

# BioSoft: Biyoistatistik/Biyoinformatik yazılımları için bulut tabanlı bir platform

Dinçer Göksülük, Phd.

Biyoistatistik Anabilim Dalı  
Erciyes Üniversitesi

16 Nisan 2021

# Sunum Planı

- 1 Giriş
- 2 BioSoft Projesi
- 3 Uygulamalar/Yazılımlar
  - Web-tabanlı uygulamalar
  - R/BIOCONDUCTOR Paketleri
- 4 TURCOSA Yazılımı (Ticari Boyut)
- 5 Değerlendirme

# Güncel bir paradoks

İstatistikçi? Biyoistatistikçi? Veri Bilimci?

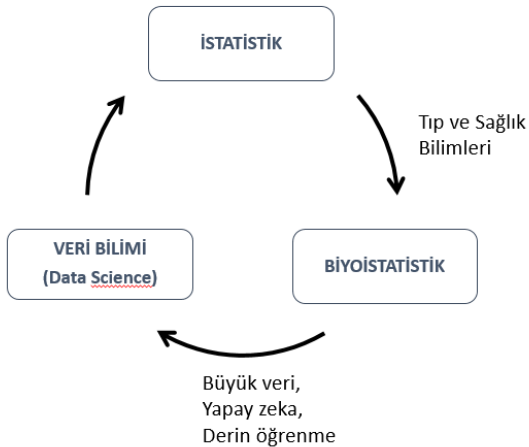


Figure: İstatistik, biyoistatistik ve veri bilimi döngüsü

# Güncel bir paradoks

İstatistikçi? Biyoistatistikçi? Veri Bilimci?

Veri analiz uzmanı (Data Analyst)

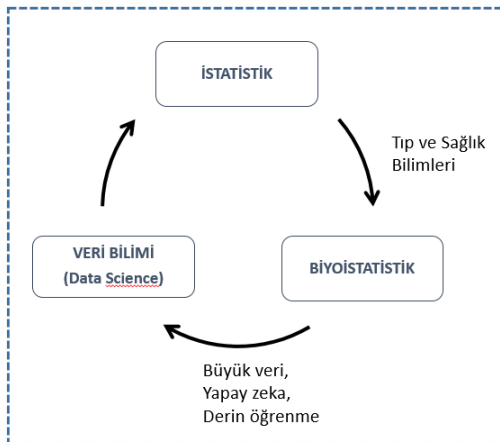
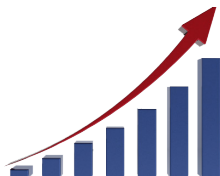


Figure: İstatistik, biyoistatistik ve veri bilimi döngüsü

# Güncel bir paradoks

İstatistikçi? Biyoistatistikçi? Veri Bilimci?



Popülerlik

- Biyoistatistik, istatistikten daha popüler
- Veri bilimi, biyoistatistikten daha popüler

**Popülerlik, bilimsel önemliliğin bir göstergesi değildir<sup>1</sup> !!!**

<sup>1</sup>Kardashian index: <https://www.doi.org/10.1186/s13059-014-0424-0>

# İstatistik/Biyoistatistik

*"The best thing about being a **statistician** is that you get to play in everyone's **backyard**."*

John Tukey (1915 – 2000)

Bizler, Biyoistatistik biliminin gelişimi ve tanınırlığına katkı verebilmek adına neler yaptık?

- Bilimsel yayınlar
- Projeler
- Yazılımlar (R programlama dili) – CRAN / BIOCONDUCTOR
- Web tabanlı uygulamalar

**BioSoft Projesi (Ücretsiz)**

**TURCOSA Yazılımı (Ticari)<sup>2</sup>**

---

<sup>2</sup>T.C. Sanayi Bakanlığı, TÜBİTAK, KOSGEB proje destekleri

# Motivasyon

## Araştırmacıların talepleri/eğilimleri (Motivasyon – Hedef)

- Güncel literatür
- **Kolay kullanım** (kurulum gerektirmeyen, işletim sisteminden bağımsız, vb.)
- Etkin sonuçlar (interaktif grafikler, raporlamalar, vb.)
- Ücretsiz
- **Açık kaynak kodlu**



# Motivasyon

## IBM SPSS Statistics

- Kullanıcı arayüzü (GUI) – dropdown menü
- Windows, Mac OS
- Ücretli
- Kodlar korumalı
- Dokümantasyon (kılavuz, vb.)??

## R Programlama Dili

- Kod penceresi – Programlama becerisi
- Windows, Mac OS, Linux
- Ücretsiz
- Açık kaynak kodlu
- Esnek, güçlü analiz modülleri ile ölçeklendirilebilir
- Kapsamlı dokümantasyon, kullanım kılavuzları

# BioSoft projesi nedir?

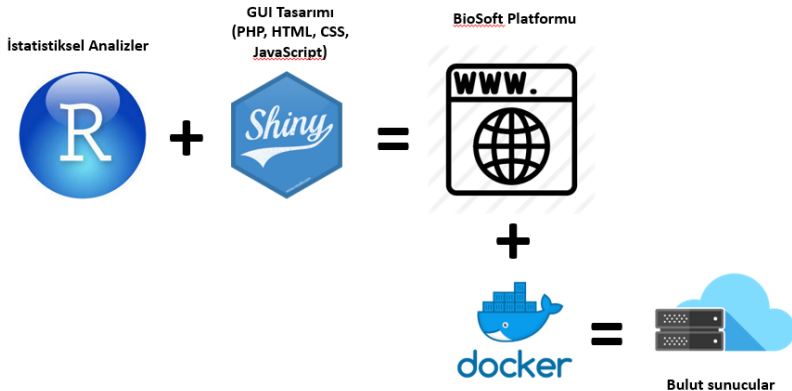
**BioSoft** (**Bi**ostatistics/**Bio**informatics **Soft**wares) platformu, 2014 yılı içerisinde hayata geçirilmiş olan

- Web-tabanlı (bulut sunucularda çalışan)
- Kullanımı kolay (GUI)
- Çoklu ortamlarda çalışabilen (tablet, telefon, bilgisayar, vb.)
- R programlama dilini kullanan
- **Açık kaynak kodlu ve ücretsiz**

uygulamalar geliştirmeyi ve paylaşmayı amaçlayan (**bilimsel fayda**) bir projedir.

# BioSoft projesi nedir?

## Sistem mimarisi



# Web sayfası...

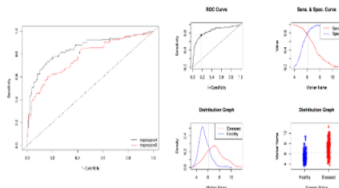
[www.biosoft.erciyes.edu.tr](http://www.biosoft.erciyes.edu.tr)

## Biosoft Project

About Tools News Publications Members

Biosoft Project is dedicated to develop free web applications for various fields of science using the **R** language environment. We aim to develop user-friendly, easy-to-use, up-to-date and comprehensive tools for the science community.

The project is started in 2014 at **Hacettepe University Department of Biostatistics** and it is managed by **Turcosa Analytics**. The main server is now moved to **Erciyes University**, and will be available through cloud servers located at Erciyes University. Biosoft is also available through a mirror server. If the main server at Erciyes University is offline, you may go to mirror [by clicking here](#).



# Mevcut uygulamalar...

## Web-tabanlı uygulamalar

- easyROC: a web-tool for ROC curve analysis
- MVN: a web-tool for assessing multivariate normality
- geneSurv: Survival Analysis for Genomics
- MLViS: machine learning-based virtual screening tool
- voomDDA: Discovery of diagnostic biomarkers and classification of RNA-Seq data
- DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen

## R/BIOCONDUCTOR paketleri

- MLSeq: Machine learning interface for RNA-Seq data

# Web Tabanlı Uygulamalar

# easyROC: a web-tool for ROC curve analysis











İki kategorili durum değişkeni olduğunda (binary outcome) uygulanan ROC analizi için geliştirilmiş bir yazılımdır.

- Tanı testinin (biyobelirteç) performansının değerlendirilmesi
- Farklı tanı testlerinin performanslarının karşılaştırılması
- En iyi kesim noktasının belirlenmesi (34 farklı kriter)
- Örneklem büyüklüğü hesaplamaları

Dincer Goksuluk, Selcuk Korkmaz, Gokmen Zararsiz and A. Ergun Karaagaoglu (2016). easyROC: An Interactive Web-tool for ROC Curve Analysis Using R Language Environment. *The R Journal*, 8:2, pages 213-230.

# easyROC: a web-tool for ROC curve analysis

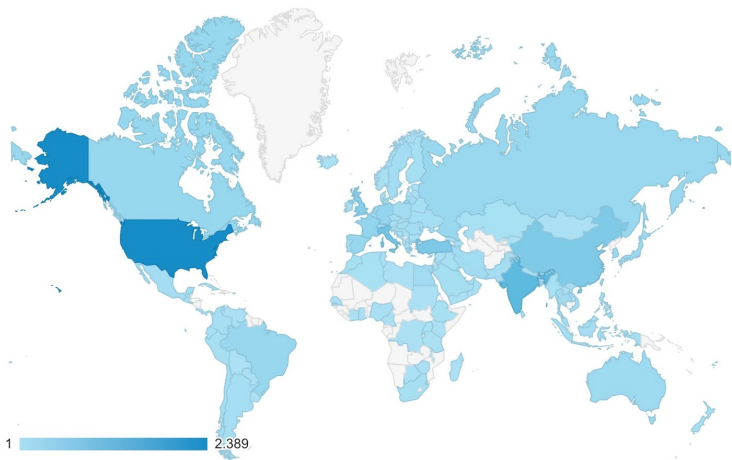
- Yıllık ~ 25,000 kullanım
- 172 atıf (~ 110 Web of Science)

| Ülke ?  | Edinme  |   |   |
|---|---|---|---|
|   | Kullanıcılar ? ↓                                  | Yeni Kullanıcılar ?                               | Oturum ?  |
|   | <b>11.609</b><br>Toplam Yüzdesi: %100,00 (11.609) | <b>11.521</b><br>Toplam Yüzdesi: %100,12 (11.507) | <b>17.600</b><br>Toplam Yüzdesi: %100,00 (17.600) |
| 1.  United States  | <b>2.389</b> (%20,53)                             | <b>2.378</b> (%20,64)                             | <b>3.611</b> (%20,52)                             |
| 2.  India          | <b>1.115</b> (%9,58)                              | <b>1.112</b> (%9,65)                              | <b>1.516</b> (%8,61)                              |
| 3.  Italy          | <b>751</b> (%6,45)                                | <b>747</b> (%6,48)                                | <b>1.186</b> (%6,74)                              |
| 4.  Turkey         | <b>641</b> (%5,51)                                | <b>639</b> (%5,55)                                | <b>924</b> (%5,25)                                |
| 5.  China          | <b>626</b> (%5,38)                                | <b>607</b> (%5,27)                                | <b>922</b> (%5,24)                                |
| 6.  United Kingdom | <b>497</b> (%4,27)                                | <b>491</b> (%4,26)                                | <b>643</b> (%3,65)                                |
| 7.  France         | <b>327</b> (%2,81)                                | <b>320</b> (%2,78)                                | <b>618</b> (%3,51)                                |
| 8.  Germany        | <b>311</b> (%2,67)                                | <b>302</b> (%2,62)                                | <b>467</b> (%2,65)                                |
| 9.  Canada         | <b>272</b> (%2,34)                                | <b>270</b> (%2,34)                                | <b>420</b> (%2,39)                                |
| 10.  Spain         | <b>268</b> (%2,30)                                | <b>265</b> (%2,30)                                | <b>448</b> (%2,55)                                |



# easyROC: a web-tool for ROC curve analysis

Kullanım İstatistikleri



# MVN: a web-tool for assessing multivariate normality










Çok değişkenli istatistiksel analizlerin temel varsayımlarından birisi olan “çok değişkenli normallik” varsayımını test etmek amacıyla geliştirilmiş bir yazılımdır.

- Çok değişkenli normallik kontrolü
  - Testler (Mardia, Royston, Henze-Zirkler)
  - Grafiksel yaklaşımlar
- Çok değişkenli aykırı değerlerin tespiti ve aykırı değerlerden arındırılmış veri setinin elde edilmesi
- Tek değişkenli normallik kontrolü
- Tanımlayıcı istatistikler

Korkmaz S, Goksuluk D, Zararsiz G. (2014). MVN: An R Package for Assessing Multivariate Normality. The R Journal, 6(2):151-162.

# MVN: a web-tool for assessing multivariate normality

- Yıllık  $\sim 4,000$  kullanım
- **696** atıf ( $\sim 320$  Web of Science)

| Ülke ?  | Edinme  |   |   |
|---|---|---|---|
|   | Kullanıcılar ? ↓                                | Yeni Kullanıcılar ?                             | Oturum ?  |
|   | <b>1.796</b><br>Toplam Yüzdesi: %100,00 (1.796) | <b>1.776</b><br>Toplam Yüzdesi: %100,23 (1.772) | <b>2.619</b><br>Toplam Yüzdesi: %100,00 (2.619) |
| 1.  Turkey         | <b>378</b> (%20,94)                             | <b>374</b> (%21,06)                             | <b>549</b> (%20,96)                             |
| 2.  United States  | <b>299</b> (%16,57)                             | <b>299</b> (%16,84)                             | <b>348</b> (%13,29)                             |
| 3.  Germany        | <b>98</b> (%5,43)                               | <b>99</b> (%5,57)                               | <b>144</b> (%5,50)                              |
| 4.  Sweden         | <b>96</b> (%5,32)                               | <b>96</b> (%5,41)                               | <b>105</b> (%4,01)                              |
| 5.  Spain          | <b>57</b> (%3,16)                               | <b>57</b> (%3,21)                               | <b>82</b> (%3,13)                               |
| 6.  United Kingdom | <b>54</b> (%2,99)                               | <b>54</b> (%3,04)                               | <b>64</b> (%2,44)                               |
| 7.  India          | <b>54</b> (%2,99)                               | <b>52</b> (%2,93)                               | <b>145</b> (%5,54)                              |
| 8.  Canada         | <b>49</b> (%2,71)                               | <b>48</b> (%2,70)                               | <b>59</b> (%2,25)                               |
| 9.  Australia      | <b>47</b> (%2,60)                               | <b>47</b> (%2,65)                               | <b>69</b> (%2,63)                               |
| 10.  Brazil        | <b>43</b> (%2,38)                               | <b>43</b> (%2,42)                               | <b>59</b> (%2,25)                               |

# geneSurv: Survival Analysis for Genomics

Genetik alanında sağkalım analizleri:

- Sağkalım modellemeleri
  - Kaplan-Meier, Cox regression, Penalized Cox regression, Random Survival Forests
- Biyobelirteçler için en iyi kesin değeri
- Genetik bilgilerin (gen ekspresyonu) modelleme aşamasına dahil edilmesi

Korkmaz, S., Goksuluk, D., Zararsiz, G., & Karahan, S. (2017). geneSurv: an interactive web-based tool for survival analysis in genomics research. Computers in biology and medicine, 89, 487-496.

# MLViS: a web tool for machine learning-based virtual screening in early-phase of drug discovery and development

İlaç olabilme özelliği olan moleküllerin tespitine yönelik algoritmalar içeren bir yazılım:

- Makine öğrenmesi algoritmaları ile sınıflama (drug vs. non-drug)
- PubChem veri tabanına erişim
  - Moleküllerin kimyasal yapılarının değerlendirilmesi
  - Molekül yapılarına yönelik grafiklerin çizilmesi
- Kümeleme analizleri

Korkmaz S, Zararsiz G, Goksuluk D (2015) MLViS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. PLoS ONE 10(4): e0124600.

# Diğer web-tabanlı uygulamalar

- DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen

Zararsiz G, Goksuluk D, Korkmaz S, Ozturk A, Akyildiz HY (2016) "Statistical learning approaches in diagnosing patients with nontraumatic acute abdomen" Turkish Journal of Electrical Engineer and Computer Science 24, 3685-3697

- voomDDA: Discovery of diagnostic biomarkers and classification of RNA-Seq data

Zararsiz, G., Goksuluk, D., Klaus, B., Korkmaz, S., Eldem, V., Karabulut, E., & Ozturk, A. (2017). voomDDA: discovery of diagnostic biomarkers and classification of RNA-seq data. PeerJ, 5, e3890.

- BioVar: an online biological variation analysis tool

Korkmaz, Selçuk, Zararsız, Gökmen, Göksülük, Dinçer, Senes, Mehmet, Sönmez, Cem and Yucel, Dogan. "BioVar: an online biological variation analysis tool" Turkish Journal of Biochemistry, vol. 45, no. 5, 2020, pp. 479-489.

# R/BIOCONDUCTOR Paketleri

# MLSeq: Machine Learning Interface for RNA-Seq Data

**MLSeq** yazılımı R programlama dilinde kodlanmış ve R/BIOCONDUCTOR ağında paylaşılmış bir kütüphanedir.

- Biyoinformatik analizler
- Gen ekspresyon verilerinin sınıflaması
  - Mikrodizilim (microarrays)
  - RNA dizileme (RNA-Seq)
- 80+ farklı sınıflama algoritması (Machine learning)
  - Mevcut makine öğrenme algoritmaları
  - Yeni önerilen algoritmalar (voomDDA, sparse NBLDA, vb.)



# MLSeq: Machine Learning Interface for RNA-Seq Data

- BIOCONDUCTOR ağında RNA-dizileme verilerinin sınıflaması amacıyla yayınlanan ilk ve en kapsamlı pakettir.
- BIOCONDUCTOR ağına Türkiye'den gönderilen ve türk araştırmacılar tarafından geliştirilen ilk pakettir.
- Uluslararası platformlarda müfredata alınmıştır.
  - Kongre, sempozyum
  - Workshop, Kurs, Eğitimler
- Bilimsel etkileri:
  - İki doktora ve bir yüksek lisans tezi
  - Avrupa Moleküler Biyoloji Laboratuvarı (EMBL) ekibi çalışmalarına etki – DESeq2 <sup>3</sup> paketinde kullanılan yöntemlerin iş akışını etkilemiştir.

---

<sup>3</sup>BIOCONDUCTOR ağında yer alan 1974 paket içerisinde en fazla indirilen 27. pakettir.

# MLSeq: Machine Learning Interface for RNA-Seq Data

## MLSeq

platforms all rank 335 / 1974 support 0 / 0 in Bioc 7 years  
build ok updated before release dependencies 124

DOI: [10.18129/B9.bioc.MLSeq](https://doi.org/10.18129/B9.bioc.MLSeq)



## Machine Learning Interface for RNA-Seq Data

Bioconductor version: Release (3.12)

This package applies several machine learning methods, including SVM, bagSVM, Random Forest and CART to RNA-Seq data.

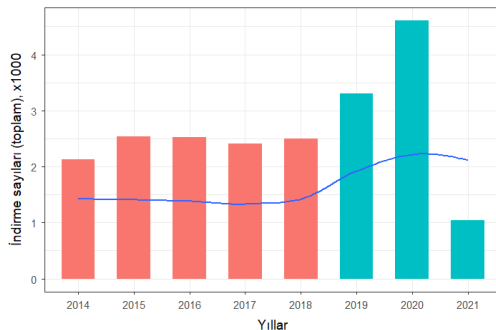
Author: Gokmen Zararsiz [aut, cre], Dincer Goksuluk [aut], Selcuk Korkmaz [aut], Vahap Eldem [aut], Izzet Parug Duru [ctb], Ahmet Ozturk [aut], Ahmet Ergun Karaagaoglu [aut, ths]

Maintainer: Gokmen Zararsiz <gokmenzararsiz at hotmail.com>

Citation (from within R, enter `citation("MLSeq")`):

Goksuluk D, Zararsiz G, Korkmaz S, Eldem V, Zararsiz GE, Ozcetin E, Ozturk A, Karaagaoglu AE (2019). "MLSeq: Machine learning interface for RNA-sequencing data." *Computer Methods and Programs in Biomedicine*, **175**, 223–231. doi: [10.1016/j.cmpmb.2019.04.007](https://doi.org/10.1016/j.cmpmb.2019.04.007), <http://www.sciencedirect.com/science/article/pii/S0169260718318728>.

# MLSeq: Machine Learning Interface for RNA-Seq Data



- 21,000 indirme (11,150 farklı kullanıcı)
- 18 atıf (5, Web of Science)

# Diğer R/BIOCONDUCTOR paketleri

- MVN: Multivariate Normality Tests – CRAN
  - **URL:** <https://cran.r-project.org/web/packages/MVN/index.html>
  - **Kılavuz:** <https://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf>
- NBLDA: Negative Binomial Linear Discriminant Analysis
  - **URL:** <https://cran.r-project.org/web/packages/NBLDA/index.html>
  - **Kılavuz:** <https://cran.r-project.org/web/packages/NBLDA/NBLDA.pdf>

# TURCOSA yazılımı

- Bulut tabanlı istatistik analiz yazılımı
  - Türkiye'nin ik milli istatistik yazılımı
  - Dünyada bir çok özelliği ile ilkler arasında olan istatistik analiz sistemi
  - T.C. Sanayi Bakanlığı, TÜBİTAK, KOSGEB proje destekleri

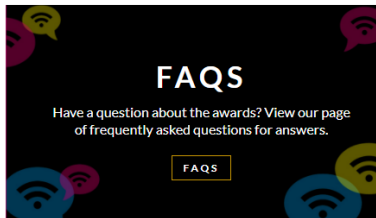
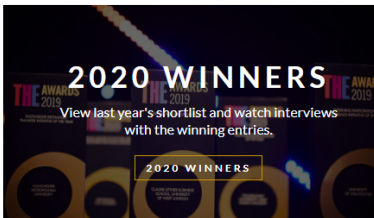
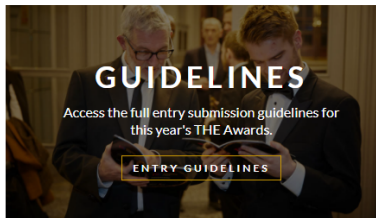
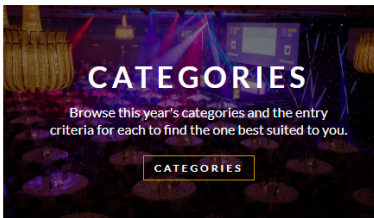


[www.turcosa.com.tr](http://www.turcosa.com.tr)

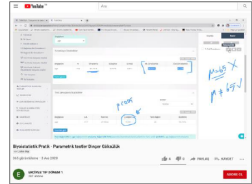
# TURCOSA yazılımı

## Times Higher Education (THE) Awards 2021

### Technological or Digital Innovation of the Year



# Eğitimde TURCOSA (Pandemi Dönemi)



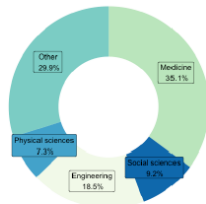
# Sayılarla TURCOSA



Over 5,000 students/researchers are actively using TURCOSA in their analysis. However, it is available for more than 100K users within four universities.



Approx. 1000 active projects created including more than 200K analysis performed (i.e. 20 analysis per project on the average).



TURCOSA is mostly preferred by physicians because of ease-of-use and comprehensive modules specifically developed for medical sciences (e.g. survival analysis, method comparison, ROC analysis, etc.) with a high user-feedback.

Overall rating

★★★★★ 4.7 / 5

## What is in TURCOSA?



60,000+ lines of codes



65 analysis modules within 10 panels



System validation with over thousands of tests



... and more





# Genel değerlendirme – BioSoft/TURCOSA

- **BioSoft**, yılda  $\sim 50,000$  defa ziyaret edilen bir platformdur.
- Bilimsel çalışmaların yazılımlar/uygulamalar ile sunulması yöntemlerin kullanılabilirliği arttırmaktadır.
- Web tabanlı uygulamalar çalışmalara olan ilgiyi ve atıf potansiyelini arttırmaktadır.
- Önerilen yöntemlerin yazılım/uygulama olarak ürüne dönüştürülmesi çalışmanın daha iyi dergilerde yayınlanmasına katkı sağlayabilir.
- Geliştirilen uygulamalar/yazılımlar ticari ürünlere dönüştürülebilir.

# Genel değerlendirme – Neden R?

- Giderek daha fazla popüler hale gelen bir yazılım
- Güçlü bir geliştirici ağı
- Güçlü analiz araçları
  - Kullanıcı
  - Geliştirici
- Bir çok yazılım ile entegrasyon
  - Web uygulamaları
  - Raporlama (TeX, Markdown, vb.) – **Reproducible reports**
  - Kitap/Makale yazımı
  - Veri tabanı (ORACLE, MySQL, vb.)
- **BioSoft + TURCOSA**
- ...

## Çalışma Ekibimiz (Alfabetik Sıra)

|                                  |  |
|----------------------------------|--|
| Ahmet ÖZTÜRK, Prof. Dr.          | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Erciyes Üniversitesi            |
| Dinçer GÖKSÜLÜK, Dr. Öğr. Üye.   | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Erciyes Üniversitesi            |
| Gökmen ZARARSIZ, Doç. Dr.        | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Erciyes Üniversitesi            |
| Gözde E. ZARARSIZ, Dr. Öğr. Üye. | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Erciyes Üniversitesi            |
| Merve BAŞOL, Araş. Gör.          | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Abant İzzet Baysal Üniversitesi |
| Selçuk KORKMAZ, Doç. Dr.         | Biyoistatistik Anabilim Dalı, Tıp Fakültesi<br>Trakya Üniversitesi             |