Data-Driven Service Networks:

Models of congested service systems
(e.g. hospitals, call-centers, courts, …)

“Theompirical” Journeys in Service Systems
OR = SE/IE + OM + DS viewpoints

Avishai Mandelbaum

IEM & SEELab, Technion
link

My First Visit to Milan

- Guest of Mediolanum Bank (then approx. 5000 agents)
- Flight, Hotel, San Siro

- Two Lectures: **Management** = fine; **Team Leaders** = **adventure** for 2 reasons:

  1. Language

My First Visit to Milan

- Guest of Mediolanum Bank (then 5000 agents)
- Flight, Hotel, San Siro

- Two Lectures: Management – fine; Team Leaders – adventure for 2 reasons:
  1. Language

  2. Teaching “staffing” via “Offered-Load calculations” (averages):
  – Suppose 25 calls per min x 4 minute per call = 100 Erlangs = **Offered-Load (min-work per min)**
  – Then one needs “around” 100 agents to accommodate the offered-load:

My First Visit to Milan

- Guest of Mediolanum Bank (then 5000 agents)
- Flight, Hotel, San Siro

- Two Lectures: Management – fine; Team Leaders – adventure for 2 reasons:

 1. Language

 2. Teaching "Offered-Load calculations" (averages):

 – Suppose 25 calls per min x 4 minute per call = 100 Erlangs = **Offered-Load (min-work per min)**
 – Then one needs "around" 100 agents to accommodate the offered-load:

   - Much less: **Efficiency-Driven** (ED-regime)
   - Much more: **Quality-Driven** (QD, e.g. 200 agents implies 100/200 = 50% utilization)
   - Aim at the **QED regime** = Quality- and Efficiency-Driven, via **Square-Root Staffing**:

Number of agents = $100 + \beta\sqrt{100}$, β in [-1, 1]

$100 + 10\,\beta$ approx. in **[90 , 110]**, then **refine**

My First Visit to Milan

- Guest of Mediolanum Bank (then 5000 agents)
- Flight, Hotel, San Siro

- Two Lectures: Management – fine; Team Leaders – adventure for 2 reasons:

  1. Language

  2. Teaching “Offered-Load calculations” (averages):
  - Suppose 25 calls per min x 4 minute per call = 100 Erlangs = **Offered-Load (min-work per min)**
  - Then one needs “around” 100 agents to accommodate the offered-load:

  Can be explained also via a universal conservation law:

  Little’s Law

  L = number in system, λ = throughput-rate, **W** = time in system

  L = $\lambda \times$W (finite-horizon, long-run/steady-state)

Research Partners (35 years)

Students:
Aldor, Baron Yonit, Carmeli-Yuviler Nitzan, Carmeli Boaz, Chen Hong, Cohen Izik, Feldman Zohar, Garnett, Ghebali, Gurvich, Khudyakov, Koren, Maman, Marmor, Reich, Rosenshmidt, Shaikhet, Senderovich, Tseytlin, Yom-Tov, Zaied, Zeltyn, Zychlinski, Zohar Eti, Zviran, …

Theory:
Armony, Atar, Azriel, Chen Hong*, Cohen Izik*, Garnett*, Gurvich*, Feigin, Gal, Huang Junfei, Jelenkovic, Kaspi, Massey, Momcilovic, Reiman, Shimkin, Stolyar, Trichakis, Trofimov, Wasserkrug, Whitt, Yom-Tov*, Zeltyn*, Zhang Jiheng, Zhang Hanqin, …

Exploratory Data Analysis, Data Sources, Statistics, Projects:
Brown, Gans, Shen Haipeng*, Sakov, Zhao Linda; Zeltyn*; Ritov, Goldberg*; Gurvich*, Huang Junfei*, Liberman*; Liu Nan, Ye Han; Armony, Marmor*, Tseytlin*, Yom-Tov*; Gorfine, Ghebali*; Tezcan; Kim Song-Hee, Won Chul Cha; Feigin, Azriel*; Rafaeli; Momcilovic, Trichakis; Bunnell, Kadish, Leib; …

Industry:
Mizrahi Bank, Fleet Bank, Rambam Hospital, IBM Research, Hapoalim Bank, Pelephone Cellular, Samsung Hospital, Singapore Hospitals, Dana Farber Cancer Institute, LivePerson, Cheetah Labs,…

Technion SEE Laboratory (SEELab):
Feigin; Trofimov, Nadjharov, Gavako; Kutsy; Senderovic*, Carmeli*; Liberman*, Koren*, Plonsky*;
Research Assistants, Visitors, Postdocs, …

Note: In many Western countries, there is a short list of popular "first names," but countless "last names." In China, it is just the reverse. The list of last names is short, and the number of first names is in the billions (from chinapage.com/biography/lastname.html).

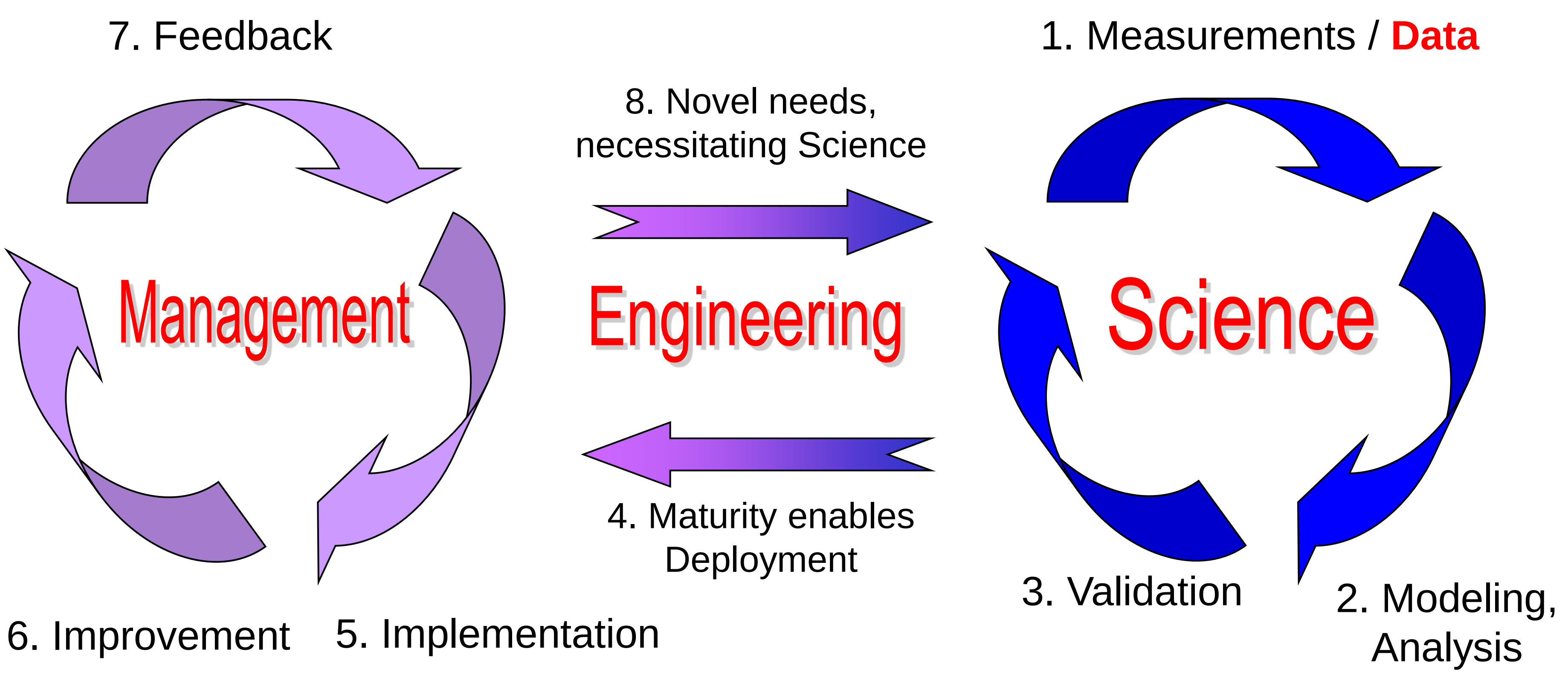
Service Science & Engineering

- Service constitutes 60-80% of the west economy
- Creating most new jobs
- Service processes are technology-intensive and data
- Require customization
  - For example, Hospitals (Emergency Room): 3M USA nurses
  - Or Call Center (IVR): 3M agents

Scope of the Service Industry

Guangzhou Railway Station, Southern China

Service Science, Engineering, Management



ER / ED Environment: Service Network

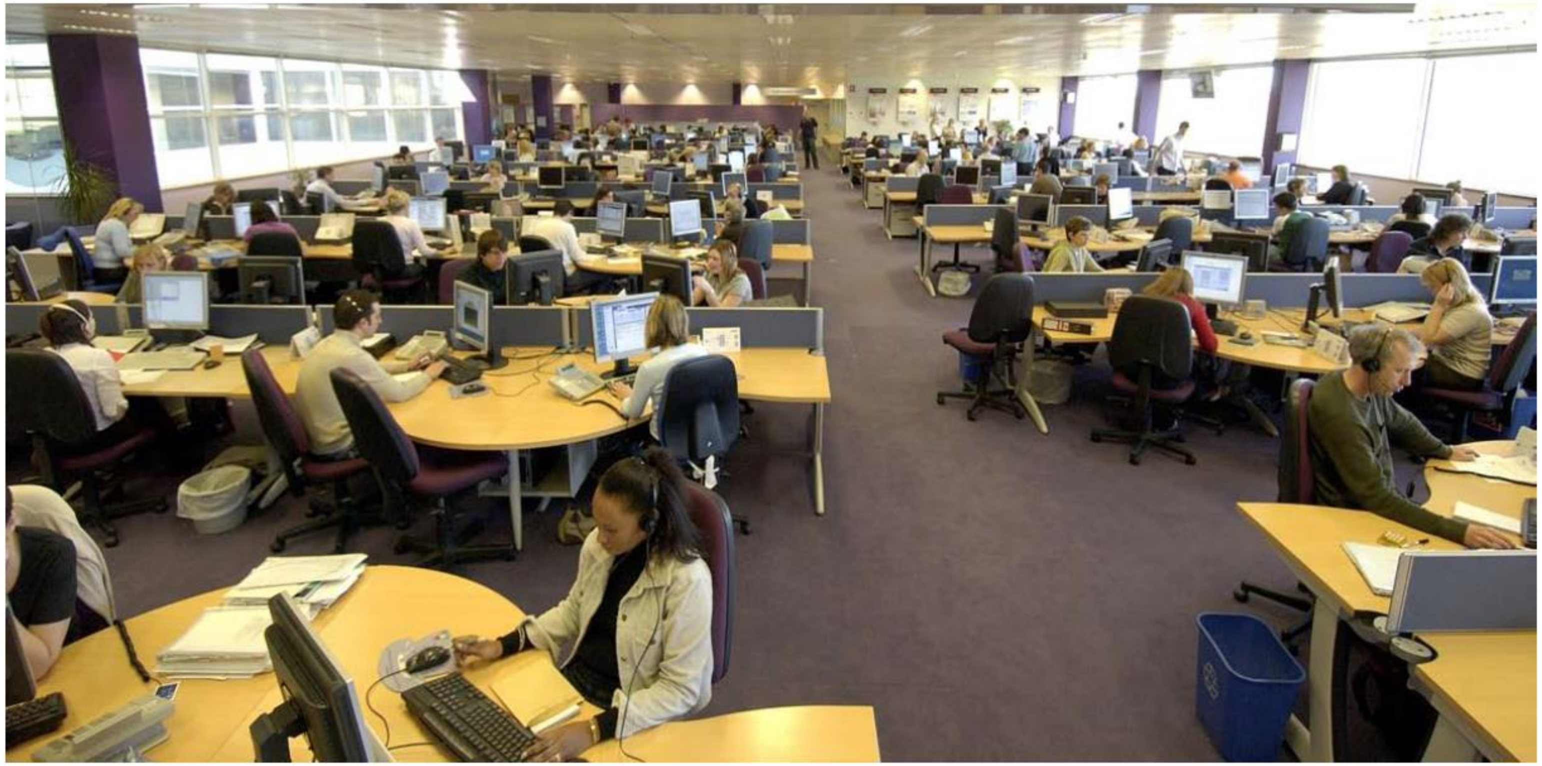
Acute (Internal, Trauma)

Walking

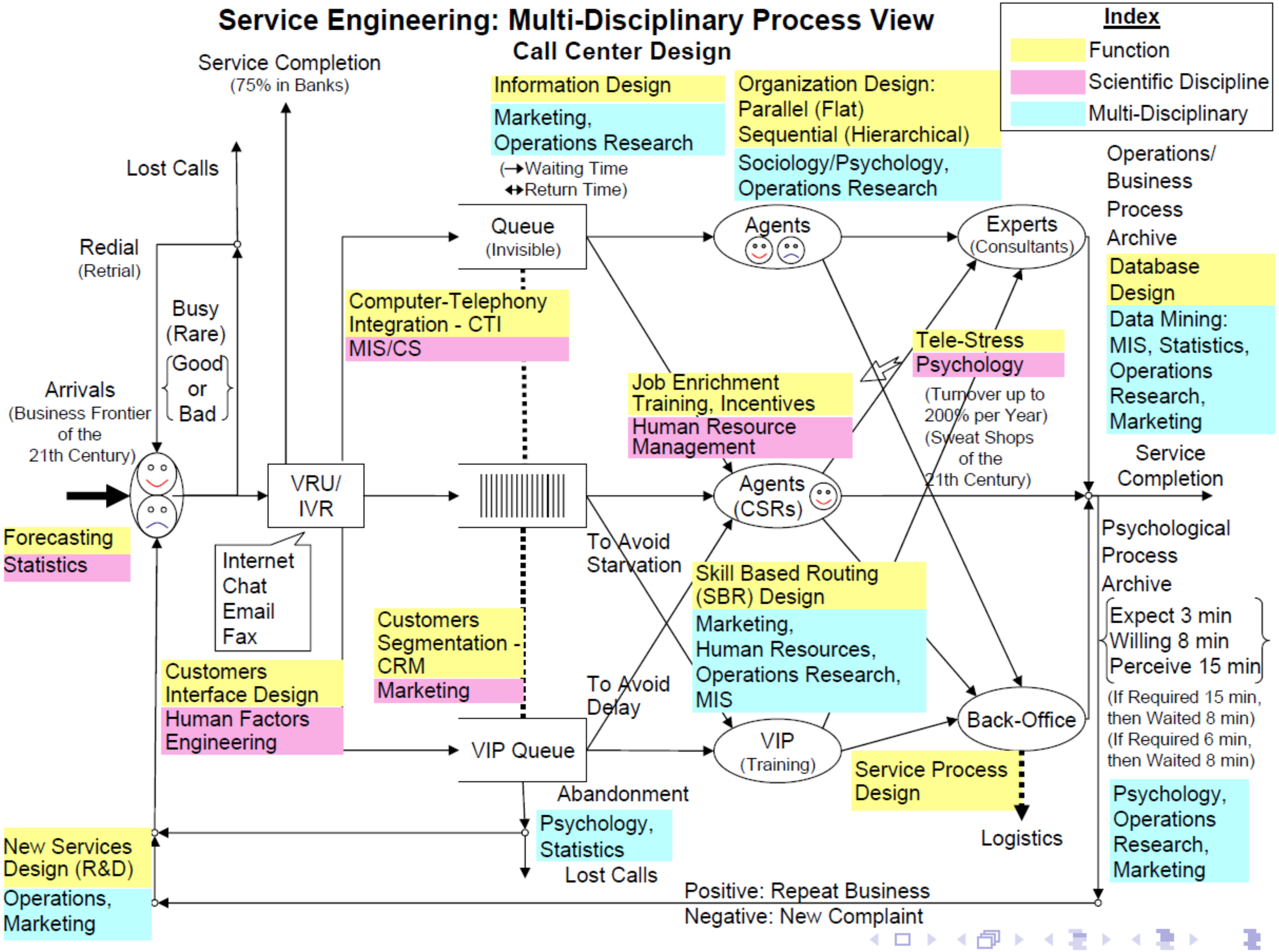
Multi-Trauma

Call-Center Environment: Service Network

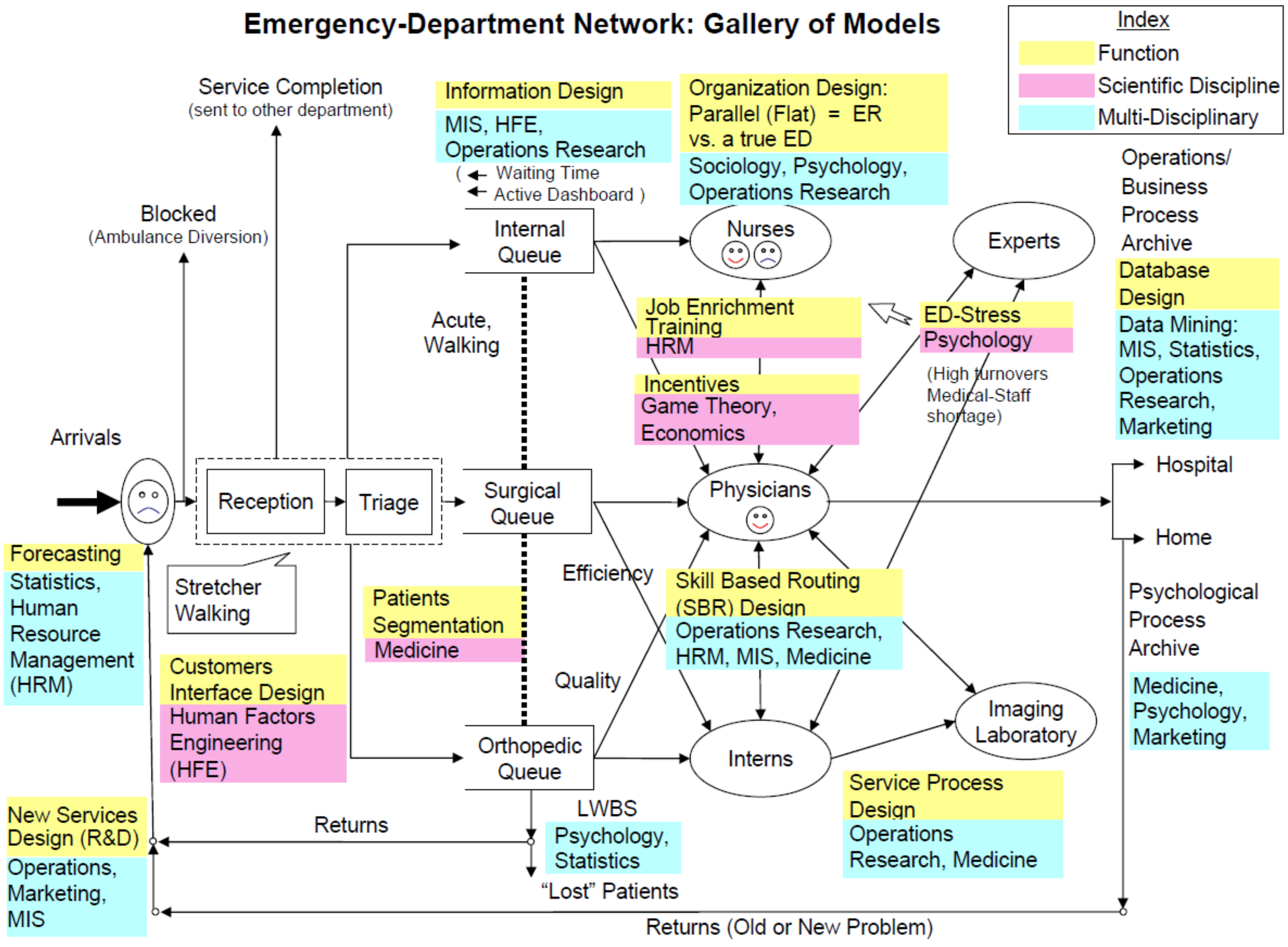
= "Fruit-flies of Hospitals": fast, low-stake, no IRB, . . .
yet highly relevant



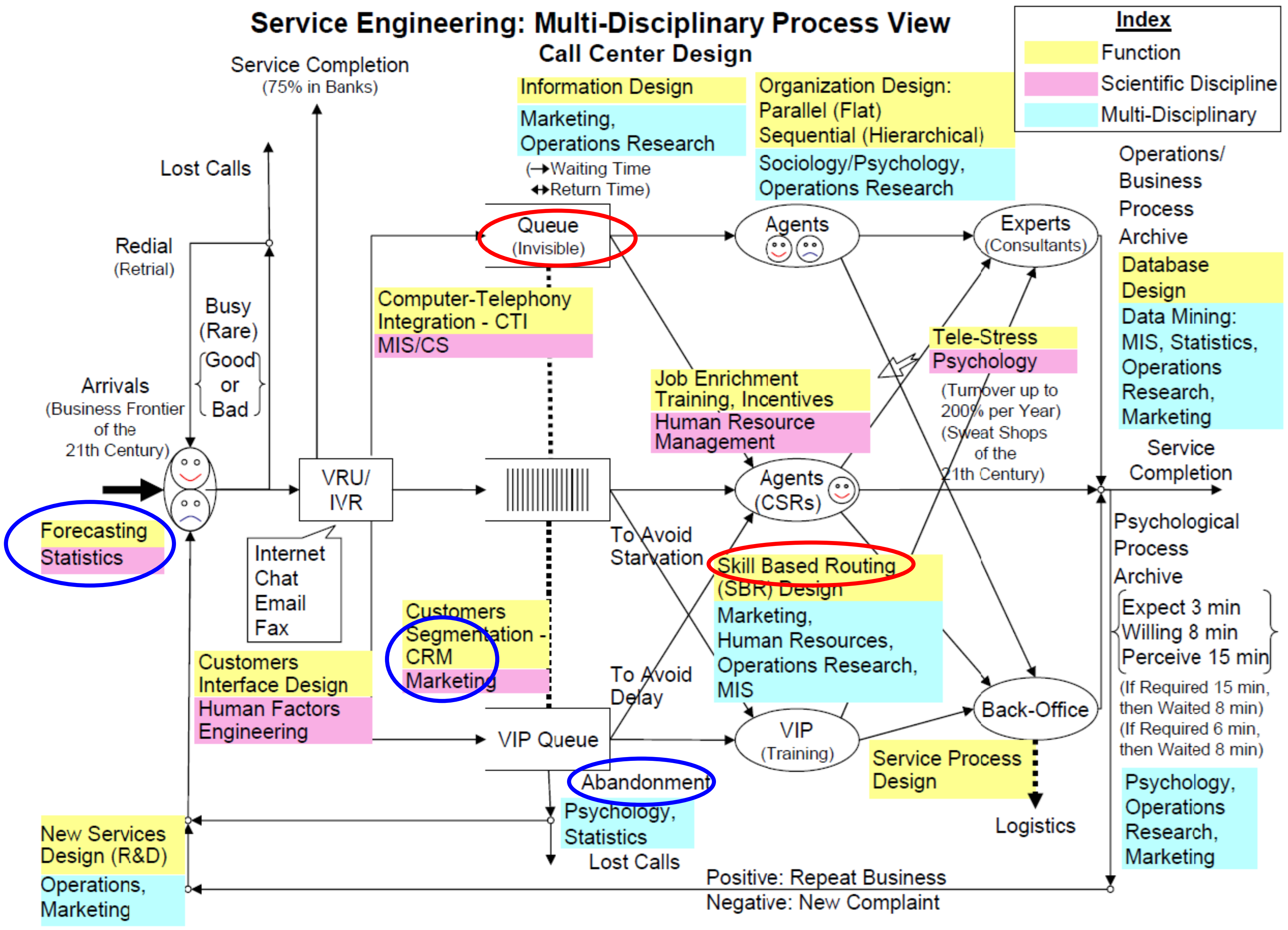
Call-Center Network: Gallery of Models



Emergency-Dept.: Multi-Disciplinary ServEng View



Call Centers = Fruit-Flies of Hospitals



Skeptic View (of Operations-Research, Service Engineering) prompted **Call Centers = Fruit-Flies of Hospitals**

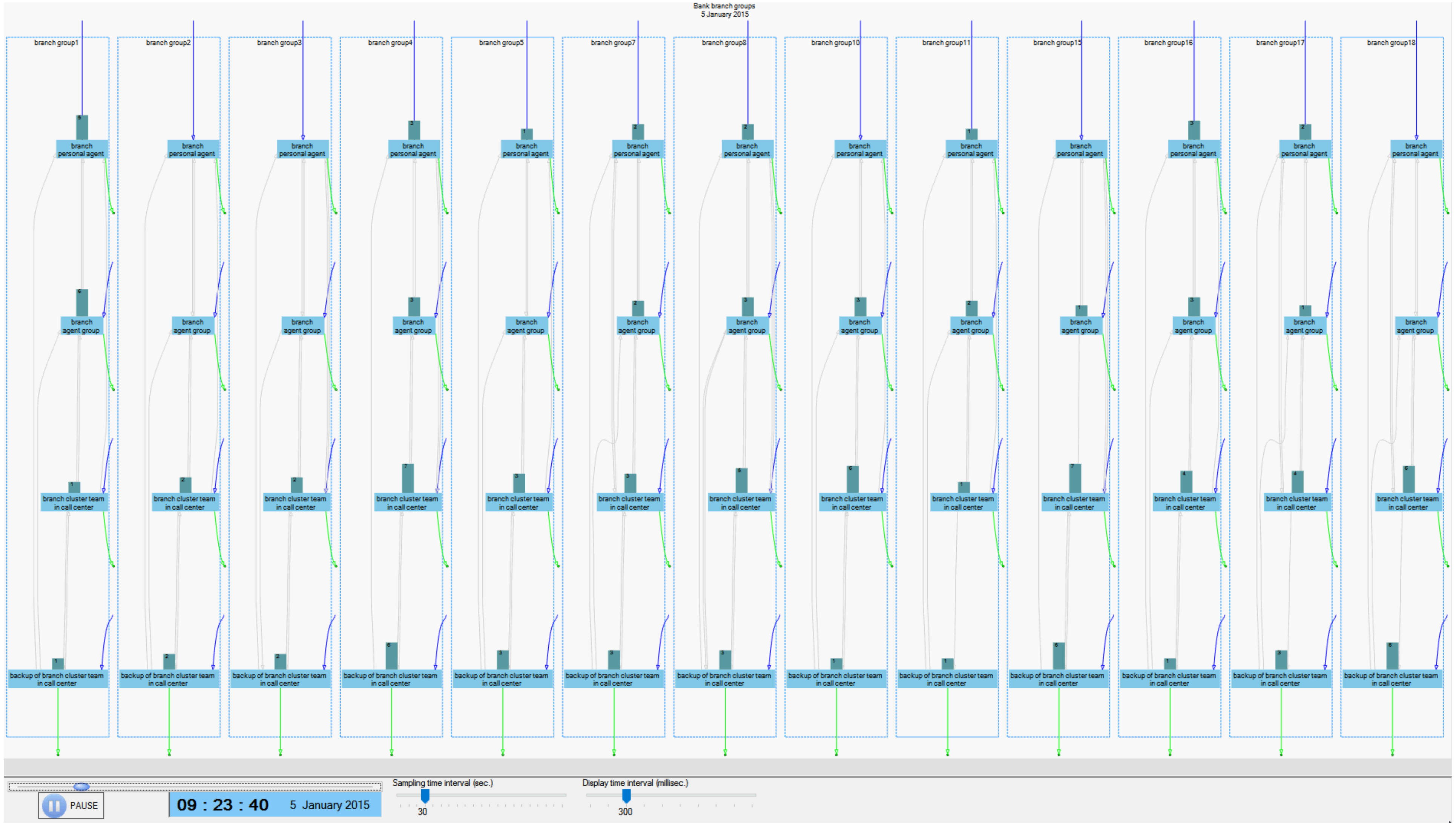
- (very) Short History of Fruit-Flies Research:

http://www.youtube.com/watch?v=bKrpnfTISaE

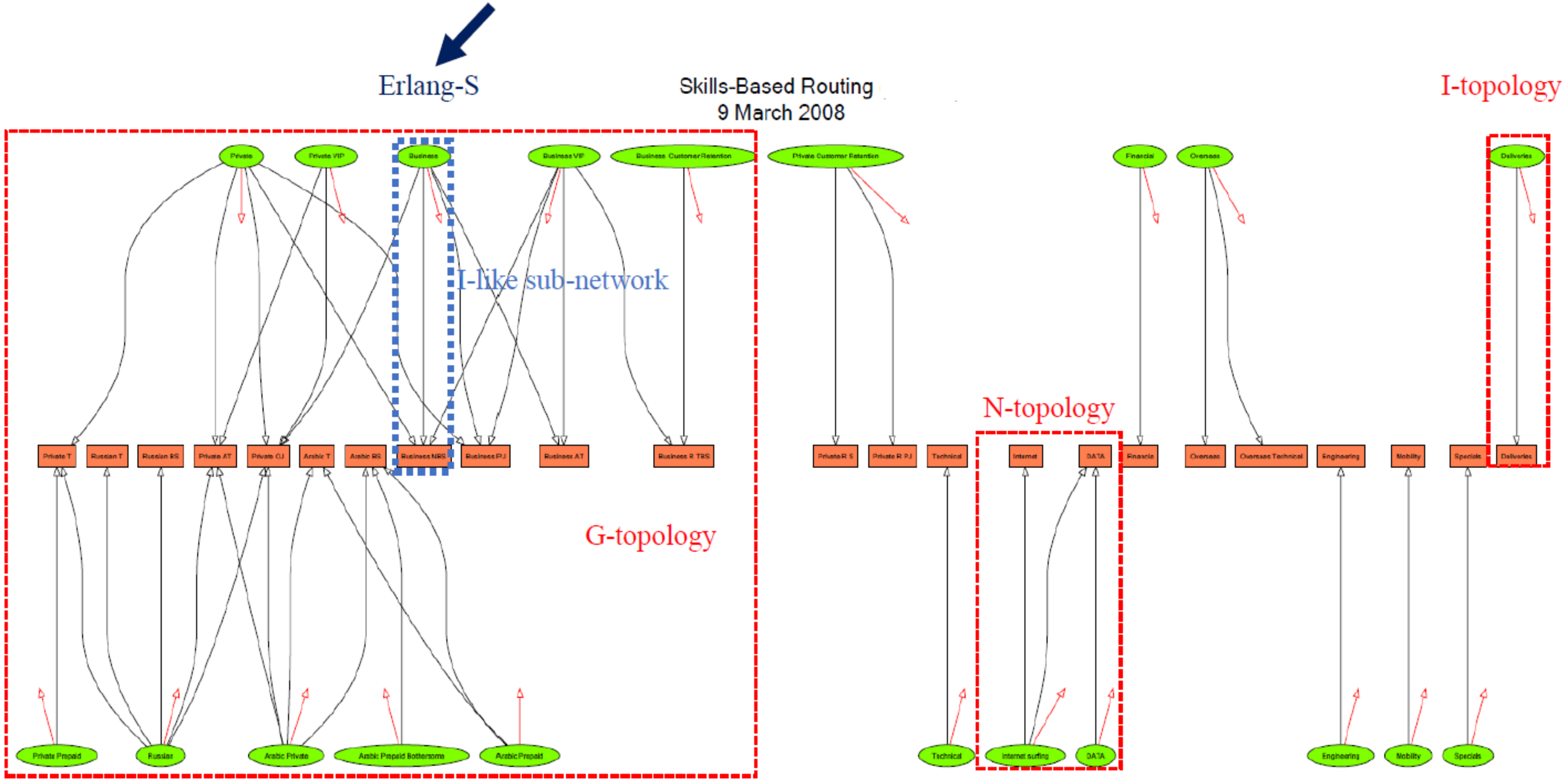
- **A** politician's view-point (also triggered the above):

http://www.youtube.com/watch?v=xao_4Y-lOdk

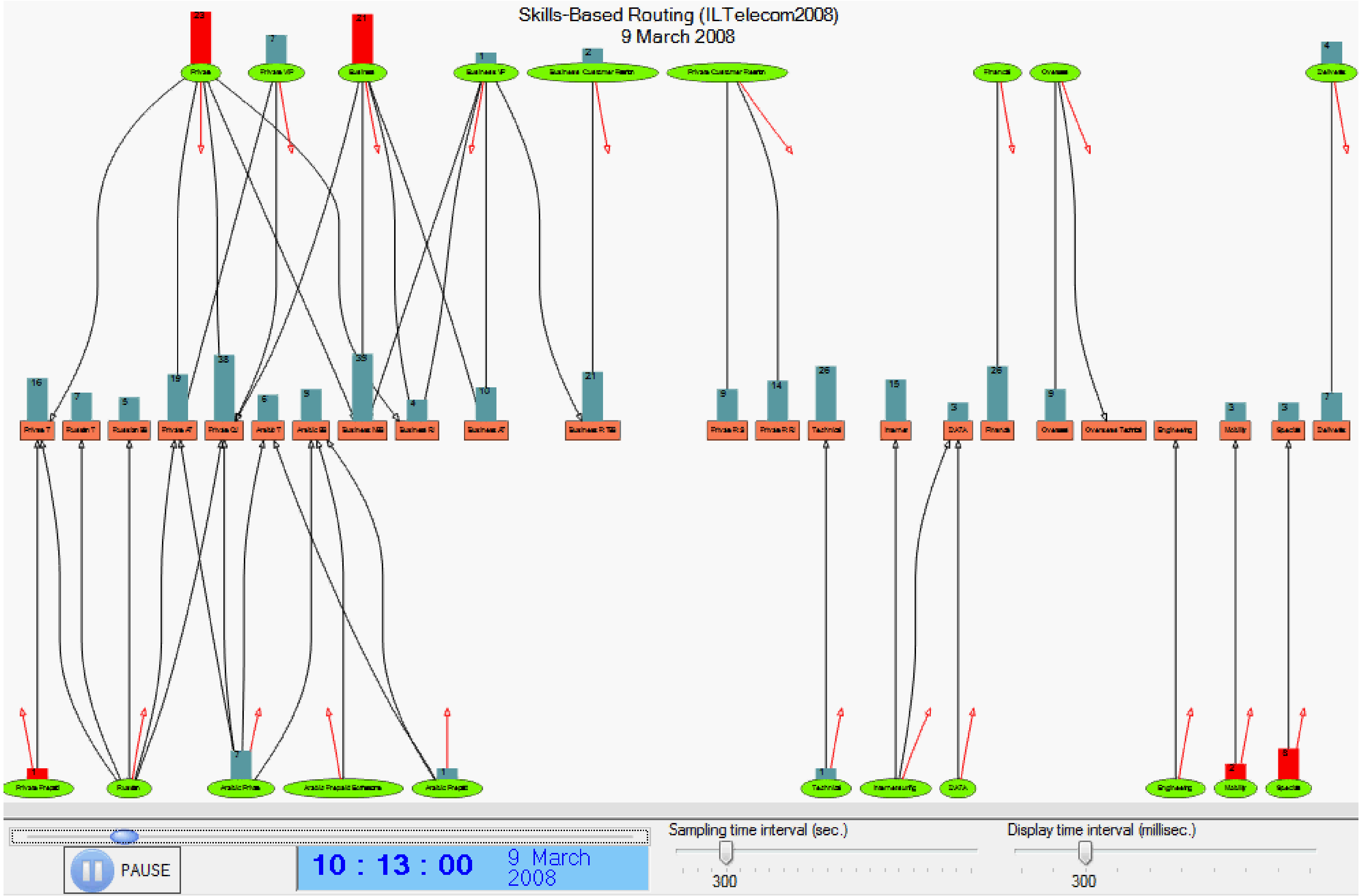
Telephone Queues: 2000 Bank-Agents, in Call-Centers + Branches (ILDU Bank)

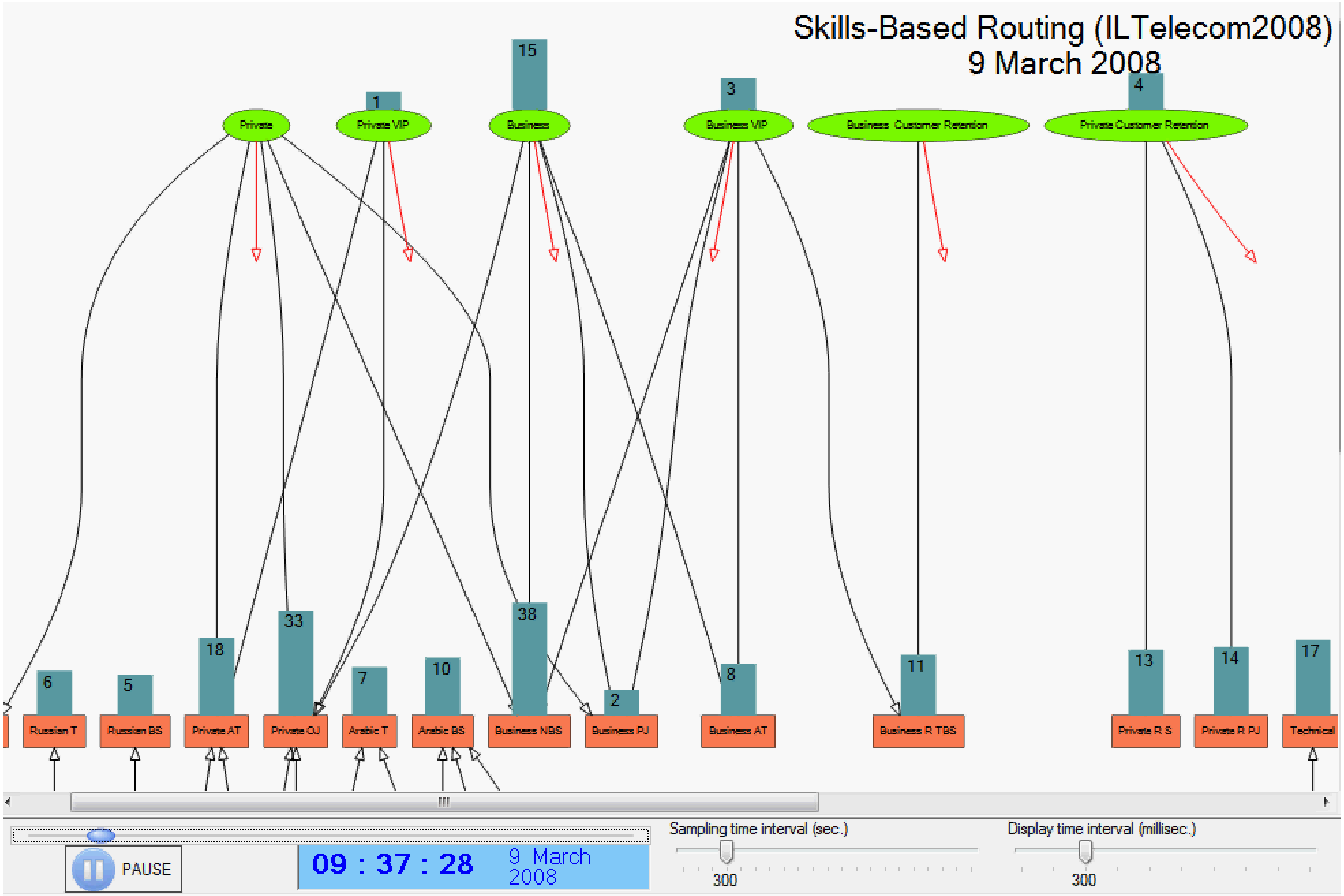


Call Center = Matching Customers & Agents (Needs & Skills)
Hospital = Patients Wards



Topology of a call center:
Server-queues are in the rectangles and customer-queues are in the ovals


Skills-Based Routing (ILTelecom2008)
9 March 2008
PAUSE
10 : 13 : 00
9 March 2008
Sampling time interval (sec.)
300
Display time interval (millisec.)
300



Skills-Based Routing (ILTelecom2008)
9 March 2008
Private
Private VIP
Business
Business VIP
Business Customer Retention
Private Customer Retention
15
1
3
4
6
5
18
33
7
10
38
2
8
11
13
14
17
Russian T
Russian BS
Private AT
Private OJ
Arabic T
Arabic BS
Business NBS
Business PJ
Business AT
Business R TBS
Private R S
Private R PJ
Technical
PAUSE
09 : 37 : 28
9 March 2008
Sampling time interval (sec.)
300
Display time interval (millisec.)
300

Skills-Based Routing in a Call Center

- **Customer Classes**
  - Marketing segregates customers according to their needs and/or importance – this determines customer priorities

- **Agent Skills**
  - Human-Resources Management assigns agent skills according to capabilities, experience (training) – this determines agent constituencies

- **Matching Class & Skill (Demand and Supply)**
  - Operations-Researchers develop matching algorithms so that customers don't wait long for an agent (service-level) and agents don't wait long for a customer (efficiency)

- **Information Infrastructure** (IS/CS)

- **Data Management** (Statisticians)

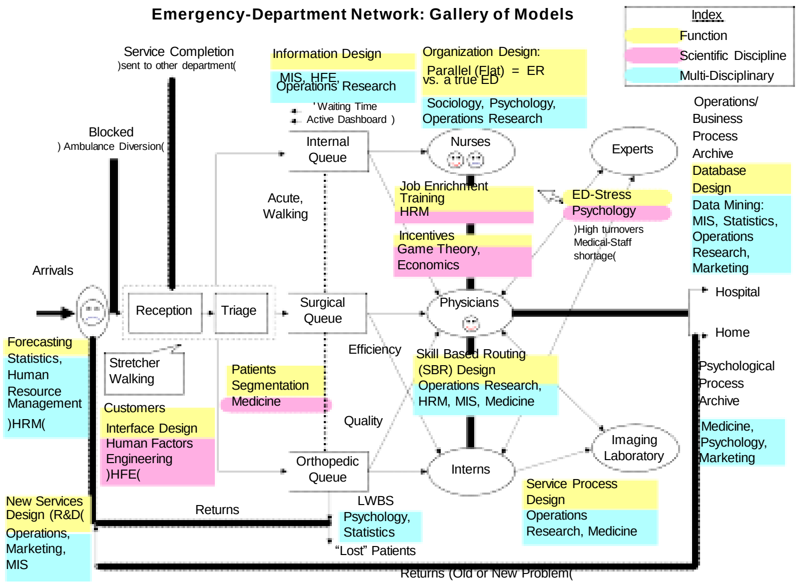
OR SBR Theoretical Support: Asymptotic Analysis in Heavy-Traffic (Stochastic Control**)**

Evidence-Based Routing in an ER/ED

- **Patient Priorities (Customer Classes)**
  - Doctors classify patients according to their urgency and/or needs – this determines customer priorities
- **Ward "Skills" (Agent Skills)**
  - Management assigns "skills" to wards according to clinical relevance and capabilities – this determines customer constituencies
- **Matching Class & Skill (Demand and Supply)**
  - Operations-Researchers develop matching algorithms so that patients don't wait long for a ward (service-level) and wards don't wait long for a patient (efficiency)
- **Information Infrastructure** (IS/CS)
- **Data Management** (Statisticians)

OR SBR Theoretical Support: Asymptotic Analysis in Heavy-Traffic (Stochastic Control)

Theses: Control of Patient Flow (Hospital Network)



- Environment-Dependent **ED Flow Design**, w/ **Marmor, Golany, Israelit**
- **Fair ED-to-IW Routing** (Patients vs. Staff), w/ **Momcilovic, Tseytlin**
- **Staffing Time-Varying Q's with Re-Entrant Customers**, w/ **Yom-Tov**
- **Queueing-Science/Congestion-Laws**, w/ **Armony, Marmor, Tseytlin, Yom-Tov; Israelit**

- **Blocking (ED to IW, IW to Geriatric-Institutions)**, w/ **Zychlinski, Cohen, Momcilovic**
- **Triage (Deadlines) vs. In-Process (Queueing)**, w/ **Huang ,Carmeli**

On Asymptotic Research of Queueing Systems

Queueing asymptotics has grown to become a central research theme in Operations Research and Applied Probability, beyond just queueing theory. Its claim to fame has been the deep insights that it provides into the dynamics of Queueing Networks (**QNets**), and rightly so:

- Kingman's invariance principle in conventional heavy-traffic
- Whitt's sample-path (functional) framework
- Reiman's network analysis via oblique reflection
- Bramson-Williams' framework for state-space collapse
- Laws' resource pooling
- Harrison's paradigm for asymptotic control (Wein; van Mieghem's $Gc\mu$)
- Dai's fluid-based stability
- Halfin-Whitt's (QED regime) ($\sqrt{}$-staffing for many-server queues)
- $P = NP$: Atar's equivalence of Preemptive and Non-Preemptive SBR; Stolyar, Gurvich
- Massey-Whitt's research of time-varying queues

Applying Queueing Asymptotics

- Has asymptotic research helped one approximate or simulate a service system more efficiently, estimate its parameters more accurately, teach it to our students more effectively, perhaps even manage the system better?

- I am of the opinion that the answer to such questions **could and should have been "yes"** more often than it has been.

- How to change this?

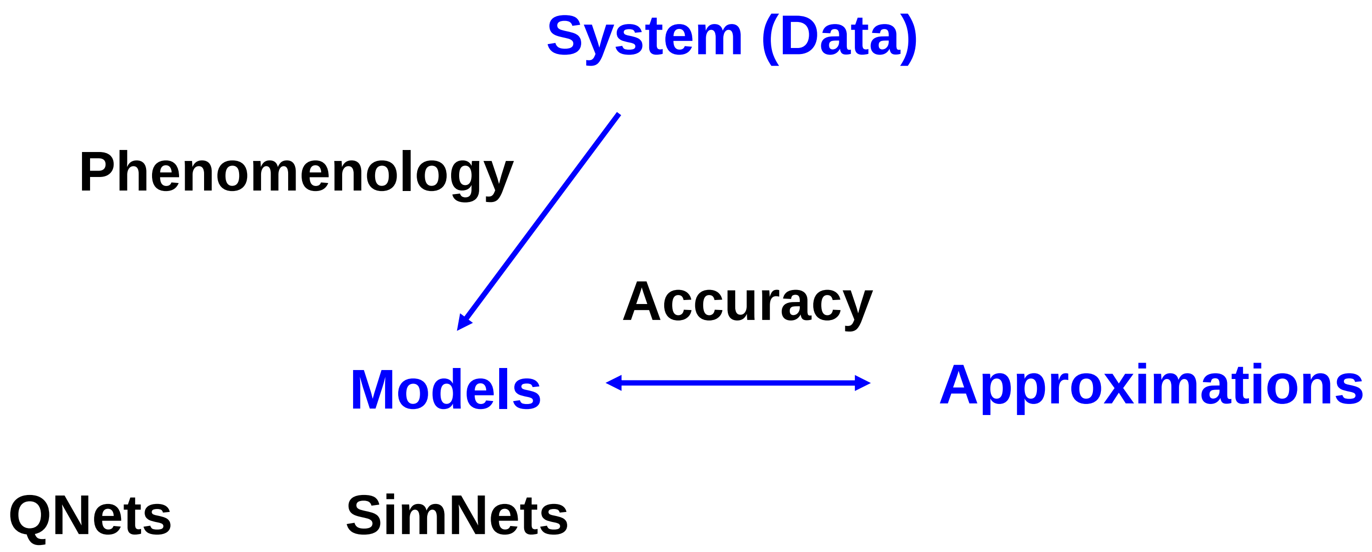
  My approach has been to **marry theory with data ("theompirical" research)**, supported by (what only in recent years I came to realize is) **Process Mining (= of Stochastic Networks**: their building-blocks, structure, protocols; flows and laws).

  Why be optimistic? Hopefully this course!

Accuracy vs. Value

Applies to any approximation scheme that approximates theoretical models by other theoretical models.

Prevalent (Asymptotic) Approximations



Data–Based Prevalent Asymptotic Approximations Models

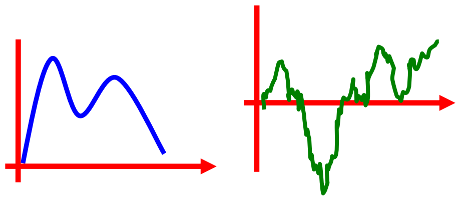
System (Data)

Phenomenology

Accuracy

Models ⟷ Approximations

QNets SimNets ⟷ FNet DNet Models



Data–Based Prevalent Asymptotic Approximations Models

System (Data)

Phenomenology

Value

Accuracy

Models

Approximations

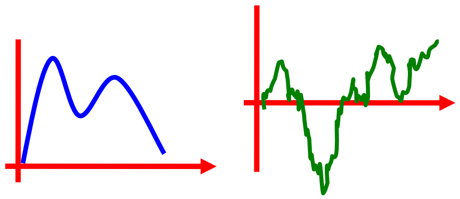
QNets

SimNets

FNet

DNet

Models



Lecture 2: Start Here

Technion SEELab

SEE = Service Enterprise Engineering

Home for my **“Theompirical” Research**

Since 2007:
Data for Research & Teaching

$1M seed: Hal & Inge Marcus
3 Researches (professionals)

- **Students, PostDocs**
- **Academic Visitors**
- **Mirror Servers**

Technion SEELab

SEE = Service Enterprise Engineering

(There is a "SEE Lab" in Bocconi: SEE = Space Economy Evolution, Born 4/6/18)

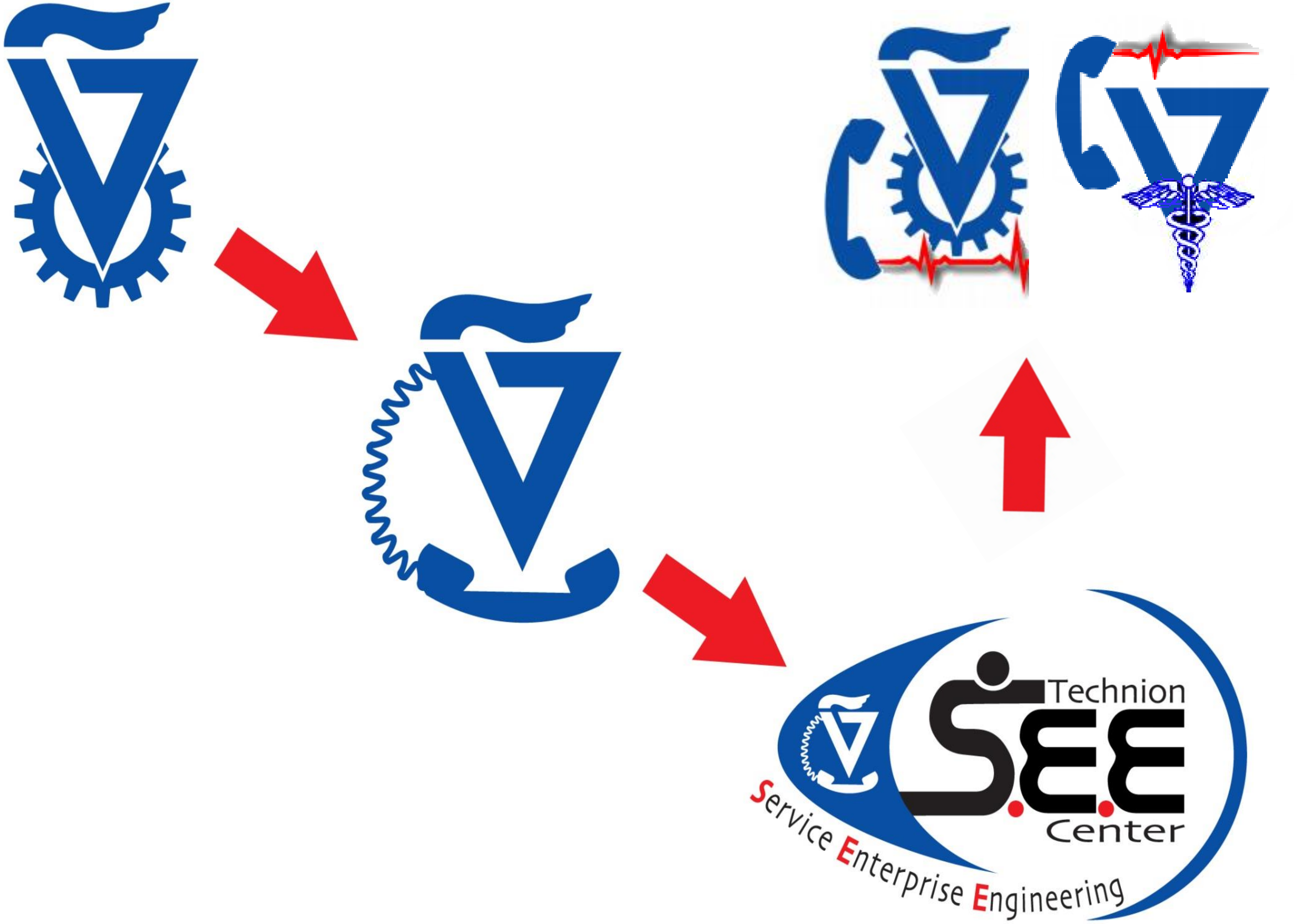
Since 2007:
Data for Research & Teaching

$1M seed: Hal & Inge Marcus
3 Researches (professionals)

- **Students, PostDocs**
- **Academic Visitors**
- **Mirror Servers**

Technion S.E.E Center

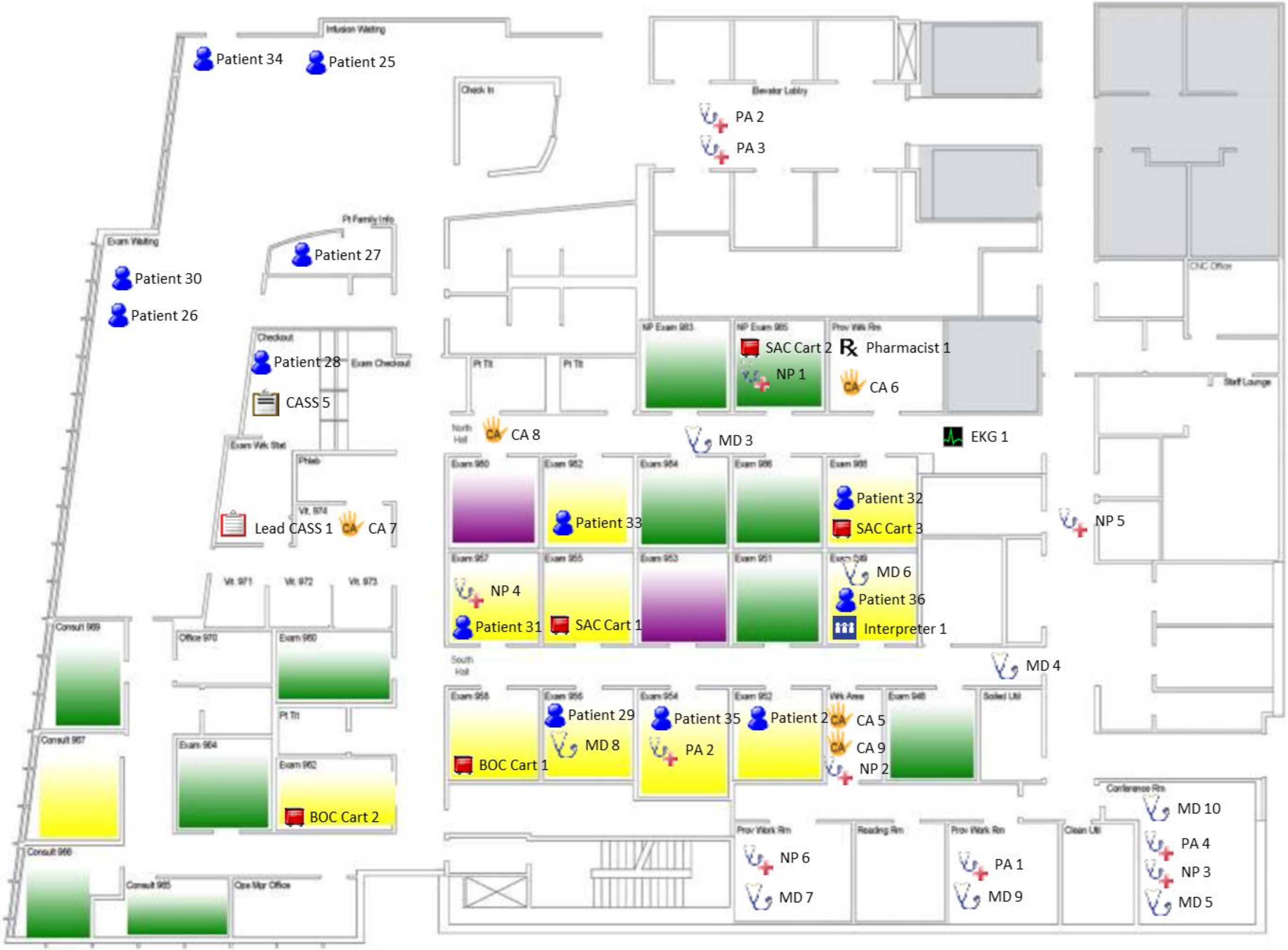
Service Enterprise Engineering

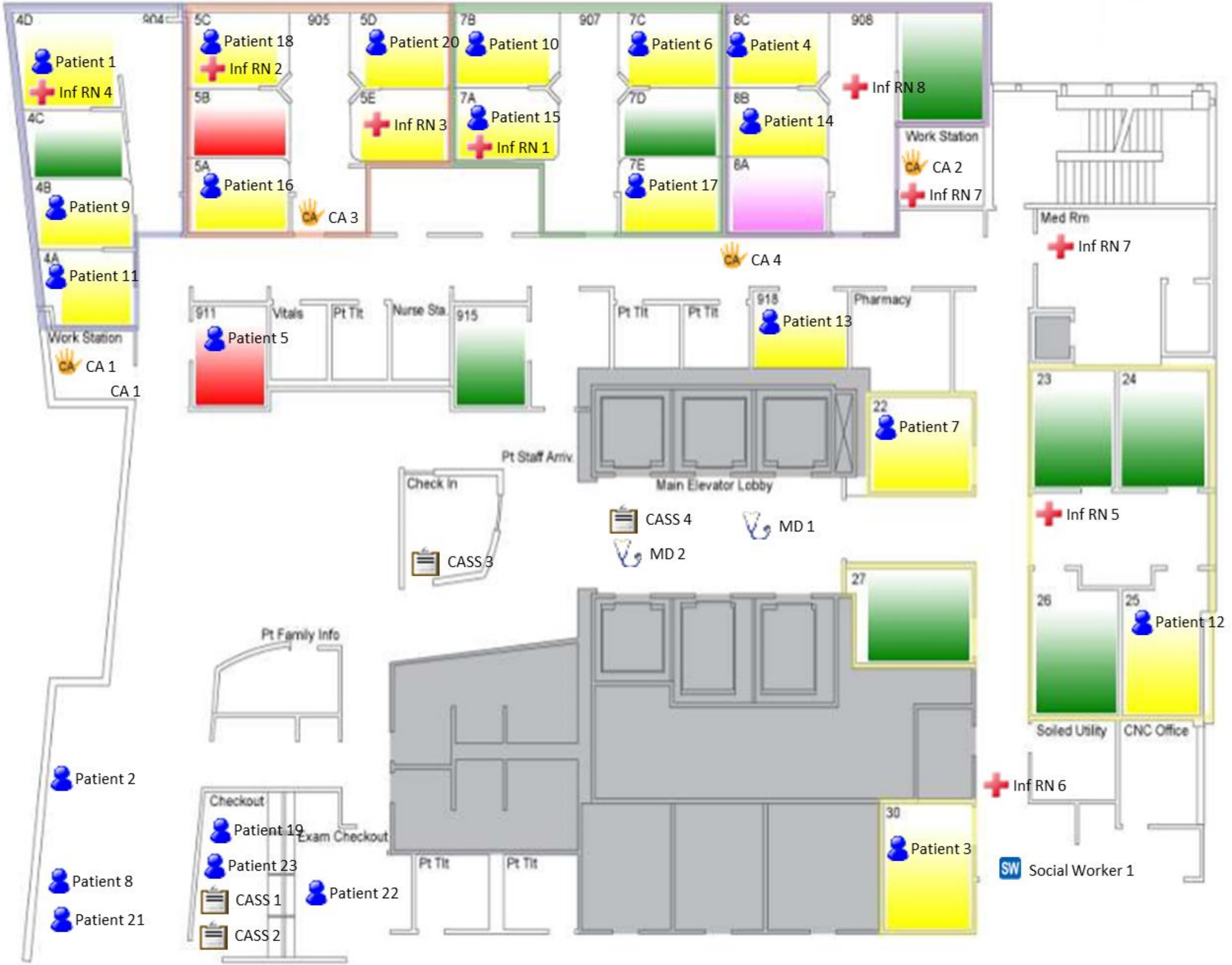

Technion
S.E.E
Center
Service Enterprise Engineering

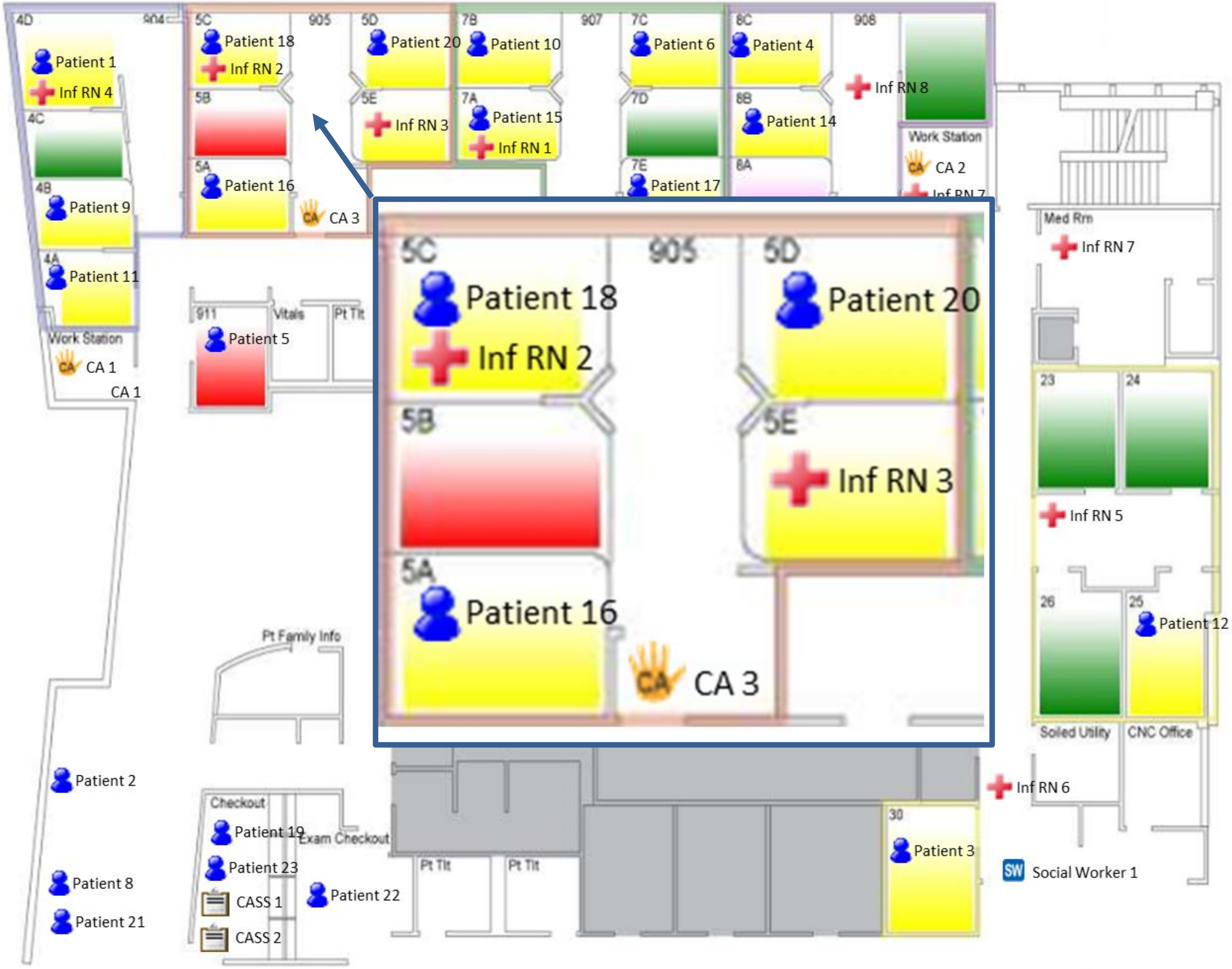
Closing the Data-Gap: from Call-Centers to Hospitals, now Banks

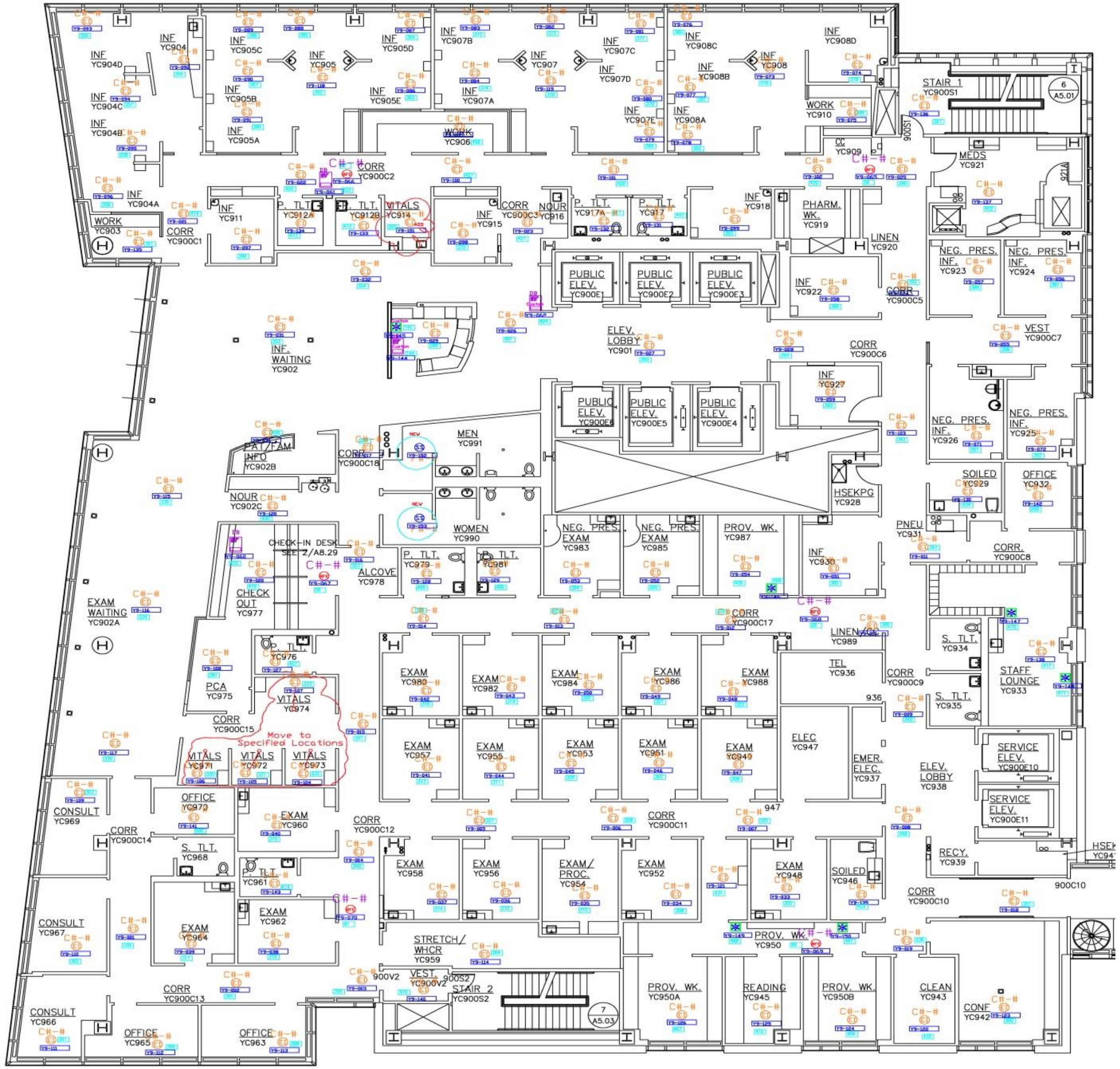
- **Large call center**:
  - 1000s of agents
  - Hundreds of thousands of calls per day
  - Data: operational, psychological, financial – **automatic** collection
- **Large hospital**:
  - 1000+ Beds
  - 1000s of patients & nurses, hundreds of doctors
  - Data: operational, clinical, financial – mostly **inaccessible (to academia)**

- **Large Bank:** "Enjoys" characteristics of both of the above

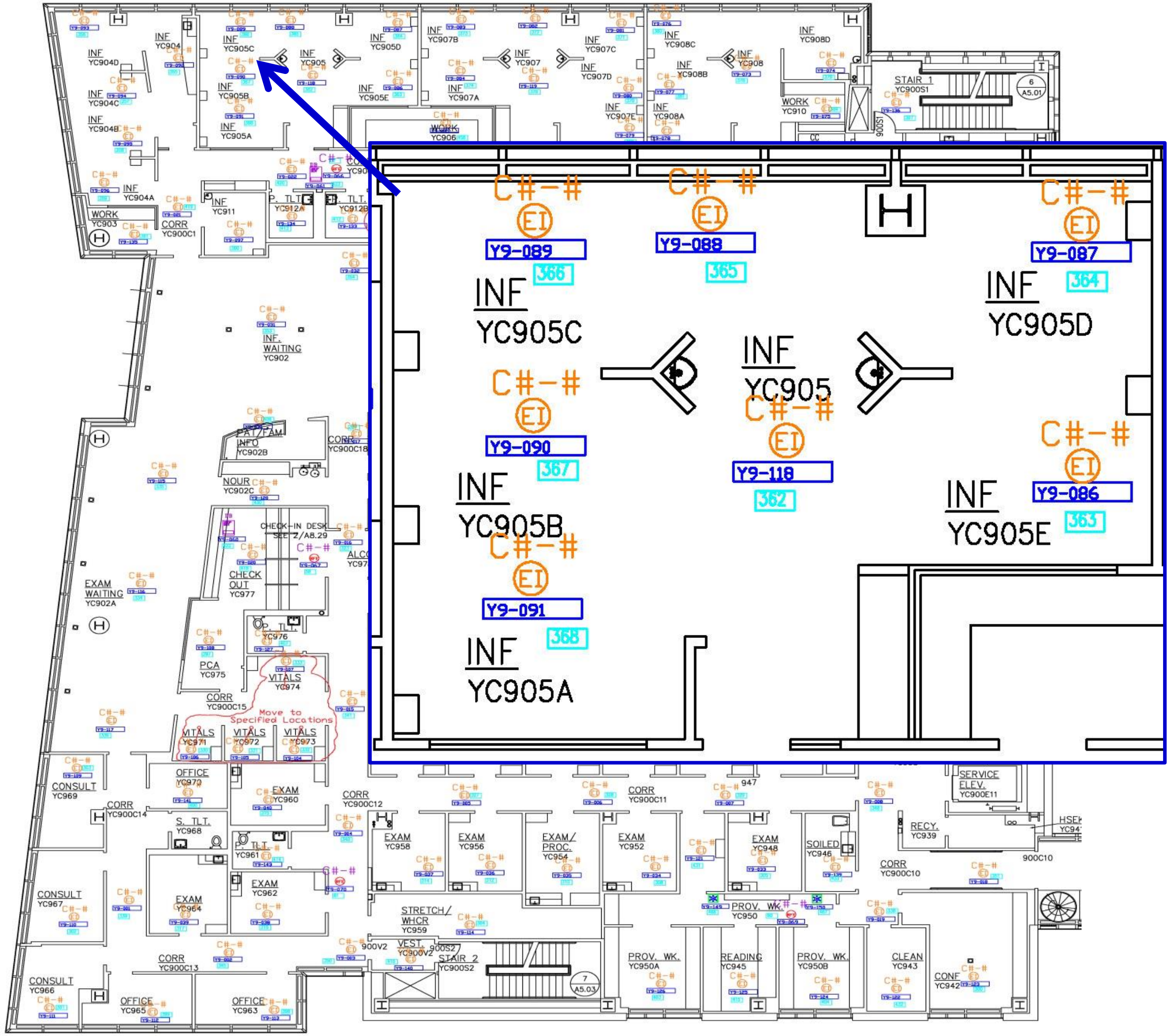

Infusion Waiting
Patient 34
Patient 25
Check In
Elevator Lobby
PA 2
PA 3
Pt Family Info
Patient 27
Exam Waiting
Patient 30
Patient 26
Checkout
Patient 28
Exam Checkout
CASS 5
NP Exam 983
NP Exam 985
SAC Cart 2
Pharmacist 1
NP 1
CA 6
CA 8
MD 3
EKG 1
Staff Lounge
Patient 32
SAC Cart 3
NP 5
Lead CASS 1
CA 7
Patient 33
MD 6
NP 4
Patient 36
Patient 31
SAC Cart 1
Interpreter 1
MD 4
Patient 29
Patient 35
Patient 2
CA 5
CA 9
MD 8
PA 2
NP 2
BOC Cart 1
BOC Cart 2
MD 10
PA 4
NP 3
MD 5
NP 6
MD 7
PA 1
MD 9


4D
904
Patient 1
Inf RN 4
4C
4B
Patient 9
4A
Patient 11
Work Station
CA 1
CA 1
5C
Patient 18
Inf RN 2
905
5B
5A
Patient 16
CA 3
5D
Patient 20
5E
Inf RN 3
7B
Patient 10
7A
Patient 15
Inf RN 1
907
7C
Patient 6
7D
7E
Patient 17
8C
Patient 4
8B
Patient 14
8A
908
Inf RN 8
Work Station
CA 2
Inf RN 7
Med Rm
Inf RN 7
CA 4
911
Patient 5
Vitals
Pt Tlt
Nurse Sta
915
Pt Tlt
Pt Tlt
918
Patient 13
Pharmacy
22
Patient 7
Pt Staff Arriv
Main Elevator Lobby
Check In
CASS 3
CASS 4
MD 1
MD 2
23
24
Inf RN 5
27
26
25
Patient 12
Soiled Utility
CNC Office
Pt Family Info
Patient 2
Checkout
Patient 19
Exam Checkout
Patient 23
CASS 1
CASS 2
Patient 22
Pt Tlt
Pt Tlt
Patient 8
Patient 21
30
Patient 3
Inf RN 6
Social Worker 1


4D
904
5C
Patient 18
Inf RN 2
905
5D
Patient 20
7B
Patient 10
907
7C
Patient 6
8C
Patient 4
908
Patient 1
Inf RN 4
4C
5B
5E
Inf RN 3
7A
Patient 15
Inf RN 1
7D
8B
Patient 14
Inf RN 8
Work Station
CA 2
5A
Patient 16
CA 3
7E
Patient 17
8A
4B
Patient 9
4A
Patient 11
Med Rm
Inf RN 7
911
Vitals
Pt Tlt
Patient 5
Work Station
CA 1
CA 1
23
24
Inf RN 5
26
25
Patient 12
Pt Family Info
Soiled Utility
CNC Office
Inf RN 6
Patient 2
Checkout
Patient 19
Exam Checkout
Patient 23
Pt Tlt
Pt Tlt
30
Patient 3
Social Worker 1
Patient 8
Patient 21
CASS 1
CASS 2
Patient 22



INF YC904D
INF YC904
INF YC905C
INF YC905
INF YC905D
INF YC907B
INF YC907
INF YC907C
INF YC908C
INF YC908
INF YC908D
STAIR 1 YC900S1
INF YC904C
INF YC904B
INF YC905B
INF YC905A
INF YC905E
INF YC907A
WORK YC906
INF YC907D
INF YC907E
INF YC908B
INF YC908A
WORK YC910
CC YC909
MEDS YC921
CORR YC900C2
INF YC904A
WORK YC903
CORR YC900C1
INF YC911
P. TLT. YC912A
P. TLT. YC912B
VITALS YC914
INF YC915
CORR YC900C3
NOUR YC916
P. TLT. YC917A
P. TLT. YC917
INF YC918
PHARM. WK. YC919
LINEN YC920
PUBLIC ELEV. YC900E1
PUBLIC ELEV. YC900E2
PUBLIC ELEV. YC900E3
INF YC922
CORR YC900C5
NEG. PRES. INF. YC923
NEG. PRES. INF. YC924
INF. WAITING YC902
ELEV. LOBBY YC901
CORR YC900C6
VEST YC900C7
PUBLIC ELEV. YC900E6
PUBLIC ELEV. YC900E5
PUBLIC ELEV. YC900E4
INF YC927
NEG. PRES. INF. YC926
NEG. PRES. INF. YC925
PAT/FAM INFO YC902B
CORR YC900C18
MEN YC991
NOUR YC902C
HSEKPG YC928
SOILED YC929
OFFICE YC932
WOMEN YC990
NEG. PRES. EXAM YC983
NEG. PRES. EXAM YC985
PROV. WK. YC987
PNEU YC931
CHECK-IN DESK SEE 2/A8.29
P. TLT. YC979
P. TLT. YC981
ALCOVE YC978
INF YC930
CORR. YC900C8
EXAM WAITING YC902A
CHECK OUT YC977
CORR YC900C17
LINEN YC989
S. TLT. YC934
P. TLT. YC976
EXAM YC980
EXAM YC982
EXAM YC984
EXAM YC986
EXAM YC988
TEL YC936
CORR YC900C9
STAFF LOUNGE YC933
PCA YC975
VITALS YC974
S. TLT. YC935
CORR YC900C15
Move to Specified Locations
EXAM YC957
EXAM YC955
EXAM YC953
EXAM YC951
EXAM YC949
ELEC YC947
EMER. ELEC. YC937
ELEV. LOBBY YC938
SERVICE ELEV. YC900E10
SERVICE ELEV. YC900E11
VITALS YC971
VITALS YC972
VITALS YC973
CONSULT YC969
OFFICE YC970
EXAM YC960
CORR YC900C12
CORR YC900C11
CORR YC900C14
S. TLT. YC968
EXAM YC958
EXAM YC956
EXAM/ PROC. YC954
EXAM YC952
EXAM YC948
SOILED YC946
RECY. YC939
P. TLT. YC961
CORR YC900C10
CONSULT YC967
EXAM YC964
EXAM YC962
STRETCH/ WHCR YC959
PROV. WK. YC950
VEST. YC900V2
STAIR 2 YC900S2
CORR YC900C13
PROV. WK. YC950A
READING YC945
PROV. WK. YC950B
CLEAN YC943
CONF YC942
CONSULT YC966
OFFICE YC965
OFFICE YC963



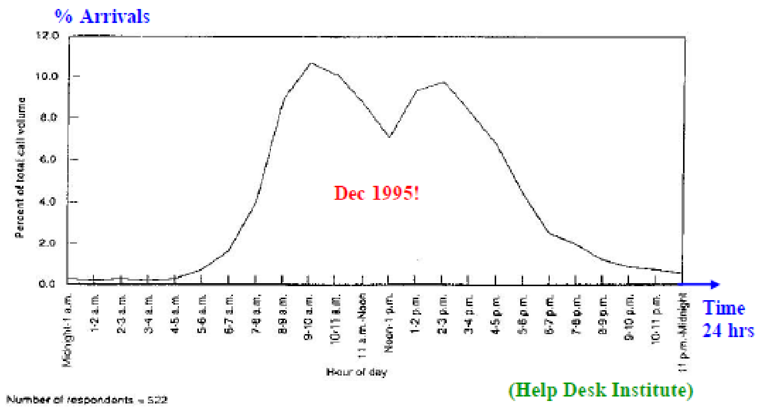
C#-#
EI
Y9-089
366
INF
YC905C
C#-#
EI
Y9-088
365
H
C#-#
EI
Y9-087
364
INF
YC905D
INF
YC905
C#-#
EI
Y9-090
367
C#-#
EI
Y9-118
362
C#-#
EI
Y9-086
363
INF
YC905B
INF
YC905E
C#-#
EI
Y9-091
368
INF
YC905A

Mining Model-Primitives / Building-Blocks

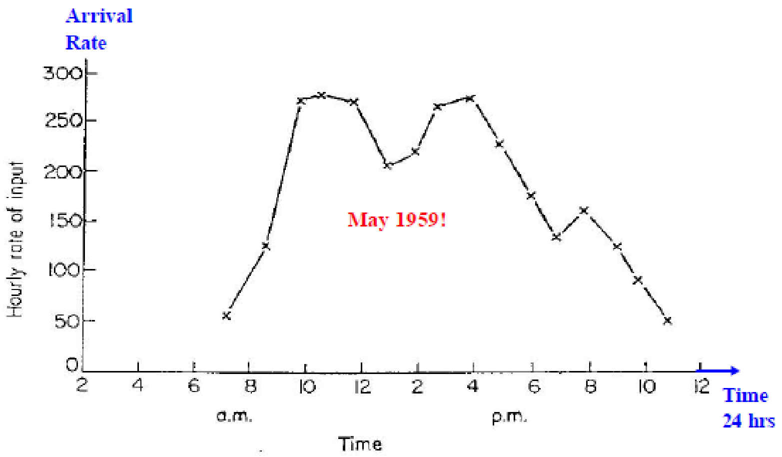
Arrivals to Service

Arrival-Rates to Three Call Centers

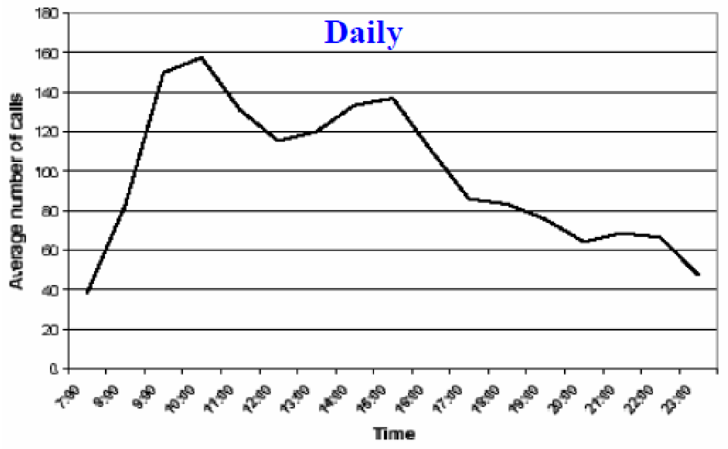
Dec. **1995** (U.S. 700 Helpdesks)



May **1959** (England)



November **1999** (Israel)

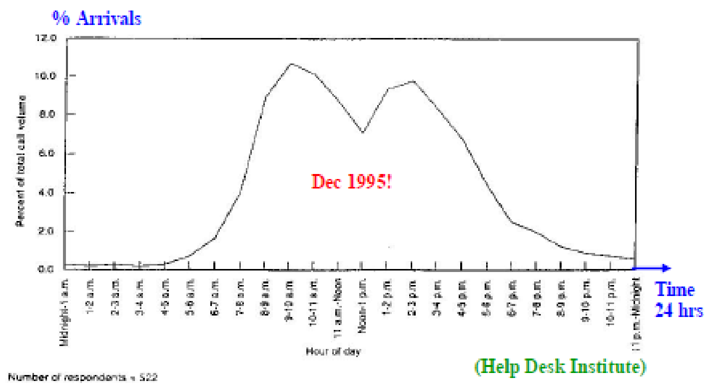


Random Arrivals "must be"
(Axiomatically)
Time-Inhomogeneous Poisson

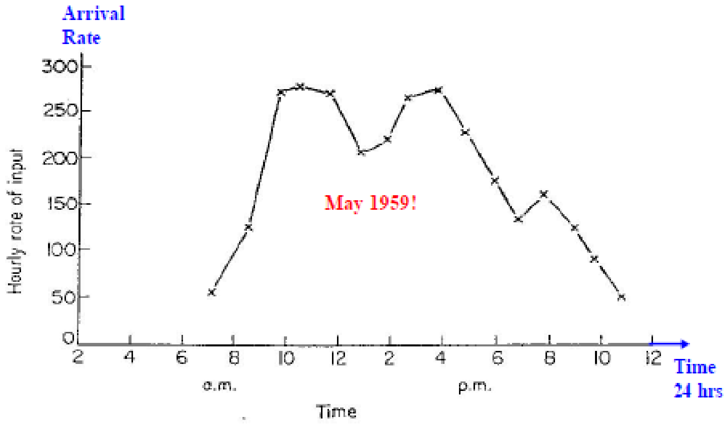
Primitives: Arrival (Rates) to Service

Why 2 Daily Peaks?

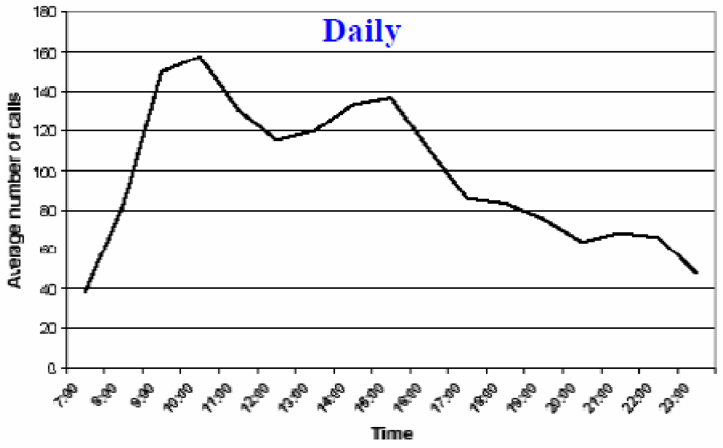
CC: Dec. **1995**, (USA, 700 Helpdesks)



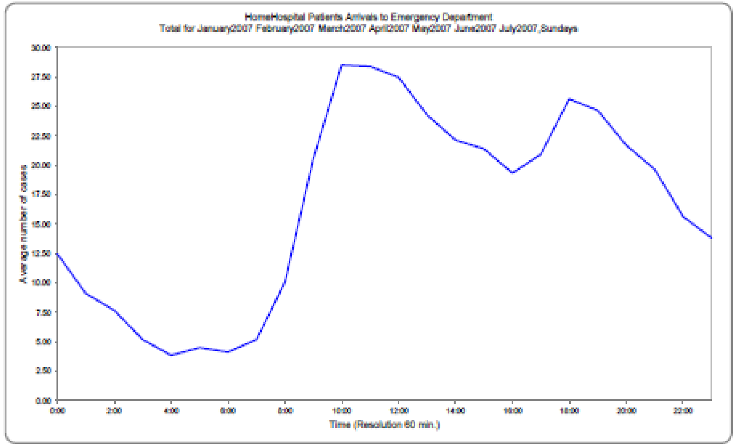
CC: May **1959** (England)



CC: Nov. 1999 (Israel)

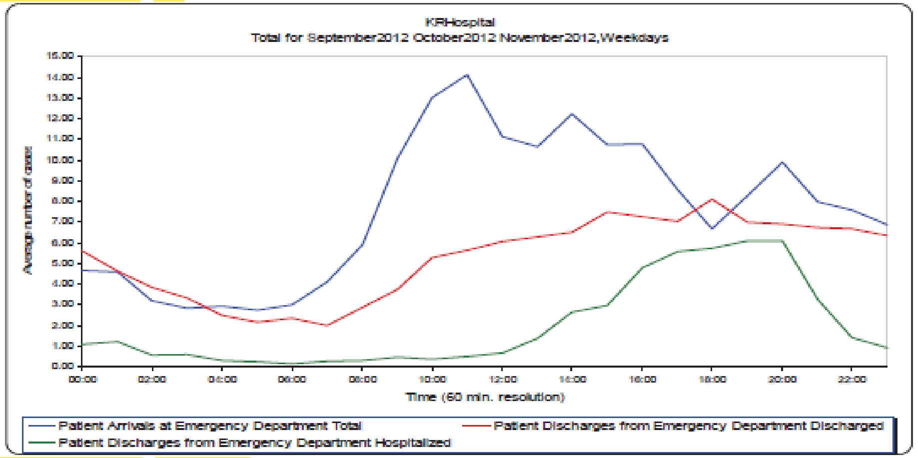


ED: Jan.–July 2007 (Israel)

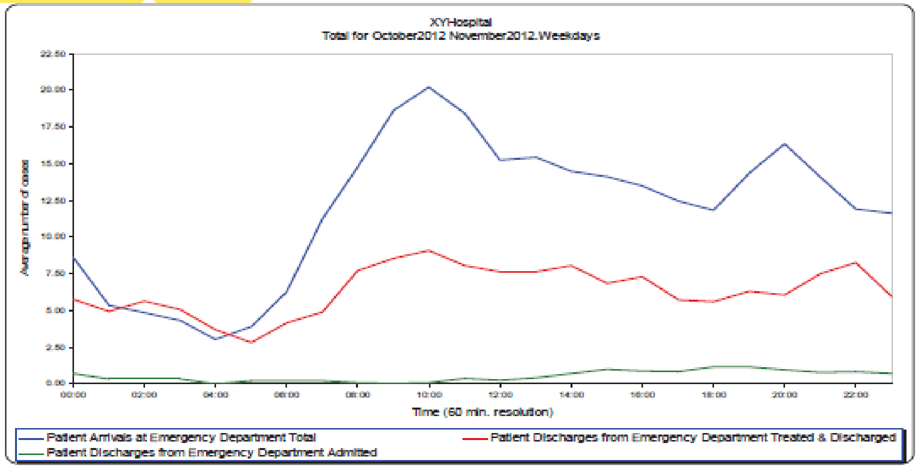


Arrival (Discharge) Rates in Korea and Singapore

KRHospital, all ED patients



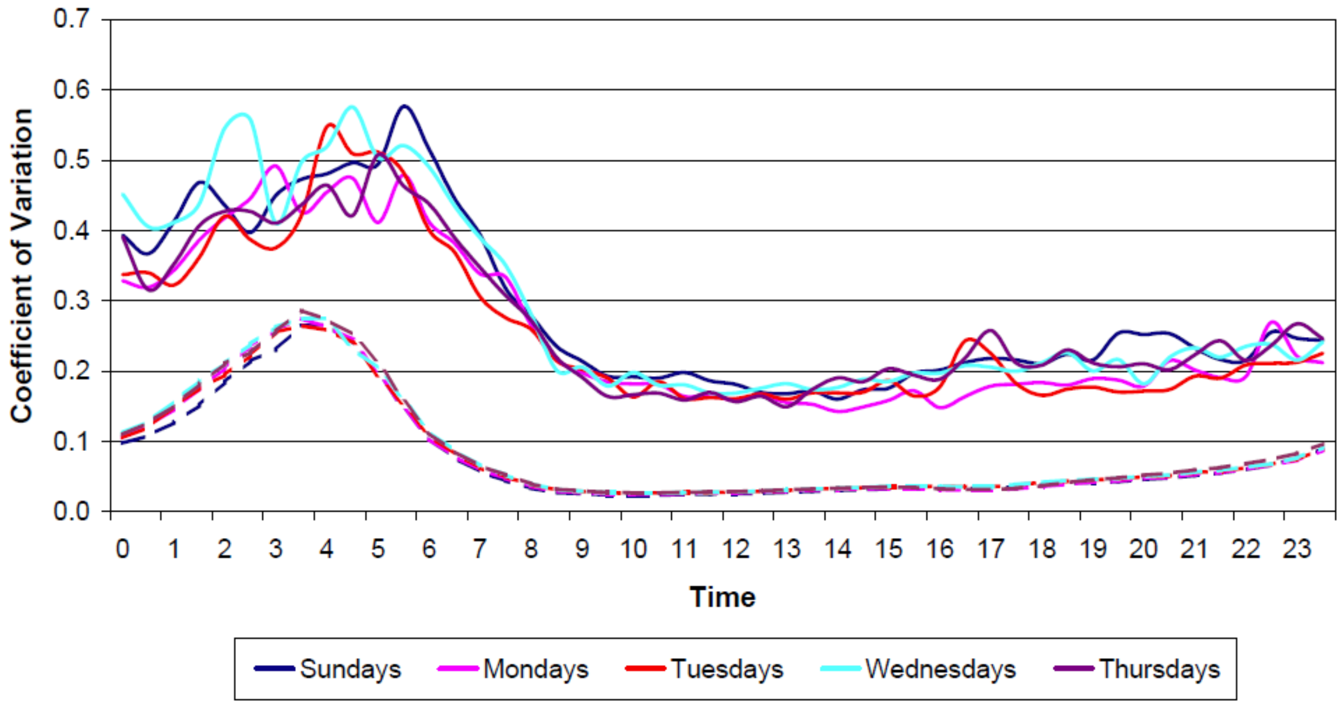
XYHospital, walking patients



Arrivals to Service: only Poisson-Relatives

Arrival-Counts: Coefficient-of-Variation (CV), per 30 min.

Israeli-Bank Call-Center, 263 regular days (4/2007 - 3/2008)

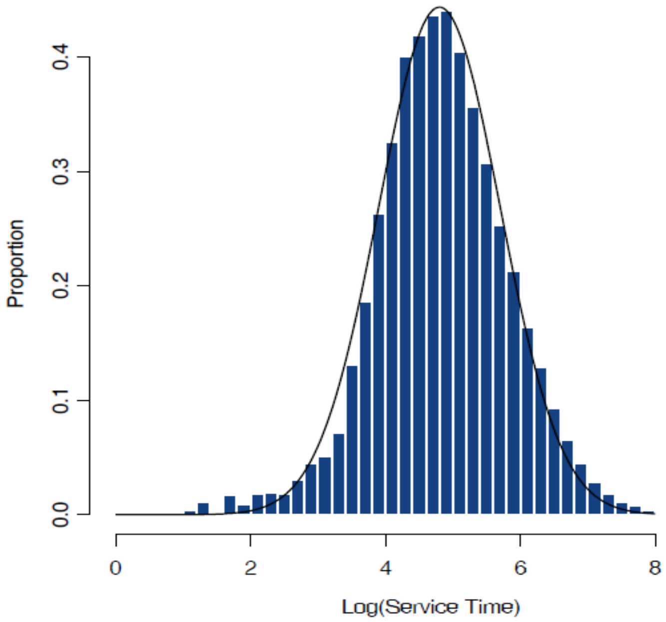


- **Poisson CV** (Dashed Line) = $1/\sqrt{\text{mean arrival-rate}}$
- Poisson CV's $\ll$ **Sampled CV's** (Solid) $\Rightarrow$ **Over-Dispersion**

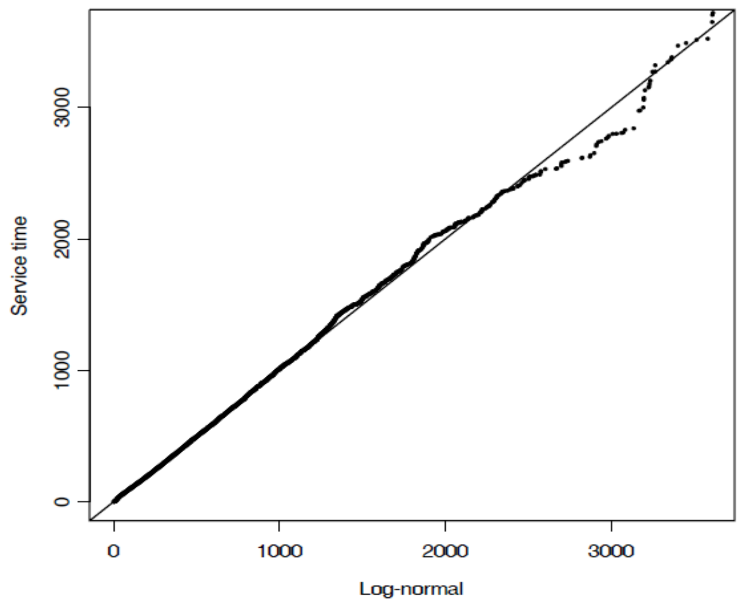
Building-Blocks: Service-Durations
e.g. Phone-Calls often LonNormal

Israeli Call Center, Nov–Dec, 1999

Log(Service Times)



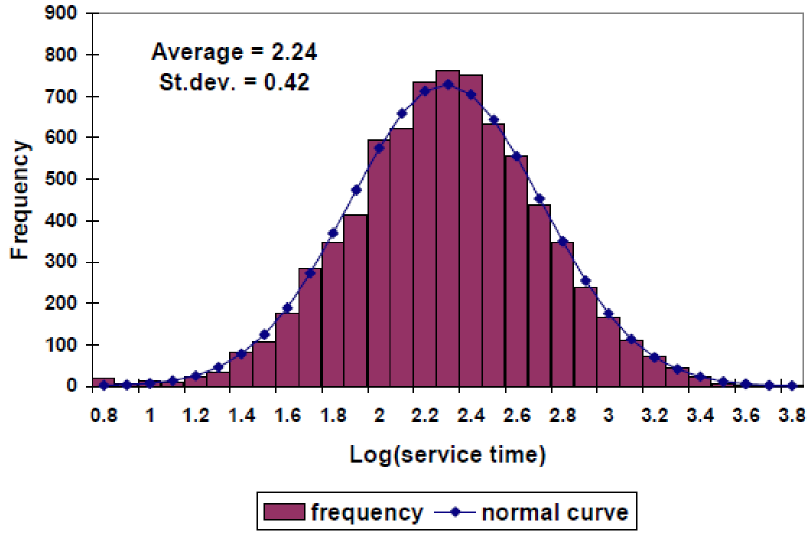
LogNormal QQPlot



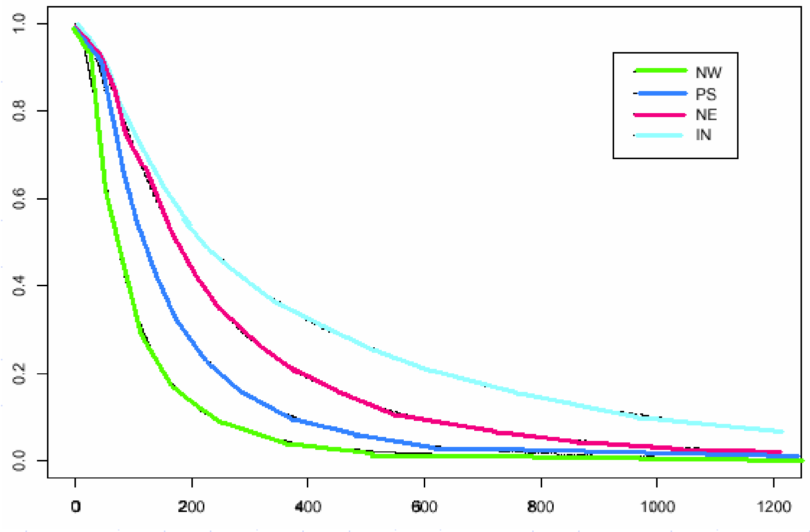
- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)

Service Durations: LogNormal Prevalent

Israeli Bank Log-Histogram



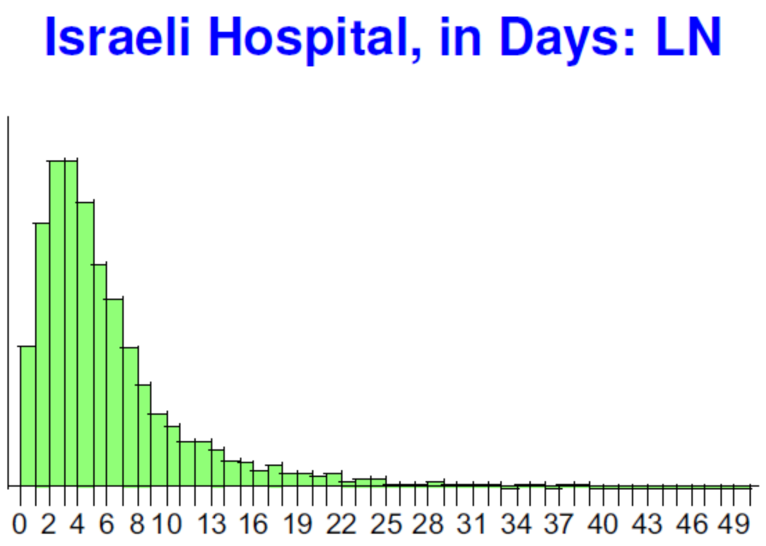
Service-Classes Survival-Functions



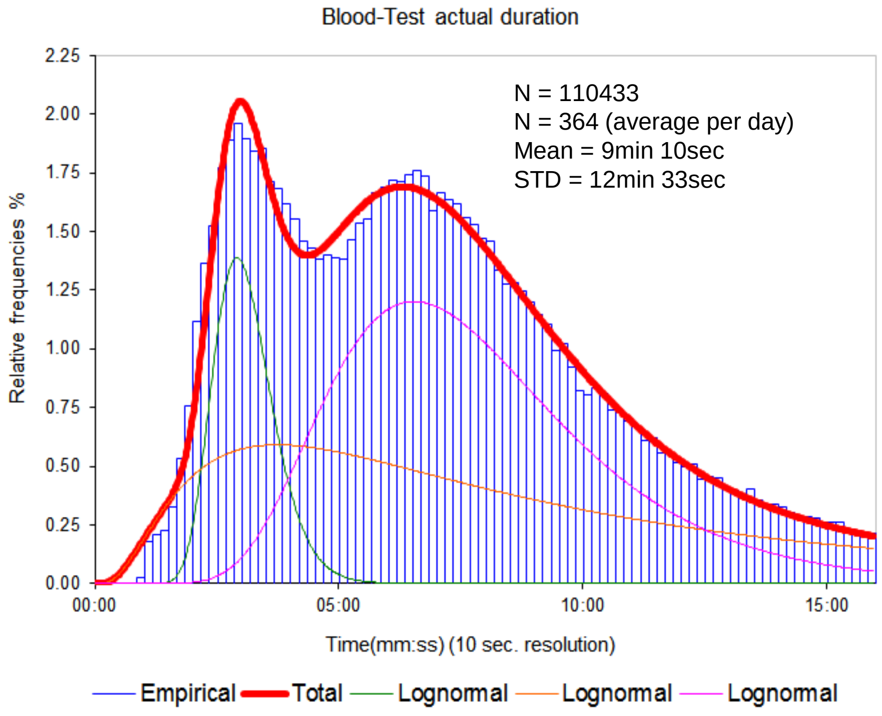
- **New** Customers: **2** min (NW);
- **Regulars**: **3** min (PS);
- **Stock**: **4.5** min (NE);
- Tech-Support: **6.5** min (IN).

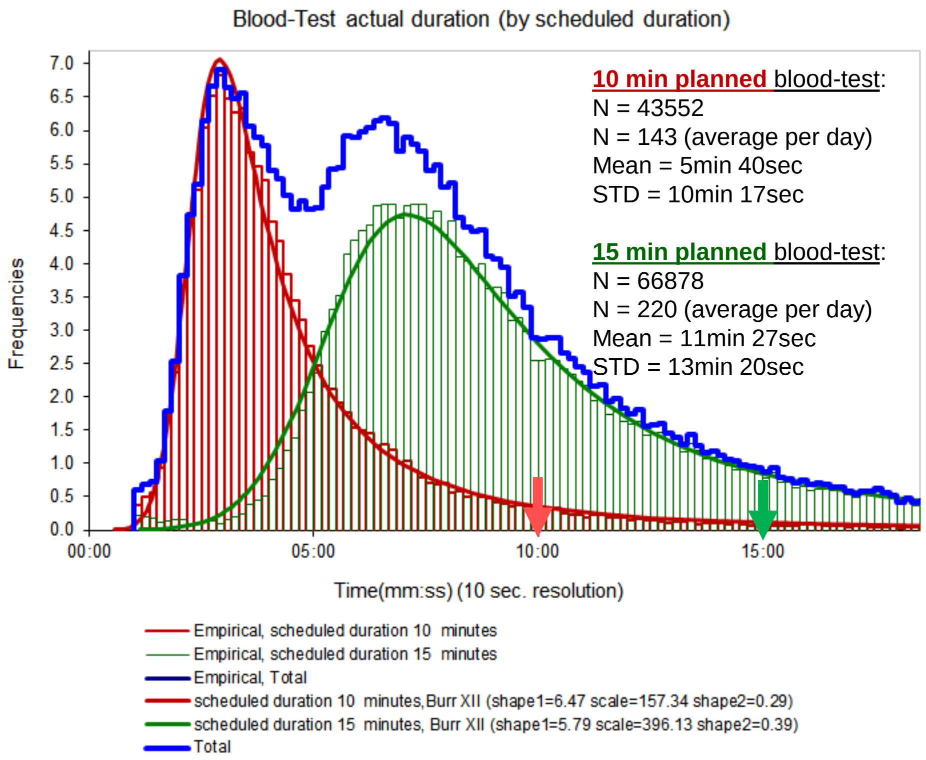
▶ Service Durations are **LogNormal (LN)** and **Heterogeneous**

Building-Blocks: Length-of-Stay in a Hospital Ward



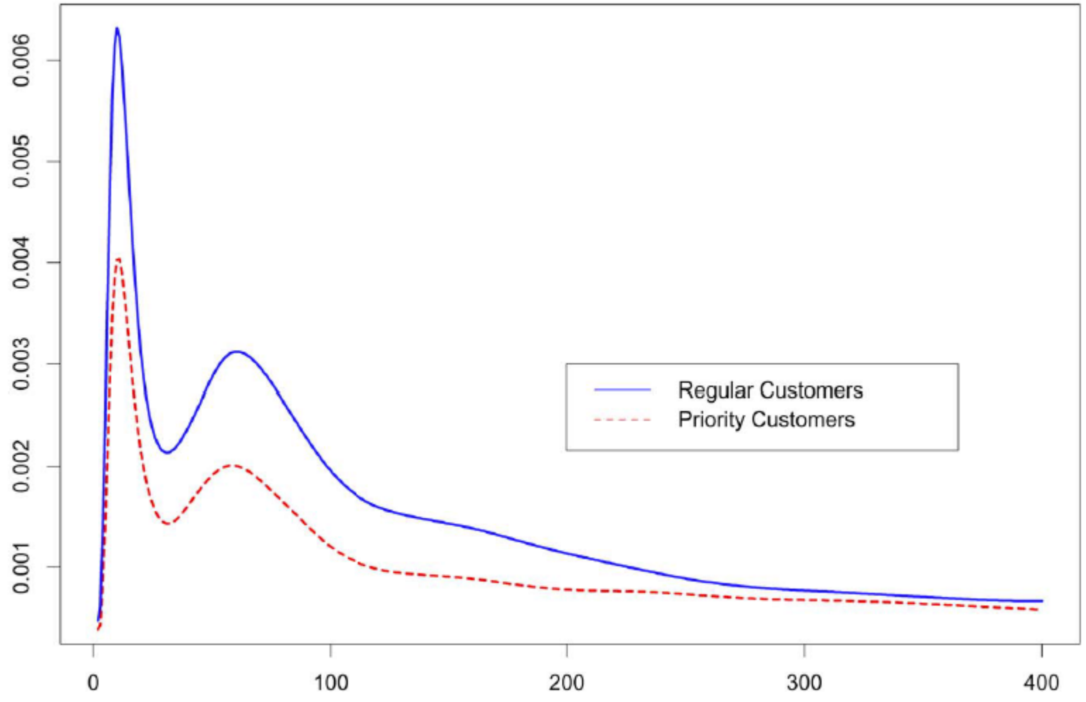
Patients treatment time: Blood-Test - Mixture Fitting and Real Components





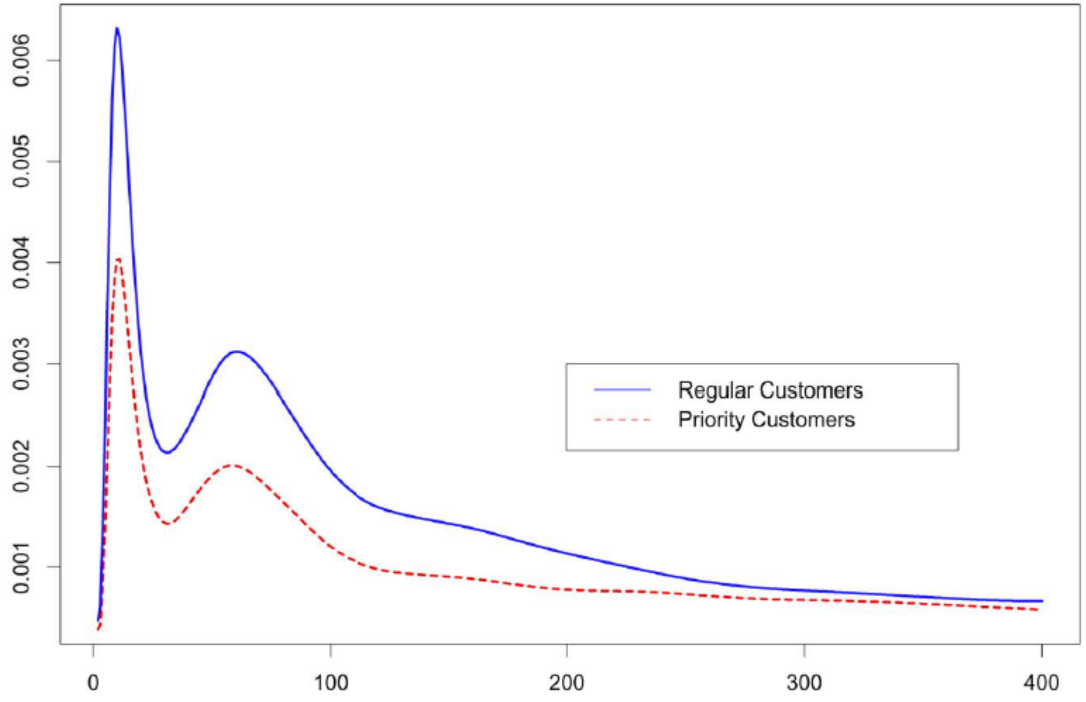
(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation
Regular over VIP Customers – Israeli Bank



(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation
Regular over VIP Customers – Israeli Bank

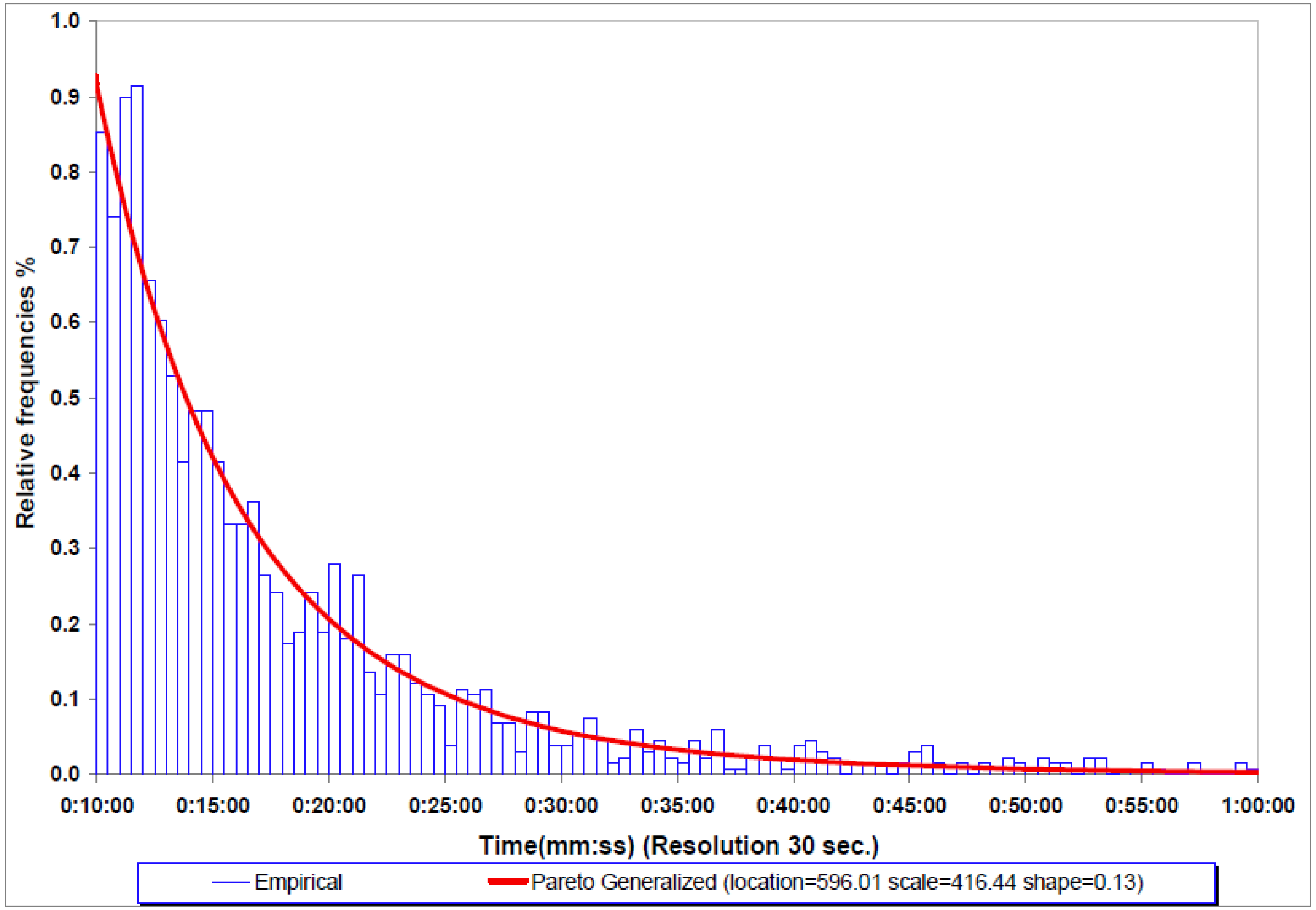


- **VIP** Customers are **more Patient** (Needy)
- **Peaks** of abandonment at times of **Announcements**
- Challenges: **Un-Censoring, Dependence (vs. KP), Smoothing** - requires **Call-by-Call Data**

Primitives: (Im)Patience

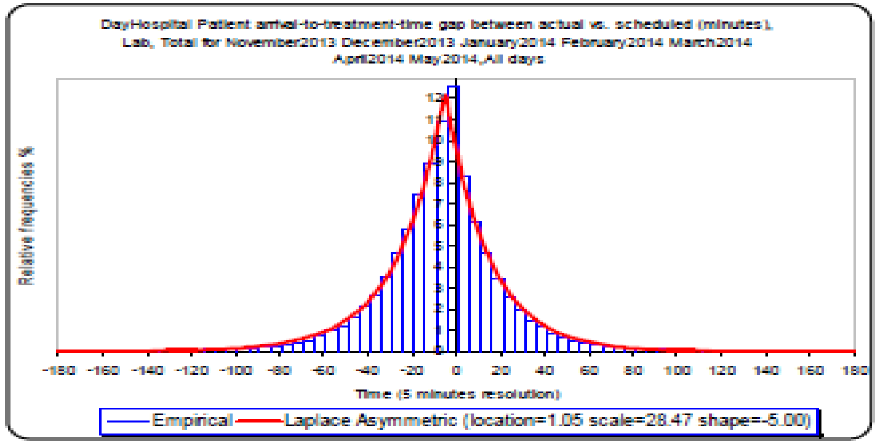
Israeli Bank: Uncensored 13,000 Customers, 24/11/2008

Patience $\geq 10min$: Why Pareto Tail?

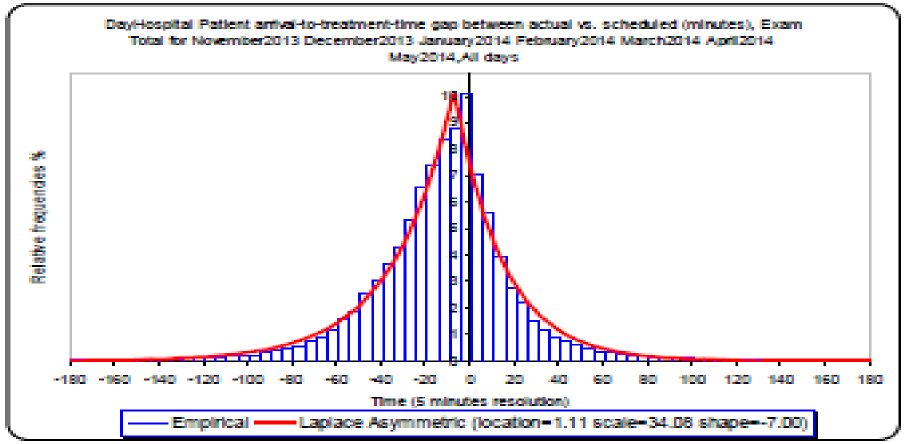


Primitives: Punctuality
Planned vs. Actual Arrival to Service
@ Stations 1, 2, 3 in a Hospital

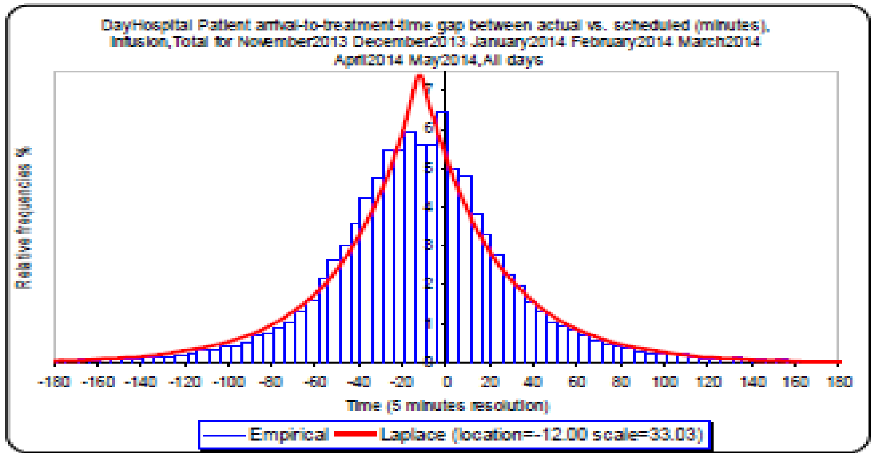
Lab (mean = -7 min)



Exam (mean = -12 min)



Infusion (mean = -12 min)



Technion S.E.E Center Service Enterprise Engineering

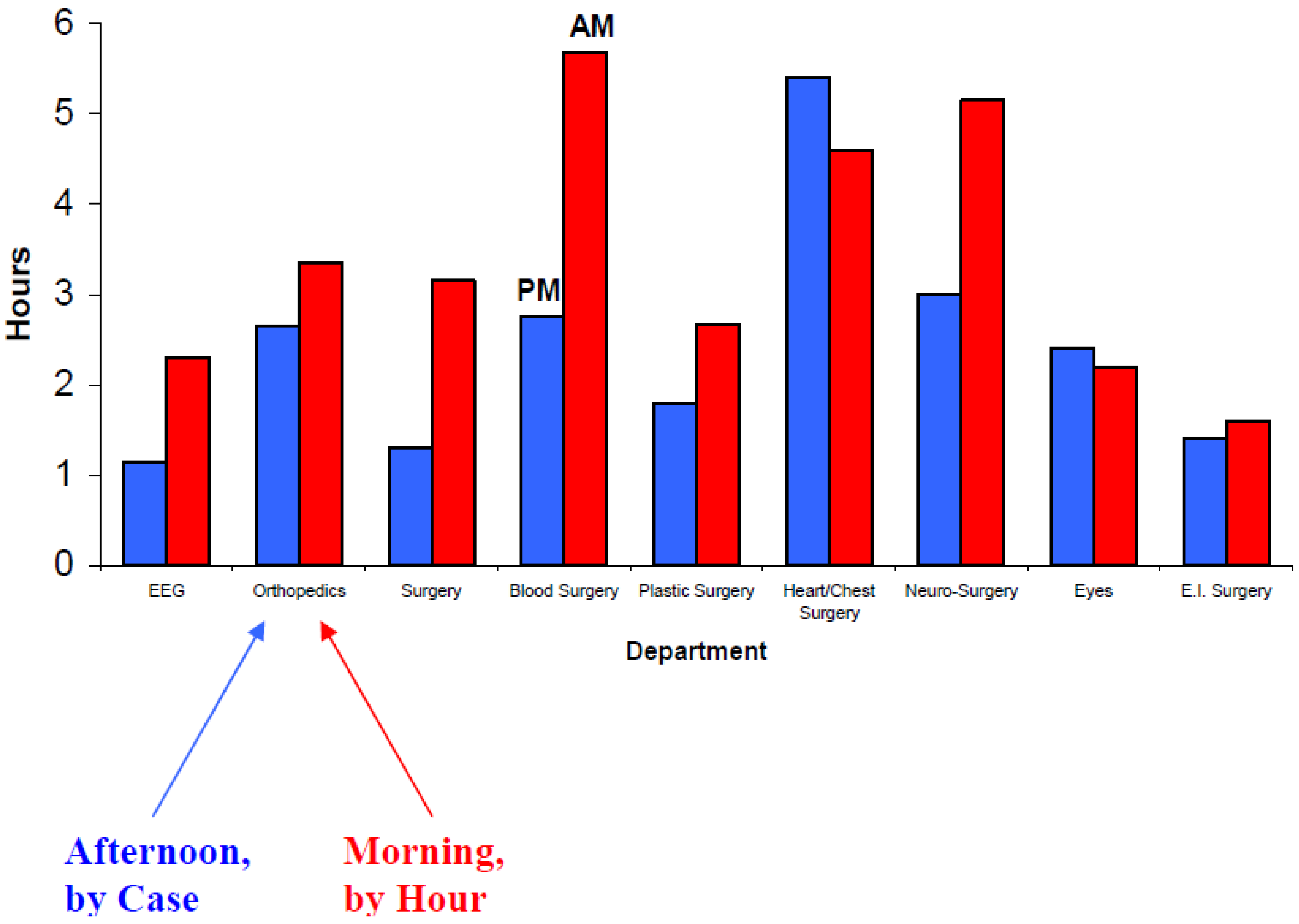
Mining Service Protocols - Examples (Behavioral OR)

- **Incentive-driven** protocols (even averages... , even doctors, …)
- FCFS (data = heavy-traffic theory)
- Customers: Priorities while waiting
- Servers: “Priorities” while serving
- Management: Discharge from Hospital
- **Data not enough (DFCI Pharmacy: FCFS default, random in peak)**

Research: First steps (with Senderovich; Liberman & Meilijson in TAU)

Interesting Averages: The Human Factor, or Even "Doctors" Can Manage

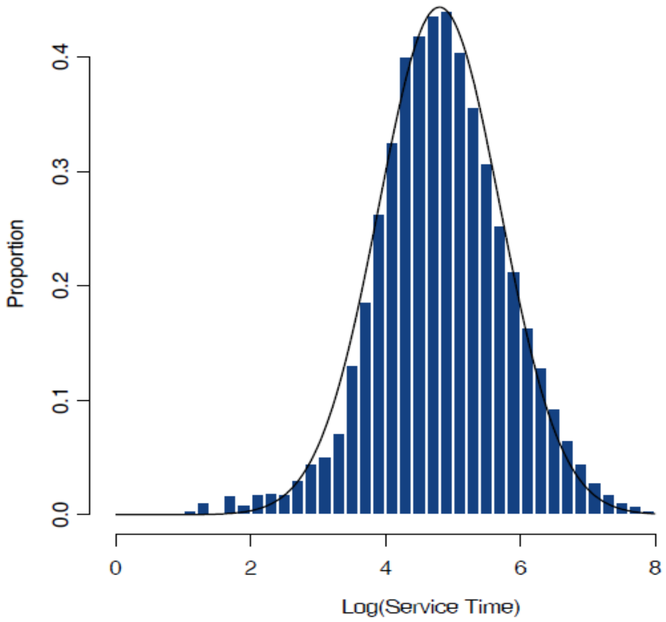
Operations Time - Morning (by Hour) vs. Afternoon (by Case):



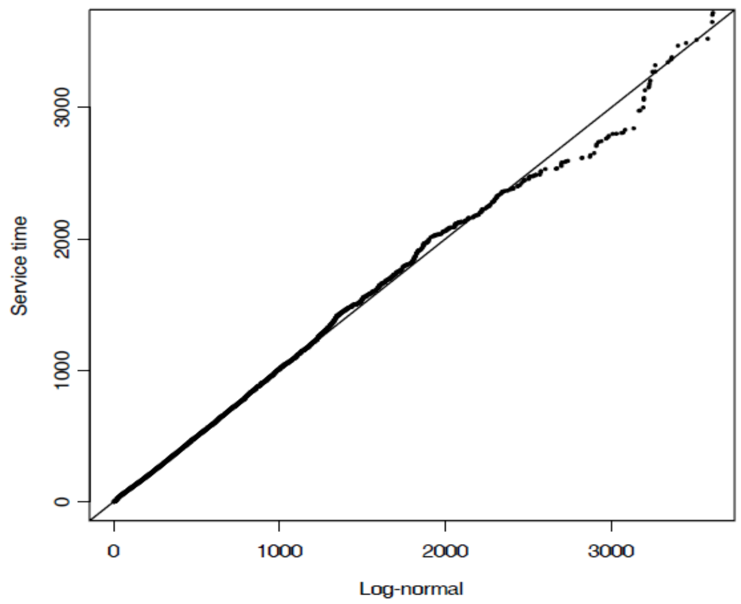
Durations: Phone Calls (2nd Surprise)

Israeli Call Center, Nov–Dec, 1999

Log(Service Times)



LogNormal QQPlot



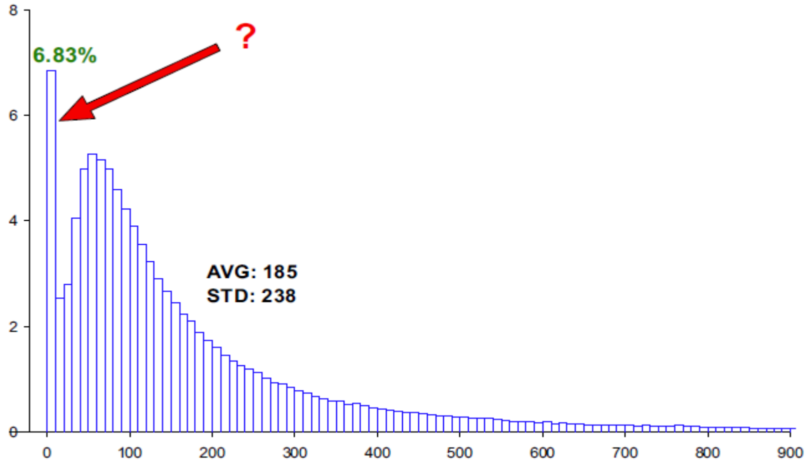
- **Practically Important**: (mean, std)(log) characterization
- **Theoretically Intriguing**: Why LogNormal ? Naturally multiplicative but, in fact, also **Infinitely-Divisible** (Generalized Gamma-Convolutions)

Building-Blocks: Service-Duration Histograms

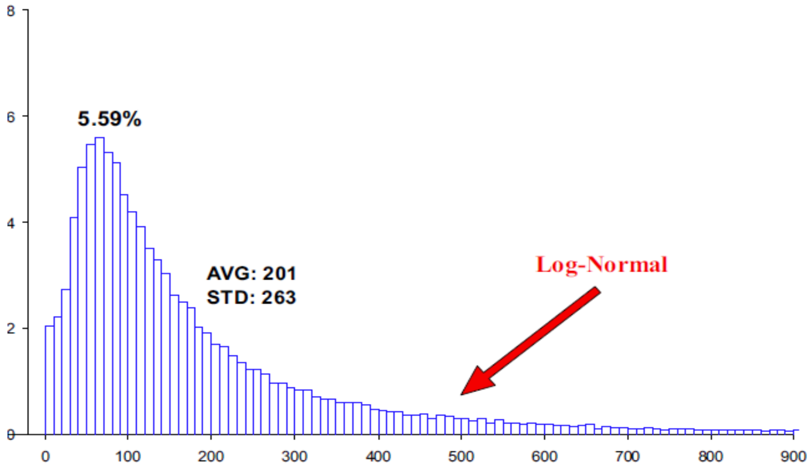
Histogram of Service-Duration in an Israeli Call Center, 1999

Why short services? Why LogNormal?

January-October

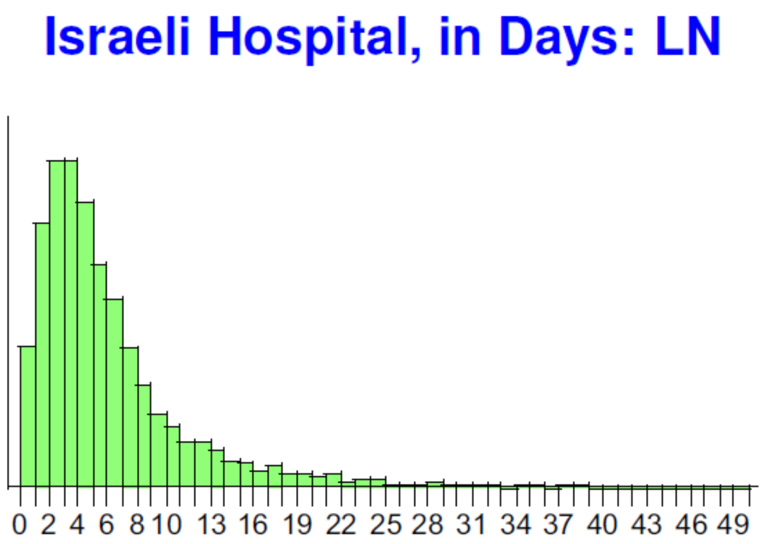


November-December



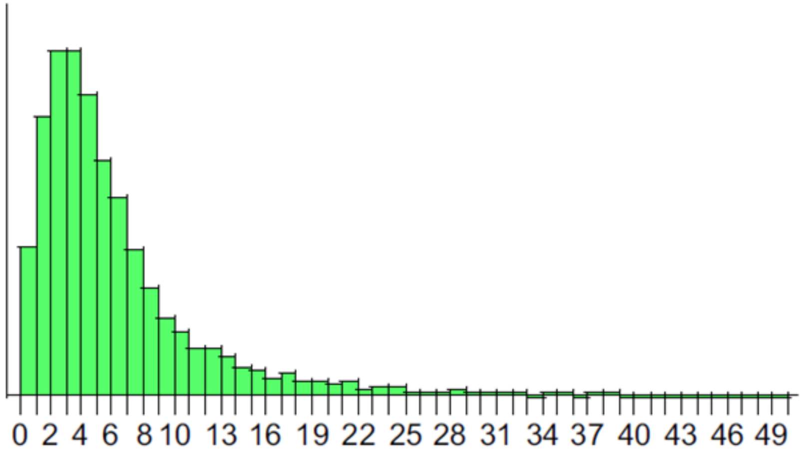
- January-October: **6.8% Short-Services** ($\leq$ 10 seconds) ?
- November-December: **LogNormal** durations (common) ?

Building-Blocks: Length-of-Stay in a Hospital Ward

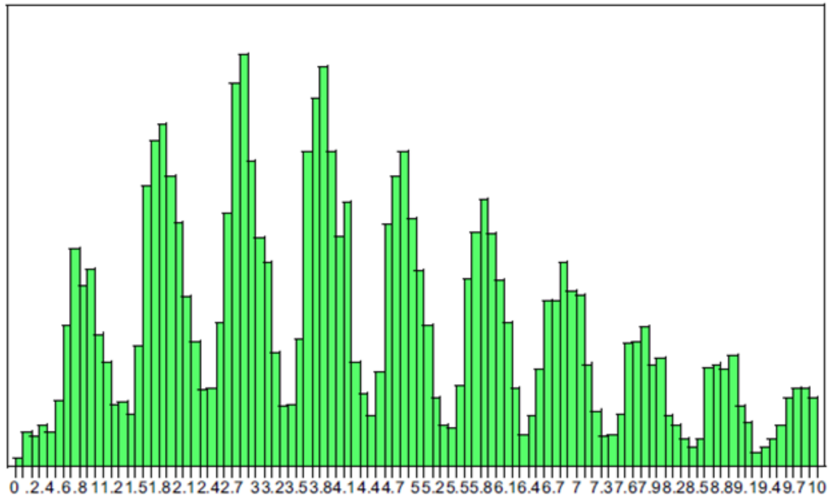


Protocols: LOS in Hospitals - Beyond LogNormal

Israeli Hospital, in Days: LN

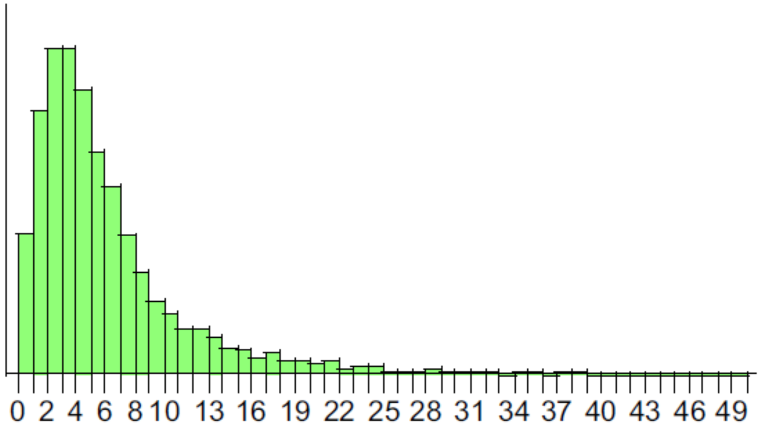


In Hours: 2 Time Scales, Mixture

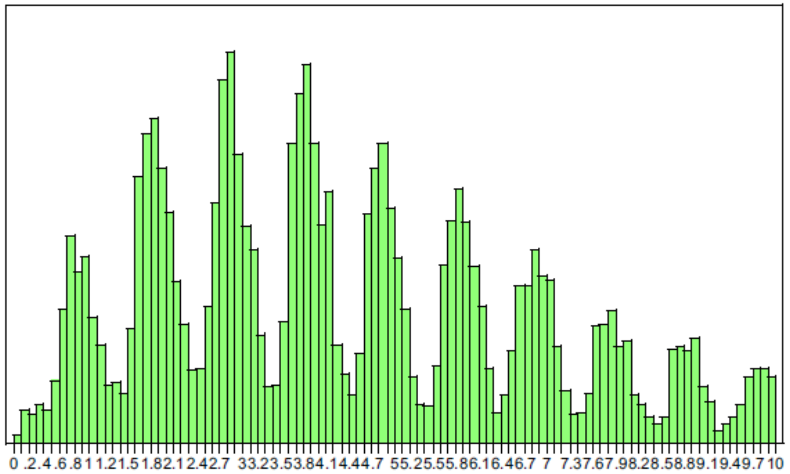


Length-of-Stay in a Hospital: Story in 2 Time-Scales

Israeli Hospital, in Days: LN

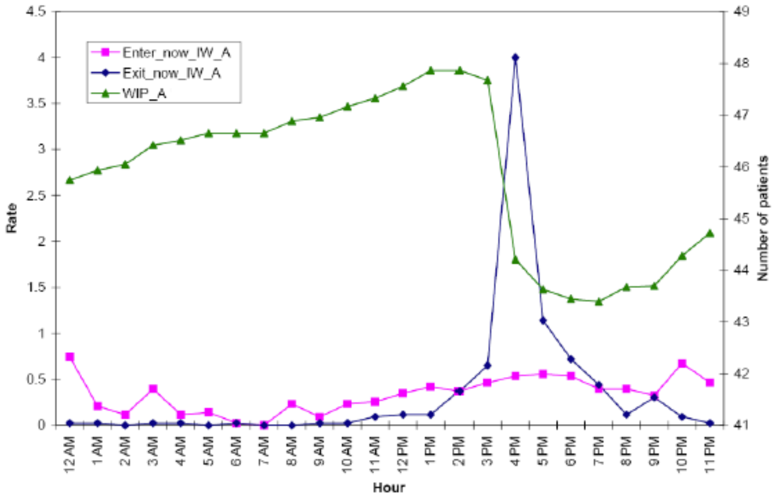


Israeli Hospital, in Hours



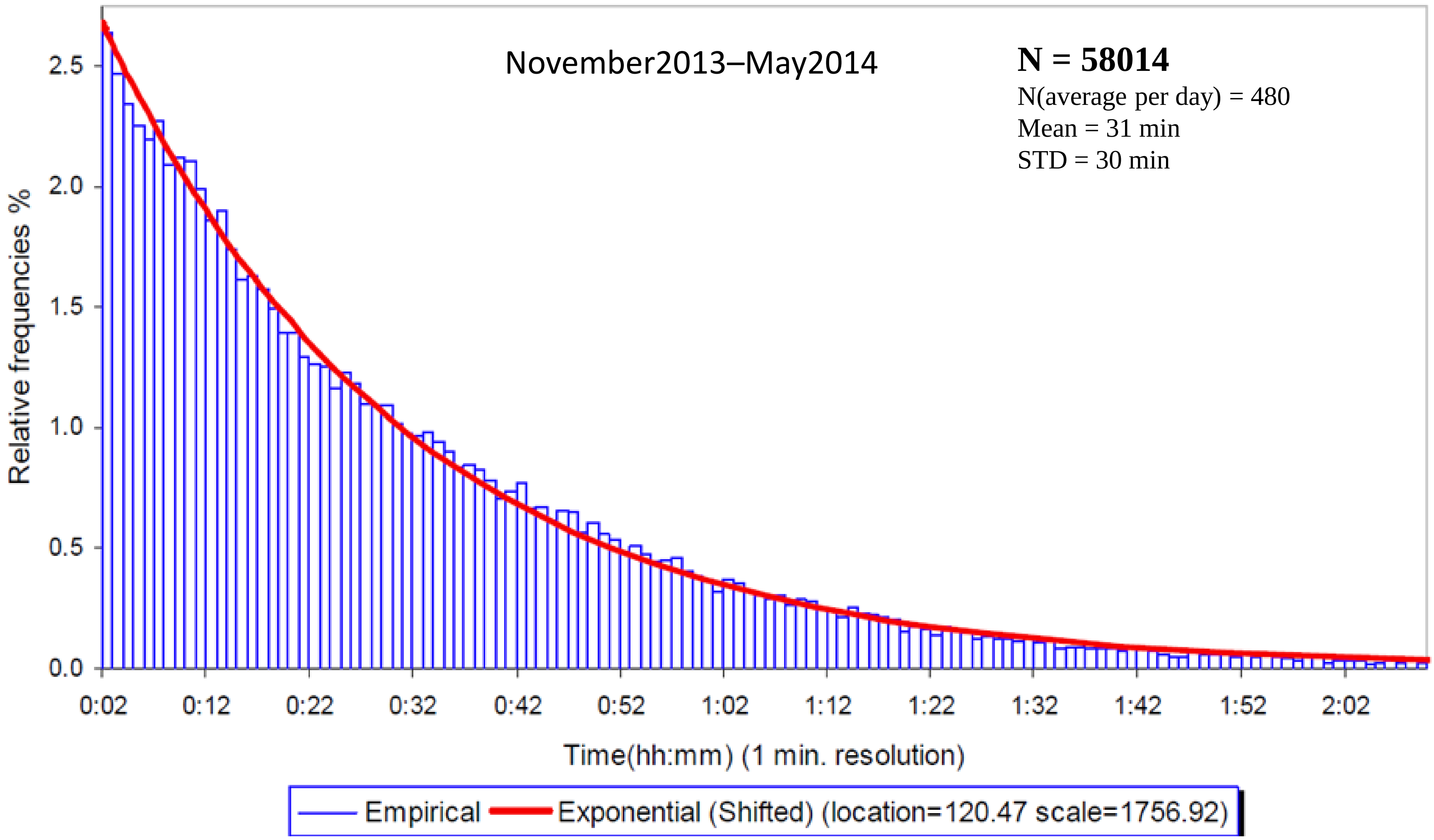
"Explanation": Patients released around **3pm** (1pm in Singapore)

Why Bother ?
Staffing, Bed Management, . . .

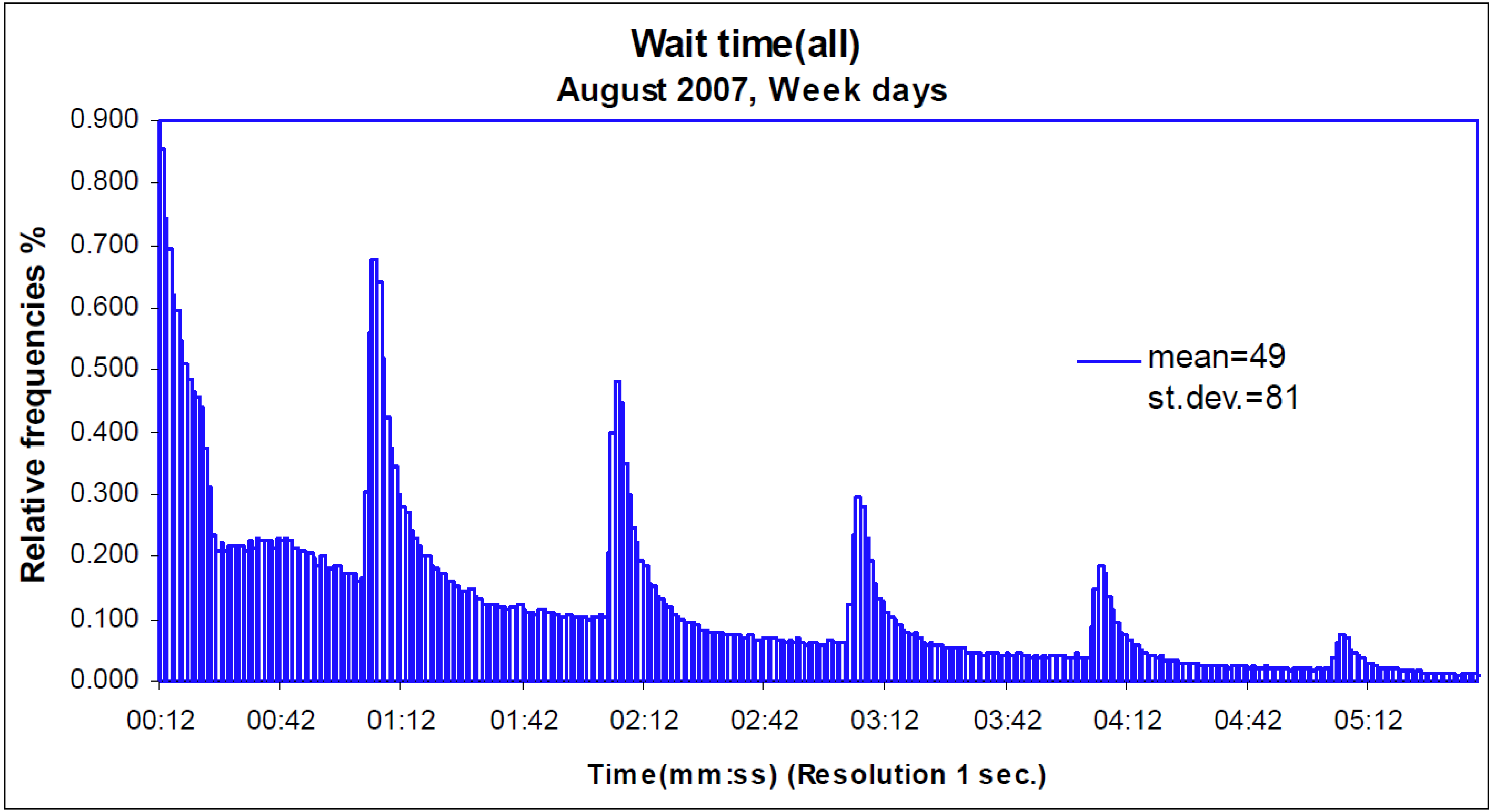


Protocols: via Waiting-Time for Physician-Exam
Theory of FCFS Single-Server Queue, in Heavy-Traffic: must be Exponential

(Kingman's Invariance theorem, 60's, under 2nd moments)

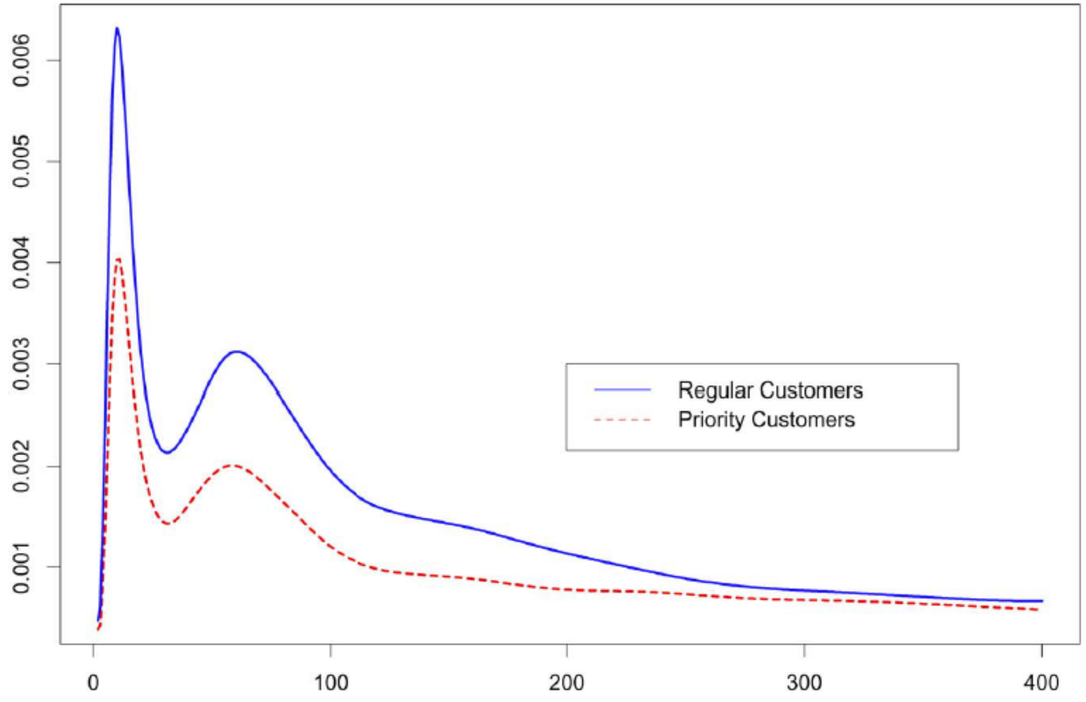


Protocols: via Waiting-Time for Phone Call
Histogram: Peaks every minute, due to Dynamic Priorities



(Im)Patience while Waiting (Palm 1943-53)

Hazard Rate of (Im)Patience Distribution $\propto$ Irritation
Regular over VIP Customers – Israeli Bank



Patience: Customers

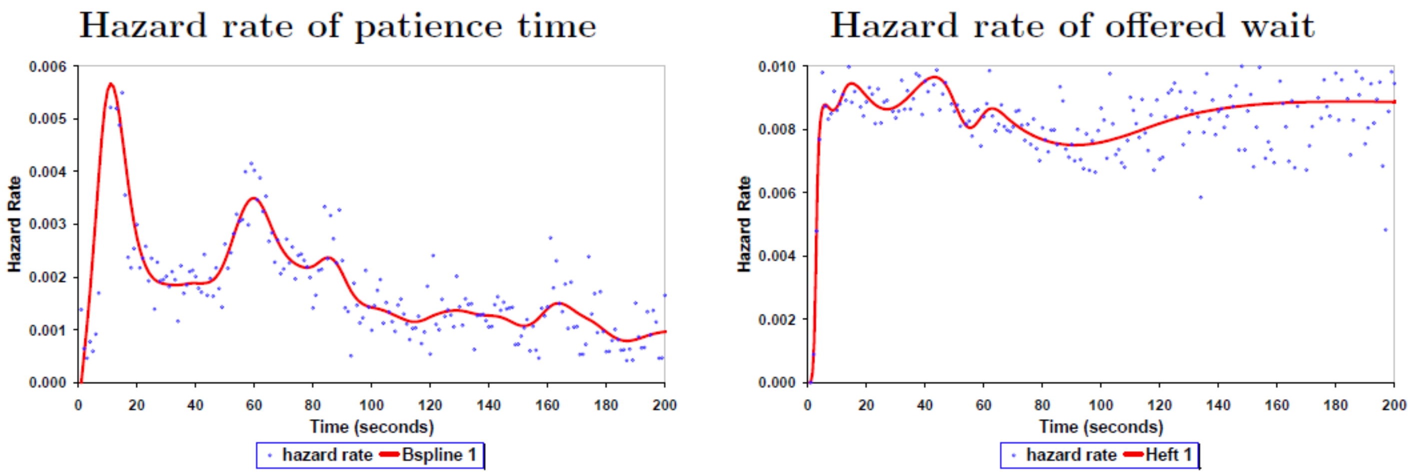


Figure 1: Patience and offered wait in an Israeli call center

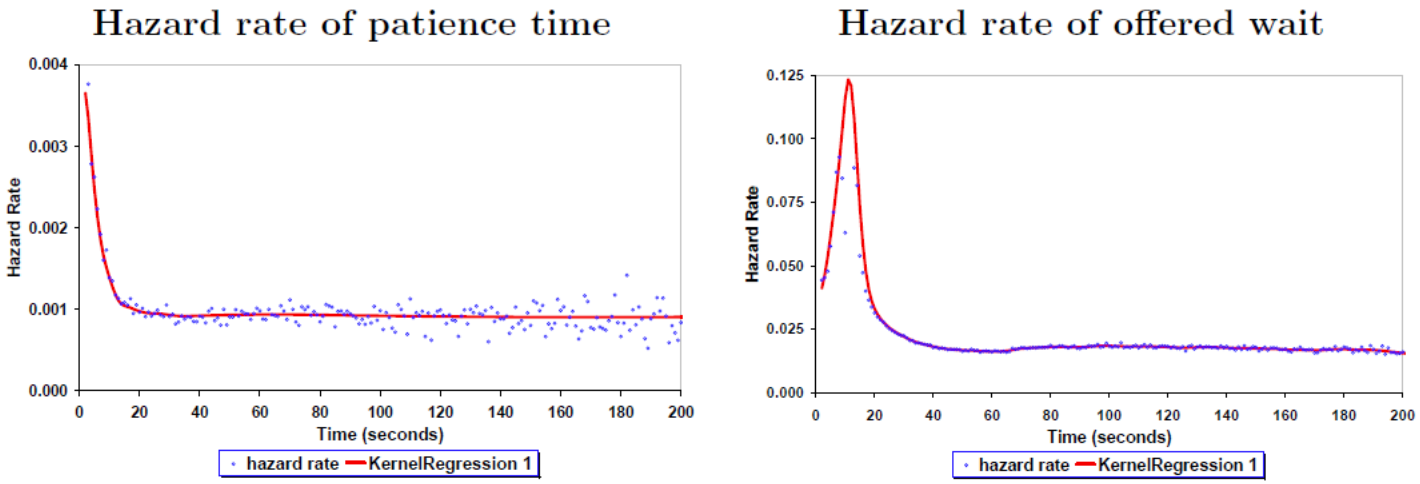


Figure 2: Patience and Offered Wait in a U.S. Call Center

Patience: Customers

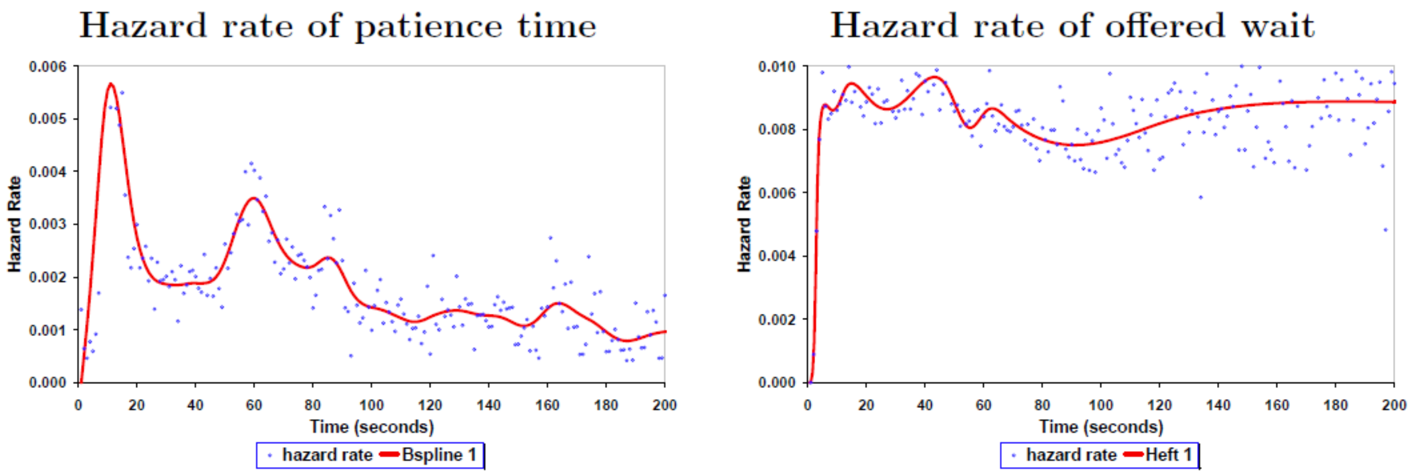


Figure 1: **Patience and offered wait in an Israeli call center**

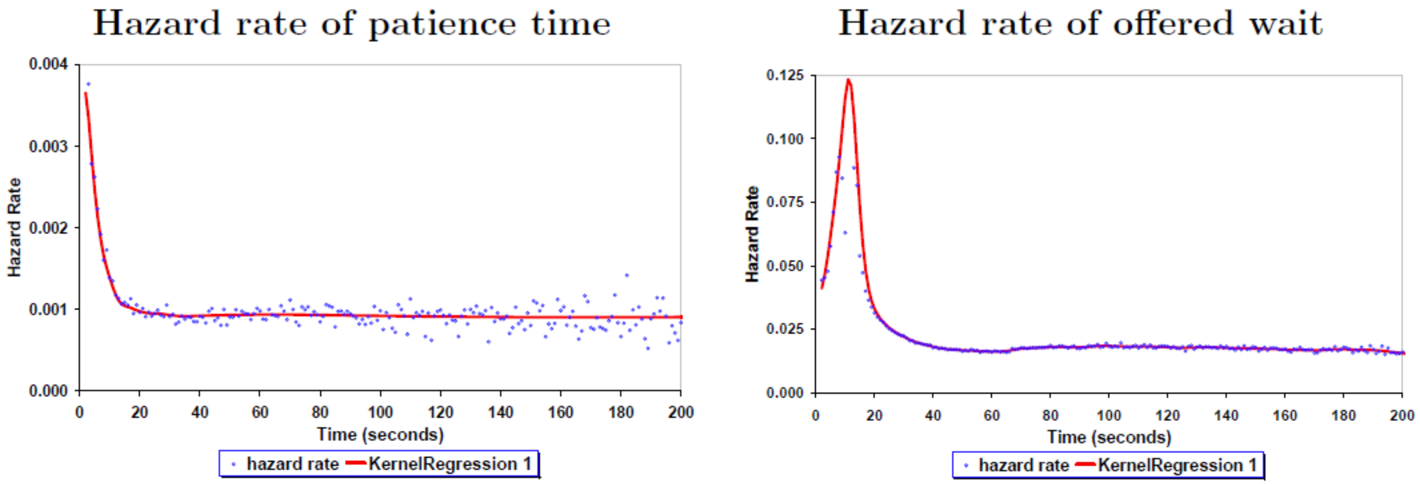
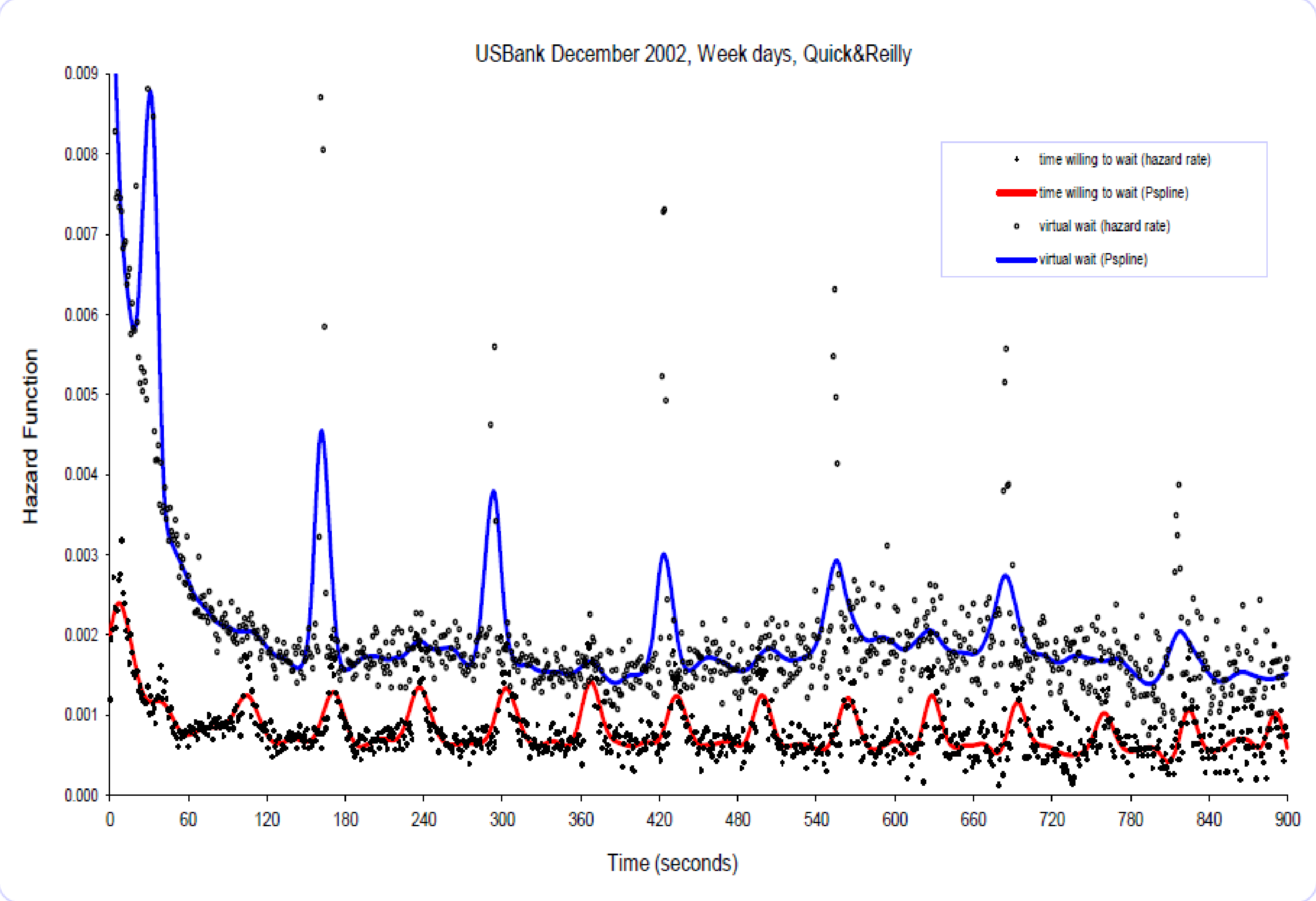


Figure 2: **Patience and Offered Wait in a U.S. Call Center**

Protocols + Psychology

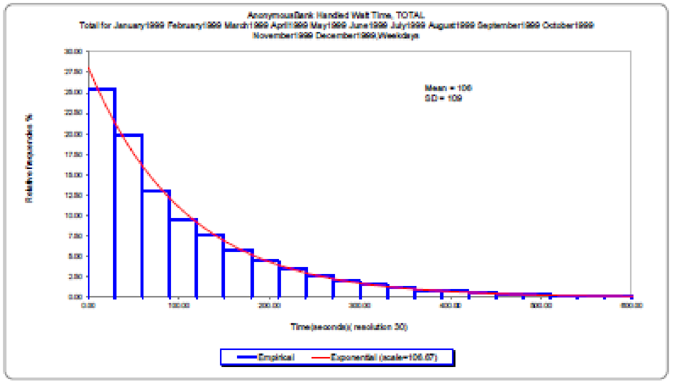
Patient Customers, Announcements, Priority Upgrades



Protocols: Waiting Time in a Call Center

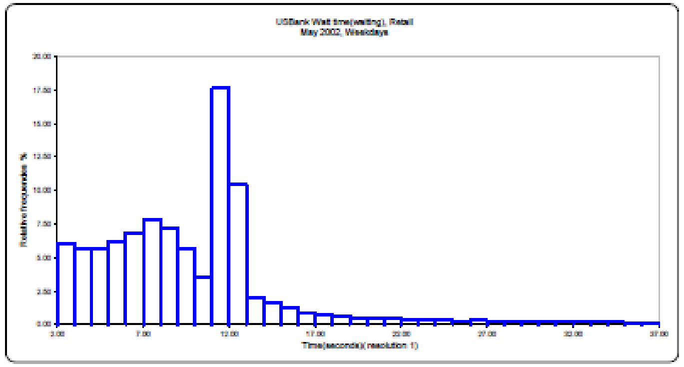
Exponential in Heavy-Traffic (min.)

Small Israeli Bank



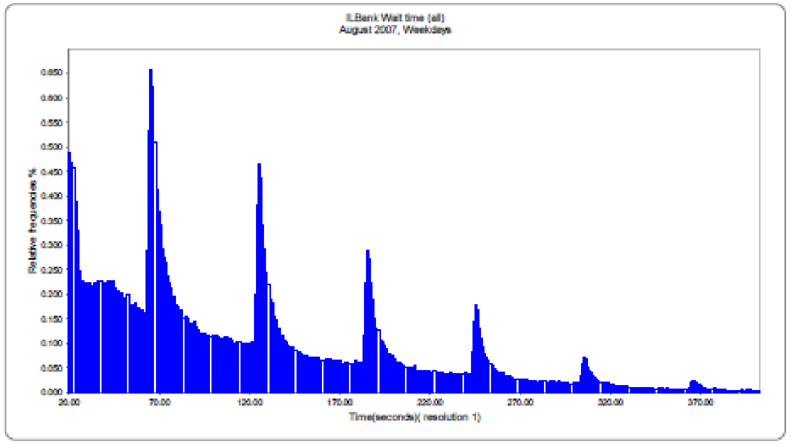
Routing via Thresholds (sec.)

Large U.S. Bank

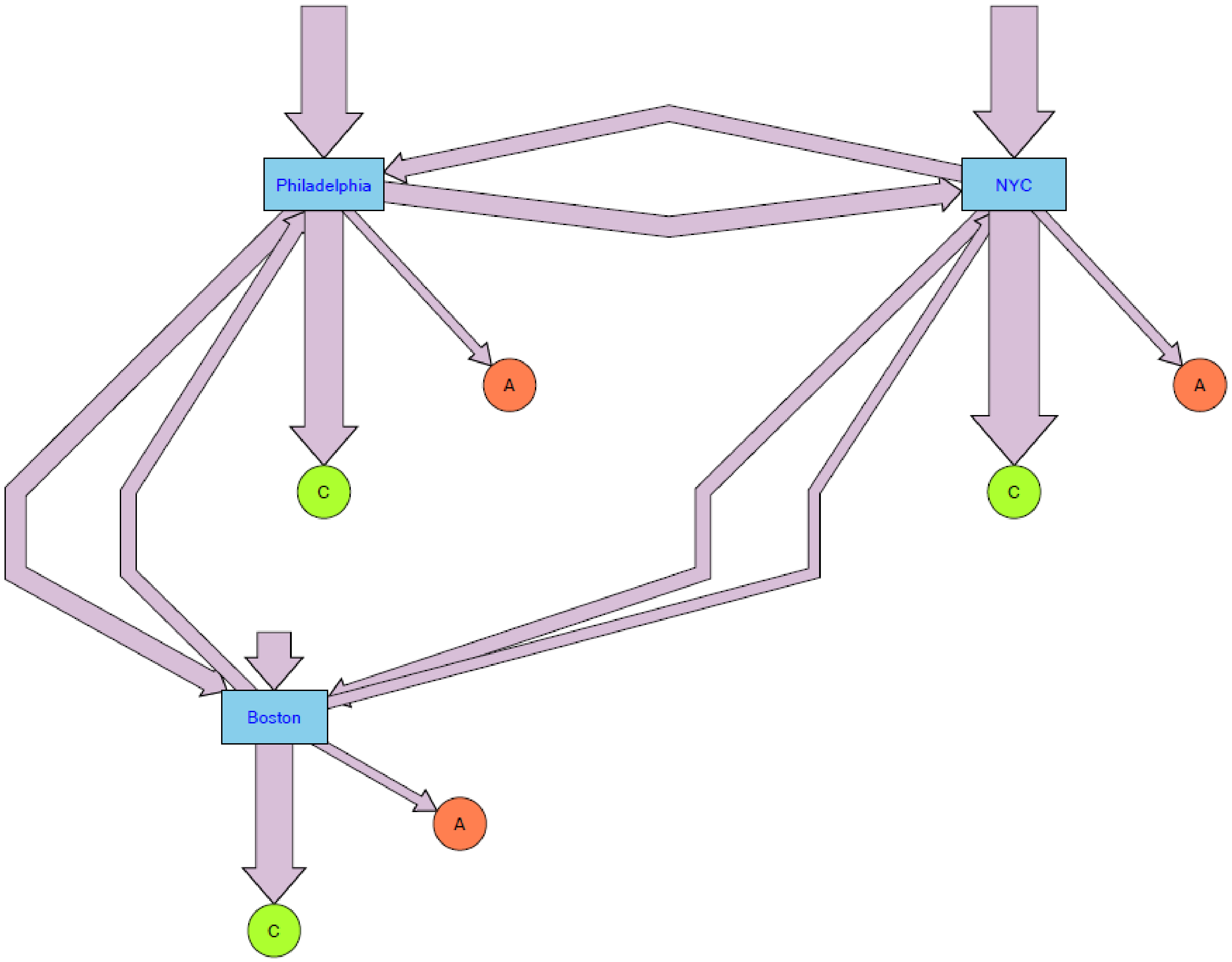


Scheduling Priorities (sec.) **[compare Hospital LOS (hours)]**

Medium Israeli Bank



Protocol Mining: Snapshots of Connectivity



Ultimate Research Goal

e.g. Specific Emergency-Department, with ample reliable data

- Real-time: **control of** patient-flow (bottlenecks); status **info** and **prediction**
- Short-term: on Monday, set Tuesday's **staffing** levels (or next week's); real **cost** of care for the individual patient (vs. mean/negotiated costs)
- Long-term: **capacity** allocation, facility/triage **design**; **social** network (e.g. correlated w/ outcomes); transformative **changes** (Epic); congestion **laws**

Research Goal (within reach)

e.g. Specific Emergency-Department, with ample reliable data

- Real-time: control of patient-flow (bottlenecks); status info and prediction
- Short-term: on Monday, set Tuesday’s staffing levels (or next week’s); real cost of care for the individual patient (vs. mean/negotiated costs)
- Long-term: capacity allocation, facility/triage design; social network (e.g. correlated w/ outcomes); transformative change (Epic); congestion laws

All above will be enabled by parsimonious (robust) models, created in real-time **mining** (semi- or un-supervised) of ED **processes** and **models** (empirical, simulation = **SimNets**, mathematical = **QNets**, **FNets**, **DNets**,…)

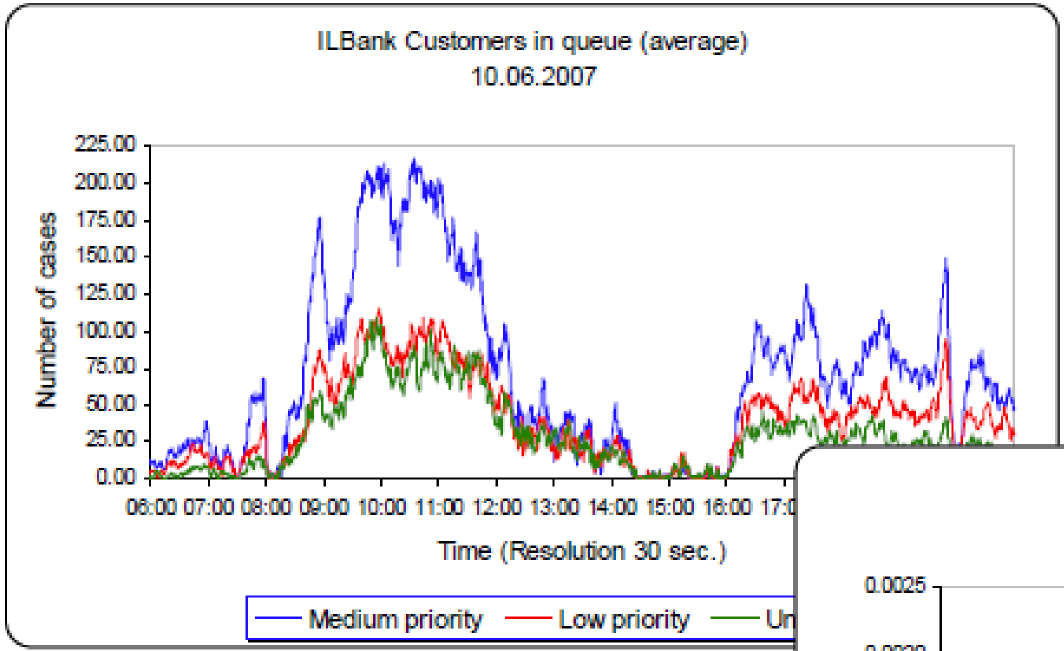
2 Prerequisites: Data, Models

- Data: The language of
  - ✓ Multi-disciplinary research (e.g. OR + Psychology + CS/DS)
  - ✓ Academia-Industry partnerships (e.g. university & bank, or hospital, or court)

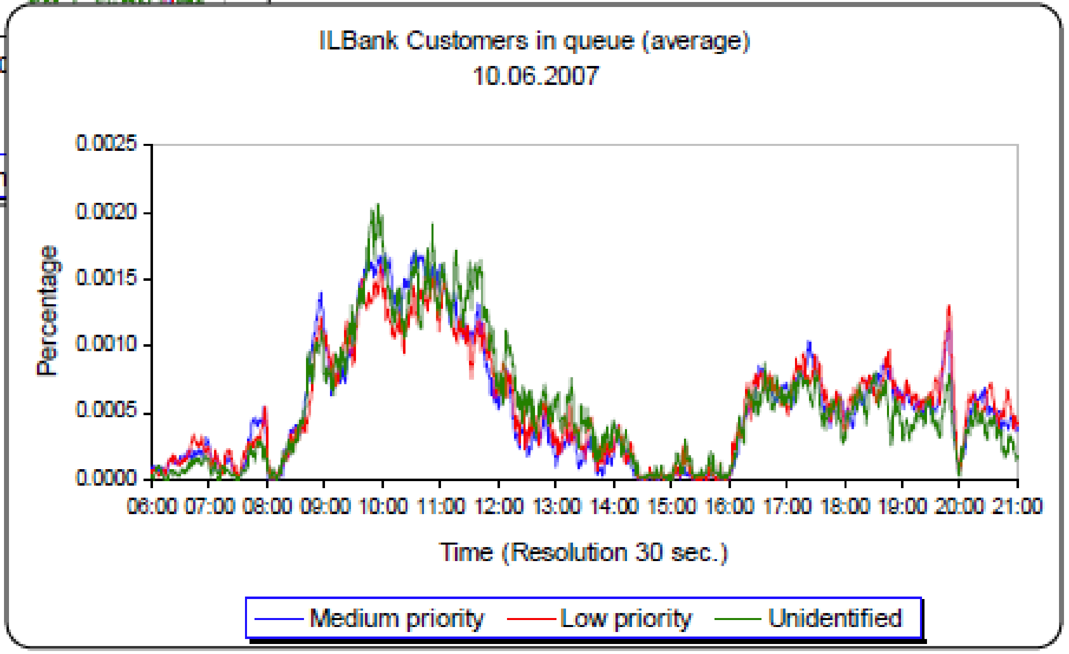
- Models: Simple models at the service of complex realities (not too simple)
  - ✓ Stochastic networks: Empirical, Analytical (QNets, FNets, DNets), SimNets
  - ✓ Insights often rooted in deep mathematics (even Little's Law)

Dynamics: Parsimonious Models (Congestion Laws)

3 Queue-Lengths at 30 sec. resolution (ILBank, 10/6/2007)



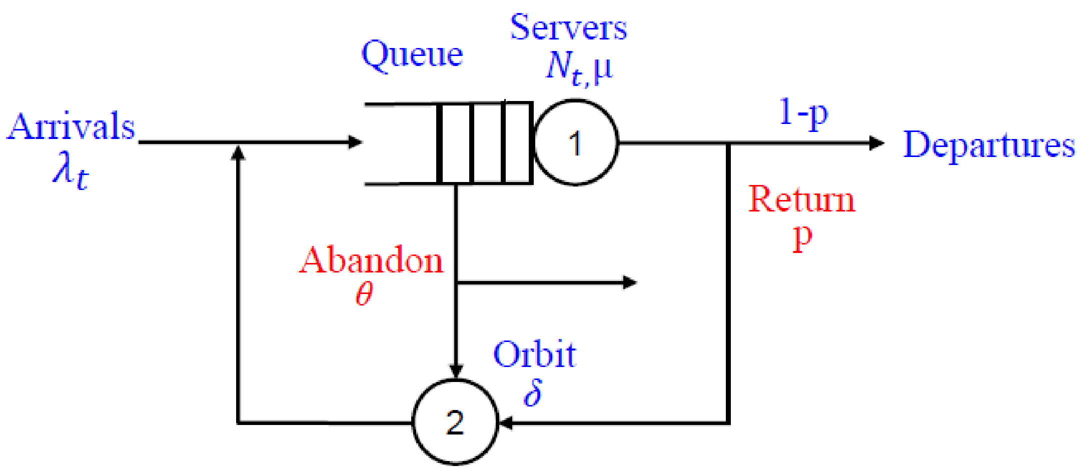
Queue "Shape"



- Area normalized to 100%
- State-Space Collapse

Model Selection: As Simple as Possible but Not Simpler

Service with Retrials and Abandonment; w/ Massey, Reiman, Stolyar



Laws & Models: Data-Based Erlang A/R/S, following B/C

- **Little's Law (Steady-State, Transient), State-Space Collapse,…**

- **Erlang-B (Blocking) and Erlang-C (?)**
  - **Erlang**, Agner Krarup: Queueing Theory was born in 1909, in his paper "The Theory of Probabilities and **Telephone Conversations**"

- **2. Erlang-A**
  - **Abandonment**: While waiting for service, does service-value dominate residual-wait-cost

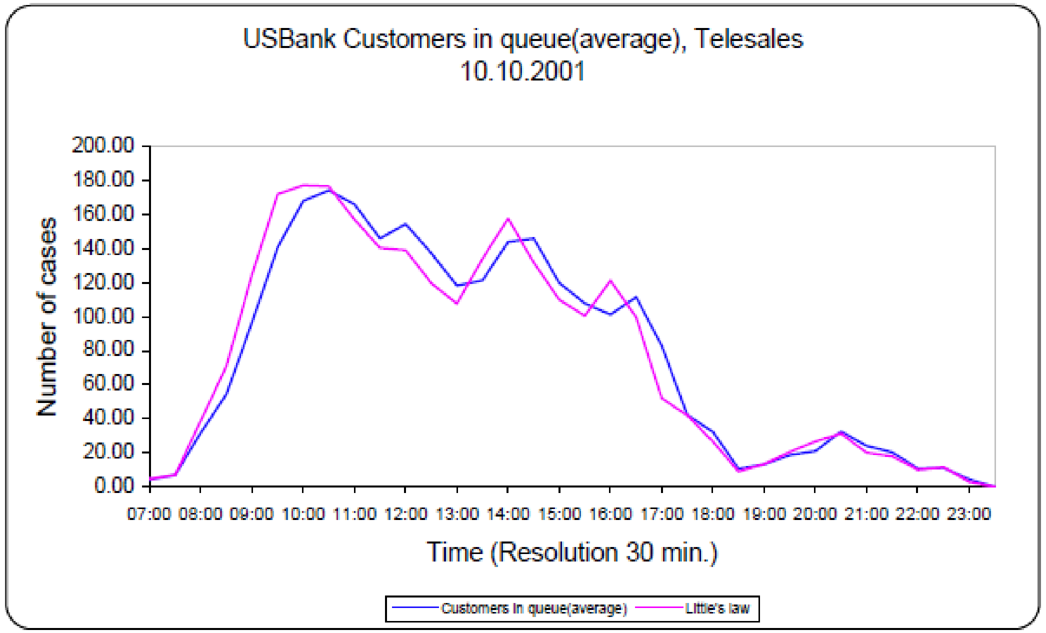
- **1. Erlang-R**
  - **Return/Feedback**: Customers often return to service (positively, negatively, just needing)

- **3. Erlang-S**
  - **Servers**: Challenging to manage, and model, no less so than customers

Above: Simple (Parsimonious) models of complex realities, yet not too simple (Robust)

Little's Law $L = \lambda \times W$, in a Time-Varying Environment

Time-Gap: **# in System** lags behind **Little / 30 min**

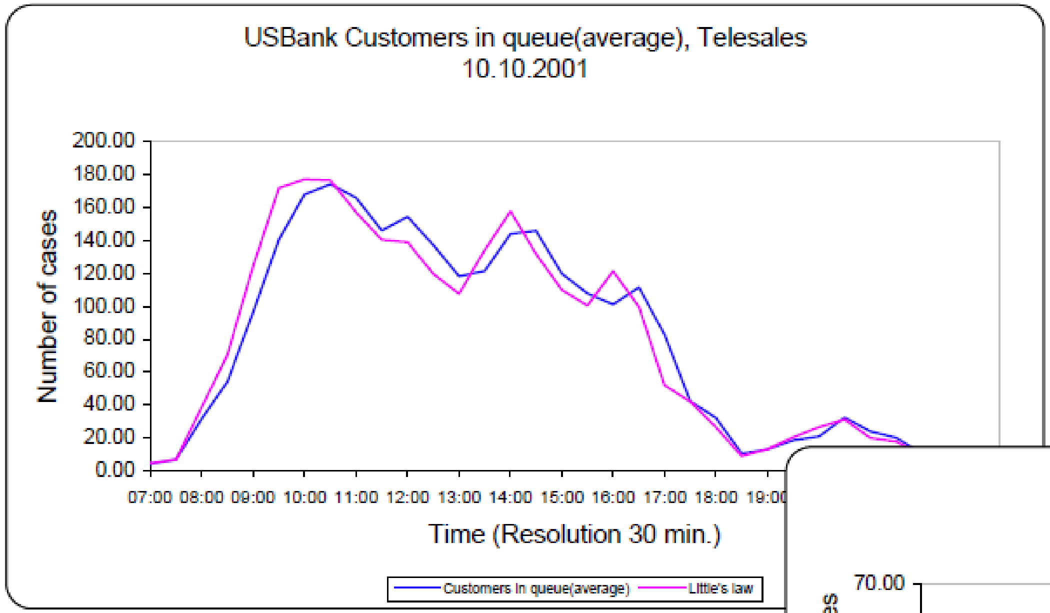


Call Center:

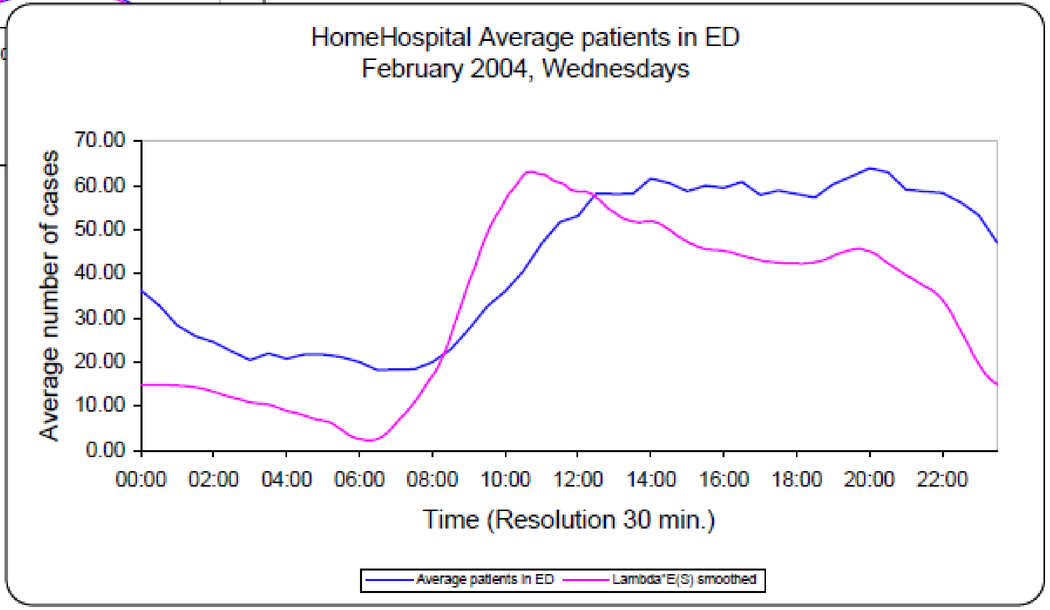
$\Rightarrow$ **Piecewise Steady-State**

Little's Law $L = \lambda \times W$, in a Time-Varying Environment

Time-Gap: **# in System** lags behind **Little / 30 min**



Call Center:
$\Rightarrow$ **Piecewise Steady-State**



Emergency Dept:
$\Rightarrow$ **Time-Varying Transient**

$$EL(t) = \tilde{\lambda}(t) \times EW,$$

$$\tilde{\lambda}(t) = E\lambda(t - W_e).$$

(Bertsimas, Mourtzinou;
Fralix, Riano, Serfozo)

Ideally for each model:

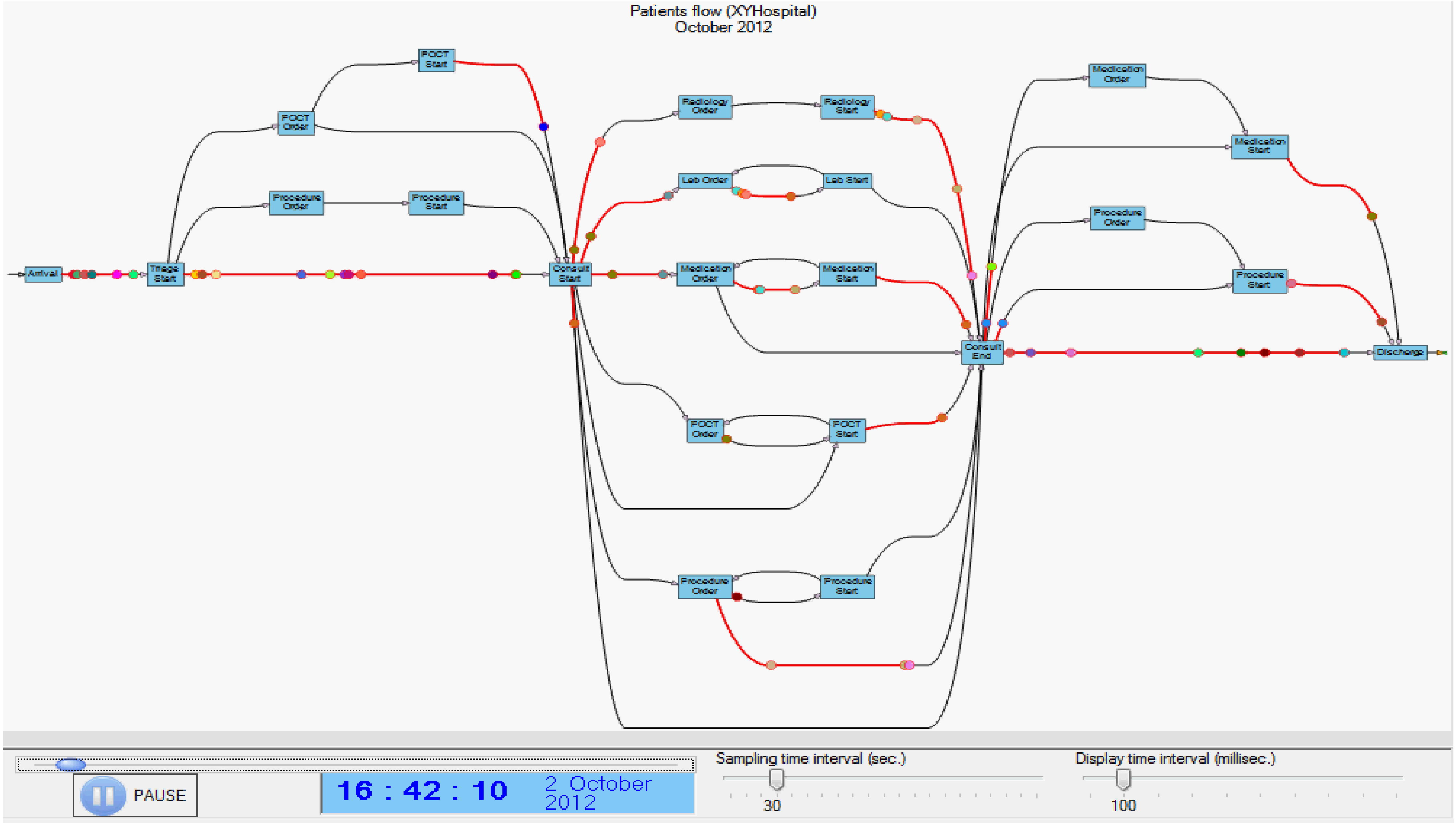
1. Motivating Phenomenon, via Data

2. Informal Description of a Model that captures the phenomenon

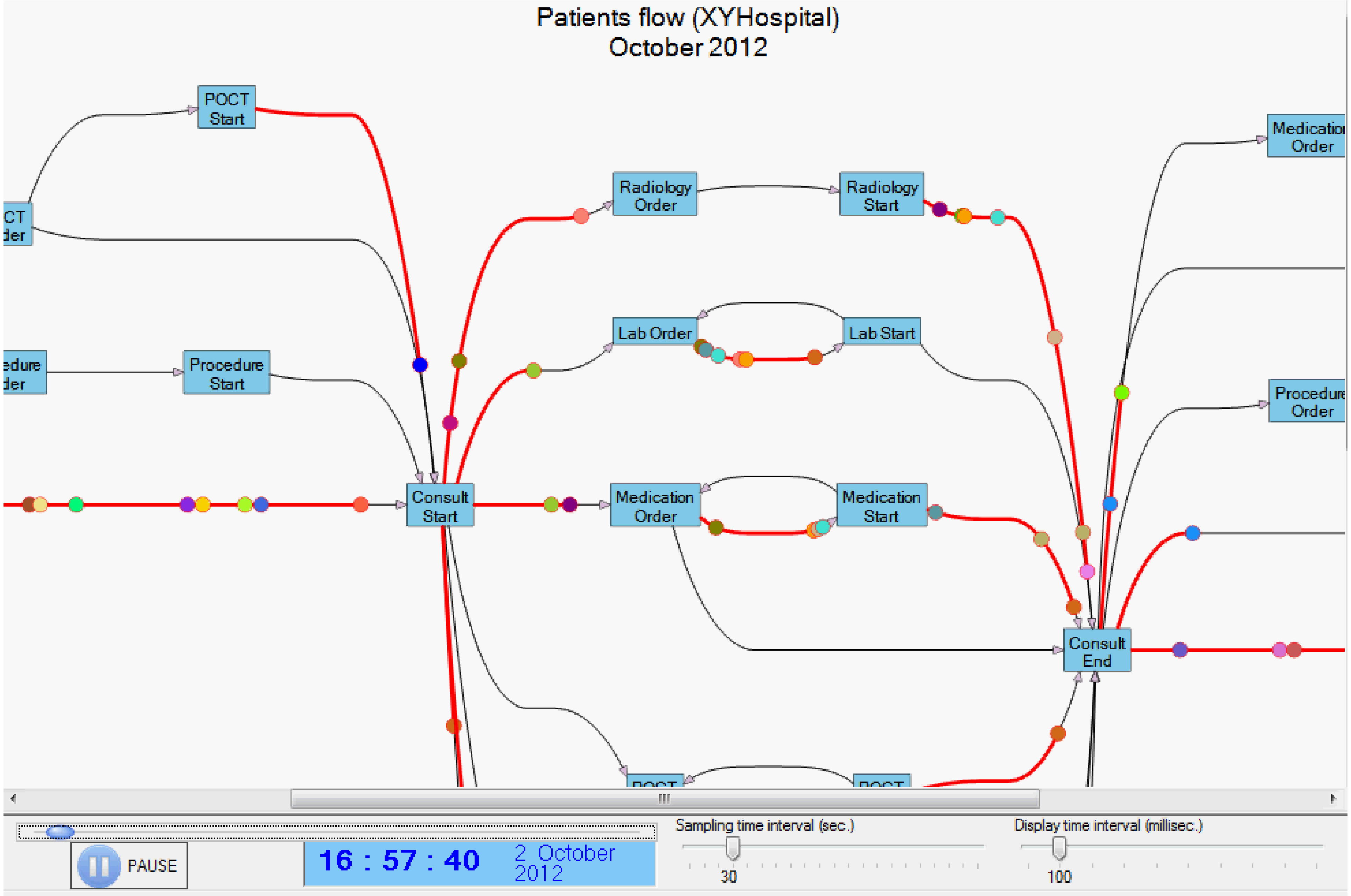
3. Example of Application(s)

4. Model Expressiveness (Strength)

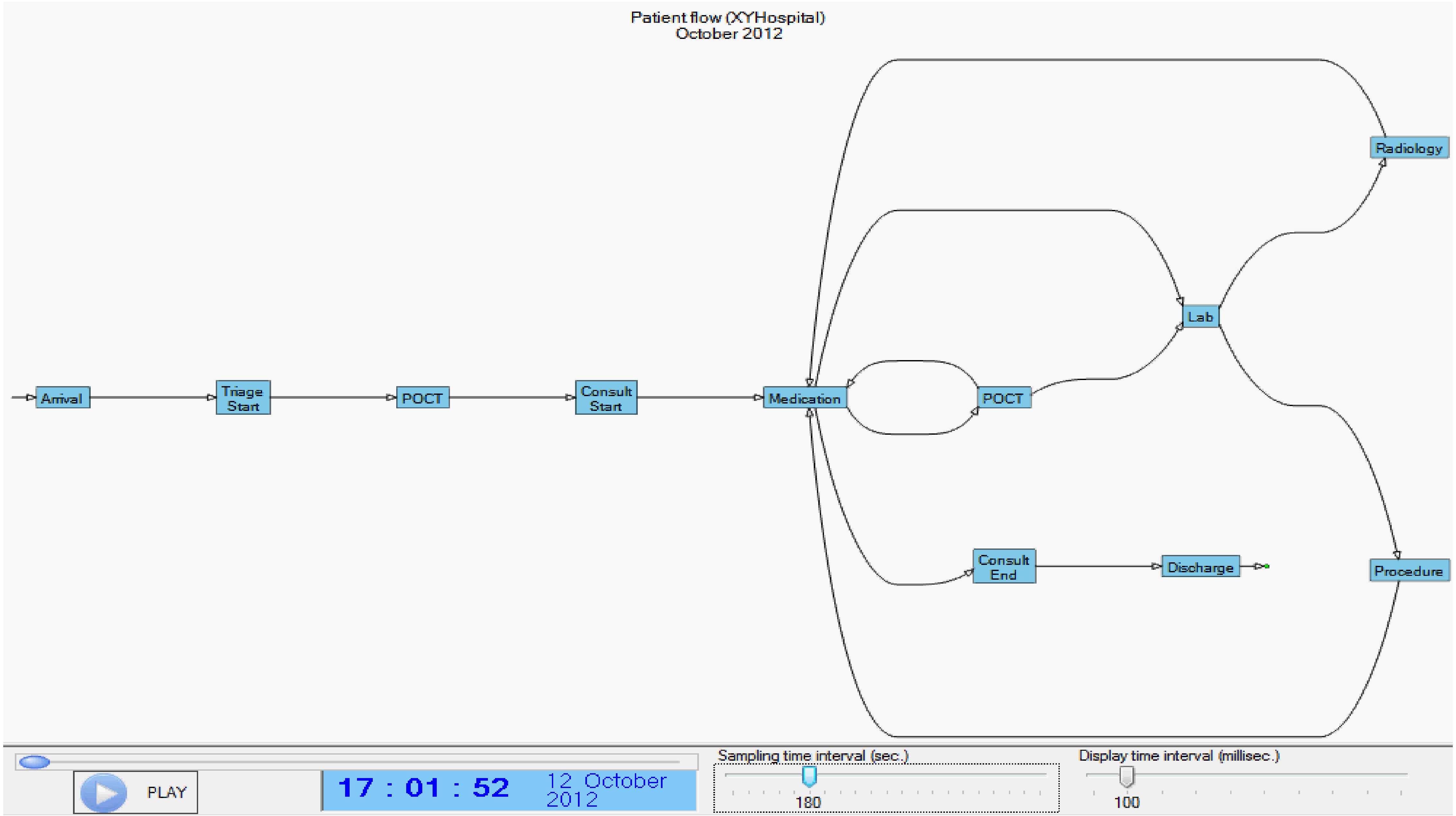
5. Insights, Extensions



Patients flow (XYHospital)
October 2012
Arrival
Triage Start
FOCT Order
FOCT Start
Procedure Order
Procedure Start
Consult Start
Radiology Order
Radiology Start
Lab Order
Lab Start
Medication Order
Medication Start
FOCT Order
FOCT Start
Procedure Order
Procedure Start
Consult End
Medication Order
Medication Start
Procedure Order
Procedure Start
Discharge
PAUSE
16 : 42 : 10
2 October 2012
Sampling time interval (sec.)
30
Display time interval (millisec.)
100

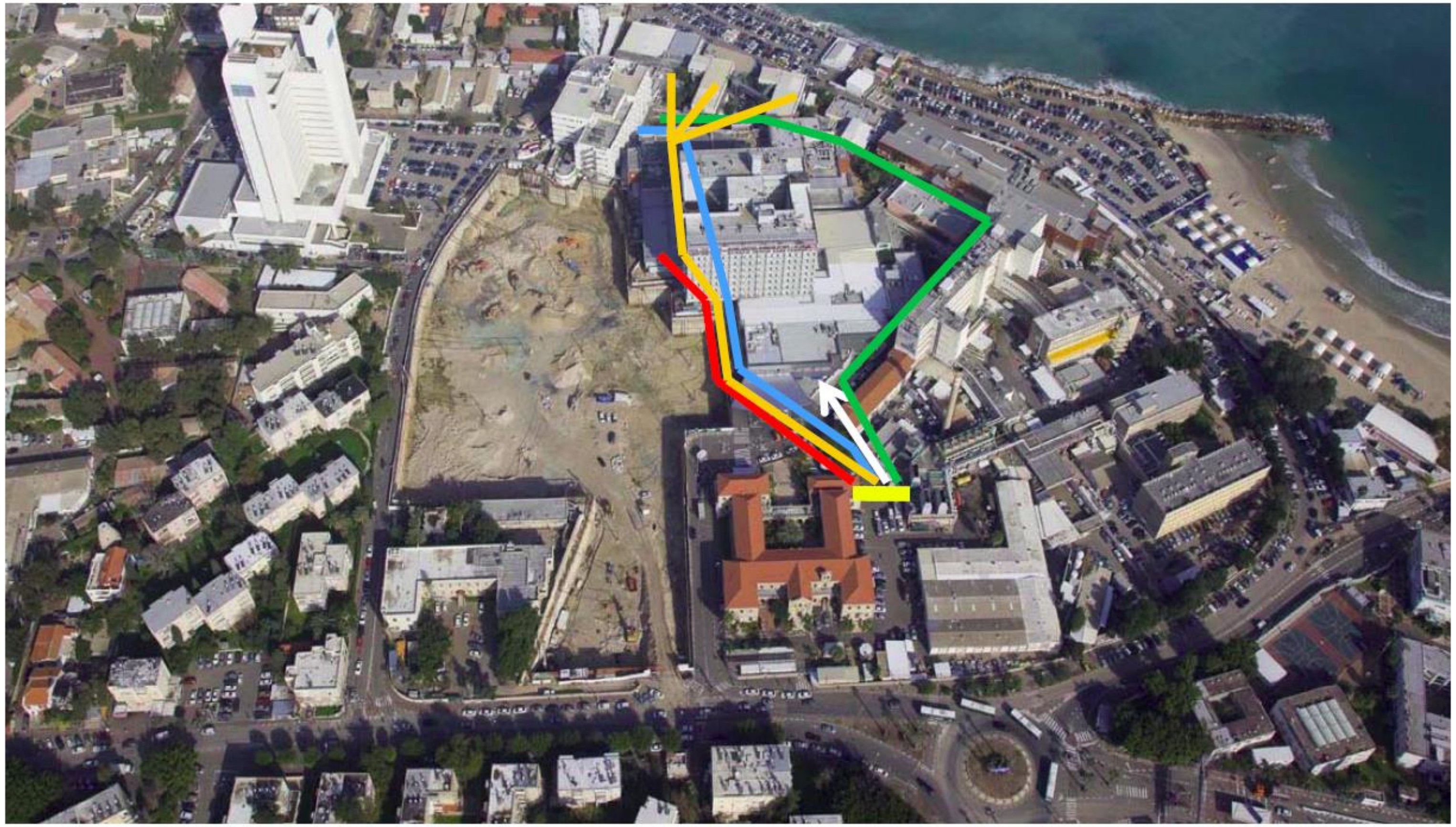


Patients flow (XYHospital)
October 2012
POCT Start
CT
der
edure
der
Procedure Start
Radiology Order
Radiology Start
Lab Order
Lab Start
Consult Start
Medication Order
Medication Start
Consult End
Medicatio Order
Procedure Order
PAUSE
16 : 57 : 40
2 October 2012
Sampling time interval (sec.)
30
Display time interval (millisec.)
100

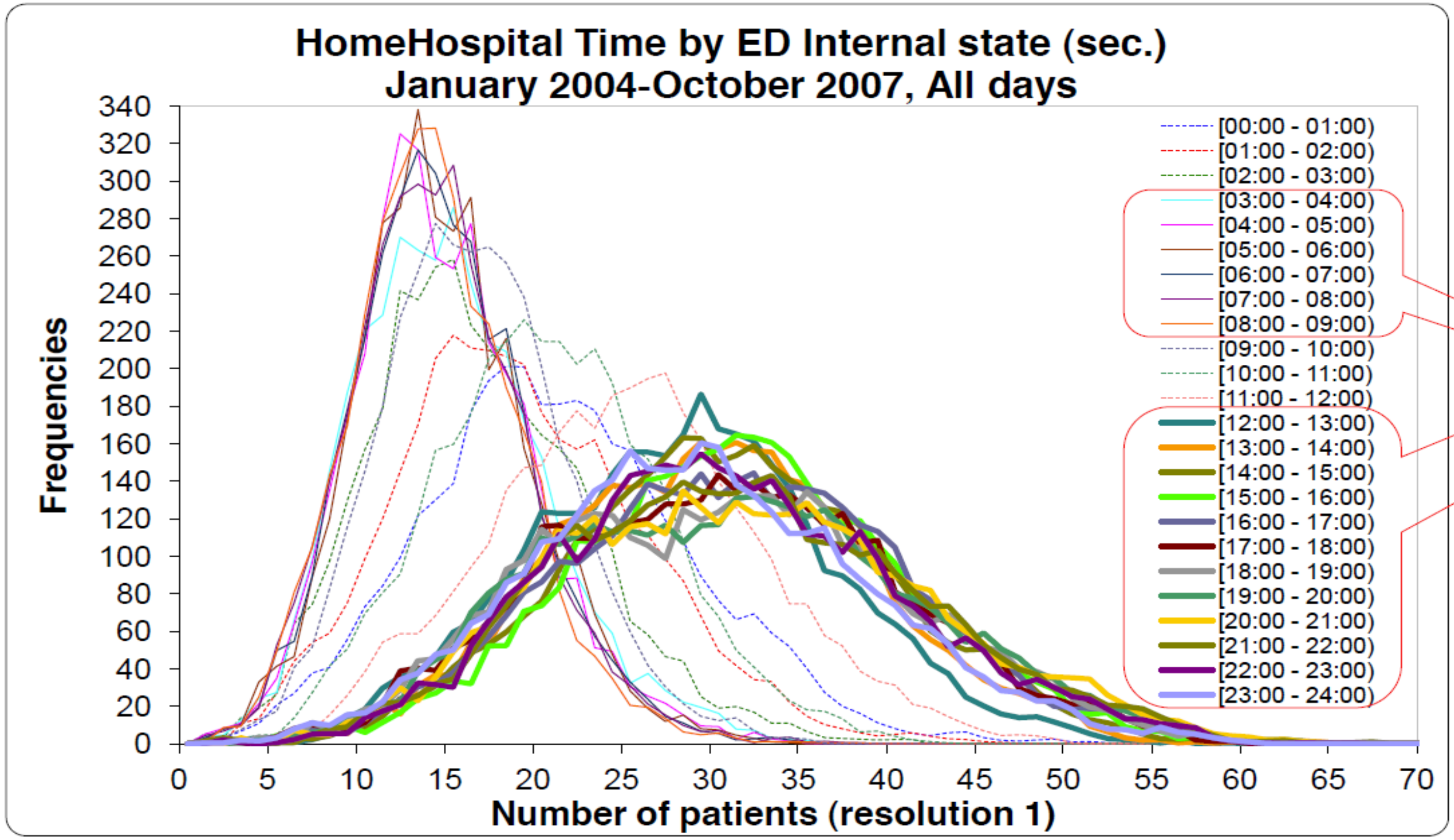

Patient flow (XYHospital)
October 2012
Arrival
Triage Start
POCT
Consult Start
Medication
POCT
Lab
Radiology
Procedure
Consult End
Discharge
PLAY
17 : 01 : 52
12 October 2012
Sampling time interval (sec.)
180
Display time interval (millisec.)
100

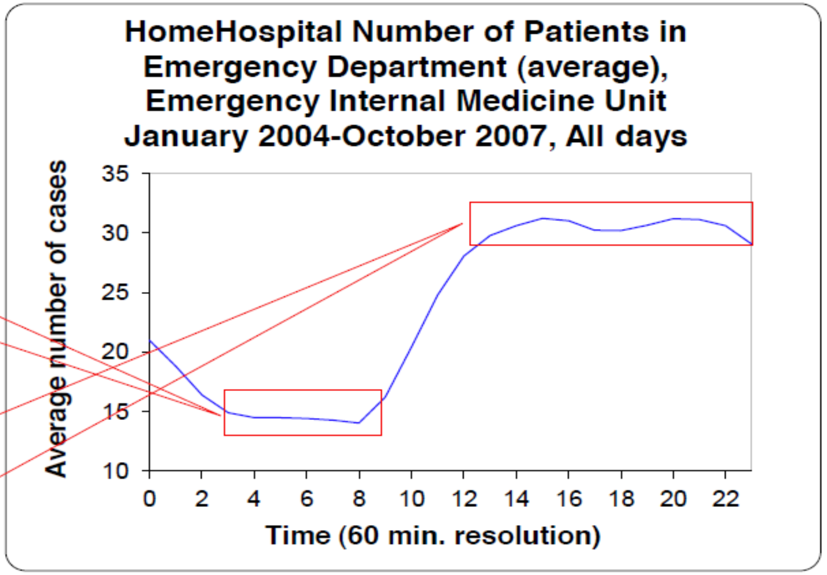
eg. RFID-Based Data: Mass Casualty Event (MCE)

Drill: Chemical MCE, Rambam Hospital, May 2010



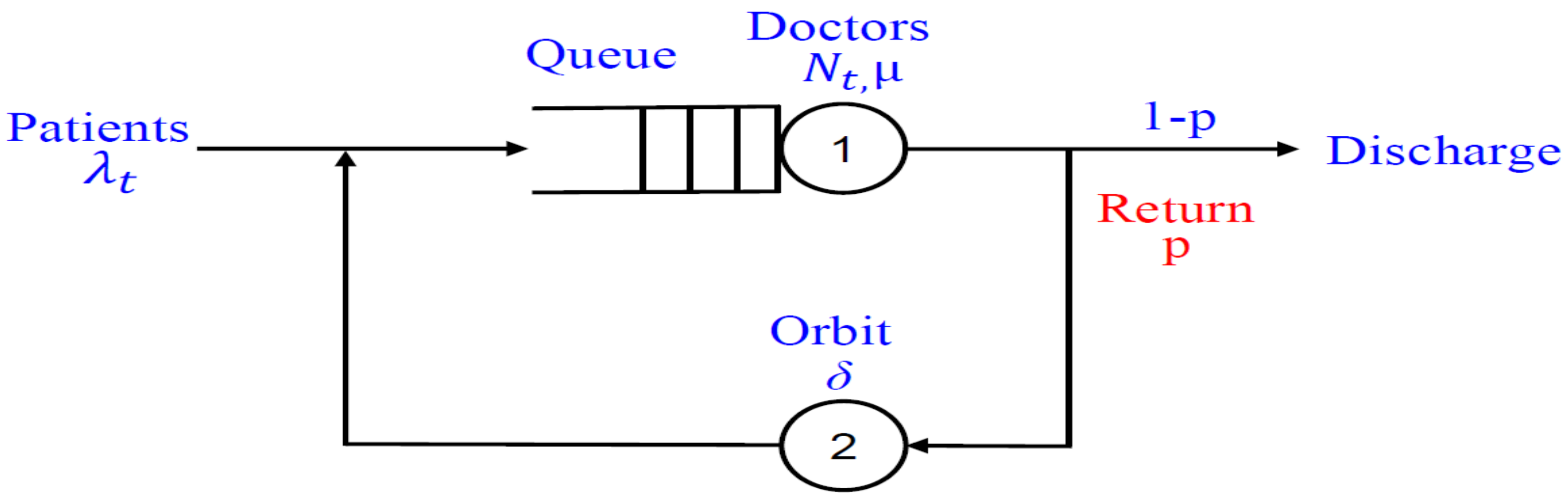
Focus on **severely wounded** casualties ($\approx$ 40 in drill)
Note: 20 observers support real-time control (helps validation)





Internal ED Occupancy histogram (left) and Average Census (right), by hour of the day

Erlang-R ↔ Fluid Model, w/ Galit Yom-Tov



Functional Strong Law of Large Numbers, for a 2-station QNet. BUT

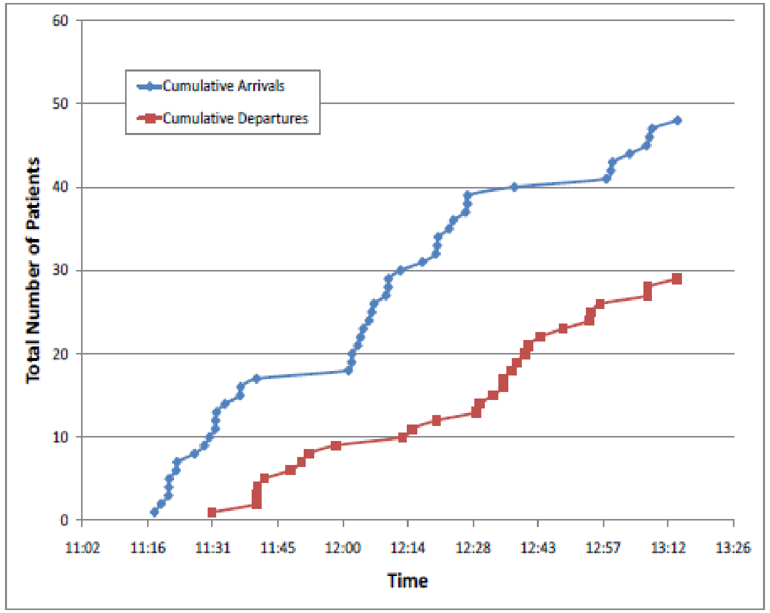
FNet = ODE: derived **directly** (no QNet), spreadsheet "solution"

$$\frac{d}{dt} q_t^1 = \lambda_t - \mu \cdot (q_t^1 \wedge N_t) + \delta \cdot q_t^2$$
$$\frac{d}{dt} q_t^2 = p \cdot \mu \cdot (q_t^1 \wedge N_t) - \delta \cdot q_t^2$$

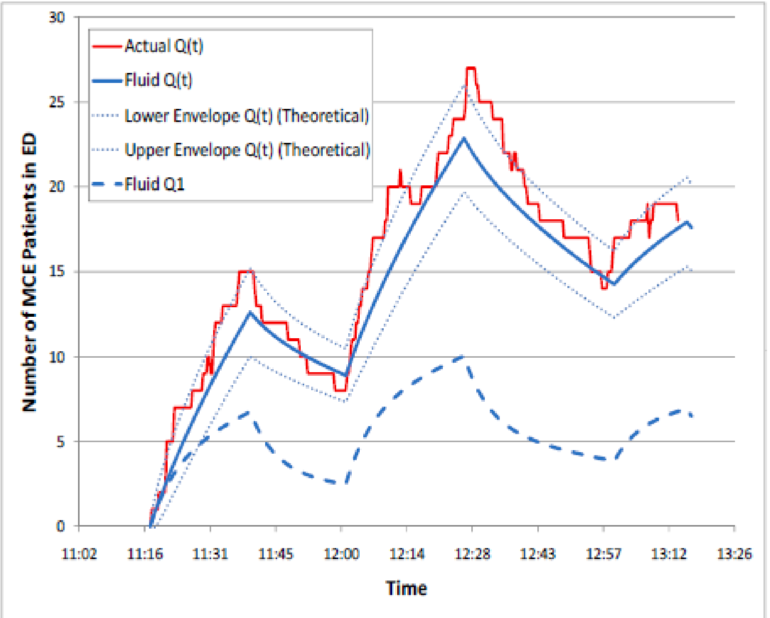
Erlang-R Value: FNet vs. Data

Chemical MCE Drill (Israel, May 2010, 11:00-13:00)

Arrivals & **Departures** (RFID)



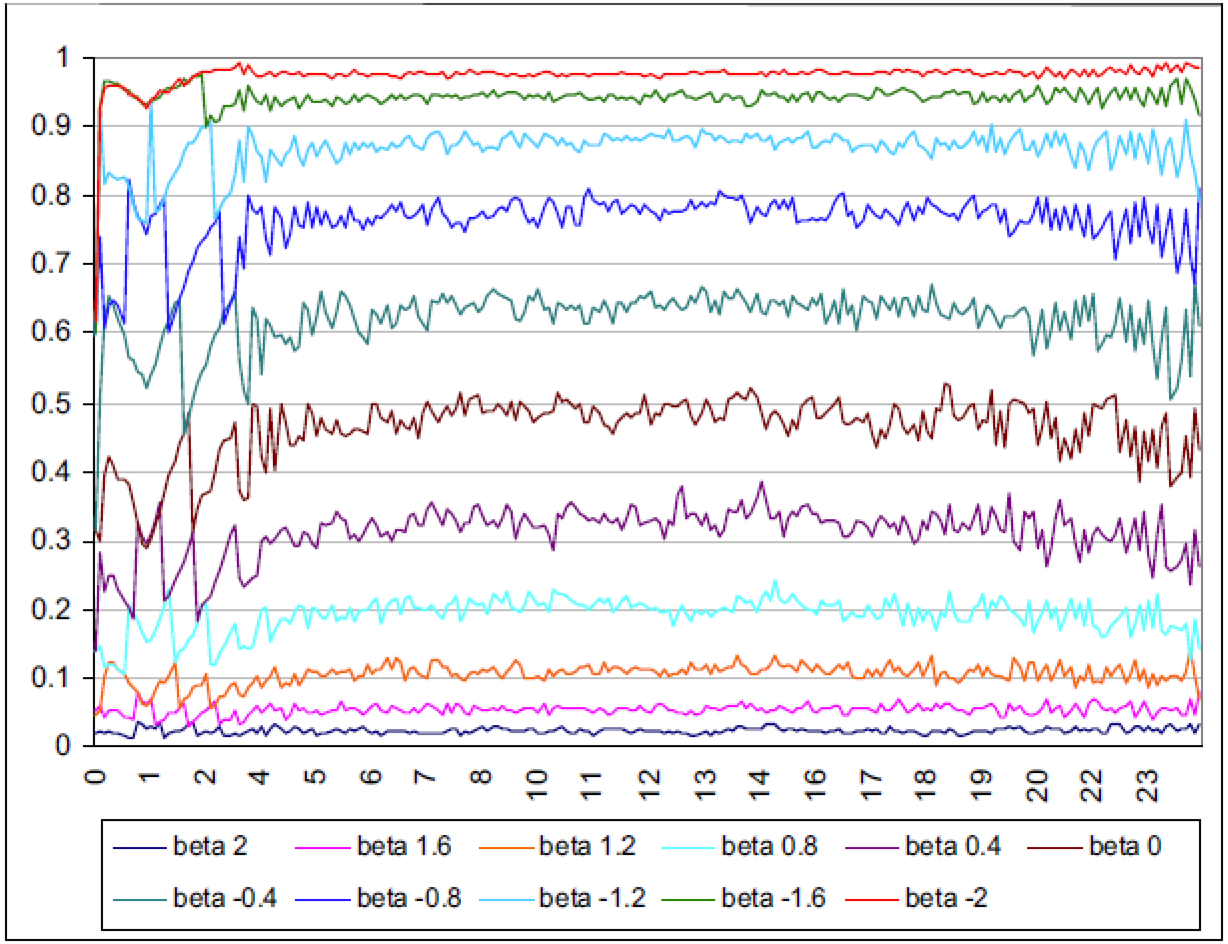
Erlang-R (**Fluid** , **Diffusion**)



- **Recurrent/Repeated** services in Chemical MCE: injection every 15/30/60 min
- **Fluid** = ODE
- **Diffusion** (confidence band), via F. Central Limit Theorem: Usefully narrow

Time-Stable Performance of Time-Varying Systems

Delay Probability = As in the **Stationary Erlang-A** / **R**



A Data-Based Framework, or “Erlang-R in the ED"

System = e.g. Emergency Department

- **QNet** = Erlang-R (time-varying 2-station Jackson; w/ Yom-Tov)
- **FNets** = 2-dim dynamical system (Massey & Whitt)
- **DNets** = 2-dim Markovian Service Net (w/ Massey and Reiman)
- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)

A Data-Based Framework, or “Erlang-R in the ED"

System = e.g. Emergency Department

- **QNet** = Erlang-R (time-varying 2-station Jackson; w/ Yom-Tov)
- **FNets** = 2-dim dynamical system (Massey & Whitt)
- **DNets** = 2-dim Markovian Service Net (w/ Massey and Reiman)
- **SimNet** = Customized ED-Simulator (Marmor & Sinreich)

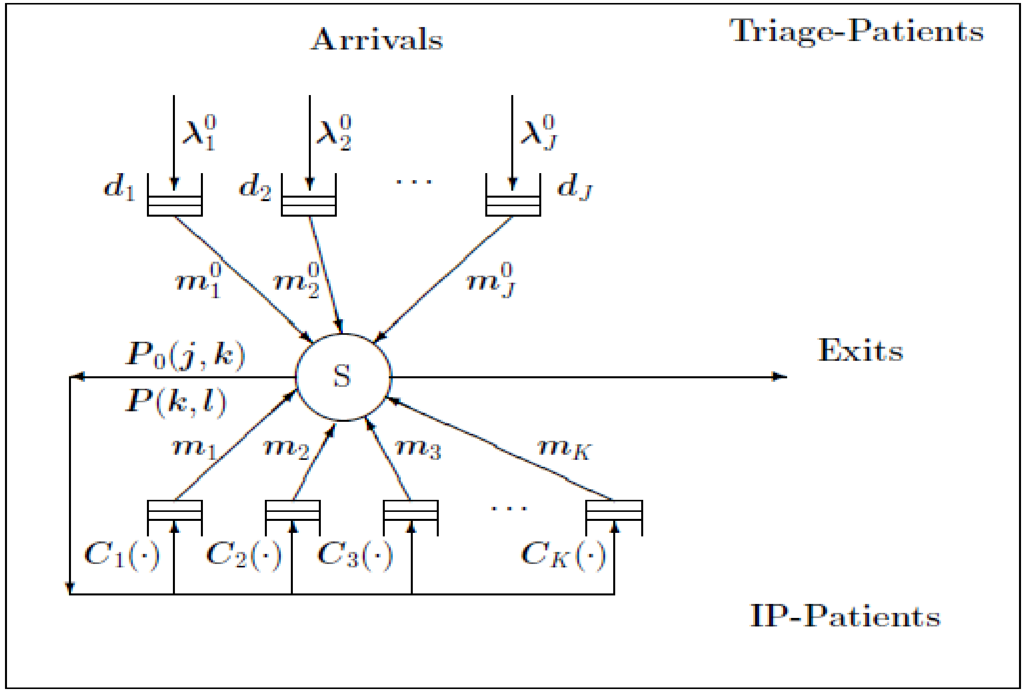
Framework: Mining (all) ServNets from Data

- **MCE ED**: FNet $\Rightarrow$ Census, DNet = Confidence band
  Performance Analysis, Prediction
  Validated against Data

- **Normal ED**: FNet $\Rightarrow$ Physician offered-load $\Rightarrow$ $\sqrt{}$-Staffing
  Staffing to stabilize operational performance
  Validated against SimNet

ED Patient Flow: The Physicians View

with **J. Huang, B. Carmeli, S. Israelit**

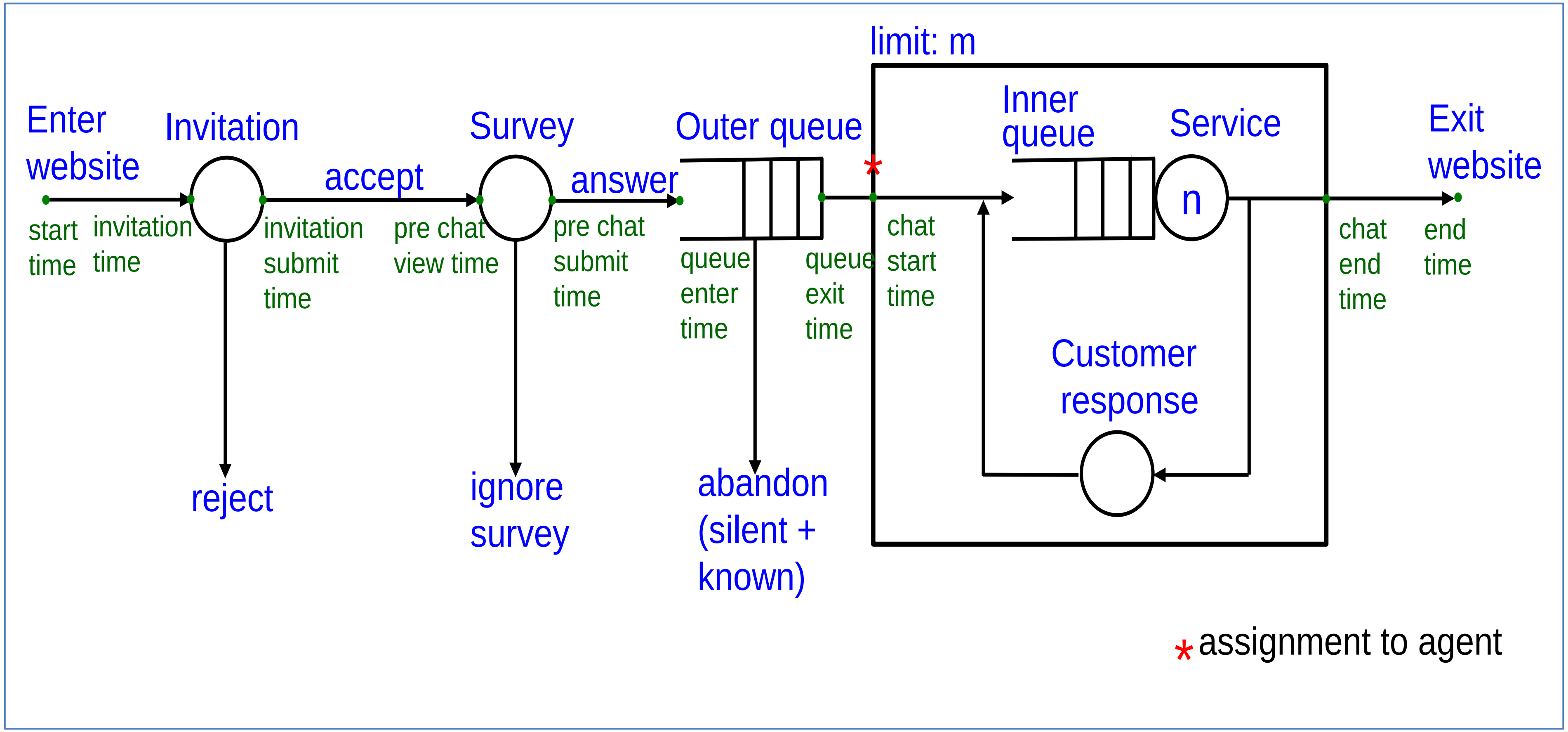


Goal: Adhere to **Triage-Constraints**, then **release In-Process** Patients

Following **Plambeck, Kumar, Harrison (2001): Throughput-time constraints**

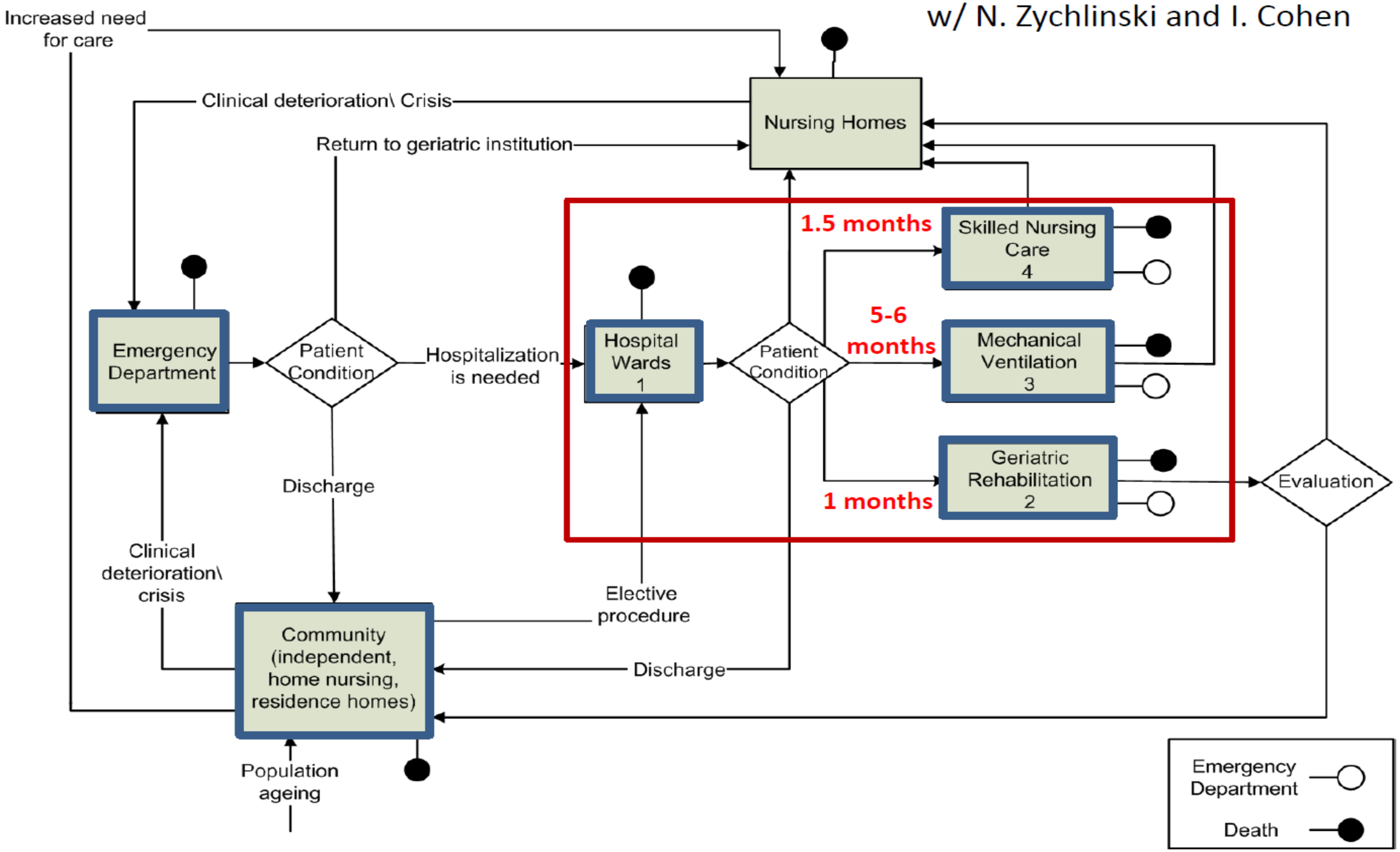
Online Chats: from Internet-Prompts to Chat-Sessions

Server with Multiple Returning Customers
Galit Yom-Tov, Anat Rafaeli, graduate (OB) students

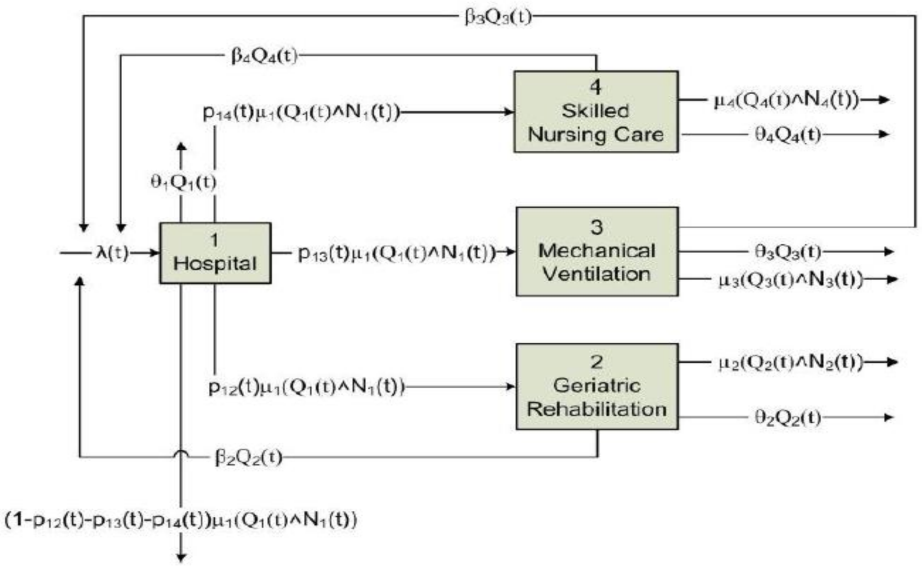


Patient Flow: ED to Wards to Nursing/Geriatric Institutions

w/ N. Zychlinski and I. Cohen



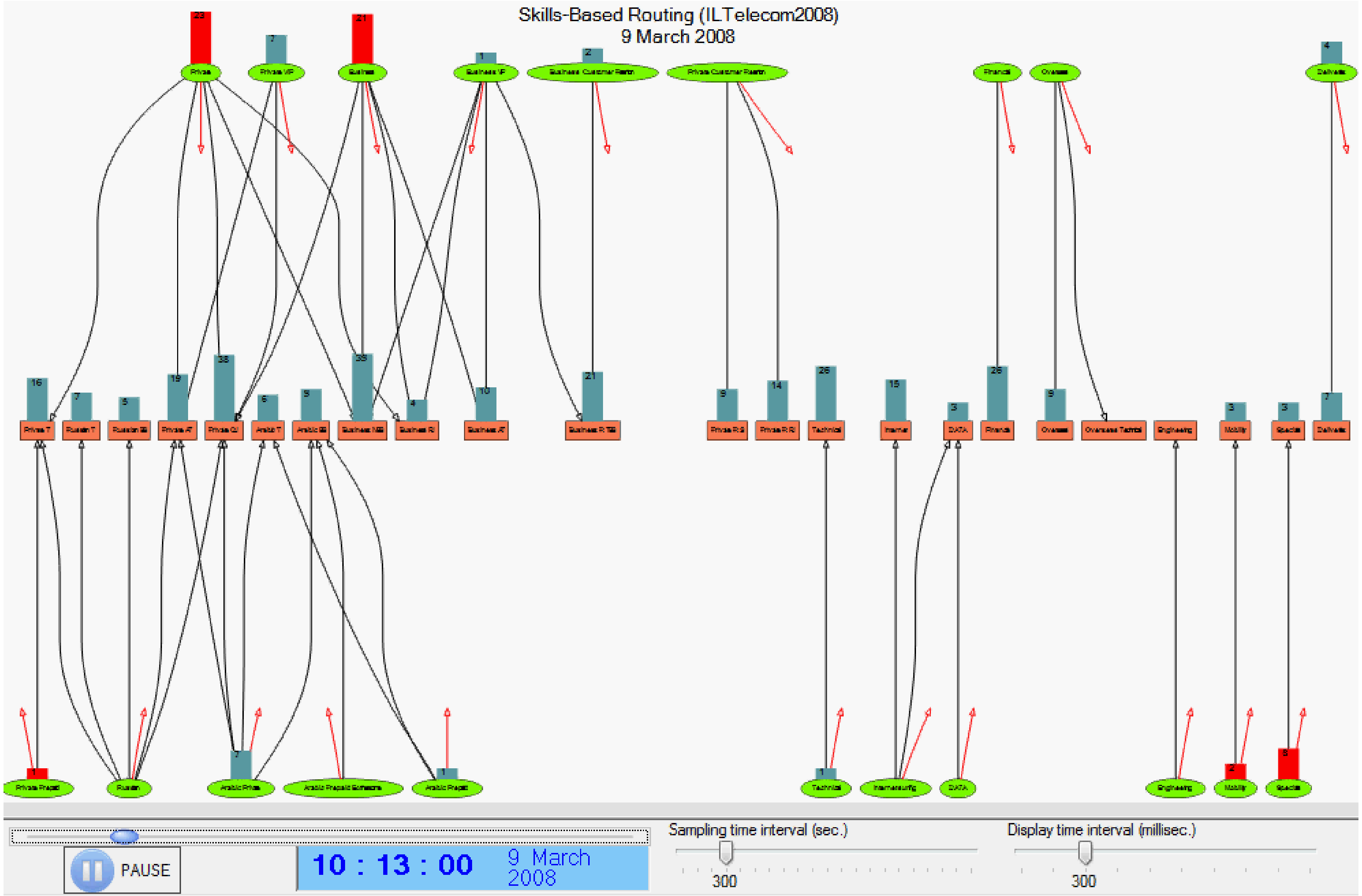
A Fluid Model



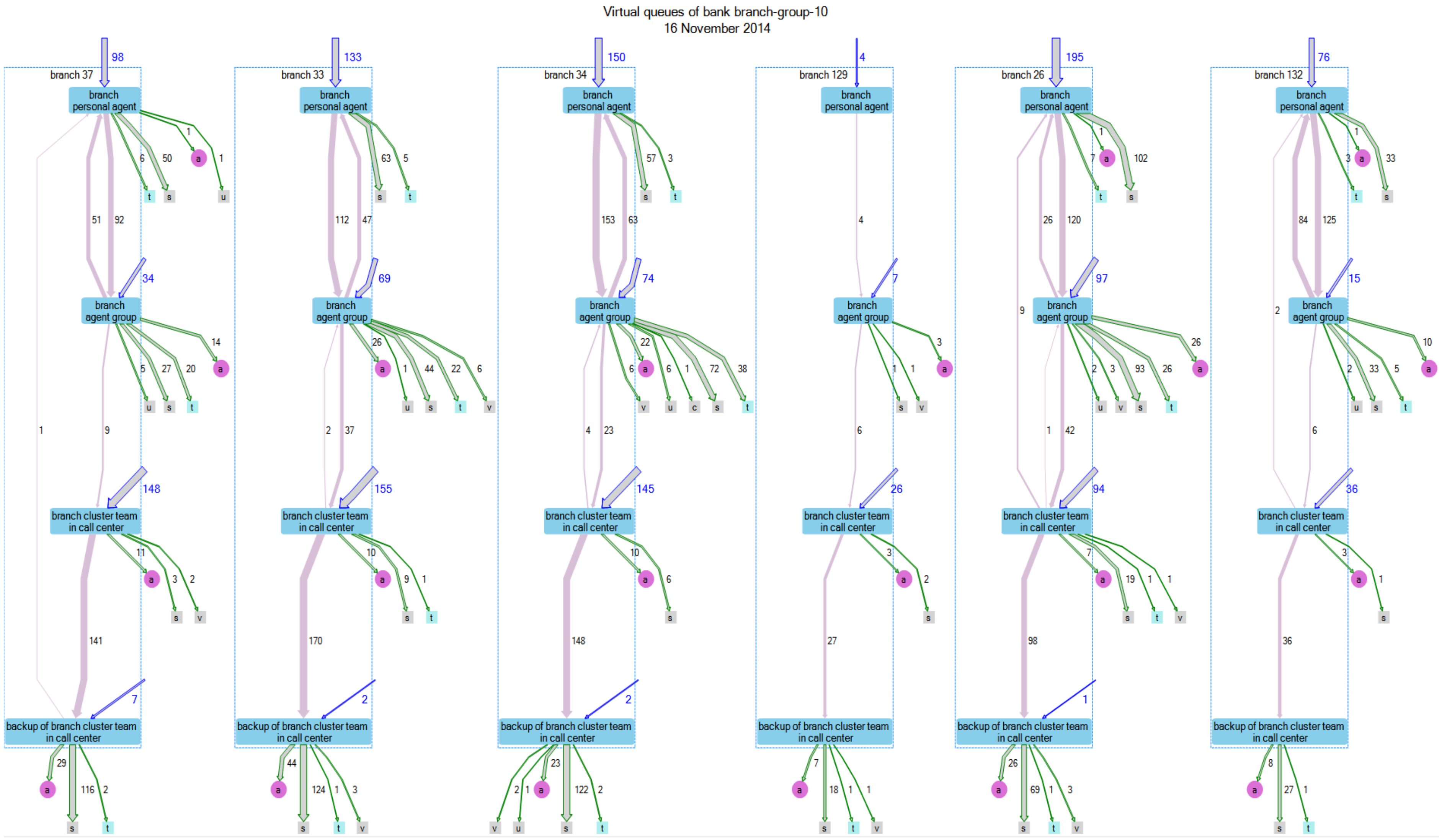
Stations 2,3 and 4

$$\dot{Q}_2(t) = (p_{12}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_2 - Q_2(t))^+) -$$
$$\theta_2 Q_2(t) - \beta_2 Q_2(t) - \mu_2 \cdot (Q_2(t) \wedge N_2)$$
$$\dot{Q}_3(t) = (p_{13}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_3 - Q_3(t))^+) -$$
$$\theta_3 Q_3(t) - \beta_3 Q_3(t) - \mu_3 \cdot (Q_3(t) \wedge N_3)$$
$$\dot{Q}_4(t) = (p_{14}(t)\mu_1 \cdot (Q_1(t) \wedge N_1)) \wedge (N_4 - Q_4(t))^+) -$$
$$\theta_4 Q_4(t) - \beta_4 Q_4(t) - \mu_4 \cdot (Q_4(t) \wedge N_4).$$

3


Skills-Based Routing (ILTelecom2008)
9 March 2008
PAUSE
10 : 13 : 00
9 March 2008
Sampling time interval (sec.)
300
Display time interval (millisec.)
300

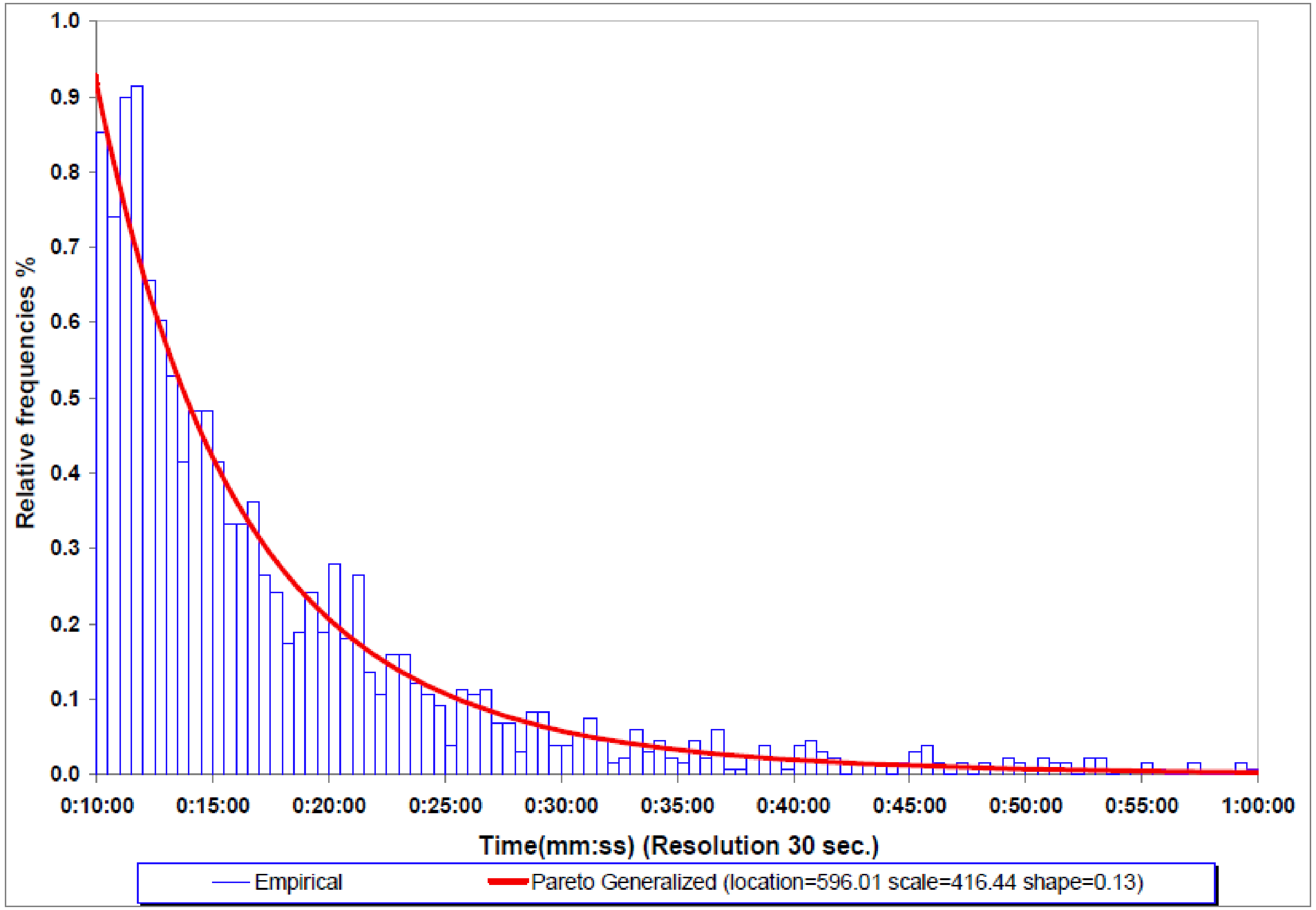
ILDU Banking: **Single Branch-Group-10 Virtual Queues - note the Abandonments**



Primitives: (Im)Patience

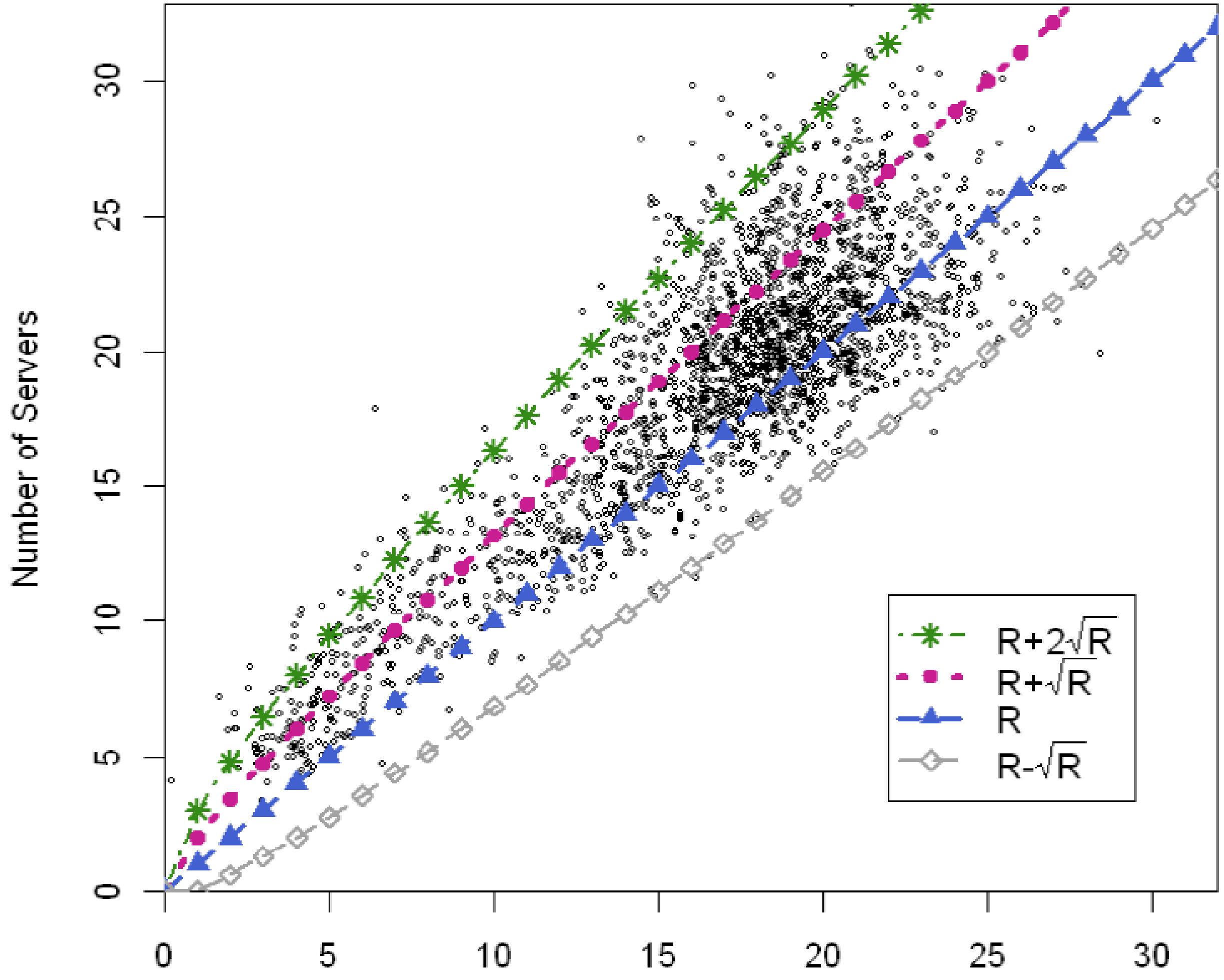
Israeli Bank: Uncensored 13,000 Customers, 24/11/2008

Patience $\geq 10min$: Why Pareto Tail?



Beyond Fluid: #Agents vs. Offered-Load ($N \approx R + \beta\sqrt{R}$)

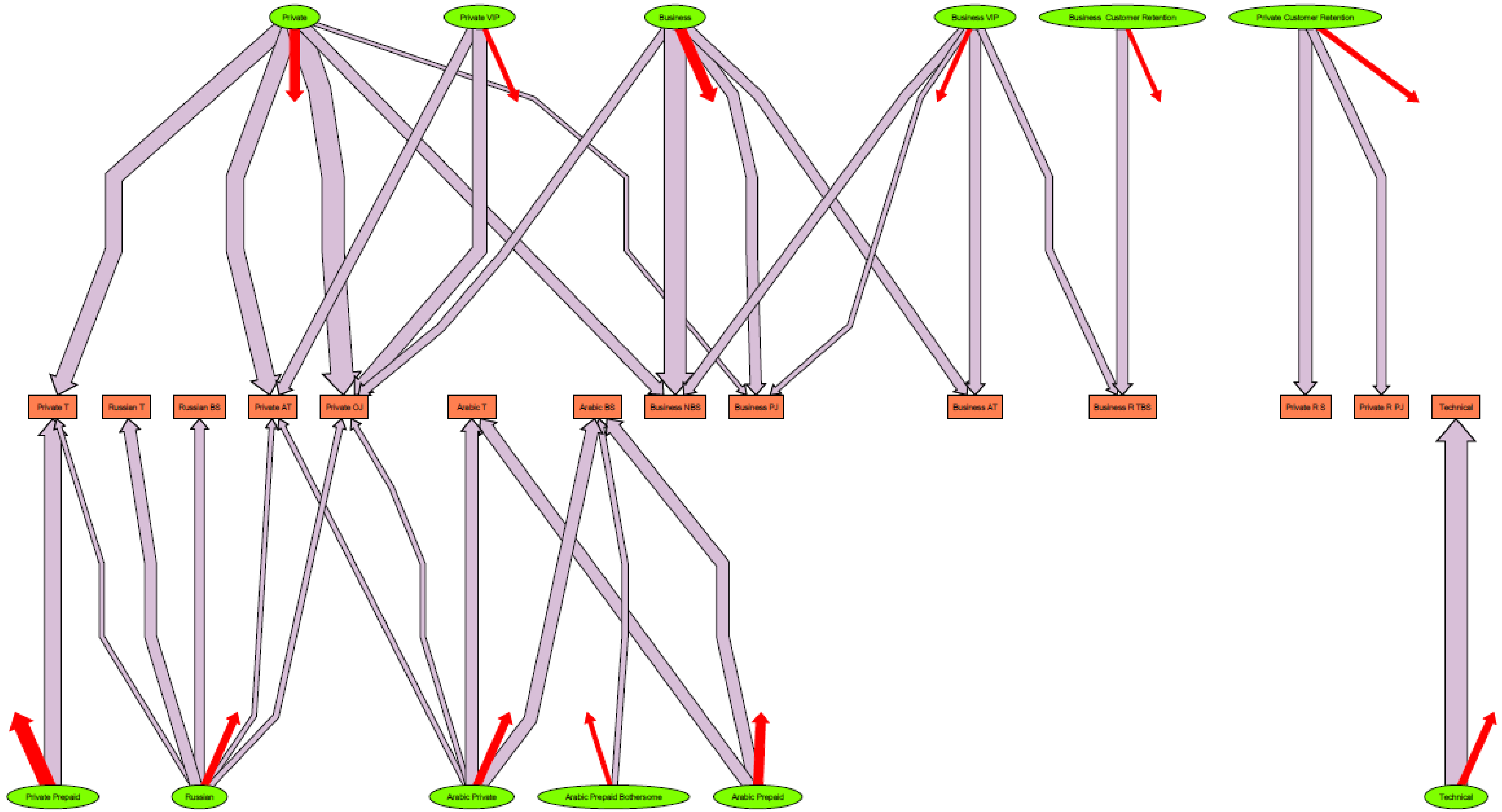
IL Telecom; June-September, 2004 (2205 30min intervals, over 13 weeks, week-days)



e.g. Offered-load R $\overset{avg}{=}$ 5 calls per min $\times$ 3.2 min per call = 16 Erlangs

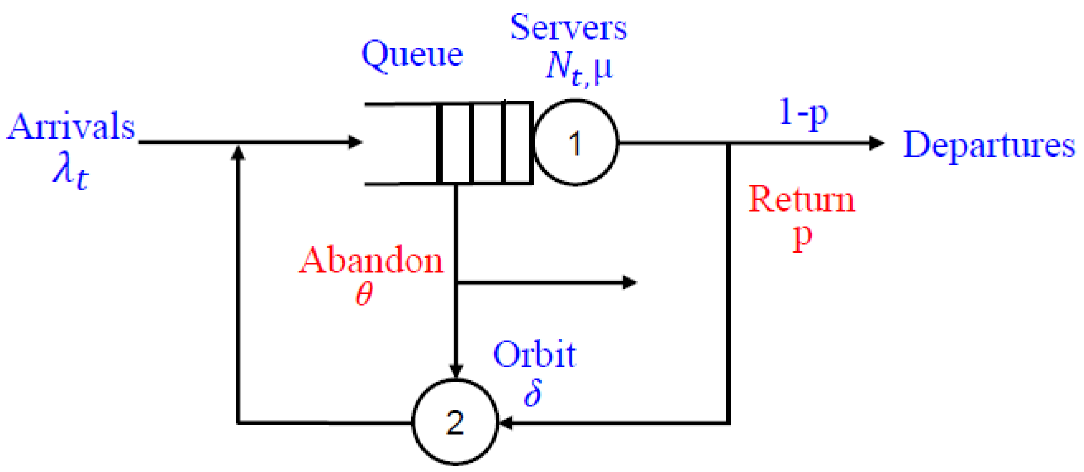
Impatient Customers - Isolate or Aggregate

ILTelecom 9/3/2008



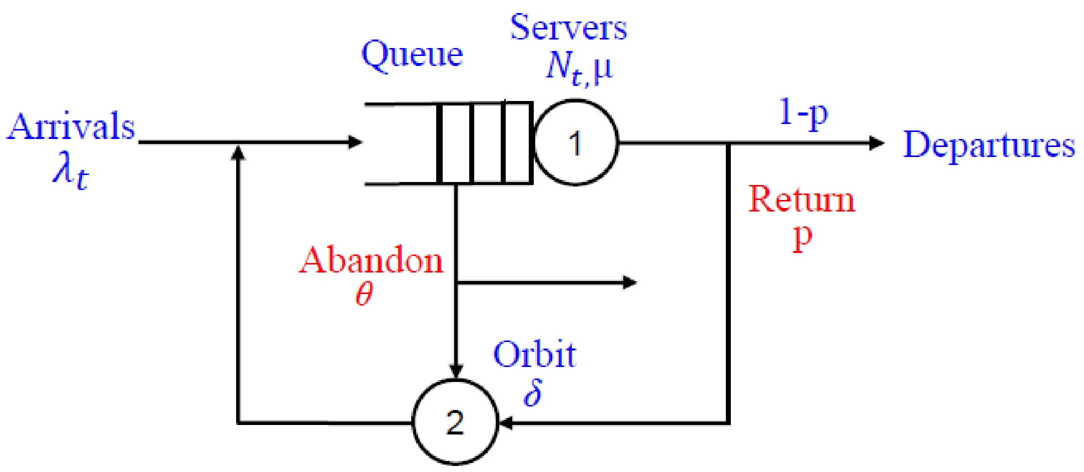
Model Selection: As Simple as Possible but Not Simpler

Service with Retrials and Abandonment; w/ Massey, Reiman, Stolyar



Model Selection: As Simple as Possible but Not Simpler

Service with Retrials and Abandonment; w/ Massey, Reiman, Stolyar

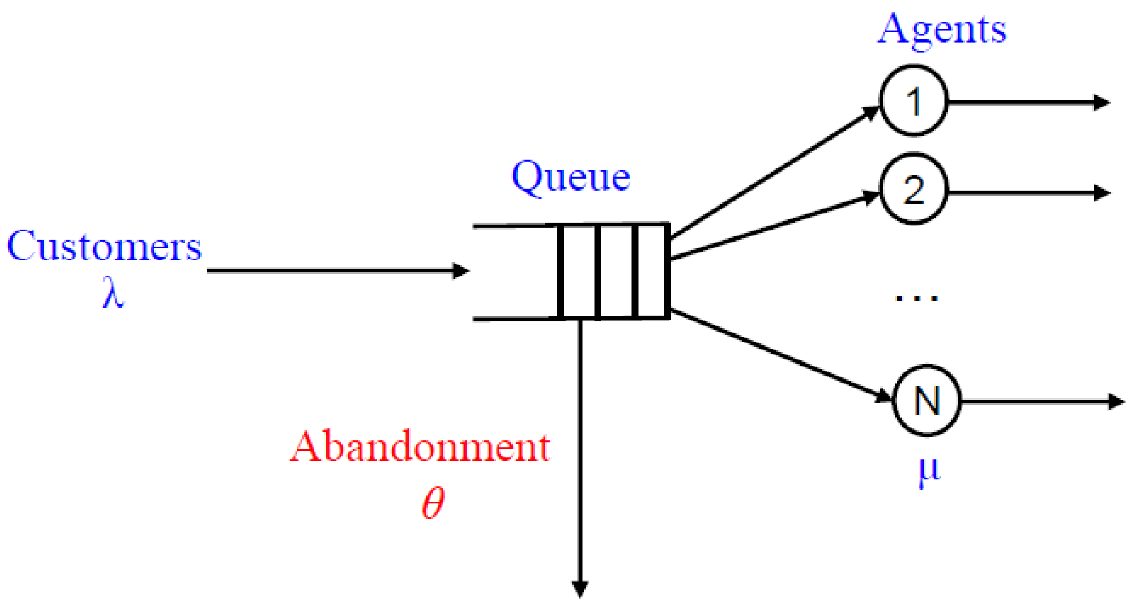


- Call centers: Visit durations naturally measured in minutes
  - Arrival rates are “constant" during visit
  - Returns occur hours after visit

$\Rightarrow$ “Select" Base Model (of 1/2 hour):

Stationary, Abandonment

A Basic Staffing Model: Erlang-A



w/ O. Garnett

"Birth & Death" Queue = M/M/N + M (Palm 1940's):

- λ – **Arrival** rate (Poisson)
- μ – **Service** rate (Exponential; $E[S] = \frac{1}{\mu}$)
- θ – **Patience** rate (Exponential, $E[\text{Patience}] = \frac{1}{\theta}$)
- N – Number of **Servers** (Agents).

Erlang-A: Is it Relevant?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).

- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- · · · , and more.

Question: **Is Erlang-A Relevant?**

Erlang-A: Is it Relevant?

Experience:

- Arrival process **not pure Poisson** (time-varying, σ^2 too large)
- Service times **not Exponential** (typically close to LogNormal)
- Patience times **not Exponential** (behavior-dependent).
- Building Blocks need **not be independent** (eg. long wait associated with long service; **w/ M. Reich & Y. Ritov**)
- Customers and Servers **not homogeneous** (classes, skills): **w/ R. Atar, G. Shaikhet; R. Atar, I. Gurvich,** . . .
- Customers return for service (after busy, abandonment; dependently: **P. Khudiakov, R. Ghebali, M. Gorfine, P. Feigin**)
- · · · , and more.

Question: **Is Erlang-A Relevant? Robust enough?** **YES !**

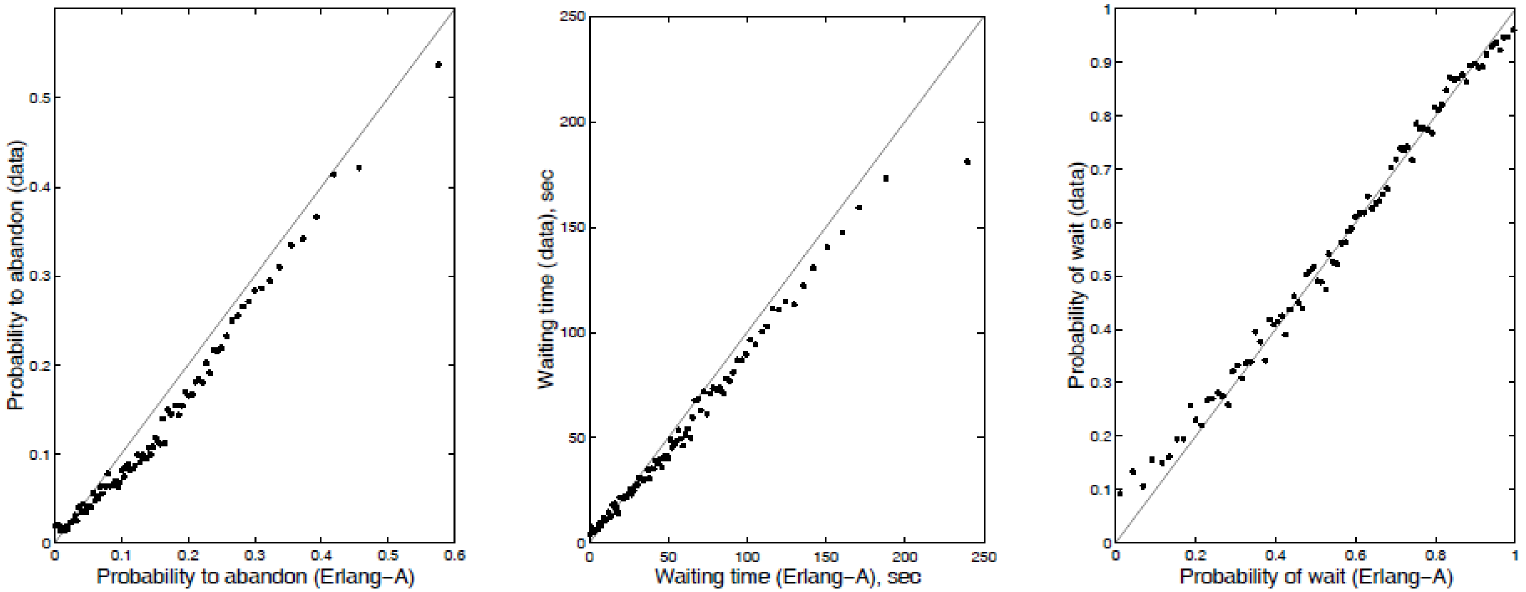
- **Practice**: Staffing engine of Work-Force Management software
- **Theory**: Theoretical engine of Operational Regimes

  QD, **ED**, **QED**

Erlang-A: Fitting a Simple Model to a Complex Reality

Hourly Performance vs. Erlang-A Predictions (1 year)

% Abandon | E[*Wait*] | %$\{Wait > 0\}$



- Empirically-Based & Theoretically-Supported Estimation of (Im)Patience: $\hat{\theta} = P\{Ab\}/E[W_q]$)
- Small Israeli Bank (more examples in progress)
- Hourly performance vs. Erlang-A predictions, 1 year: aggregated groups of 40 similar hours

Example of a Theorem (QED)

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

Yet, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait "**seconds**" for **minutes** service;
- **Transportation**: Search "**minutes**" for **hours** parking;

Prerequisite II: Models (Diffusion/QED's Q's)

Traditional Queueing Theory predicts that **Service-Quality** and **Servers' Efficiency must** be traded off against each other.

For example, **M/M/1** (single-server queue): **91%** server's utilization goes with

$$\textbf{Congestion Index} = \frac{E[Wait]}{E[Service]} = \mathbf{10},$$

and only 9% of the customers are served immediately upon arrival.

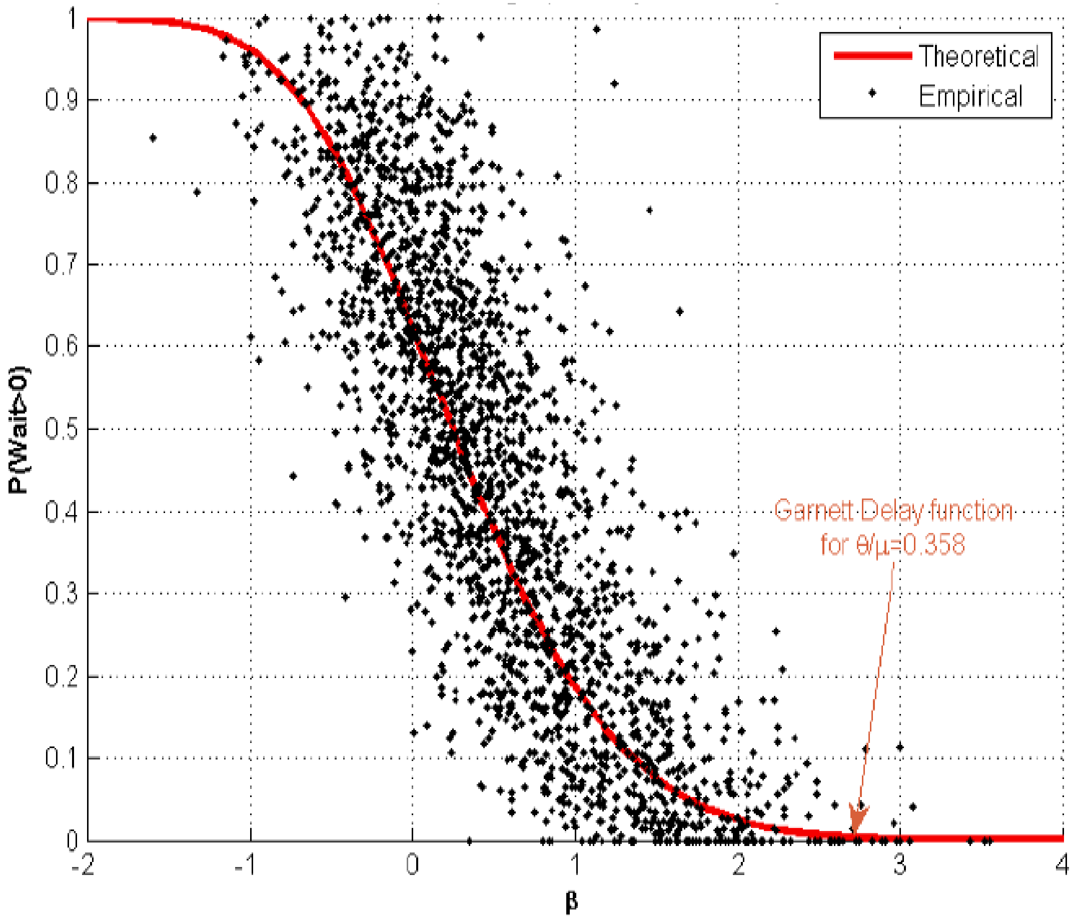
Yet, **heavily-loaded** queueing systems with **Congestion Index = 0.1** (Waiting one order of magnitude less than Service) are prevalent:

- **Call Centers**: Wait "**seconds**" for **minutes** service;
- **Transportation**: Search "**minutes**" for **hours** parking;
- **Hospitals**: Wait "**hours**" in ED for **days** hospitalization in IW's;

and, moreover, a significant fraction are not delayed in queue. (For example, in well-run call-centers, **50%** served "immediately", along with over **90%** agents' utilization, is not uncommon) **?** **QED**

Erlang-A Value: DNet $P(W_q > 0)$ vs. Data

IL Telecom; June-September, 2004 (2205 30min intervals, weekdays)



- Approximations, w/ Patience $\approx 3\times$ Service-Duration ($\mu/\theta \approx 3$)

QED Theory (Erlang '13; Halfin & Whitt '81; w/Garnett & Reiman '02)

Consider a sequence of **steady-state** M/M/N + M queues, $N = 1, 2, 3, \ldots$
Then the following points of view are **equivalent**, as $N \uparrow \infty$:

- **QED** $\quad \%\{\text{Cust Wait} > 0\} \approx \alpha, \qquad 0 < \alpha < 1;$

  or $\quad \%\{\text{Serv Idle} > 0\} \approx 1 - \alpha$

- **Customers** $\quad \{\text{Abandon}\} \approx \frac{\gamma}{\sqrt{N}}, \qquad 0 < \gamma;$

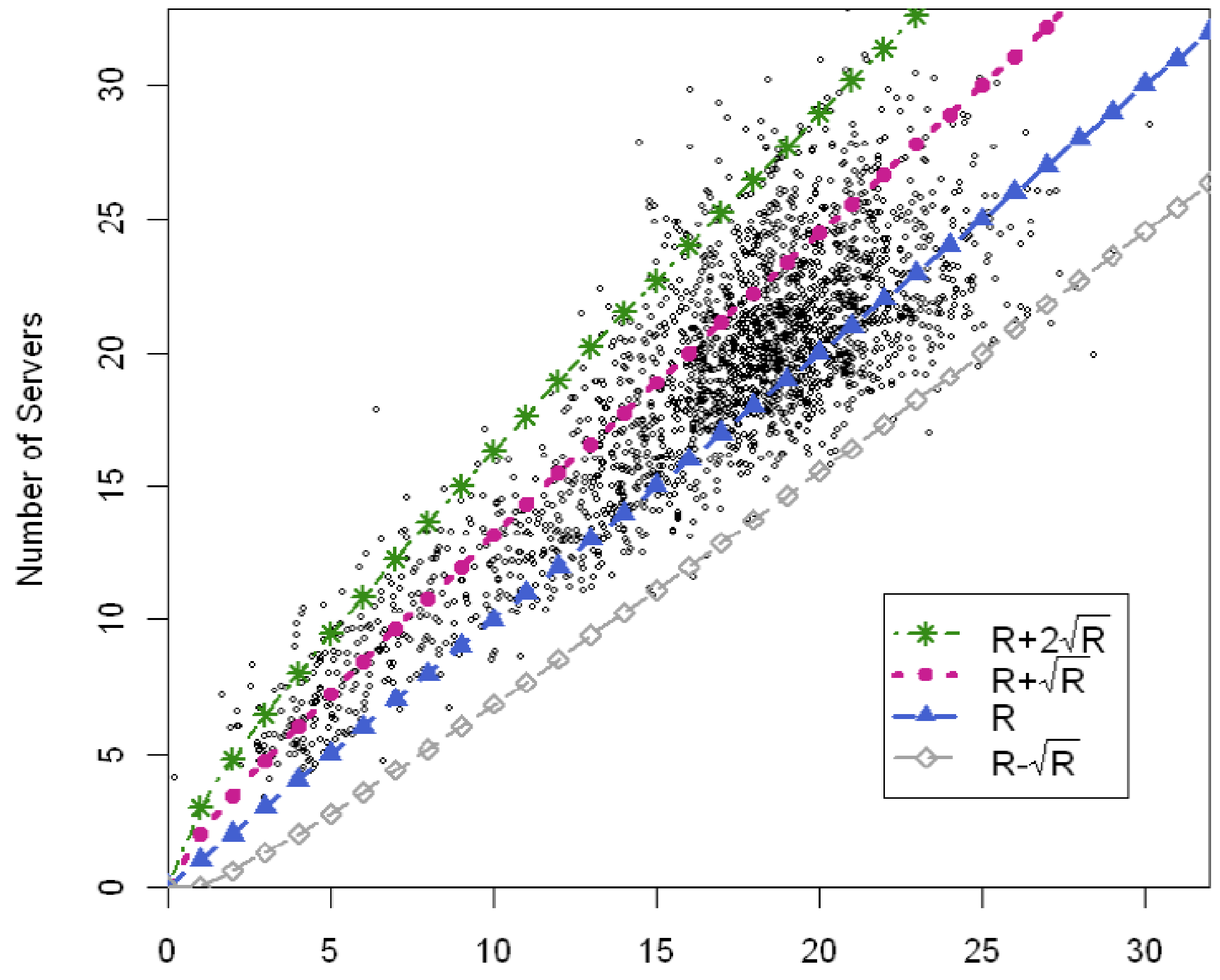
- **Agents** $\quad \text{OCC} \approx 1 - \frac{\beta+\gamma}{\sqrt{N}} \qquad -\infty < \beta < \infty;$

- **Managers** $\quad N \approx R + \beta\sqrt{R}, \quad R = \lambda \times E(S) \quad$ not small;

Here **R = Offered Load**
eg. $\quad R = 25$ call/min. $\times$ 4 min./call $= 100$

Beyond Fluid: #Agents vs. Offered-Load ($N \approx R + \beta\sqrt{R}$)

IL Telecom; June-September, 2004 (2205 30min intervals, over 13 weeks, week-days)



e.g. Offered-load R $\overset{avg}{=}$ 5 calls per min $\times$ 3.2 min per call = 16 Erlangs

Erlang-A: QED Approximations (Examples)

Assume **Offered Load** R not small ($\lambda \to \infty$).

Let $\hat{\beta} = \beta\sqrt{\frac{\mu}{\theta}}$, $h(\cdot) = \frac{\phi(\cdot)}{1 - \Phi(\cdot)}$ = hazard rate of $\mathcal{N}(0, 1)$.

- **Delay Probability:**

$$\mathrm{P}\{W_q > 0\} \approx \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} .$$

- **Probability to Abandon:**

$$\mathrm{P}\{\mathrm{Ab}|W_q > 0\} \approx \frac{1}{\sqrt{N}} \cdot \sqrt{\frac{\theta}{\mu}} \cdot \left[h(\hat{\beta}) - \hat{\beta}\right] .$$

- $\boldsymbol{P\{Ab\} \propto \mathrm{E}[W_q]}$, both order $\frac{1}{\sqrt{N}}$:

$$\frac{\mathrm{P}\{\mathrm{Ab}\}}{\mathrm{E}[W_q]} = \theta \quad (\approx g(0) > 0).$$

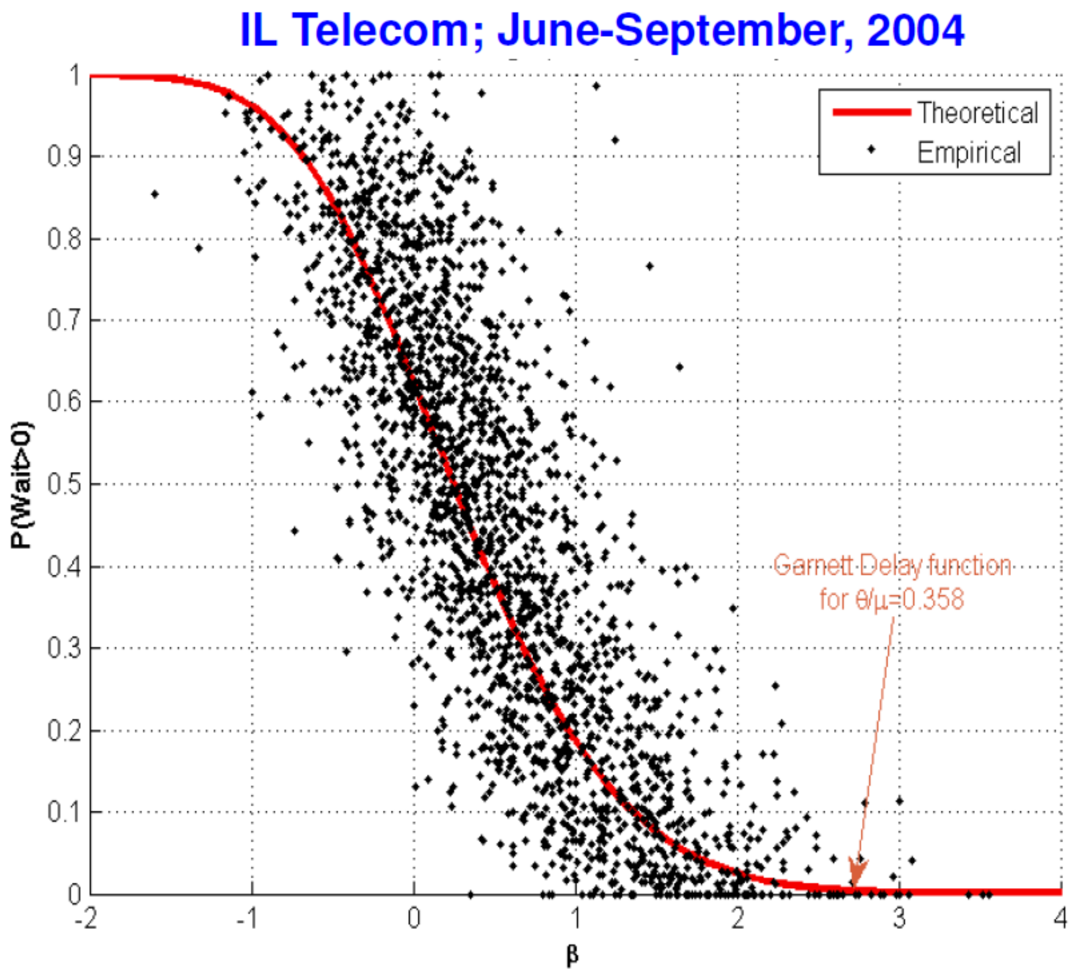
Asymptotic Landscape: 9 Operational Regimes, and then some

Erlang-A, w/ I. Gurvich & J. Huang

Erlang-A μ & θ fixed	Conventional scaling			Many-Server scaling			NDS scaling		
	Sub	Critical	Over	QD	QED	ED	Sub	Critical	Over
Offered load per server	$\frac{1}{1+\delta}$	$1-\frac{\beta}{\sqrt{n}}$	$\frac{1}{1-\gamma}$	$\frac{1}{1+\delta}$	$1-\frac{\beta}{\sqrt{n}}$	$\frac{1}{1-\gamma}$	$\frac{1}{1+\delta}$	$1-\frac{\beta}{n}$	$\frac{1}{1-\gamma}$
Arrival rate λ	$\frac{\mu}{1+\delta}$	$\mu-\frac{\beta}{\sqrt{n}}\mu$	$\frac{\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu\sqrt{n}$	$\frac{n\mu}{1-\gamma}$	$\frac{n\mu}{1+\delta}$	$n\mu-\beta\mu$	$\frac{n\mu}{1-\gamma}$
# servers	1			n			n		
Time-scale	n			1			n		
Impatience rate	θ/n			θ			θ/n		
Staffing level	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}(1+\frac{\beta}{\sqrt{n}})$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta\sqrt{\frac{\lambda}{\mu}}$	$\frac{\lambda}{\mu}(1-\gamma)$	$\frac{\lambda}{\mu}(1+\delta)$	$\frac{\lambda}{\mu}+\beta$	$\frac{\lambda}{\mu}(1-\gamma)$
Utilization	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$	1	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{\sqrt{n}}$	1	$\frac{1}{1+\delta}$	$1-\sqrt{\frac{\theta}{\mu}}\frac{h(\hat{\beta})}{n}$	1
$\mathbb{E}(Q)$	$\frac{1}{\delta(1+\delta)}$	$\sqrt{n}g(\hat{\beta})$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$\frac{1}{\delta}\varrho_n$	$\sqrt{n}g(\hat{\beta})\alpha$	$\frac{n\mu\gamma}{\theta(1-\gamma)}$	$o(1)$	$ng(\hat{\beta})$	$\frac{n^2\mu\gamma}{\theta(1-\gamma)}$
$\mathbb{P}(Ab)$	$\frac{1}{n}\frac{1}{\delta}\frac{\theta}{\mu}$	$\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})$	γ	$\frac{1}{n}\frac{(1+\delta)}{\delta}\frac{\theta}{\mu}\varrho_n$	$\frac{\theta}{\sqrt{n}\mu}g(\hat{\beta})\alpha$	γ	$o(\frac{1}{n^2})$	$\frac{\theta}{n\mu}g(\hat{\beta})$	γ
$\mathbb{P}(W_q>0)$	$\frac{1}{1+\delta}$	≈ 1		ϱ_n	$\alpha\in(0,1)$	≈ 1	≈ 0	≈ 1	
$\mathbb{P}(W_q>T)$	$\frac{1}{1+\delta}e^{-\frac{\delta}{1+\delta}\mu T}$	$1+O(\frac{1}{\sqrt{n}})$	$1+O(\frac{1}{n})$	≈ 0		$f(T)$	≈ 0	$\frac{\bar{\Phi}(\hat{\beta}+\sqrt{\theta\mu}T)}{\bar{\Phi}(\hat{\beta})}$	$1+O(\frac{1}{n})$
Congestion $\frac{EW_q}{ES}$	$\frac{1}{\delta}$	$\sqrt{n}g(\hat{\beta})$	$n\mu\gamma/\theta$	$\frac{1}{n}\frac{(1+\delta)}{\delta}\varrho_n$	$\frac{\alpha}{\sqrt{n}}g(\hat{\beta})$	$\frac{\mu\gamma}{\theta}$	$o(\frac{1}{n})$	$g(\hat{\beta})$	$n\mu\gamma/\theta$

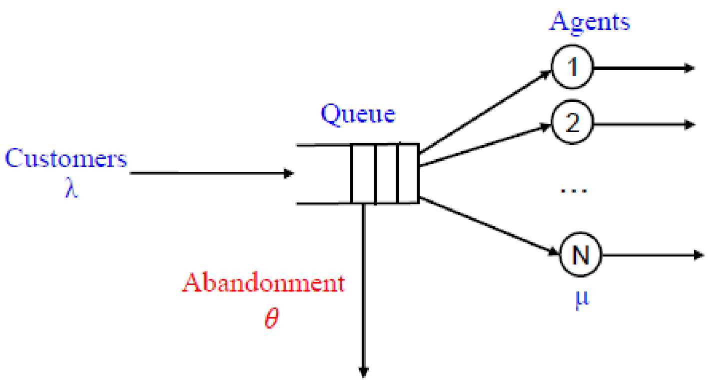
- Conventional: Ward & Glynn (03, $G/G/1+G$)
- Many-Server:
  - QED: Halfin-Whitt (81), w/ Garnett & Reiman (02)
  - ED: Whitt (04)
  - NDS: Atar (12)
- **“Missing"**: ED+QED; Hazard-rate scaling (M/M/N+G); Time-Varying, Non-Parametric; Moderate- and Large-Deviation; Networks (multi-regimes)

$P(W_q > 0)$, or "Universalizing" the QED regime



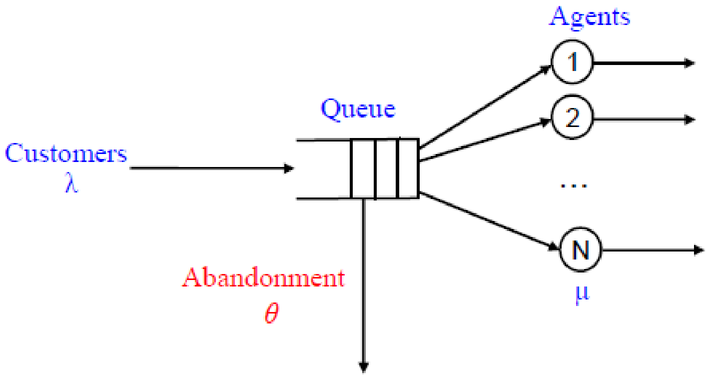
- 2205 half-hour intervals (13 summer weeks, week-days)
- Erlang-A approximations for the appropriate $\mu/\theta \approx 3$

Universal Approximations: Erlang-A (M/M/N + M)



w/ **I. Gurvich & J. Huang**

Universal Approximations: Erlang-A (M/M/N + M)

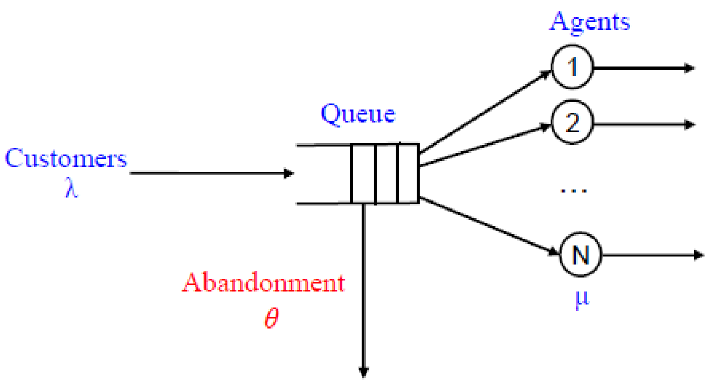


w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

Universal Approximations: Erlang-A (M/M/N + M)



w/ **I. Gurvich & J. Huang**

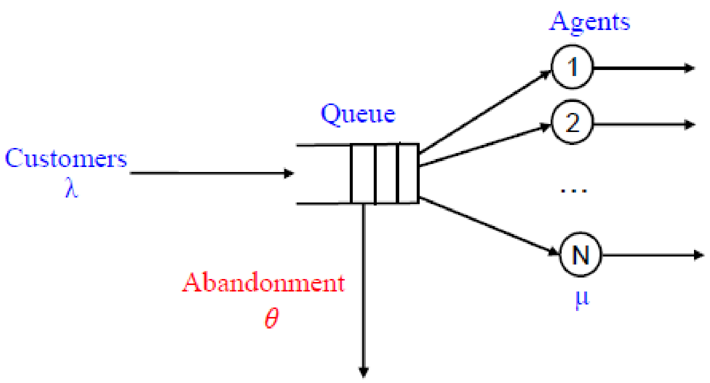
- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

Universal Approximations: Erlang-A (M/M/N + M)



w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^{+}, \quad q = 0, 1, \ldots$$

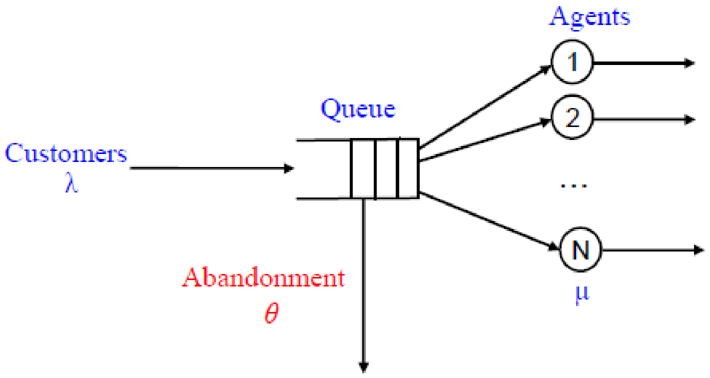
- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

- **DNet**: Universal (Stochastic) Approximation – SDE

$$dY_t = F(Y_t)dt + \sqrt{2\lambda}\, dB_t, \quad t \geq 0$$

Universal Approximations: Erlang-A (M/M/N + M)



w/ **I. Gurvich & J. Huang**

- **QNet**: Birth & Death Queue, with B - D rates

$$F(q) = \lambda - \mu \cdot (q \wedge n) - \theta \cdot (q - n)^+, \quad q = 0, 1, \ldots$$

- **FNet**: Dynamical (Deterministic) System – ODE

$$dx_t = F(x_t)dt, \quad t \geq 0$$

- **DNet**: Universal (Stochastic) Approximation – SDE

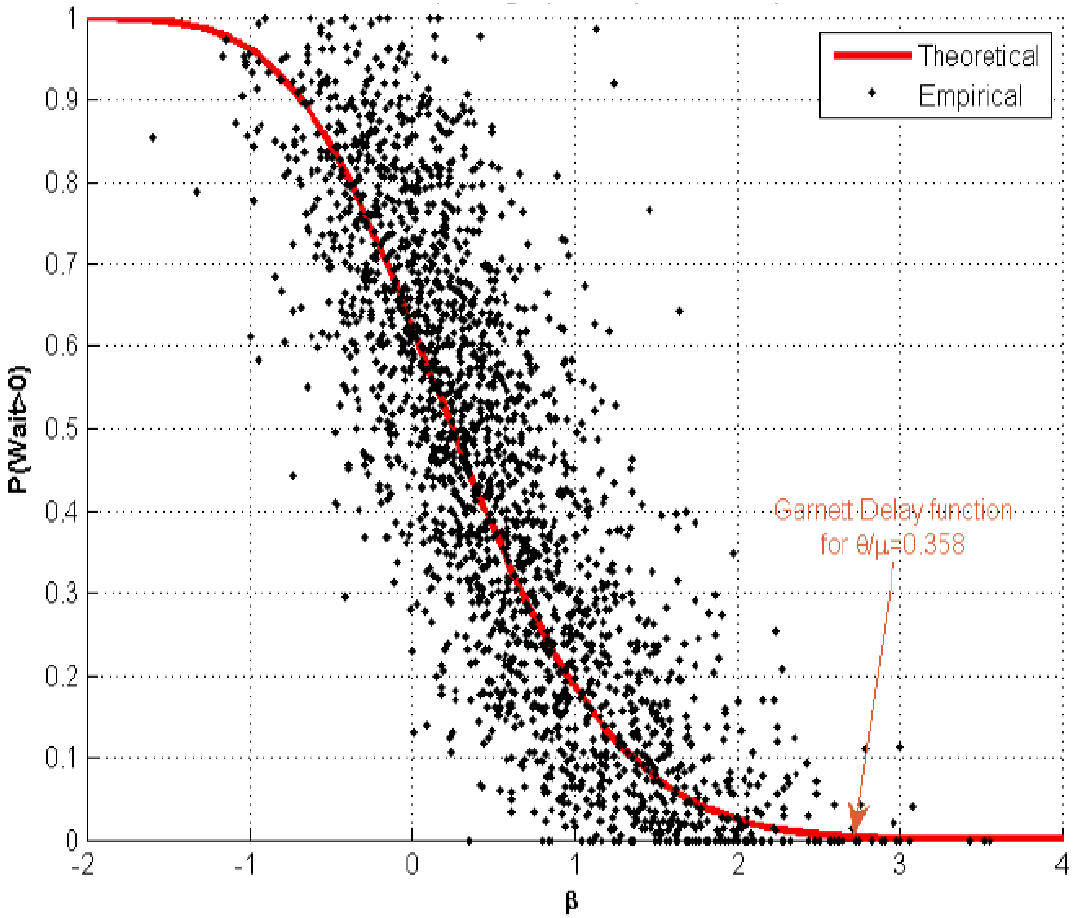
$$dY_t = F(Y_t)dt + \sqrt{2\lambda}\, dB_t, \quad t \geq 0$$

eg. $\mu = \theta : \quad \dot{x} = \lambda - \mu \cdot x, \quad Y =$ OU process

Parsimonious (Tractable, Robust), Accurate, Valuable

Erlang-A Value: DNet $P(W_q > 0)$ vs. Data

IL Telecom; June-September, 2004 (2205 30min intervals, weekdays)



- Approximations, w/ Patience $\approx 3\times$ Service-Duration ($\mu/\theta \approx 3$)

Accuracy: DNet vs. QNet

- Δ^λ is the "balancing" state, obtained by solving

$$\lambda = \mu(n \wedge \Delta^\lambda) + \theta(\Delta^\lambda - n)^+.$$

Solution: $\Delta^\lambda = \frac{\lambda}{\mu} - \left(\frac{\lambda}{\mu} - n\right)^+ \left(1 - \frac{\mu}{\theta}\right).$

Specifically: **QD** = $\frac{\lambda}{\mu}$; **ED** = $n + \frac{1}{\theta}(\lambda - n\mu)$; **QED** = $n + \mathcal{O}(\sqrt{\lambda})$)

- Centered processes (excursions):

$$\tilde{Q}^\lambda(\cdot) = Q(\cdot) - \Delta^\lambda, \quad \tilde{Y}^\lambda(\cdot) = Y(\cdot) - \Delta^\lambda.$$

Theorem: For f bounded by an m-degree polynomial ($m \geq 0$):

$$\mathbb{E}f(\tilde{Q}^\lambda(\infty)) - \mathbb{E}f(\tilde{Y}^\lambda(\infty)) = \mathcal{O}(\sqrt{\lambda}^{m-1}).$$

- **Accurate**: more than heavy-traffic *limits*

Simplicity: Why 2λ?

- Semi-martingale representation of the B&D process:
  Fluid + Martingale
- Predictable quadratic variation:

$$\int_0^t [\lambda + \mu(Q_s \wedge n) + \theta(Q_s - n)^+] ds$$

- In steady-state, arrival rate $\equiv$ departure rate:

$$\lambda = \mathbb{E}[\mu(Q_s \wedge n) + \theta(Q_s - n)^+]$$

- Expectation of the predictable quadratic variation:

$$\mathbb{E} \int_0^t [\lambda + \mu(Q_s \wedge n) + \theta(Q_s - n)^+] ds = 2\lambda t$$

- **Simple** $\Rightarrow$ **Tractable, Robust**: $\text{dMartingale}_t \approx \sqrt{2\lambda} \cdot \text{dBrownian}_t$

Reconciling Time-Varying and Steady-State Models

- **Rigid** (fixed) staffing level during a time-varying shift:
  Doomed to alternate between overloading and underloading
- **Flexible** staffing:
  Can design **time-varying staffing** that achieves, **at all times, Steady-State performance**
  via Square-Root Staffing (Modified Offered-Load)

Reconciling Time-Varying and Steady-State Models

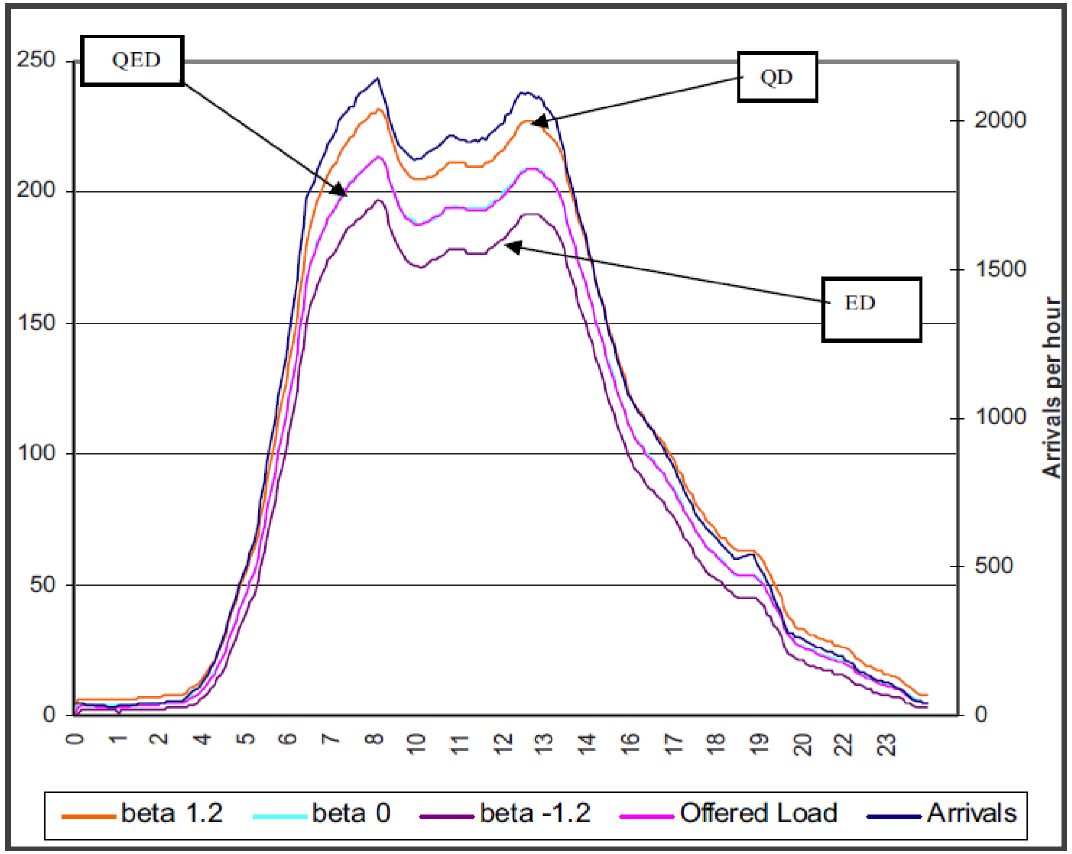
- **Rigid** (fixed) staffing level during a time-varying shift:
  Doomed to alternate between overloading and underloading
- **Flexible** staffing:
  Can design **time-varying staffing** that achieves, **at all times, Steady-State performance**
  via Square-Root Staffing (Modified Offered-Load)
- **History**:
  - Jennings, M., Reiman, Whitt (1996): Emergence of the phenomenon, with infinite-server heuristics
  - Feldman, M., Massey, Whitt (2008): Stabilize delay probability with QED staffing, with little theory
  - Liu and Whitt (2012): Stabilize abandonment probability, with ED theory
  - w/ Huang, Gurvich (ongoing): QED theory

Why Does Erlang-A Work? Time-Varying Arrival Rates

Square-Root Staffing: $N_t = R_t + \beta\sqrt{R_t}\,, \quad -\infty < \beta < \infty$

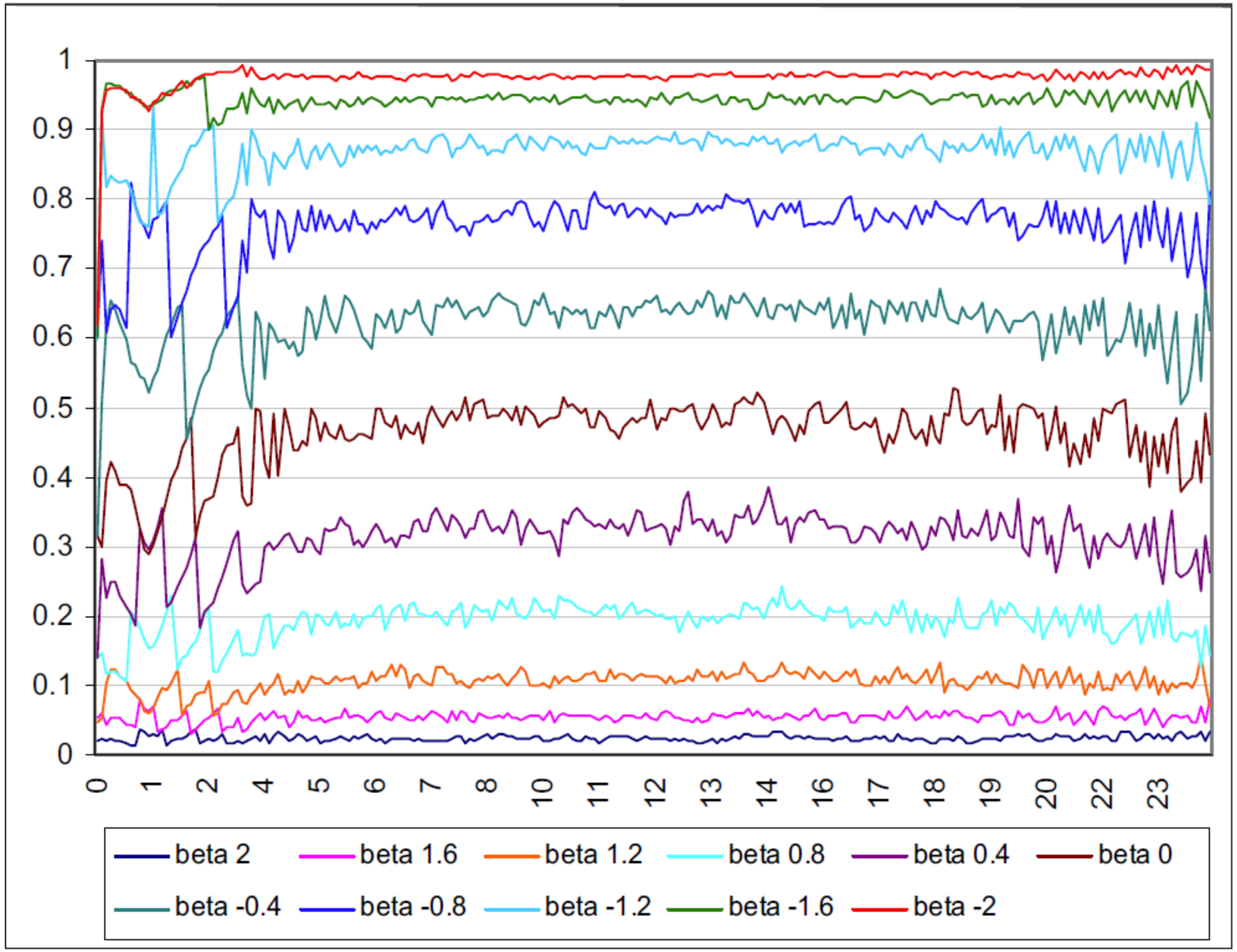
What is R_t, the **Offered-Load** at time t ? $\quad (R_t \neq \lambda_t \times E[S])$

Arrivals, Offered-Load and Staffing



Time-Stable Performance of Time-Varying Systems

Delay Probability = As in the **Stationary Erlang-A** (Garnett)



Calculating the Offered-Load $R(t)$, Theoretically

- Offered-Load Process: $\boldsymbol{L(\cdot)}$ = **Least** number of **servers** that guarantees **no delay**.
- **Offered-Load** Function $\boldsymbol{R(t)} = E[L(t)]$, $t \geq 0$.
  Think $M_t/G/N_t^? + G$ vs. $\boldsymbol{M_t/G/\infty}$: **Ample-Servers**.

Four (all useful) representations, capturing "**workload before t**":

$$R(t) = E[L(t)] = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t - u)du = E\Big[A(t) - A(t - S)\Big] =$$

$$= E\left[\int_{t-S}^{t} \lambda(u)du\right] = E[\lambda(t - S_e)] \cdot E[S] \approx \dots .$$

- $\{\boldsymbol{A(t)}, t \geq 0\}$ Arrival-Process, rate $\boldsymbol{\lambda(\cdot)}$;
- $\boldsymbol{S}$ ($\boldsymbol{S_e}$) generic Service-Time (Residual Service-Time).
- Relating L, λ, S ("W"): **Time-Varying Little's Formula**.
  Stationary models: $\lambda(t) \equiv \lambda$ then $R(t) \equiv \boldsymbol{\lambda} \times \boldsymbol{E[S]}$.

Abandonment: Further Applications

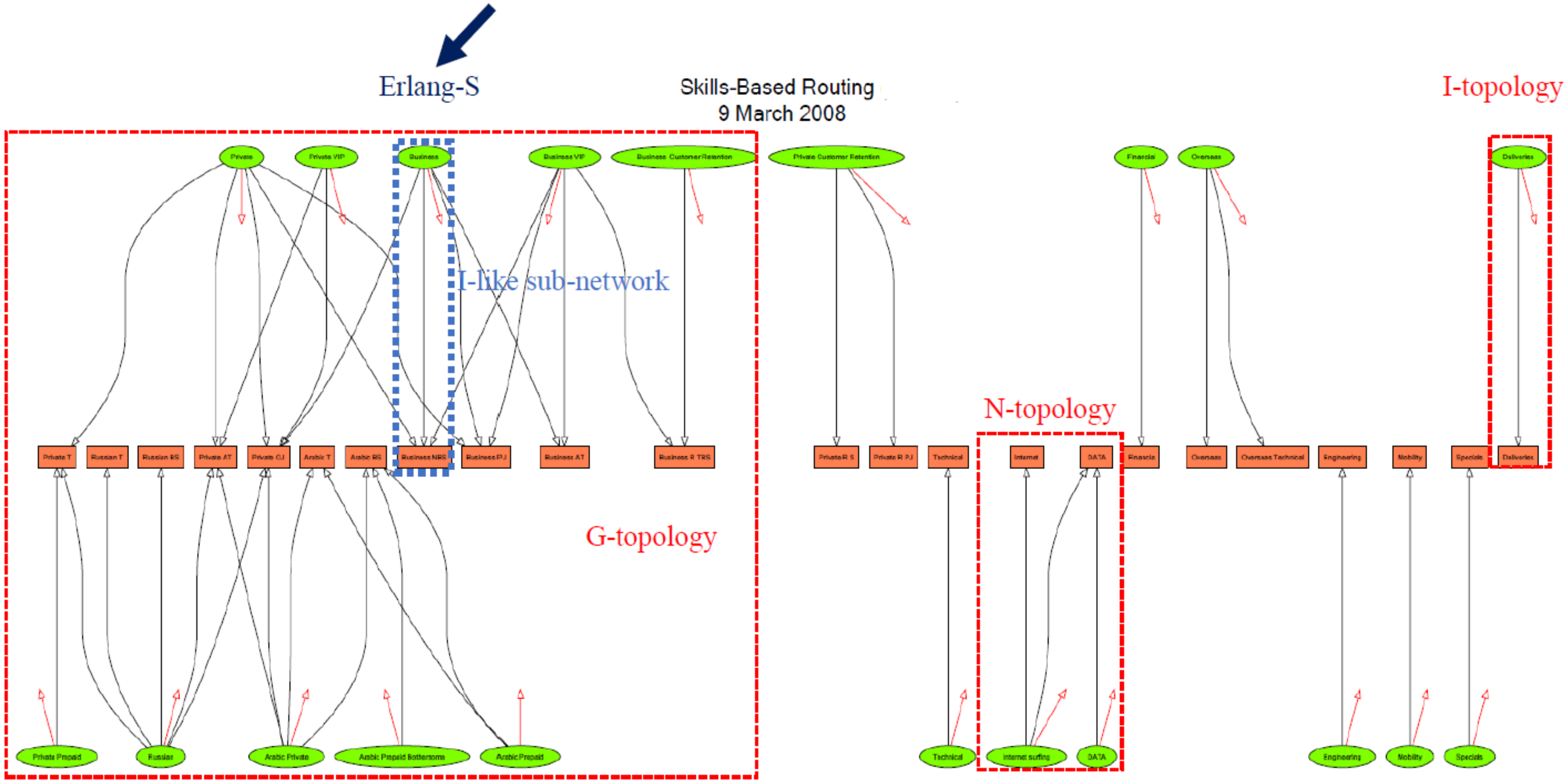
- **Personalized Queueing**: e.g. when Impatience has been estimated (personalized), exactly or approximately, choosing who to serve next will benefit from taking this information into account (e.g. Shortest-Patience-First); **w/ P. Momcilovic**
- **CCs:** Control of congestion via announcement / information that monitors the queue**, w/ Junfei Huang, Hanqin Zhang, Jiheng Zhang**
- **IVRs** (Self-Services): identify via Mixture Fitting, **w/ N. Carmeli, H. Kaspi**
- **EDs**: Left-Without-Being-Seen. Theory exists if All-Unknown – Current-Status Data. Otherwise, namely when having both Right-Censoring and Current-Status data, theory developed **w/ Y. Yefenof, Y. Goldberg, Y. Ritov**
- **Chats:** Before or within session (registered vs. silent abandonment)

Laws & Models: Data-Based Erlang A/R/S, following B/C

- **Little's Law (Steady-State, Transient), State-Space Collapse,…**
- **Erlang-B (Blocking) and Erlang-C (?)**
  - **Erlang**, Agner Krarup: Queueing Theory was born in 1909, in his paper "The Theory of Probabilities and **Telephone Conversations**“
- **2. Erlang-A**
  - **Abandonment**: While waiting for service, does service-value dominate residual-wait-cost
- **1. Erlang-R**
  - **Return/Feedback**: Customers often return to service (positively, negatively, just needing)
- **3. Erlang-S**
  - **Servers**: Challenging to manage, and model, no less so than customers

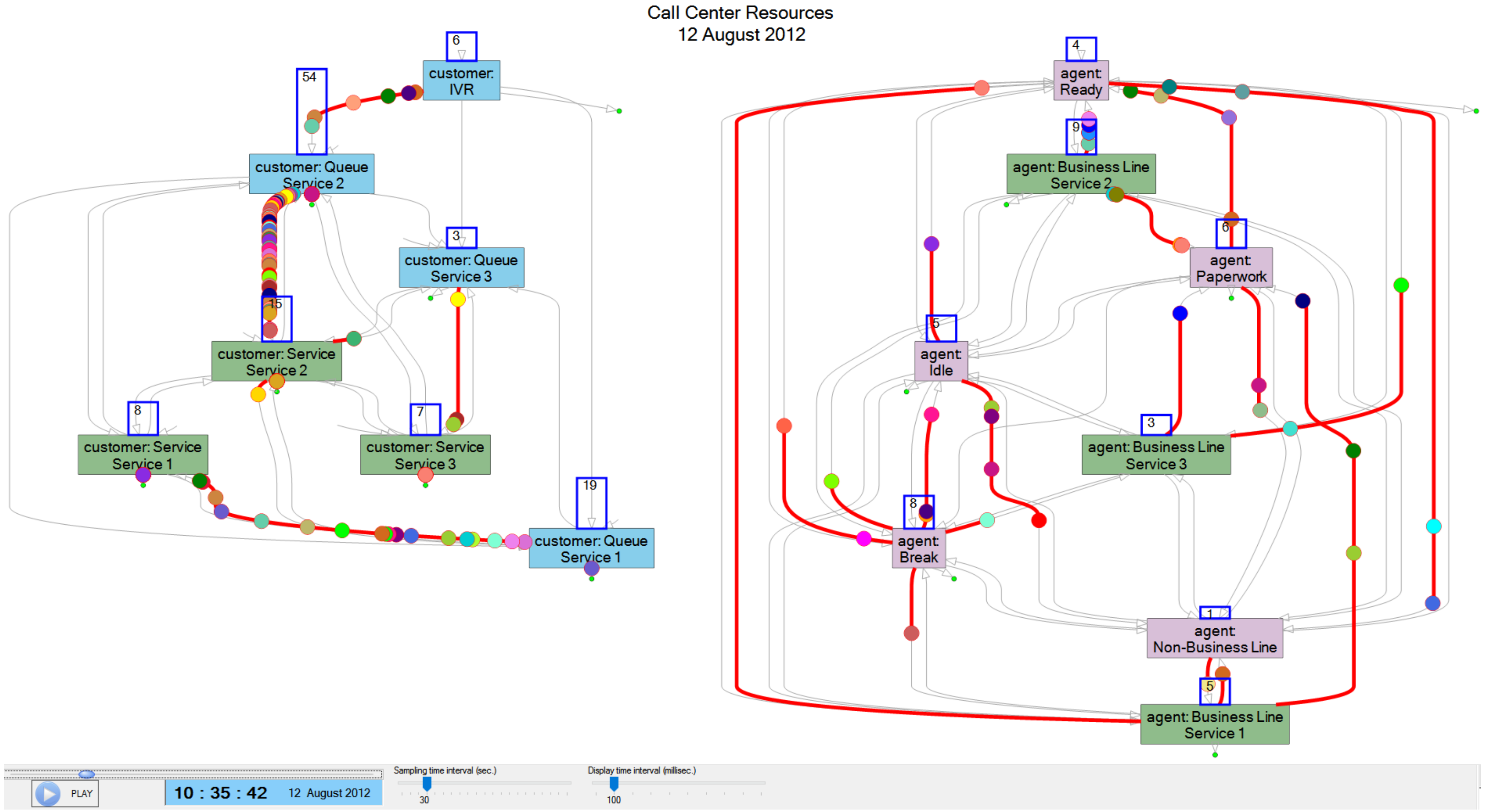
Above: Simple (Parsimonious) models of complex realities, yet not too simple (Robust)

Erlang-S: Servers in Qnets – Aggregate or Zoom   w/ D. Azriel, P. Feigin

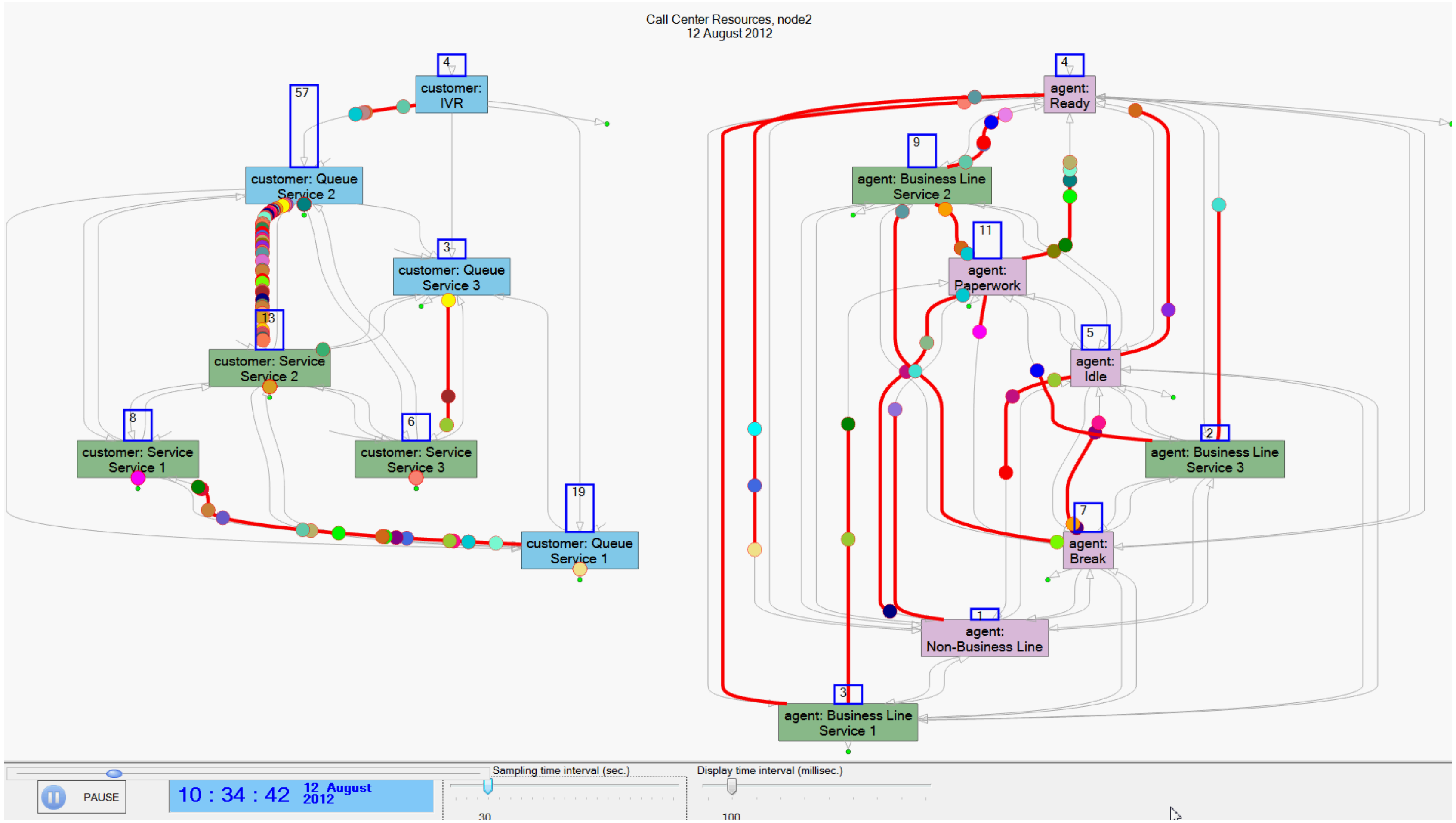


Topology of a call center:
Server-queues are in the rectangles and customer-queues are in the ovals

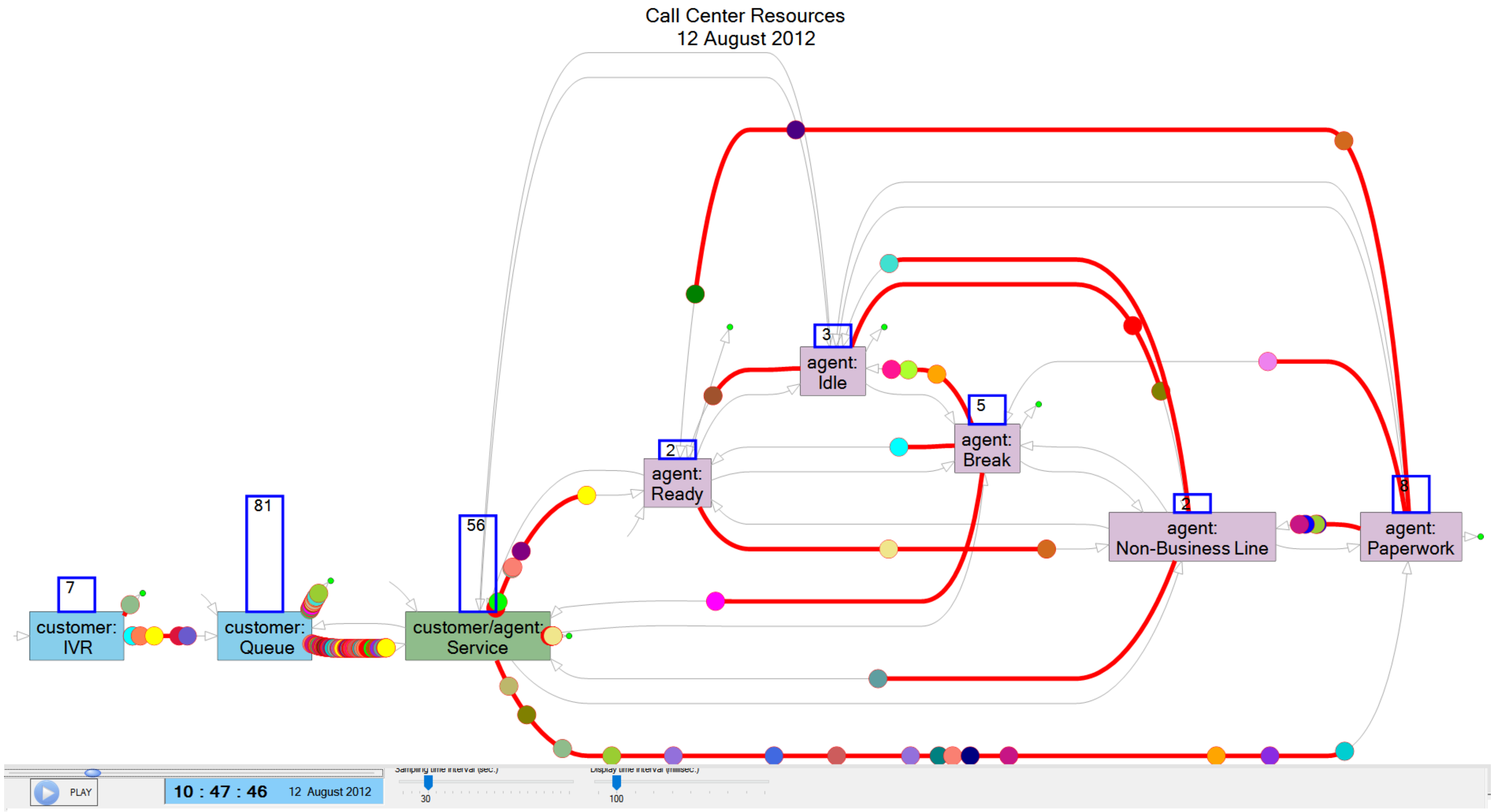
Telephone Services: Customers and Servers - Symmetric Viewpoint



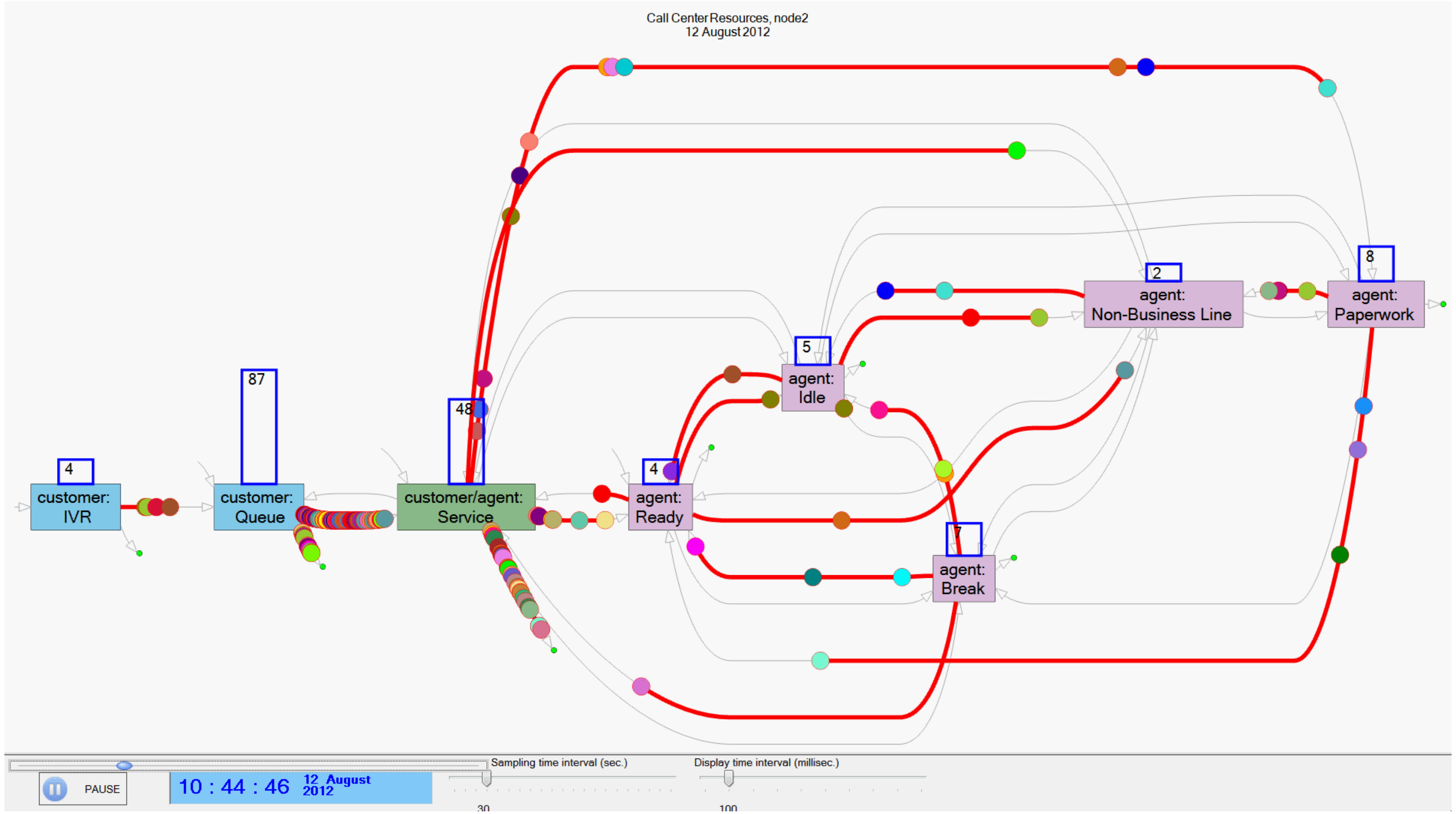
Telephone Services: Customers and Servers Symmetric Viewpoint



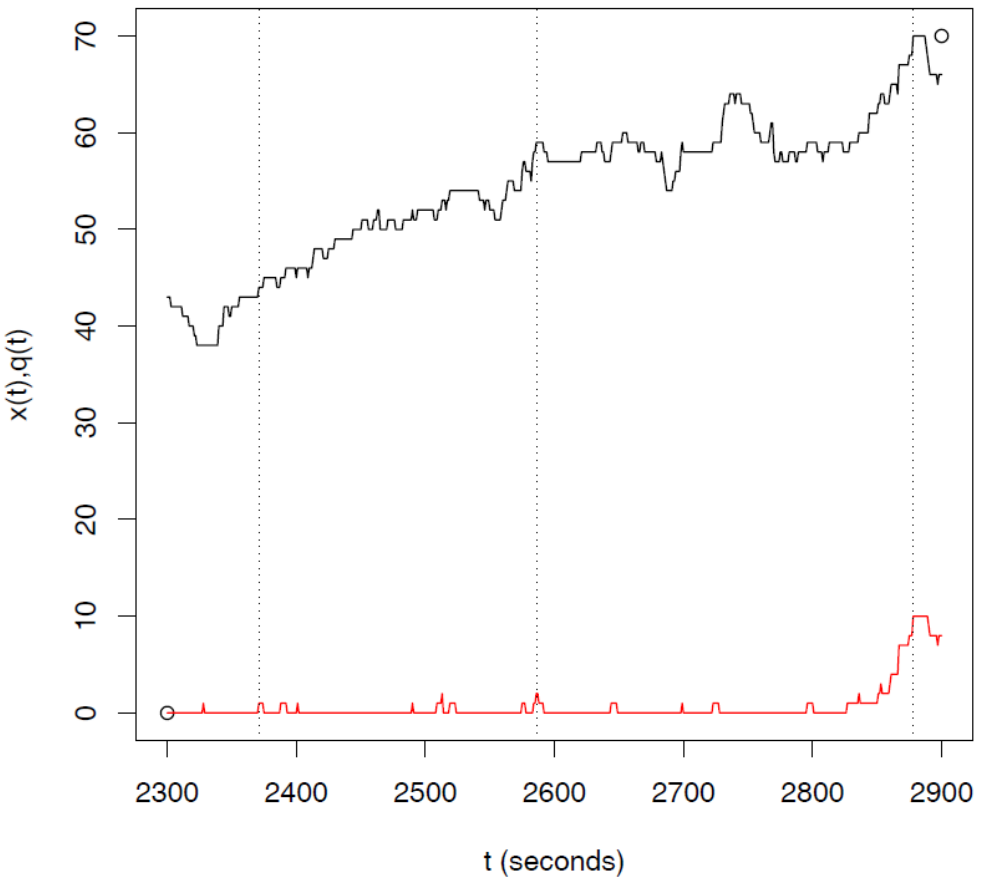
Telephone Services: Customers and Servers – Erlang-S, or **Resource-Driven Activity Networks**



Telephone Services: Customers and Servers - Erlang-S, or Resource-Driven Activity Networks



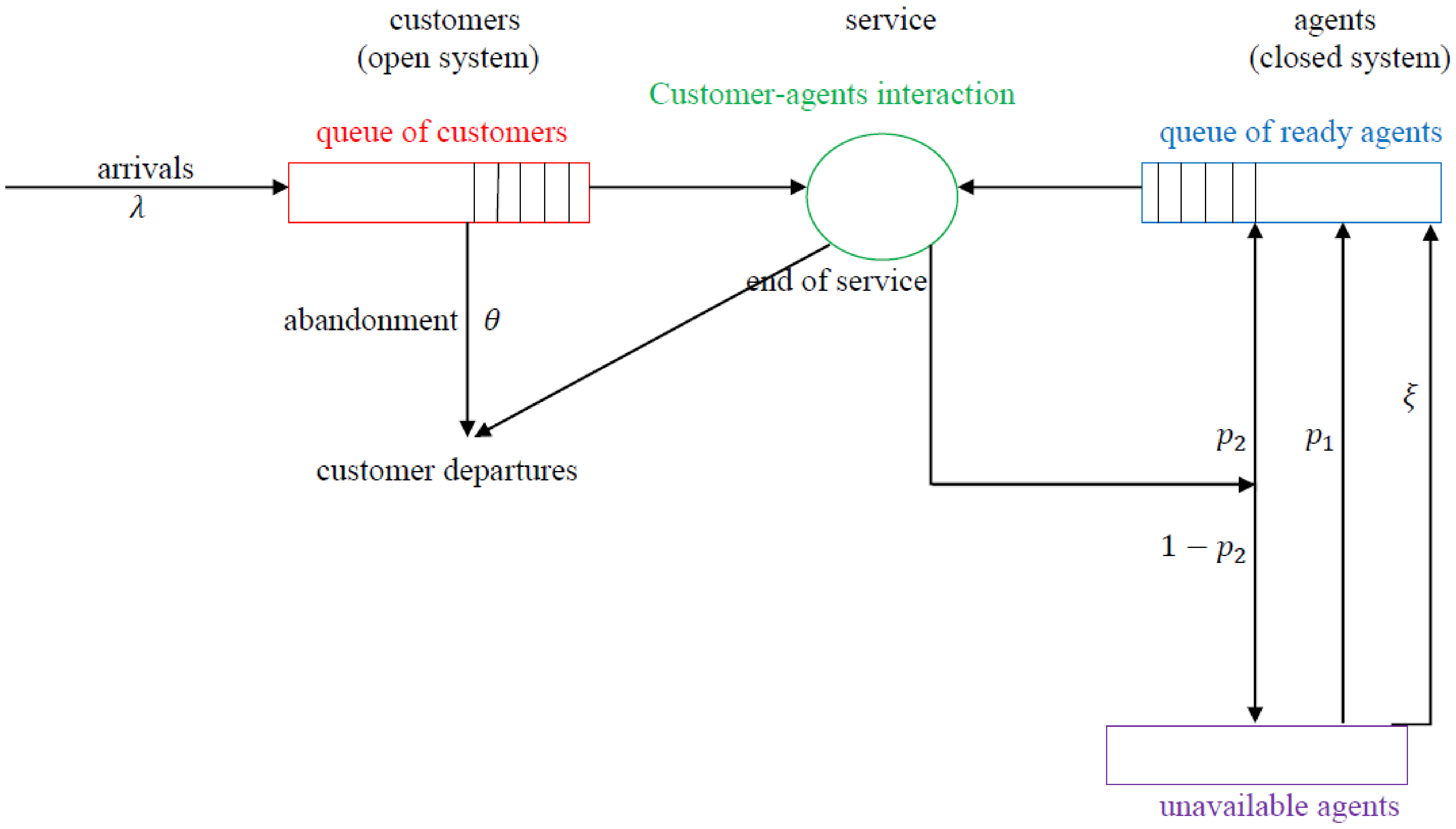
Number of Servers: Present (=Constant) vs. Available (Random)



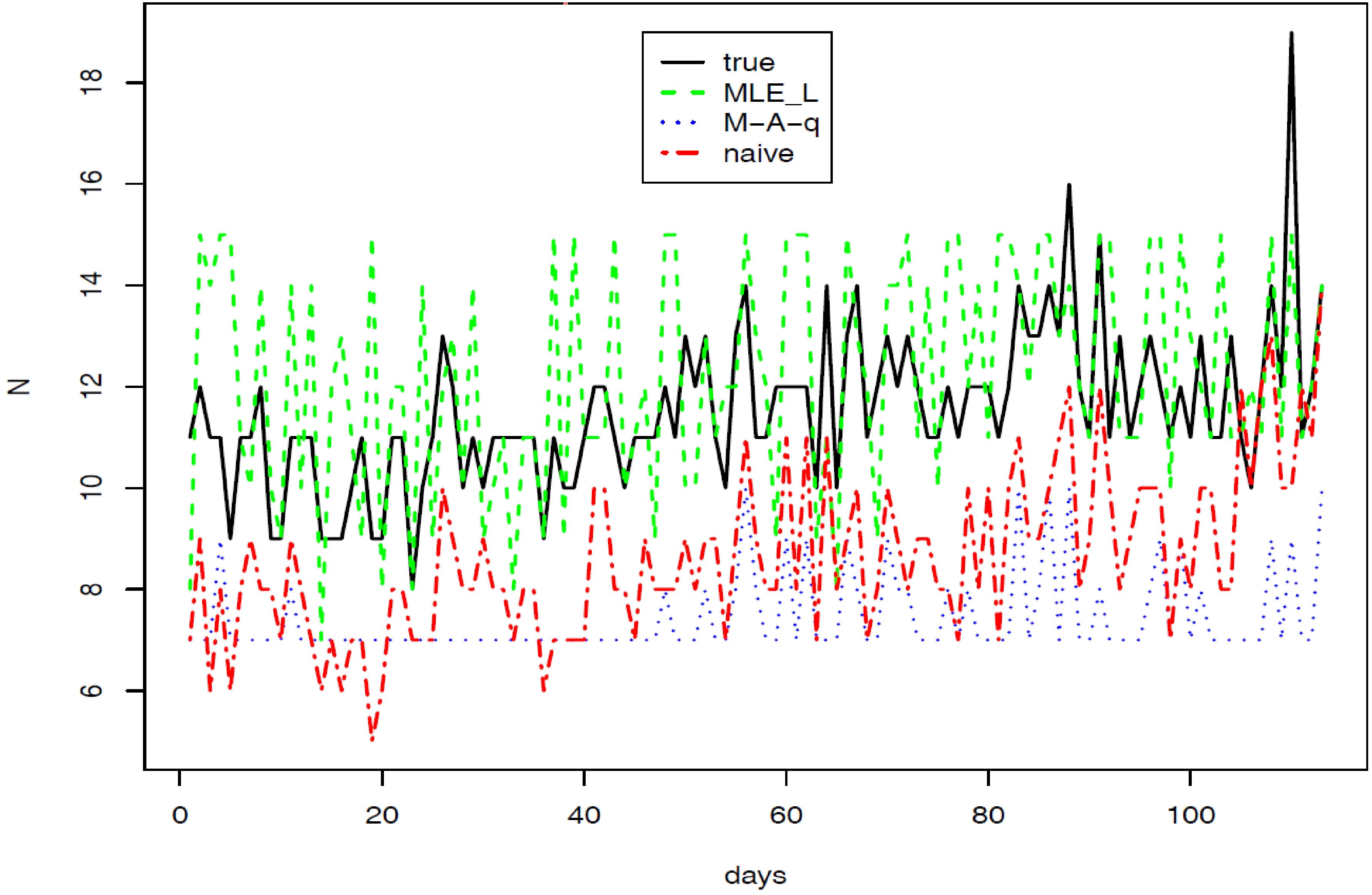
x(t), q(t) in real data (U.S. Bank; telesales; 12/1/2002)

n(t) = (43,57,60) @*t* = (10:36:54, 10:43:07, 10:47:23):
17 agents became available within **11** minutes

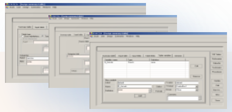
Erlang-S w/ David Azriel & Paul Feigin



Estimating the Number of Present Servers
2/1/2005 – 27/6/2005

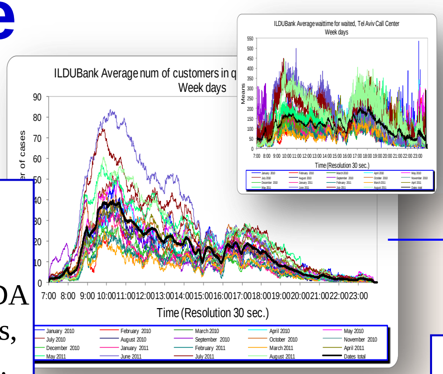


SEELab: Research, Teaching, Practice



SEE Summary Tables
- Cross Tabulations
- Time Series

SEEStat
- Universal Design
- Online & RealTime EDA
- Tools: Fitting, Mixtures, Smoothing, Survival,…



SEE Server
- Free for Academic Use
- Register + Open Access (U.S. Bank, Bank Anonymous, Home Hospital)

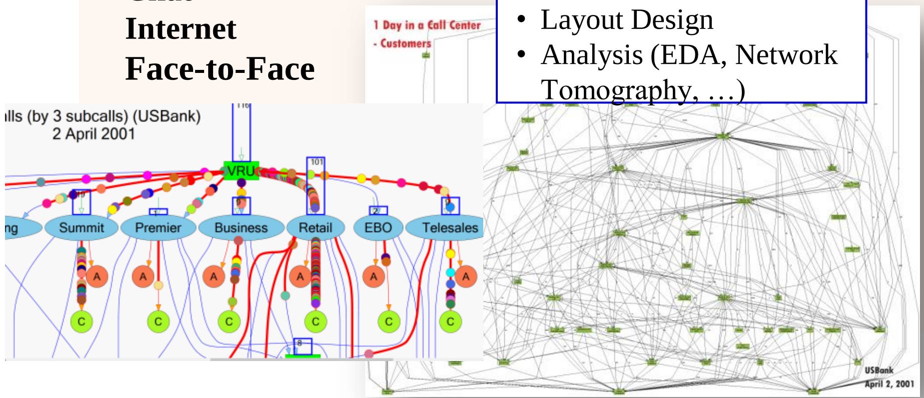
Research
- Theoretical, Empirical
- OR/OM/IE/CS/Process Mining
- Graduate & honor students
- OCR (IBM+Rambam+Technion)
- Hospital: RTLS+Appointments
- Operations+Emotions (e.g. Chat)
- Bank Warehouse: in planning

Data Source
- ACD, CRM, RTLS, Clicks…

Call-Centers
Hospitals
Banks
Chat
Internet
Face-to-Face

SEE Data
- Cleaning
- Universal Schema

SEEGraph/Nimations
- Structural Inference
- Layout Design
- Analysis (EDA, Network Tomography, …)



Teaching/Mentoring
- Service Engineering Course
- Mini-courses on Service Engineering at Stanford, Columbia, Wharton, NUS, HKUST, …
- Training routines (basic, data-bases, graphics)
- Data for other courses: Data-Mining, Time-Series,…
- Undergraduate & graduate Projects

SEE (International) Outreach
- Hosting PhDs/Scholars
- Workshops (MSOM,…)
- Mirror Servers: HKUST,…
- Scholars Worldwide

Practice
- Design and improve data collection
- Hosting data partners: workshops, data sessions
- Open Bi-Directional Channel

Data Cleaning: MCE with RFID Support

Data-base				Company report		comment
Asset id	order	Entry date	Exit date	Entry date	Exit date	
4	1	1:14:07 PM		1:14:00 PM		
6	1	12:02:02 PM	12:33:10 PM	12:02:00 PM	12:33:00 PM	
8	1	11:37:15 AM	12:40:17 PM	11:37:00 AM		exit is missing
10	1	12:23:32 PM	12:38:23 PM	12:23:00 PM		
12	1	12:12:47 PM	12:35:33 PM		12:35:00 PM	entry is missing
15	1	1:07:15 PM		1:07:00 PM		
16	1	11:18:19 AM	11:31:04 AM	11:18:00 AM	11:31:00 AM	
17	1	1:03:31 PM		1:03:00 PM		
18	1	1:07:54 PM		1:07:00 PM		
19	1	12:01:58 PM		12:01:00 PM		
20	1	11:37:21 AM	12:57:02 PM	11:37:00 AM	12:57:00 PM	
21	1	12:01:16 PM	12:37:16 PM	12:01:00 PM		
22	1	12:04:31 PM	12:20:40 PM			first customer is missing
22	2	12:27:37 PM		12:27:00 PM		
25	1	12:27:35 PM	1:07:28 PM	12:27:00 PM	1:07:00 PM	
27	1	12:06:53 PM		12:06:00 PM		
28	1	11:21:34 AM	11:41:06 AM	11:41:00 AM	11:53:00 AM	exit time instead of entry time
29	1	12:21:06 PM	12:54:29 PM	12:21:00 PM	12:54:00 PM	
31	1	11:40:54 AM	12:30:16 PM	11:40:00 AM	12:30:00 PM	
31	2	12:37:57 PM	12:54:51 PM	12:37:00 PM	12:54:00 PM	
32	1	11:27:11 AM	12:15:17 PM	11:27:00 AM	12:15:00 PM	
33	1	12:05:50 PM	12:13:12 PM	12:05:00 PM	12:15:00 PM	wrong exit time
35	1	11:31:48 AM	11:40:50 AM	11:31:00 AM	11:40:00 AM	
36	1	12:06:23 PM	12:29:30 PM	12:06:00 PM	12:29:00 PM	
37	1	11:31:50 AM	11:48:18 AM	11:31:00 AM	11:48:00 AM	
37	2	12:59:21 PM		12:59:00 PM		

- Imagine **"Cleaning" 60,000+ customers per day** (call centers) !
- "Psychology" of Data Trust and Transfer (e.g. 2 years till transfer)

Event-Logs in a Call Center (Bank Anonymous)

A Data Sample (Excel worksheet)

vru+line	call_id	customer_id	priority	type	date	vru_entry	vru_exit	vru_time	q_start	q_exit	q_time	outcome	ser_start	ser_exit	ser_time	server
AA0101	44749	27644400	2	PS	990901	11:45:33	11:45:39	6	11:45:39	11:46:58	79	AGENT	11:46:57	11:51:00	243	DORIT
AA0101	44750	12887816	1	PS	990905	14:49:00	14:49:06	6	14:49:06	14:53:00	234	AGENT	14:52:59	14:54:29	90	ROTH
AA0101	44967	58660291	2	PS	990905	14:58:42	14:58:48	6	14:58:48	15:02:31	223	AGENT	15:02:31	15:04:10	99	ROTH
AA0101	44968	0	0	NW	990905	15:10:17	15:10:26	9	15:10:26	15:13:19	173	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44969	63193346	2	PS	990905	15:22:07	15:22:13	6	15:22:13	15:23:21	68	AGENT	15:23:20	15:25:25	125	STEREN
AA0101	44970	0	0	NW	990905	15:31:33	15:31:47	14	00:00:00	00:00:00	0	AGENT	15:31:45	15:34:16	151	STEREN
AA0101	44971	41630443	2	PS	990905	15:37:29	15:37:34	5	15:37:34	15:38:20	46	AGENT	15:38:18	15:40:56	158	TOVA
AA0101	44972	64185333	2	PS	990905	15:44:32	15:44:37	5	15:44:37	15:47:57	200	AGENT	15:47:56	15:49:02	66	TOVA
AA0101	44973	3.06E+08	1	PS	990905	15:53:05	15:53:11	6	15:53:11	15:56:39	208	AGENT	15:56:38	15:56:47	9	MORIAH
AA0101	44974	74780917	2	NE	990905	15:59:34	15:59:40	6	15:59:40	16:02:33	173	AGENT	16:02:33	16:26:04	1411	ELI
AA0101	44975	55920755	2	PS	990905	16:07:46	16:07:51	5	16:07:51	16:08:01	10	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44976	0	0	NW	990905	16:11:38	16:11:48	10	16:11:48	16:11:50	2	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44977	33689787	2	PS	990905	16:14:27	16:14:33	6	16:14:33	16:14:54	21	HANG	00:00:00	00:00:00	0	NO_SERVER
AA0101	44978	23817067	2	PS	990905	16:19:11	16:19:17	6	16:19:17	16:19:39	22	AGENT	16:19:38	16:21:57	139	TOVA
AA0101	44764	0	0	PS	990901	15:03:26	15:03:36	10	00:00:00	00:00:00	0	AGENT	15:03:35	15:06:36	181	ZOHARI
AA0101	44765	25219700	2	PS	990901	15:14:46	15:14:51	5	15:14:51	15:15:10	19	AGENT	15:15:09	15:17:00	111	SHARON
AA0101	44766	0	0	PS	990901	15:25:48	15:26:00	12	00:00:00	00:00:00	0	AGENT	15:25:59	15:28:15	136	ANAT
AA0101	44767	58859752	2	PS	990901	15:34:57	15:35:03	6	15:35:03	15:35:14	11	AGENT	15:35:13	15:35:15	2	MORIAH
AA0101	44768	0	0	PS	990901	15:46:30	15:46:39	9	00:00:00	00:00:00	0	AGENT	15:46:38	15:51:51	313	ANAT
AA0101	44769	78191137	2	PS	990901	15:56:03	15:56:09	6	15:56:09	15:56:28	19	AGENT	15:56:28	15:59:02	154	MORIAH
AA0101	44770	0	0	PS	990901	16:14:31	16:14:46	15	00:00:00	00:00:00	0	AGENT	16:14:44	16:16:02	78	BENSION
AA0101	44771	0	0	PS	990901	16:38:59	16:39:12	13	00:00:00	00:00:00	0	AGENT	16:39:11	16:43:35	264	VICKY
AA0101	44772	0	0	PS	990901	16:51:40	16:51:50	10	00:00:00	00:00:00	0	AGENT	16:51:49	16:53:52	123	ANAT
AA0101	44773	0	0	PS	990901	17:02:19	17:02:28	9	00:00:00	00:00:00	0	AGENT	17:02:28	17:07:42	314	VICKY
AA0101	44774	32387482	1	PS	990901	17:18:18	17:18:24	6	17:18:24	17:19:01	37	AGENT	17:19:00	17:19:35	35	VICKY
AA0101	44775	0	0	PS	990901	17:38:53	17:39:05	12	00:00:00	00:00:00	0	AGENT	17:39:04	17:40:43	99	TOVA

- Unsynchronized transition times, consistently

SEELab: Environment for graphical EDA

Operational histories (customers, servers) at the **individual-transaction level**, e.g.

1. *Bank Anonymous Call-Center: **1 year, 350K calls by 15 agents** - in 2000, **which paved the way to:**
2. *U.S. Bank Call-Center : **2.5 years, 220M calls, 40M by 1000 agents**
3-4. Israeli Cellular Company: **2.5 years, 110M calls, 25M calls by 750 agents;** ILBank (IVR, SBR): 2 years
5. Back to Bank Anonymous: Call-Center and more - **from January 2010, daily-deposit at a SEESafe**

6. Service Engineering **internet site**: click-stream data (2 years)

7. *Home (Rambam) Hospital : **4 years, 1000 beds**, inter-ward patient flow

8. Emergency Departments (ED) patient flow:
- 5 EDs in Israel: 1-2 years, **late David Sinreich**, ED arrivals & LOS
- ED in Seoul: 2 months, **K. Song-Hee & W. Cha**, pilot
- ED in Singapore: 2 years, pilot

***Open & Free for (reproducible) research and teaching**

SEELab: Environment for graphical EDA

Operational histories (customers, servers) at the **individual-transaction level**, e.g.

1. *Bank Anonymous Call-Center: **1 year, 350K calls by 15 agents** - in 2000, **which paved the way to:**
2. *U.S. Bank Call-Center : **2.5 years, 220M calls, 40M by 1000 agents**
3-4. Israeli Cellular Company: **2.5 years, 110M calls, 25M calls by 750 agents;** ILBank (IVR, SBR): 2 years
5. Back to Bank Anonymous: Call-Center and more - **from January 2010, daily-deposit at a SEESafe**

6. Service Engineering **internet site**: click-stream data (2 years)

7. *Home (Rambam) Hospital : **4 years, 1000 beds**, inter-ward patient flow

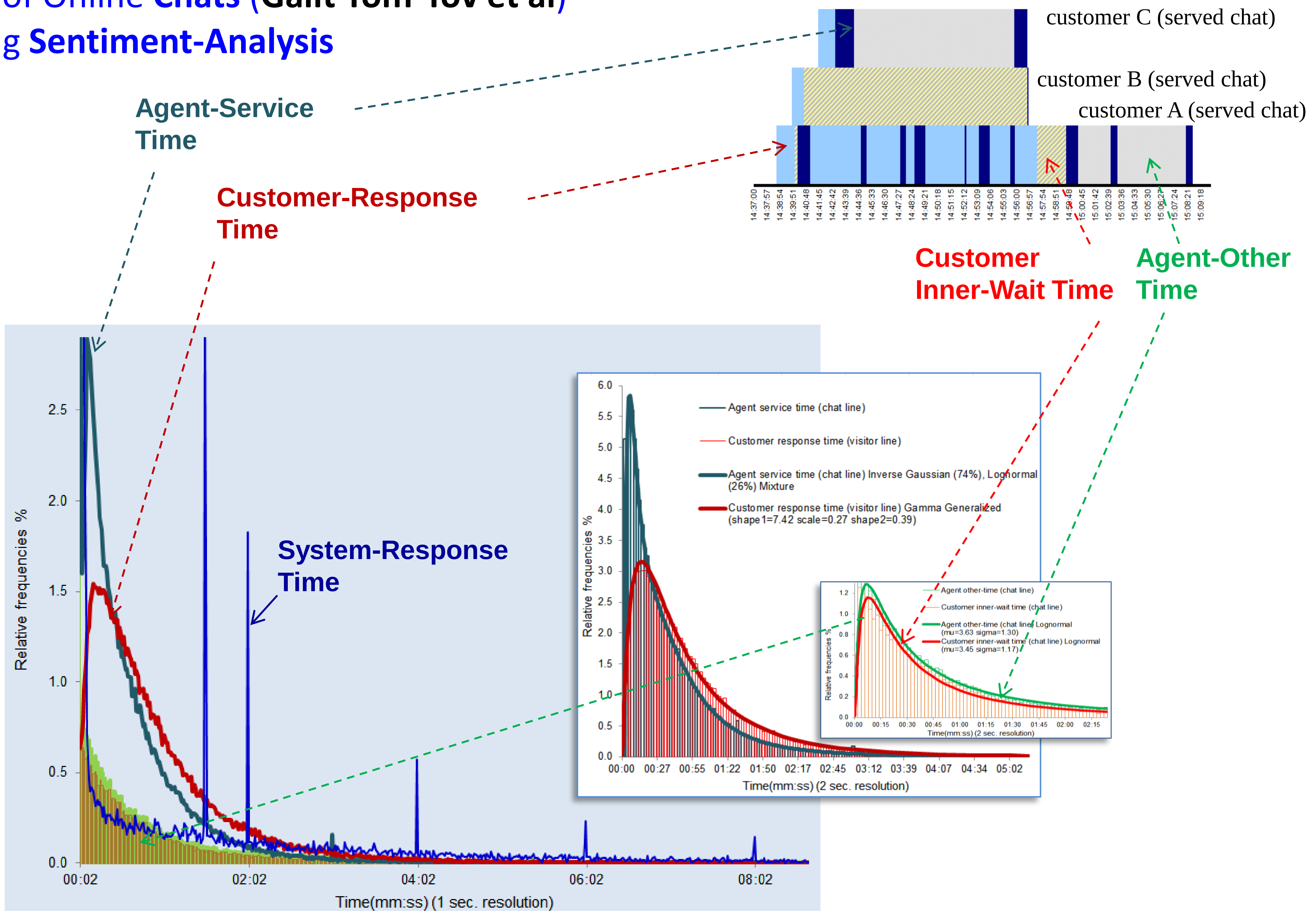
8. Emergency Departments (ED) patient flow:
- 5 EDs in Israel: 1-2 years, **late David Sinreich**, ED arrivals & LOS
- ED in Seoul: 2 months, **K. Song-Hee & W. Cha**, pilot
- ED in Singapore: 2 years, pilot

9. U.S. Ambulatory Hospital RTLS (Real-Time Location System): Since November 2013
- **250K events/day (1GB/week): 1000 patients, 350 staff (1500 tagged entities), every 3 sec's**
- Infrastructure: **900 readers (sensors) over ceilings of 7 (now 8) clinical floors**
- Both actual and planned (**appointment book** of resources: staff, patients, rooms)

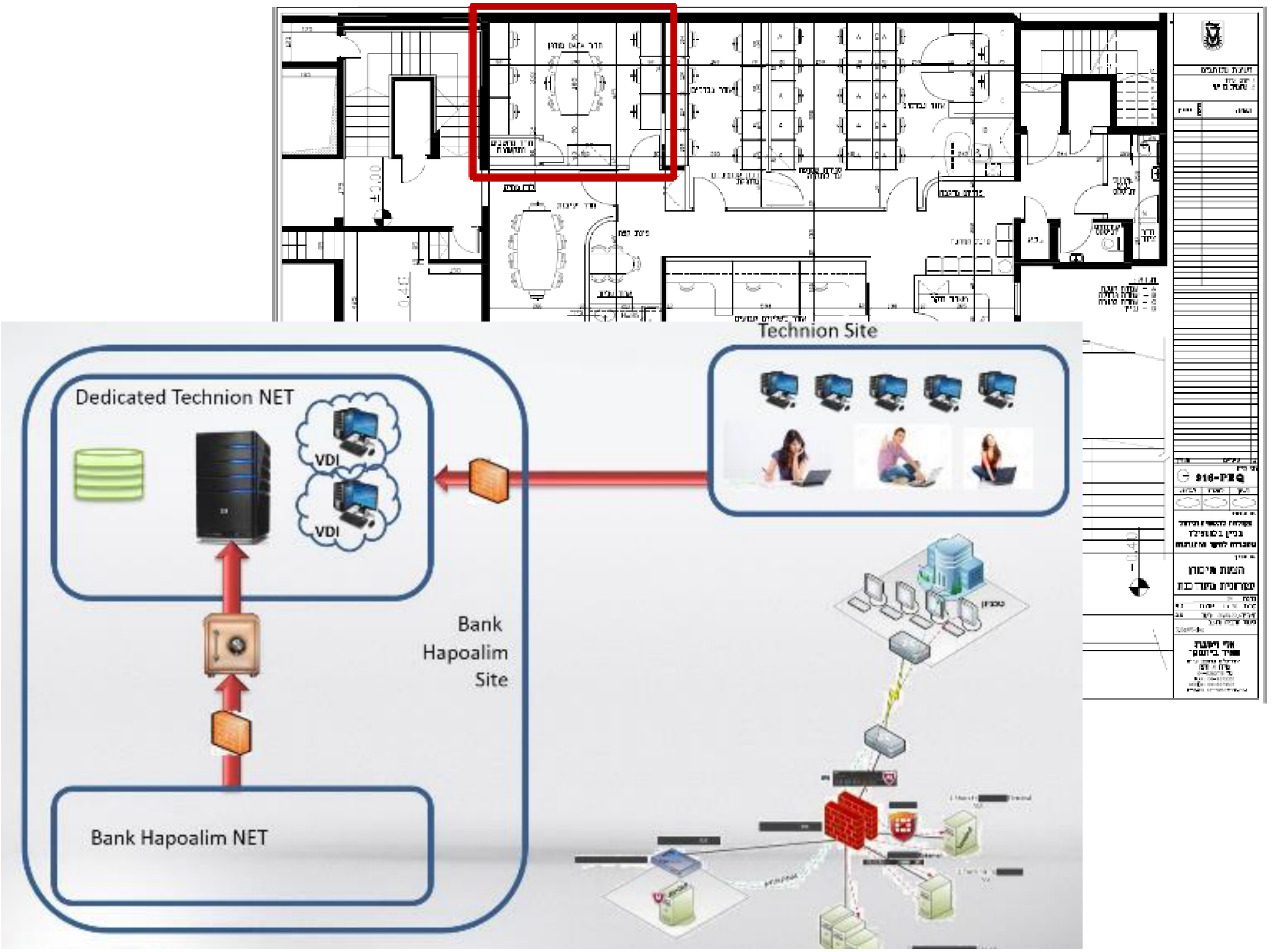
10-13:
Chat Services (Europe); ILBank Warehouse; Smart-City Simulator (Haifa, Boston,…); Justice System

***Open & Free for (reproducible) research and teaching**

Anatomy of Online Chats (Galit Yom-Tov et al)
Supporting Sentiment-Analysis



Data-Room (Secured): Accessing 80% of Bank Warehouse



- T-PADS מעבדה: איך, מי, איפה, ומתי
- מיקוד ראשוני Data Science & Engineering
- אפשרות הרחבה ל Blockchain, IoT, וכו'
  - 5 חוזהX5
  - בעלות משותפתIP -
  - פרסומים משותפים בהסכמה
  - דרך פרויקטים, לאימות המחקר POC
- סטטוס: 1ח, ועדת היגוי - 6ח, יום אירוע מרוכז- 1ש

Smart-City “Simulator” (40K sq ft “playground”, 10K offices)

Emergency services

Traffic lights

Lightening

Smart parking

Water, sewage

Energey systems

600 מ"ר משרדים + **2.5 דונם** של "מגרש משחקים" + 2 דונם הרחבה עתידית

מתהווה: שת"פ עם מערכת המשפט

שיתוף פעולה יאפשר תמיכה מחקרית במטרות הבאות (דוגמאות בלבד):

- ❖פיתוח מפה תהליכית המתארת הליכי טיפול בתיקי בתי המשפט, תוך איתור צווארי בקבוק תפעוליים ורעיונות לקיצור משכי הליכים.
- ❖ניתוח מדיניות חלוקת התיקים והתיעדוף הקיימים, במטרה לקצר תורי תיקים ושיפור חלוקתם (תורת התורים, תורת המשחקים,...).
- ❖בחינה וחקר (עידון) המדד "משקלות תיקים" (עומס תפעולי=Offered-Load), ויותר (למשל רגשי, דוגמת 2 מחלקות יולדות ברמב"ם, לאיזון עומסים "הוגן").
- ❖תכנון "יומנים" של שופט, עו"ד, תובע ונתבע (**Appointment Scheduling**)

The Skeptics

ERC: on using data to motivate theory (as in say Physics, Biology,…)

How can breakthrough mathematics come out of so much data?

NSF: on funding data-collection and maintenance in OR research grants

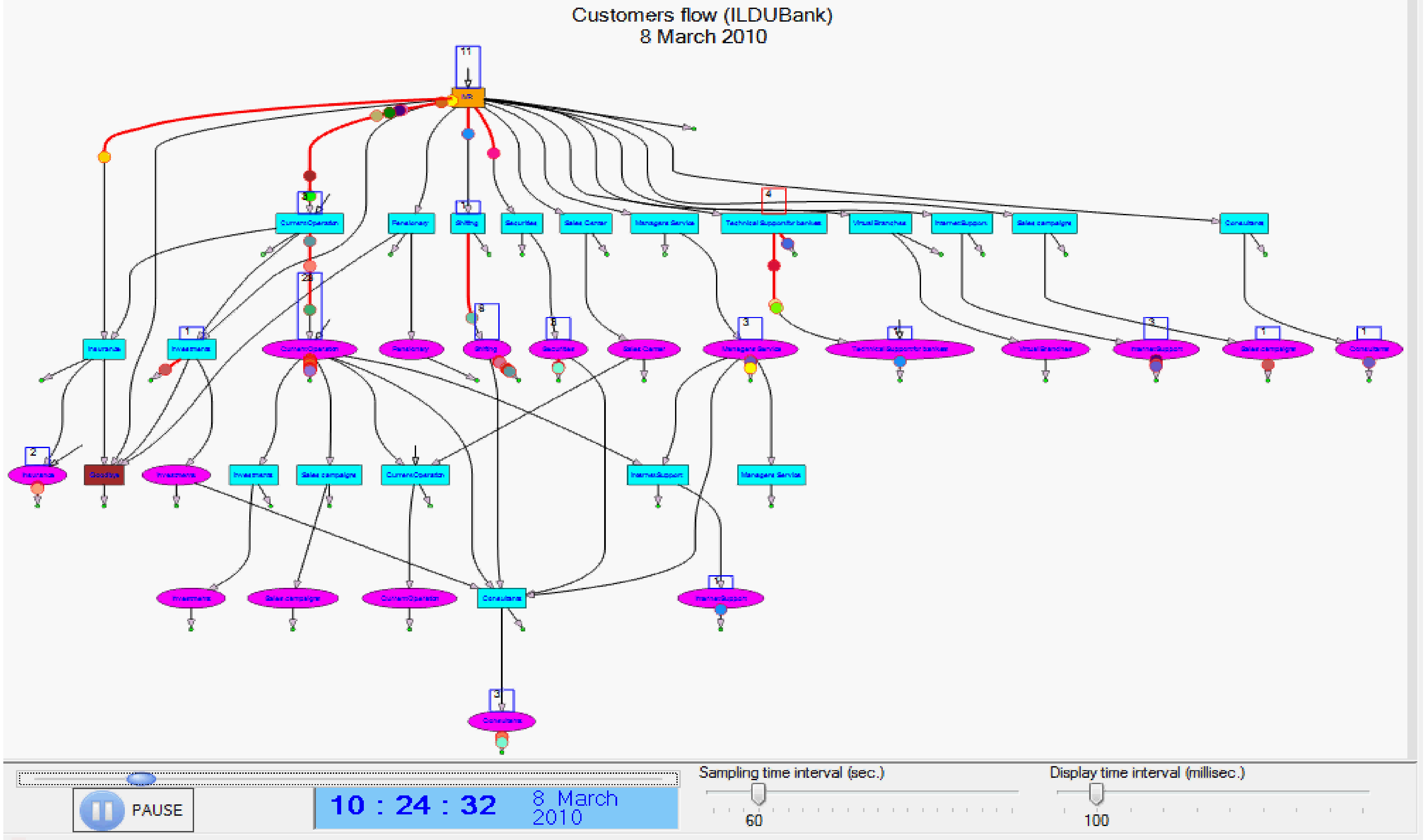
Finding an interested industry-partner w/ data $\Rightarrow$ problem solved

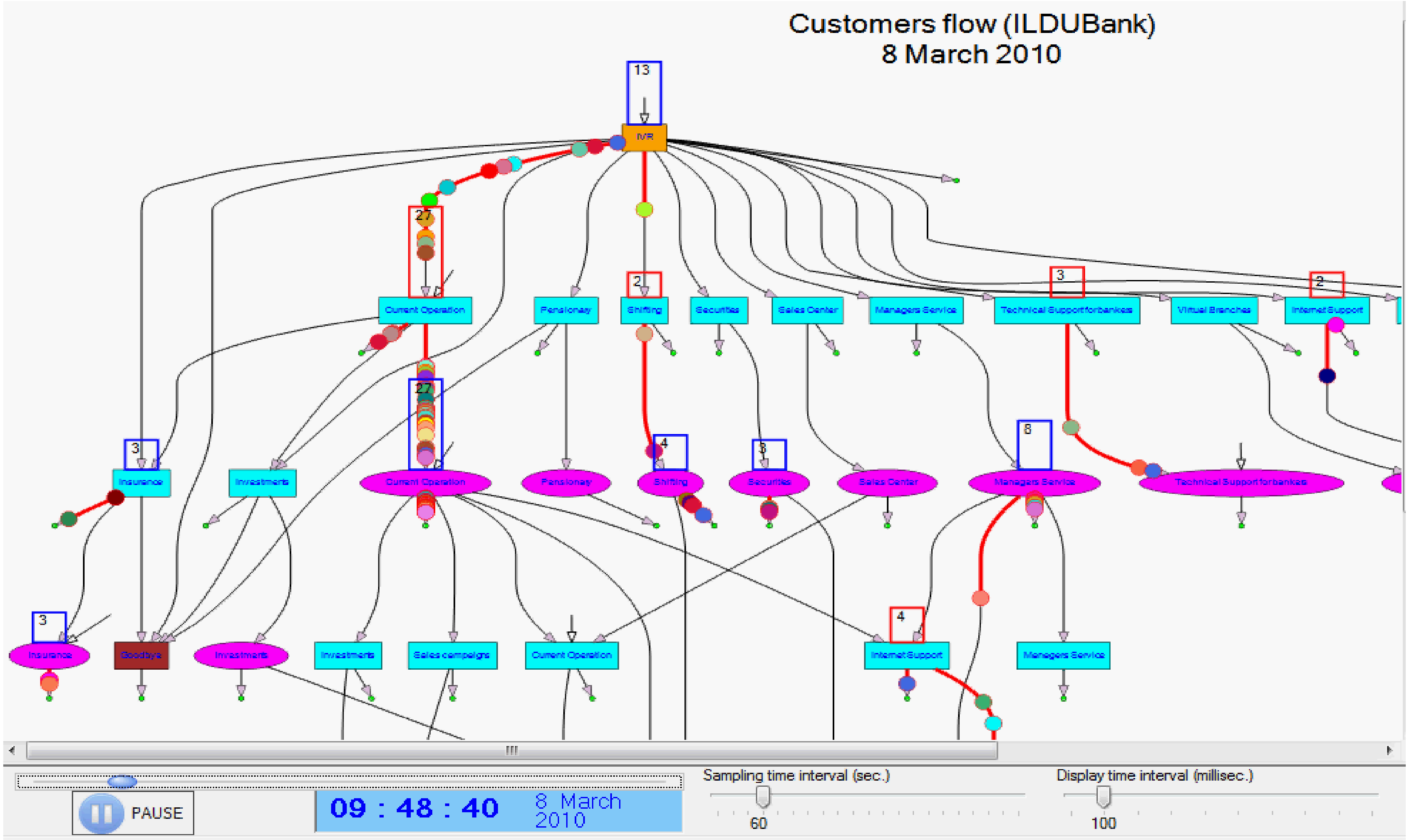
ISF: on measuring judicial workload (JW)

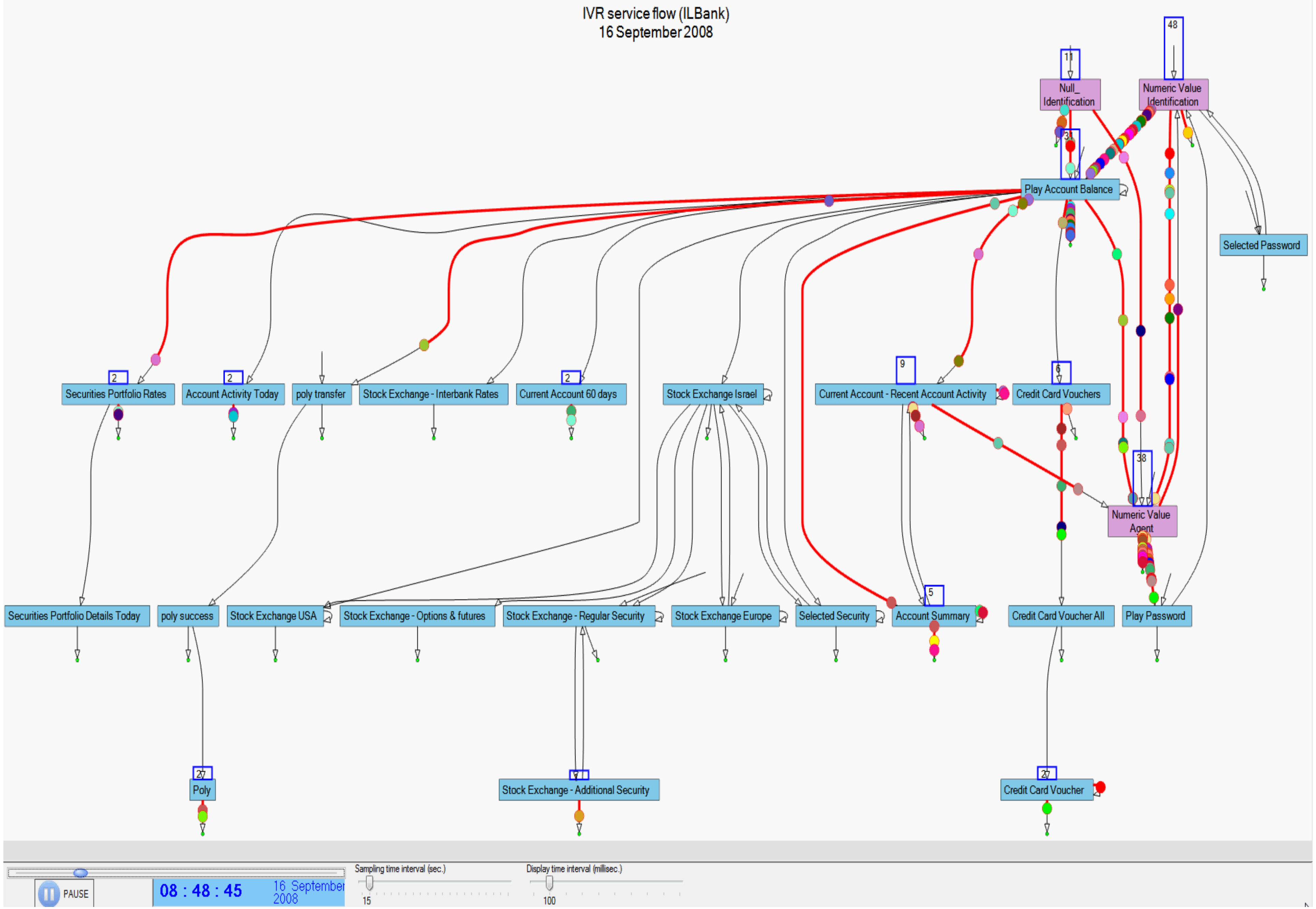
For the most part, … the applicant proposes to quantify the unquantifiable and solve the unsolvable, namely JW.

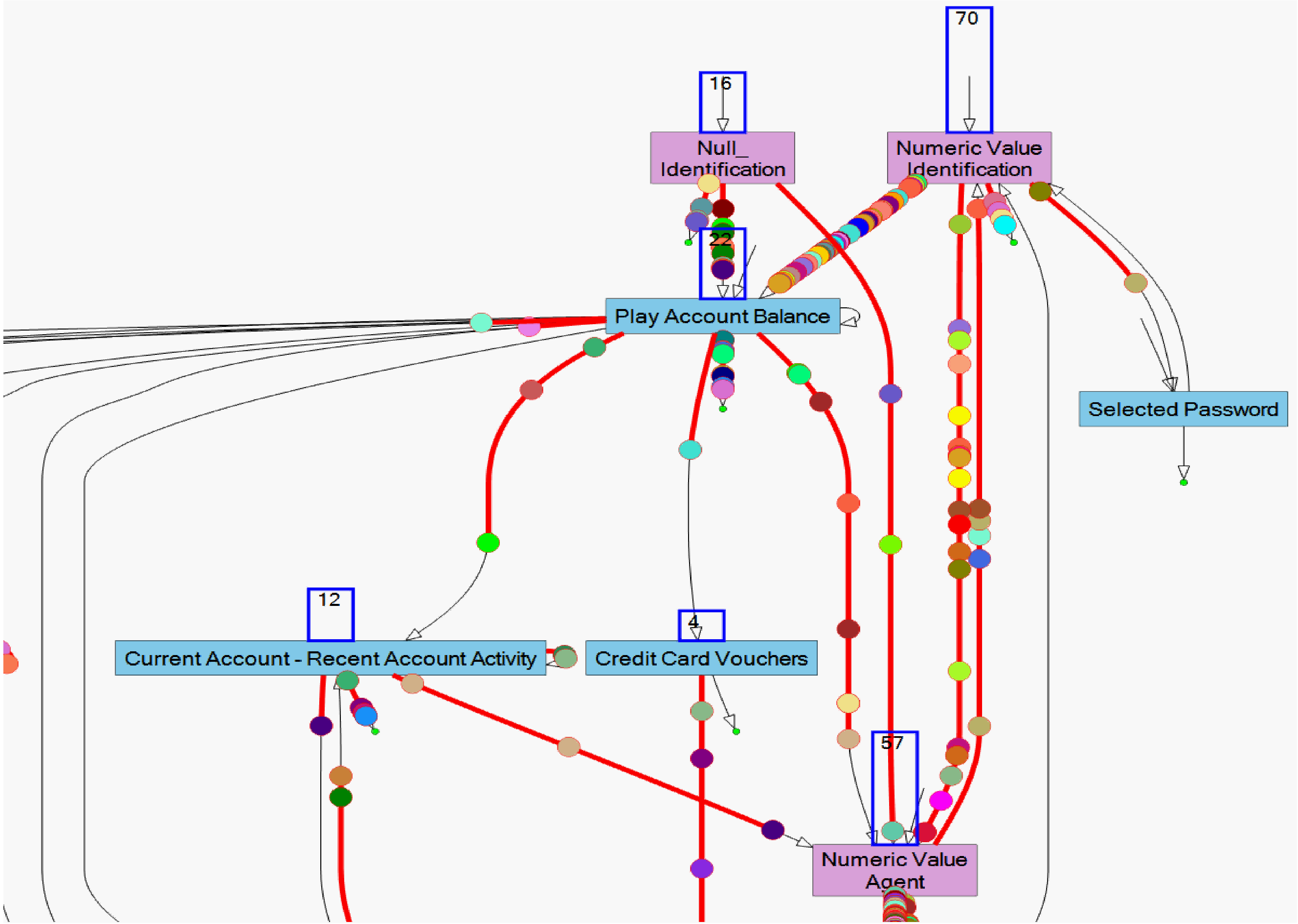
At least since the eighteenth century, there are continuous …

Still: Nurse-workload in maternity wards: operational+emotional(+cognitive)


Customers flow (ILDUBank)
8 March 2010
PAUSE
10 : 24 : 32
8 March 2010
Sampling time interval (sec.)
60
Display time interval (millisec.)
100

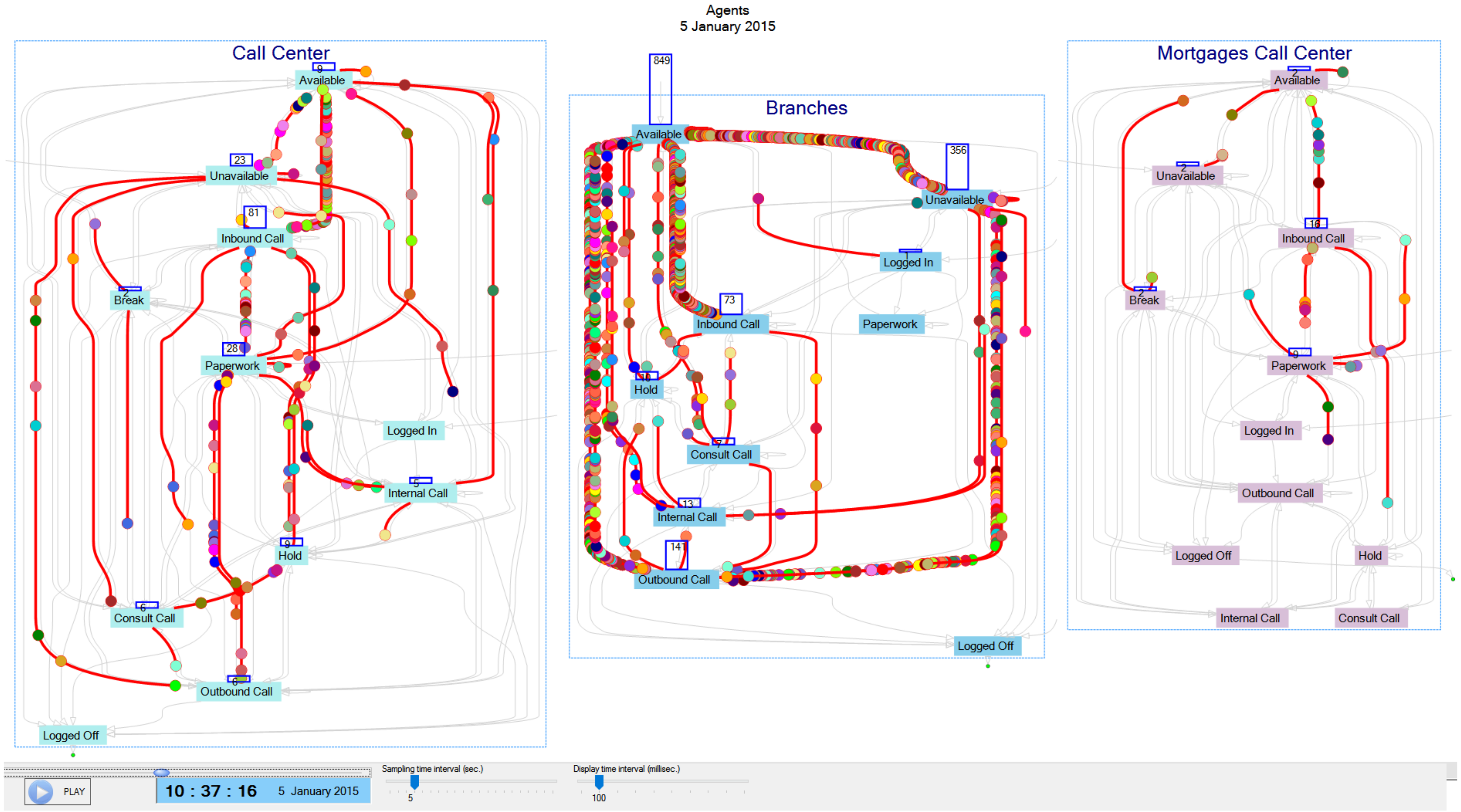

Customers flow (ILDUBank)
8 March 2010
IVR
Current Operation
Pensionary
Shifting
Securities
Sales Center
Managers Service
Technical Support for bankers
Virtual Branches
Internet Support
Insurance
Investments
Goodbye
Sales campaigns
PAUSE
09 : 48 : 40
8 March 2010
Sampling time interval (sec.)
60
Display time interval (millisec.)
100


IVR service flow (ILBank)
16 September 2008
Null_ Identification
Numeric Value Identification
Play Account Balance
Selected Password
Securities Portfolio Rates
Account Activity Today
poly transfer
Stock Exchange - Interbank Rates
Current Account 60 days
Stock Exchange Israel
Current Account - Recent Account Activity
Credit Card Vouchers
Numeric Value Agent
Securities Portfolio Details Today
poly success
Stock Exchange USA
Stock Exchange - Options & futures
Stock Exchange - Regular Security
Stock Exchange Europe
Selected Security
Account Summary
Credit Card Voucher All
Play Password
Poly
Stock Exchange - Additional Security
Credit Card Voucher
PAUSE
08 : 48 : 45
16 September 2008
Sampling time interval (sec.)
15
Display time interval (millisec.)
100

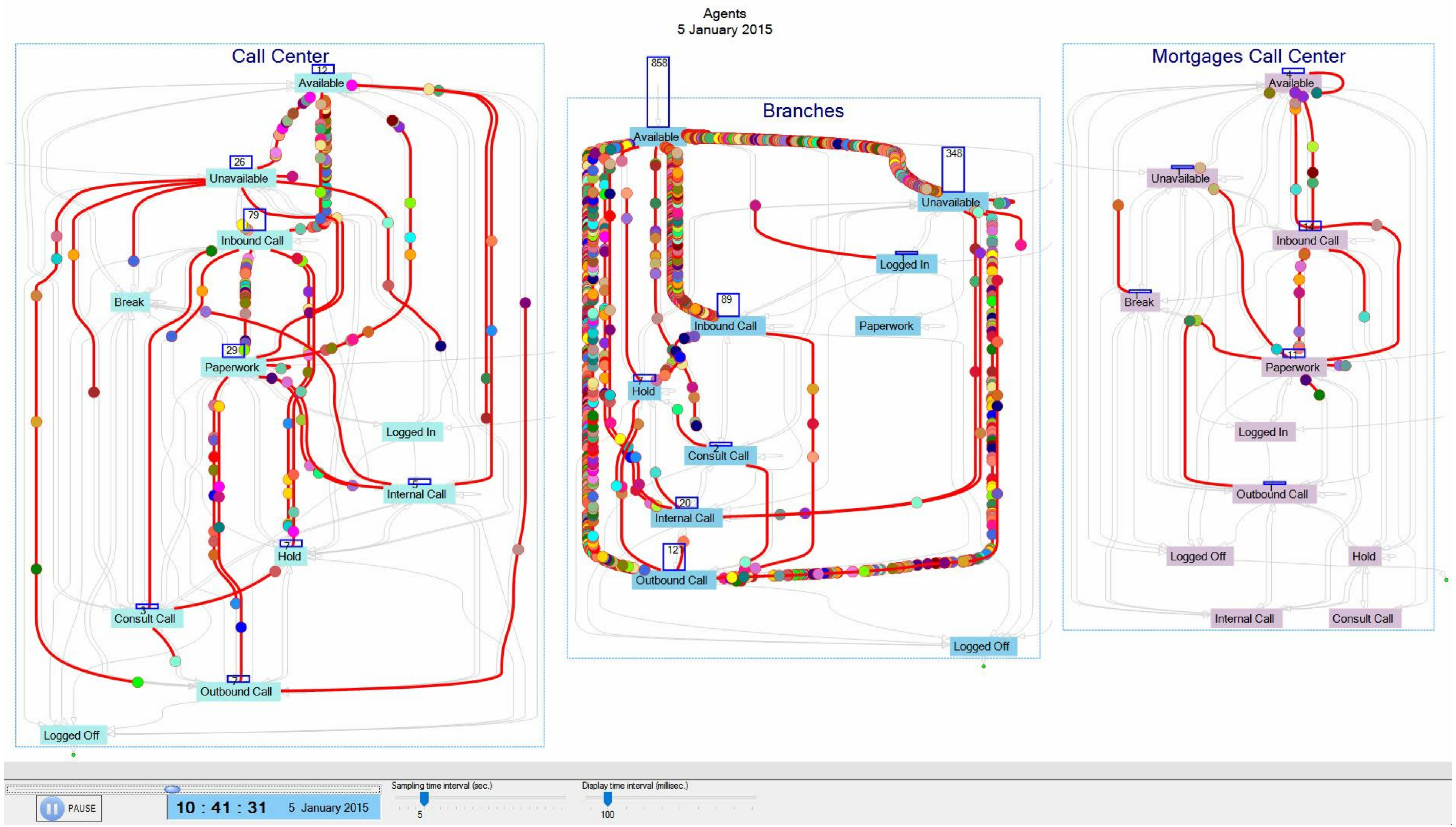


70
16
Null_ Identification
Numeric Value Identification
22
Play Account Balance
Selected Password
12
Current Account - Recent Account Activity
4
Credit Card Vouchers
57
Numeric Value Agent

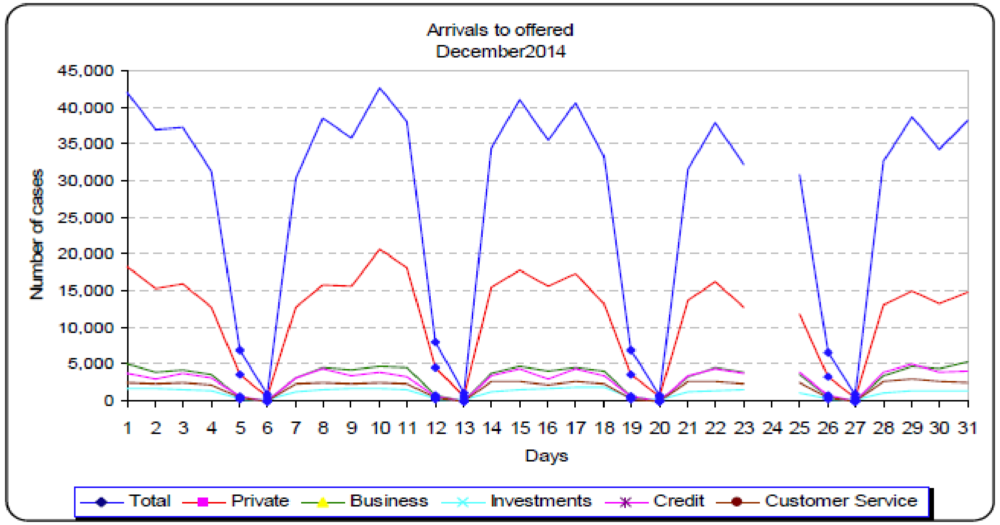
Agents (2000): Branches (1700) and Call Centers - Commercial (270), Mortgages (50)

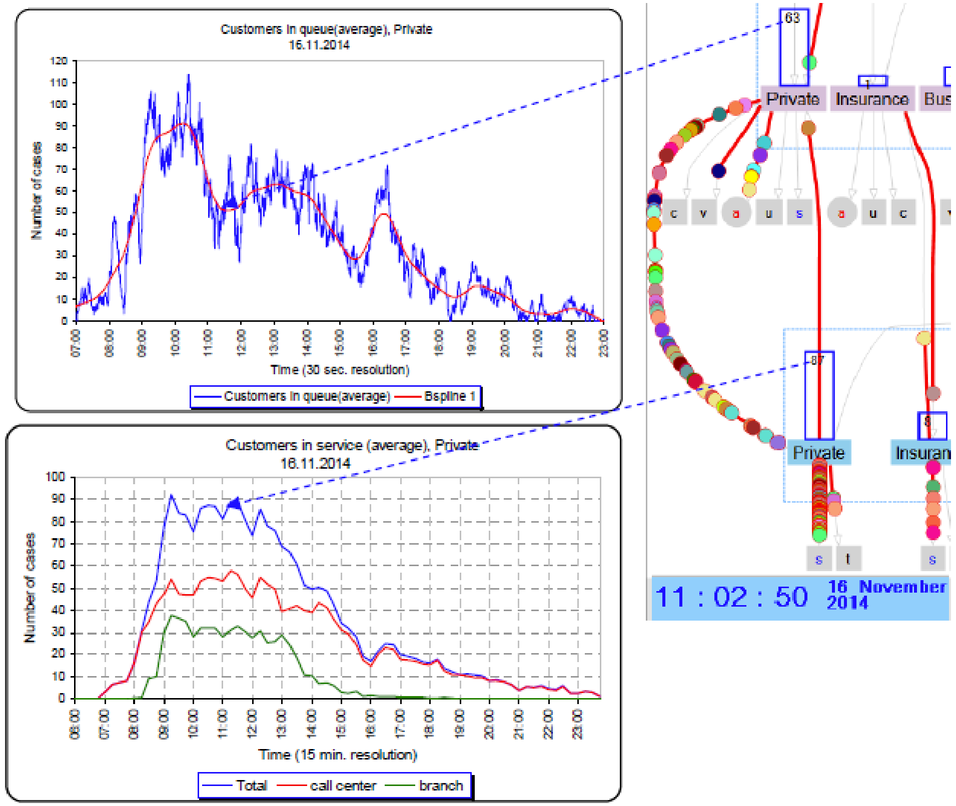


Agents: Branches, Commercial and Mortgages Call Centers



ILDU Banking:
branches and call centers
Service Types

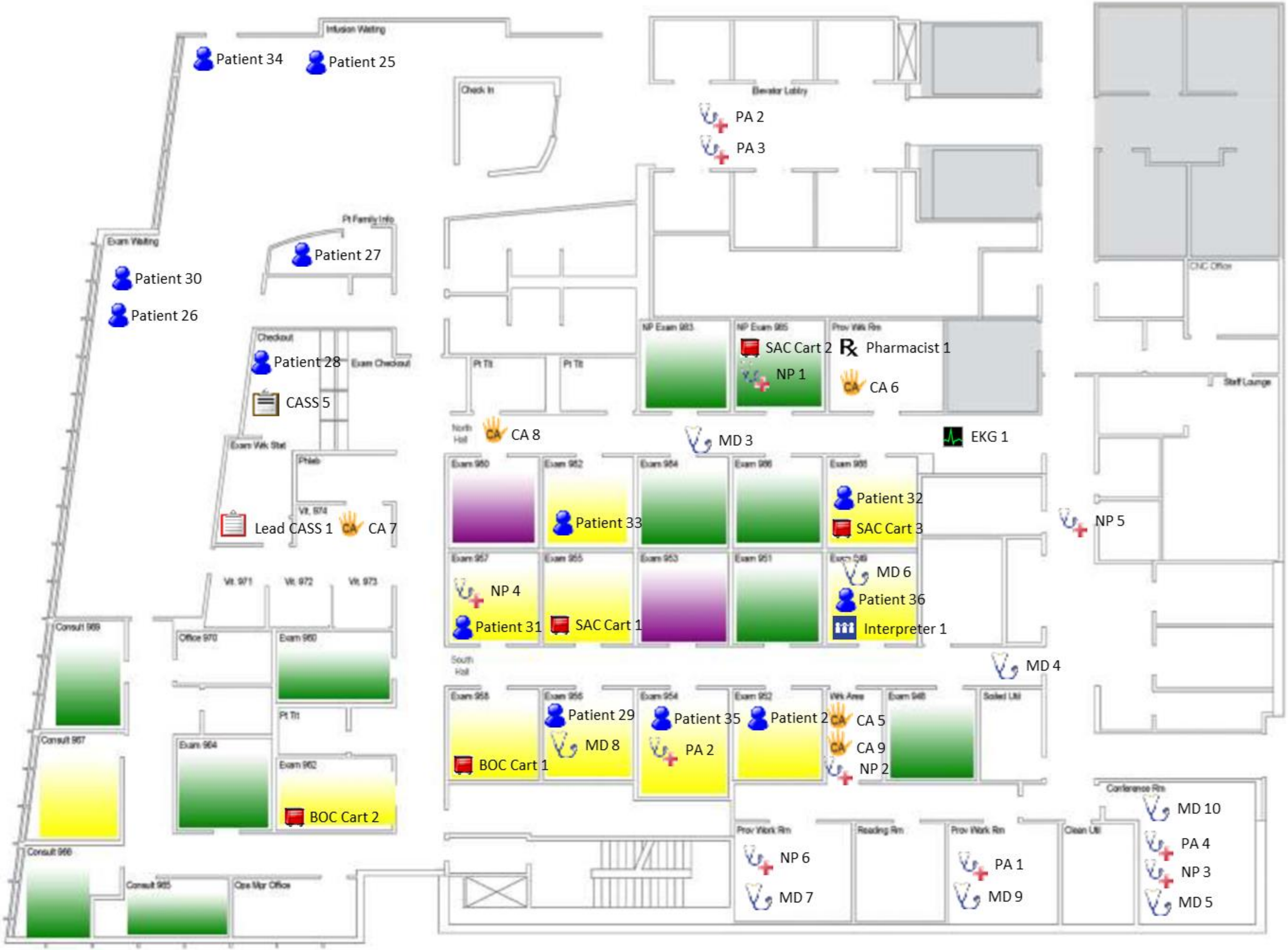




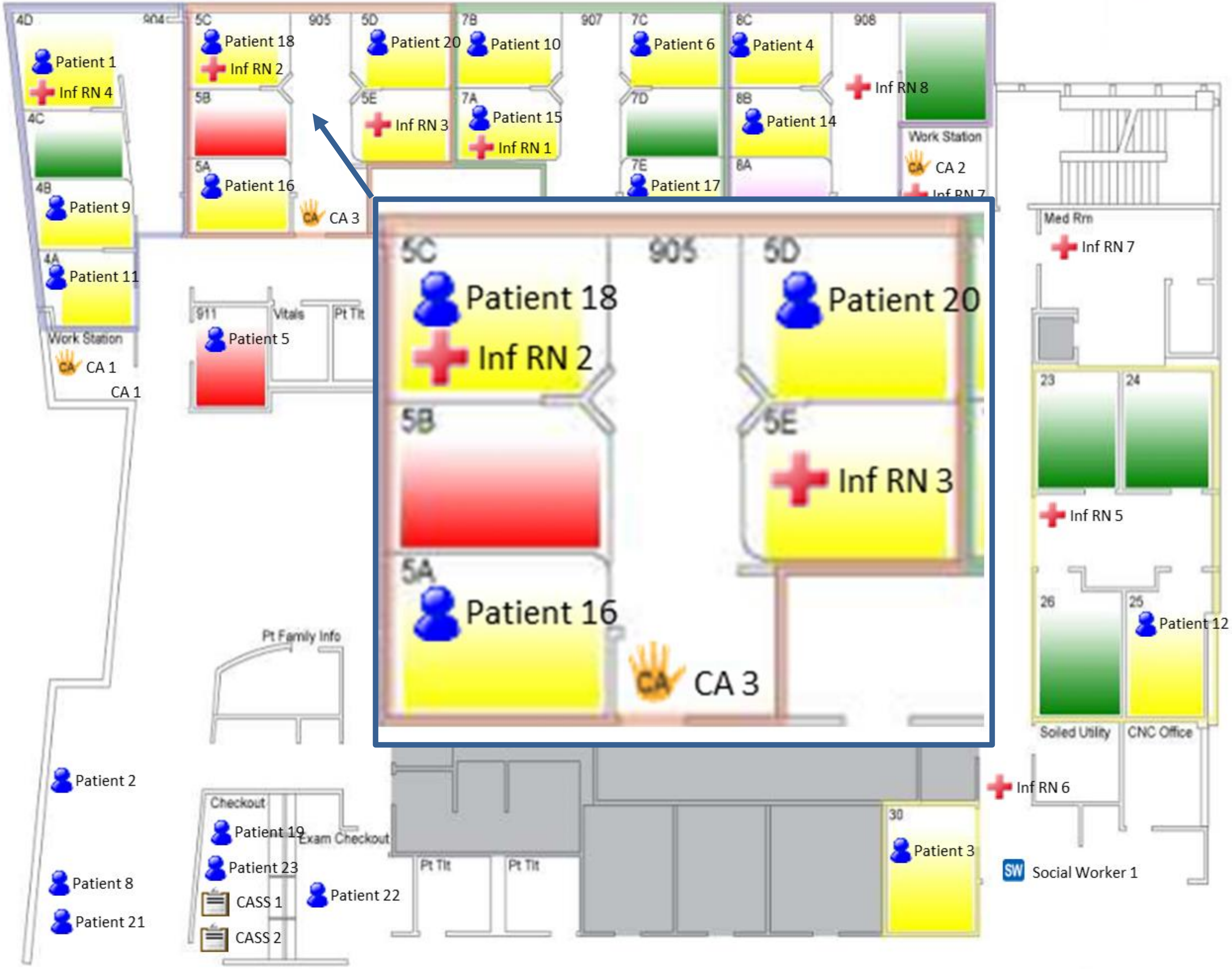
Closing the Data-Gap: from Call-Centers to Hospitals, now Banks

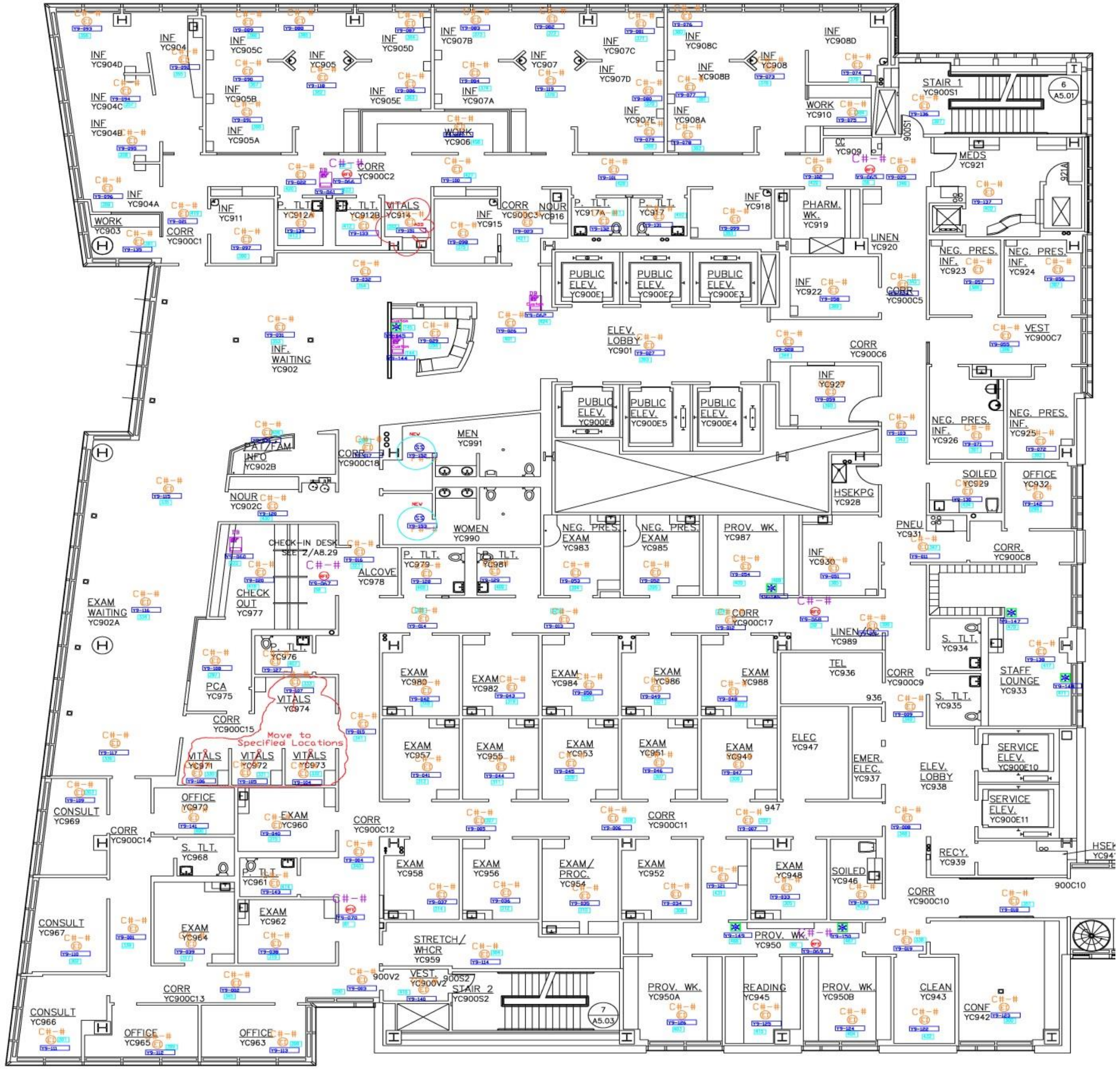
- **Large call center**:
  - 1000s of agents
  - Hundreds of thousands of calls per day
  - Data: operational, psychological, financial – **automatic** collection
- **Large hospital**:
  - 1000+ Beds
  - 1000s of patients & nurses, hundreds of doctors
  - Data: operational, clinical, financial – mostly **inaccessible (to academia)**

- **Large Bank:** “Enjoys” characteristics of both of the above

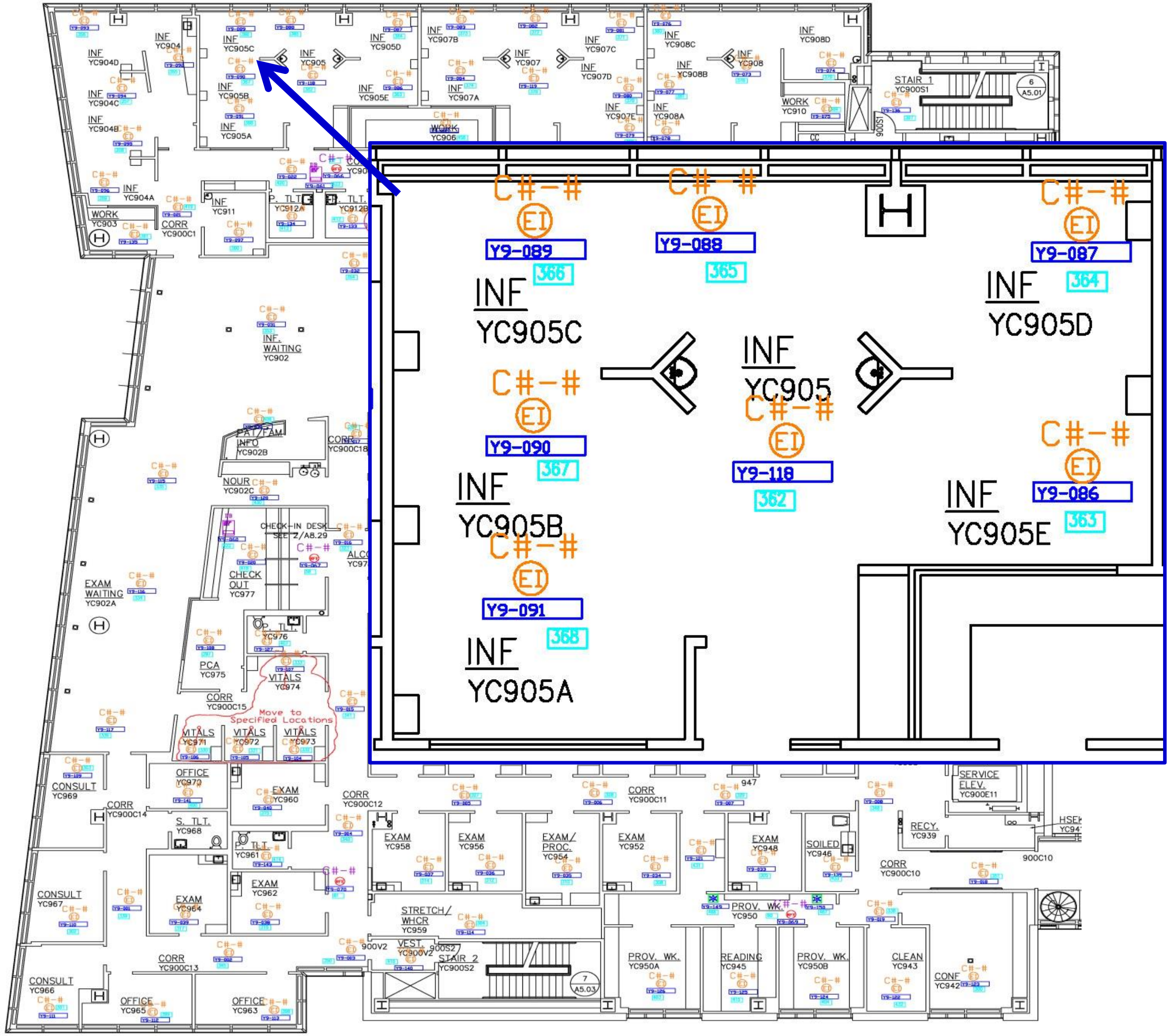

Patient 34
Patient 25
PA 2
PA 3
Patient 27
Patient 30
Patient 26
Patient 28
CASS 5
SAC Cart 2
Pharmacist 1
NP 1
CA 6
CA 8
MD 3
EKG 1
Patient 32
SAC Cart 3
Patient 33
Lead CASS 1
CA 7
NP 5
MD 6
Patient 36
Interpreter 1
NP 4
Patient 31
SAC Cart 1
MD 4
Patient 29
Patient 35
Patient 2
CA 5
CA 9
NP 2
MD 8
PA 2
BOC Cart 1
BOC Cart 2
MD 10
PA 4
NP 3
MD 5
NP 6
MD 7
PA 1
MD 9


4D
904
Patient 1
Inf RN 4
4C
4B
Patient 9
4A
Patient 11
Work Station
CA 1
CA 1
5C
Patient 18
Inf RN 2
905
5B
5A
Patient 16
CA 3
5D
Patient 20
5E
Inf RN 3
7B
Patient 10
7A
Patient 15
Inf RN 1
907
7C
Patient 6
7D
7E
Patient 17
8C
Patient 4
8B
Patient 14
8A
908
Inf RN 8
Work Station
CA 2
Inf RN 7
CA 4
Med Rm
Inf RN 7
911
Patient 5
Vitals
Pt Tlt
Nurse Sta
915
Pt Tlt
Pt Tlt
918
Patient 13
Pharmacy
22
Patient 7
Pt Staff Arriv
Main Elevator Lobby
Check In
CASS 3
CASS 4
MD 1
MD 2
23
24
Inf RN 5
27
26
25
Patient 12
Soiled Utility
CNC Office
Pt Family Info
Patient 2
Checkout
Patient 19
Exam Checkout
Patient 23
Patient 22
CASS 1
CASS 2
Patient 8
Patient 21
Pt Tlt
Pt Tlt
30
Patient 3
Inf RN 6
Social Worker 1


4D
904
Patient 1
Inf RN 4
4C
4B
Patient 9
4A
Patient 11
Work Station
CA 1
CA 1
5C
Patient 18
Inf RN 2
905
5D
Patient 20
7B
Patient 10
907
7C
Patient 6
8C
Patient 4
908
Inf RN 8
5B
5E
Inf RN 3
7A
Patient 15
Inf RN 1
7D
8B
Patient 14
Work Station
CA 2
5A
Patient 16
CA 3
7E
Patient 17
8A
Med Rm
Inf RN 7
911
Vitals
Pt Tlt
Patient 5
23
24
Inf RN 5
26
25
Patient 12
Soiled Utility
CNC Office
Pt Family Info
Inf RN 6
Patient 2
Checkout
Patient 19
Exam Checkout
Patient 23
Pt Tlt
Pt Tlt
30
Patient 3
Social Worker 1
Patient 8
Patient 21
CASS 1
CASS 2
Patient 22



INF
YC904D
INF
YC904C
INF
YC904B
INF
YC904A
WORK
YC903
CORR
YC900C1
INF
YC904
INF
YC905C
INF
YC905
INF
YC905B
INF
YC905A
INF
YC905E
INF
YC905D
INF
YC907B
INF
YC907
INF
YC907A
WORK
YC906
INF
YC907C
INF
YC907D
INF
YC907E
INF
YC908C
INF
YC908B
INF
YC908A
INF
YC908
INF
YC908D
WORK
YC910
CC
YC909
STAIR 1
YC900S1
MEDS
YC921
CORR
YC900C2
INF
YC911
P. TLT.
YC912A
P. TLT.
YC912B
VITALS
YC914
INF
YC915
CORR
YC900C3
NOUR
YC916
P. TLT.
YC917A
P. TLT.
YC917
INF
YC918
PHARM.
WK.
YC919
LINEN
YC920
PUBLIC
ELEV.
YC900E1
PUBLIC
ELEV.
YC900E2
PUBLIC
ELEV.
YC900E3
INF
YC922
CORR
YC900C5
NEG. PRES.
INF.
YC923
NEG. PRES.
INF.
YC924
VEST
YC900C7
INF.
WAITING
YC902
ELEV.
LOBBY
YC901
CORR
YC900C6
INF
YC927
PUBLIC
ELEV.
YC900E6
PUBLIC
ELEV.
YC900E5
PUBLIC
ELEV.
YC900E4
NEG. PRES.
INF.
YC926
NEG. PRES.
INF.
YC925
PAT/FAM
INFO
YC902B
CORR
YC900C18
MEN
YC991
NOUR
YC902C
WOMEN
YC990
HSEKPG
YC928
SOILED
YC929
OFFICE
YC932
PNEU
YC931
NEG. PRES.
EXAM
YC983
NEG. PRES.
EXAM
YC985
PROV. WK.
YC987
INF
YC930
CORR.
YC900C8
CHECK-IN DESK
SEE 2/A8.29
P. TLT.
YC979
P. TLT.
YC981
ALCOVE
YC978
CHECK
OUT
YC977
EXAM
WAITING
YC902A
CORR
YC900C17
LINEN
YC989
P. TLT.
YC976
PCA
YC975
VITALS
YC974
CORR
YC900C15
EXAM
YC980
EXAM
YC982
EXAM
YC984
EXAM
YC986
EXAM
YC988
TEL
YC936
CORR
YC900C9
S. TLT.
YC934
STAFF
LOUNGE
YC933
S. TLT.
YC935
Move to
Specified Locations
VITALS
YC971
VITALS
YC972
VITALS
YC973
EXAM
YC957
EXAM
YC955
EXAM
YC953
EXAM
YC951
EXAM
YC949
ELEC
YC947
EMER.
ELEC.
YC937
ELEV.
LOBBY
YC938
SERVICE
ELEV.
YC900E10
SERVICE
ELEV.
YC900E11
CONSULT
YC969
OFFICE
YC970
EXAM
YC960
CORR
YC900C14
CORR
YC900C12
CORR
YC900C11
S. TLT.
YC968
P. TLT.
YC961
EXAM
YC958
EXAM
YC956
EXAM/
PROC.
YC954
EXAM
YC952
EXAM
YC948
SOILED
YC946
RECY.
YC939
CORR
YC900C10
CONSULT
YC967
EXAM
YC964
EXAM
YC962
STRETCH/
WHCR
YC959
PROV. WK.
YC950
CORR
YC900C13
VEST
YC900V2
STAIR 2
YC900S2
PROV. WK.
YC950A
READING
YC945
PROV. WK.
YC950B
CLEAN
YC943
CONF
YC942
CONSULT
YC966
OFFICE
YC965
OFFICE
YC963


C#-#
EI
Y9-089
366
C#-#
EI
Y9-088
365
C#-#
EI
Y9-087
364
INF
YC905C
INF
YC905D
INF
YC905
C#-#
EI
Y9-090
367
C#-#
EI
Y9-118
362
C#-#
EI
Y9-086
363
INF
YC905B
INF
YC905E
C#-#
EI
Y9-091
368
INF
YC905A

Applications in DFCI

Control: rooms status, physicians location, long wait times

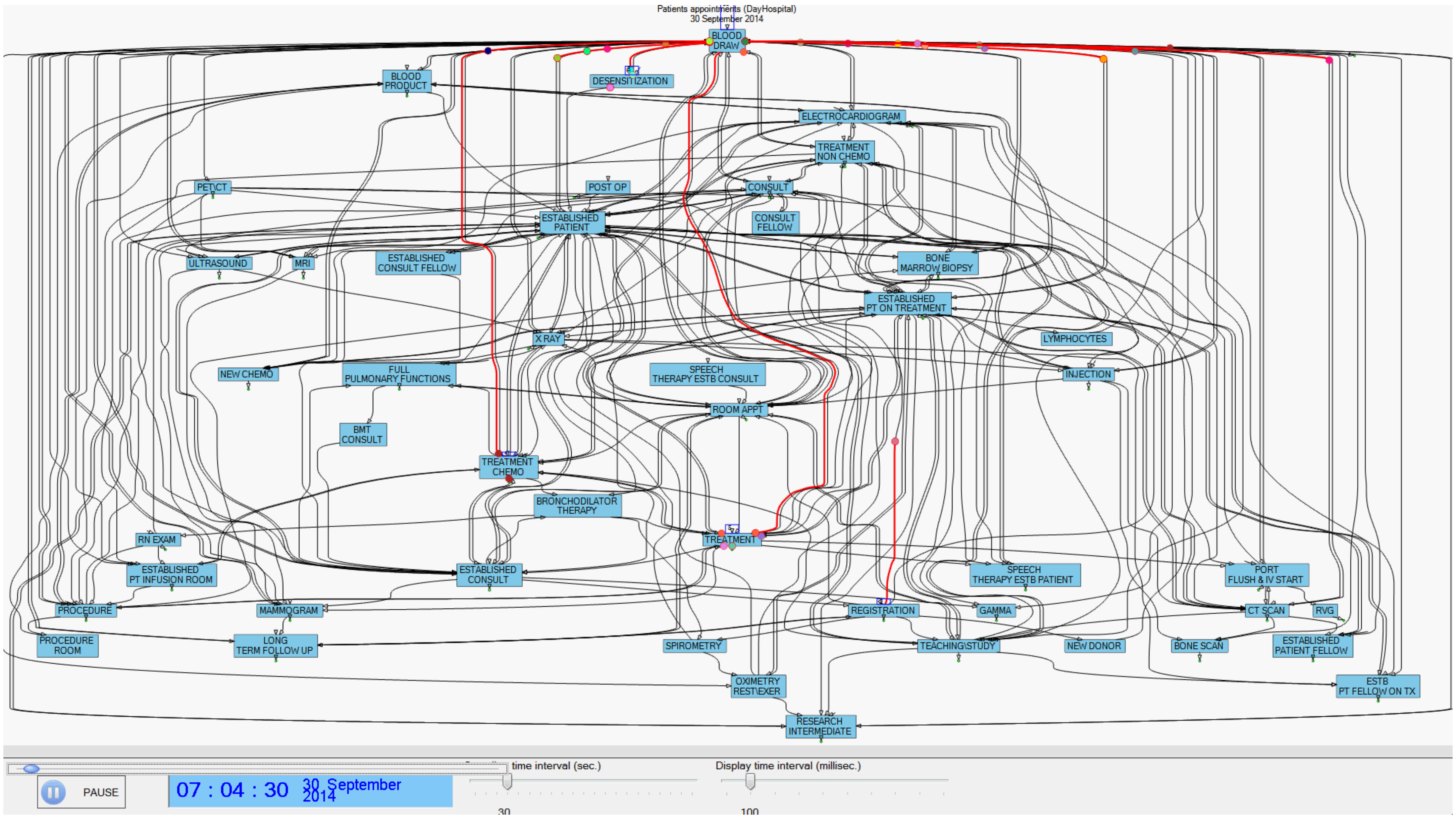
Planning: number infusion chairs, load-balancing among floors

Management: evidence-based

Motivating improvement: room for physician vs. room for patient

BUT how about
Time & Motion Studies of Resources (IEM 21st century), or
Mining Social Network(s), or
…

Elaine
Next: CT Scan At: 15:30
Geting there
Time left: about 5 hours
CT Scan
Blood test
Exam 1
Study
Blood test
Exam 2
X-Ray
Blood test
X-Ray
Idle state: Seeing the planned schedule before anything started The next activity highlighted up front
Elaine
Next: CT Scan At: 15:30
Go to: Clinical center 10
Back to full schedule
Next activity screen
Elaine
Next: Blood test At: 16:15
Go to: Clinical center 6
Back to full schedule
Change screen on button click
Next activity screen: Switch to it is automatic after previous activity was completed
Elaine
Next: Blood test At: 16:15
Geting there
Time left: about 4 hours
Idle state: Seeing the planned schedule after one activity was completed. The next activity highlighted up front
Elaine
Waiting for blood test results
More info
Time left: about 4 hours
Idle state: Seeing the planned schedule after two activities were completed. The next activity highlighted up front but
Elaine
Waiting for blood test results
More info
Time left: about 4 hours
Idle state: Seeing the planned schedule after two activities were done The next activity highlighted up front but time unknown. Selecting a later activity to get more details
Elaine
Later: Exam Planned at: 17:00
Location at: Clinical center 9a
Back to full schedule
Later activity screen
Elaine
Blood test results back
Next: Exam 1 At: 16:30
Getting there
Time left: about 4 hours
Idle state: Seeing the planned schedule after two activities were completed. Status of recent activity updated. The next activity highlighted up front with time and location.



Patients appointments (DayHospital)
30 September 2014
BLOOD DRAW
BLOOD PRODUCT
DESENSITIZATION
ELECTROCARDIOGRAM
TREATMENT NON CHEMO
PET\CT
POST OP
CONSULT
CONSULT FELLOW
ESTABLISHED PATIENT
ULTRASOUND
MRI
ESTABLISHED CONSULT FELLOW
BONE MARROW BIOPSY
ESTABLISHED PT ON TREATMENT
X RAY
LYMPHOCYTES
NEW CHEMO
FULL PULMONARY FUNCTIONS
SPEECH THERAPY ESTB CONSULT
INJECTION
ROOM APPT
BMT CONSULT
TREATMENT CHEMO
BRONCHODILATOR THERAPY
RN EXAM
TREATMENT
ESTABLISHED PT INFUSION ROOM
ESTABLISHED CONSULT
SPEECH THERAPY ESTB PATIENT
PORT FLUSH & IV START
PROCEDURE
MAMMOGRAM
REGISTRATION
GAMMA
CT SCAN
RVG
PROCEDURE ROOM
LONG TERM FOLLOW UP
SPIROMETRY
TEACHING\STUDY
NEW DONOR
BONE SCAN
ESTABLISHED PATIENT FELLOW
OXIMETRY REST\EXER
ESTB PT FELLOW ON TX
RESEARCH INTERMEDIATE
time interval (sec.)
Display time interval (millisec.)
PAUSE
07 : 04 : 30
30 September 2014
30
100

Appointment-Driven Services (Networks)

e.g. Healthcare, Courts, Projects, MSEs, …, even Banks now

- Ample research (1000's papers, books) since the 50's; **theory = 1-server**
- Significant in heavy-traffic hence begs for heavy-traffic theory, yet
- Aware of 1 example: **Atar, Armony, Honnappa** (2017): Single server, via "Asymptotically optimal appointment schedules with customer no-shows" (submitted)
- Ongoing: **with Momcilovic, Trichakis, DFCI partners; Huang** "Data-Driven Appointment-Scheduling Under Uncertainty: … Infusion…" (submitted)
  - QED Appointments (ongoing)
  - Ultimately: DFCU = Network in heavy-traffic (start AAH, then mix with QED)

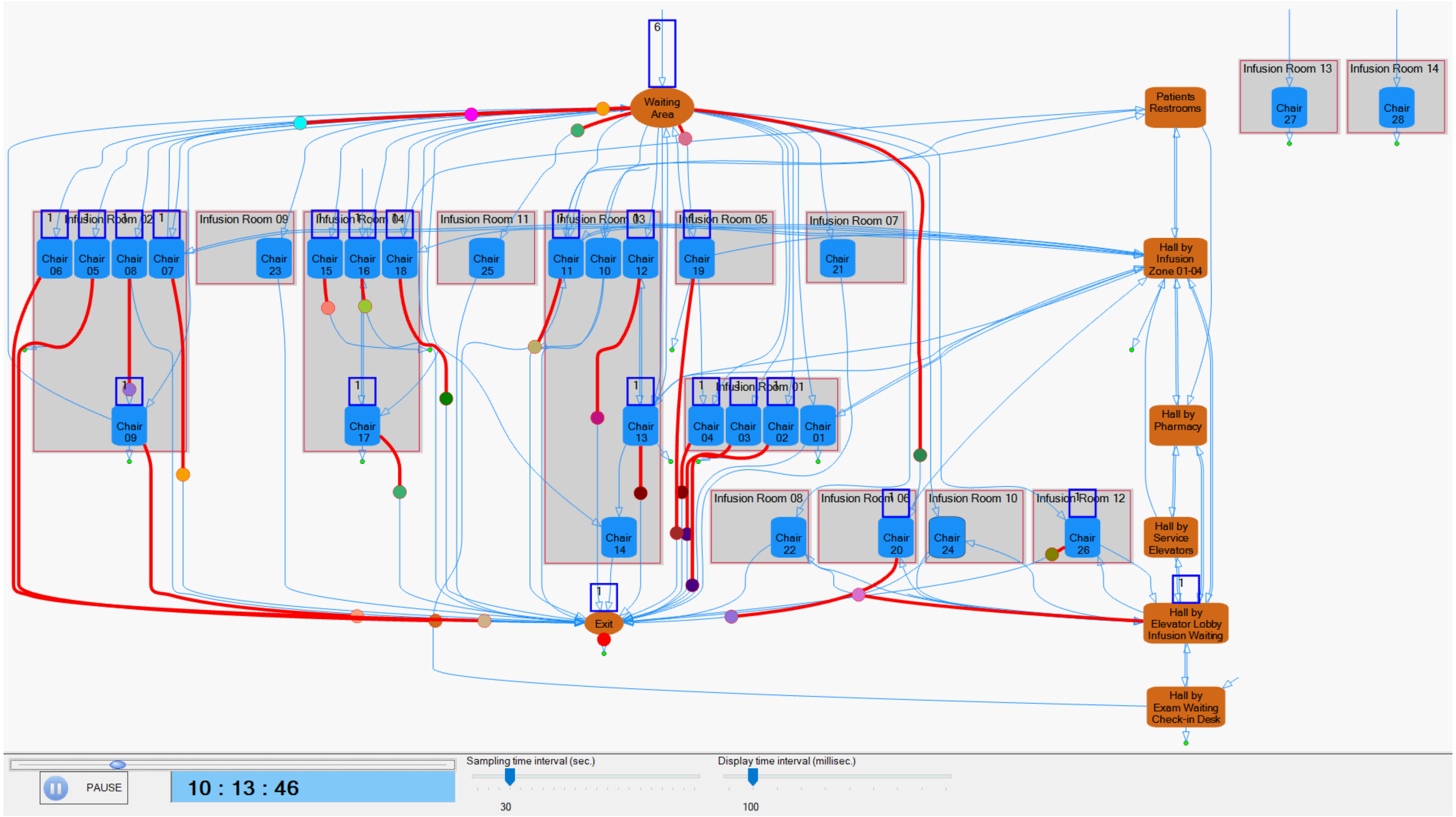
Appointments dominate healthcare systems.
Such systems are intrinsically stochastic, yet appointment systems (too) often view them deterministically:

DFCI Sample Infusion Schedule on Sep 3, 2014

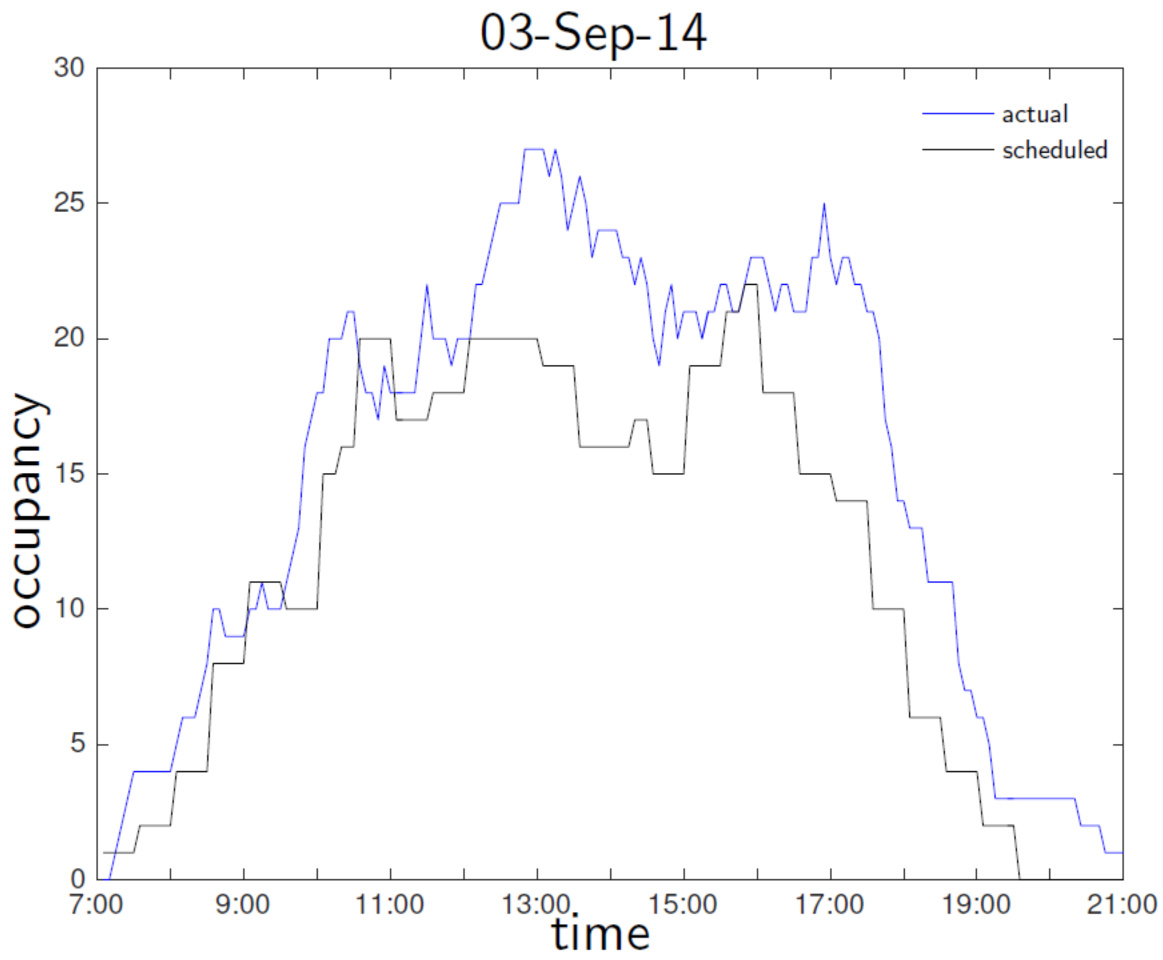
time	patient_id	duration (min)	link_flag	floor_id	disease_center
15:00	01	60	unlinked	9	breast onc.
12:30	02	120	unlinked	9	breast onc.
10:30	03	180	linked	9	genitourinary onc.
12:30	04	60	linked	9	breast onc.
12:00	05	120	linked	9	genitourinary onc.
07:00	06	60	unlinked	9	breast onc.
⋮	⋮	⋮	⋮	⋮	⋮

w/ P. Momcilovic & N. Trichakis: Develop methodologies for appointment scheduling, in **multi-server** environments, that take into account **stochastic** punctuality and service-durations

Patients Flow: N^{th} Floor, Infusion Unit w/ P. Momcilovic & N. Trichakis

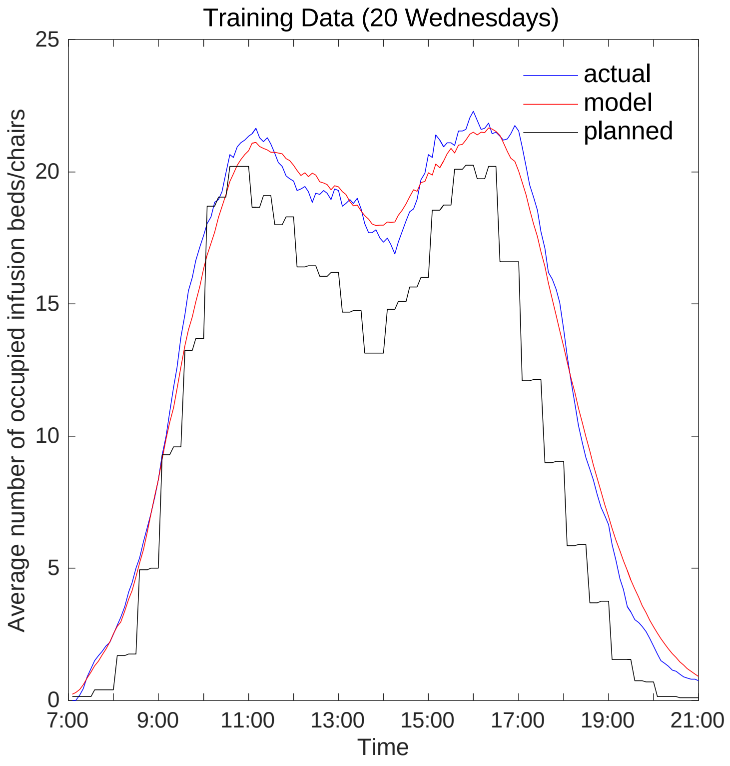


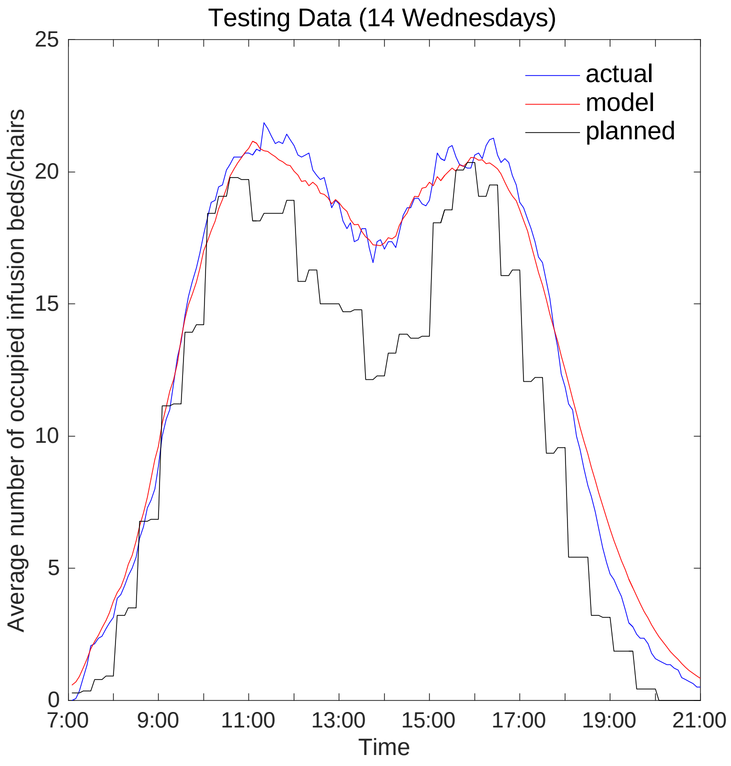
Scheduled *vs* Actual



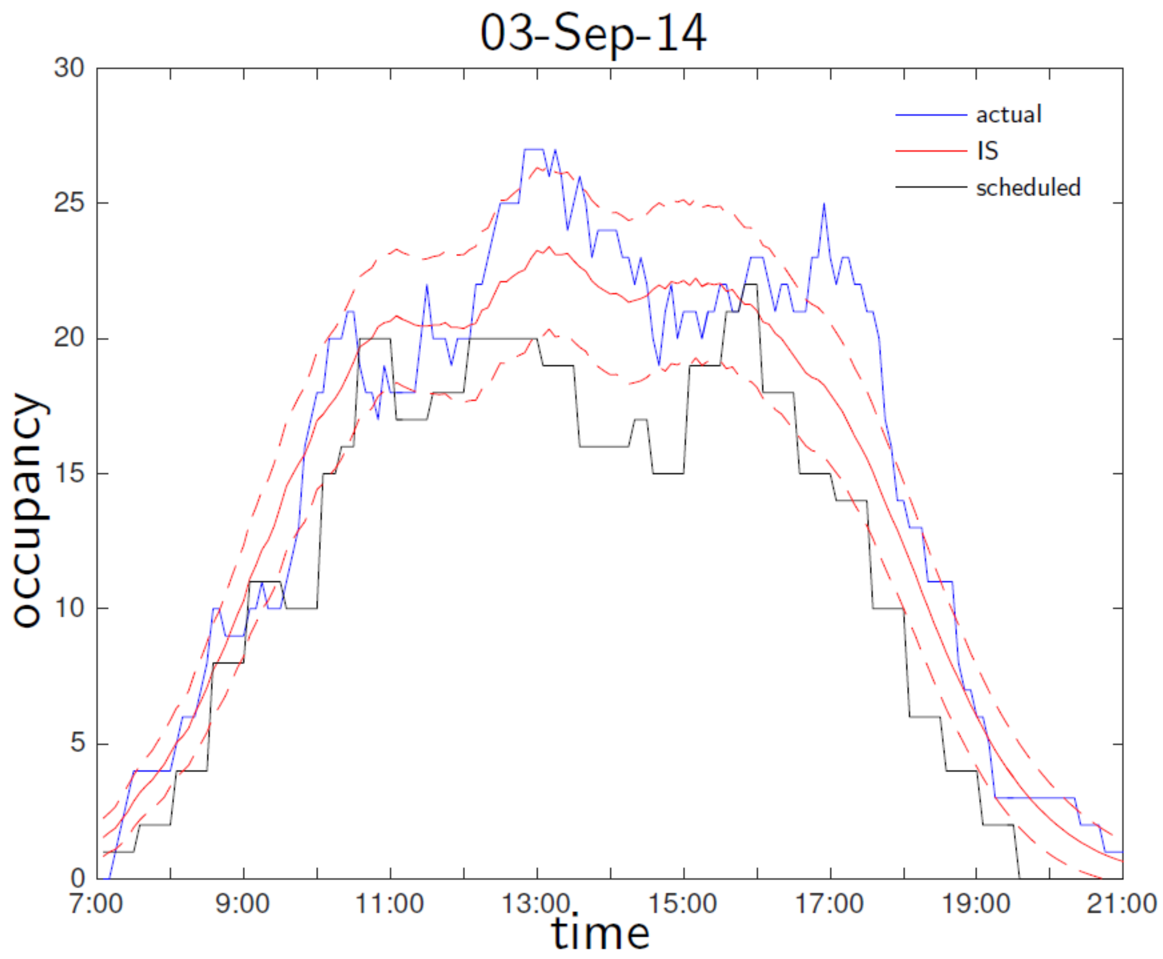
Approach & Validation

Appointment scheduling in **multi-server** environments: accounting for **stochastic** punctuality and service-durations, via infinite-server models (offered-load)





Validation with DFCI Data



Infinite-Server (IS) Approximation

- ith customer arrives $a_i + P_i$, leaves $a_i + P_i + D_i$ $(c_t = \infty)$
- $Z_i(t) := 1_{\{a_i+P_i \le t < a_i+P_i+D_i\}}$ "presence" indicator
- $Z(t) = \sum_{i=1}^{n} Z_i(t)$

$$\mathbb{E}Z(t) = \sum_{i=1}^{n} \mathbb{E}\tilde{F}_i(t - a_i - P_i) =: \sum_{i=1}^{n} \Omega_i(t)$$

$$\text{Var}(Z(t)) = \sum_{i=1}^{n} \Omega_i(t)\ (1 - \Omega_i(t))$$

where $\tilde{F}_i(x) := 1_{\{x \ge 0\}}(1 - F_i(x))$

IS Solution Approach

- recall: $Z = \{Z(t)\}$ occupancy process
- sample path cost

$$Q(Z) := \int_{-\infty}^{\infty} r(Z(t) - c_t)\,\mathrm{d}t + \tilde{\gamma} \int_{T}^{\infty} Z(t)\,\mathrm{d}t,$$

  $r(\cdot)$ cost function, *e.g.*, $(\cdot)^+$

- CLT approximation $\tilde{Z}(t) := \mathbb{E}Z(t) + \xi(t)\,\sqrt{\mathrm{Var}(Z(t))}$ ($\xi(t)$ std normal)

$$\min_{0 \le a \le T} \mathbb{E}Q(\tilde{Z})$$

Comparison with DFCI Practice

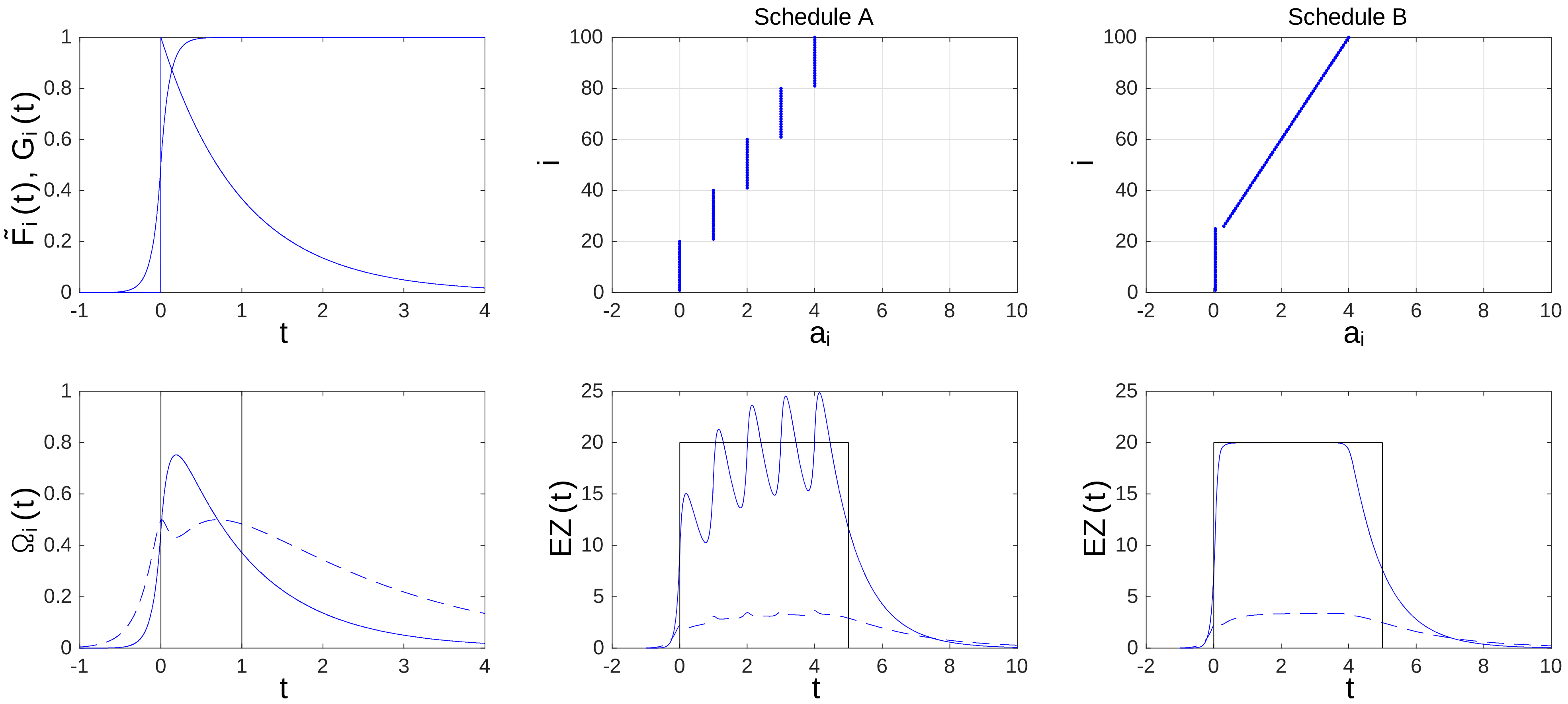
- infusion service on floors 8, 9
- $c_t \approx 25$, $n \approx 90$
- compare IS with "means-based" scheduling
- consider 85% and 95% utilization levels

- Reduce costs (waiting, overtime) by approximately **30%**

- Hopefully a pilot soon

Appointment Scheduling Matters:
Means-Based (Prevalent) vs. (Reasonable/almost Optimal) Alternative

- 100 customers, 20 servers: $\{a_i\}_i$ appointment times
- **Exponential** service times with mean=1; **Laplace** distribution of punctuality
- Compare 2 schedules:



Observations on “Academia-Industry Partnerships”
(not MSR or the Original-Bell-Labs)

- Intriguingly: RTLS successful at a **leading research** hospital?
- Intriguingly: initiative of **Physicians** at DFCI (as opposed to managers)?
- **Conjectures?**

Some Observations on "Academia-Industry Partnerships"
(not MSR or the Original-Bell-Labs)

- Intriguingly: RTLS successful at a **leading research** hospital?
- Intriguingly: initiative of **Physicians** at DFCI (as opposed to managers)?
- **Conjectures?**

Partnership with "Strong" partners, who appreciate Research/Evidence-Based-Management, and it is naturally based on Knowledge partnering with Data

Data enables SYMMETRIC Partnerships, caters to goals of both partners (even IP)

The Technion SEE Center / Laboratory

Data-Based Service Science / Engineering