Statistical Utilization of Complex Data: An Application in a Pilot Fatigue Study

or

Evidence-Based Pilot Fatigue Risk Management at Boeing

Christopher M. Gast, Ph.D.
Kim Craig
Rob Grube
Mike Muhm, MD
Lisa Thomas, Ph.D.

Boeing Research & Technology, Applied Mathematics, Applied Statistics

June 24, 2014

Introduction

- NASA Research indicates that **fatigue** is at least a contributing factor in approximately 20% of all accidents or incidents
- From 1993 to 2008, **fatigue** was associated with over 250 fatalities in air carrier accidents being investigated by NTSB
  - Feb. 2008: *"NTSB: Both Pilots Asleep on Hawaii Flight"*
  - Oct. 2009: *"Northwest's Wayward Flight: Sleeping Pilots?"*
  - Mar. 2012: *"Sleep Deprivation Blamed for JetBlue Pilot's March Meltdown"*

And, more recently: *"Focus on Chicago Transit Authority crash falls on operator fatigue, braking system"*

Introduction

- Airlines use Fatigue Risk Management Systems (FRMS) for multiple reasons:
  - **Safety**: FRMS intends to reduce the risk of accidents and incidents related to fatigue
  - **Efficiency**: Schedule crews more efficiently
  - **Regulatory compliance**: Airlines must comply with flight/duty time limits, established by the FAA

- In January of 2014, the Federal Aviation Administration enacted new rules on flight, duty and rest for commercial airline pilots
  - Airlines may earn exemptions from certain regulations if they provide an approved, valid, scientifically-based FRMS

Introduction

Current systems take a "strategic" perspective (as opposed to "tactical").

Example: Jeppesen CrewAlert software, plus the Boeing Alertness Model

Introduction

- Boeing (among others) is interested in studying fatigue from the *tactical* perspective.
- Can we estimate a pilot's *current* state of fatigue or alertness?

Introduction

- Boeing (among others) is interested in studying fatigue from the *tactical* perspective.
- Can we estimate a pilot's *current* state of fatigue or alertness?

- How do we define fatigue?
- How do we measure it?
- Can we accurately predict future fatigue?
- Is there a difference between low-workload (boring) and high-workload fatigue?
- Can we associate level of fatigue with performance degradation?

Crew Measurement Experiment - Design

- To help answer these questions, Boeing conducted a "Crew Measurement Experiment"

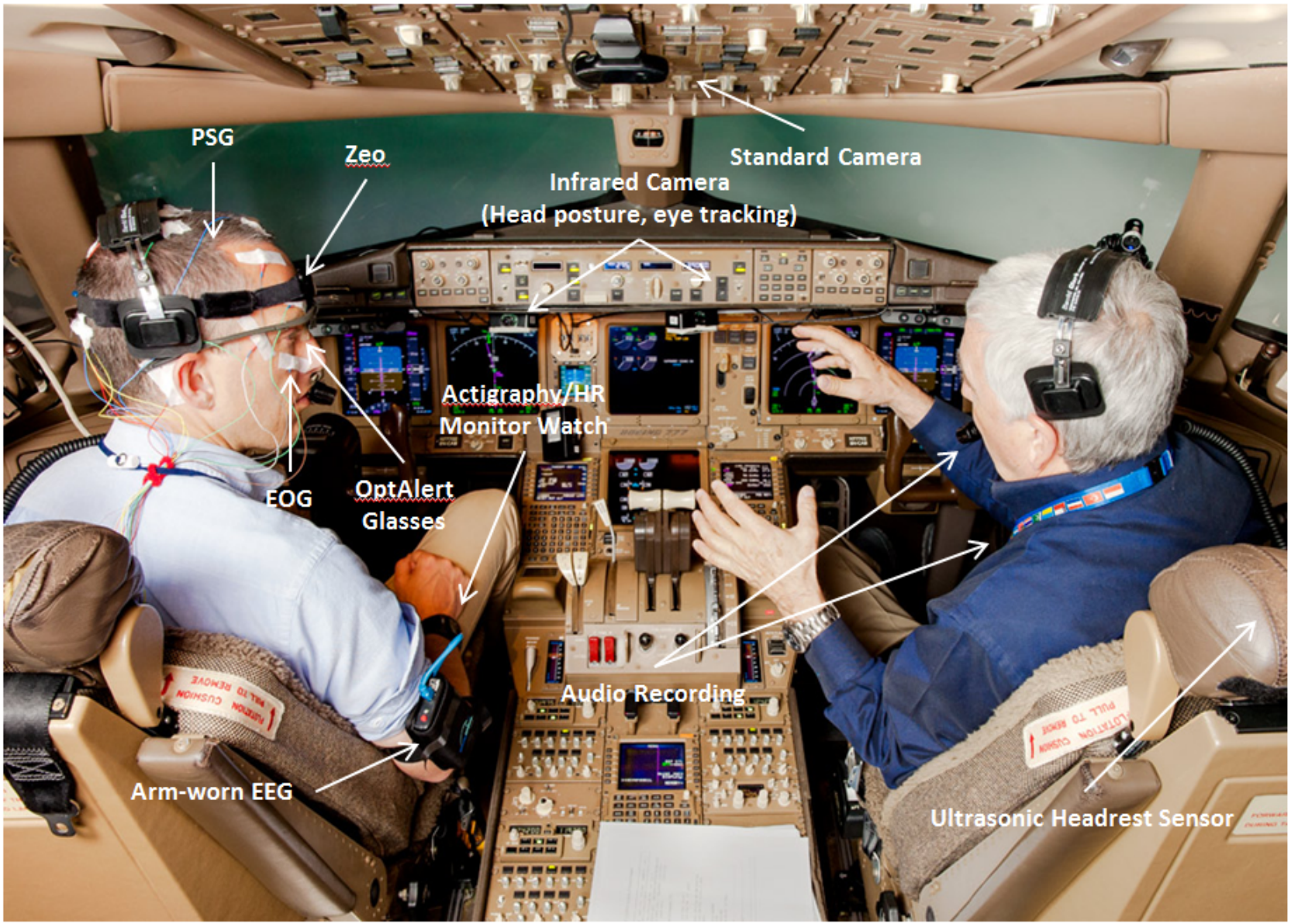
Embedded Measures Experiment - Design

- Single-airline, 777 (long-haul) pilots with common home base, planned 32 crews collected (16 collected to date).
  - 777 pilots tend to be older, more experienced, male (50-62 years).
- 2 weeks normal duty $\Rightarrow$ 72+ hours of rest $\Rightarrow$ simulator flights
- Crossover study design:
  - Each pair of pilots flies either 4 short-haul flights, or 1 long-haul flight in a day.
  - One day is "rested" the other day is "fatigued" (flying overnight after all-day activities). Within short-haul/long-haul designation, half of the pilots fly rested-first, half fly fatigued-first
  - Flights are mostly "standard", but with minor opportunities for error included (weather delay, passenger in labor, etc.)

Embedded Measures Experiment - Data

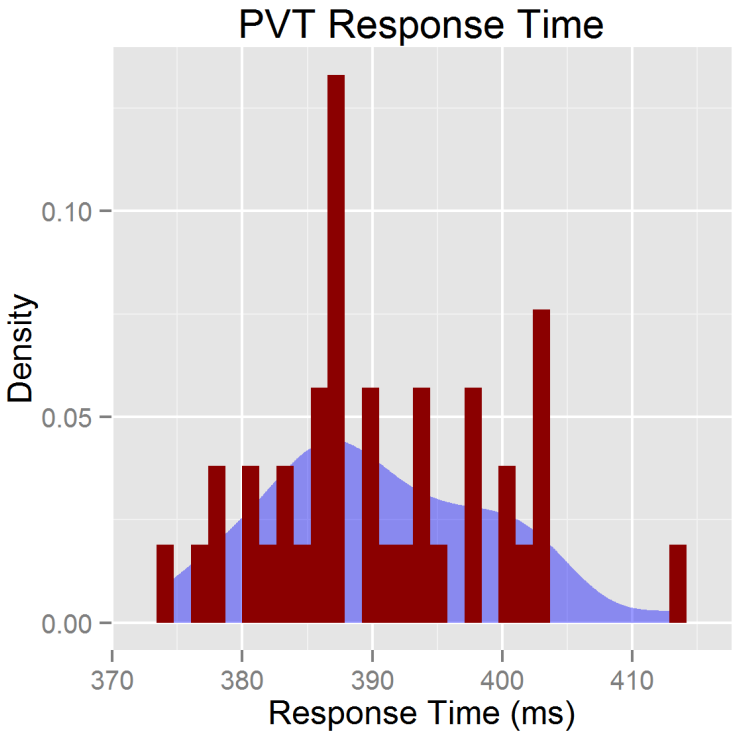
- EEG
  - PSG
  - Zeo
  - BodyWave
- ECG
- EOG
- % Eye Closure
- Head posture
- Blink rate
- Facial features
- Voice analysis
- BAM
- PVT
- KSS
- ERP
- Subjective scales
- Actigraphy

Embedded Measures Experiment - Data

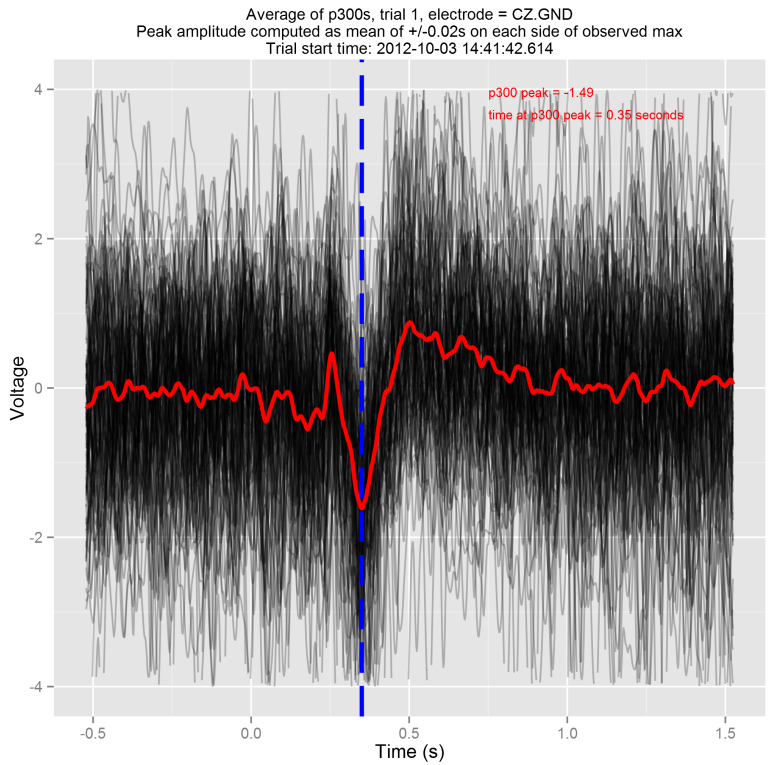


Embedded Measures Experiment - Data

In addition to these highly-sampled data sources, a few discrete measures are also collected



Psychomotor Vigilance Task



Event-related Potential

Embedded Measures Experiment - Data

- Many sensors, many measurements (each), high sampling rate
  - EEG: 8 electrode pairs, 6 measurements each
  - OptAlert: 13 measurements
  - Zeo: 8 measurements
  - ECG: 15 measurements (HRVB)
  - et cetera

Embedded Measures Experiment - Data

- Many sensors, many measurements (each), high sampling rate
  - EEG: 8 electrode pairs, 6 measurements each
  - OptAlert: 13 measurements
  - Zeo: 8 measurements
  - ECG: 15 measurements (HRVB)
  - et cetera
- Many observations, correlated in time, each may or may not be informative at a given time point

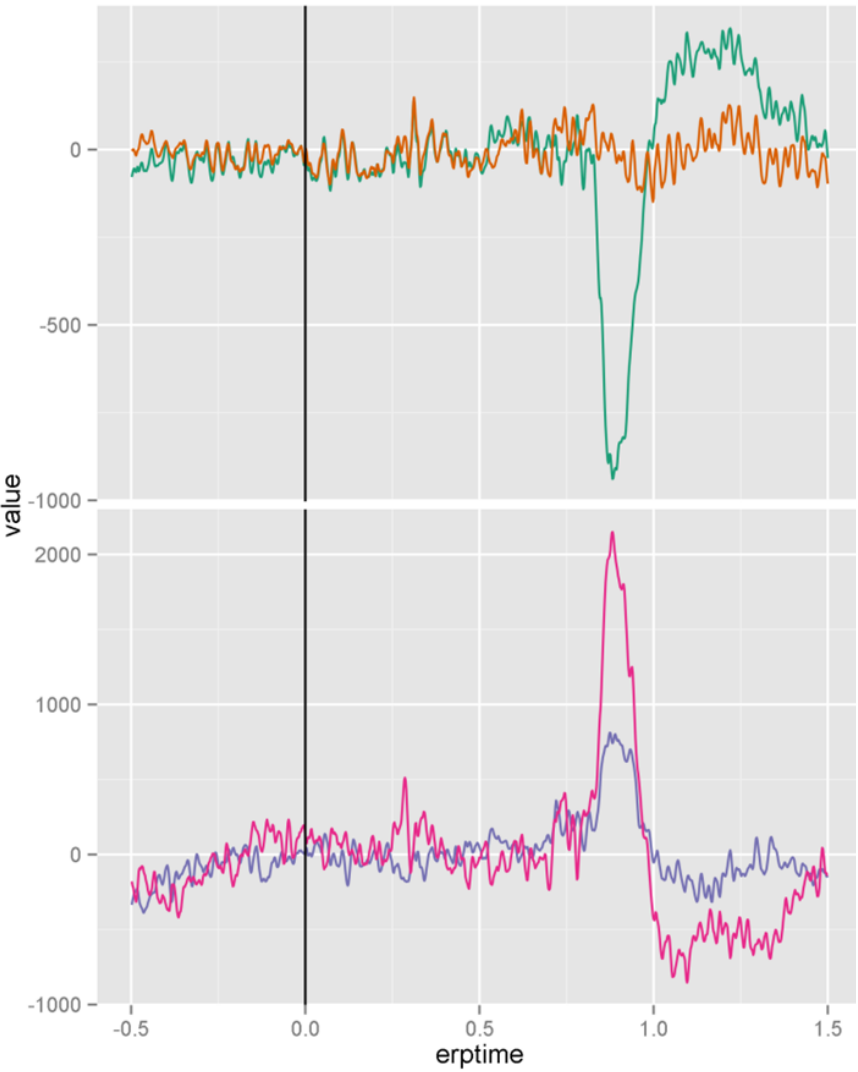
Embedded Measures Experiment - Data

- Many sensors, many measurements (each), high sampling rate
  - EEG: 8 electrode pairs, 6 measurements each
  - OptAlert: 13 measurements
  - Zeo: 8 measurements
  - ECG: 15 measurements (HRVB)
  - et cetera
- Many observations, correlated in time, each may or may not be informative at a given time point

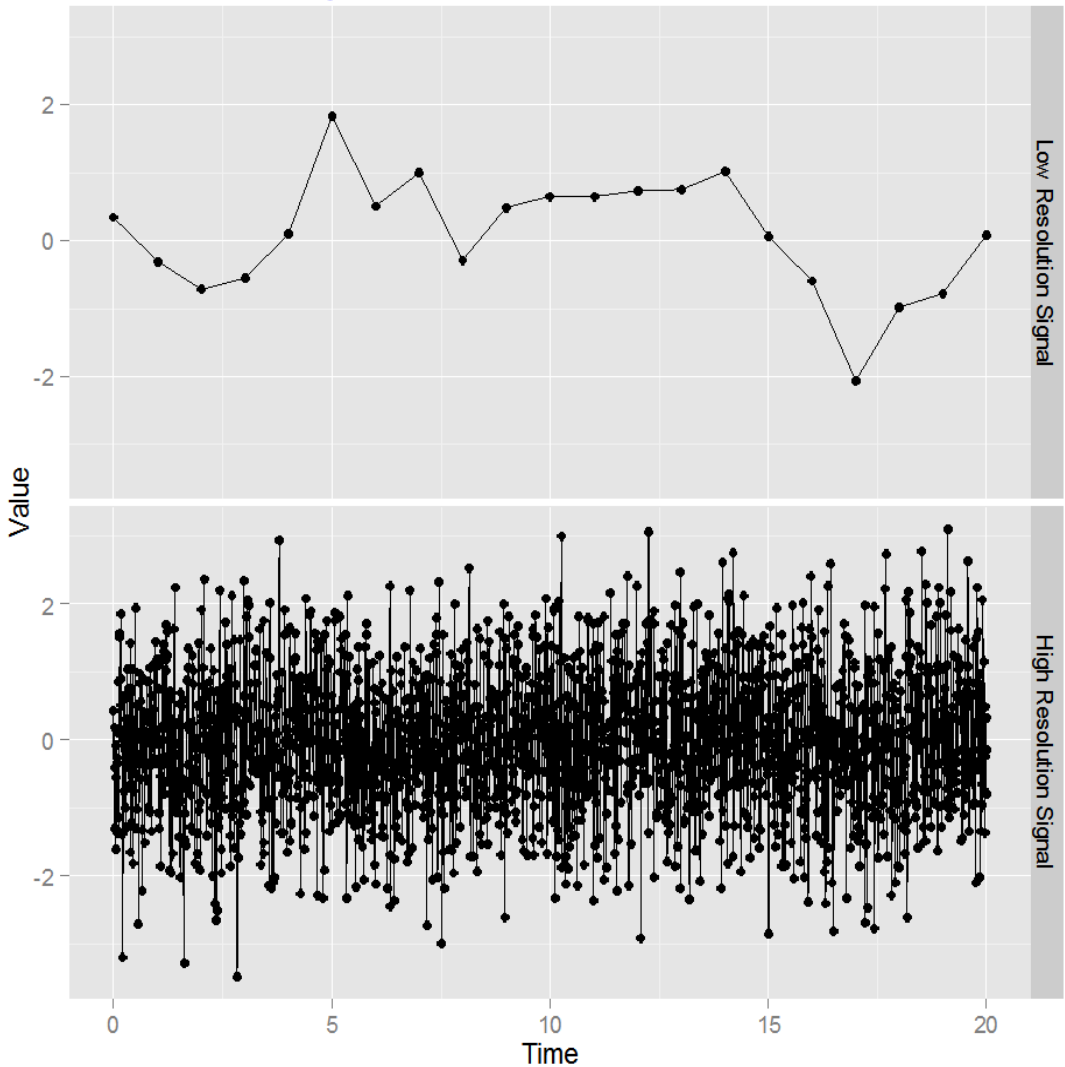
- Challenge: How to efficiently and effectively use these measurements to produce an estimate of fatigue/alertness?

Embedded Measures Expt. - Data Prep

EEG Correction with EOG



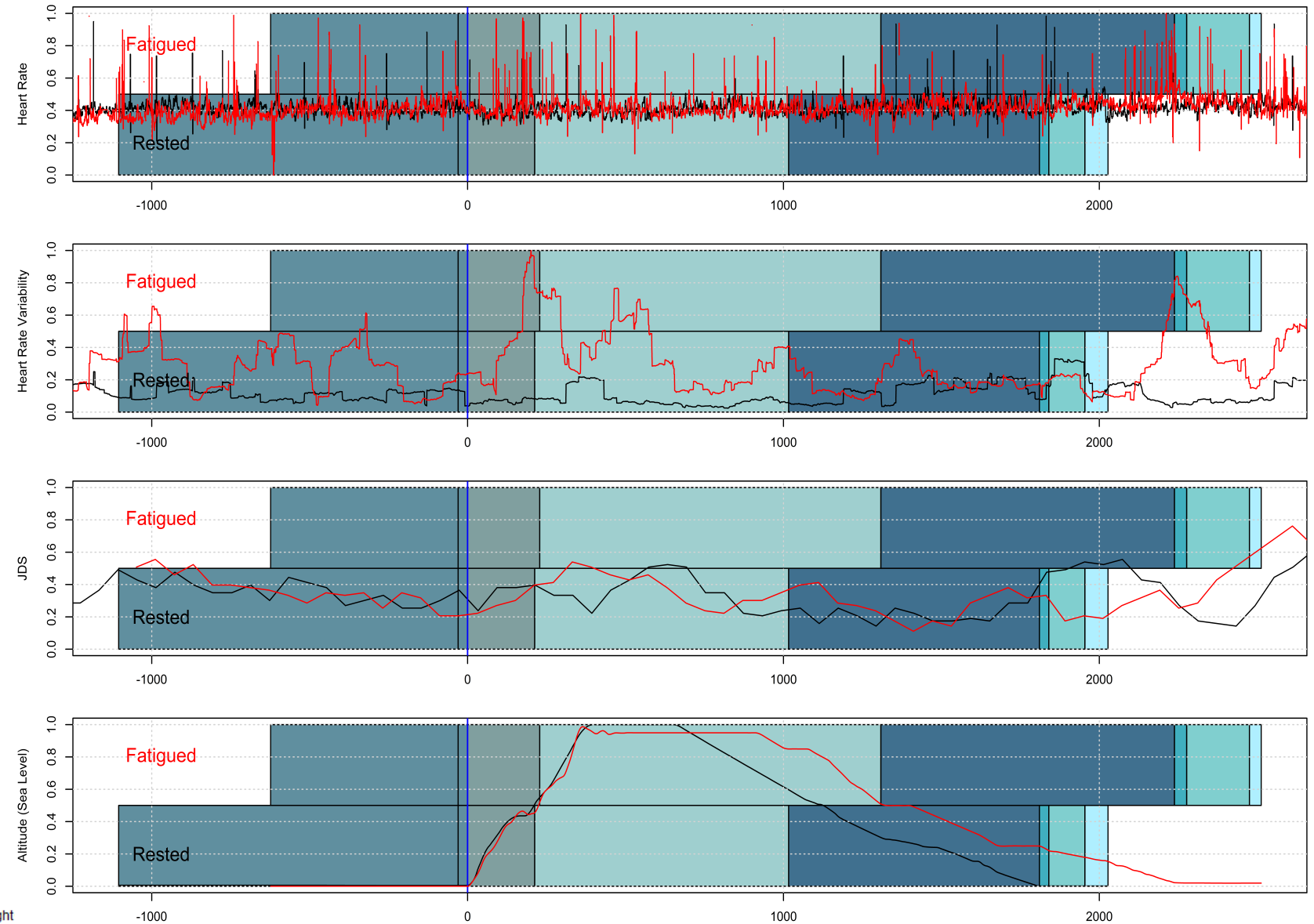
Signal Interpolation



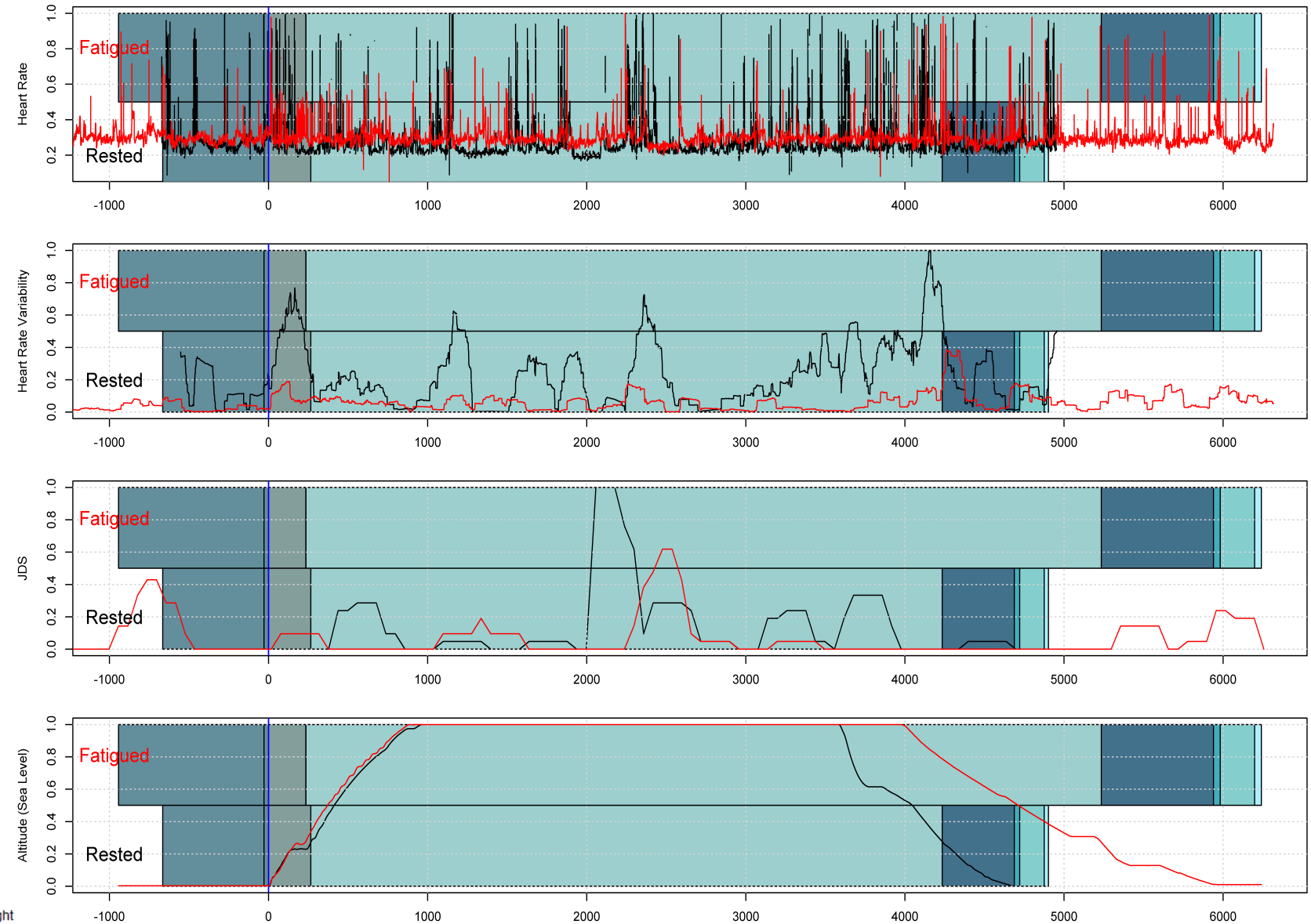
Embedded Measures Expt. - Fatigue Indicators

At this point, I'd like to start talking about statistical methods, but there's a road block (intellectual property), so instead, I'll talk about the challenges we've encountered, and general approaches to challenges of this nature. Then I'll share some results.

What Doesn't Work - Pilot 1



What Doesn't Work - Pilot 2



Embedded Measures Expt. - Fatigue Indicators

- In practice, many of these frequently-sampled signals are reduced to something at a lower sampling rate
  - EEG may be reduced to delta, alpha, theta, beta, sigma power bands at, say, 1/12Hz
  - OptAlert's 13 measurements are optionally automatically reduced to the JDS, at 1/60Hz

Embedded Measures Expt. - Fatigue Indicators

- In practice, many of these frequently-sampled signals are reduced to something at a lower sampling rate
  - EEG may be reduced to delta, alpha, theta, beta, sigma power bands at, say, 1/12Hz
  - OptAlert's 13 measurements are optionally automatically reduced to the JDS, at 1/60Hz
- In these "individual measurements", there is often a lot of noise
  - EEG is occluded by artifact (motion)
  - OptAlert measurements will change if a pilot is reading a chart or turning to talk to his co-pilot

Embedded Measures Expt. - Fatigue Indicators

- In practice, many of these frequently-sampled signals are reduced to something at a lower sampling rate
  - EEG may be reduced to delta, alpha, theta, beta, sigma power bands at, say, 1/12Hz
  - OptAlert's 13 measurements are optionally automatically reduced to the JDS, at 1/60Hz
- In these "individual measurements", there is often a lot of noise
  - EEG is occluded by artifact (motion)
  - OptAlert measurements will change if a pilot is reading a chart or turning to talk to his co-pilot
- How much do we trust each observation to be a "measurement" as opposed to considering it an "estimate"?
- If each is an "estimate", perhaps the estimate can be improved by considering its neighboring observation(s).

Embedded Measures Expt. - Fatigue Indicators

- In addition, what if a pilot's fatigue/alertness state is not only a function of (or indicated by) the current epoch's data? What if it is a combination of instantaneous data and a cumulative component over some unknown duration of recent data?

Embedded Measures Expt. - Fatigue Indicators

- In addition, what if a pilot's fatigue/alertness state is not only a function of (or indicated by) the current epoch's data? What if it is a combination of instantaneous data and a cumulative component over some unknown duration of recent data?

- So we should have a method for considering the possibility that historical data from one variable may be combined with current or historical data from another variable to provide a good indication of fatigue.
  - Smoothing, aggregation, etc. Additional dimension(s) requiring tuning.
  - Naturally, this greatly expands the search for a predictive model

Embedded Measures Expt. - Response Variable

- Fatigue is difficult to define
- It has both mental and physical components, and is manifest both mentally and physically
- But if we're trying to build a model to predict fatigue, we need known cases of fatigue, and known cases of non-fatigue.

Embedded Measures Expt. - Response Variable

- Fatigue is difficult to define
- It has both mental and physical components, and is manifest both mentally and physically
- But if we're trying to build a model to predict fatigue, we need known cases of fatigue, and known cases of non-fatigue.
- Additional difficulty: We're trying to study fatigue in an *operational* context
  - This is not a sleep study. In real life, pilots will drink coffee. In real life, pilots will have varying degrees of sleep quality prior to duty.
  - While we tried to control variables as much as was reasonable (no duty prior to study flights, in bed at certain times, awake at certain times), fatigue/alertness is very difficult to control.

Embedded Measures Expt. - Response Variable

- Fatigue is difficult to define
- It has both mental and physical components, and is manifest both mentally and physically
- But if we're trying to build a model to predict fatigue, we need known cases of fatigue, and known cases of non-fatigue.
- Additional difficulty: We're trying to study fatigue in an *operational* context
  - This is not a sleep study. In real life, pilots will drink coffee. In real life, pilots will have varying degrees of sleep quality prior to duty.
  - While we tried to control variables as much as was reasonable (no duty prior to study flights, in bed at certain times, awake at certain times), fatigue/alertness is very difficult to control.
- Fundamentally, fatigue is influenced by personal differences. The same duration of sleep will produce a higher level of alertness in one person than another, due to complex biological differences.

Embedded Measures Expt. - Response Variable

We have a couple of different options:

1. Use our experimental conditions to define fatigue:
   - At the beginning of their rested flight, pilots are as rested as they will be during our experiment
   - At the end of their fatigued flight, they are as fatigued as they will be during our experiment
   - We may then sample epochs of data from each of these periods, and train a classifier on this response variable

Embedded Measures Expt. - Response Variable

We have a couple of different options:

1. Use our experimental conditions to define fatigue:
   - At the beginning of their rested flight, pilots are as rested as they will be during our experiment
   - At the end of their fatigued flight, they are as fatigued as they will be during our experiment
   - We may then sample epochs of data from each of these periods, and train a classifier on this response variable
2. Use self-report to define fatigue:
   - Assumption: In a study such as this, a pilot has little incentive to hide fatigue level, so they will be honest about the fatigue they feel they are experiencing
   - We can use a variety of "validated" self-report scales of fatigue to train a model to predict fatigue level. We choose the Karolinska Sleep Scale (1-9).
   - Note that this is only useful if we assume there *is* incentive during normal operations for pilots to hide their level of fatigue (perhaps pilots fear some kind of reprisal for refusing to fly when they're fatigued).
   - Otherwise, we're just producing an estimate of something the pilots could tell us instantly with no variability.

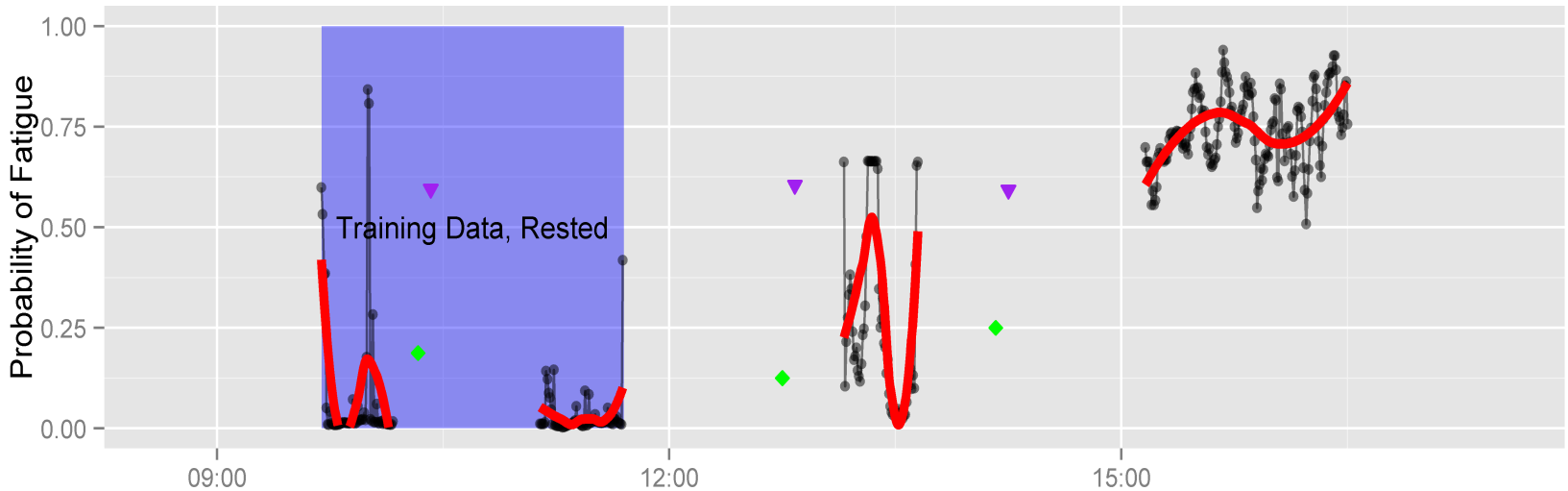
Methods

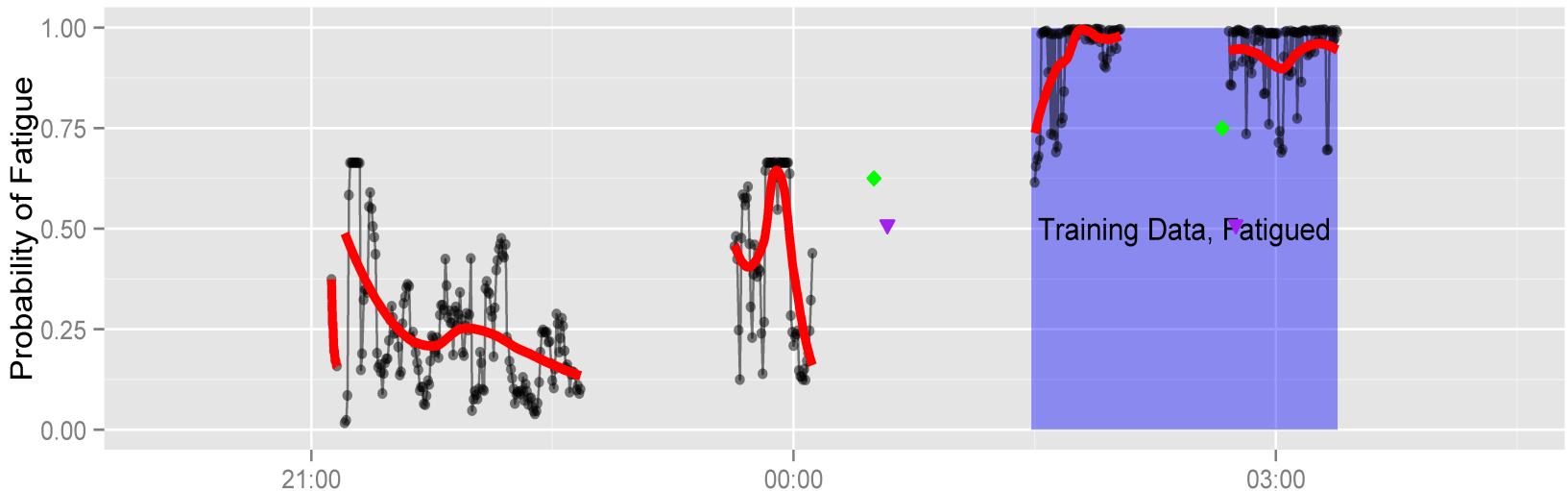
- We have hundreds of thousands of observations on thousands of variables, collected over two flight sessions
- Clearly some form of dimensionality-reduction is important
- Candidates are kernel-based forms of PCA, PLS, or methods that incorporate parameter estimate shrinkage
  - Choice of this method would be a good candidate for optimization

Methods

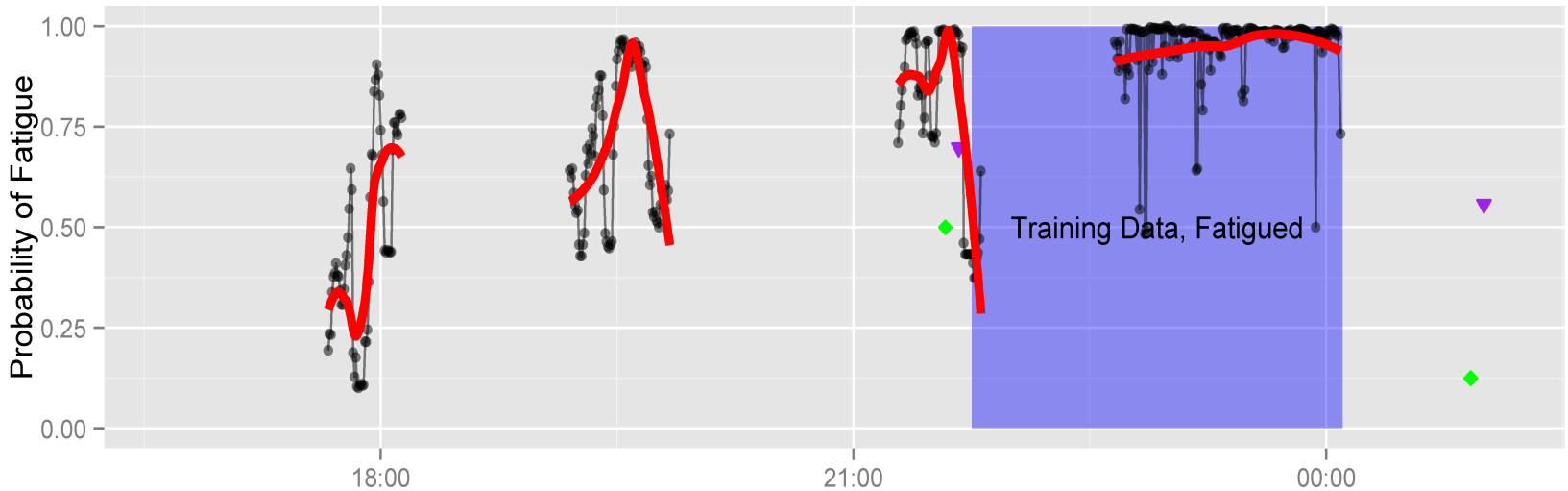
- We have hundreds of thousands of observations on thousands of variables, collected over two flight sessions
- Clearly some form of dimensionality-reduction is important
- Candidates are kernel-based forms of PCA, PLS, or methods that incorporate parameter estimate shrinkage
  - Choice of this method would be a good candidate for optimization
- After making that decision, the problem becomes a classification or regression-type problem, based on the nature of the response variable selected.
- The following results consider the classification form of the problem, with two classes (fatigued vs. not fatigued), trained on an investigator-chosen section of data suspected to represent these two classes. (Combination of experimental conditions and PVT/KSS results.)
  - These do not represent final results.

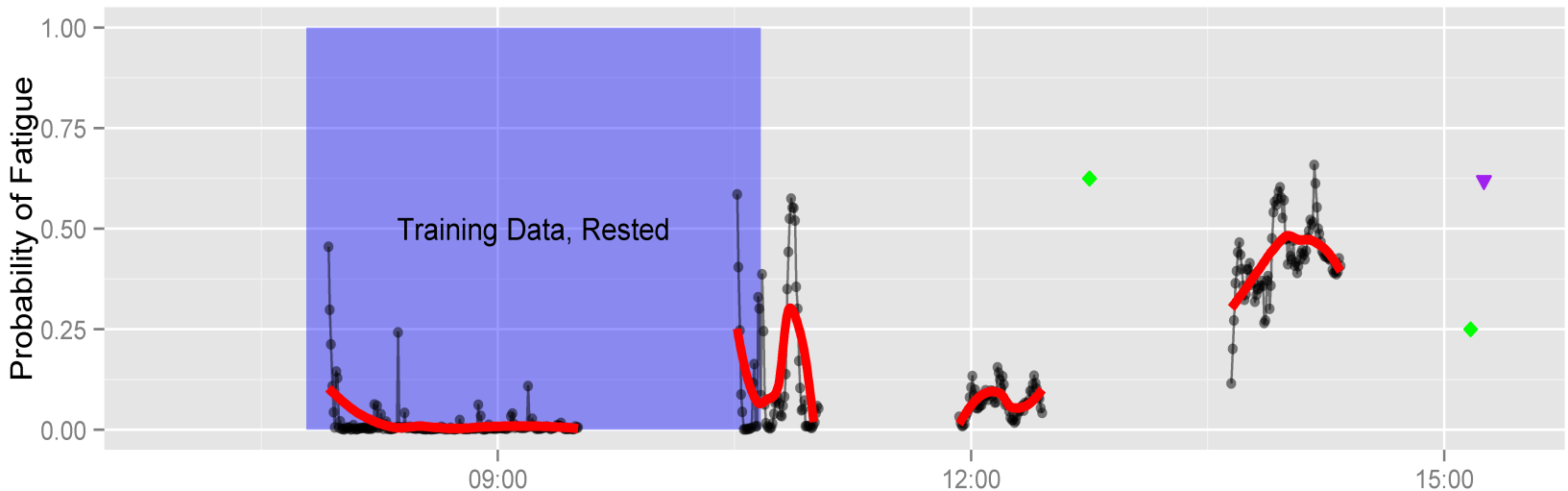
Results: Predicted Fatigue Level



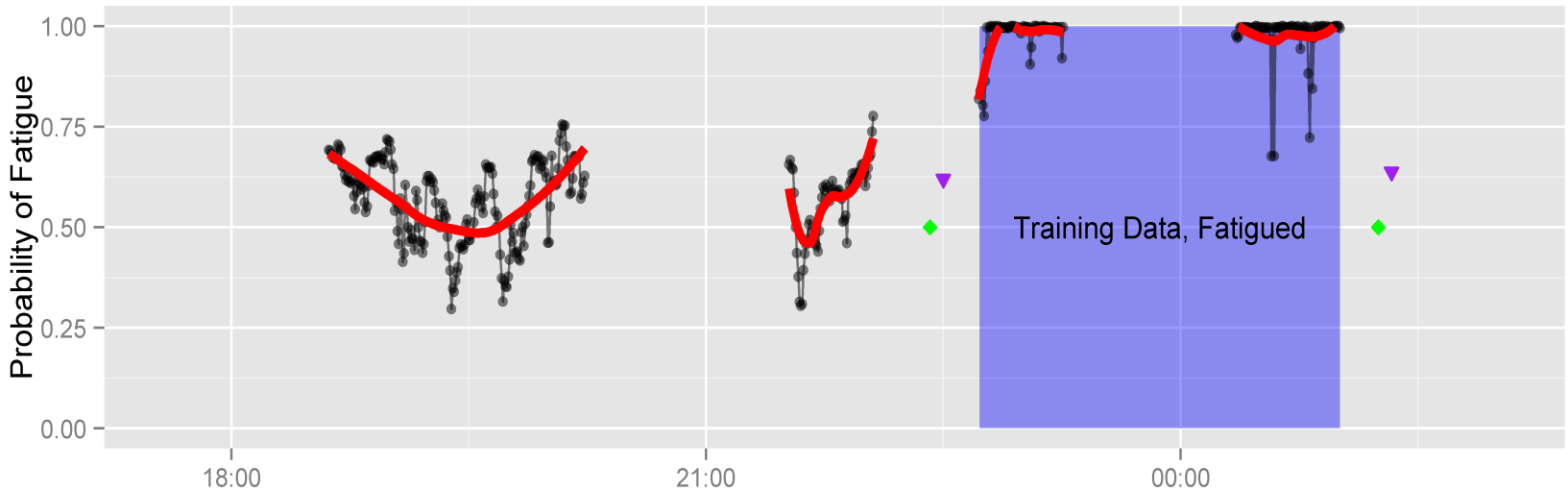


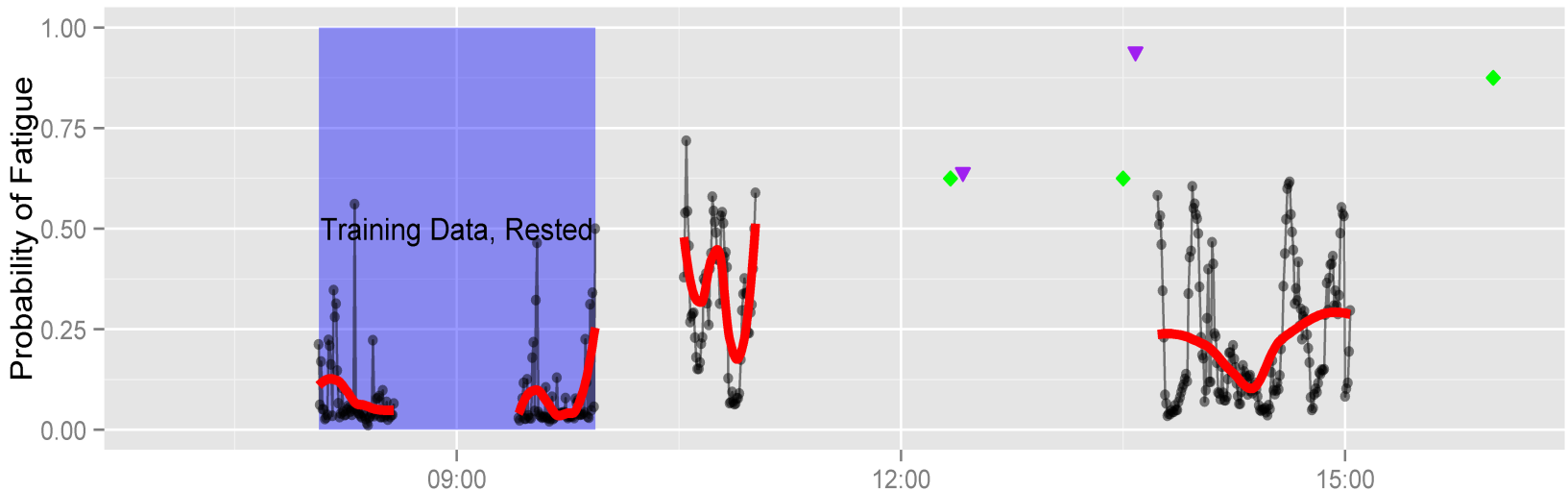
Results: Predicted Fatigue Level



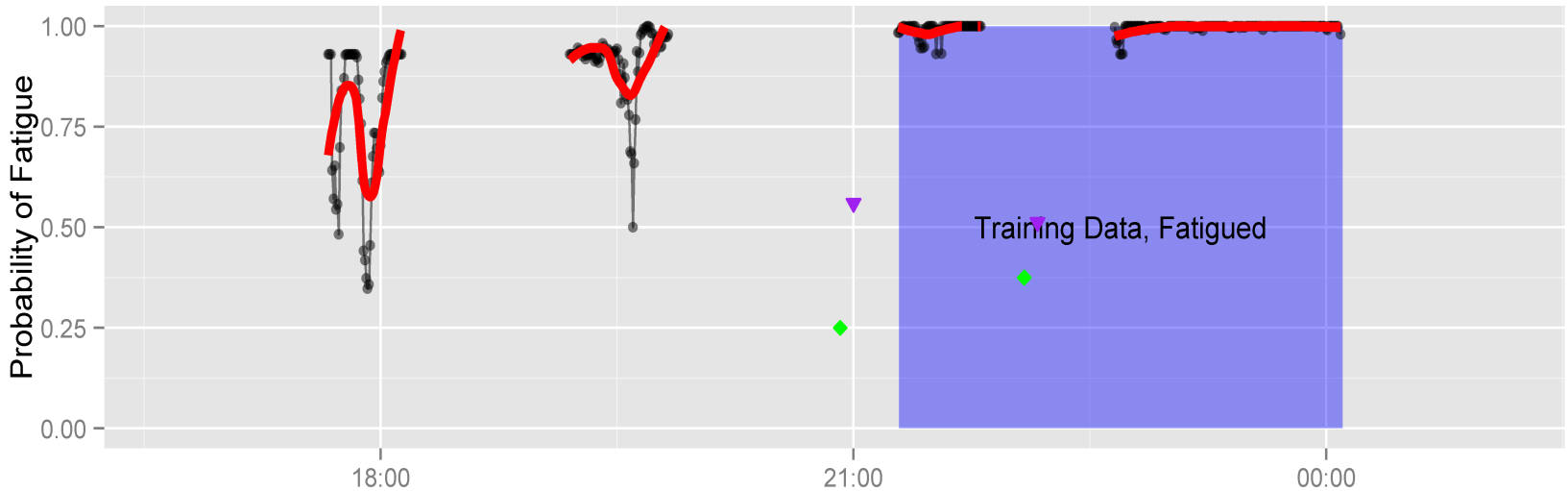


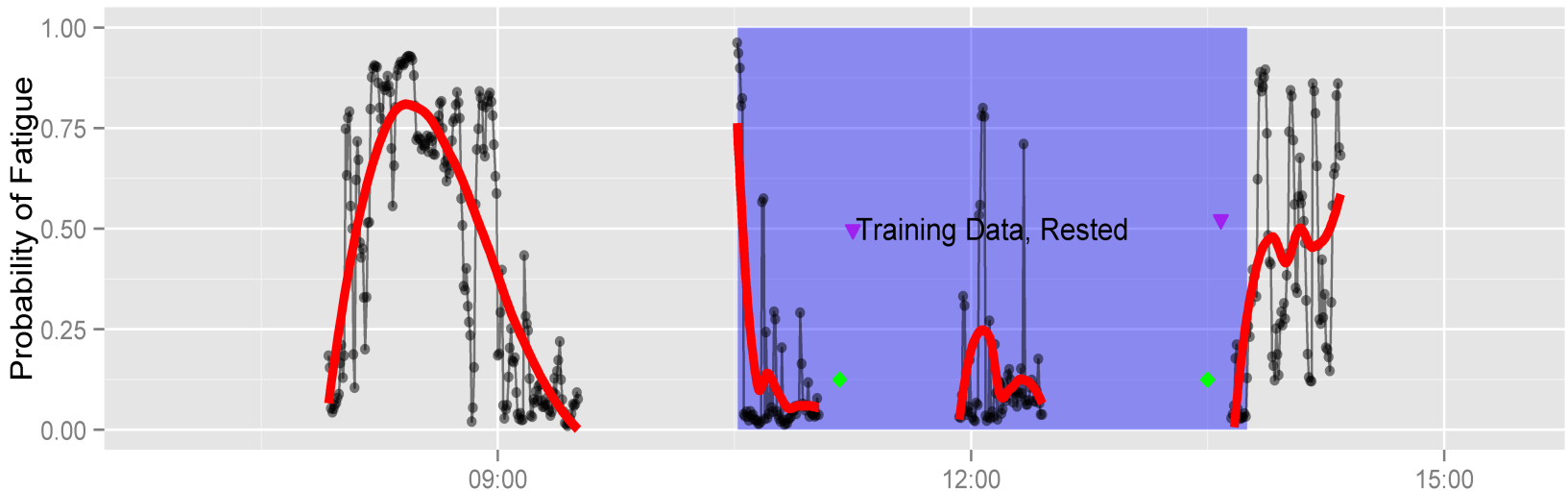
Results: Predicted Fatigue Level



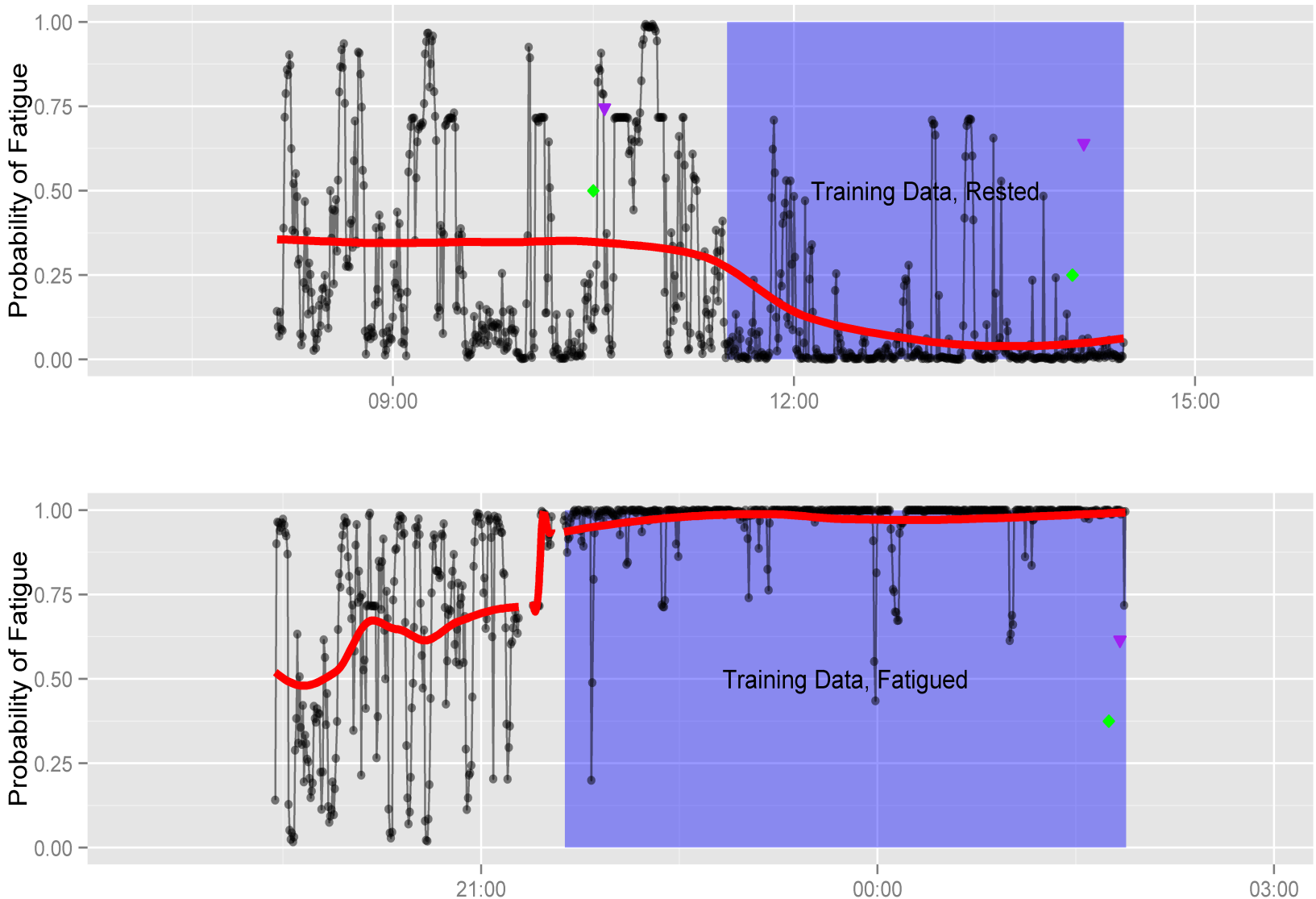


Results: Predicted Fatigue Level

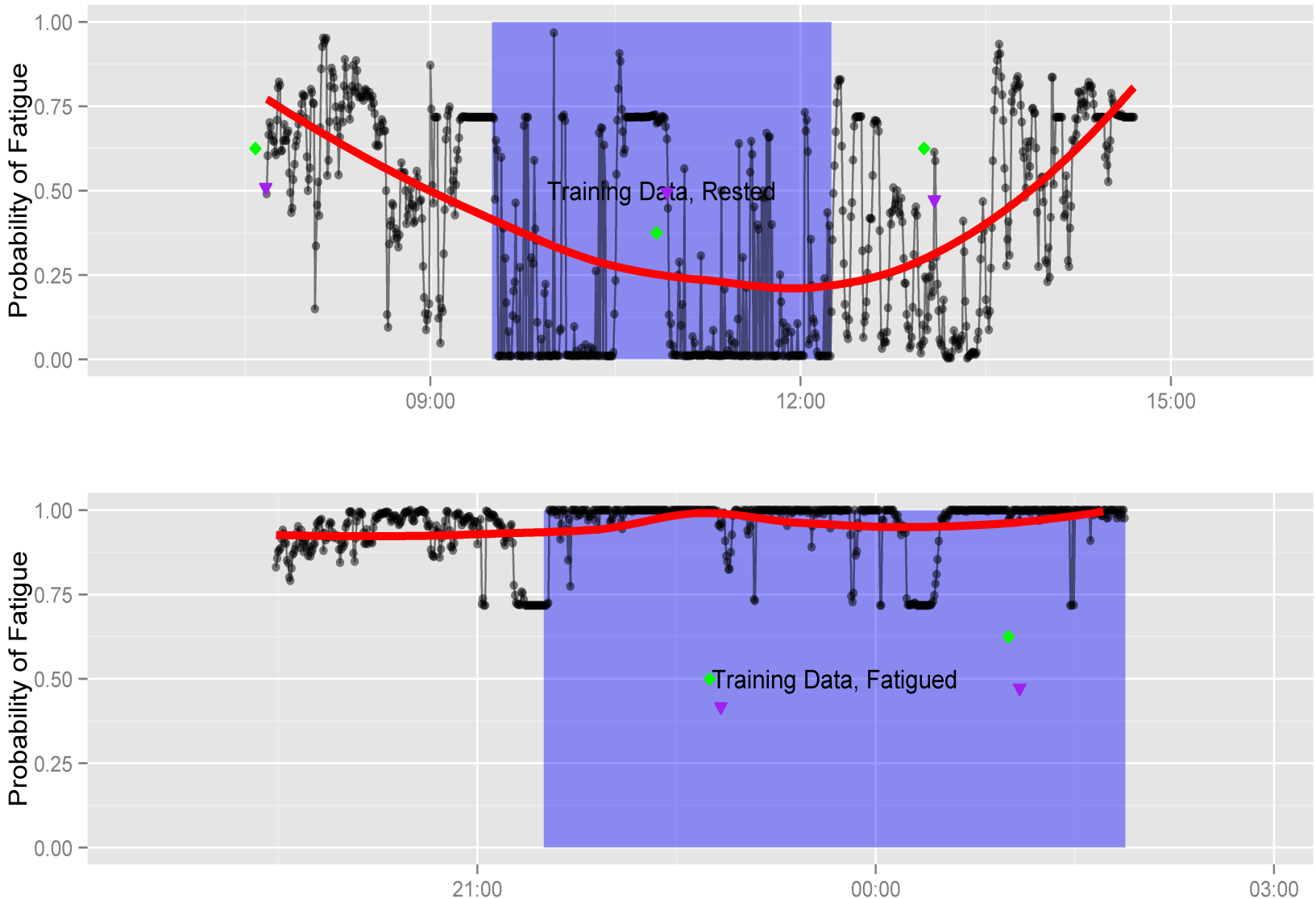




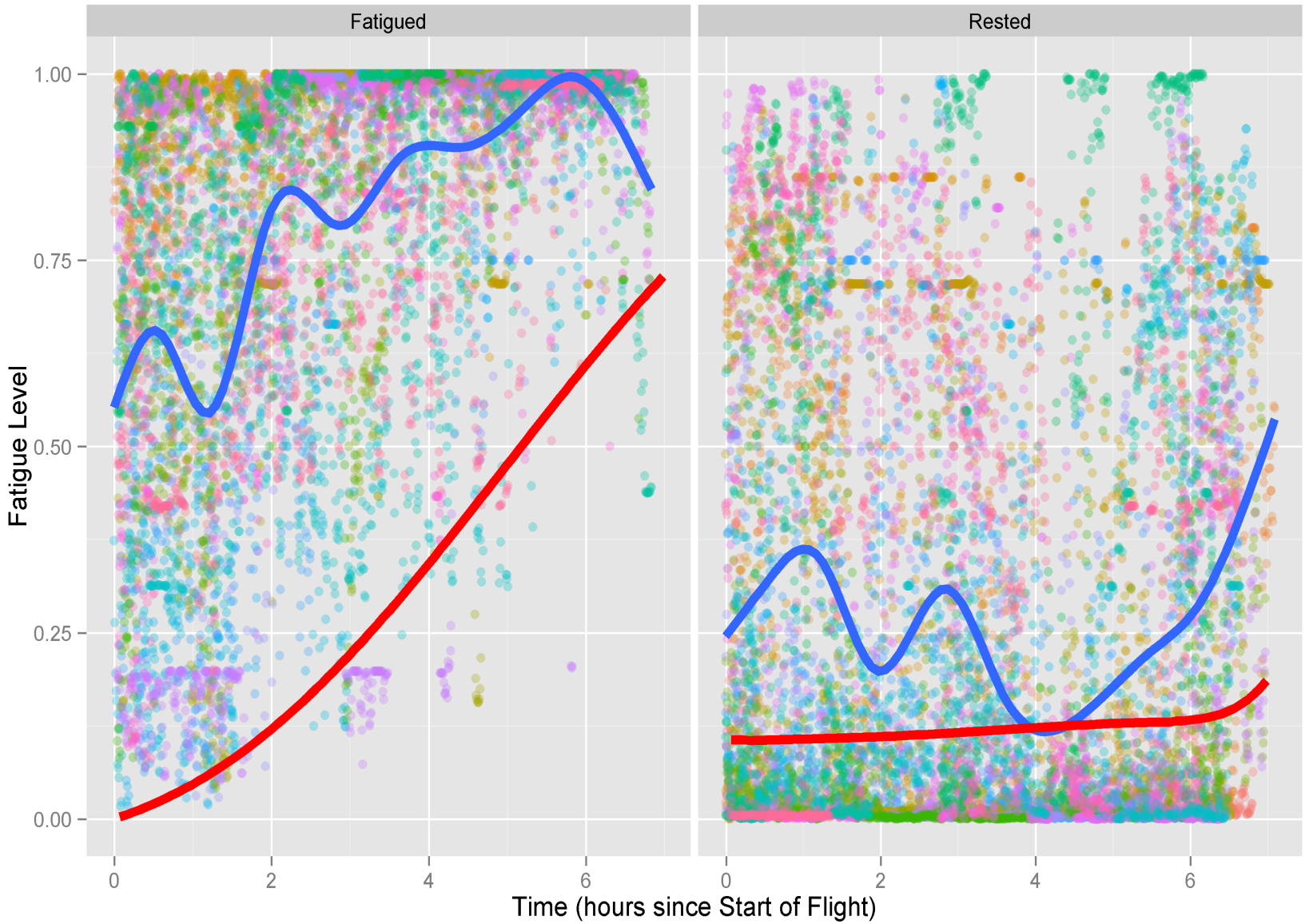
Results: Predicted Fatigue Level



Results: Predicted Fatigue Level



Results: Predicted Fatigue Level + BAM (red)



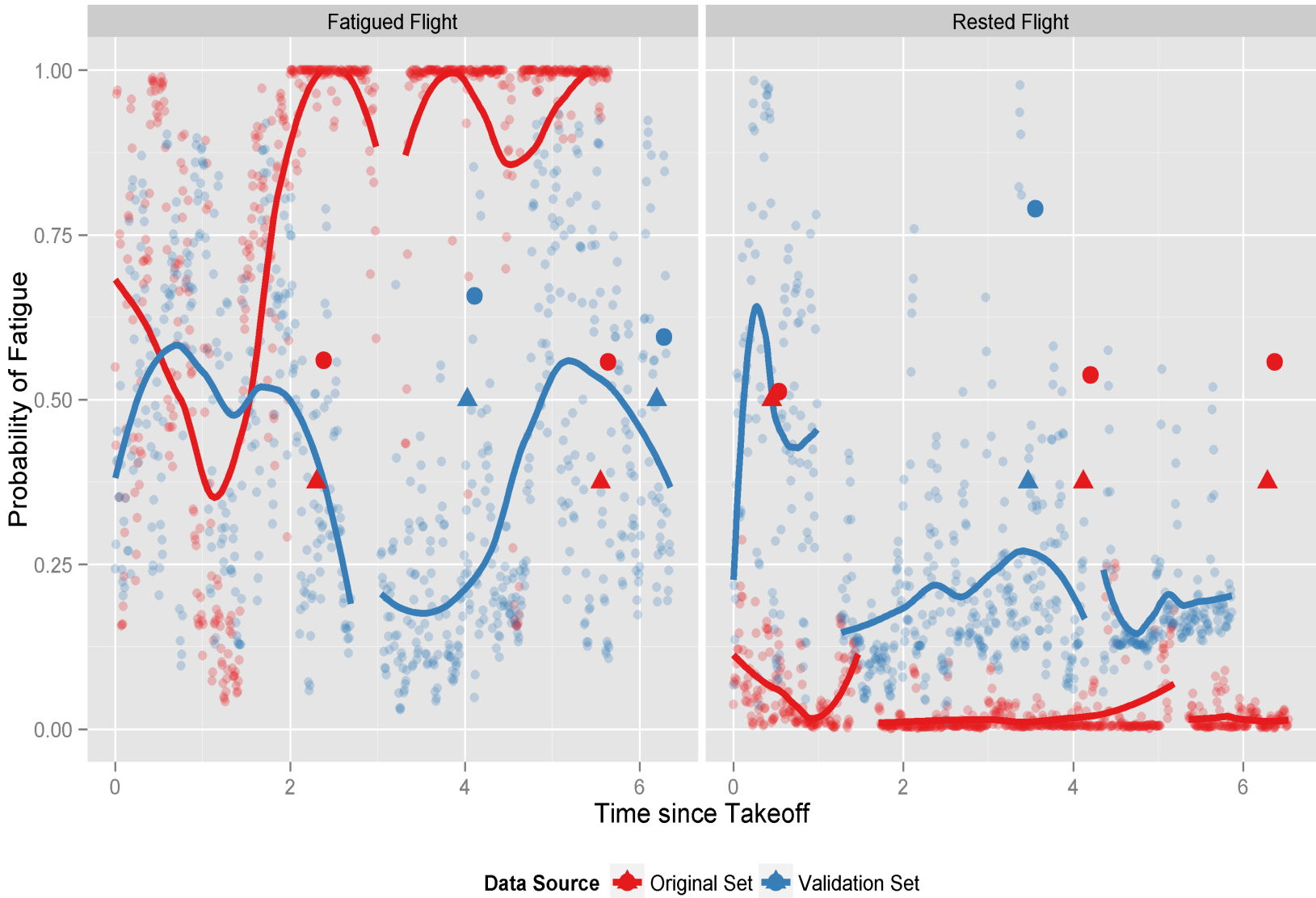
Results

- Some results look good, some results don't. Why?
  - In some cases, an electrode was discovered to have malfunctioned (probably accidentally removed by the pilot while adjusting the headset).
  - In other cases, no obvious cause for poor predictions was discovered.

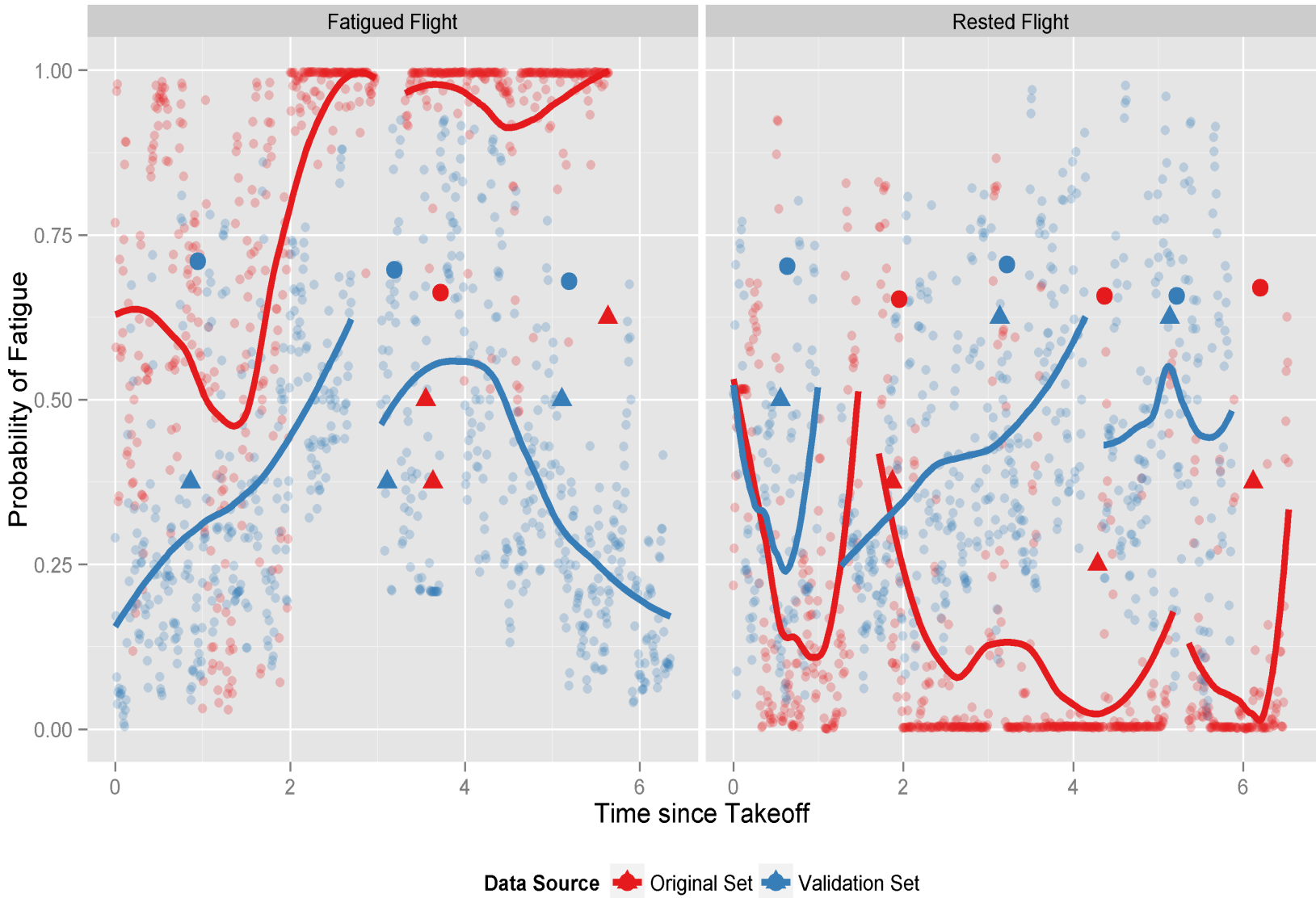
Results

- Some results look good, some results don't. Why?
  - In some cases, an electrode was discovered to have malfunctioned (probably accidentally removed by the pilot while adjusting the headset).
  - In other cases, no obvious cause for poor predictions was discovered.
- There is a suspicion that some of the predictive ability is captured by device setup.
  - Since rested data may be collected on day 1, and fatigued data may be collected on day 2, perhaps the equipment setup (electrode attachment, where they choose to wear their eyeglasses, tightness of the headband/armband EEG), which is performed independently on each day, contributes to the observed difference between rested and fatigued.
- We may investigate this possibility by asking participants to return for another round of data collection, with the hope that either
  1. The new data will corroborate the previous results, OR
  2. The new data will help point to which device (or other data feature) that is causing the malfunctioning predictions

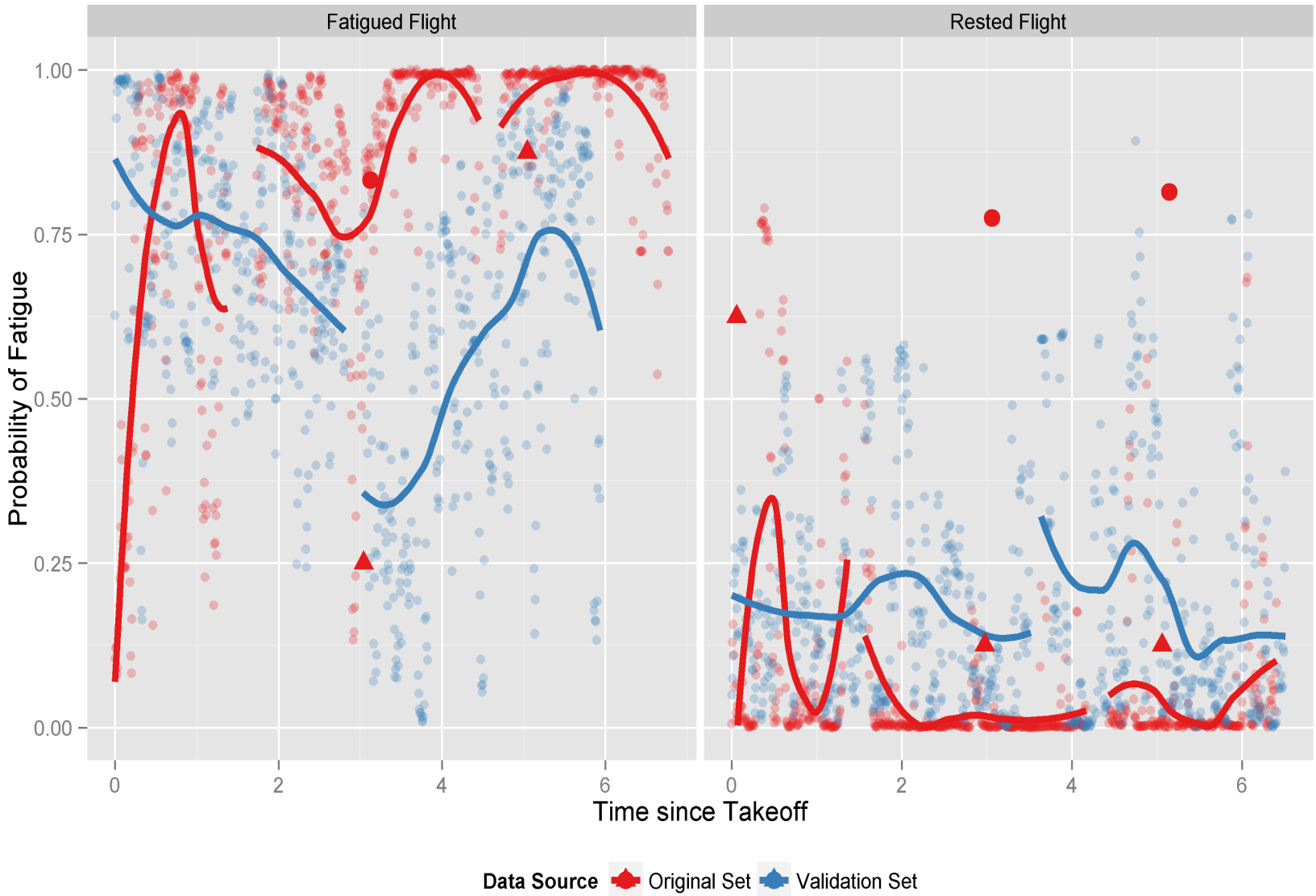
Results: Predicted Fatigue Level, Pilot A



Results: Predicted Fatigue Level, Pilot B



Results: Predicted Fatigue Level, Pilot C



Conclusions

- If fatigue is a smooth, continuous, latent variable, then it appears we are having trouble estimating it well, *all the time*. This could be due to
  - Inadequate experimental data (not fatigued enough)
  - Incorrect sensors used (melatonin? respiration rate? skin conductivity?)
  - Inadequate resolution and stability of physiological sensors
  - Something else

Conclusions

- If fatigue is a smooth, continuous, latent variable, then it appears we are having trouble estimating it well, *all the time*. This could be due to
  - Inadequate experimental data (not fatigued enough)
  - Incorrect sensors used (melatonin? respiration rate? skin conductivity?)
  - Inadequate resolution and stability of physiological sensors
  - Something else
- What's left to do:
  - Rigorous comparison of fatigue predictions to PVT/KSS results
    - Some people refer to PVT as the "gold standard" in fatigue level estimation
  - Investigate other means of producing fatigue level estimates
  - Train predictive model directly to KSS or PVT (ignore experimental conditions)
  - Collect more (and better) data

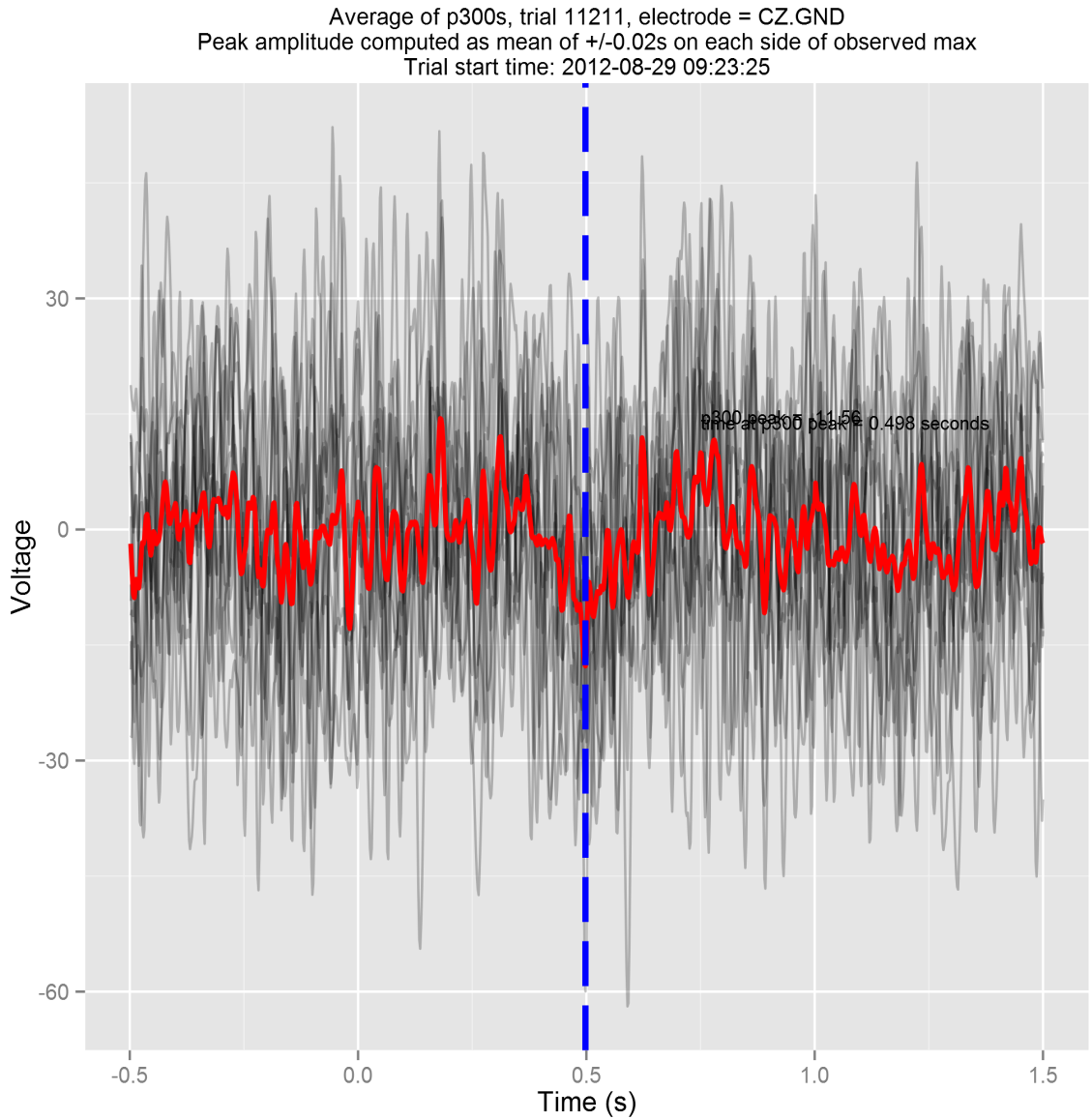
Conclusions

- If fatigue is a smooth, continuous, latent variable, then it appears we are having trouble estimating it well, *all the time*. This could be due to
  - Inadequate experimental data (not fatigued enough)
  - Incorrect sensors used (melatonin? respiration rate? skin conductivity?)
  - Inadequate resolution and stability of physiological sensors
  - Something else
- What's left to do:
  - Rigorous comparison of fatigue predictions to PVT/KSS results
    - Some people refer to PVT as the "gold standard" in fatigue level estimation
  - Investigate other means of producing fatigue level estimates
  - Train predictive model directly to KSS or PVT (ignore experimental conditions)
  - Collect more (and better) data
  - Is (estimated) fatigue level associated with performance degradation?

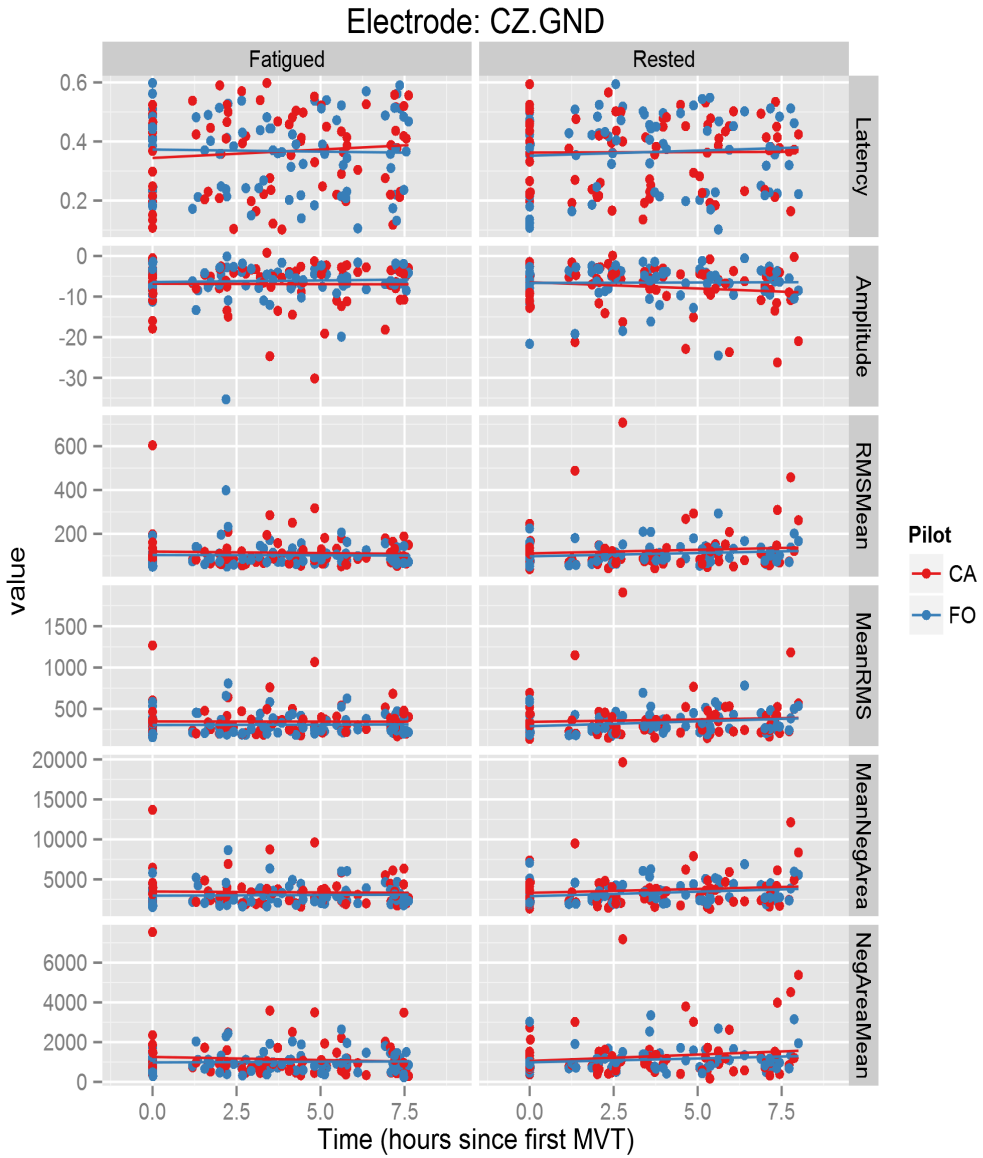
The End

Questions?

Event-Related Potential (VEP)



Event-Related Potential (VEP)



Event-Related Potential (VEP)