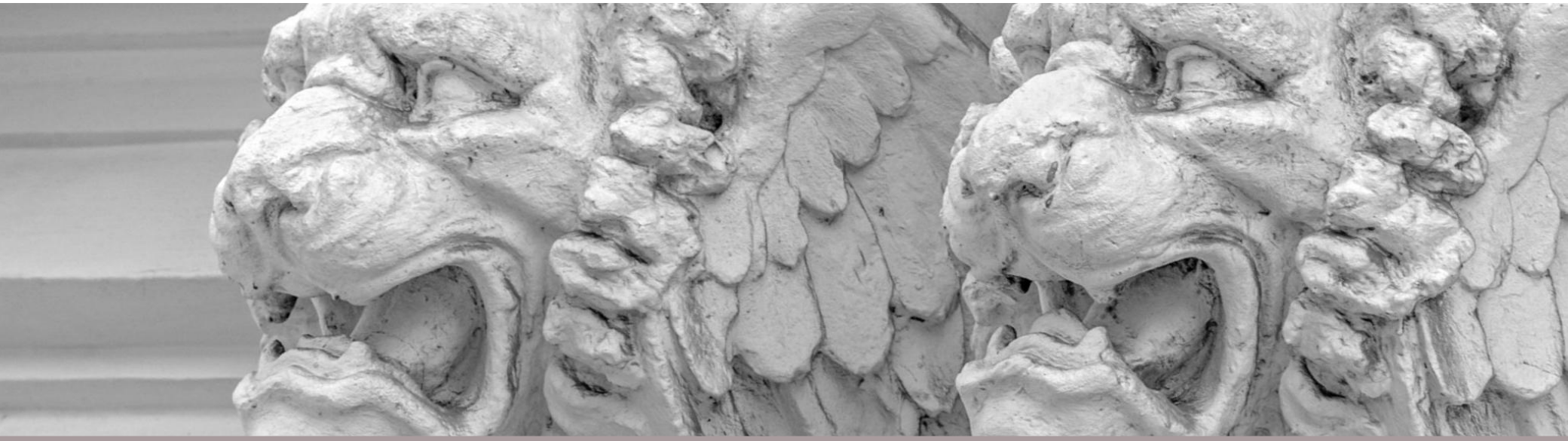
Bank of Russia
The Central Bank of the Russian Federation



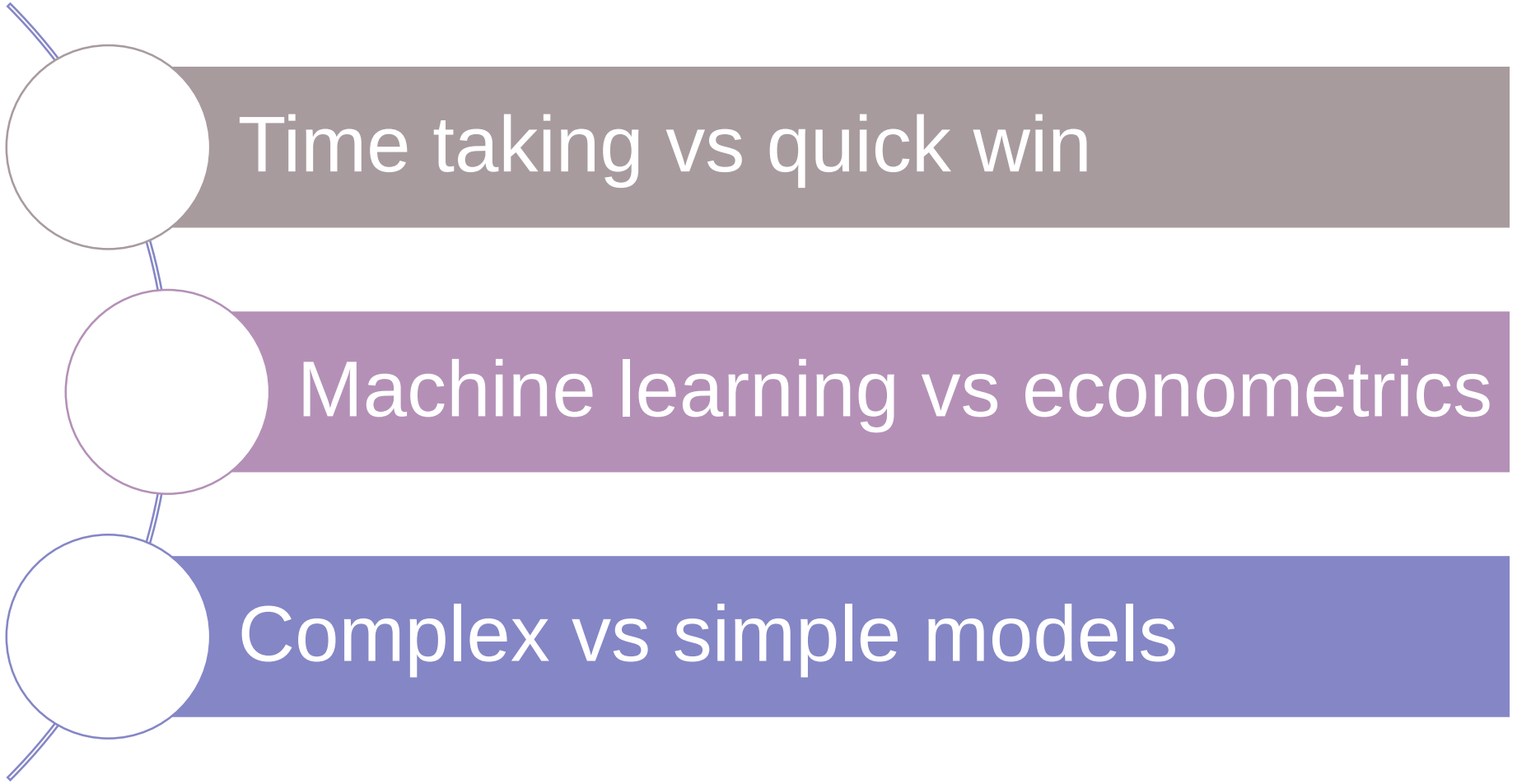
Assessing mortgage risks using big data

Evgeny Rumyantsev

Financial Stability Department

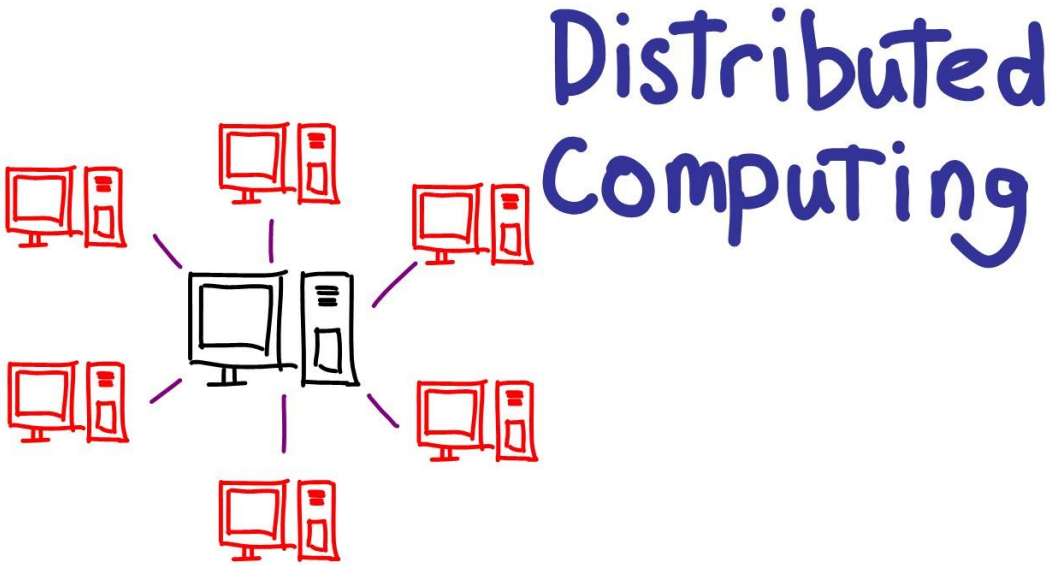
September 2018

Big data and machine learning challenges



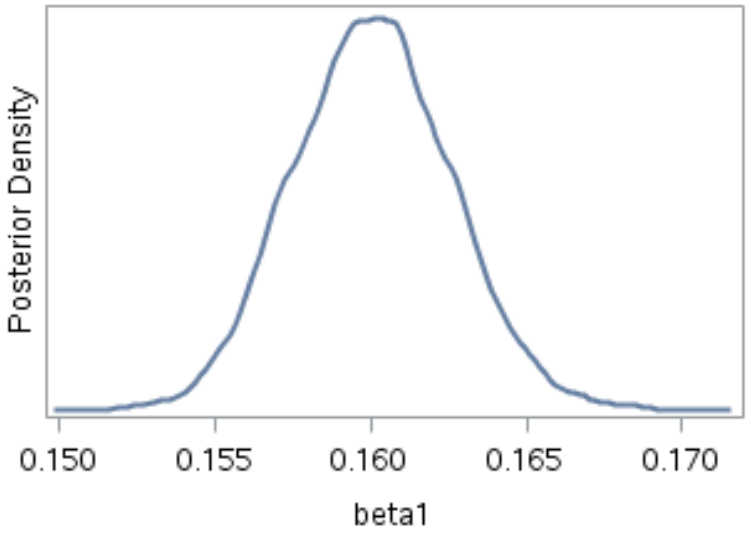
Time taking vs quick win

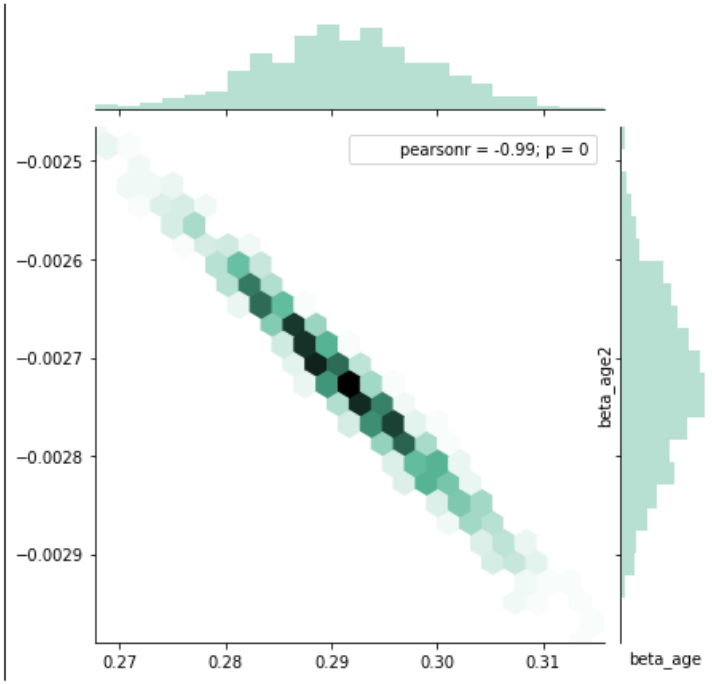
Parallel computing or distributed processing of large data sets across clusters or computers



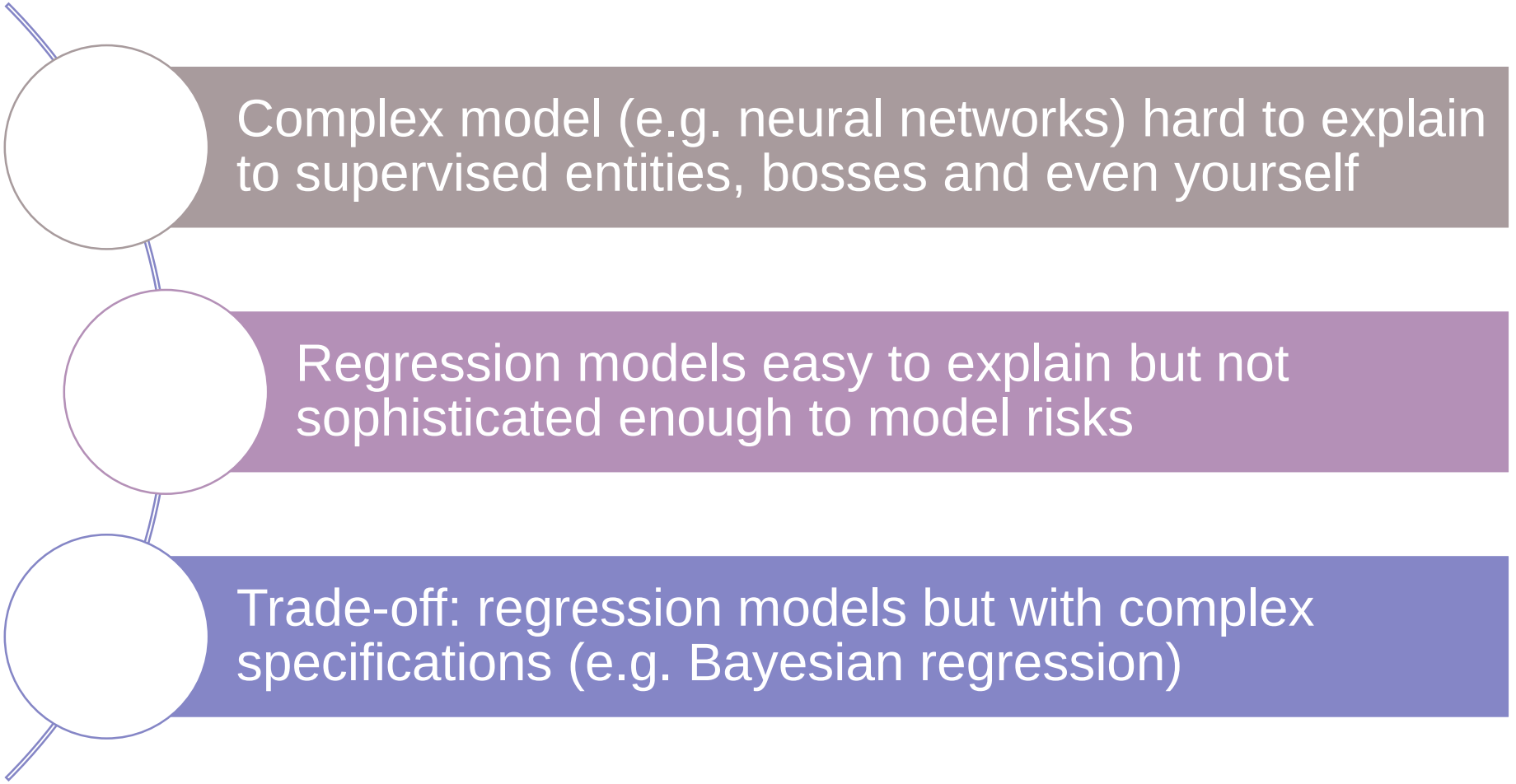
Machine learning vs econometrics

- ML includes larger variety of models (clustering, regression models (also hierarchical), neural networks etc.)
- No normality assumptions (use appropriate distribution for modeling even constructed by yourself and get non-normal posterior distr.)
- Deal with correlated covariates using Bayesian MCMC approach and take away posterior distr.
- Work with distr., not point estimates





Complex vs simple models



Bank of Russia

The Central Bank of the Russian Federation

Mortgage credit risk assessment in BoR

Data for mortgage credit risk analysis

Participants: six banks (covering more than 80% of the whole mortgage market).

Aggregated monthly data since 2007 till 2018: portfolio, issuance, transitions (over 70 mln. records).

Covariates: issuance date, state (region), collateral, currency, Current Loan-to-Value ratio, Payment-to-Income ratio, #household members, days past due.

Data template for loans migration

Report date	Vintage date	Currency	Collateral	State (region)	Prev. segment Index	Curr. segment Index	Loans outstanding	# Loans
01.01.2007	01.01.2007	Local	Real estate	1	LTV', PTI', CTI', DPD'	LTV, PTI, CTI, DPD		
…	…	Foreign	Claims Under Share Construction Participation Agreements	…				
01.01.2018	01.12.2017			11				

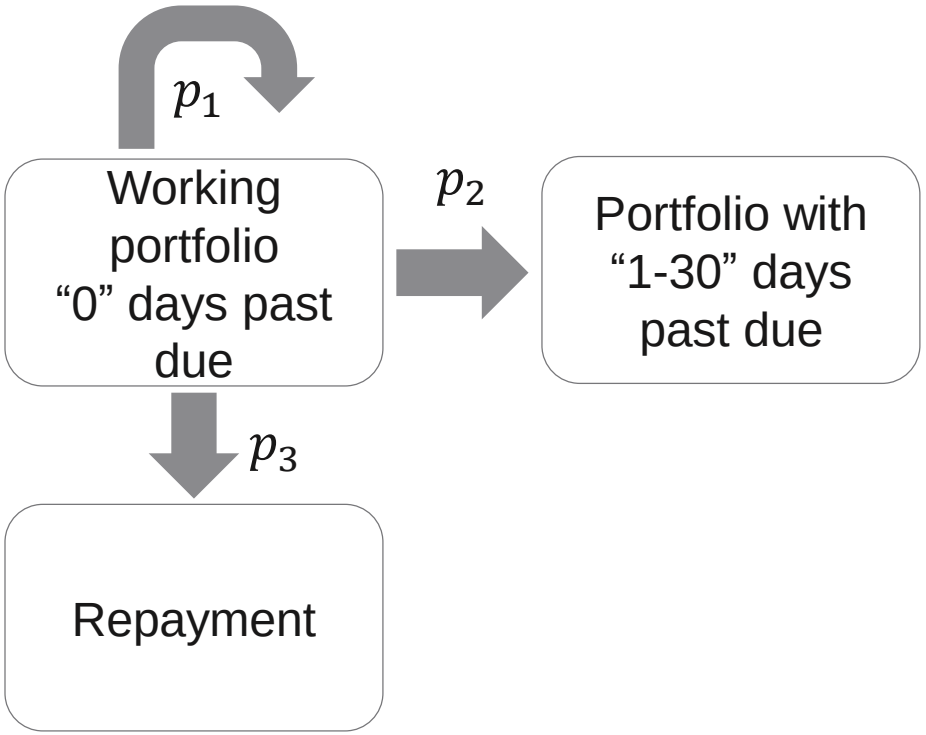
Notes:

$$\text{Consumption}-\text{to}-\text{Income (CTI)} = \frac{\#\ \text{household members} * \text{minimal allowence}}{\text{Income}}$$

LTV, PTI, CTI takes discrete values from 0 to 100% with step 10 p.p.

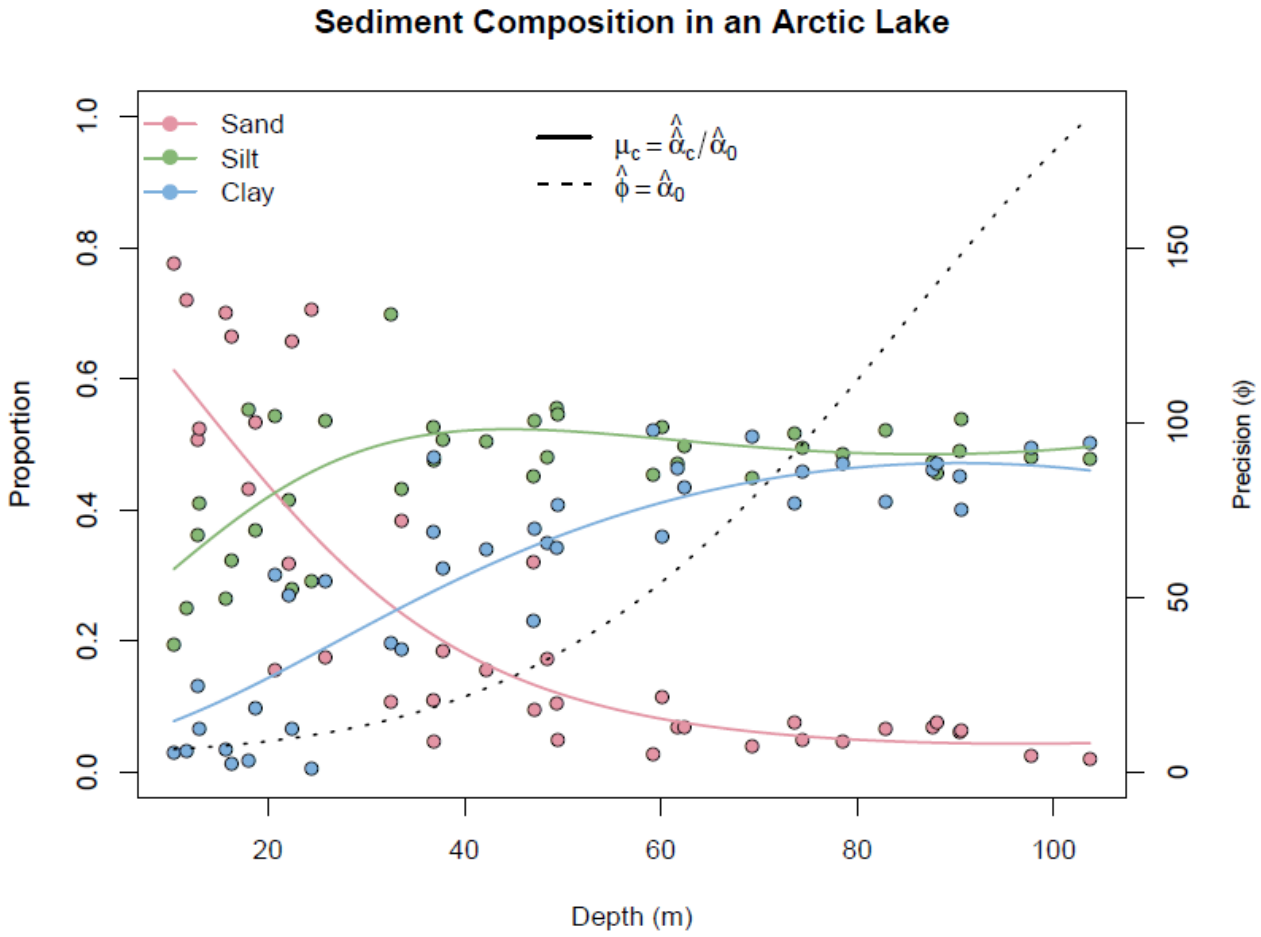
Days past due (DPD): "0", "1-30", "31-60", "61-90", "91-120", "121-150", "151-180",
"180-360", "360+" and "write off", "sell", "repayment".

Credit portfolio modeling



$$\sum_{i=1}^{3} p_i = 1$$

Compositional data



We use Dirichlet regression to deal with composite data (see M.J. Maier 2014)

$$f(\boldsymbol{p}|\boldsymbol{\mu},\varphi) = \frac{1}{B(\boldsymbol{\mu}\varphi)}\prod_{i=1}^{C} p_i^{(\mu_i\varphi-1)}, \qquad \mu_i \in (0;1), \forall i, \qquad \sum_{i=1}^{C}\mu_i = 1,$$

$$\sum_{i=1}^{C} p_i = 1\,, p_i \in (0;1), \forall i \qquad \varphi > 0, \qquad E[p_i] = \mu_i$$

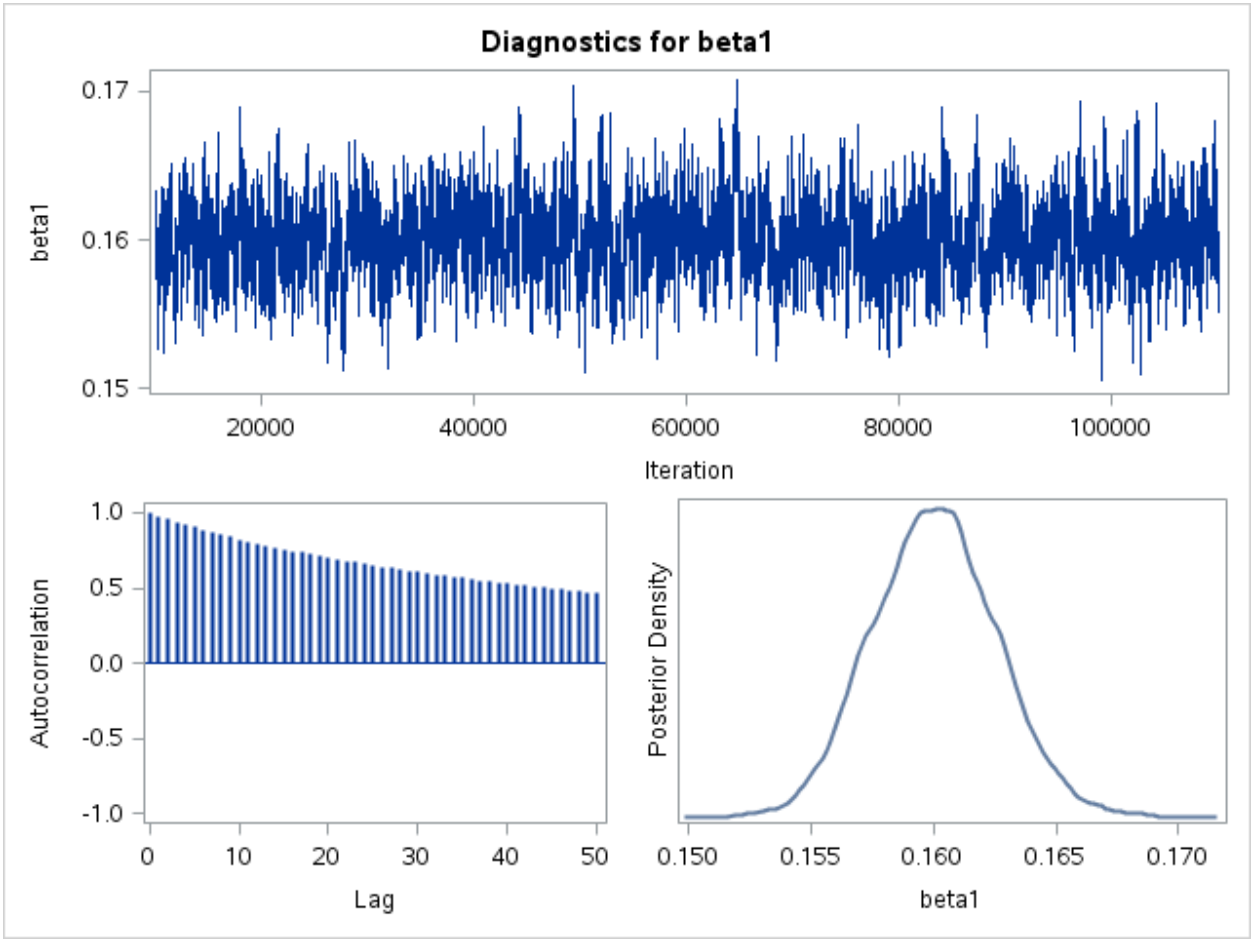
Multinomial logistic function for μ_i and exponential for φ :

$$E[p_i] = \mu_i = \frac{\exp(\boldsymbol{X}\boldsymbol{\beta}_i)}{\sum_{j=1}^{C}\boldsymbol{X}\boldsymbol{\beta}_j}, \ \boldsymbol{\beta}_C = \mathbf{0} \qquad \textit{Roll-rate}$$

$$\varphi = \exp(\boldsymbol{X}\boldsymbol{\gamma}_i) \qquad \textit{Precision parameter}$$

$$\boldsymbol{X} = (Current\ LTV, Discounted\ PTI,$$
$$Discounted\ Consumption\ to\ Income, Unemployement)$$

Bayesian regression model with MCMC procedure for parameters estimation (SAS proc MCMC)



Transition matrix approach

Portfolio is 4-d matrix with dimensions: LTV, PTI, Consumption-to-Income, Days past due.

$$Pfl_t(LTV, PTI, CTI, PDD)$$

Transition matrix is 8-d array with dimensions: LTV*, PTI*, CTI*, Days past due*, LTV, PTI, Consumption-to-Income, Days past due.

$$Prob(LTV', PTI', CTI', DPD' \rightarrow LTV, PTI, CTI, DPD)$$

Transition rule:

$$Pfl_{t+1}(LTV, PTI, CTI, DPD) =$$

$$\sum_{LTV,PTI,CTI,DPD} Prob(LTV', PTI', CTI', DPD' \rightarrow LTV, PTI, CTI, DPD) Pfl_t(LTV', PTI', CTI', DPD')$$

Discounted PTI and CTI on nominal income growth and inflation

$$PTI_t = \frac{PTI_{t-1}}{1 + i_t}$$

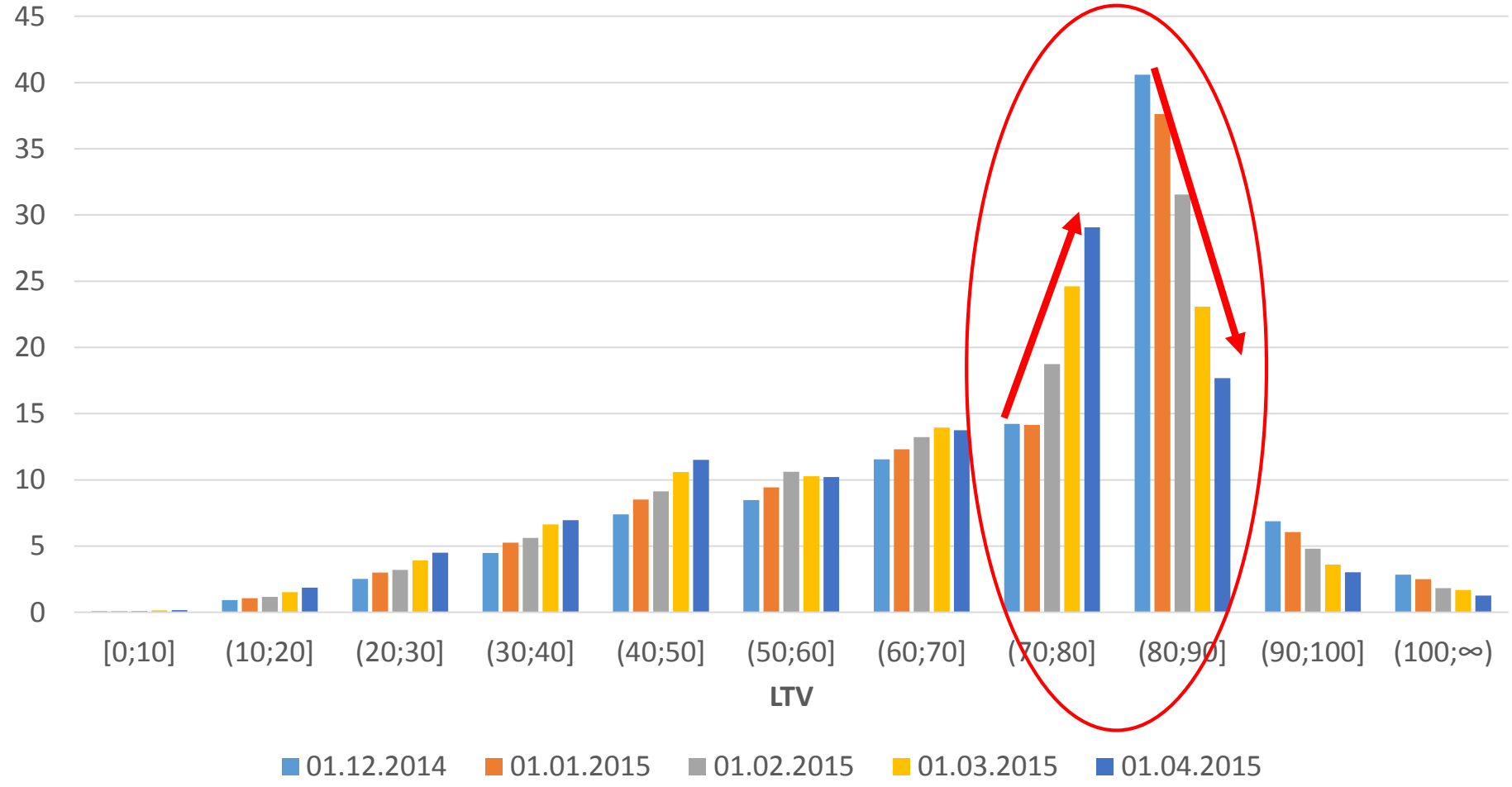
$$CTI_t = \frac{CTI_{t-1}(1 + \pi_t)}{1 + i_t}$$

i_t - nominal income growth rate by regions
π_t - inflation rate by regions

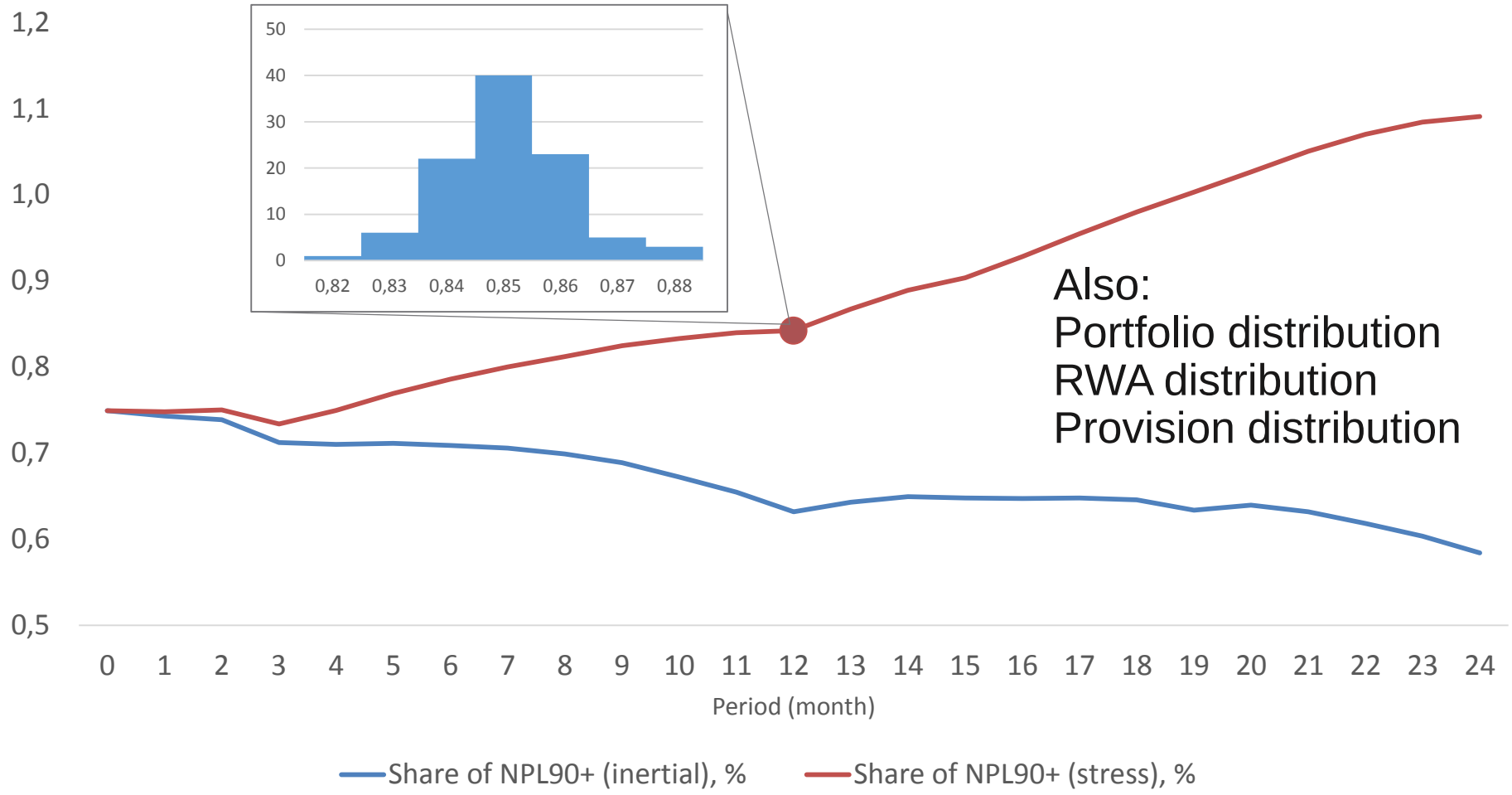
It gives Discounting matrix (4-d array):

$$Disc(PTI', CTI' \rightarrow LTV, PTI)$$

Banks constrain origination of new loans with low down payment



Example of stress-test results



Conclusions

- Big data may help to identify and estimate more sophisticated risks (behavioral aspects, concentration of risks in specific segments)
- The results of estimation may be used in macroprudential policy and under supervision
- Data scientist need to find a trade-off between simple and sophisticated models