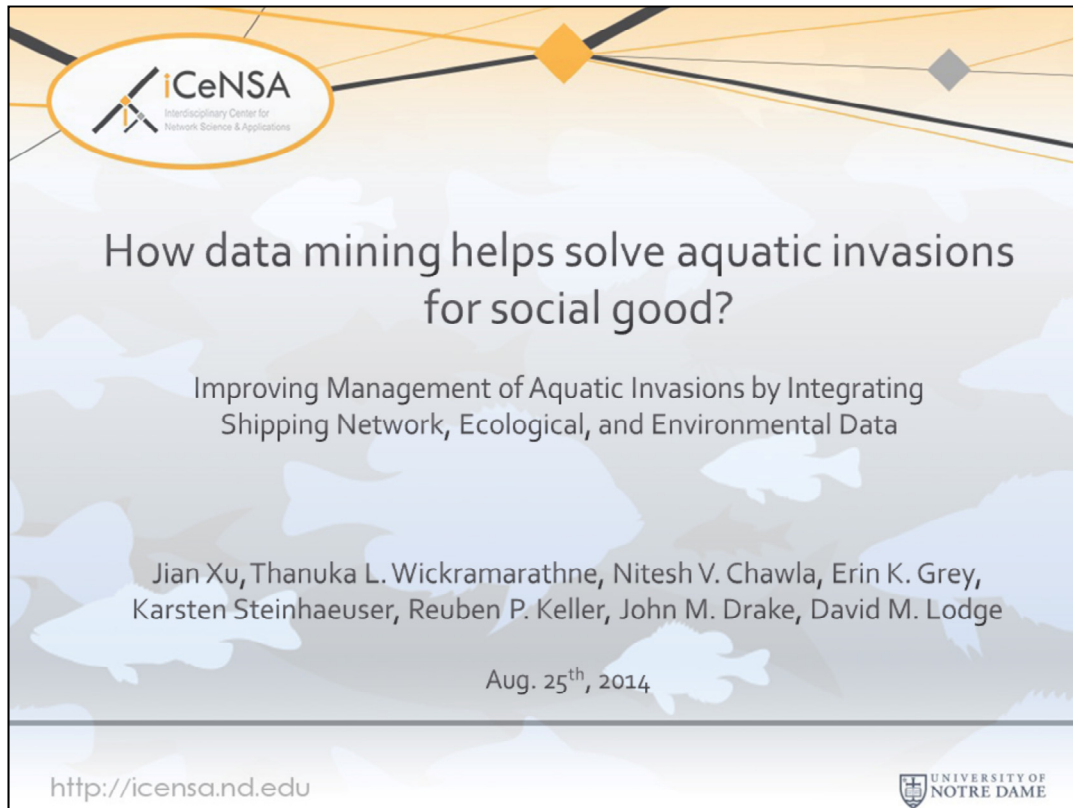




Good afternoon, my **name** is Jian, from University of Notre Dame.

My presentation is: **how** data mining helps solve aquatic invasions for social good.

Perhaps the topic about **aquatic invasions** is not frequently seen in KDD,  
So first of all I will show a **short video** about how the Asian Carp invasion is happening in the United States.



## The **problem** of invasive species



Video courtesy of David Lodge and Milwaukee Journal Sentinel

<http://icensa.nd.edu>

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)



Aug 25, 2014

How data mining helps solve aquatic invasions for social good?

2



## The **problem** of invasive species

- Non-native species that **adversely** affect habitats and bioregions
- **Top three** drivers of global environmental change
- Outcompete or prey on native species
- Annual damage and control costs of all invasive species in the U.S.: **\$120 Billion**
  - An invisible tax that **everyone** has to pay!



<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 3

Quite impressive, isn't it?

This is **not** a good phenomena because Asian Carps are **not delicious** (joke).

Invasive species, **namely** non-native species that **adversely** affect habitats and bioregions,

are among the **top three** drivers of global environmental change.

They cause substantial **economic and environmental harm** by outcompeting or preying on native species.

In fact, the annual damage and control costs of all invasive species in the United States is as high as 120 billion dollars,

Which is an invisible tax that **everyone** has to pay!



# The problem of **aquatic invasions**

- Impact of aquatic invasions
  -  Increased diseases
    - *In humans (e.g. cholera)*
    - *In aquaculture species (e.g. fish virus)*
  -  Losses of wild caught fisheries
    - *E.g. comb jelly invasion of the Black sea*

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 4

**Aquatic** invasions has **additional negative impacts**, including the increased diseases in humans and aquaculture species, And the losses of wild caught fisheries.



## Ship-borne invasive species problem

- Global shipping is responsible for **69%** of known aquatic invasions
- Species get translocated by means of:
  -  Ballast water uptake / discharge
  -  Biofouling (species attached on surfaces of ships)
- Annual loss to the Great Lakes regional economy due to ship-borne aquatic invasive species: **\$800 million**

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME  
Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 5

Meanwhile, **global shipping** is responsible for 69% of known aquatic invasions

Species get translocated during **ballast** water uptake and discharge, and by **attaching** themselves on the surfaces of ships.

The annual loss to the Great Lakes regional economy due to ship-borne aquatic invasive species is 800 million dollars, and much more globally.

We **cannot wait** any longer, and **actions** must be in place.



## Demand of management

- Global shipping: responsible for **90%** of global trade
- Optimized for trade considerations but not aquatic invasions
- Expensive and cumbersome regulations are not acceptable
- Need a **thorough understanding** of ship-borne invasion risks to devise policies that are:
  - Feasible
  - Effective
  - Long-term benefits to human welfare

<http://icensa.nd.edu>

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)

 UNIVERSITY OF NOTRE DAME

Aug 25, 2014

How data mining helps solve aquatic invasions for social good?

6

However, global shipping is responsible for **90%** of global trade, and it is optimized for **trade** considerations, not aquatic invasions.

**Expensive** and **cumbersome** are not acceptable.

So, before actions can be taken, a **thorough understanding** of ship-borne invasion risks is necessary to devise policies that are feasible, effective, and can bring long-term benefits to human welfares.

**iCeNSA**  
Interdisciplinary Center for  
Network Science & Applications

## Challenges of risk assessment

Ship traffic

Ship-borne invasion

Environment

Ecoregion

Ballast discharge

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 7

Unfortunately, invasion risk assessment is **not trivial**.

Ship-borne invasions are a result of a **complex interplay** of ship traffic, ballast uptake and discharge dynamics, various environmental and biological variables, and so on.



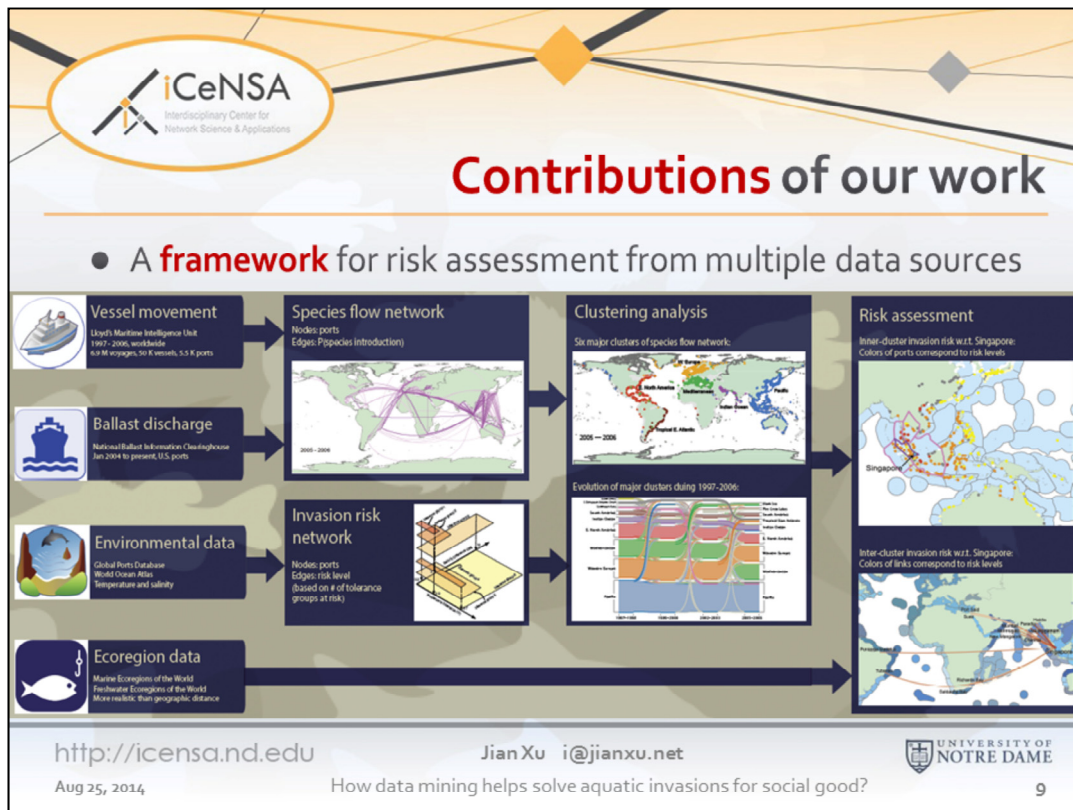
## Challenges of risk assessment

- Poorly quantified governing relationships
  - Inadequate characterizations of invasion processes
  - Lack of crucial data
  - Large uncertainties associated with available data
- The probabilistic approach:
  - Multiple simplifying assumptions
  - Unable to model invasion mechanisms
    - *e.g. "stepping-stone" process*

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 8

The few studies that have attempted to quantify invasion risks via **probabilistic approaches** have relied on multiple **simplifying** assumptions, which hampered the usefulness of such estimate.



Therefore, we present here a framework for addressing the ship-borne species invasion problem from multiple data sources.



## Contributions of our work

- **Crucial knowledge** on how patterns in global shipping network affect aquatic invasions

**Species flow network**  
Nodes: ports  
Edges: Pespacies introduction



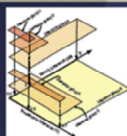
2005-2006

**Clustering analysis**  
Six major clusters of species flow network:

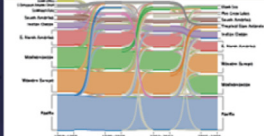


2005-2006

**Invasion risk network**  
Nodes: ports  
Edges: risk level (based on # of tolerance groups at risk)



**Evolution of major clusters during 1997-2006:**



<http://icensa.nd.edu>

Jian Xu i@jianxu.net

 UNIVERSITY OF NOTRE DAME

Aug 25, 2014
How data mining helps solve aquatic invasions for social good?
10

By experimenting on **real** global shipping traffic data for 1997-2006, we have discovered **crucial** knowledge on how patterns in global shipping network affect aquatic invasions,

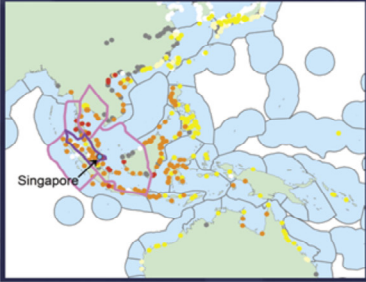


## Contributions of our work

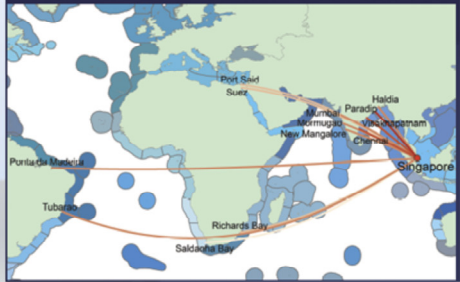
- Knowledge used to devise effective and economical invasive species **management strategies**

**Risk assessment**

Inner-cluster invasion risk w.r.t. Singapore:  
Colors of ports correspond to risk levels



Inter-cluster invasion risk w.r.t. Singapore:  
Colors of links correspond to risk levels



<http://icensa.nd.edu>

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)

UNIVERSITY OF NOTRE DAME

Aug 25, 2014

How data mining helps solve aquatic invasions for social good?

11

and I will illustrate how such knowledge can be used to devise **effective** and **economical** management strategies.



## Contributions of our work

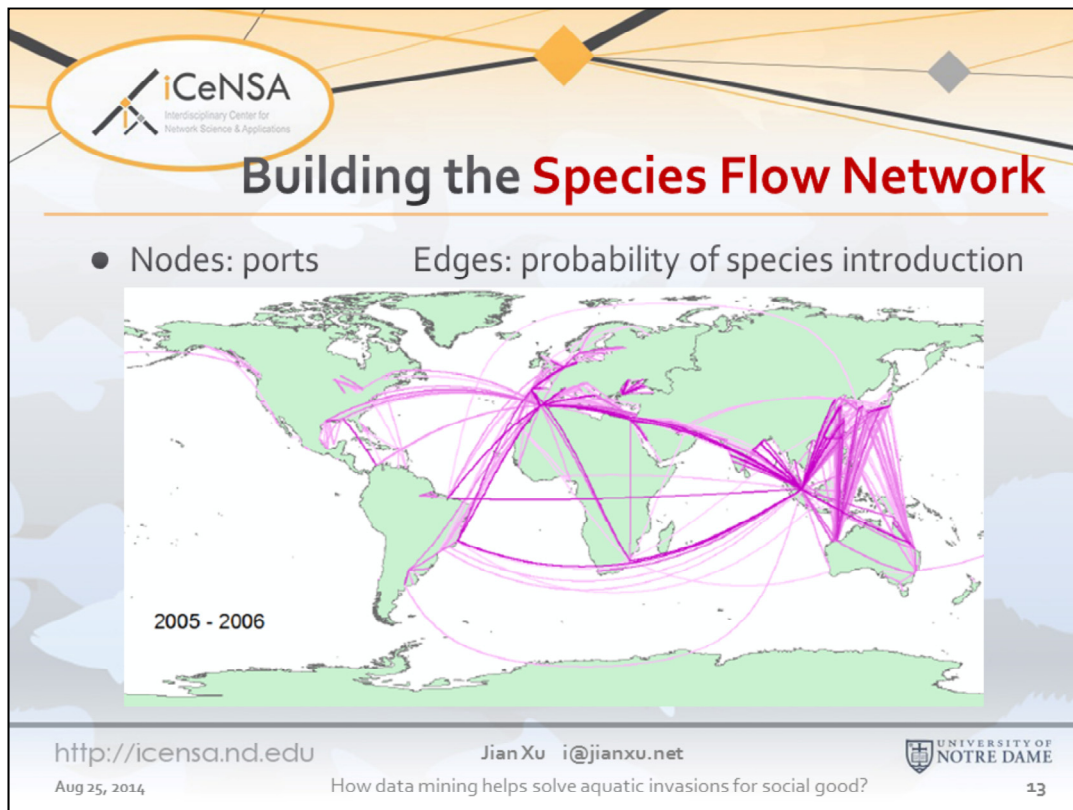
- Data mining for **social good**



<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 12

Finally, our work shows how data mining can be used for solving crucial, yet very complex problems towards **social good**.



To begin with, the basic idea is to find **patterns** of species flow in order to identify **key ports and shipping routes**,

Where **interventions** would be the most effective in **stopping** invasions through the entire shipping network.

The vessel movement data and ballast discharge data can be utilized to **build a network** to represent species flow among ports.



## Building the Species Flow Network

- Probability of species introduction for every vessel movement
 

Probability of vessel  $v$  introducing species from port  $i$  to port  $j$

$$p_{ij}^{(v)} = \underbrace{\rho_{ij}^{(v)}}_{\text{Mgmt efficacy}} \underbrace{(1 - e^{-\lambda D_{ij}^{(v)}})}_{\text{Ballast discharge}} \underbrace{e^{-\mu \Delta t_{ij}^{(v)}}}_{\text{Mortality}}$$
- Aggregate movements between ports as edge weight
 

$$w_{ij} = 1 - \prod_{\substack{r \in DB \\ r=v:n_i \rightarrow n_j}} (1 - p_{ij}^{(v)})$$
- Details in (Seebens et al., Ecology Letters, Apr 2013)

<http://icensa.nd.edu>
Jian Xu [i@jianxu.net](mailto:i@jianxu.net)
UNIVERSITY OF NOTRE DAME

Aug 25, 2014
How data mining helps solve aquatic invasions for social good?
14

For every vessel movement, the **probability** of species introduction is estimated above,

**considering** the impact of ballast discharge, species mortality during the trip, management efficacy, and so on.

Then the **aggregated** probability between every pair of ports is the edge weight of species flow network.



## Properties of species flow network

- Scale-free network for degree  $\geq 139$ 
  - Some ports connect to many other ports (e.g. Singapore)
- Average path length = 3
  - Only **two intermediate stops** for species to be carried from an arbitrary port to another
  - High risk of invasions in the absence of risk management

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 15

### Some **quick observations**.

The network is **scale-free** and there are ports connected to many other ports such as **Singapore**.

As a consequence, species can be carried from **any** port to another in the world within **two intermediate stops** on average.

This indicates a generally high risk of invasions in the **absence** of risk management.



## Clustering analysis

- Why
  - Complex networks: too complex to derive useful inferences
  - Clusters (modules + interactions) are useful for giving a **big picture**
  - Easier identification of **risk corridors**
  - Naturally handles the “**stepping-stone**” process
- How
  - **MapEquation** decomposes network with respect to **flow-dynamics**
    - Rosvall et al., PNAS 2008
  - Can handle weighted, directed network
  - Flow of information resembles flow of species

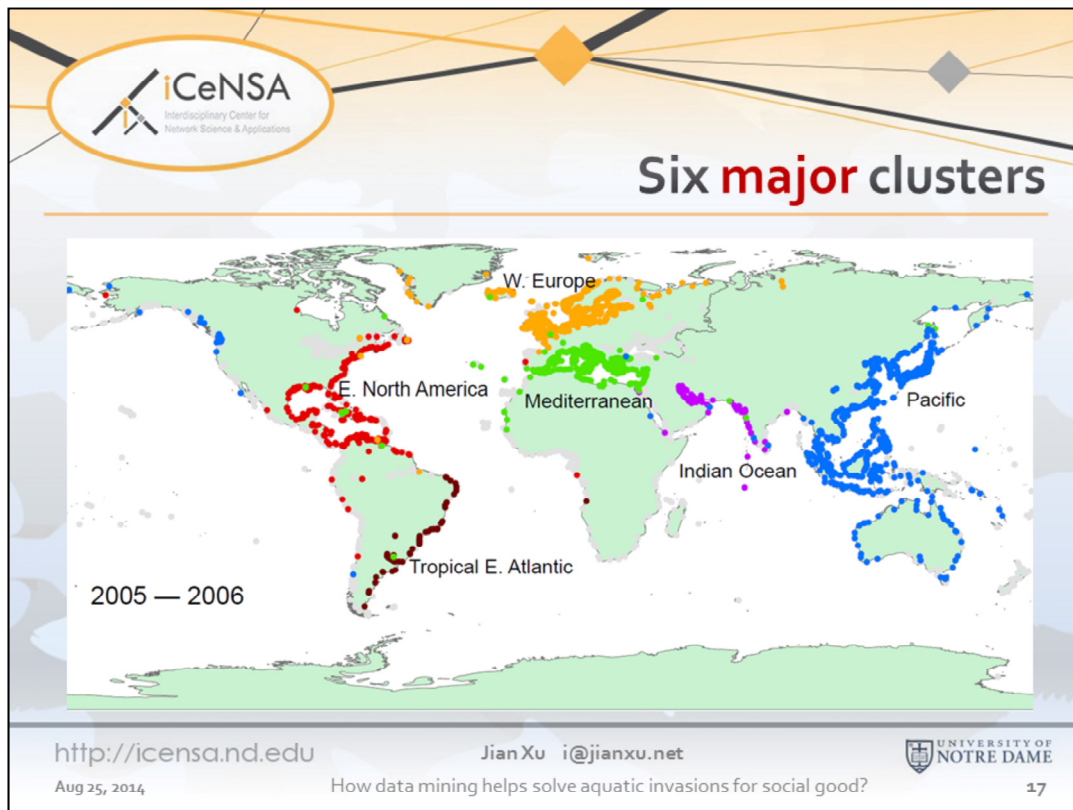
<http://icensa.nd.edu> Jian Xu i@jianxu.net UNIVERSITY OF NOTRE DAME  
Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 16

However, the network itself is still **too complex** to derive useful inferences.

Therefore, a **decomposition** that represents such complex networks via **modules** and their **interactions** can be very useful.

With clusters, it is easier to identify **risk corridors** connecting the main components, and the “**stepping-stone**” process in species propagation can be handled naturally.

Here we use **MapEquation** for clustering, because it is based on **flow dynamics** and resembles the flow of species very well.

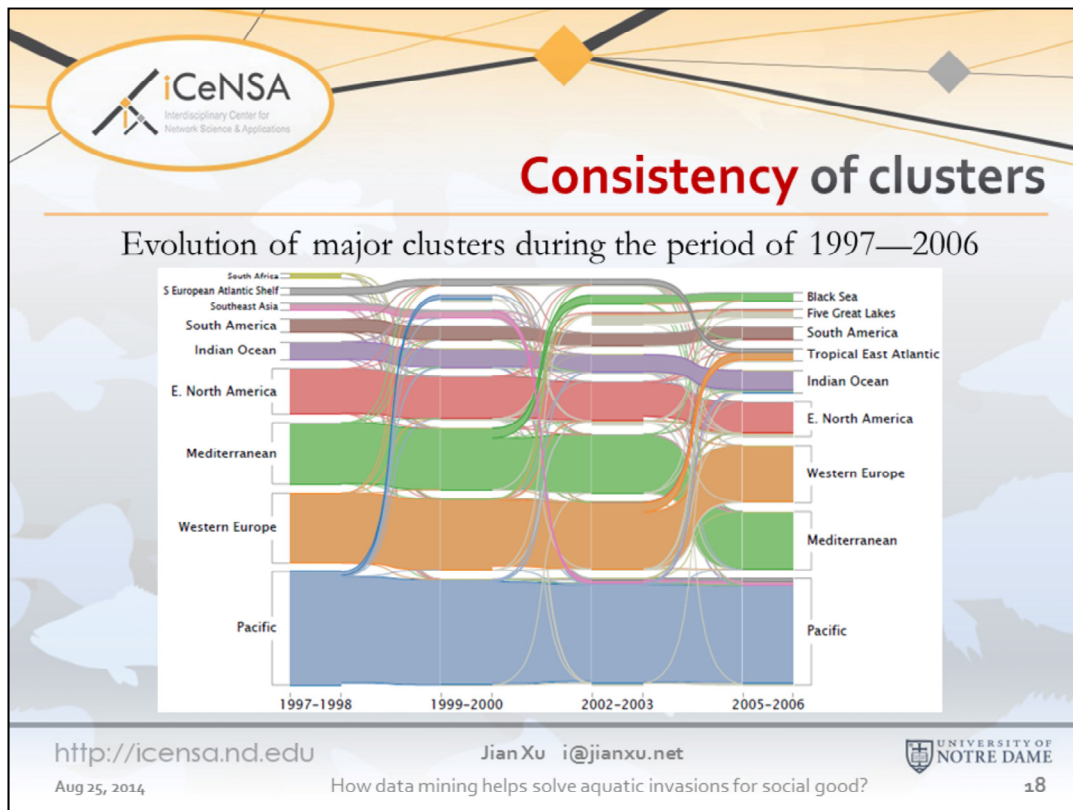


Six **major** clusters shown in colors correspond to a significant proportion of total species flow among ports.

We notice that **San Francisco Bay**, which is the blue point here, belongs to the Pacific Cluster,

which means San Francisco has lots of species exchange with other ports in the Pacific Cluster,

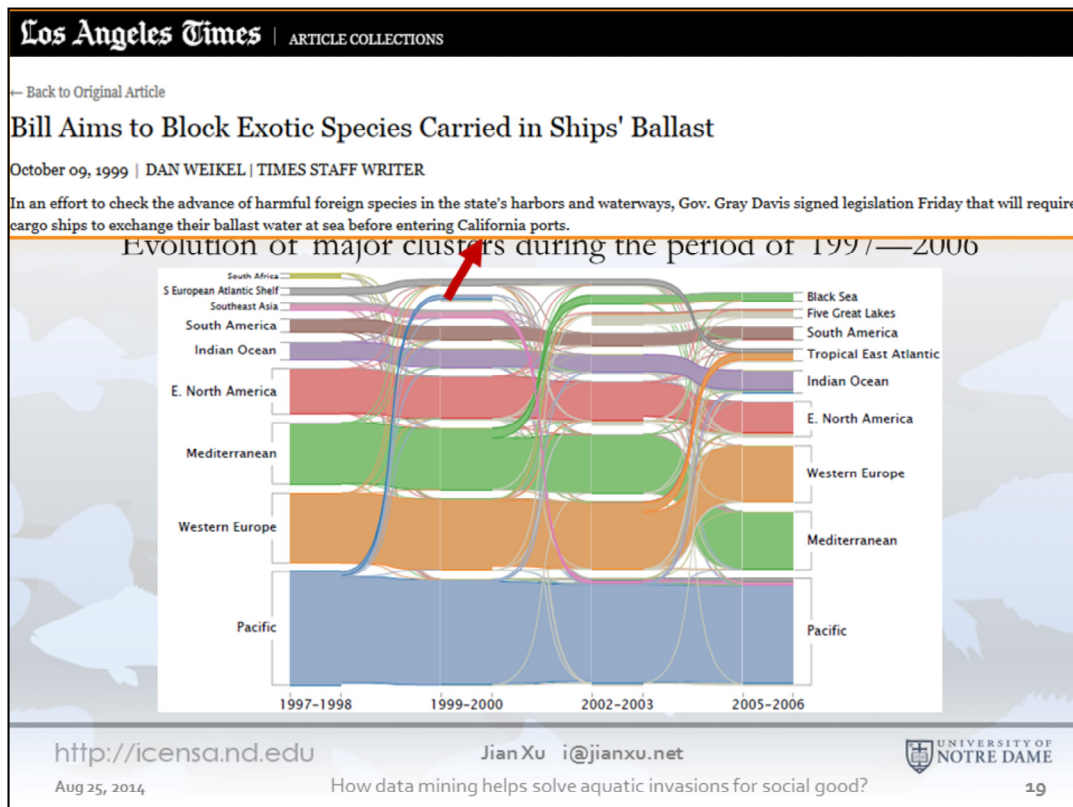
And that explains why lots of Asian species can be found in San Francisco Bay.



Moreover, these major clusters **continue** to exist over the duration studied;

for a given cluster, while some ports leave and join over time, the **vast majority** of the ports continue to remain in the same cluster.

This provides a solid foundation for devising management strategies **targeting** clusters and inter-cluster connections.



In addition, the evolution of clusters can reveal important information on how changes in vessel movement and ballast discharge patterns will affect species flow dynamics.

For instance, a **new** smaller cluster is formed in 1999--2000 by 21 ports in California that previously belonged to the Pacific cluster in 1997--1998.

This is the result of a new regulation of ballast water management in California in 1999, and is **captured** in our chart immediately.

**iCeNSA**  
Interdisciplinary Center for  
Network Science & Applications

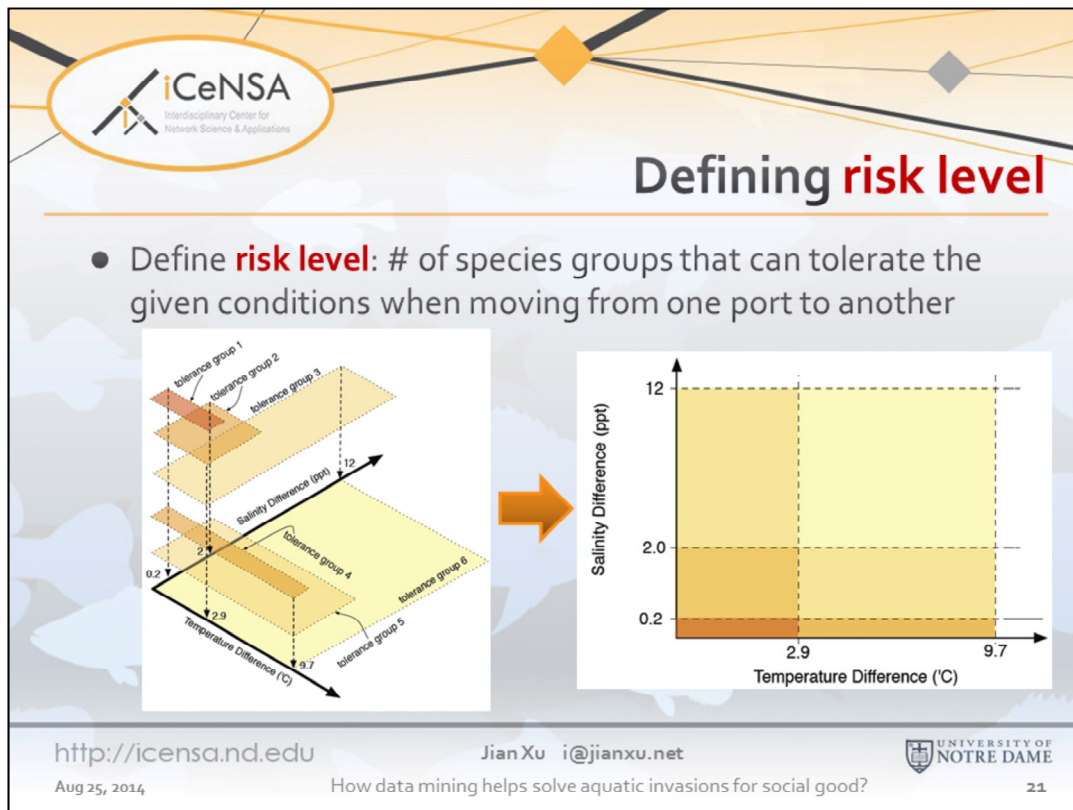
## Invasion **risk** modeling

- Non-indigenous species invasion
  - Be **translocated** from one place to another
    - *Discussed in species flow network*
  - Be **non-indigenous**
    - *Movements between non-contiguous ecoregions*
  - Be able to **survive** and establish in its new environment
    - *Can tolerate environmental differences (temperature, salinity, etc.)*

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME  
Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 20

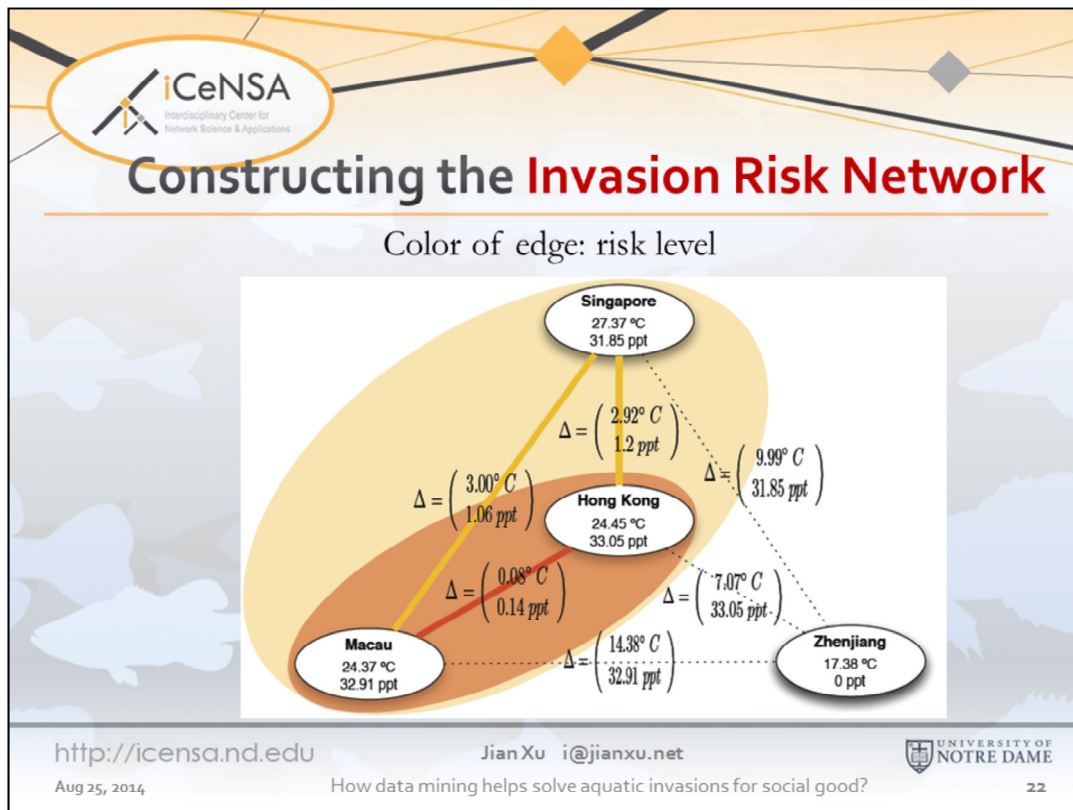
Here, we shift our attention from inter-cluster species flow to intra-cluster non-indigenous species invasion risk,

For an exchange of species to become an invasion,  
the introduced species must move between **non-contiguous ecoregions**,  
And can tolerate **environmental differences**.



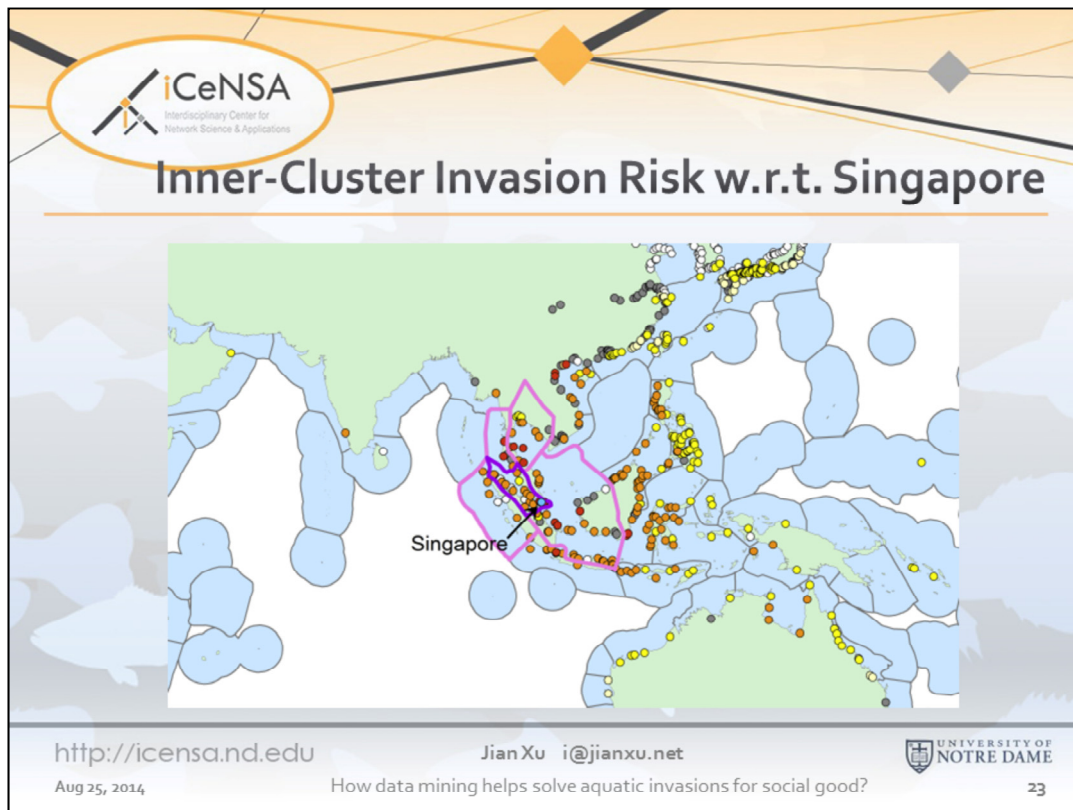
Invasion risk can be estimated by the number of species groups that can tolerate the given conditions when moving from one port to another.

Risk between any two ports is then quantified as an index created by overlapping the species tolerance groups, ranging from 0 to 6.



An Invasion Risk Network is built for every major cluster in Species Flow Network.

Clustering can help detect groups of ports that have similar environmental conditions.



Let's look at the inner-cluster invasion risk with regard to Singapore.

If Singapore is at risk, then the ports in the same cluster, the ones with colors outside the purple ecoregions will be at risk, where the colors show the risk level of ports.



## Managing **inter-cluster** exchanges

- **Weak** connections between different clusters
- If inter-cluster pathways are controlled
  - Species exchange can be restricted to ports **within** clusters
  - Species flow will be combinatorially reduced
- Inter-cluster connections are relatively **constant**
  - Good for devising long-term management strategies

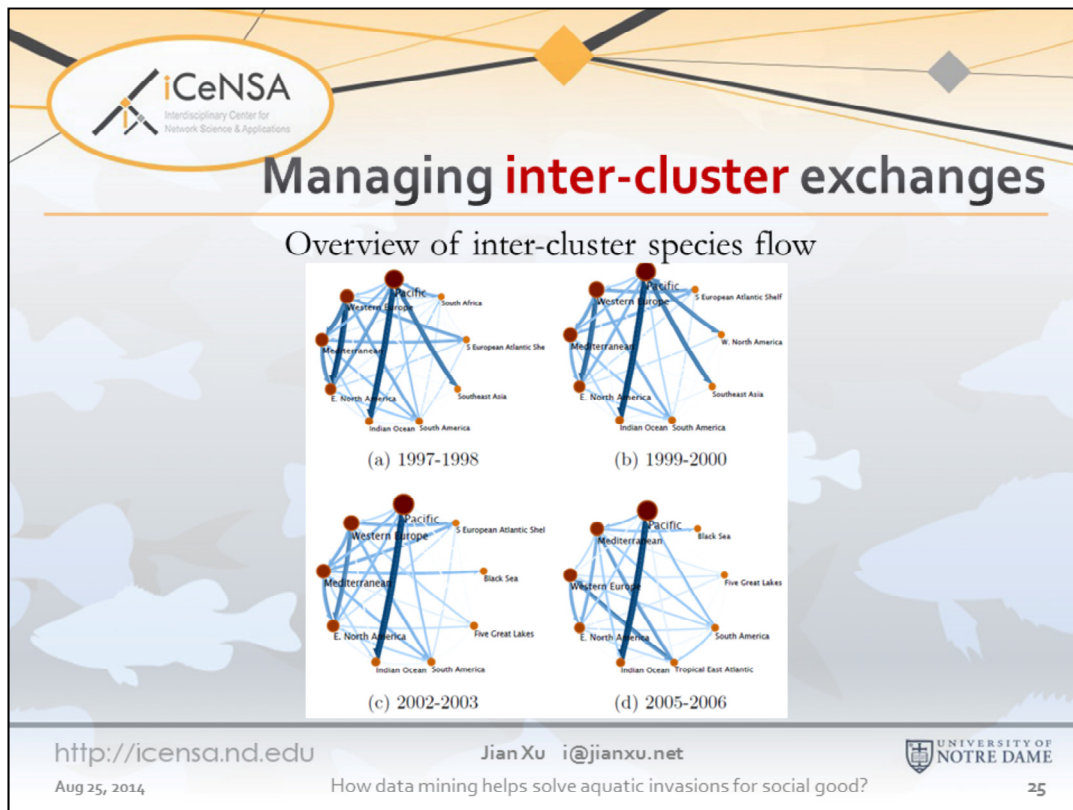
<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 24

Let's come to the **management** part.

By the nature of clustering, the connections between different clusters are weak.

if inter-cluster pathways are controlled, species flow would be combinatorially reduced.



The inter-cluster species flow being relatively **constant** is ideal for devising long-term management strategies.



## Case study: **Pacific Cluster**

- 818 ports
- **37,596** inter-cluster connections
- Singapore alone is responsible for
  - **26%** of all inter-cluster flow
  - **Highest 10** inter-cluster flows are all associated with Singapore
  - Needs our first attention!

<http://icensa.nd.edu> Jian Xu [i@jianxu.net](mailto:i@jianxu.net) UNIVERSITY OF NOTRE DAME

Aug 25, 2014 How data mining helps solve aquatic invasions for social good? 26

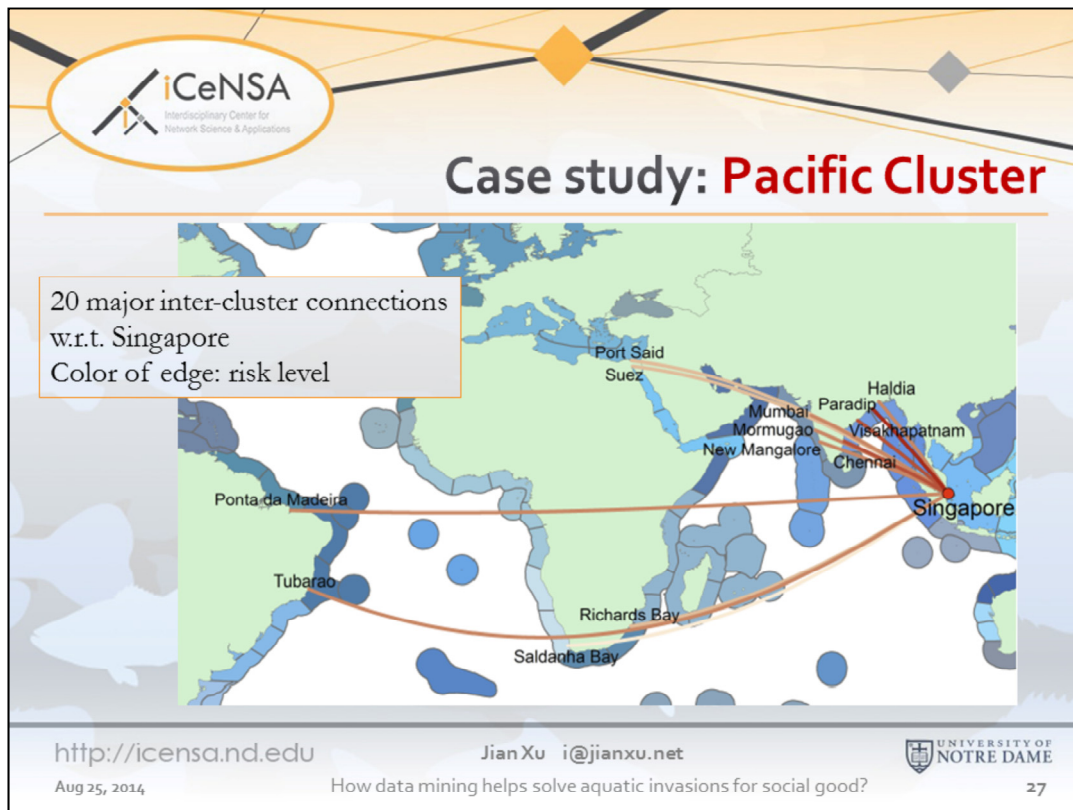
Let's do a case study on Pacific Cluster.

It has 818 ports and 37 thousand inter-cluster connections.

However, Singapore alone is responsible for 26% of all inter-cluster flow,

And the highest 10 inter-cluster flows are **all** associated with Singapore.

Thus, Singapore needs our **first attention**.



The major inter-cluster connections with regard to Singapore is shown here, and the colors correspond to risk levels.



## Ports need our first **attention**

Major inter-cluster contributors in 2005—2006:

| Cluster       | Port          |
|---------------|---------------|
| Pacific       | Singapore     |
| Mediterranean | Gibraltar     |
| W. Euro       | Rotterdam     |
| E. N. America | New York      |
| Indian Ocean  | Mormugao      |
| S. America    | Tubarao       |
| Great Lakes   | Seven Islands |
| Black Sea     | Istanbul      |
| W. N. America | Long Beach    |

Practices that reduce species movements through these ports would:

- Potentially **reduce** a large portion of inter-cluster species flow
- Strengthen geographic **segregation** of species
- Provide a **baseline** for measuring effectiveness of risk reduction efforts

<http://icensa.nd.edu>

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)

 UNIVERSITY OF NOTRE DAME

Aug 25, 2014

How data mining helps solve aquatic invasions for social good?

28

This table lists ports corresponding to the highest inter-cluster flow in big clusters.

Any practice that reduce species movements through these ports would potentially reduce a large proportion of inter-cluster species flow.



## Vessel type based management

- Frequent **inter-cluster** travelers
  - **Container carriers** contribute to 24% of all inter-cluster trips in 2005—2006
    - *Though they are not the most active vessel type*
  - Others include bulkers, crude oil tankers, refrigerated general cargo ships, combined bulk and oil carriers
- Frequent **intra-cluster** travelers
  - **Passenger carriers** stay within clusters for 97.6% of their trips
  - Barge ships stay for 98% of their trips




Photos from Wikipedia

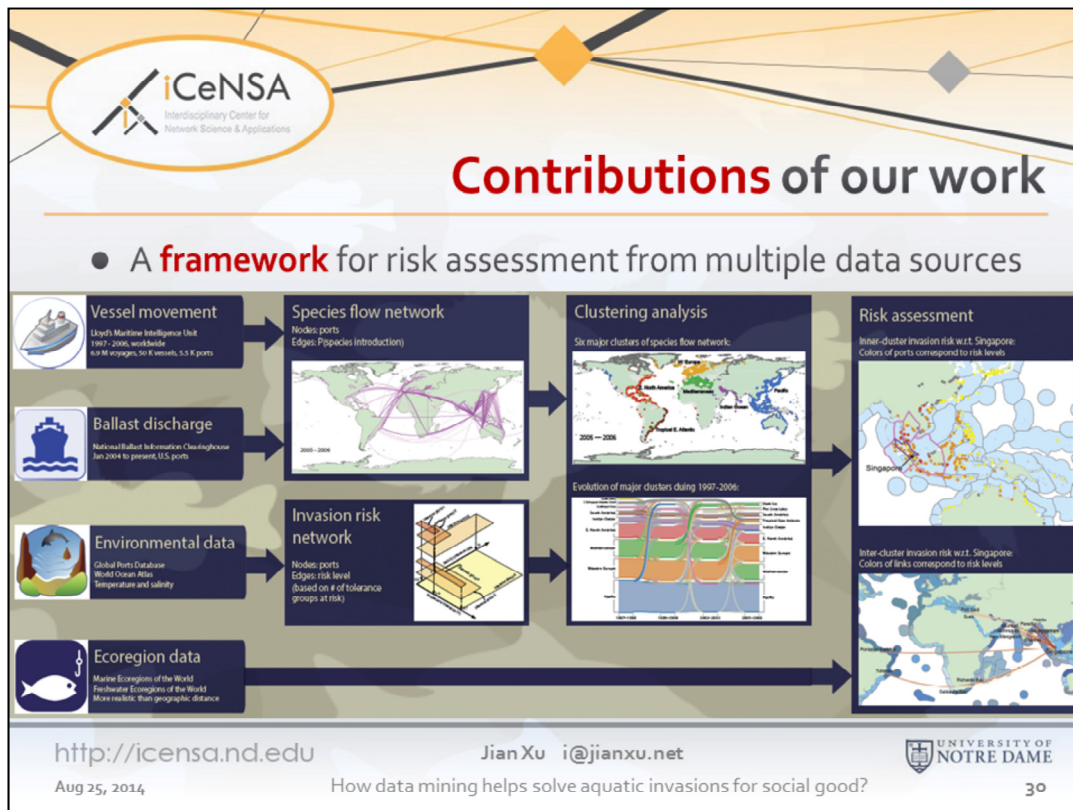
<http://icensa.nd.edu>
Jian Xu [i@jianxu.net](mailto:i@jianxu.net)
UNIVERSITY OF NOTRE DAME

Aug 25, 2014
How data mining helps solve aquatic invasions for social good?
29

Different types of vessels have different typical behaviors.

Some types of vessels, such as container carriers, although they are not the most active type of vessel, they contribute to 24% of all inter-cluster trips, which definitely need our attention.

Knowledge like this is also valuable for developing effective management strategies.

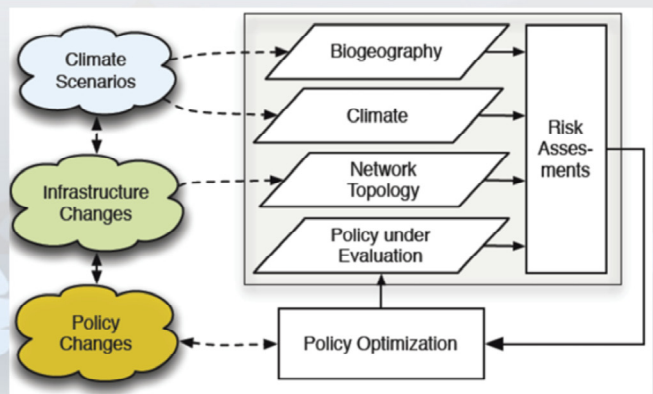


In summary, we have built a framework for risk assessment from multiple data sources,  
have crucial observations for species flow patterns,  
And there are already some take-aways for governments and organizations.



# Deployment of framework

Use of discovered knowledge in a potential deployed setting for invasion risk assessment with respect to changing climate, policy and infrastructure:



<http://icensa.nd.edu>  
Aug 25, 2014

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)  
How data mining helps solve aquatic invasions for social good?

  
31

In a potentially deployment scenario, the risk assessment part can form a feed-back loop with **changing** climate, policy and infrastructure.



## Future work

- Modeling numerous aspects of invasive species
  - Elaborate methods to generate species distributions
  - Mathematical modeling of invasions
  - Evaluation of the impact of global climate change on invasions
  - Received NSF funding for 2015—2019
- Implement a Management Transition Board
  - Our research findings will be effectively communicated with the authorities
  - And their feedback will be utilized to drive our research to be both effective and useful to the society

<http://icensa.nd.edu>

Jian Xu [i@jianxu.net](mailto:i@jianxu.net)

 UNIVERSITY OF NOTRE DAME

Aug 25, 2014

How data mining helps solve aquatic invasions for social good?

32

We are preparing to launch a detailed project looking at **numerous aspects** of invasive species.

We are also going to implement a **Management Transition Board**, where our research findings will be effectively **communicated** with the authorities

and their **feedback** will be utilized to drive our research to be both effective and useful to the **society**.



## Acknowledgement: **co-authors**

- Thanuka L. Wickramaratne *Notre Dame, iCeNSA, ECI*
- Nitesh V. Chawla *Notre Dame, iCeNSA, ECI*
- Erin K. Grey *Notre Dame, ECI*
- Karsten Steinhaeuser *University of Minnesota*
- Reuben P. Keller *Loyola University Chicago*
- John M. Drake *University of Georgia*
- David M. Lodge *Notre Dame, ECI*

<http://icensa.nd.edu>      Jian Xu [i@jianxu.net](mailto:i@jianxu.net)      

Aug 25, 2014      How data mining helps solve aquatic invasions for social good?      33

Thanks again for all co-authors



## Acknowledgement: **grants**

  
UNIVERSITY OF NOTRE DAME



  
UNITED STATES AIR FORCE

  
UNITED STATES AGENCY  
ENVIRONMENTAL PROTECTION AGENCY



 **NOAA**

  
Great Lakes  
RESTORATION

<http://icensa.nd.edu>      Jian Xu   [i@jianxu.net](mailto:i@jianxu.net)      

Aug 25, 2014      How data mining helps solve aquatic invasions for social good?      34

Thanks for organizations sponsoring our research



Thank you!

For latest research progress please visit:



**Jian Xu**  
i@jianxu.net

<http://icensa.nd.edu>



The slide features a background with a light blue and white fish pattern. At the top, there are orange and black geometric shapes. The iCeNSA logo is prominently displayed in the top left and middle right. The text 'Thank you!' is centered in a large, dark font. Below it, the text 'For latest research progress please visit:' is followed by the iCeNSA logo. The name 'Jian Xu' and email 'i@jianxu.net' are centered below the text. At the bottom left is the URL 'http://icensa.nd.edu' and at the bottom right is the University of Notre Dame logo.

And thank **you** for listening!