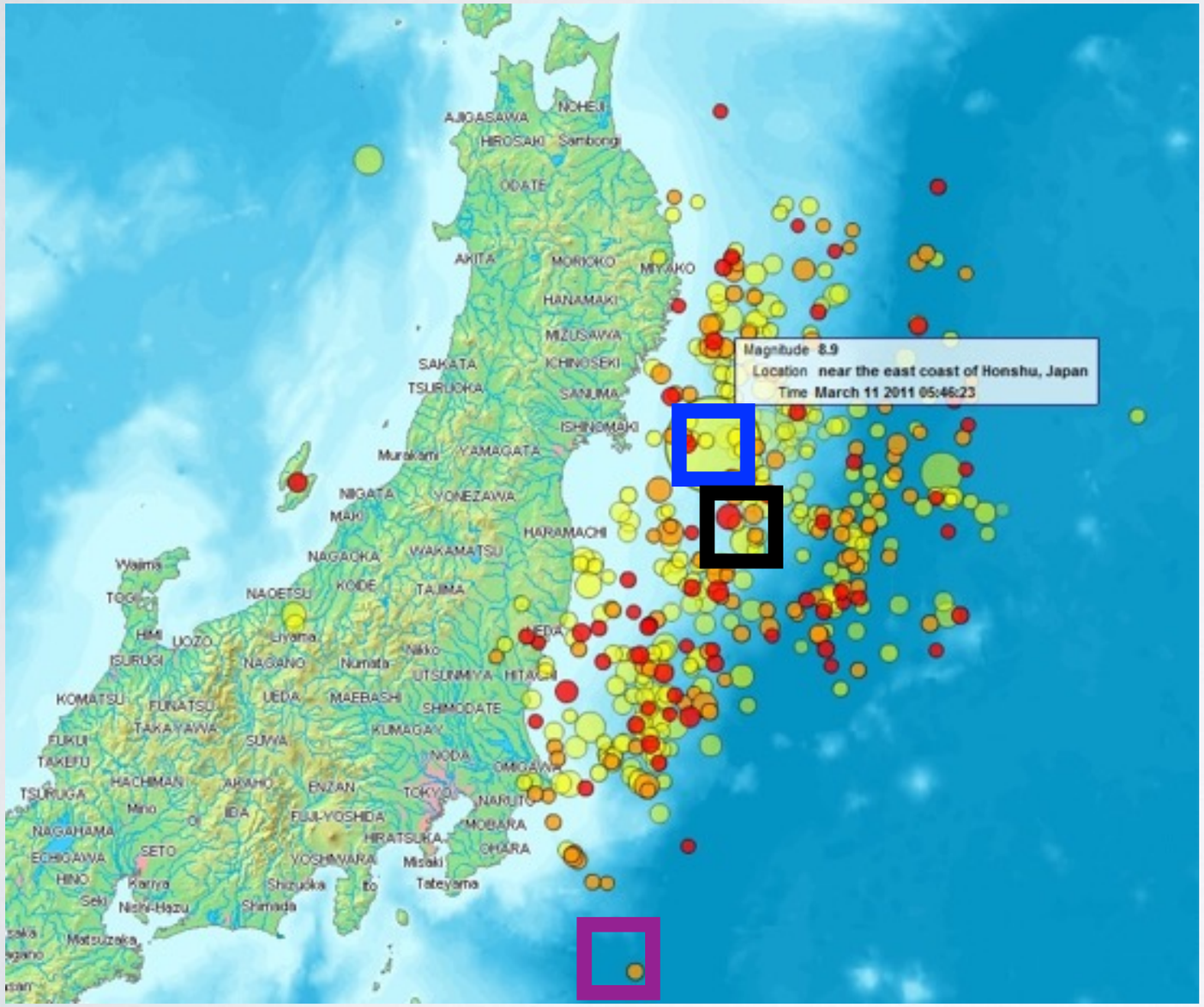


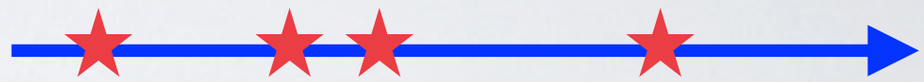


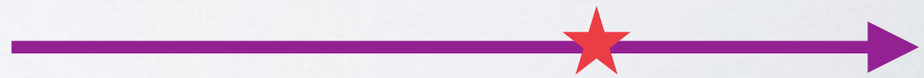


EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

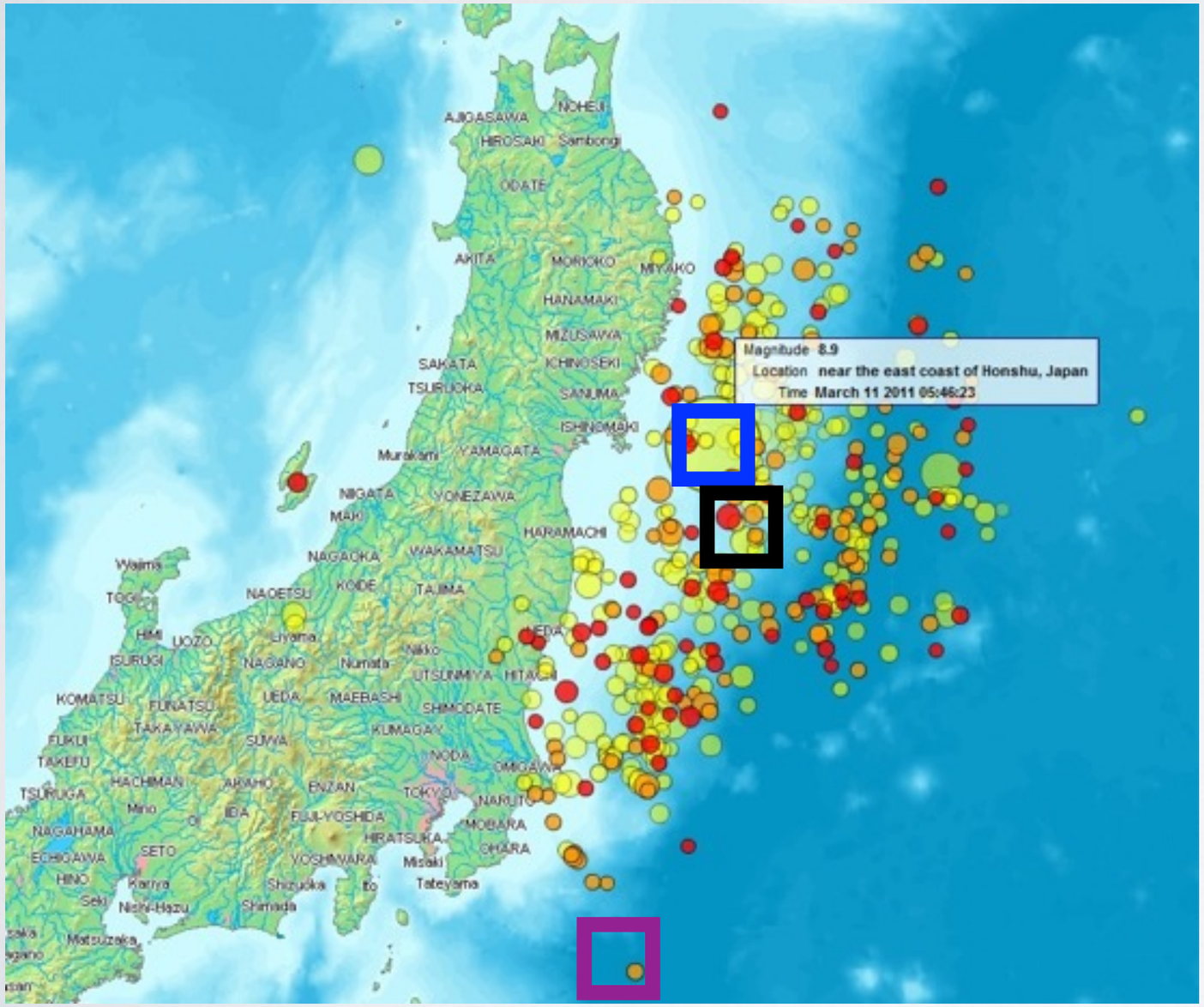


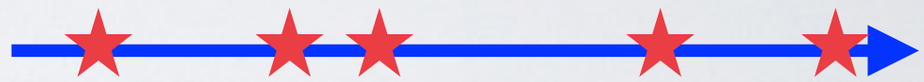


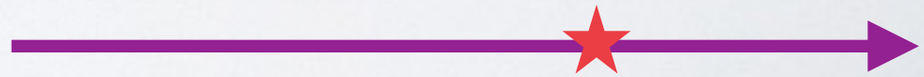


EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:

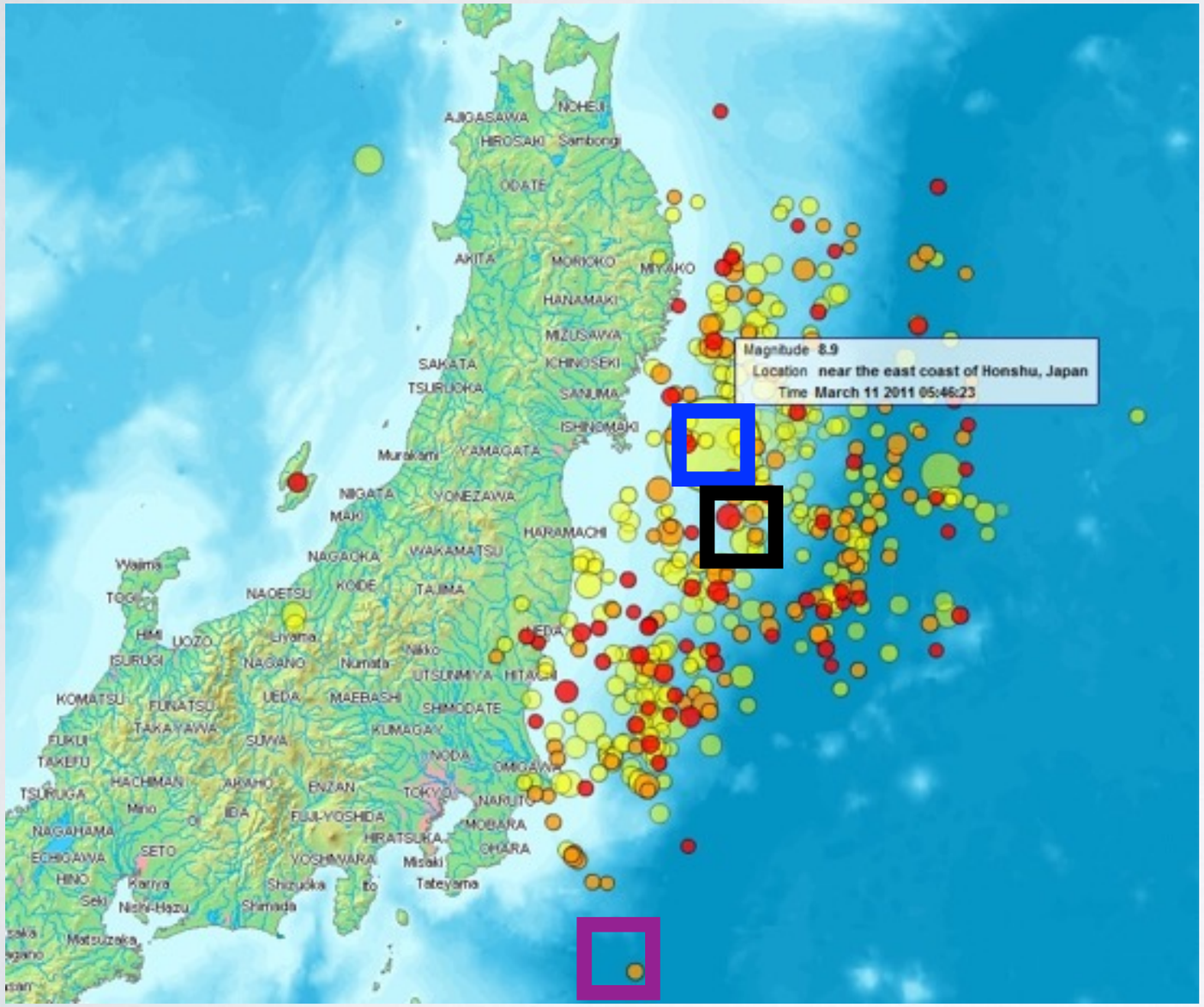


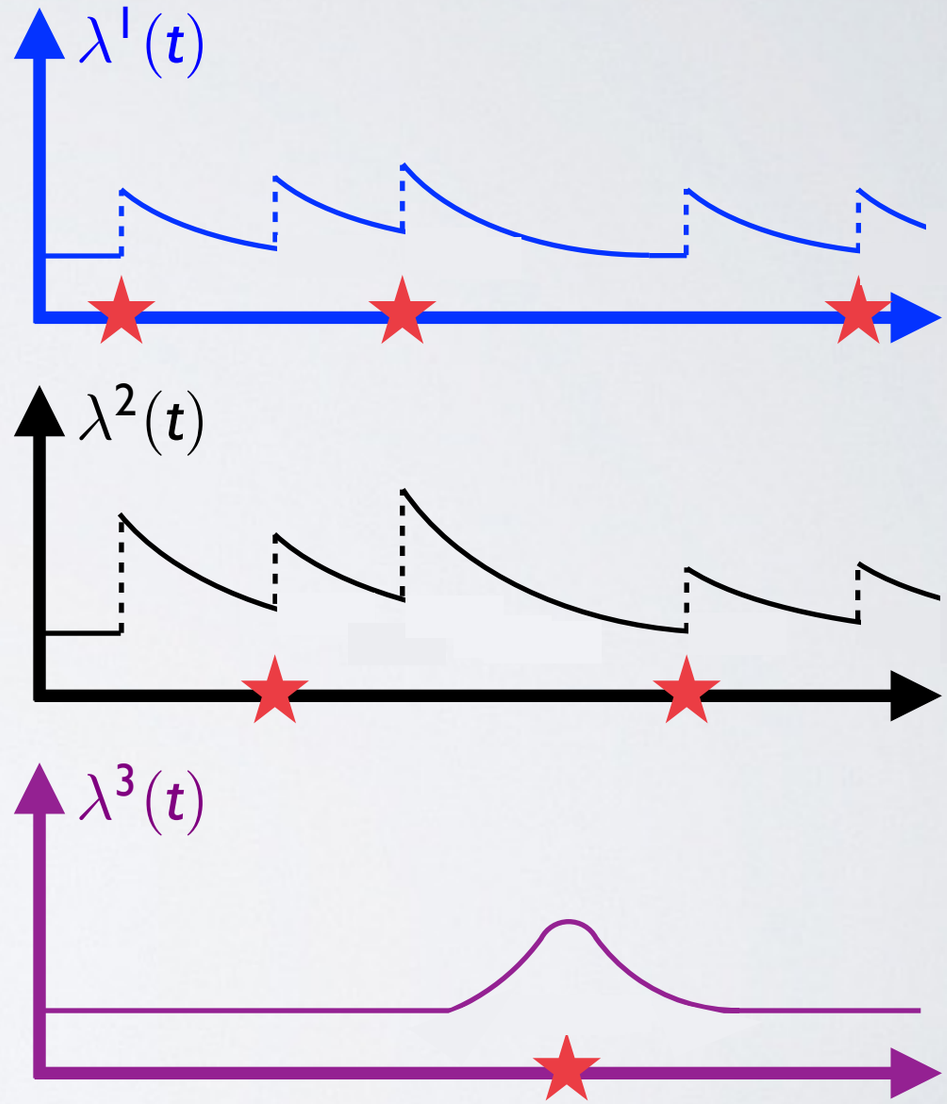




EARTHQUAKES

- For several processes, the arrival rate depends on their history.
- Earthquakes and aftershocks:





HAWKES PROCESS

- The vector of stochastic intensities $[\lambda^1(t), \ldots, \lambda^D(t)]$ associated with the multivariate point process N_t is defined as

$$\lambda^i(t) = \lim_{h \to 0} \frac{\mathbb{P}(N^i_{t+h} - N^i_t = 1 | \mathcal{F}_{t^-})}{h}.$$

- The multivariate point process N_t is a Hawkes point process if the stochastic intensities can be written as

$$\lambda^i(t) = \mu^i + \sum_{j=1}^{D} \sum_{\tau_j \in Z^j} \phi^{i \leftarrow j}(t - \tau_j),$$

where $\mu^i \in \mathbb{R}^+$ is an exogenous intensity, and $\phi^{i \leftarrow j}$ are positive, integrable and supported in $\mathbb{R}^+$ functions called kernels, encoding the impact of an event of *j* on the activity of *i*.

EXISTING ESTIMATION METHODS

Two main classes of estimation methods for Hawkes kernels:

EXISTING ESTIMATION METHODS

Two main classes of estimation methods for Hawkes kernels:

- the parametric one:
  - parametrization of the kernels (usually $\phi^{i \leftarrow j}(t) = \alpha_{ij} \exp(-\beta t)$)
  - parameters obtained by maximizing the likelihood
    - fast and concave log-lik. if exponential but not flexible, nor realistic

EXISTING ESTIMATION METHODS

Two main classes of estimation methods for Hawkes kernels:

- the parametric one:
  - parametrization of the kernels (usually $\phi^{i \leftarrow j}(t) = \alpha_{ij} \exp(-\beta t)$)
  - parameters obtained by maximizing the likelihood
    - fast and concave log-lik. if exponential but not flexible, nor realistic
- the non-parametric one:
  - various methods: maximize the likelihood with EM-algorithm, solve Wiener-Hopf equations, minimize contrast function.
  - flexible but computationally expensive when $D \gtrsim 10$

HAWKES CAUSALITY

- Instead of trying to estimate the kernels $\phi^{i \leftarrow j}$, we focus on the direct estimation of their *integrals*:

$$g^{ij} = \int_0^\infty \phi^{i \leftarrow j}(t)\,dt \qquad\qquad G = [g^{ij}].$$

HAWKES CAUSALITY

- Instead of trying to estimate the kernels $\phi^{i \leftarrow j}$, we focus on the direct estimation of their *integrals*:

$$g^{ij} = \int_0^\infty \phi^{i \leftarrow j}(t)\, dt \qquad\qquad G = [g^{ij}].$$

- Using Hawkes process representation as a Poisson cluster process, we introduce the (unobserved!) counting process $N_t^{i \leftarrow j}$ that counts the number of events of i whose direct estimator is an event of j. Then,

$$\mathbb{E}[N_t^{i \leftarrow j}] = g^{ij} \mathbb{E}[N_t^j].$$

HAWKES CAUSALITY

- Instead of trying to estimate the kernels $\phi^{i \leftarrow j}$, we focus on the direct estimation of their *integrals*:

$$g^{ij} = \int_0^\infty \phi^{i \leftarrow j}(t)\, dt \qquad\qquad G = [g^{ij}].$$

- Using Hawkes process representation as a Poisson cluster process, we introduce the (unobserved!) counting process $N_t^{i \leftarrow j}$ that counts the number of events of i whose direct estimator is an event of j. Then,

$$\mathbb{E}[N_t^{i \leftarrow j}] = g^{ij} \mathbb{E}[N_t^j].$$

- The integral g^{ij} represents the average number of events of type i triggered by one event of type j.

INTEGRATED CUMULANTS

- A recent paper (S. Jovanovic, 2014) related the integrated cumulants of the process to $R = (I - G)^{-1}$. The three first orders:

$$\Lambda^i = \sum_{m=1}^{D} R^{im} \mu^m$$

$$C^{ij} = \sum_{m=1}^{D} \Lambda^m R^{im} R^{jm}$$

$$S^{ijk} = \sum_{m=1}^{D} \left(R^{im} R^{jm} C^{km} + R^{im} C^{jm} R^{km} + C^{im} R^{jm} R^{km} - 2\Lambda^m R^{im} R^{jm} R^{km} \right)$$

- We designed asymptotically unbiased estimators of the integrated cumulants above, see my thesis for details.

CUMULANTS MATCHING

- **Idea:** matching the theoretical cumulants with the empirical ones enables the estimation of G without any parametric modeling, and without estimating the kernels $\phi^{i \leftarrow j}$.

CUMULANTS MATCHING

- **Idea:** matching the theoretical cumulants with the empirical ones enables the estimation of G without any parametric modeling, and without estimating the kernels $\phi^{i \leftarrow j}$.

- To recover G, we minimize the following least-square criterion using AdaGrad (a SGD-like algorithm):

$$\mathcal{L}(R) = (1-\kappa)||S_c(R) - \widehat{S}_c||_2^2 + \kappa||C(R) - \widehat{C}||_2^2.$$

- Our estimator is defined as:

$$\widehat{G} = I - (\arg\min_R \mathcal{L}(R))^{-1}.$$

CUMULANTS MATCHING

- **Idea:** matching the theoretical cumulants with the empirical ones enables the estimation of G without any parametric modeling, and without estimating the kernels $\phi^{i \leftarrow j}$.

- To recover G, we minimize the following least-square criterion using AdaGrad (a SGD-like algorithm):

$$\mathcal{L}(R) = (1 - \kappa)||S_c(R) - \widehat{S}_c||_2^2 + \kappa||C(R) - \widehat{C}||_2^2.$$

- Our estimator is defined as:

$$\widehat{G} = I - (\arg\min_R \mathcal{L}(R))^{-1}.$$

- Our method (called *NPHC*) scales better than state-of-the-art methods, is robust towards the kernel shape and directly outputs the kernel integral.

CONSISTENCY OF THE ESTIMATOR

Theorem: Suppose N_t is observed on $\mathbb{R}^+$ and assume that

- $g_0(R) = 0$ if and only if $R = R_0$,
- $R \in \Theta$, which is a compact set,
- the spectral norm of the kernel norm matrix satisfies $||G_0|| < 1$,
- $H_T \to \infty \quad$ and $\quad H_T^2/T \to 0$.

CONSISTENCY OF THE ESTIMATOR

Theorem: Suppose N_t is observed on $\mathbb{R}^+$ and assume that

- $g_0(R) = 0$ if and only if $R = R_0$,
- $R \in \Theta$, which is a compact set,
- the spectral norm of the kernel norm matrix satisfies $||G_0|| < 1$,
- $H_T \to \infty \quad \text{and} \quad H_T^2/T \to 0.$

Then,

$$\widehat{G_T} = I - \left(\arg\min_{R\in\Theta} \mathcal{L}_T(R)\right)^{-1} \xrightarrow{\mathbb{P}} G_0.$$

CONSISTENCY OF THE ESTIMATOR

Sketch of the proof:

- We design unbiased estimators $\widehat{C}^{(T)}$ and $\widehat{S}^{(T)}$ of $\int_{[-H_T,H_T]} C(t)dt$ and $\int_{[-H_T,H_T]^2} S(t,t')dtdt'$.

CONSISTENCY OF THE ESTIMATOR

Sketch of the proof:

- We design unbiased estimators $\widehat{C}^{(T)}$ and $\widehat{S}^{(T)}$ of $\int_{[-H_T,H_T]} C(t)dt$ and $\int_{[-H_T,H_T]^2} S(t,t')dtdt'$.

- To make Generalized Method of Moments' proof work, it is sufficient to prove that

$$||\widehat{X}^{(T)} - X|| \xrightarrow{\mathbb{P}} 0 \qquad \text{for} \qquad X \in \{\Lambda, C, S\}.$$

CONSISTENCY OF THE ESTIMATOR

Sketch of the proof:

- We design unbiased estimators $\widehat{C}^{(T)}$ and $\widehat{S}^{(T)}$ of $\int_{[-H_T,H_T]} C(t)dt$ and $\int_{[-H_T,H_T]^2} S(t,t')dtdt'$.
- To make Generalized Method of Moments' proof work, it is sufficient to prove that

$$||\widehat{X}^{(T)} - X|| \xrightarrow{\mathbb{P}} 0 \qquad \text{for} \qquad X \in \{\Lambda, C, S\}.$$

- We prove the previous statement.

CAPTURE ORDER BOOK DYNAMICS WITH HAWKES CAUSALITY

FINANCIAL TRANSACTIONS

- For a single asset, hundreds of thousands financial transactions happen everyday, and datasets describing the process event by event are available (but not free).

FINANCIAL TRANSACTIONS

- For a single asset, hundreds of thousands financial transactions happen everyday, and datasets describing the process event by event are available (but not free).

- Bowsher recognized in 2007 the flexibility and the simplicity of using Hawkes processes to model the joint dynamics of trades and mid-price changes of the NYSE.

FINANCIAL TRANSACTIONS

- For a single asset, hundreds of thousands financial transactions happen everyday, and datasets describing the process event by event are available (but not free).

- Bowsher recognized in 2007 the flexibility and the simplicity of using Hawkes processes to model the joint dynamics of trades and mid-price changes of the NYSE.

- One of the core question in financial statistics is to understand order book dynamics. Previous representations with Hawkes processes were low-dimensional because of their estimation method's complexity.

FINANCIAL TRANSACTIONS

- For a single asset, hundreds of thousands financial transactions happen everyday, and datasets describing the process event by event are available (but not free).

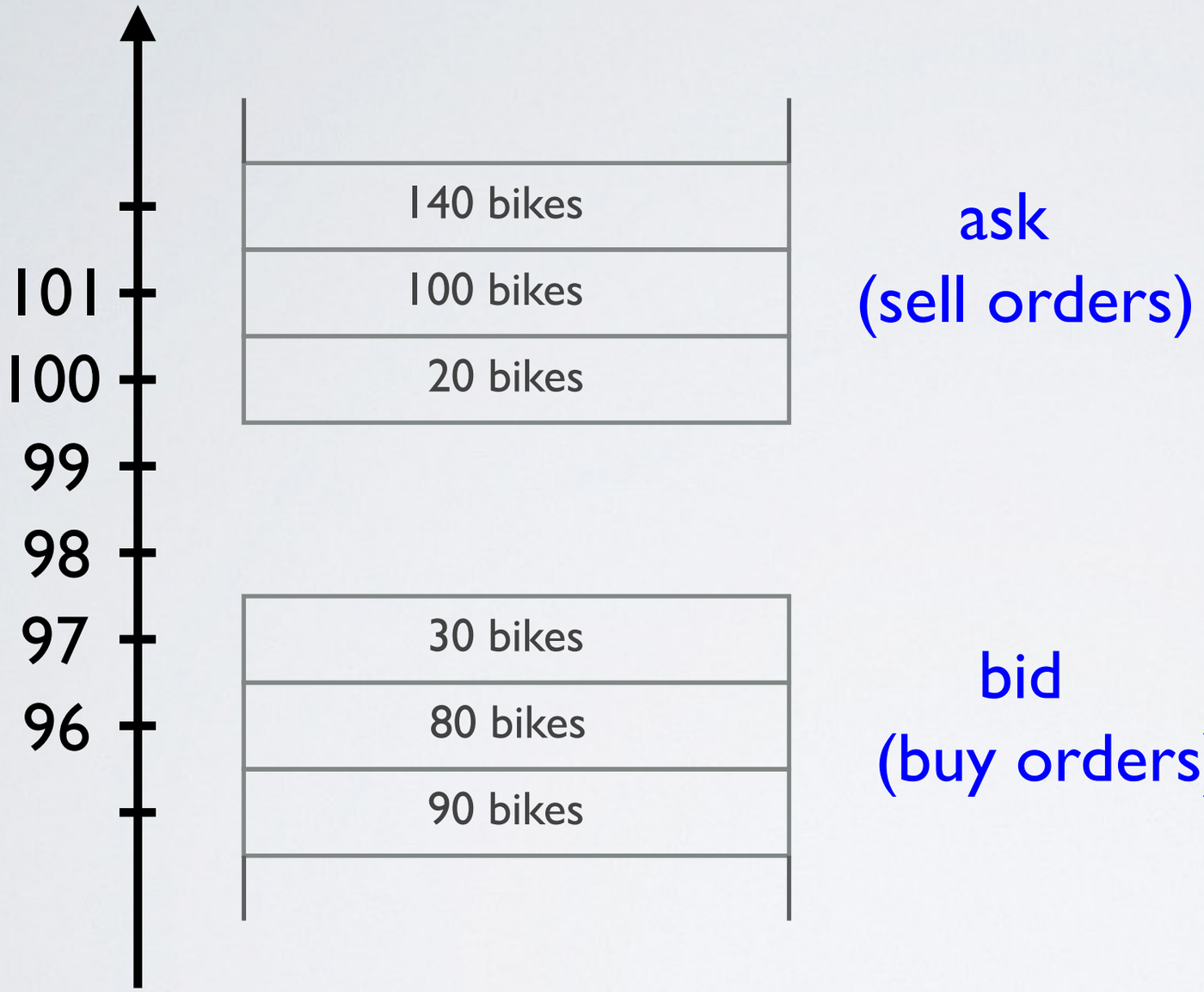
- Bowsher recognized in 2007 the flexibility and the simplicity of using Hawkes processes to model the joint dynamics of trades and mid-price changes of the NYSE.

- One of the core question in financial statistics is to understand order book dynamics. Previous representations with Hawkes processes were low-dimensional because of their estimation method's complexity.

- Can NPHC provide us a novel picture of order book dynamics?

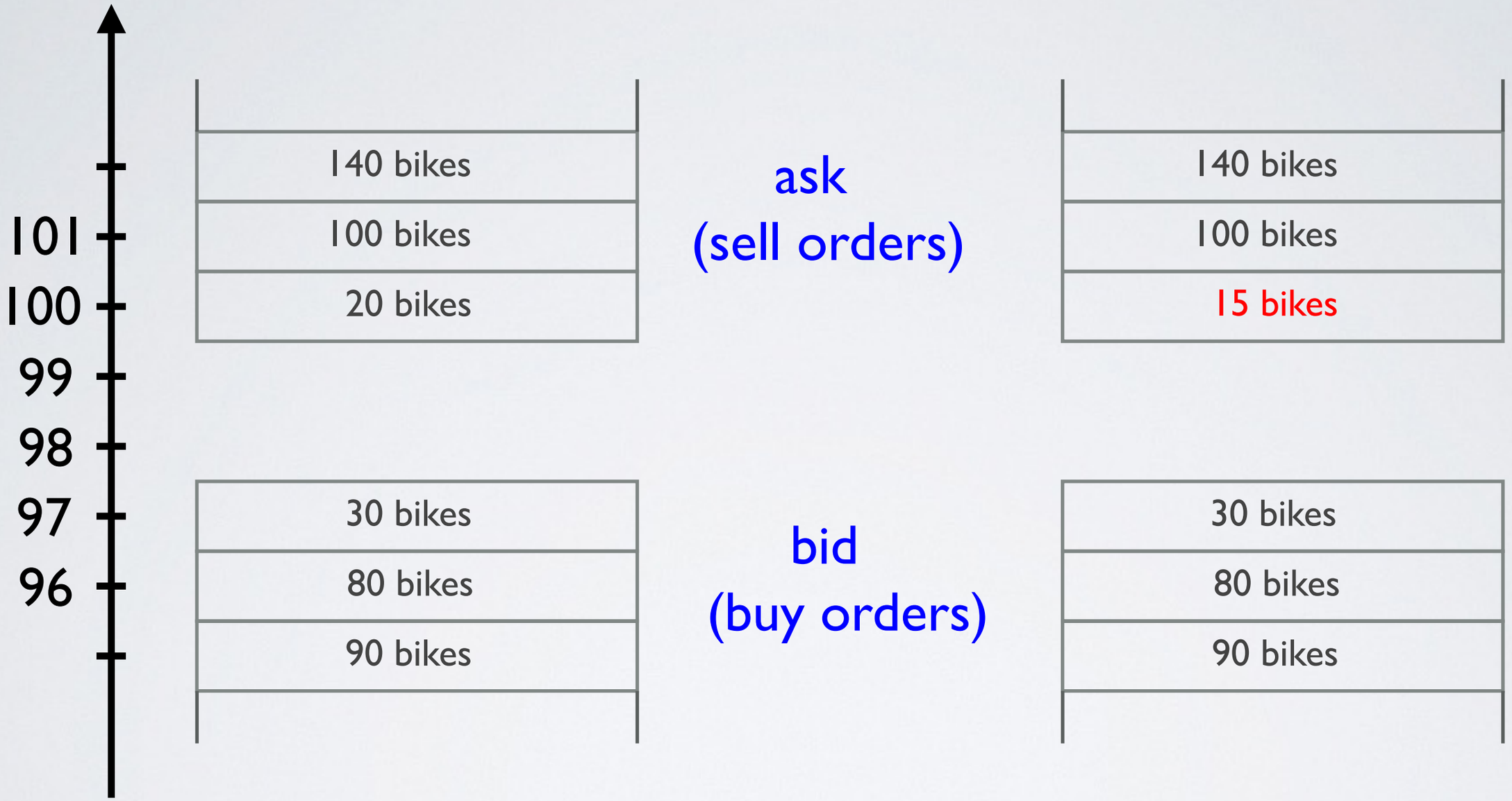
LIMIT ORDER BOOK 101

- Example: bikes on craiglist / leboncoin



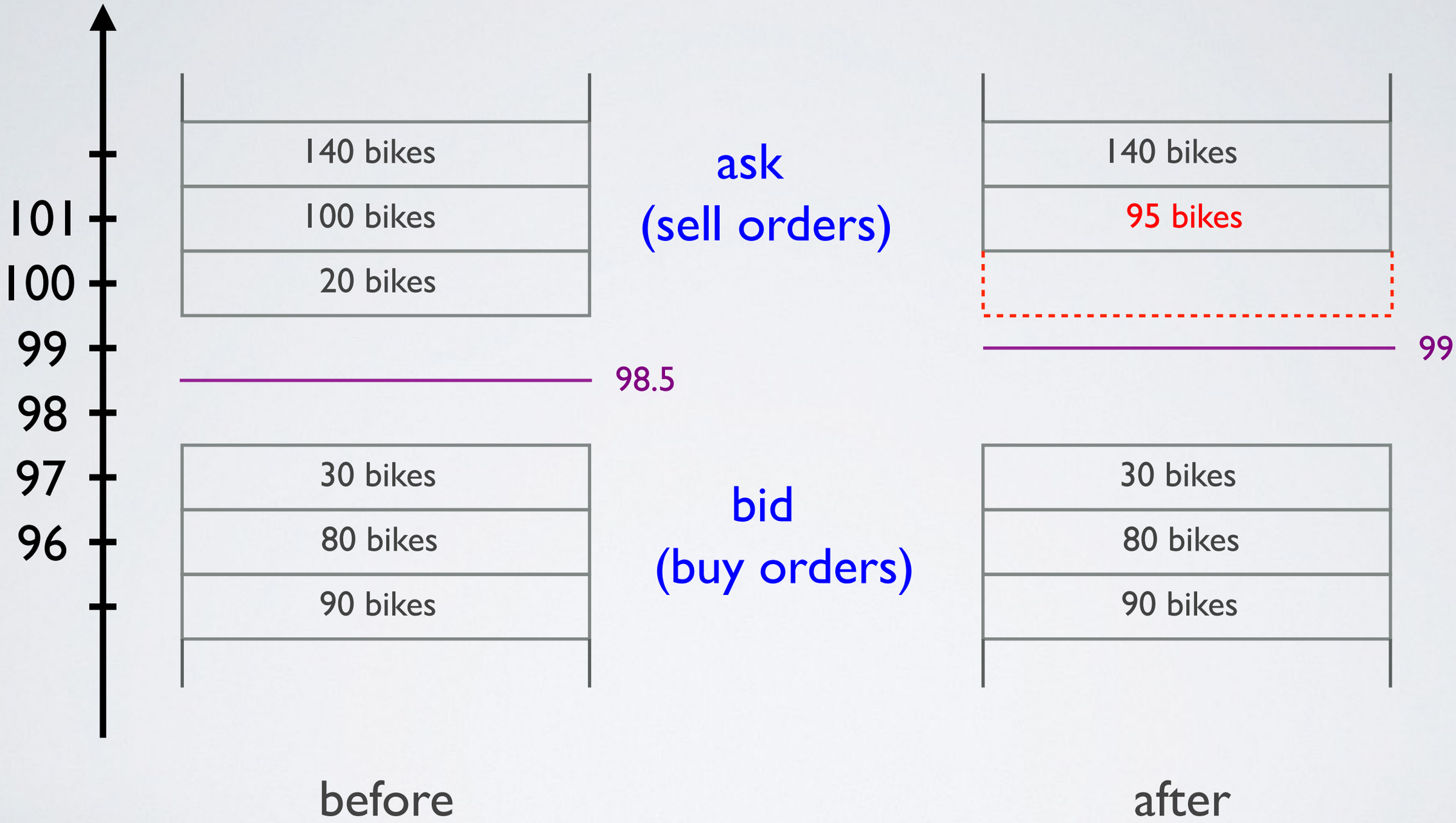
LIMIT ORDER BOOK 101

- Market order at the ask that does not move the mid-price



LIMIT ORDER BOOK 101

- Market order at the ask that moves the mid-price

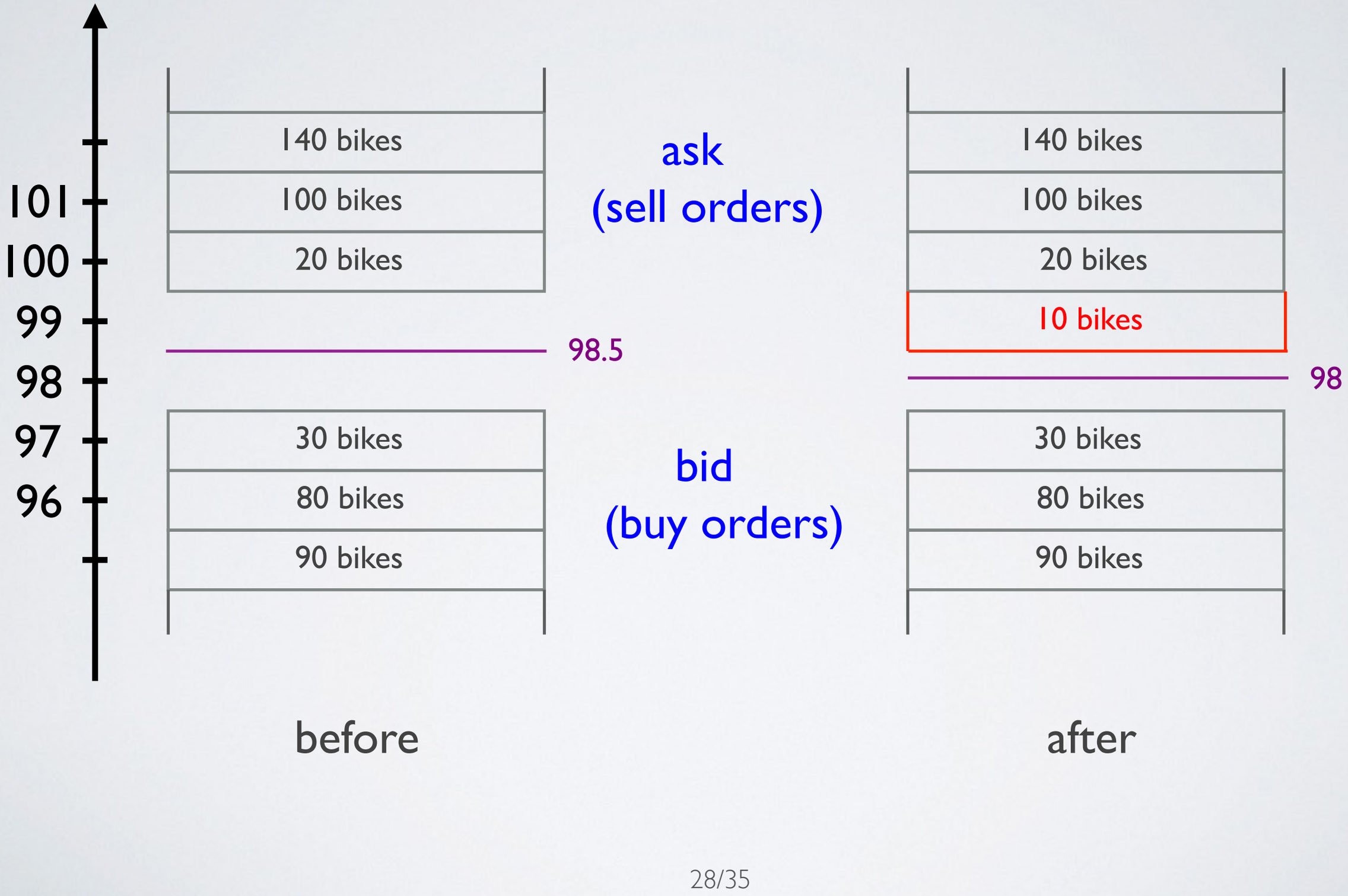


LIMIT ORDER BOOK 101

- Limit order at the ask that does not move the mid-price

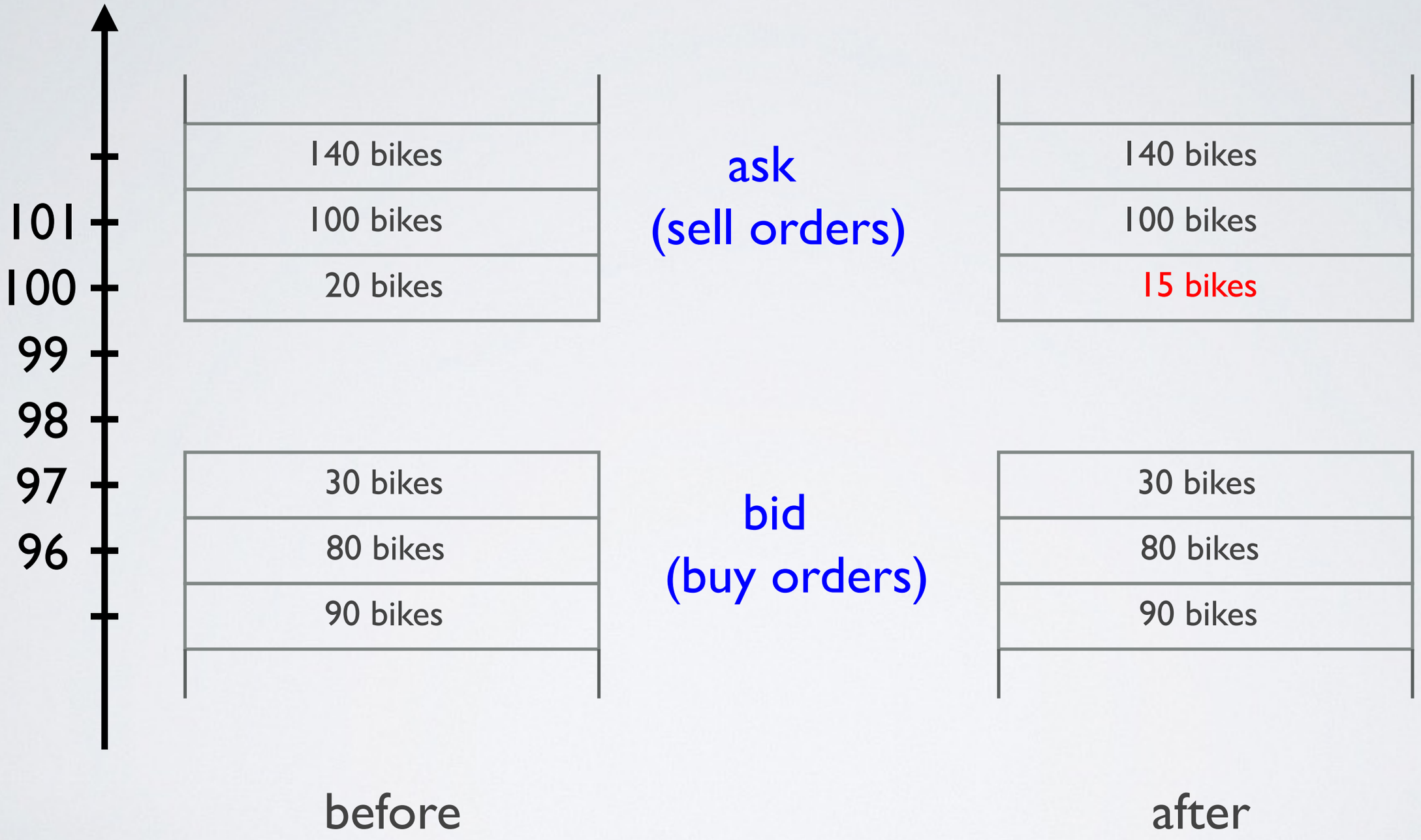
LIMIT ORDER BOOK 101

- Limit order at the ask that moves the mid-price



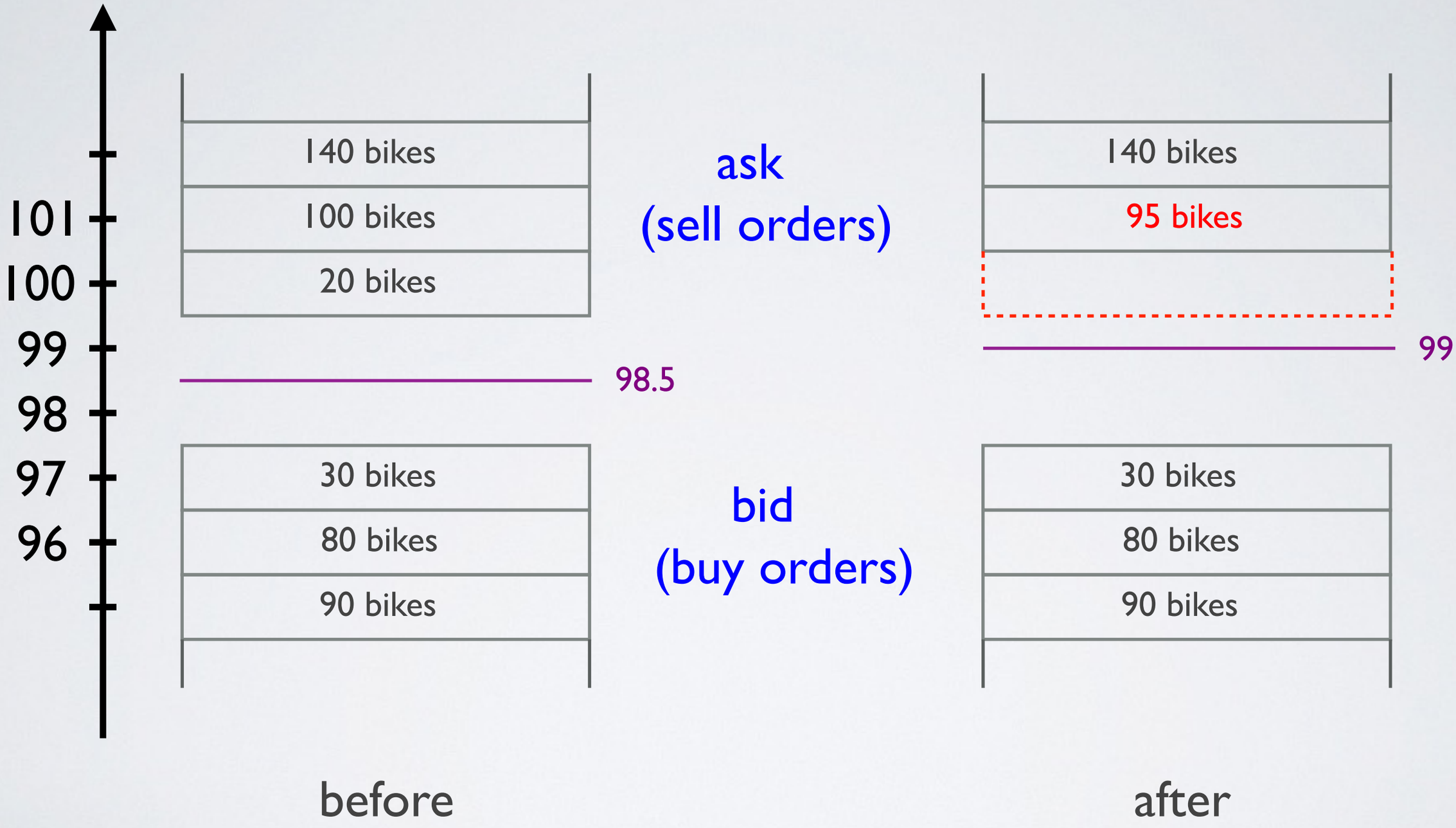
LIMIT ORDER BOOK 101

- Cancel order at the ask that does not move the mid-price



LIMIT ORDER BOOK 101

- Cancel order at the ask that moves the mid-price

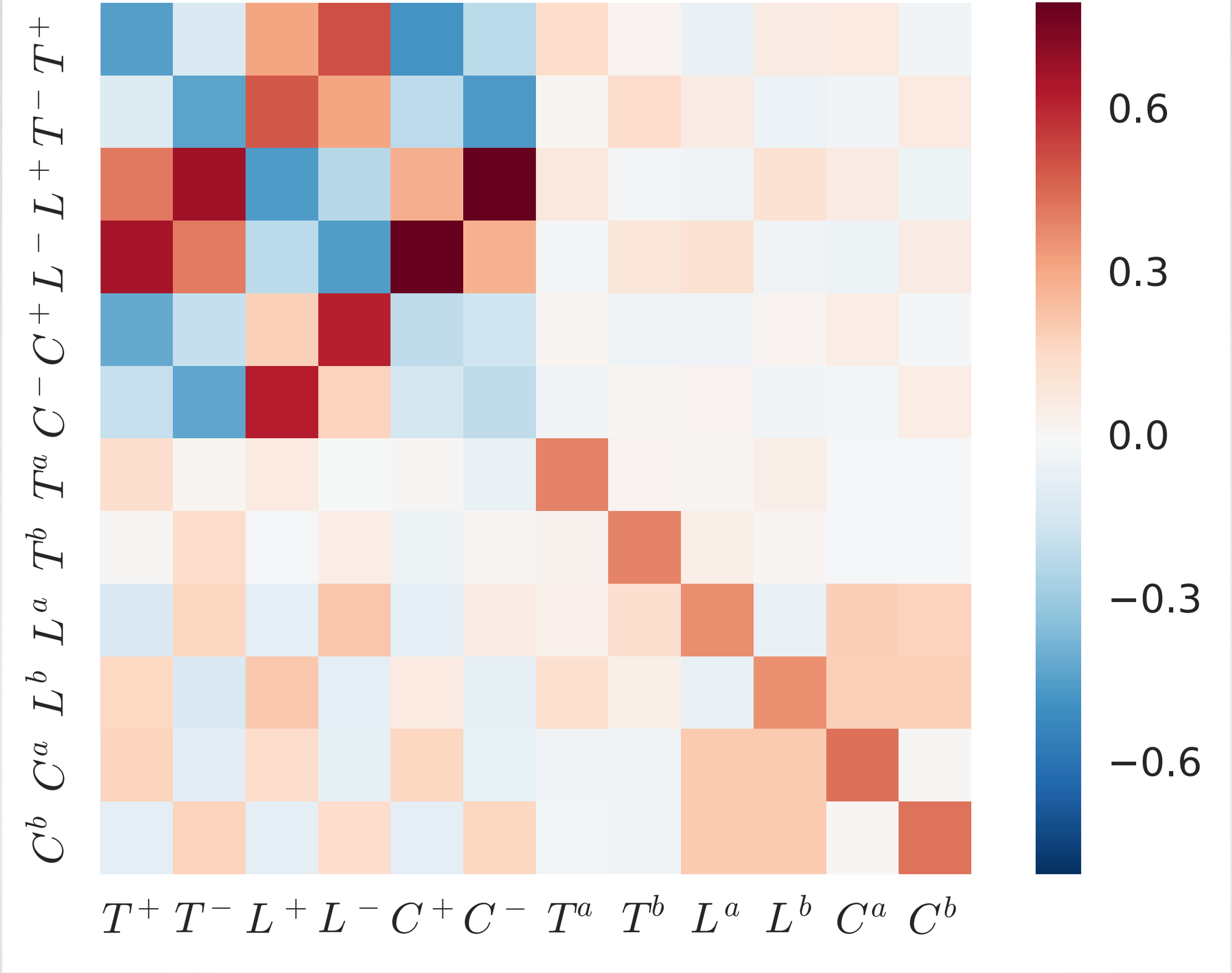


LIMIT ORDER BOOK MODEL

- We consider the following 12-dimensional point process to model the dynamic of an order book:

$$N_t = \left(T_t^+, T_t^-, L_t^+, L_t^-, C_t^+, C_t^-, T_t^a, T_t^b, L_t^a, L_t^b, C_t^a, C_t^b\right).$$

- Each dimension counts the number of events before t:
  - $T^+ (T^-)$: upwards (downwards) mid-price move triggered by a market order.
  - $L^+ (L^-)$: upwards (downwards) mid-price move triggered by a limit order.
  - $C^+ (C^-)$: upwards (downwards) mid-price move triggered by a cancel order.
  - $T^a (T^b)$: market order at the ask (bid) that does not move the mid-price.
  - $L^a (L^b)$: limit order at the ask (bid) that does not move the mid-price.
  - $C^a (C^b)$: cancel order at the ask (bid) that does not move the mid-price.


T^+ T^- L^+ L^- C^+ C^- T^a T^b L^a L^b C^a C^b
T^+ T^- L^+ L^- C^+ C^- T^a T^b L^a L^b C^a C^b
0.6
0.3
0.0
−0.3
−0.6

MULTI-ASSET MODEL

- We scale up the model to account on events of two assets simultaneously, and unveil a precise structure of the high-frequency cross-asset dynamics.

- We consider a 16-dimensional model, made of two 8-dimensional models of the form:

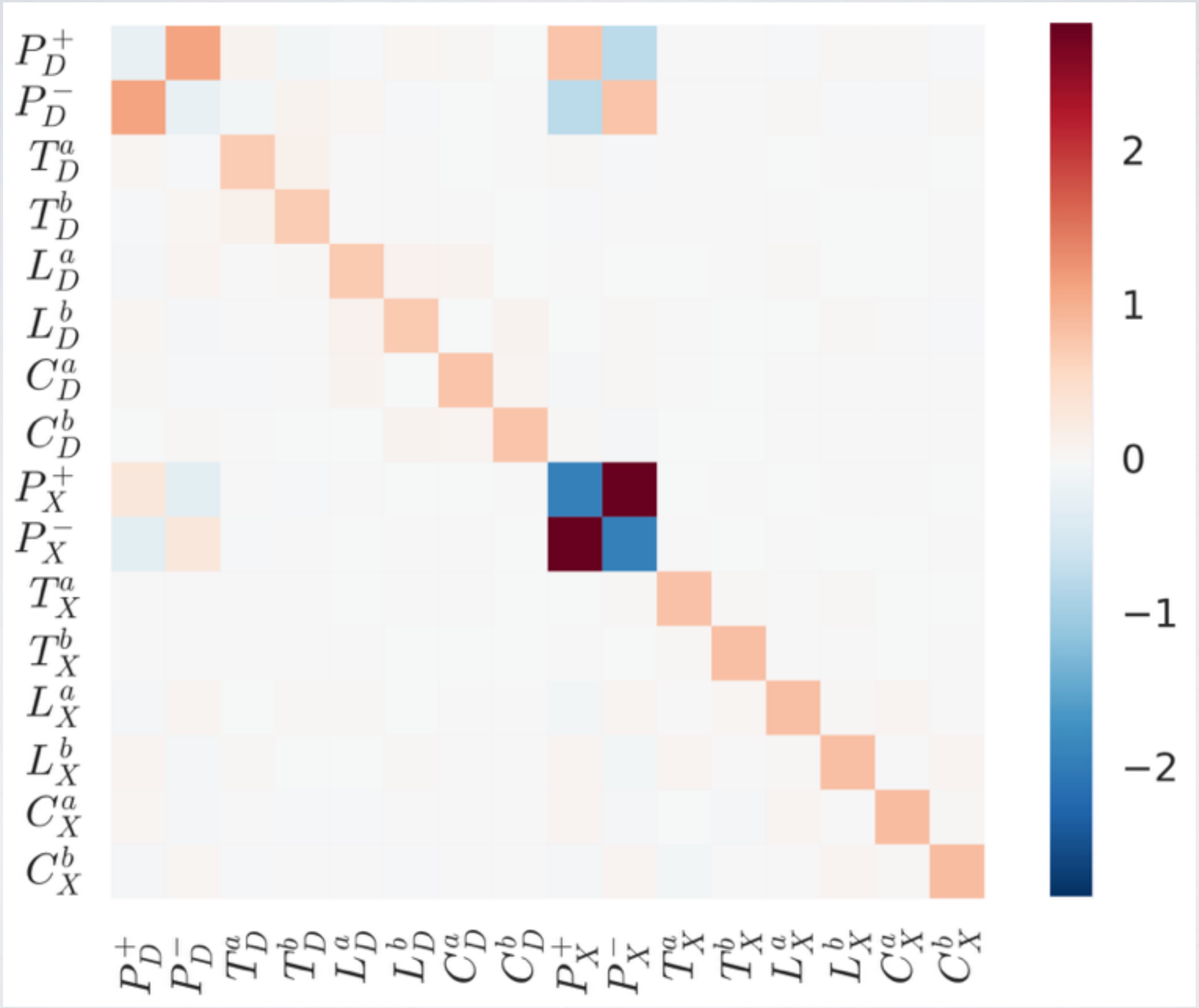
$$N_t = (P_t^+, P_t^-, T_t^a, T_t^b, L_t^a, L_t^b, C_t^a, C_t^b),$$

where $P^+ (P^-)$ count upwards (downwards) mid-price move triggered by any order.

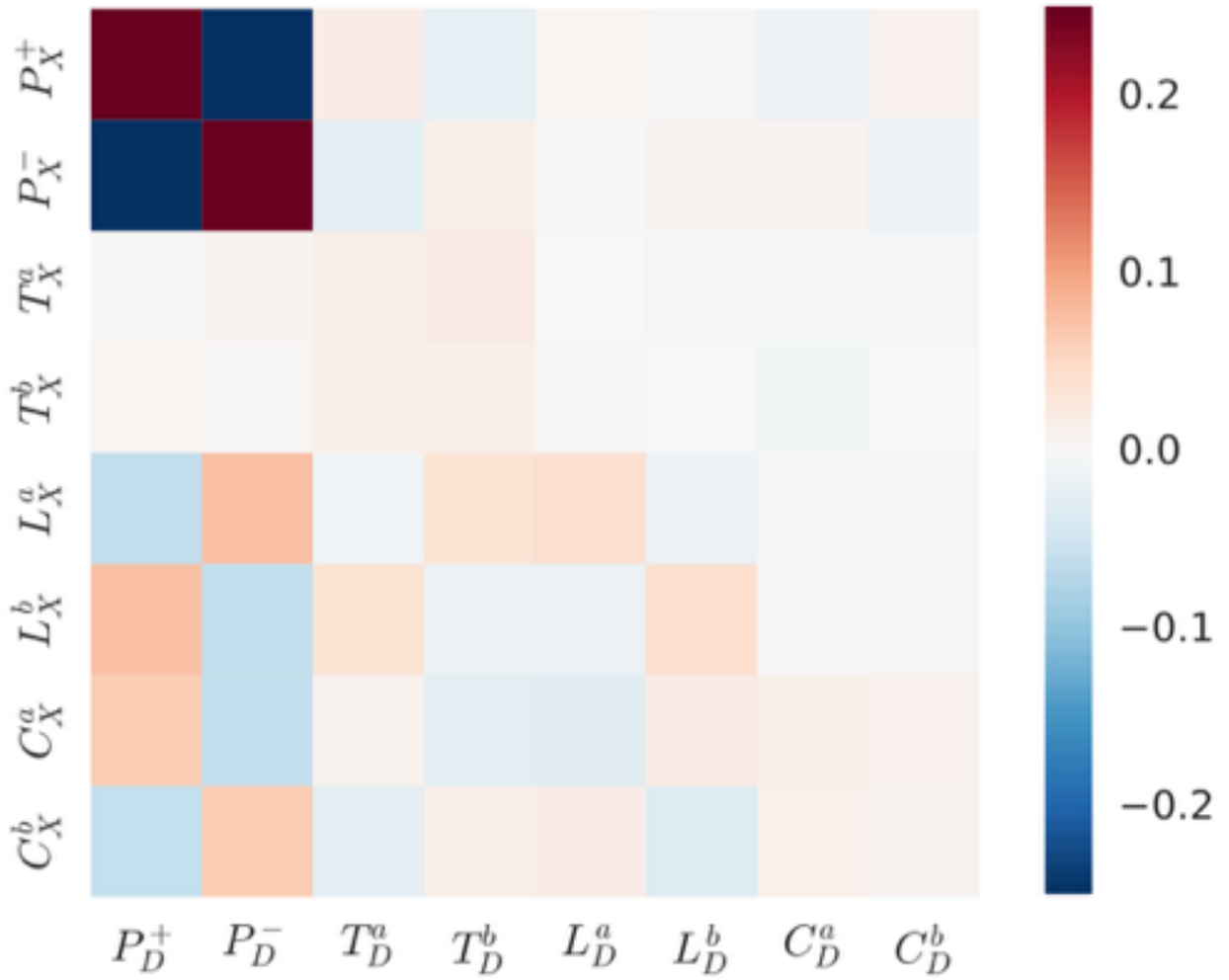
- We consider two very correlated assets: the futures on the DAX (German stock index) and the Euro Stoxx 50 (Eurozone stock index).

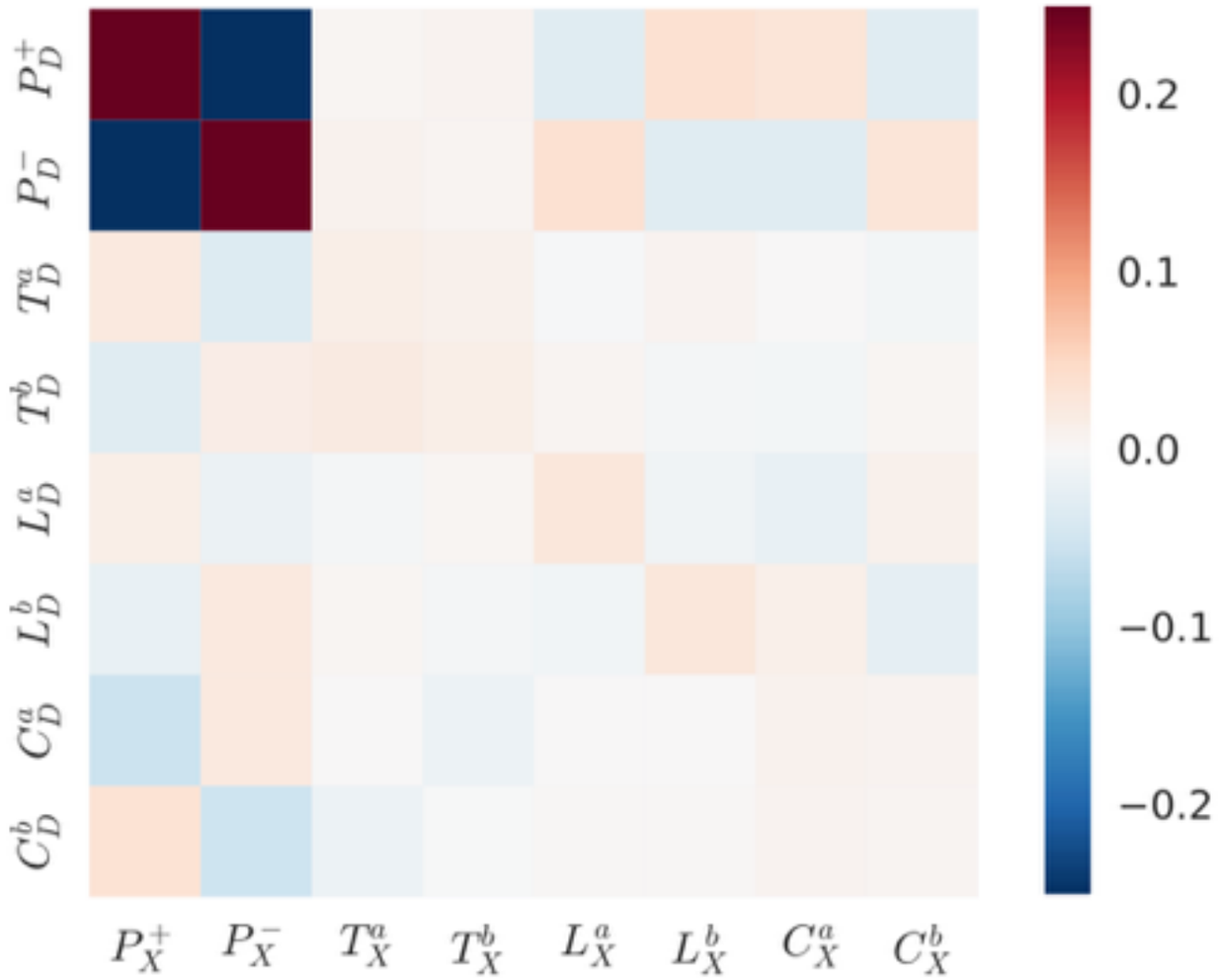
- DAX has a small tick size (= minimal price movement), and Euro Stoxx a large tick size.

MULTI-ASSET MODEL



MULTI-ASSET MODEL





QUESTIONS ?