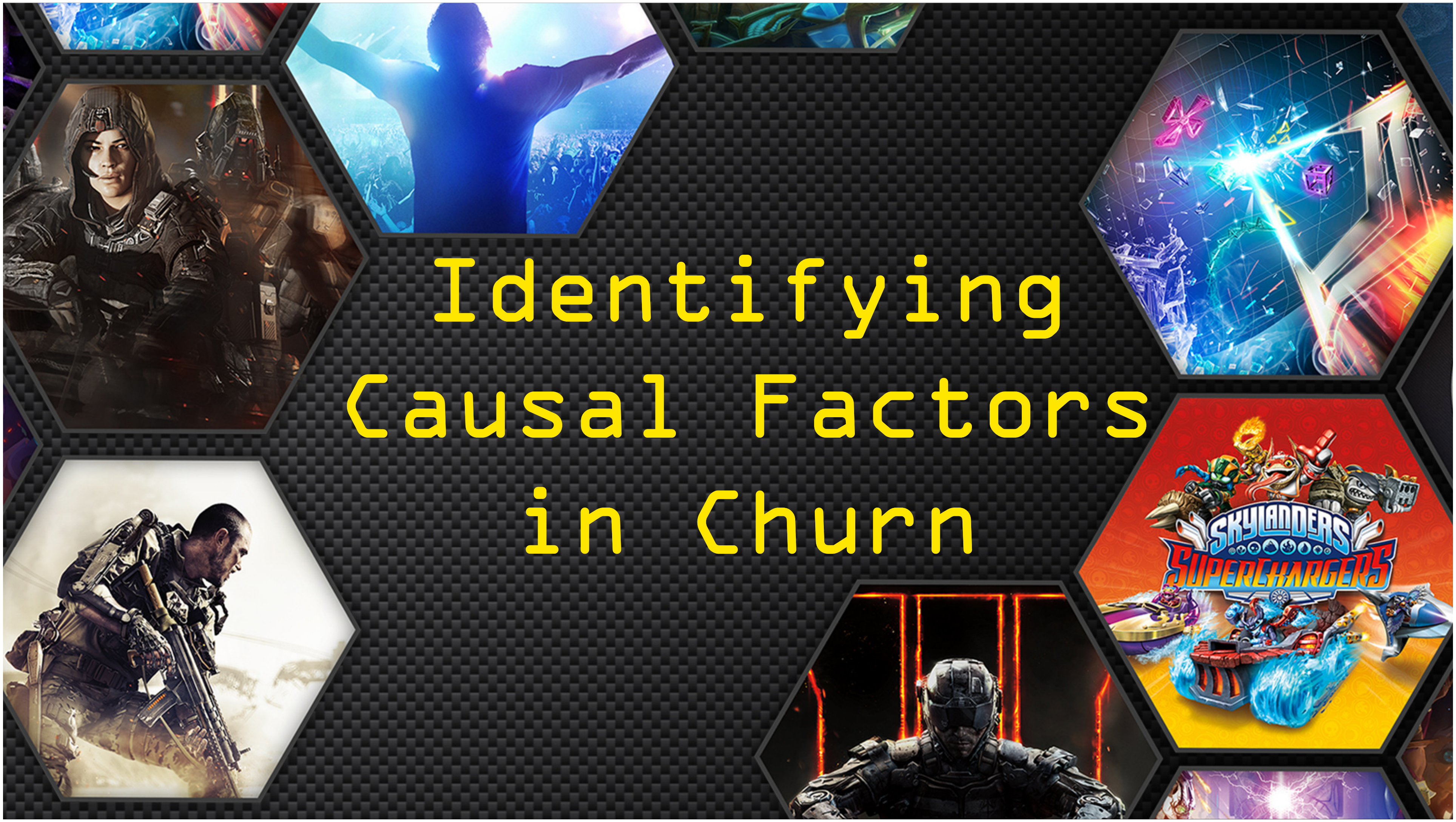
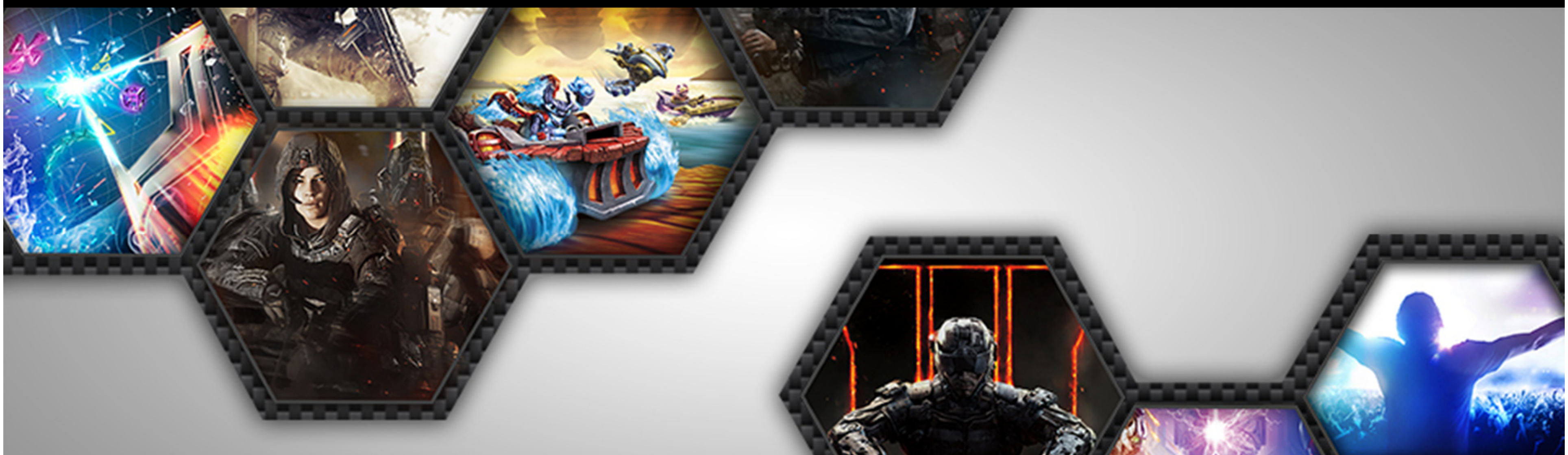

Identifying
Causal Factors
in Churn
SKYLANDERS
SUPERCHARGERS

Overview of Core Insights

Alan Burke – Manager, Analytic Services



ACTIVISION

Scenario 1 – Ad Campaign

Scenario 1 Background

- Your company's user acquisition team launches a new ad campaign to drive acquisition.
- Conversion rates are good, but are you acquiring quality players?
- Create a fast identifiable model for churn on limited data.
- Data is hypothetical, but the scenario is based on real world experiences.

Requirements

- Estimate the effect of the campaign within 10 days of launch.
- Use churn rates as a proxy for quality of player.

Why use a survivor model?

- Every player (data point) is valuable to adding to the accuracy of the model.
- Right censoring is a major obstacle.
- Implementation if fairly straight forward and quick.

Why not use something easier?

1. Why not compare mean time-to-event between your groups using a t-test or linear regression?

-- ignores censoring

2. Why not compare proportion of events in your groups using risk/odds ratios or logistic regression?

--ignores time

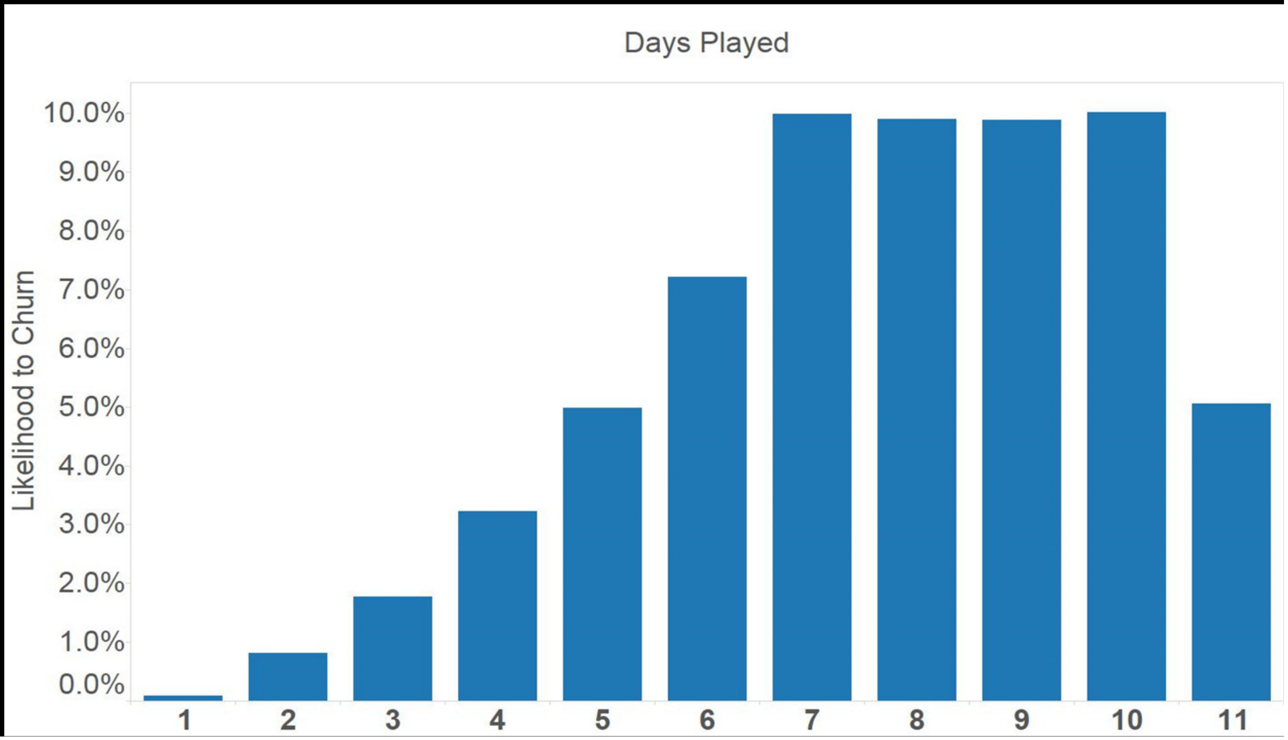
Introduction to survival distributions

- T_i the event time for an individual, is a random variable having a probability distribution.
- Different models for survival data are distinguished by different choice of distribution for $T_{i.}$

Probability density function: f(t)

In this example, the longer players play, the more likely they are to churn each day, except for day 11.

Hypothetical data:



Probability density function: f(t)

The probability of the failure time occurring at exactly time t (out of the whole range of possible t's).

$$f(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

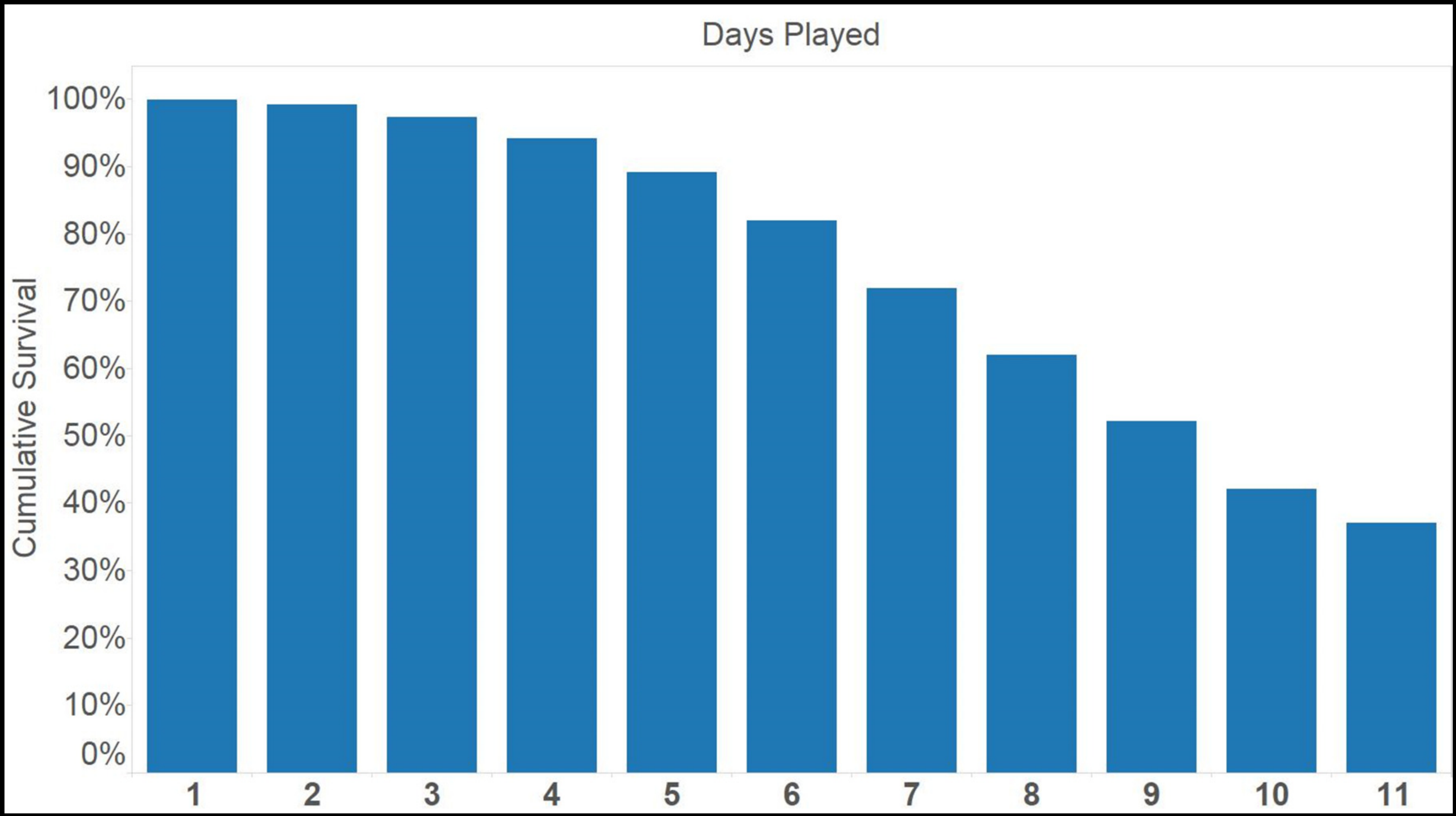
Cumulative distribution function: F(t)

At a given point in time t, what is the likelihood that failure has occurred.

$$F(t) = P(T \leq t) = \int_{-\infty}^{t} f(u)\,du$$

ACTIVISION ▸ STUDIOS

Survival function: S(t)



ACTIVISION STUDIOS

Survival function: S(t)

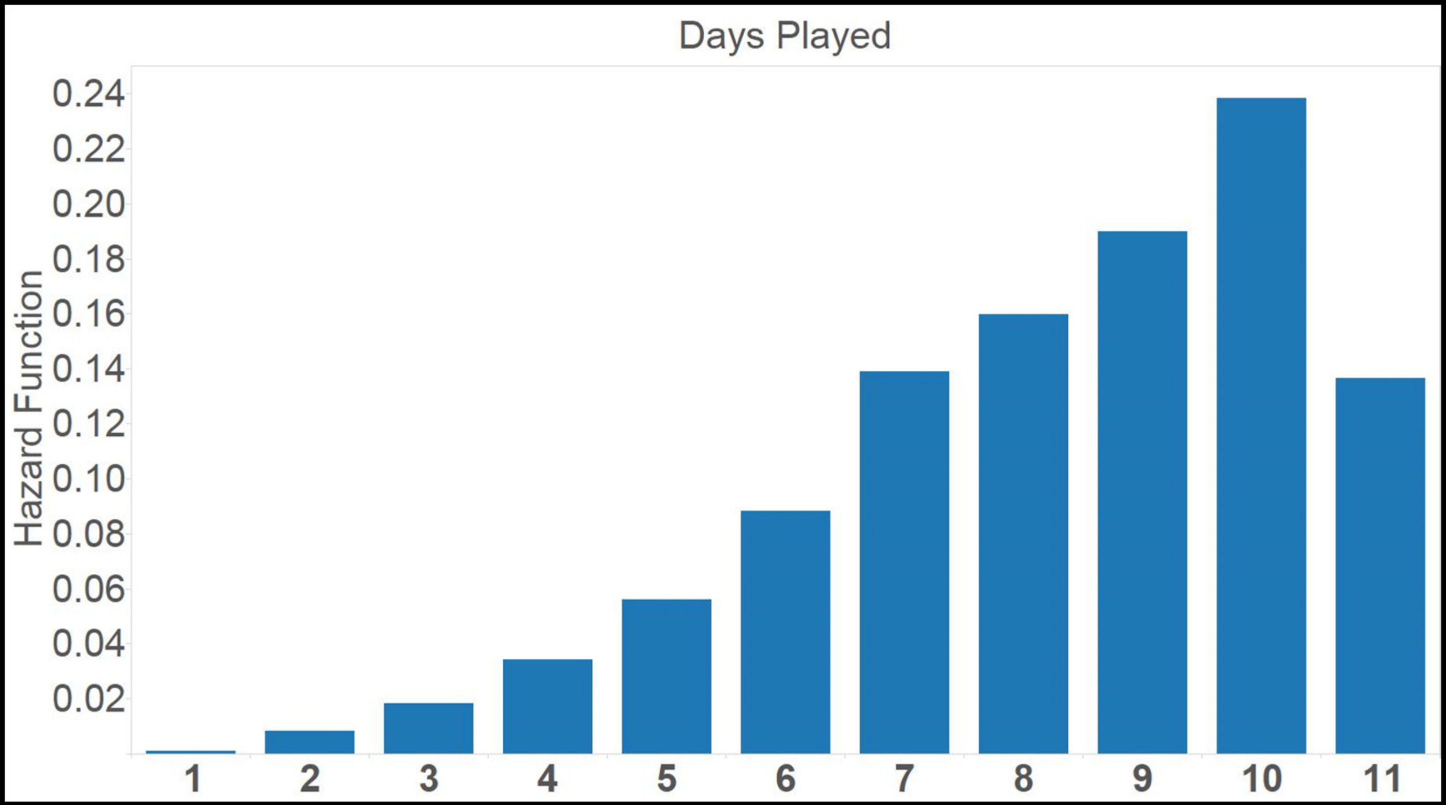
The goal of survival analysis is to estimate and compare survival experiences of different groups.

Survival experience is described by the cumulative survival function:

$$S(t) = 1 - P(T \le t) = 1 - F(t)$$

Example: If t=5 days, S(t=5) = probability of still playing beyond 5 days.

Hazard Function



Hazard rate is an instantaneous incidence rate.

Hazard function

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t}$$

<u>In words:</u> the probability that ***if you keep playing to t***, you will churn in the next instant.

Deriving hazard function from density and survival functions:

$$h(\mathrm{t}) = \frac{f(t)}{S(t)}$$

Derivation (Bayes' rule):

$$h(t)dt = P(t \le T < t + dt / T \ge t) = \frac{P(t \le T < t + dt \ \& \ T \ge t)}{P(T \ge t)} = \frac{P(t \le T < t + dt)}{P(T \ge t)} = \frac{f(t)dt}{S(t)}$$

ACTIVISION STUDIOS

Relating these functions

(a little calculus just for fun…):

Hazard from density and survival : $h(t) = \dfrac{f(t)}{S(t)}$

Survival from density : $S(t) = \int_t^{\infty} f(u)\,du$

Density from survival : $f(t) = -\dfrac{dS(t)}{dt}$

Density from hazard : $f(t) = h(t)\,e^{\left(-\int_0^t h(u)\,du\right)}$

Survival from hazard : $S(t) = e^{\left(-\int_0^t h(u)\,du\right)}$

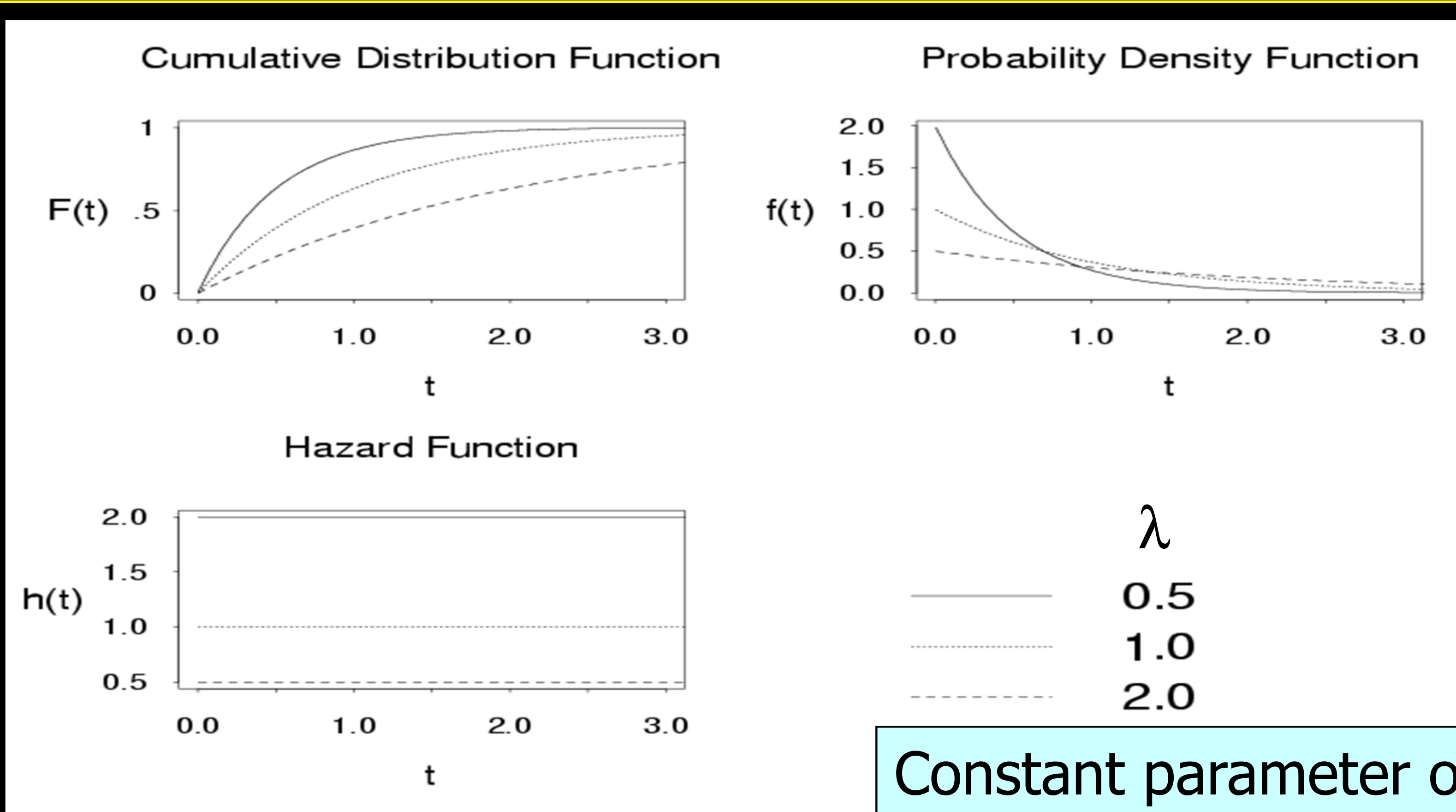
Hazard from survival : $h(t) = -\dfrac{d}{dt}\ln S(t)$

ACTIVISION STUDIOS

Examples: common functions to describe survival

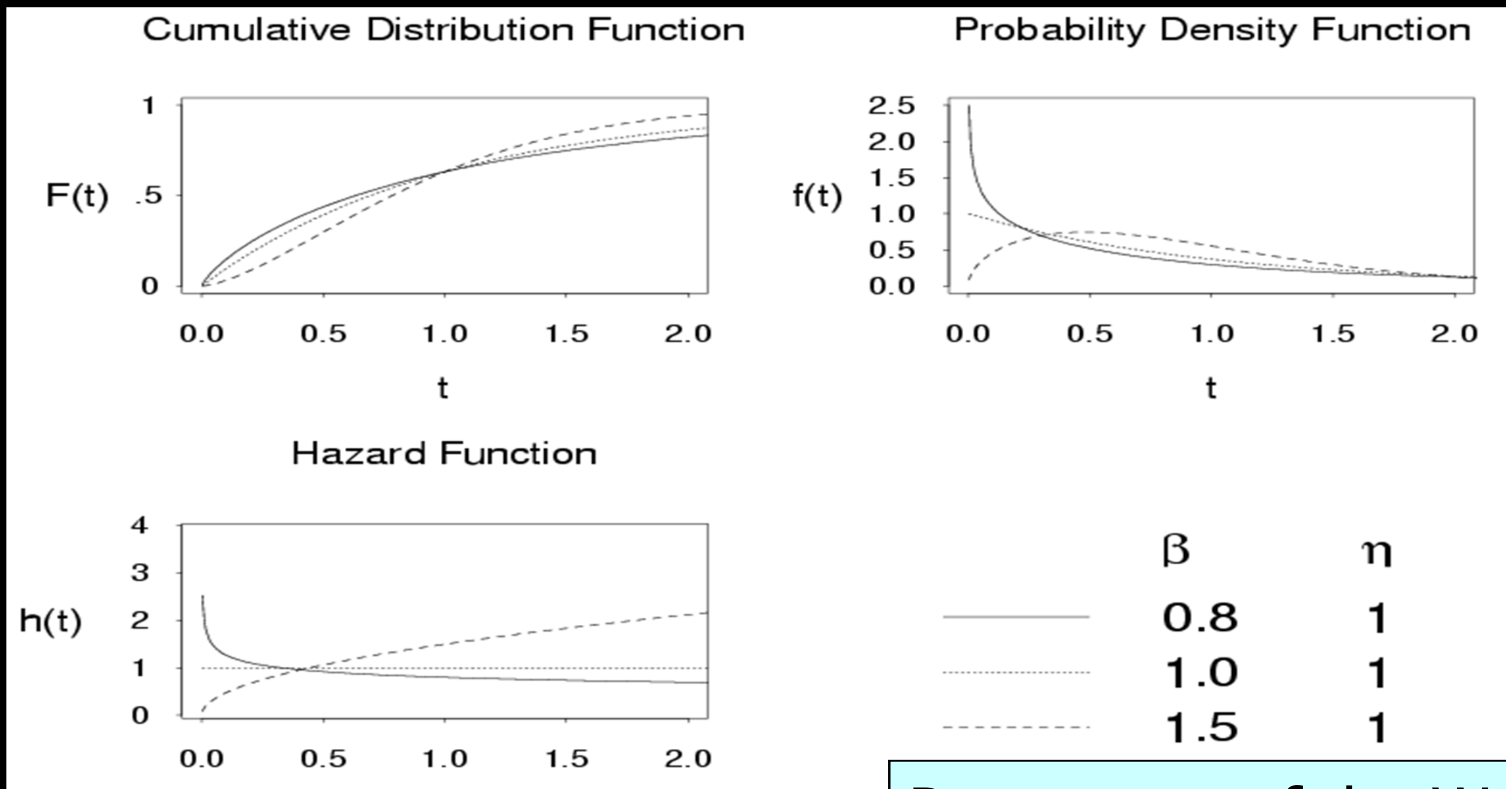
- Exponential (hazard is constant over time, simplest!)
- Weibull (hazard function is increasing or decreasing over time)

Functions for exponential distributions:



Constant parameter of the exponential distribution

Functions for Weibull distributions:



Parameters of the Weibull distribution

Parametric regression techniques

- Model the underlying hazard/survival function.
- Assume that the dependent variable (time-to-event) takes on some known distribution, such as Weibull, exponential, or lognormal.
- Estimates *parameters* of these distributions (e.g., baseline hazard function).
- Estimates covariate-adjusted hazard ratios.
  - A hazard ratio is a ratio of hazard rates
- *Or, you estimate the covariates of the survival function, which we are going to do in this case.*

Many times we care more about comparing groups than about estimating absolute survival.

The model: parametric regression

Components:

- A baseline hazard function (which may change over time).
- A linear function of a set of k fixed covariates.

Exponential model assumes fixed baseline hazard that we can estimate.

$$\log h_i(t) = \mu + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Weibull model models the baseline hazard as a function of time. Two parameters (shape and scale) must be estimated to describe the underlying hazard function over time.

$$\log h_i(t) = \mu + \alpha \log t + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

ACTIVISION ▸ STUDIOS

Ad Churn Data

Start	End	Churned	DaysPlayed	Baidu	Tencent	Other	Age	FirstSessionLength
5	7	1	2	0	0	1	30	69
1	9	1	8	0	1	0	40	104
1	2	0	1	1	0	0	60	105
4	10	1	6	0	1	0	40	133
3	12	1	9	0	1	0	50	118
1	7	0	6	1	0	0	60	110
4	10	0	6	0	1	0	40	171
5	11	0	6	0	1	0	60	126
0	3	0	3	0	0	1	50	133
0	3	1	3	0	1	0	30	84
3	13	1	10	0	1	0	10	131

Exponential Model

```
survreg(formula = Surv(DaysPlayed, Churned) ~
Baidu+Tencent+FirstSessionLength+Age, dist = "exponential")
                      Value Std. Error    z         p
(Intercept)         1.18025   0.017063 69.2  0.00e+00
Baidu               0.40787   0.010726 38.0  0.00e+00
Tencent             0.11250   0.008739 12.9  6.32e-38
FirstSessionLength 0.00272   0.000105 25.9 1.44e-147
Age                 0.01661   0.000351 47.3  0.00e+00

Scale fixed at 1

Exponential distribution
Loglik(model)= -212158.6   Loglik(intercept only)= -214884.3
        Chisq= 5451.41 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 4
n= 104856
```

ACTIVISION ▸ STUDIOS

Weibull Model

```
survreg(formula = Surv(DaysPlayed, Churned) ~
Baidu+Tencent+FirstSessionLength+Age, dist = "weibull")
                       Value Std. Error       z        p
(Intercept)         2.098805   4.50e-03  466.85 0.00e+00
Baidu               0.012933   2.48e-03    5.21 1.84e-07
Tencent             0.003131   2.03e-03    1.54 1.24e-01
FirstSessionLength  0.000143   2.54e-05    5.62 1.95e-08
Age                 0.000690   9.01e-05    7.65 1.96e-14
Log(scale)         -1.456626   3.02e-03 -482.26 0.00e+00

Scale= 0.233

Weibull distribution
Loglik(model)= -146039.7   Loglik(intercept only)= -146101
        Chisq= 122.62 on 4 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 15
n= 104856
```

ACTIVISION ▸ STUDIOS

Likelihood Ratios

```
Model 1: Surv(DaysPlayed, Churned) ~ Amazon + Google + FirstSessionLength +
    PurchasePrice
Model 2: Surv(DaysPlayed, Churned) ~ Amazon + Google + FirstSessionLength +
    PurchasePrice
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -212159
2   6 -146040  1 132238  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ACTIVISION STUDIOS

Weibull New Ad

```
survreg(formula = Surv(DaysPlayed, Churned) ~ NewAd +
Baidu+Tencent+FirstSessionLength+Age, dist = "weibull")

                        Value Std. Error       z        p
(Intercept)          2.098137   5.33e-03  393.98 0.00e+00
NewAd               -0.095127   2.12e-03  -44.81 0.00e+00
Baidu                0.012497   2.85e-03    4.39 1.12e-05
Tencent             -0.002237   2.33e-03   -0.96 3.37e-01
FirstSessionLength   0.000121   2.94e-05    4.11 3.91e-05
Age                  0.000502   1.03e-04    4.88 1.08e-06
Log(scale)          -1.233473   2.82e-03 -437.16 0.00e+00

Scale= 0.291

Weibull distribution
Loglik(model)= -180633.5   Loglik(intercept only)= -181713.9
        Chisq= 2160.92 on 5 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 11
n= 104856
```

Results from Scenario 1

- The hazard rate is not constant and the Weibull model fits the data better.
- The New Ad attracted players who averaged 9.1% shorter retention than players acquired through other ads.
- To get ROI, you would compare the acquisition costs for the new ad campaign to the lower lifetime value numbers for players due to the higher churn rates.

Scenario 2 – Patching an RPG

ACTIVISION STUDIOS

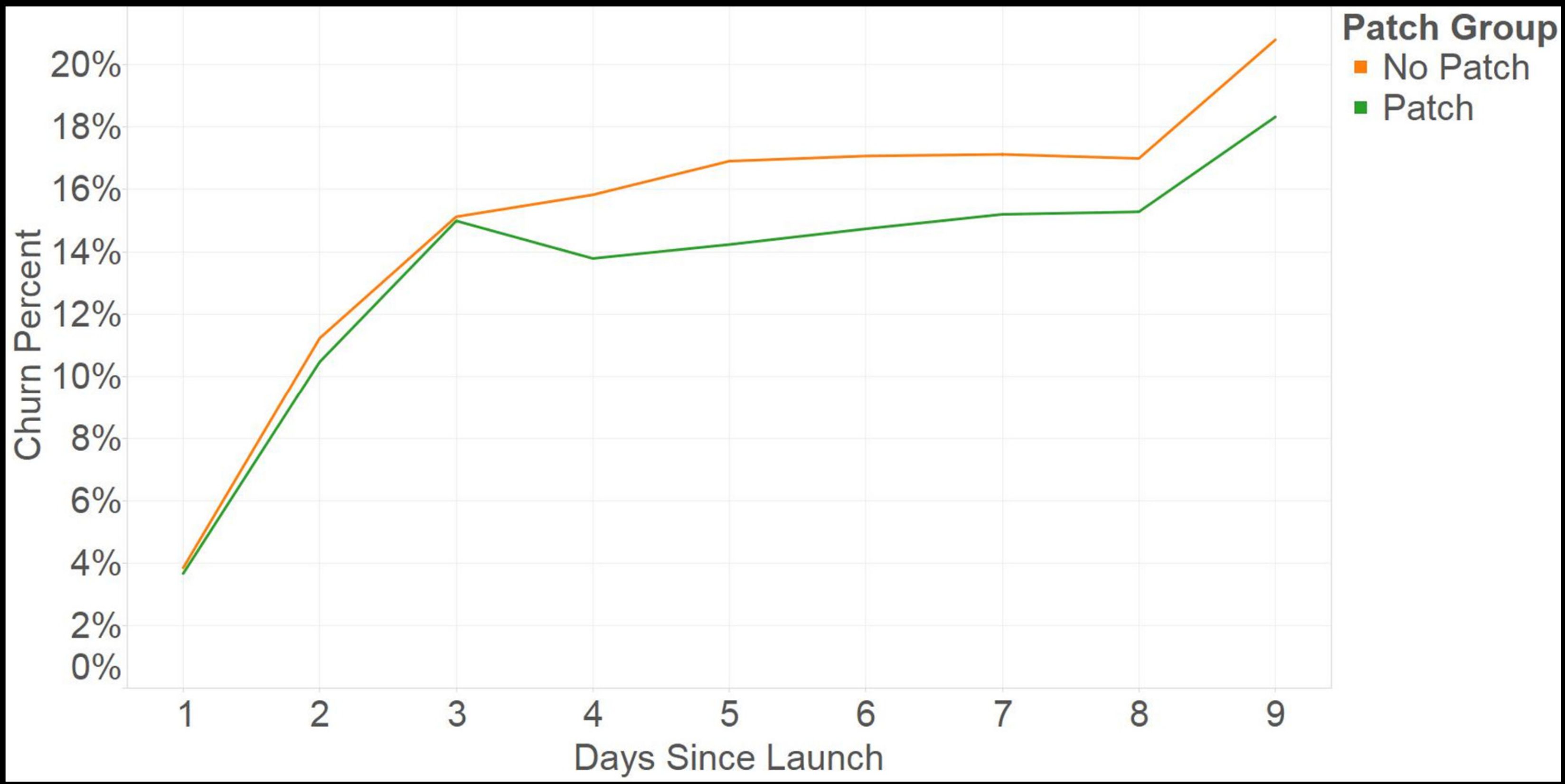
Scenario 2 Background

- A patch is distributed on day 4 to 50% of your player population.
- The patch has a UI change to make warriors more prominent as a class choice compared to wizards.
- There are overall graphics enhancements.
- The patch had a critical bug fix for all players.
- Data is hypothetical, but the scenario is based on real world experiences.

Patch Churn Data

start	stop	age_of_account	patch	bug	minutes_played	is_wizard	churn
4	5	2	1	0	47.90703801	1	0
9	10	2	1	1	22.2085228	0	0
6	7	6	0	0	54.67300914	1	0
7	8	2	0	0	42.19317015	0	0
8	9	3	0	0	47.61779281	0	0
4	5	3	0	0	35.71537856	1	0
8	9	2	1	0	42.3742449	0	1
1	2	1	0	0	34.90920097	0	0
1	2	1	0	0	38.2912341	1	0

Patch Deployed on Day 4



ACTIVISION ▸ STUDIOS

Cox Proportional-Hazard Model

```
coxph(formula = Surv(start, stop, churn) ~ bug + is_wizard +
    patch + age_of_account + minutes_played + cluster(quit_seed),
    data = patch_data)

  n= 382440, number of events= 59227

                    coef  exp(coef)   se(coef)  robust se       z Pr(>|z|)
bug            1.3921964  4.0236780  0.0112569  0.0110634 125.838   <2e-16 ***
is_wizard      0.0830510  1.0865972  0.0086099  0.0092643   8.965   <2e-16 ***
patch         -0.0828029  0.9205325  0.0089436  0.0094945  -8.721   <2e-16 ***
age_of_account 0.1582036  1.1714047  0.0023148  0.0024072  65.720   <2e-16 ***
minutes_played -0.0196435 0.9805482  0.0005978  0.0005983 -32.832   <2e-16 ***
```

ACTIVISION ▸ STUDIOS

Takeaways?

- Patch improves retention by 8%?
- But wizards are far more likely to churn, with or without the patch?

Improved - Detecting Specific Patch Problems

```
coxph(formula = Surv(start, stop, churn) ~ bug + is_wizard +
    patch + age_of_account + minutes_played + wizard_patch +
    cluster(quit_seed), data = patch_data)

  n= 382440, number of events= 59227

                     coef  exp(coef)   se(coef)  robust se       z Pr(>|z|)
bug             1.3777800  3.9660871  0.0113080  0.0111103 124.009   <2e-16 ***
is_wizard      -0.0054967  0.9945184  0.0109240  0.0115514  -0.476    0.634
patch          -0.1666453  0.8464998  0.0109573  0.0112832 -14.769   <2e-16 ***
age_of_account  0.1560963  1.1689388  0.0023195  0.0024141  64.659   <2e-16 ***
minutes_played -0.0196420  0.9805496  0.0005978  0.0005983 -32.830   <2e-16 ***
wizard_patch    0.2353473  1.2653481  0.0176368  0.0188478  12.487   <2e-16 ***
```

Results from Scenario 2

- A bug fix lowered bug incidence by 5% for all players.
- A new bug was introduced that raised bug incidence for wizard players by 10%.
- The patch for warriors decreased risk of churn by 15%.
- The patch for wizards increased risk of churn by 7%.

Conclusion

- Survival model is very interpretable.
- Determining causal factors is possible with proper testing.
- Correlation factors can still be useful a world of incomplete information.

ACTIVISION STUDIOS

Questions?

- Contact info: alan.burke@activision.com
- Blog: http://activisiongamescience.github.io/
- The code used in this presentation will be available via the blog after 3/23/2016.