



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Forecasting Time Series of Inhomogeneous Poisson Processes

with Applications to Call Center Workforce Management

Haipeng Shen

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

June 3, 2009

Outline

1 Motivation

2 Methods

3 Application

First Direct call center

First Direct (Larrece et al., 1997)



A more modern call center



A sweatshop call center???



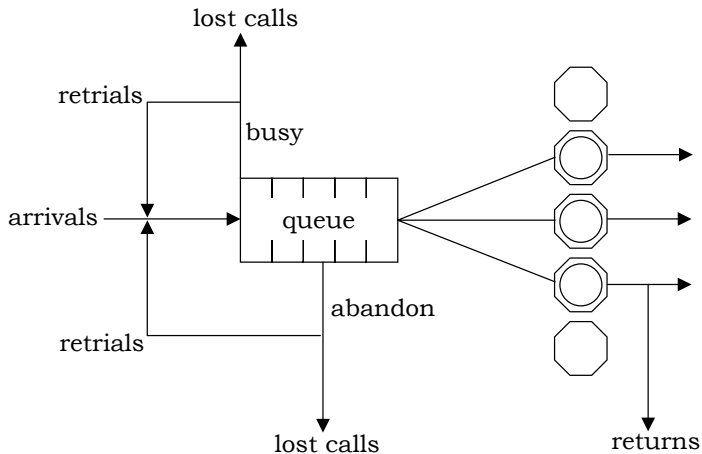
A vast call center world

- 4 million call center agents in US, 800 thousands in UK, 500 thousands in Canada and 500 thousands in India
- Call center costs exceeded \$300 billion worldwide
- 70% of the cost for human resource

A wider perspective

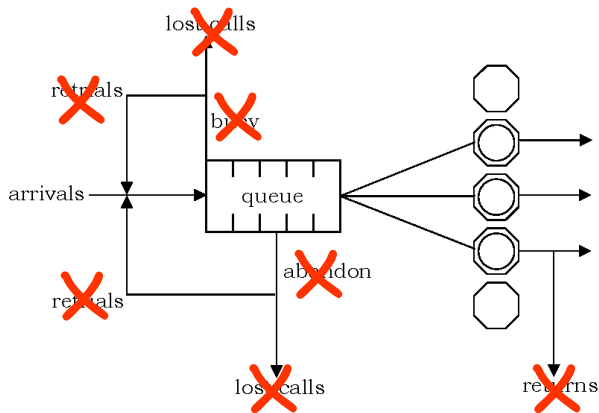
- Multi-disciplinary research area
- Service Enterprise Engineering (NSF)
- Service Science, Management, and Engineering (IBM)
- Service Science - “computer science of the 21st century”

Queueing model for a single call center



Gans, Koole and Mandelbaum (2003)

The $M/M/N + \infty$ model or the Erlang-C model



- no blocking, abandonment, or retrials
- fixed arrival rate λ_j and service rate μ_j for time period j
- exponential inter-arrival and service times

“Standard” model for call center workforce management

- 1 Forecast **offered load** (e.g., by the 1/2-hour)

$$\{R_j = \lambda_j / \mu_j : j = 1, \dots, m\}$$

where λ_j : arrival rate, μ_j : service rate.

- 2 Find minimum numbers of agents to make QoS constraint

$$s_j = \min\{s \mid P\{\text{Delay} \leq T\} \geq 1 - \epsilon\}$$

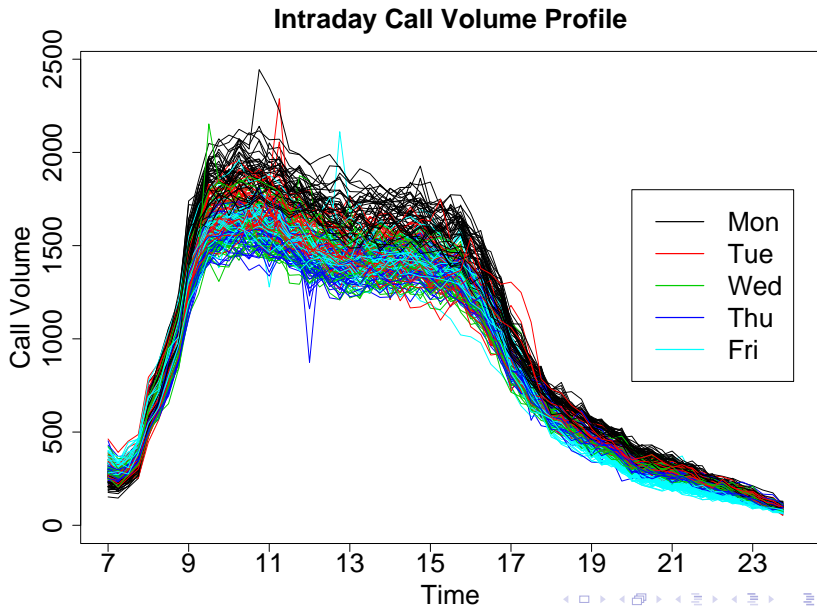
where in the common 80-20 rule, $T = 20$ seconds and $\epsilon = 20\%$.

- 3 Find minimum cost assignment of agents to schedules

$$\min\{cN \mid AN \geq s; N \geq 0; N \text{ integer}\}$$

where A : 0-1 schedule matrix, N : # of agents for each schedule, c : schedule cost.

However, arrival rates are not known with certainty



The research problem

- For a given day, call arrivals follow an inhomogeneous Poisson process
 - ▶ Day-to-day time series dependence
 - ▶ Within-day dependence
 - ▶ Seasonal effects
- Two forecasting scenarios:
 - ▶ **Day-to-day forecasting** of future daily arrival rate profile
 - ▶ **Within-day updating** of existing forecast
- Perform stochastic scheduling (with recourse) using the distributional forecasts
- Apply the approach to large-scale real systems

The data

- Bank with a network of 4 call centers in northeast US
- 300K calls/day, 60K/day seeking agents, 1K agents in peak hours
- 210 weekdays between January 6th and October 24th, 2003
- Call volumes during every quarter hour between 7am and midnight
- For each day, 68-dimensional vector of Poisson variables
- Across days, time series of vectors

Statistical model

- $\mathbf{y}_{(i)} = (y_{i1}, \dots, y_{im})^T$: arrival *count profile* of the *i*th day
- $\boldsymbol{\lambda}_{(i)} = (\lambda_{i1}, \dots, \lambda_{im})^T$: arrival *rate profile* of the *i*th day
- $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(n)}, \dots$: *m*-dimensional vector time series
- Time series of vectors of Poisson variables
- High dimension

Dimension reduction

- K -factor model

$$\begin{cases} \mathbf{y}_{(i)} \sim \text{Poisson}(\boldsymbol{\lambda}_{(i)}), & i = 1, \dots, n, \\ g(\boldsymbol{\lambda}_{(i)}) = \beta_{i1}\mathbf{f}_1 + \dots + \beta_{iK}\mathbf{f}_K \equiv \mathbf{F}\boldsymbol{\beta}_{(i)}, \end{cases}$$

- ▶ g : link function
- ▶ generalized linear models

- Time series model for $\boldsymbol{\beta}_{(i)}$

- Intraday feature vectors $f_1, \dots, f_K \in R^m$ ($K \ll m$)

- ▶ summarize intraday call arrival patterns
- ▶ reveal dominant intraday arrival features

Model estimation

- K -factor model

$$\begin{cases} \mathbf{y}_{(i)} \sim \text{Poisson}(\boldsymbol{\lambda}_{(i)}), & i = 1, \dots, n, \\ \mathbf{g}(\boldsymbol{\lambda}_{(i)}) = \beta_{i1}\mathbf{f}_1 + \dots + \beta_{iK}\mathbf{f}_K \equiv \mathbf{F}\boldsymbol{\beta}_{(i)}, \end{cases}$$

Time series model for $\boldsymbol{\beta}_{(i)}$

- Two-step procedure
 - ▶ extracting $\boldsymbol{\beta}_{(i)}$ and \mathbf{f}_j using alternating maximum likelihood (AML)
 - ▶ building time series model for $\boldsymbol{\beta}_{(i)}$
- Univariate time series models for components of $\boldsymbol{\beta}_{(i)}$

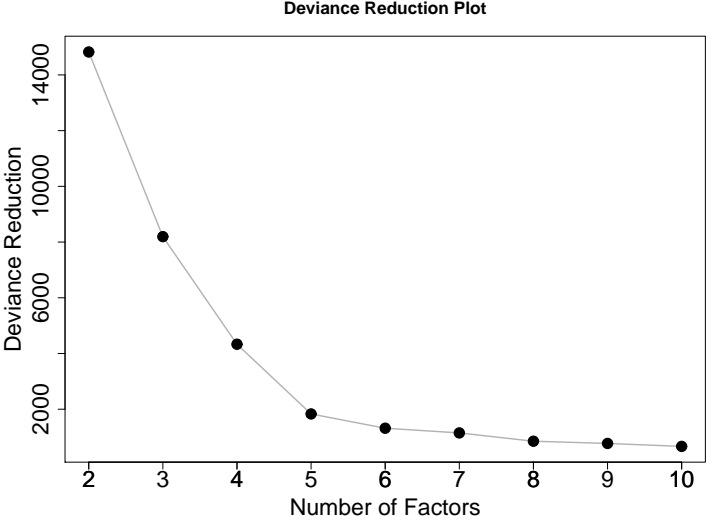
Forecasting future rate profile

- Using historical count data $\mathbf{y}^{(i)}$, $i = 1, \dots, n$, to forecast future rate profile $\lambda_{(n+h)}$ ($h > 0$)
- from the factor model

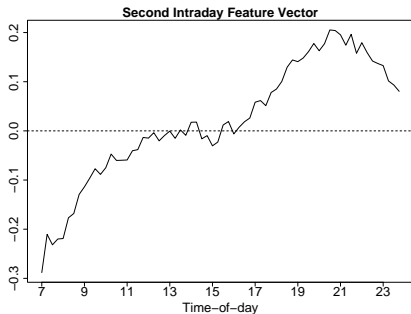
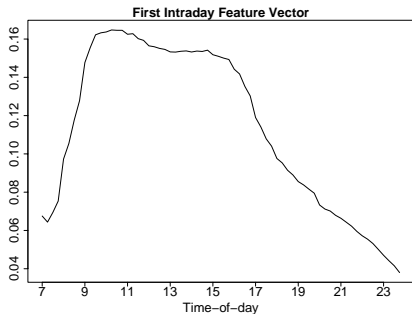
$$\lambda_{(n+h)} = g^{-1}(\beta_{n+h,1}\mathbf{f}_1 + \dots + \beta_{n+h,K}\mathbf{f}_K).$$

- ▶ $\mathbf{f}_1, \dots, \mathbf{f}_K$ by AML
- ▶ $(\beta_{i,1}, \dots, \beta_{i,K})$ by AML
- ▶ time series forecasts of $\beta_{(n+h)} = \{\beta_{n+h,1}, \dots, \beta_{n+h,K}\}^T$
- Interval and distributional forecasts
 - ▶ bootstrap the time series models

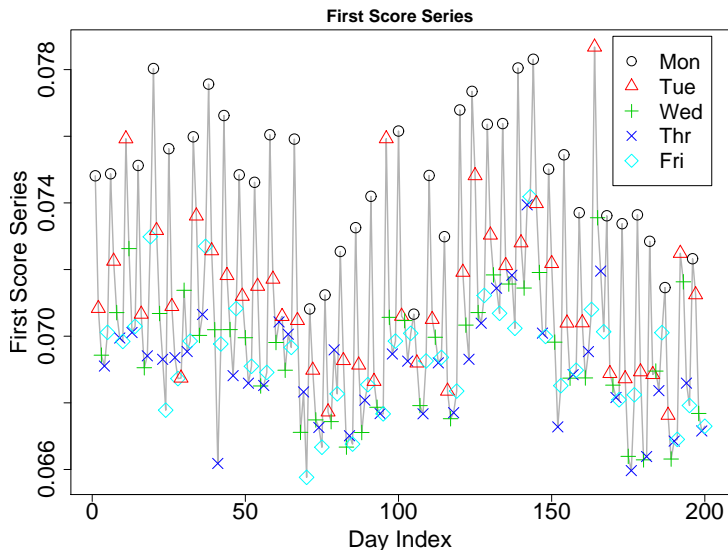
Deviance reduction plot



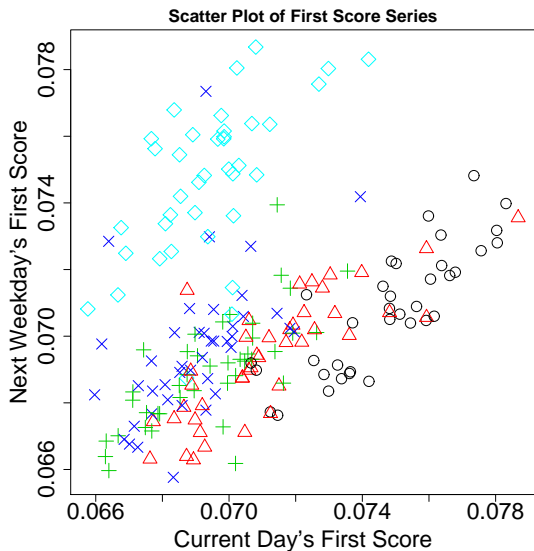
First two intraday factors



Time series plot of first factor score series



Lag-1 scatter plot of first factor score series



Model building

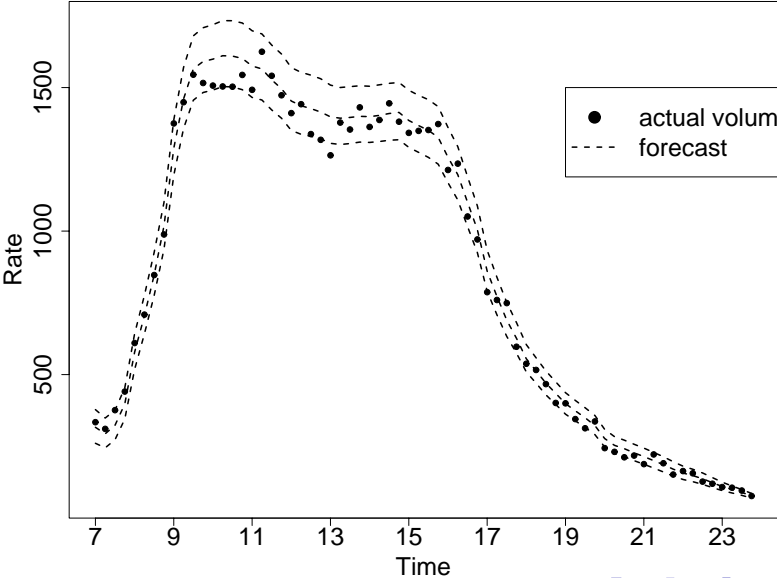
- Varying-coefficient AR(1) model

$$\beta_{i1} = a_1(d_{i-1}) + b_1\beta_{i-1,1} + \epsilon_{i1},$$

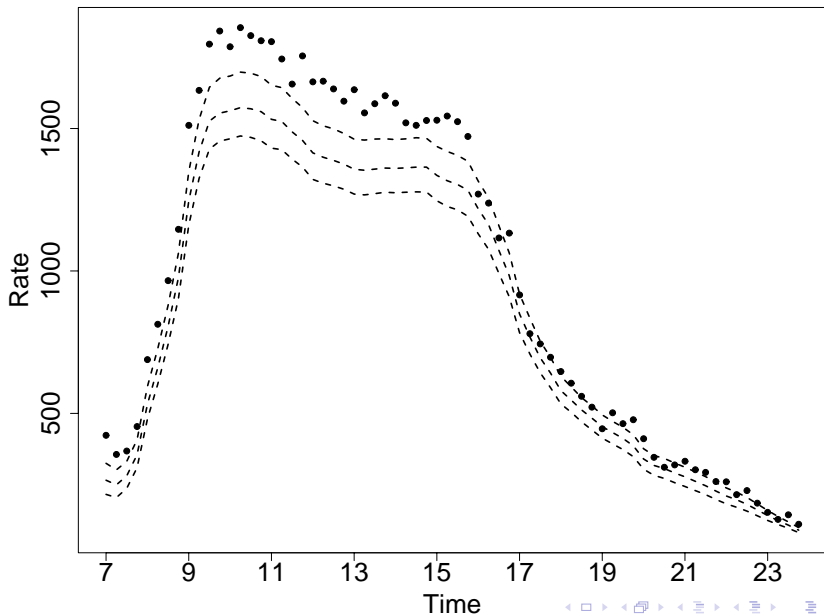
d_{i-1} denotes the day-of-the-week of day $i - 1$

- Varying slope does not provide any significant improvement
- Similar models for additional factors

Distributional arrival-rate forecasts (often) work well



Night-before forecasts can sometimes be off



Dynamic updating: the problem

- Early part of the day: $\mathbf{y}_{(n+1)}^e, \lambda_{(n+1)}^e$
- Latter part of the day: $\mathbf{y}_{(n+1)}^l, \lambda_{(n+1)}^l$
- Two sets of information available
 $\{\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(n)}\}$ and $\mathbf{y}_{(n+1)}^e$
- **Goal:** to obtain updated forecast of $\lambda_{(n+1)}^l$
- Time series forecast does not use new info in \mathbf{y}_{n+1}^e

$$\hat{\lambda}_{n+1}^{l,TS} = \mathbf{g}^{-1}(\hat{\beta}_{n+1,1}^{TS} \mathbf{f}_1^l + \dots + \hat{\beta}_{n+1,K}^{TS} \mathbf{f}_K^l)$$

- How to use the new info?
 - ▶ update $\hat{\beta}_{n+1,k}^{TS}$

Dynamic updating

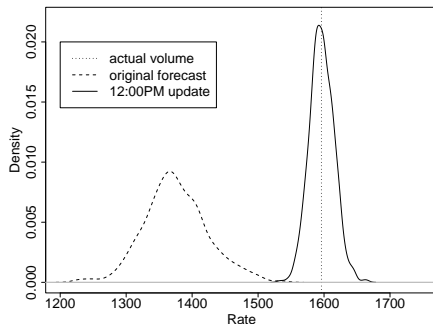
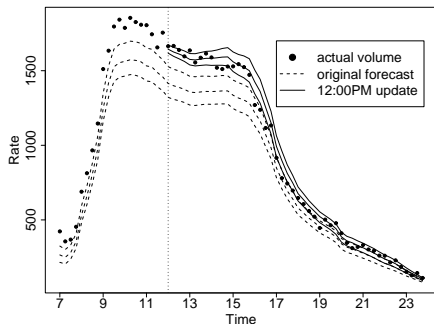
- Estimate $\beta_{(n+1)}$ using info from $\mathbf{y}_{(n+1)}^e$

$$\begin{cases} \mathbf{y}_{(n+1)}^e \sim \text{Poisson}(\boldsymbol{\lambda}_{(n+1)}^e), \\ g(\boldsymbol{\lambda}_{(n+1)}^e) = \mathbf{F}^e \boldsymbol{\beta}_{(n+1)}. \end{cases}$$

- Incorporate info from time series forecasts
- Minimize *penalized likelihood* wrt $\beta_{(n+1)}$

$$-2 \log \text{lik}\{\mathbf{y}_{(n+1)}^e\} + \omega \|\boldsymbol{\beta}_{(n+1)} - \hat{\boldsymbol{\beta}}_{(n+1)}^{\text{TS}}\|^2$$

Forecast updates can significantly reduce error and uncertainty



Implications for workforce management

- Distribution of future $\lambda_{(n+h)}$ determined from the forecast
- With distributions for $\lambda_{(n+h)}$, solve a stochastic program (SP)
- Considering forecast updates, solve a SP with recourse
 - ▶ Changing staffing assignments
 - ★ send agents home early ... → reduce cost
 - ★ call in part-time agents ... → better achieve QoS measure

We test six scheduling schemes

- Two schemes with no updating
 - ▶ one scenario = IP \diamond
 - ▶ 100 scenarios = SP100 \blacklozenge
- Two schemes with an afternoon update of the original schedule
 - ▶ one scenario = UP \square
 - ▶ 100 scenarios = UP100 \blacksquare
- Two schemes that update an original schedule with recourse
 - ▶ one scenario = RP \circ
 - ▶ 100 scenarios = RP100 \bullet

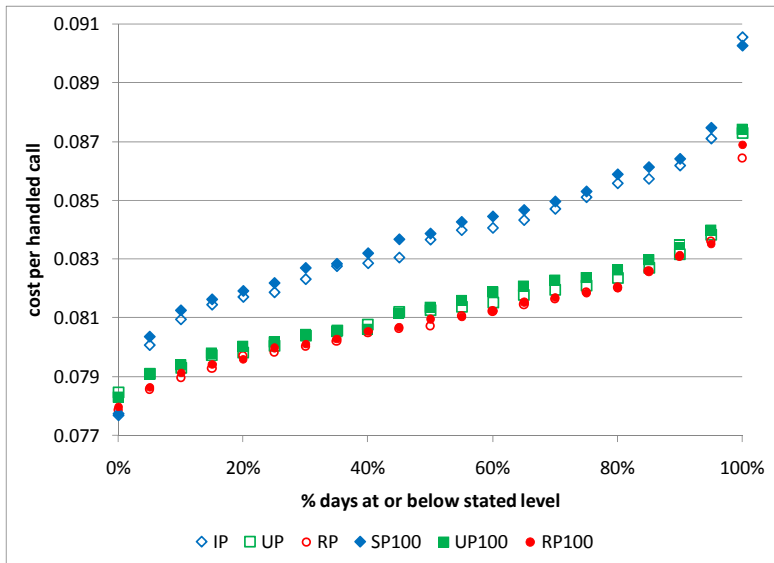
Testing the value of the scheduling schemes

- 1 Preliminary forecast using previous n days of data
- 2 Solve 4 scheduling problems based on initial forecast
 - ▶ IP \diamond and SP100 \blacklozenge
 - ▶ 1st phase of RP \circ and RP100 \bullet
- 3 Update forecast based on 1st part of day
- 4 Update solutions based on revised forecast
 - ▶ IP \Rightarrow UP \square and SP100 \Rightarrow UP100 \blacksquare
 - ▶ 2nd phase of RP \circ and RP100 \bullet
- 5 Simulate using schedules and actual arrival counts

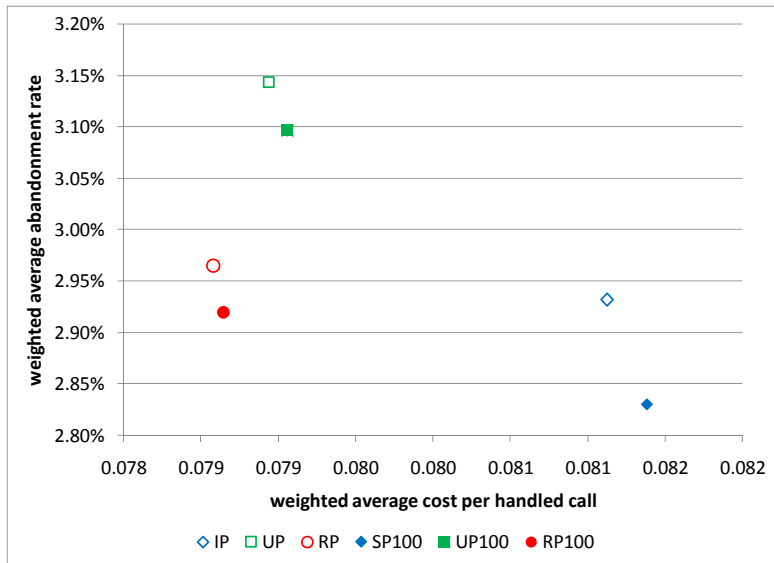
One set of empirical tests

- The same network of four large retail-banking call centers in US
 - ▶ Schedule updates at 11am
- Shift structure and costs
 - ▶ 262 feasible daily schedules (7 and 9-hour shifts, with breaks)
 - ★ cost of 1 per agent per 1/2-hour interval
 - ▶ 4,973 potential recourse actions (with 1/2-hour costs)
 - ★ send home (-0.75), overtime (1.5), call in (2.0)
- Arrival data, forecasts, and QoS target
 - ▶ Last 100 days as testing set
 - ▶ Forecasts based on previous (rolling) 110 days of data
 - ▶ Target expected abandonment rate of 3% across scenarios

Updating systematically lowers cost per call



RP and RP100 saving 3.2%–3.5% vs IP and SP100



Collaborators

- Arrival rate forecasting/updating
 - ▶ Jianhua Huang (Texas A&M, Statistics)
- Scheduling
 - ▶ Noah Gans (Wharton, OM)
 - ▶ Yong-Pin Zhou (U Washington, OM)

Acknowledgement of Support

- NSF grant DMS-0606577 (Statistics)
- NSF grant CMMI-0800575 (Service Enterprise Engineering)

Questions?

